Democratic and Popular Republic of Algeria
Ministry of Higher Education and Scientific Research

University of Abderrahmane Mira Bejaia
Faculty of Natural Sciences
Computer science department

Graduation thesis

To obtain an academic master degree in Advanced Information Systems

Sentiment analysis for health crisis management in smart cities

Presented by Misses :

ITMACENE Ouardia.
OUBEKKOU Milissa.

Evaluated By:

Professor AMROUN Kamal: **President**
Dr CHIBANI Samia Wife SADOUKI : **Examiner**
Dr EL BOUHISSI Houda Wife BRAHAMI : **Superviser**
Mr ZIANE Amine (Doctorand) : **Guest**

Promotion 2022/2023

Thanks

And there we go!, The best things have an ending.
First of all, we thank Almighty God for giving us the will and perseverance to carry out this work.
We would like to take this opportunity to express our deep gratitude and appreciation to our funder.
Mrs. EL BOUHISSI Houda, for her valuable ad vices and guidance throughout our research. It would never have seen the light of day without her encouragement, patience and even humor.
We also sincerely thank the jury members for the interest shown in our research, who agreed to review our work and enrich it with their suggestions.
We would like to express our sincere thanks to those who helped us and contributed to the preparation of this dissertation and to the success of this wonderful academic year. We thank all teachers for their commitment, patience and contributions to our education.
Finally, thanks to all the family and friends who have supported us throughout this thesis .

Dedications

To my dear mother,

No dedication can express my respect, my eternal love and my consideration for the sacrifices you have made for my education and my well being.I appreciate all the support and love you have given me since my childhood.may this modest work be the fulfillment of your long-formulated wishes, the fruit of your innumerable sacrifices; may god grants you health and happiness.

In memory of my father

This work is dedicated to my father, who i loosed too early, who always pushed and motivated me in my studies,i hope that, from the world that is yours now, you appreciate this humble gesture as a sign of gratitude from a girl who loves you .

To my dear sister, Thilelli and my dear brother, Amazigh

To my dear binom ouardia,

To my dear grandmother, To my dear maternal aunties and their husbands especially yasmina, To my dear cousins especially feriel, kamilia and thiziri, To my dear uncle ali, hakim and djamel who helped me a lot, To my best friends, who i miss cylia and menad so much, To my dear mounir, who has always been there for me To my dear friends, ouahiba, dihia and leticia, To my dear zaina and her family, To all those who helped me or were there for me.

Millisa

Dedications

I humbly dedicate this modest work to my dear parents, whose sacrifices and encouragement have been an inexhaustible source of motivation.May God protect them

And my dear grandparents, especially Djedi Makhlouf for his encouragement and support throughout my educational career.

Would also like to express my sincere thanks and deep gratitude to my lovely sisters, Amel and Loundja, as well as to my brother Makhlouf and the entire Merzouk family. Your love and support have been essential pillars in the realization of this project.

Would also like to express my gratitude to my best friends, Kamilia, Kenza and Lyna, and to my partner Melissa, for their constant presence and support. Your precious friendship has been a source of comfort throughout this journey.

Ouardia

Résumé

Le machine learning devient nécessaire, ce dernier consiste à créer des systèmes qui apprennent ou améliorent les performances en fonction des données qu'ils traitent, c'est un outil d'aide à la décision grâce à son pouvoir de prédiction .

Dans notre projet, nous nous concentrerons sur l'analyse des sentiments dans les réseaux sociaux, plus précisément sur la plate-forme Twitter, dans le contexte de la pandémie de coronavirus. Notre objectif principal sera de déterminer la tonalité émotionnelle des discours des utilisateurs en classifiant leurs messages dans trois catégories principales : positif, neutre et négatif .

Nous utiliserons des techniques d'apprentissage automatique et de traitement du langage naturel pour classifier les tweets nous combinerons les techniques d'apprent-issage automatique , telles que le modèle Long Short-Term Memory (LSTM) avec l'algorithme Elephant Herding Optimization (EHO),

Mots clés : analyse de sentiments ,classification, coronavirus, opinions, pandémie, réseaux sociaux,twitter,EHO,LSTM,NLP.

Abstract

Machine learning becomes necessary. consists in creating systems that learn or improve performance according to the data they process. It is a decision-making tool thanks to its predictive power.

In our project, we will be focusing on the analysis of of sentiment in social networks, more specifically on the Twitter platform, in the context of the coronavirus pandemic. Our main objective will be to determine the emotional tone of users' discourse by classifying their messages into three main categories: positive, neutral and negative .

We will use machine learning and natural language processing techniques to classify tweets. We will combine Long Short-Term Memory (LSTM) model with Elephant Herding Optimization (EHO) algorithm, .

keywords: sentiment analysis ,classification, coronavirus,opinions, pandemic, social networks, twitter ,EHO,LSTM,NLP

Table of Contents

List of Figures

List of Tables

List of Algorithms

Abreviations list

BERT	Bidirectional Encoder Representations from Transformers
CSV	Comma-separated values
DP	Deep Learning
EHO	elephant herding optimization
IA	artificial intelligence
JCDL	Joint Conference on Digital Libraries
KNN	k-Nearest Neighbors
LR	Linear Regression
LSTM	Long-Short Term Memory
ML	Machine learning
NB	Naive Bayes
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
RNN	Recurrent Neural Network
SATD	administrative seizure to third party holder
SVM	Support Vector Machine ou Machine à vecteurs de support
$TF-IDF$	Term Frequency Inverse Document Frequency
XGB	EXtreme Gradient Boosting

Chapter 1

General Introduction

1.1 Introduction

Nowadays, Artificial Intelligence (AI) has become increasingly prevalent in our daily lives, with a wide range of applications, such as personal assistants like Siri and Alexa, industrial robots, safety monitoring systems, and healthcare applications. Research on AI is also ongoing, with ambitious projects aimed at creating machines capable of understanding natural language, reasoning logically, and simulating human consciousness.

Sentiment analysis is a branch of AI,which is a data mining method that uses natural language processing (NLP) techniques to extract information about people's opinions, feelings and emotions from texts such as social media messages, blogs and online comments.

Sentiment analysis is the examination of people's opinions, feelings, evaluations, appreciations, attitudes, emotions and personal preferences towards entities such as products, services, organizations, individuals, issues, events, subjects and their attributes. It

aims to get users feelings about specific events or topics, which makes it possible to complete the textual analysis of feelings. Sentiment analysis is a powerful tool for understanding people's perceptions and reactions to different elements, which can help businesses, organizations and decision makers make informed decisions. Using advanced techniques and algorithms, sentiment analysis allows valuable information to be extracted from large amounts of textual data, making it easier to understand trends, dominant opinions and overall feelings of a given audience. This approach is widely used in areas such as marketing, online reputation management, business intelligence, crisis management, and many more. Sentiment analysis provides a valuable perspective on how people perceive and respond to different aspects of our world, allowing for more informed decision-making and a better understanding of user needs and expectations.

With the advent of COVID-19, sentiment analysis has become even more important in smart cities. Lockdowns and radical changes in our way of life have had a significant impact on the mental health and well-being of citizens. Smart cities can use sentiment analysis to understand how people perceive the situation and how they adapt to the changes taking place. Data from the sentiment analysis can help authorities make informed decisions on how to support communities most affected by the pandemic.

1.2 Problematic

Sentiment analysis can be applied to gain a comprehensive understanding of Twitter users' opinions and emotions regarding COVID-19. By analyzing the content of tweets, it becomes possible to distinguish between negative, positive, and neutral feelings.

To perform sentiment analysis, natural language processing techniques can be employed to extract key features from tweets, such as words, phrases, or emojis, which indicate sentiment. Machine learning algorithms can then be trained using labeled data to classify tweets into negative, positive, or neutral categories.

By examining the sentiment of tweets, it becomes feasible to identify the prevailing public sentiment towards COVID-19. This information can be valuable for various purposes, including monitoring public opinion, tracking the effectiveness of public health measures, identifying areas of concern, and evaluating the impact of communication strategies.

1.3 Objectives and contributions

The aim of this thesis is to detect the sentiments of Twitter users and their positive, negative or neutral opinions on covid-19 in smart cities. To achieve our goal, we use different techniques and approaches.

Towards evaluating citizen sentiments for smart city services,

our research findings make the following contributions.

• We reviewed the most important works related to sentiment analysis in smart cities.

• We developed an effective deep convolutional network architecture for tweet sentiment analysis.

• We employ Herding Elephant Optimization (EHO) algorithm to choose the best features for deep learning algorithms

1.4 Methodology

In particular, our work is based on the following steps:

1.Present Research and analysis step: a state of the art of the different technologies and methods proposed in the framework of the analysis of feelings and comparison of the advantages and disadvantages of each approach proposed in each paper.

2. Solution proposal step: propose an efficient solution to the problem.

3. Implementation and experimentation step: how the proposed system works.

Chapter 2

Sentiment analysis and smart cities

2.1 Introduction

The development of information and communication technologies has led to the emergence of smart cities. These cities integrate advanced digital systems to improve residents' quality of life and optimize resource management. One of the key areas of the smart city concept is sentiment analysis, a technique aimed at understanding and measuring people's emotions and opinions on a large scale. Sentiment analysis plays an essential role in the context of smart cities, enabling authorities to understand citizens' opinions and feelings.

This contributes to a better quality of life, informed decision-making and more participative governance.

2.2 Sentiment analysis

2.2.1 Definition

In Merriam-Webster's Collegiate Dictionary, a feeling is defined as an attitude, thought or judgment encouraged by a sensation.

Sentiment analysis: also referred to as opinion mining extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, appraisal extraction, is an natural language processing approach (NLP) that identifies the emotional tone behind a body of text. This is a popular way for organizations to determine and categorize opinions about a product, service or idea, at the beginning of the 2000s. The origins of sentiment analysis refer to the sciences of psychology, sociology and anthropology, which focus on human emotions[1][2].

2.2.2 Sentiment Analysis Types

There are several types of sentiment analysis so we will cover the most important ones[1][2].

•Detailed sentiment analysis:

Instead only talking about the negativity or positivity of the sentence, a good example is Google's 5-star rating system. It's nice when people take the time to write a basis for their star ratings, but if you only have stars to analyze, you can read them like this:

5 tars = very positive.

4 stars = positive.

3 stars = Neutral.

2 stars = negative.

1 star = very negative.

Some systems have also given different polarity classifications by identifying whether the positive or negative feeling is associated with a particular feeling, such as anger, sadness, or worries (feelings) or happiness, love, or excitement (Positive feelings).

•Emotion detection :

Most of the emotion detection systems are based on the use of sentiment lexicons or complex machine learning algorithms, this type of sentiment analysis helps identify the emotions that customers express in their comments, ranging from joy and satisfaction to anger and frustration etc.

For example, sites like The Athletic allow readers to comment on articles, but also offer a simpler ”what do you think of this story” feedback option.

•Aspect-Based Sentiment Analysis:

In this type, the results are more detailed, interesting and precise because the aspect-based analysis examines in detail the information contained in a text.

For example, a customer may often visit a company’s website to initiate a service call via a CHATBOT. To determine where they should be directed, the BOT will ask ”How can I help you?” Customers will enter things like ”The picture on my TV is too dark” or ”I need to cancel my streaming subscription because there are too many errors”, such content is an obvious problem with the product or service and should be flagged as negative and corrected to improve the user experience.

2.2.3 Sentiment analysis challenges

Sentiment analysis, also known as opinion mining, involves analyzing text data to determine the sentiment or subjective information expressed within it. While sentiment analysis offers valuable insights into public opinion, it also presents several challenges that need to be addressed. Some of the key challenges in sentiment analysis include:[2].

•Context and Polarity:

Algorithms struggle to understand context. If humans can understand the context of an interaction, this can be an obstacle for an algorithm. Therefore, the algorithm will need to be configured to include a context component for messagesFor example:

If the question is ”What did you like?” the first answer will be positive and the second negative, But if the question is “What didn’t you like?” the meaning of the two answers changes completely. Pre-processing or post-processing will therefore be important so that the machine understands the context that may have caused certain responses. Nevertheless, it remains a difficult task.

•Determine subjectivity and tone:

Tone analysis can be simple or complex, depending on the words used. Human interactions can be implicit or explicit, and subjective or objective, which is difficult for algorithms to judge. If we look at these two examples:

Ex 1: The wallpaper is beautiful.

Ex 2: The wallpaper is white.

We can estimate that the feeling is positive for the first sen-

tence and neutral for the second. All predicates (adjectives, verbs, nouns, etc.) should not be treated in the same way when analyzing the sentiment in a sentence. Here, the term “beautiful” is much more subjective than the term “white”.

•Identify sarcasm and irony:

People express their negative feelings using positive words, which can be difficult for machines to detect without having a thorough understanding of the context in which a feeling was expressed. For example, if we take the answer to the question:

“Did you enjoy your experience on our site?”.

”Yes of course! There are no bugs!”

Here, at first glance, it would seem that the answer is yes. However, we could very well see irony in it and understand the opposite. The problem is that we have no textual clue that can help the machine learn or, at least, question the real feeling behind this sentence.

•Neutral posts:

Another issue is neutral posts, which are not categorized. How does the algorithm handle neutral messages?.

2.2.4 Sentiment analysis application

The importance of sentiment analysis exists in many fields, and a number of applications have emerged in this context. Let’s briefly mention few applications[1]:

Policy:

Before a new law is made, politicians try to get the opinion of social media users on the law.

Economy:

The customer asks for the opinion of other people who are using the product before buying it, companies can know the opinion of customers on their products or services to make changes.

Education:

Sentiment analysis helps teachers and schools take corrective action.

2.3 Smart cities

2.3.1 Definition

In this part we will discuss the concept of smart cities because it tends to be the model that is the most common and least understood by no specialist audiences, for this becoming a smart city means bringing together all available technologies and resources develop city centers in smart and coordinated manner once integrated, livable and sustainable.

According to ICLEI, a smart city is one that willing to make a difference challenging conditions for a healthy and happy community that may affect global, environmental, economic and social trends bring[3].

The concept of smart cities dates back to the 1960s and 1970s, when the Community Analysis Bureau began using computer databases, cluster analysis and infrared aerial photography to collect data,

generate reports and direct resources to areas where they needed to fight the most to avoid potential disasters and poverty reduction. Since then, three different generations of smart cities have emerged[4].

Smart Cities 1.0 are spearheaded by technology providers. This generation is focused on the implementation of technology in cities, even though communities do not fully understand the potential impact of technology or the impact it may have on everyday life[4].

Smart City 2.0, on the other hand is led by cities. In the second generation, forward thinking leaders within the community help define the future of the city and how it can be shaped using smart technology and other innovations[4].

Cities around the world are at various stages of developing and implementing smart technologies. However, there are a few that are ahead of the curve and leading the way to creating fully smart cities. These include[4]:

- Barcelona, Spain.
- Columbus, Ohio, USA.
- Dubai, United Arab Emirates.
- Hong Kong, China.
- Kansas City, Missouri, USA.
- London, England.
- Melbourne, Australia.
- New York City, New York, USA.
- Reykjavik, Iceland.
- San Diego, California, USA.
- Singapore.

• Tokyo, Japan.
• Toronto, Canada.
• Vienna, Austria.

2.3.2 Advantages

A smart city is a sign of development and entrepreneurship, which in turn represents a major advancement for the world and the country in which it is located. Their main advantages include [5]:

- Effective decision-making based on data.
- Create a safe community.
- Improve urban transportation.
- Improve the environment through various systems.
- Optimize time in hospitals and public services.
- Evolution of the Internet of Things (IoT).
- Realize new business opportunities.
- Create services that better meet the needs of citizens.
- Lower economic and natural input costs .

2.3.3 Disadvantages

However, despite the many benefits they offer, smart cities also have some disadvantages which are[5]:

- Significant capital investment is required for the technology.
- Dependence on technical service companies.
- Real estate gets more expensive because it’s harder to build and execute.

- A wider technology gap is opening between smart cities and other cities.

- E-waste has increased significantly .

2.3.4 Smart City Challenges

Smart cities have many advantages, but there are also challenges that need to be overcome. These include government officials, allowing broad citizen participation. The private and public sectors also need to coordinate with residents so that everyone can make a positive contribution to the community.

Smart city projects must be transparent and accessible to citizens through open data portals or mobile applications. This allows residents to engage with the data and perform personal tasks such as paying bills, finding efficient transportation, and assessing energy usage in the home.

This all requires a reliable and secure data collection and storage system to prevent hacking or misuse. Smart city data also needs to be anonymized to avoid privacy concerns.

The biggest challenge may be connectivity, because of the thousands or even millions of IoT devices that need to connect and work together. This allows services to be pooled together and continuously improved as demand increases.

In addition to technology, smart cities must also consider social factors, creating a cultural fabric that is attractive to residents and provides a sense of home. This is especially important for cities that are being built from the ground up and need to attract residents

[5].

2.4 Conclusion

Sentiment analysis can play an important role in smart cities by measuring citizens' opinions and feelings about various topics such as municipal services, quality of life, security, etc. This information can be used to inform policy decisions and improvements to the city. Moreover, sentiment analysis can also be used to monitor the reputation of the city and detect problems early. Finally, by combining sentiment analysis data with other data sources, cities can improve the efficiency of their services and strengthen their engagement with citizens.

In this chapter, we have presented the fundamental concepts of sentiment analysis with a special focus on smart cities and their concepts as well, and some examples of them.

The next chapter is devoted to the state of art with details on the studied papers, methodologies and tools.

Chapter 3

State of the art

3.1 Introduction

Over the past three years, humanity has faced a series of major challenges due to the emergence of potentially fatal viruses and diseases. One of the most striking examples is the global outbreak of the disease known as Corona or COVID-19. This virus, which was initially identified in 2019, has spread rapidly around the world, with considerable health, economic and social consequences.

In this chapter, our focus is on reviewing existing works that specifically address sentiment analysis for the management of health crises. We aim to provide a comprehensive overview of the research conducted in this area and compare the different approaches based on established criteria.

We start by conducting an extensive literature search, including academic papers, conference proceedings, and relevant reports from reputable sources. Our search is guided by specific keywords related to sentiment analysis, health crises, and management strategies.

After gathering a substantial number of potential works, we perform a rigorous screening process to select the most relevant ones. We consider factors such as the publication's relevance to our research topic, the quality of the methodology employed, and the significance of the findings.

Once the final set of works is determined, we proceed to analyze them in detail. For each selected work, we provide a summary that highlights the key aspects of the approach used, including data sources and sentiment analysis techniques. Additionally, we identify the specific health crises addressed in each study and discuss the management strategies proposed.

To compare the different works, we establish specific criteria that are relevant to sentiment analysis for health crisis management. These criteria may include accuracy of sentiment classification, scalability of the approach, real-time analysis capabilities, integration with other data sources, and practical applicability in crisis situations. We assign scores or rankings to each work based on how well they meet these criteria.

3.2 Related works

In this section, we present the main research works related to sentiment analysis in the context of COVID-19:

Zunera and al.[6], are speaking about analysis of sentiments in the period of COVID-19 using Social media like Twitter (twitters) and use a data set COVIDSenti consists of 90,000 tweets which

is divided into three subsets of data. (COVIDSENTI A contains tweets regarding the measures taken by government authorities to protect people and COVIDSENTI B mainly concerns four topics COVID-19 disasters, maintaining social distance, confinement and staying at home, The COVIDSENTI C is a collection of tweets about COVID-19 cases, outbreaks and advice to stay at home), the proposed approach is divided into four phases.

1- Data preprocessing

2- keyword trend analysis

3- Feature extraction

4- Ranking methods

Then they proposed other methods based on precision and used different conventional methods: count-victories, TF-IDF, different models based on word embeddings like Word2Vec, fast Text and Glove.

Nifula and al . [7], this paper proposes an approach based on Data Acquisition (Dataset-I Dataset-II was obtained from Kaggle). They applied the analysis with machine learning and deep learning with different algorithms (logistic regression, support vector machine, decision tree, random forest, Naïve Bayes, k-nearest neighbors, and XGBoost.) at the end of this article a comparison is made between the real cases that have been declared in hospitals and that have to be published on Twitter.

chandra and al.[8], they have presented a study with new deep learning models for sentimental analysis during the rise of COVID-

19 infections, uses twitters (10,000 tweets and emotional symbols (EMOJI)) from India. They are proposed a method with LSTM and BERT language models and for the analysis done with deep learning follows a series of steps:

1. Extracting tweets.
2. Pre-processing tweets.
3. Model Development and Training Using LSTM, BD-LSTM and BERT.
4. Prediction using selected COVID-19 data for LSTM, BD-LSTM and BERT models with GLOVE integration.

Mohammed and al.[9] ,The study of sentiment analysis in this paper aims to understand public health by analyzing the sentiment and topic modeling of Indonesian public conversations on Twitter about the COVID-19. It proposes an approach based on Tweet Data Acquisition and Pre-processing, the data in this method was obtained from the Twitter streaming API.They applied text mining and machine learning with its four different algorithms (Gaussian Naïve Bayes, Multinomial Naïve Bayes, Support Vector Machine and Random Forest), the result with these algorithms are interpretable and understandable. From the discussion, we can conclude that according to Twitter data, most Indonesian citizens have a bigger concern on economic problems rather than health problems.

Chhinder and Anand [10] , this paper aims to analyze people's opinions towards corona virus pandemic from all over the world

using machine learning techniques on Twitter about the COVID-19. They have applied Bigram, Unigram, Bow and Tfidf features for effective feature extraction and machine learning Techniques. Furthermore, different classifiers: Naïve Bayes, Support Vector machine, logistic regression and random forest were used to compare the different models performance. From the result we can conclude that people are vigilant and hope of reducing the effects of corona virus.

Priya and al . [11], in this paper they have applied machine learning algorithm (Naïve Bayes Classification). From the result we can conclude that people are very aware of government policies, safety measures, their symptoms and precautions to take during this time. They followed and maintained social distance and sanitizing methods very well.

Malak and al.[12], the study of sentiment analysis in this paper aims to analyze Twitter data to examine public attitudes, concerns, and thoughts about the COVID-19 pandemic. It proposes an approach based on Tweet Data Acquisition and Pre-processing, the data in this method was obtained from Twitter. They have applied a set of ML classification algorithms such as SVM, RF and XGB to classify the tweets as positive, negative, or neutral.

They examined the performance of SVM, RF, and XGBoost for feature extraction using N-grams and TF-IDF. They also analyzed the effect of using binary classification (positive or negative) and tertiary classification (positive, negative or neutral). Research

shows that XGBoost outperforms SVM and RF in analyzing sentiment as it achieves 90 accuracy in binary class datasets with unigrams and bigrams. Also, they say that binary classification performs better than multiclass classification.

Phil [13], the study of sentiment analysis aims to know the polarity of Nepali people. The data were collected from 21 May 2020 to 31 May 2020 using the Twitter API and the Tweepy library for Python. To collect 615 tweets from people whose location was determined to be Nepal they applied the TextBlob library of python sentiment analysis.

They have applied Naïve Bayes model. The results show that whole 58percentage of people published positive tweets, while only 15 percentage of the tweets were negative. Nevertheless the neutral tweets were about 27percentage. From this study, we can say that people's reactions when they post their feelings on social media change from day to day.

Alamoodi and al.[14] , the article is a chapter book of expert system with application, was published April 1, 2021 .they are giving the importance of social media platforms, because people use their applications and spend excessive hours on these media, especially in the period of epidemics and disease outbreaks.

They used three methods, the first Lexicon-based model: more than 80,000 tweets with Apache HADOOP an analysis of feelings positive and negative and neutral, and the 2nd approach is the Models based on ML (decision trees, k-nearest neighbor, support

vector machines and naive bayes, link regression) and the SVM approach with different result for each disease, The third method Hybrid models uses ML, SVM, and NB techniques. At the end COVID 19 an infectious disease remains unclear as its literature and cases proliferate massively; therefore, it is almost impossible to report updated information. Moreover, accurate information can only be obtained at the end of the pandemic.

Dharmendra and al.[15], present some Machine Learning Techniques like, Random Forest Classifier and Multinomial Naive Bayes (MNB), Logistic Regression, Support Vector Machine, Based on a Review of Related Work; they propose an approach for analyzing and predicting Twitter data with the algorithm called Sentiment Analysis of Twitter social media Data (SATD) and the three types of feelings. Neutral Sentiment, Positive Sentiment, Negative Sentiment. ,and the five ML models such as Random Forest Classifier, Multinomial NB Classifier, Logistic Regression, Support Vector Machine and Decision Tree.

In conclusion, the article demonstrates how effective machine learning is in analyzing feelings about COVID-19 social media data. This approach could be used to track changing user attitudes and opinions about the pandemic and help identify new issues in the response to COVID-19.

3.3 Comparative table

In this section, we compared the proposed approaches, which relate to the analysis of sentiments in health crises in a table. This table shows six columns.

- Column 1 “Approach “: presents the names of the writers.
- Column 2: present the dataset uses .
- Column 3 “Accuracy”: the result that is obtained from each of the approaches.
- Column4”Technique”: Present the machine learning techniques and algorithms used.
- Column 5: Inconvenient of the technique used .
- Column 6: Advantage of the technique used.

Approach	Dataset	Accuracy	Technique	Inconvenient	Advantage
Zunera and all[6]	-COVID-Sent -COVID-SENTI A -COVID-SENTI B	**Gain(Glove)** COVIDSenti=0.97 % **Gain(XGB):** COVIDSenti=4.41 % **Gain(multi-depth):** COVIDSenti=2.58 %	KNN LR Ensemble	1-Sentiment analysis in a single social network. 2-lack of precision.	1-The technique used simple and easy. 2-Does not require a lot of data or model
Zunera and all[6]	COVIDSent COVID-SENTI A COVID-SENTI B COVID-SENTI C	XGB(vectoriseur): COVIDSenti=89.81 % XGB(TF-IDF):precision COVIDSenti= 88.46 % XGB(word2vec): COVIDSenti= 97.17 %	KNN LR Ensemble	1- Sentiment analysis in a single social network (twitter).	1-Add the precision. 2- The technique used is simple and easy.
Nifula and all [7]	DATASET-I DATASET-II	ML=The accuracy is more than 90 % in Random Forest with algorithms bag-of-words and TF-IDF DL= The accuracy is more than 99 % in word2vec with algorithms CNN and LSTM	-L.R - -NB -k-NN - Decision Terr -RF -XG word-2vec	1-Sentiment analysis in a single social network. 2-very difficult to choose the structure. 3- The analysis is done just in some state of USA.	1-easy and efficient training of different models thanks to already labeled data. 2-efficiently captures the semantic and arithmetic properties of a word.
chandra end all [8]	Senwave	60 % singular feeling 5 % of tweets have two sentiments attached to them 14 % have no feelings attached to them	word2-vec	1- the analysis is done just in some state of India. 2- very far from perfect.	1-The dimensionality of the vector space to be constructed. 2- quick to train and run.
Mohammed and all [9]	Text documents Twitter streaming API	Support Vector Machine = the accuracy is 81 %. Naive Bayes = the accuracy is 74 % Random Forest= the accuracy is 68 % K-NN =the accuracy is 66 %	-K-NN. - SVM. - Naive Bayes. - RF.	1- SVM algorithm is not suitable for large data sets. 2-SVM does not perform very well when the data set has more noise.	1-SVM is more effective in high dimensional spaces. 2-SVM is relatively memory efficient.

Table 3.1: State of the art of related works

Approach	Dataset	Accuracy	Technique	Inconvenient	Advantage
Chhinder and Anand [10]	NLTK library Tweeter's data set Tweeter Scraper	-Support vector machine=the accuracy is 94.16 %. -Logistic regression= the accuracy is 91.52 %. -Random forest classifier= the accuracy is 90.13 %. -Naive bayes= the accuracy is75.99 % .	SVM. LR. RFC. NB.	1-In cases where the number of features for each data point exceeds the number of training data samples, the SMV will underperform.	1-SVM works relatively well when there is a clear margin of separation between classes. 2-SVM is effective in cases where the number of dimensions is greater than the number of samples.
Priya and all [11]	SS-Tweet data set. Labeled data set. Tweeter data.	-Naïve Bayes= the accuracy is 70-The percentage of positive tweets is 30 %. -The percentage of negative tweets is 16 %. -The percentage of neutral tweets is 56 %.	Naïve Bayes	1-Its estimations can be wrong in some cases, so you shouldn't take its probability outputs very seriously.	1-It works quickly and can save a lot of time. 2-It is suitable for solving multi-class prediction problems.
Malak and all[12]	Twitter dataset	•SVM +(-unigram 87 %. -bigram 88 %. -unigram +TF-IDF 88 %. -bigram +TF-IDF 86 %. •RF + (-unigram 78 %. -bigram 71 %. -unigram +TF-IDF 78 %. -bigram +TF-IDF 72 %. •XGB + (-unigram 90 %. -bigram 90 %. -unigram +TF-IDF 88 %. -bigram +TF-IDF 88 %.	SVM RF XGB	1-Long training time for large data-sets. 2-A large number of trees can make the algorithm too slow for real-time predictions. 3-It does not perform so well on spare and unstructured data.	1-It has excellent generalization capability with high prediction accuracy. 2-It automates missing values present in the data. 3-It is highly flexible.
Phil [13]	the Twitter API and the Tweepy library for Python	-58 %of people published positive tweets. -15 % of the tweets were negative. -27 % of the tweets were neutral.	- Naïve Bayes model	1- It assumes that all predictors are independent, rarely happening in real life. 2- Its estimations. 3- can be wrong in some cases.	1- It doesn't require as much training data. 2- It handles both continuous and discrete data.

Table 3.2: State of the art of related work (continued)

Approach	Dataset	Accuracy	Technique	Inconvenient	Advantage
Alamoodi and all [14]	Dataset1(a) Dataset2(b)	**Dataset1(a)** : Precision Neutral Class Sentiment : NB = 0.93 % Positive Class Sentiment : Random Foerst= 0.8 %. Negative Class Sentiment : NB = 1 % **Dataset2(b)** : Precision Neutral Class Sentiment : Decision Tree=0.7 % Positive Class Sentiment : Random Foerst= 0.85 % Negative Class Sentiment : Random Foerst=0.85 %	RF, NB, LR, Support Vector, Decision Tree	1- Cannot be used to solve nonlinear problems. 2- Takes more time.	1- Quickly process large quantities.
Dharmendra and all [15]	Twint	The precision : Random Forest Classifier=0.97 % Multinomial NB Classifier= 0.98 % Logistic Regressio=0.96 % Support Vector Machine =0.97 % Decision Tree=0.98 %	SATD RF, NB, LR, Decision Tree	1-Need for human input.	1- The SADT method is fast and efficient in the analysis of large volumes of Twitter data.

Table 3.3: State of the art of related work (continued)

The aim of the study was to analyze the data and achieve improved precision, recall, and accuracy in sentiment analysis. Various approaches, including Machine Learning and semantics, were employed, considering multiple factors.

Several techniques, such as Text Mining and Web Scraping, were utilized, along with supervised learning methods like SVM, Naïve Bayes, and Hybrid SVM-CNN. This approach offers notable advantages for sentiment analysis, including easy interpretability and efficient computation of results. The utilization of different algorithms leads to enhanced performance and response time, ul-

timately increasing the robustness of the evaluation system. However, these approaches also have limitations, such as the reliance of sentiment classification on data size.

Our proposed solution leverages the Elephant Herding Optimization (EHO) algorithm in conjunction with Long Short-Term Memory (LSTM) classification to enhance sentiment analysis for tweets. By integrating these techniques, our goal is to achieve higher performance and accuracy in analyzing sentiments within the context of tweets.

3.4 Conclusion

In this chapter, we have reviewed the state of the art of the main contributions in the field of sentiment analysis. We have synthesized all related work and presented it in a table, highlighting the main points of each approach. For each work, we have added a brief paragraph summarizing its main points. This literature review enables us to situate our research in relation to existing work, and to identify gaps and opportunities for our approach.

In the next chapter, we will detail our specific approach and the various steps we will follow. We will explain in detail our methodology, the techniques we will use and our approach to sentiment analysis.

Chapter 4

Contributions

4.1 Introduction

Sentiment Analysis is that branch of Natural Language Processing (NLP) that analyses textual data online by returning information on user opinions. It has a wide range of applications.

Sentiment analysis from data streams aims to identify user attitudes, emotions, and opinions from text in real time. In our case, it allows researchers and authorities to understand how people feel about the Covid-19 health crisis or simply about the corona virus.

Sentiment Analysis in smart cities is to integrate technological innovation with people's daily lives, promoting sustainability and quality of life.

The aim is to monitor the level of citizen's satisfaction. In this way, it is possible to make more analyses that are accurate and implement strategies that improve services and quality of life. Over time, various views related to the outbreak have been discussed on social networking platforms such as Twitter and Facebook.

The objective of this project is to help fill the void by reviewing the state of the art, challenges and opportunities of sentiment

analysis platforms, architectures and applications for the smart city application domain. Furthermore, the aim is to propose novel approach software supported for sentiment analysis in smart cities for healthcare crisis.

In this chapter, we present in detailed our approach for sentiment analysis in smart cities. We describe the different steps involved in extracting and processing tweets to classify them according to their polarity.

4.2 Approach steps

The COVID-19 pandemic has affected millions of lives around the globe and as a major issue affecting the health of people around the world; we need to know how people feel about this pandemic.

Scientists and researchers have been working hard since the beginning of the health crisis to find out the concerns and emotions of the people and how they feel about COVID-19. Since the beginning of this crisis, people's opinions about it are diverging on social networks, some are worried and concerned about the consequences of this pandemic, and others are peaceful and calm.

The objective of our work is to propose an approach based on a classification algorithm to analyze the public's feelings on social networks around the world towards the COVID-19 pandemic in order to better understand the public's attitude and concerns about it.

Approach Architecture

This approach architecture is sketched in Figure 4.1 , and it involves various layers to process the information. The pipeline begins at the data crawling layer and finishes in the final sentiment analysis process.

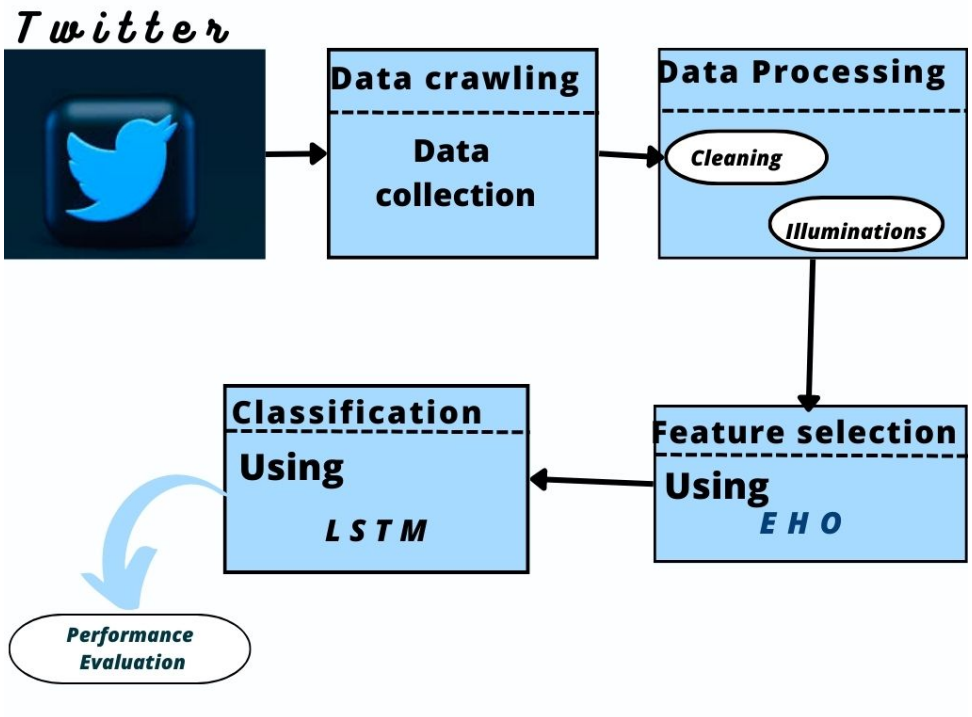

Figure 4.1: System architecture

Our approach involves mainly four steps. The first step entitled “Data collection”, which includes tweet crawling from various data sources. The second step named “Data processing” involves a series of modifications to the data collected, in 5 sub-steps (tokenization, normalization, stemming, lemmatization and vectorization), then the third stage, known as ”feature selection”, which consists of using EHO to choose the best features and parameters. Finally, the fourth stage called “classification” which involves using Long Short-Term Memory (LSTM) to classify the data into

categories neutral, positive and negative.

4.2.1 Data crawling

The first step is to collect the data from various data sources. Data crawling, also called ”web scraping” or ”spidering”, is the process that automatically extract data from the Internet. Using automated systems (“bots”) to extract data has many practical applications. Popular services such as search engines, price comparison websites, or news aggregators are essentially huge data crawling operations.

The purpose of the data collection step is to ensure that informative and reliable data are collected for statistical analysis in order to make data-driven research decisions [16].

The data used for sentiment analysis are comments, opinions and tweets expressing users’ opinions and feelings. the source of this data is social media, the tweeter platform [17]

4.2.2 Data processing

Data processing is the main step of our proposal. It occurs when data is collected and transformed into usable information. It is important that data processing is performed correctly so as not to negatively impact the final product or data output. Therefore, cleaning the tweets is very important to procee it. This step involves the following phases.

Tokenization

Tokenization refers to breaking a string of characters into smaller words called tokens. Tokenization includes identifying nouns, verbs, adverbs, and adjectives, among others. Grouping words with the same meaning is one of the processes in NLP [18].

For example, feel, feels, and feeling are considered one word, not distinct words [11].

The first phase of and cleaning textual data includes a set of operations which aim to define the objects of analysis: this is the tokenization of documents, which consists of recognizing basic textual units that can be words, but also letters, syllables, sentences or sequence of these elements. Each document then becomes an ordered or unordered list of basic concepts: tokens. We move from data schemes that list documents (and possibly assign them to authors) to schemes that associate documents with a set of attributes that are their unifying elements (letters, words, syllables, sentences) [19]. We consider two types of tokenization :

Word-based tokenization: Words are like atoms of natural language. They are the smallest unit of meaning. Verbatim tokenization of text makes it possible to identify the most frequently occurring words.

Segmentation by sentence: If a sentence is partitioned, we analyse the relationship between the words, which it contains to increase context.

The table 4.1 below represents an example of the tokenisation

phase; here we report the tweet before and after the tokenization.

Tweet before tokenization	tweet after tokenization
Advice Talk to your neighbors family to exchange, Coronavirus Australia: Woolworths to give, My food stock is not the only one which is empty, Me, ready to go at supermarket during Covid, As news of the region as first confirmed COVID case.	'Advice','Talk','to', 'your','neighbors','to','exchange',',', 'Coronavirus','Australia', 'Woolworths', 'to', 'give',',', 'My','food','stock','is', 'empty',',', 'Me',' ',', 'ready', 'to', 'go', 'at', 'supermarket',' during' 'Covid',',','As', 'news', 'of', 'region', 'as', 'first', 'confirmed' ,'COVID', 'case'.

Table 4.1: Tweets before and after the normalization

Normalization

Normalization is the step of removing words that occur in large numbers but are considered meaningless (empty words). A stop word list is a set of words that are widely used in different languages. The stop words are removed from many text mining-related applications due to their usage is too general, permitting the user to focus on other more important words[20].

The table4.2 below represents an example of the normalization phase, here we report the tweet before and after the normalisation.

Tweet before normalization	Tweet after normalization
‘Advice’,’Talk’,’to’, ‘your’,’neighbors’,’to’,’exchange’,’,’, ‘Coronavirus’,‘Australia’, ‘Woolworths’, ‘to’, ‘give’,’,’, ‘My’,‘food’,’stock’,’is’, ‘empty’,’,’, ‘Me’,’ ,’, ‘ready’, ‘to’, ‘go’, ‘at’, ‘supermarket’,’ during’ ‘Covid’,’,’,‘As’, ‘news’, ‘of’, ‘region’, ‘as’, ‘first’, ‘confirmed’ ,’COVID’, ‘case’.	Advice’,’Talk’, ’neighbors’,’exchange’, ‘Coronavirus’, ‘Australia’, ‘Woolworths’, ‘give’,‘food’,’stock’, ‘empty’, ‘ready’, ‘go’,‘supermarket’,’ during’ ‘Covid’, ‘news’, ‘region’, ‘first’, ‘confirmed’ ,’COVID’, ‘case’.

Table 4.2: An example of a tweet (Tweets before and after the normalization)

Stemming

The purpose of stemming is to assembly together many forms of a word as a single word. The idea is to remove suffixes, prefixes and other words in order to keep only their origin .

The resultant word is the same but this reduction is useful for a reduction in vocabulary size in bag-of-words approaches.

For example the words ”regionalization”, ”regionalist”, ” regionalism”, ”regionalisms” will be reduced to “regionalism ” so that all these forms refer to a single token: ”regionalism” [21].

Lemmatization

In lemmatization, words are generated down to their root words. Unlike stemming, lemmatization preserves parts of speech without splitting suffixes .

For example, in English, the word ”booked” is converted to ”booking” by lemmatization, while stemming converts it to ”book”. Stemming can be used when the stem is constant across all possible shapes, but when the stem is not, lemmatization is the best choice. Lemmatization aims to obtain a similar basic ”stem” of a word, but aims to derive the real root from the dictionary, not just a shortened version of the word. For example, ”went”, ”gone”, ”goes”, ”going” are lemmatized to ”go” [22].

The table 4.3represents a tweet before and after the lemmatization step:

Tweet before lemmatization	Tweet after lemmatization
‘Advice’,’Talk’, ’neighbors’,’exchange’, ‘Coronavirus’, ‘Australia’, ‘Woolworths’, ‘give’,‘food’,’stock’, ‘empty’, ‘ready’, ‘go’,‘supermarket’,’ during’ ‘Covid’, ‘news’, ‘region’, ‘first’, ‘confirmed’ ,’COVID’, ‘case’.	‘Advice’,’Talk’, ’neighbor’,’exchange’, ‘Coronavirus’, ‘Australia’, ‘Woolworth’, ‘give’,‘food’,’stock’, ‘empty’, ‘ready’, ‘go’,‘supermarket’,’ duren’ ‘Covid’, ‘new’, ‘region’, ‘first’, ‘confirm’ ,’COVID’, ‘case’.

Table 4.3: A Tweet before and after lemmatization

Vectorization

Vectorization comprises creating a sparse matrix of all words and their number of repetitions in the document using a counting vectorizer or the TF-IDF vectorizer [23].

Vectorization is the process of converting text to digital input in raster form. Vectorization creates a document-term matrix where each cell represents the number of times a word occurs in the document, also known as term frequency (TF) .

A document-term matrix is a set of dummy variables indicating whether a given word occurs in a document. A column is dedicated to each word in the corpus. This count is proportional to the category relevance of the news headline. This means that if a given word appears multiple times in fake or real news headlines, then the given word has high predictive power in determining whether a headline is fake or real [24].

4.2.3 Feature selection

In artificial intelligence to process data and provide prediction in after training in efficient way is the biggest challenge, to overcome such problem different optimization algorithm produced by various researchers, in which it uses to maximize and minimize the function to reduce the error. It varies on model learning parameters for the computation of the target value (Y) from predictor value (x) being used.

In the neural network, there are some weights (w) and the bias (b) used as learning parameters during computations and gives output updates. The main purposed of using optimization algorithm is to minimize or maximize a loss function using gradients parameters; one of the most useful algorithms is Practical Swarm Optimization (PSO), Elephant Herding Optimization (EHO), etc...

In our proposed method, we are more focused on distinguishing the efficiency of deep learning algorithm by using EHO .

EHO algorithm is one of the latest swarm intelligence algorithms. It was proposed in 2016 by Wang, Deb, Gao, and Coelho.

Even though it is a rather new optimization algorithm, it has already been used for various applications. EHO algorithm was proposed for community detection in complex social networks [25] .

The EHO algorithm is used to optimize parameter control and selection and convergence speed in the deep learning architecture to improve the accurate classification of tweets, in other words, by using the EHO algorithm, the parameters and convergence of the deep learning architecture are optimized, thereby significantly improving the accuracy and evaluation of tweets [26].
As pack animals, elephants live in a social structure of females and calves. The elephant clan is led by a matriarch and consists of several elephants. Female members prefer to live with their families, while male members tend to live elsewhere.

They gradually become independent from the family until they completely abandon the family[25], the numbers of all elephants are shown in Figure 4.2, EHO considers the following assumptions[27].

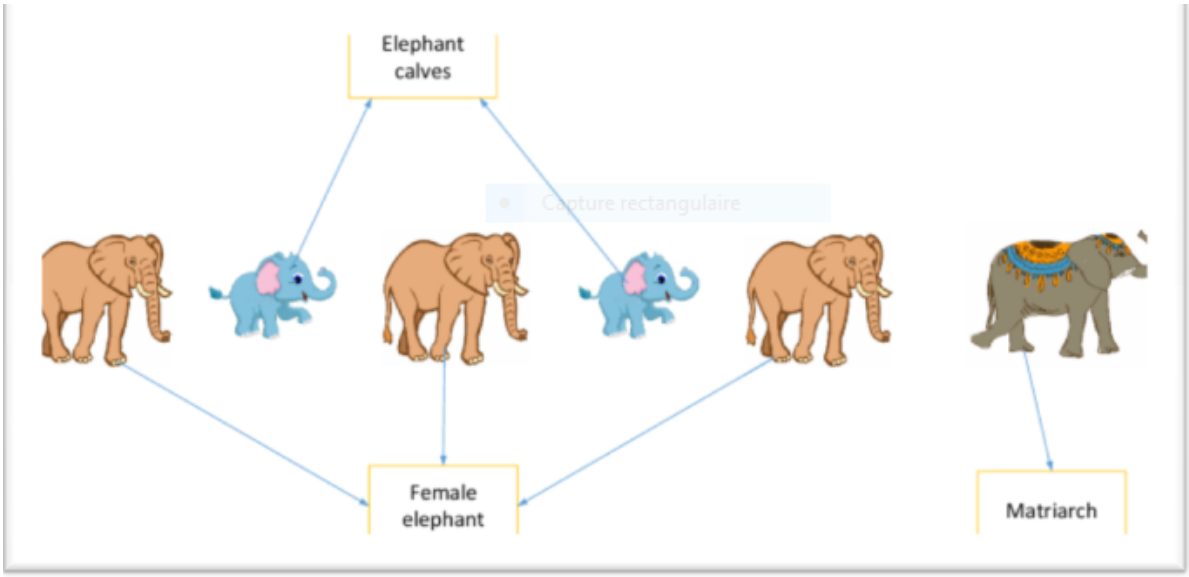

Figure 4.2: Elephants'clan members

On the basis of the natural habits of elephants, a matriarch is the leader of her clan. Thereby, the new position of each elephant Ti is influenced by the matriarch Ti. Elephant n in clan Ti can be calculated using the equation (1).

$$X_{new,Ti,n}=X_{Ti,n} + a\,(X_{best,Ti} - X_{Ti,n})\,r \qquad (1).$$

Where $X_{center\ ,Ti,n}$ and $X_{Ti,n}$ denote the new and old position of elephant n in clan Ti,respectively. $X_{best,Ti}$,is the matriarch Ti, representing the best elephant in the clan. $a \in [0, 1]$, specifies the scaling factor, $r \in [0, 1]$. The best elephant for each clan can be calculated using equation (2).

$$X_{center\ ,Ti,n} = X_{center\ ,Ti} \qquad (2)$$

Where $\beta \in [0,1]$ represents the factors that determine the influence of $X_{center\ ,Ti}$, $X_{new,Ti,n}$ on are new individuals.X_{center} , Ti is the central individual of clan Ti. For the d-th dimension, it can be calculated according to formula (3).

$$X_{\text{center, Ti, -d}} = \frac{1}{bTi} \sum_{i=1}^{bTi} \text{X,n,d} \qquad (3)$$

Where $1 \leq d \leq D$ and b_{Ti} represent the number of elephants in clan Ti. $X_{\text{Ti, n, d}}$ represent the d-th dimension of elephant individualX_{Ti} $X_{\text{center, Ti}}$ is the center of family Ti, which can be updated by formula (3).

Separating Operator

The separation process of a male elephant leaving its family group can be modeled as a separation operator. The separation operator is implemented by the elephant with the worst fitness in each generation, as shown in formula (4).

$$X_{\text{worst , Ti}} = X_{\text{min}} + (X_{\text{max}} - X_{\text{min}} + 1) \text{ rand} \quad .....(4)$$

X_{max} represents the upper limit of the individual, X_{min}represents the lower limit of the individual, and $X_{\text{worst ,Ti}}$, represents the worst individual in the clan Ti. Rand $\in$[0,1] is a random value between 0 and 1 [28].

The following figure 4.3 [29] shows the process of separation of the adult elephant from its clan:

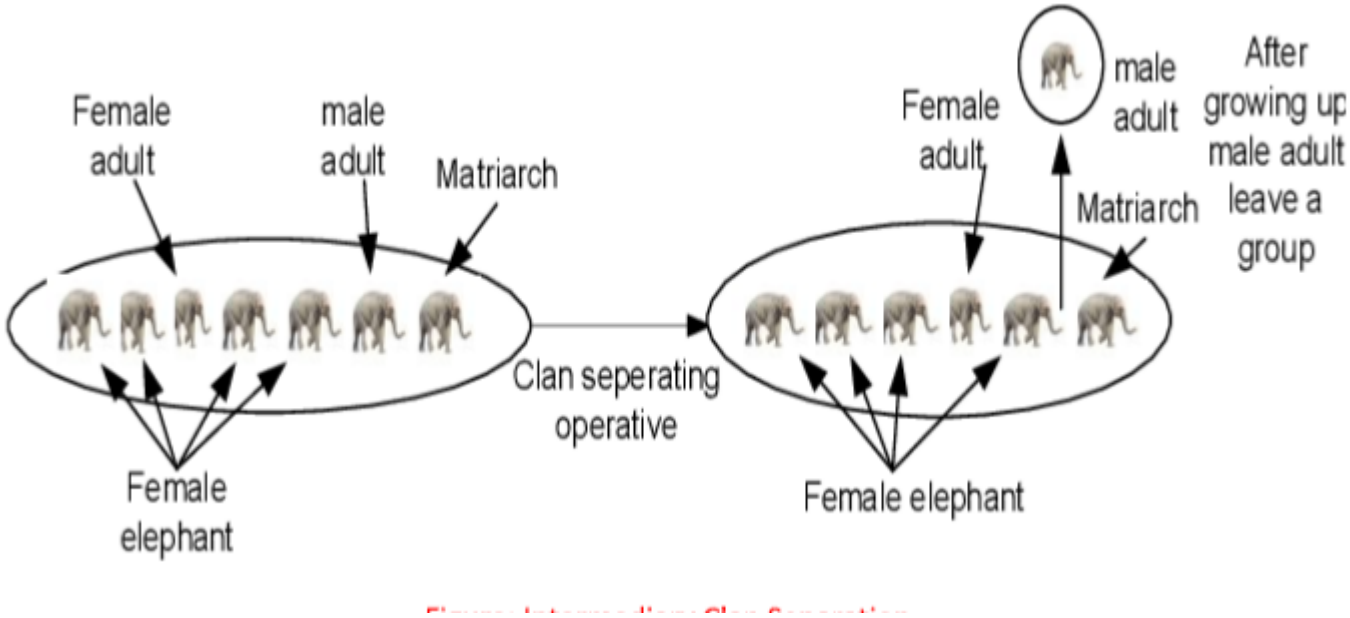

Figure 4.3: Intermediary Clan Separation.

The EHO Algorithm Code

Algorithm 1 Elephant Herding Optimization

```
Begin
Initialization. Set the initialize iterations H = 1; initialize the population R randomly; set maximum generation MaxGen
while stopping criterion is not met do
    Sort the population according to fitness of individuals.
    for all clans T_i do
        for all elephant n in the clan T_i do
            Generate x_new, T_i, n and update X_{T_i,n} by Equation 1
            if X_{T_i,n} = X_{best,T_i} then
                Generate X_new, T_i, n and update X_{T_i,n} by Equation 2
            end if
        end for
    end for
    for all clans T_i do
        Replace the worst individual c_i by Equation 4
    end for
    Evaluate each elephant individual according to its position.
    H = H + 1
end while
End
```

The goal of EHO is to simplify the control and selection of parameters to improve finding subsets of features from a larger fea-

ture pool, enabling more accurate tweet classification and scoring [30],The figure 5.7 illustrates the importance of applying EHO on the classification resultants in our appraoch.

4.2.4 Classification

In the field of sentiment analysis, it is important to improve the classification and evaluation of tweets, which are a group of different opinions, from people, about the COVID19 pandemic and its after-effects on their daily lives. To this end, this paper presents an efficient method for collecting and classifying tweets by combining LSTM and EHO, where EHO is used to optimize parameter control and selection and deep convergence speed learning architecture is used to improve the accuracy of tweet classification .

Classification is used to determine whether a document belongs to a set of predefined class documents. Automated classification systems can support the classification process largely.

With the rapid growth of Internet information, text classification has become a common and important trend in the field of information retrieval. Most approaches to text classification problems are proposed to improve the accuracy of text classifiers. For text classification tasks, documents are identified by the words that appear in the text. For this purpose, we use the LSTM algorithm which is a deep learning algorithm widely use for text classfication [26].

Deep learning is a machine learning-derived type of artificial intelligence based on artificial networks of neurons inspired by the human brain [27].

Artificial neural networks' algorithms are used to solve machine-learning problems. A neural network is a set of artificial neurons organized into layers (an input layer, an output layer, and one or more hidden layer), where each neuron in the hidden layer is a perceptron [31].

The following figure 4.4 [32] represents neural network diagram.

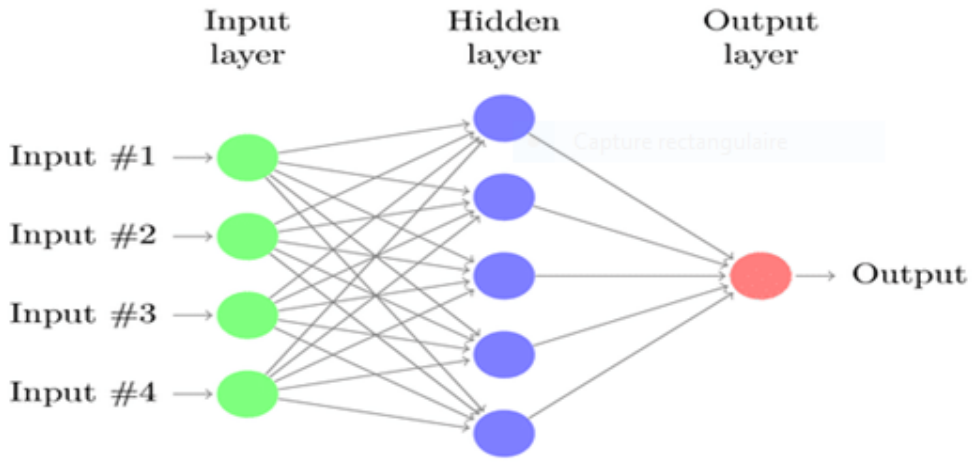

Figure 4.4: Neural network diagram.

Recurrent Neural Networks

Recurrent Neural Networks (RNNs) perform recurrent processing of information. Data can be transferred in both forward and reverse directions and is specifically designed to handle sequence data, such as word sequences for machine translation problems, audio data for speech recognition, or time series for prediction problems[33].

Long Short-Term Memory (LSTM): The standard basic RNN has an evanescent gradient problem, that is, the gradient decreases

as the number of layers increases. These networks have short-term memory and do not work well with long sequences, which require storing all the information contained in the entire sequence. For this reason, LSTM recurrent networks seem to be able to solve the evanescent gradient problem. LSTM uses three gates to store relevant long-term information and reject irrelevant information. These gates are as follows:

Γf Forget Gate: Decide which information should be discarded or saved. - Values close to 0: previous information is forgotten.

- Values close to 1: information is preserved. Γu your update gate: decides what new information ct to use to update the memory state ct. Therefore, ct is updated using Γf and Γu. Γo output gate: decides which output value is used as input to the next hidden unit.

Figure 5.1 illustrates the architecture of long short-term memory [33].

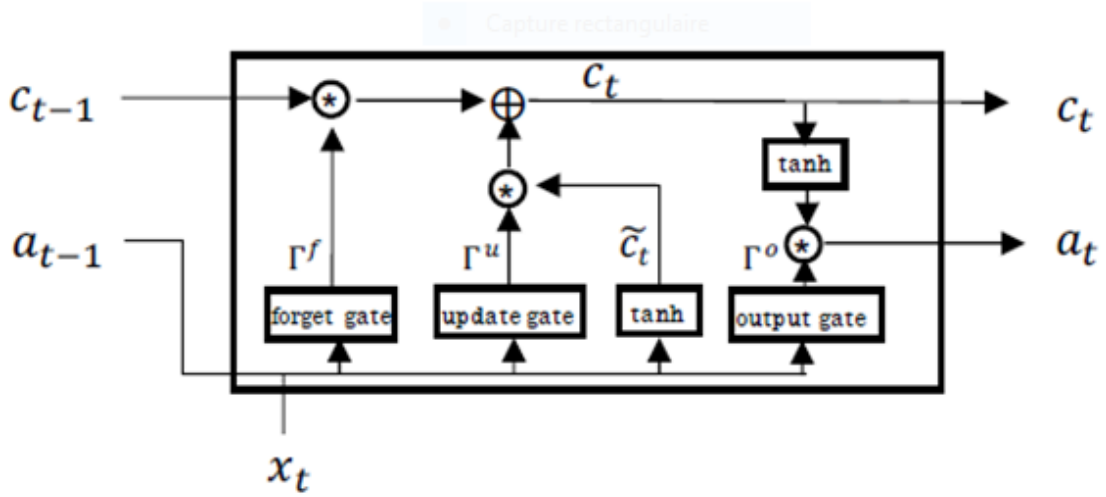

Figure 4.5: Long- and short-term memory architecture.

The information of the previous hidden unit at−1 and the infor-

mation of the current input xtcalculate all the gate values through the sigmoid σ activation function and calculate the new information çt through the tanh activation function for updating. The equations defining the LTSM elements are as follows [34]:

$$C_t = \tanh\left(W_c\,[a_{t-1}, x_t] + b_c\right)$$
$$\Gamma^u = \sigma\left(W_u\,[a_{t-1}, x_t] + b_u\right)$$
$$\Gamma^f = \sigma\left(W_f\,[a_{t-1}, x_t] + b_f\right)$$
$$\Gamma^o = \sigma\left(W_o\,[a_{t-1}, x_t] + b_o\right)$$
$$C_t = \Gamma u \times \text{ç}_t + \Gamma^f \times c_{t-1}$$
$$a_t = \Gamma^o \times \tanh\left(c_t\right)$$

Figure 4.6: LTSM elements.

Where W_u, W_f , W_o, b_u, bf and b_o are the weights and biases that determine the behavior of the gates Γu, Γf and Γo, respectively, and W_c and b_c are the weights and biases of the candidate memory cell c_t.

LSTMs solve the problem of short-term memory by introducing a specific memory cell structure. This memory cell enables the network to store and access past information over a long period of time. It is designed to retain important information and avoid the problem of gradient disappearance. Thanks to their special architecture, LSTMs can store variable-length sequences and maintain information over an extended period of time.

The LSTM learning model

Algorithm 2 LSTM learning model

```
modèle ← Séquentiel()
modele.add(Incorporation(voc_size,embedding_vector_fatures,input_length=sent_length))
modele.add(LSTM(128,input_shape(embedded_docs.shape),activation='relu', return_sequences=True))
modèle.add(Dropout(0.2))
modèle.add(LSTM(128, activation='relu'))
modèle.add(Dropout(0.2))
modèle.add(Dense(32, activation='relu'))
modèle.add(Dropout(0.2))
modèle.add(Dense(4, activation='softmax'))
modèle.compile(loss='sparse_categorical_crossentropy'o ptimizer='adam', metrics=['accuracy'])
imprimer (modèle.summary())
```

The model begins with an embedding layer that converts words into dense vectors, enabling the model to capture the semantic relationships between words. Next, two LSTM layers are added to exploit long-term memory and capture temporal dependencies in word sequences.

To avoid overlearning, dropout layers are inserted after each LSTM layer. These layers randomly deactivate certain neurons during training, thus generalizing the model and reducing the detrimental effects of overlearning.

The model ends with a dense layer with a 'relu' activation function, followed by another dropout layer and a final dense layer with a 'softmax' activation function. This last layer assigns probabilities to each output class, enabling texts to be classified according to predefined categories.

4.3 Conclusion

In this era of deep learning where models are used in well-known applications such as speech to text, real-time translation, image recognition, requires a large amount of data for the training. To process a large amount the data, model efficiency is a major concern, for this, there are few optimizations' methods being developed by a different researcher. In our research, we used the famous optimization method EHO to select the best features for the deep learning algorithm.

In this chapter, we have described the main steps of the proposed approach for sentiment analysis classification. Our system goes through various steps namely data collection, data processing, future selection and classification.

We have defined what the elephant herding optimization (EHO) algorithm is, we have defined how LSTM works and why we chose it.

In the next chapter, we will implement and evaluate our approach for tweet classification. We also present the used tools the development environment.

Chapter 5

Implementation and experiments

5.1 Introduction

In our research, we have processed English texts for sentiment analysis. This processing extracts the polarity of opinions expressed in negative, positive and neutral terms. The input data we use are tweets extracted from the dataset.

In this chapter, we will present the different aspects related to the implementation of the method we have developed, i.e. the technologies, software and languages chosen with different data sources to implement our method

5.2 Dataset description

The dataset used was extracted from Tweeter and is in CSV format, since it is easier for Python to handle this type of file in the field of sentiment analysis. The size of the record is 102,000 KB. It can be downloaded [17], and used by any analyst who needs the dataset. Only the structure and type of data contained in the dataset need to be considered.

The Dataset is composed of six (6) columns:

• UserName which is the name of the user (Ex: 3799).

• Screen Name which is the number of the tweet (Ex: 48751).

• Location which is the location of the user who posted the tweet (Ex: London).

• TweetAt which is the date of the tweet's publication (Ex: 16-03-2020).

• OriginalTweet which is the tweet in question (user's opinion).

• Sentiment which is the category of the tweet (Ex: Neutral)

	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment
0	3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral
1	3800	48752	UK	16-03-2020	advice Talk to your neighbours family to excha...	Positive
2	3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elde...	Positive
3	3802	48754	NaN	16-03-2020	My food stock is not the only one which is emp...	Positive
4	3803	48755	NaN	16-03-2020	Me, ready to go at supermarket during the #COV...	Negative
5	3804	48756	ÃT: 36.319708,-82.363649	16-03-2020	As news of the regionÂs first confirmed COVID...	Positive

Figure 5.1: Our dataset

5.3 Development environment

5.3.1 Google Colab

Google Colab, also known as Collaboratory, is gaining popularity in the field of education and research. This platform serves the purpose of spreading knowledge and research in machine learning. One of its notable features is the ability to create notebooks where users can write and execute code. Similar to Google Docs, users can add comments to the code, allowing for collaboration and sharing. This collaborative environment fosters the development of Python programming language code and encourages teamwork

in machine learning projects [35].

5.3.2 Jupyter Notebook

Jupyter Notebook is an open-source, browser-based tool that acts as a virtual laboratory notebook, supporting workflow, code, data, and visualizations to describe research processes in detail. It is machine and human readable, facilitating interoperability and scholarly communication. These notebooks can be stored in online repositories and provide links to research objects such as datasets, code, method documents, workflows, and publications located elsewhere.

Jupyter notebooks are a tool for making science more open.

Their relevance to the JCDL community lies in their interaction with multiple components of the digital library infrastructure, such as digital identifiers, persistence mechanisms, version control, records, documents, software, and publications [36].

5.4 Programming language

Python is an open source programming language developed in 1991 by programmer Guido van Rossum. It got its name from the TV show Monty Python's Flying Circus . Because it's an interpreted programming language, it doesn't need to be compiled to work. An "interpreter" program allows you to run Python code on any computer. This allows you to quickly see the results of code changes. On the other hand, this makes the language slower than compiled languages like C.

As a high-level programming language, Python allows programmers to focus on what they do, not how they do it. Therefore, writing programs takes less time than in any other language, Ideal language for beginners[37].

5.4.1 Python libraries

Pandas

Pandas: is a BSD-licensed open source library that provides powerful, easy-to-use data structures and data analysis tools for the Python programming language [38], Installation is done by opening a command shell and invoking the command: pip install pandas.

This makes it easy to edit data tables with variable and person labels. These tables are called ”data frames” (stored in CSV, TSV files, etc.) similar to data frames in R. You can easily read and write these dataframes or table files, and draw charts from these dataframes using matplotlib[39].

Numpy

It is the basic package for scientific computing in Python. This is a Python library that provides multidimensional array objects, various derived objects (such as masked arrays and matrices), and a set of routines for performing fast operations on arrays, including mathematics, logic, basic linear algebra, basic statistical operations, Random simulations and more.[38].

TensorFlow

Developed by Google researchers, Tensor Flow is an open source

machine learning, deep learning, statistical and predictive analytics tool. Like similar platforms, it is designed to simplify the development and execution of advanced analytics applications for data scientists, statisticians, and predictive modelers.

Tensor Flow software manages datasets by arranging them as computation nodes on an execution graph. Connections between nodes in a graph can represent multidimensional matrices or vectors, resulting in so-called tensors [40].

5.5 Implementation

5.5.1 Import a dataset

Prior to data preparation, it is common practice to import the dataset from a Comma-Separated Values (CSV) file, or any other data file format. Here's how to import a dataset from a CSV file:

```
# Charger les données à partir du fichier CSV
train_data = pd.read_csv('Corona_NLP_train.csv', encoding='ISO-8859-1', low_memory=False)

train_data['Sentiment'] = train_data['Sentiment'].replace('Extremely Negative', 'Negative')
train_data['Sentiment'] = train_data['Sentiment'].replace('Extremely Positive', 'Positive')

texts, labels = train_data['OriginalTweet'].values, train_data['Sentiment'].values
```

Figure 5.2: Import a dataset

5.5.2 Preparation of Datas

In the data pre-processing stage, we perform various operations to prepare our data. Figure 5.3shows the code used.

```
from keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences

# Diviser les données en ensembles d'entrainement, de validation et de test
train_texts, train_labels = texts[:40000], labels[:40000]
val_texts, val_labels = texts[40000:41000], labels[40000:41000]
test_texts, test_labels = texts[41000:], labels[41000:]

# Convertir les textes en séquences d'entiers
max_words = 10000
max_len = 100
# Convert elements in train_texts to strings
train_texts = [str(text) for text in train_texts] #plus
tokenizer = Tokenizer(num_words=max_words)
tokenizer.fit_on_texts(train_texts)



train_seqs = tokenizer.texts_to_sequences(train_texts)
val_texts = [str(text) for text in val_texts]#plus
val_seqs = tokenizer.texts_to_sequences(val_texts)
test_seqs = tokenizer.texts_to_sequences(test_texts)

# Remplir les séquences pour qu'elles aient toutes la même longueur
train_seqs = pad_sequences(train_seqs, maxlen=max_len)
val_seqs = pad_sequences(val_seqs, maxlen=max_len)
test_seqs = pad_sequences(test_seqs, maxlen=max_len)

train = train_seqs.reshape(train_seqs.shape[0], max_len, 1)
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()
train_labels = le.fit_transform(train_labels)

val_texts = val_seqs.reshape(val_seqs.shape[0], max_len, 1)
val_labels = le.fit_transform(val_labels)
```

Figure 5.3: Preparation of Datas

5.5.3 Featureselaction and classification

In this step, we applied the EHO algorithm in conjunction with LSTM to process the data.

In Figure 5.4, represent the first iteration of the EHO algorithm, elephants (representing different combinations of hyper parameters) are evaluated using a specific objective function (in this case, the ‘objective-function‘). Costs (or performance metrics) are calculated for each elephant, and the dominant elephant (the one with the lowest cost) is identified.

The positions of the elephants are then updated using mathematical operations to approximate the position of the dominant elephant. Hyper parameters are updated according to the position of the dominant elephant, then a new LSTM model is created and trained with these updated hyper parameters.

This process is repeated over several iterations, and at the end of the algorithm, the best hyperparameters are those associated with the final dominant elephant.

```python
# Définir la fonction objectif pour optimiser les hyperparamètres
max_features = 1
history = []
# Boucle principale de l'algorithme EHO
for it in range(num_iter):
    print("literation numero :", it)
    # Calculer les coûts pour chaque éléphant
    costs = np.array([objective_function(elephants[i, :]) for i in range(pop_size)])
    # Trouver l'indice de l'éléphant dominant
    index = np.argmin(costs)
    # Mettre à jour les positions des éléphants
    for i in range(pop_size):
        if i == index:
            continue
        r = np.random.rand(3)
        elephants[i, :] = (elephants[i, :] + r * (elephants[index, :] - elephants[i, :]))
    # Mettre à jour les hyperparamètres
    layer_size = int(elephants[index, 0])
    learning_rate = elephants[index, 1]
    num_epochs = int(elephants[index, 2])
    # Créer et entrainer le modèle avec les hyperparamètres mis à jour
    model = Sequential()
    model.add(Embedding(input_dim=max_words, output_dim=128, input_length=max_len))
    model.add(LSTM(layer_size))
    model.add(Dense(1, activation='sigmoid'))
    optimizer = Adam(lr=learning_rate)
    model.compile(loss='binary_crossentropy', optimizer=optimizer, metrics=['accuracy'])
    model.fit(train, train_labels, validation_data=(val_texts, val_labels), epochs=1, batch_size=64, verbose=0)
```

Figure 5.4: Code of EHO in conjunction with LSTM

5.6 Evaluation

In this step we display a Classification report, which generally presents these measures for each class in the dataset, allowing you to evaluate the model's performance [41] [42].

5.6.1 Precision

Precision measures how accurate the classifier is. Higher precision means fewer false positives, while lower precision means more false positives.This is often not consistent with recall, since an easy way to increase precision is to decrease recall

Accuracy = True Positives / (True Positives + False Positives)

5.6.2 F1-score

The F1 score is a measure of test accuracy. Both the precision and recall of the test are consideredwhen calculating the score. F-Score is the harmonic mean of precision and recall. Hers how your system works.

F1 score = [2 * (precision * recall) / (precision + recall)]

5.6.3 Recall

Recall measures the completeness or sensitivity of a classifier. Higher recall means fewer falsenegatives, while lower recall means more false negatives. Improving recall usually leads to a decrease in precision, since the larger the sample space, the harder it is to be precise

Alerts = True Positives / (True Positives + False Negatives).

5.6.4 Support

Support is the actual number of occurrences of the class in the dataset. It doesn't vary between models, it's just a diagnostic performance evaluation process.

The precision of all classes

The figure 5.5 shows the accuracy score of all the dataset and the code used.

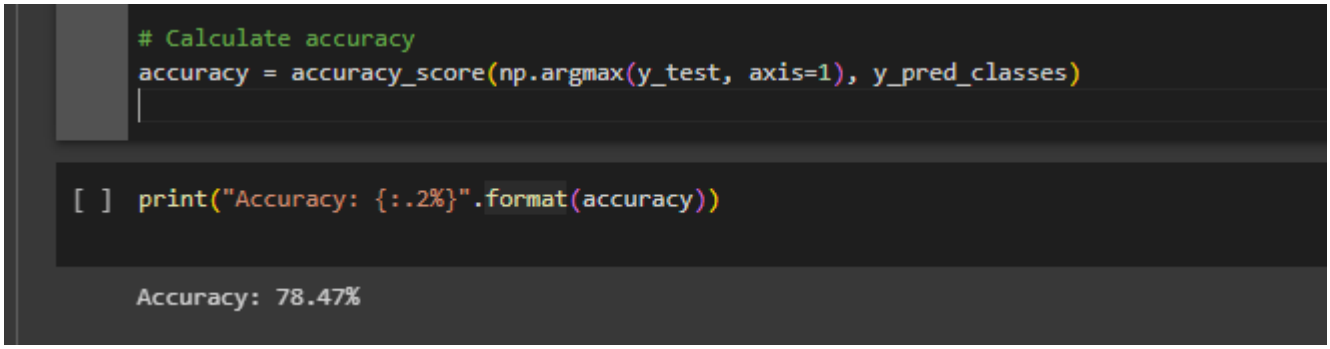

Figure 5.5: Accuracy

The figure 5.6 shows precision, recall and F score for all sentiments such that:

0 represent neutral feelings.

1 represent negative feelings .

2 represent positive feelings.

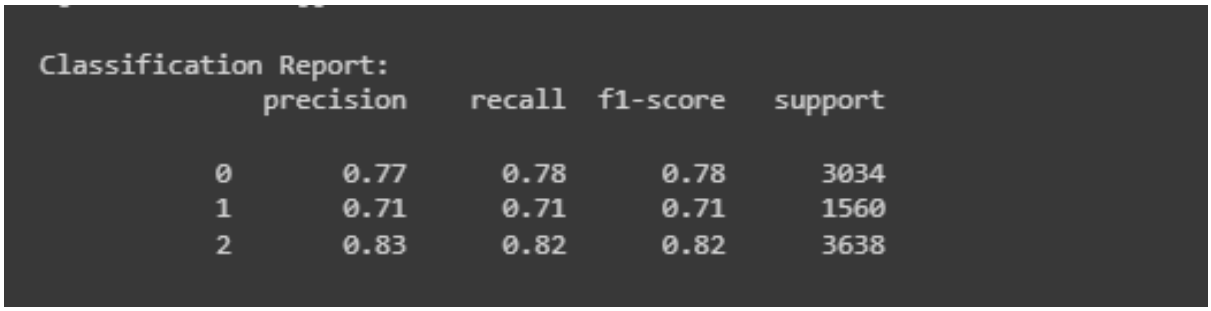

Figure 5.6: Results

5.7 Comparison end Discussions

Comparison:

To explain the importance of the eho algorithm in data processing and in NLP we have calculated the accuray without the algorithm and the figure 5.7 represents the results obtained.

```
# Calculate accuracy
accuracy = accuracy_score(np.argmax(y_test, axis=1), y_pred_classes)
print("Accuracy: {:.2%}".format(accuracy))
```

```
Accuracy: 18.50%
```

Figure 5.7: Accuray without EHO

Discussions:

In our approach, the combination of LSTM with EHO yields precision results of 78.47 %, an F-score of over 71 %, recall of 71 % and support of 3638 , indicating that the model performs well in terms of identifying positive instances and maintains a good balance.

The model's performance can be considered quite good.

5.8 Conclusion

In this chapter, we have implemented our LSTM-based approach using the EHO algorithm. We have carried out experiments and obtained valid and promising results.

In conclusion, our implementation of the LSTM approach with EHO has provided validated and encouraging results for sentiment analysis in tweets. This paves the way for future research and applications in the field of sentiment and public opinion analysis in social media.

Chapter 6

General conclusion

Sentiment analysis provides a comprehensive understanding of Twitter users, opinions, and sentiment related to COVID-19. By analyzing the content of tweets, it is possible to distinguish between negative, positive and neutral sentiment.

For sentiment analysis, natural language processing techniques can be used to extract key featuresfrom tweets, such as: words, phrases, or emoji that represent emotions. The labeled data can thenbe used to train a machine learning algorithm to classify tweets as negative, positive, or neutral.

By examining the sentiment in tweets, it is possible to determine the general public sentiment towards COVID-19. This information is valuable for a variety of purposes, including monitoring publicopinion, tracking the effectiveness of public health interventions, identifying problem areas, andassessing the impact of communication strategies.

At the end of this project, we remind you that the purpose of this work was to determine the feelingsof Twitter users and their positive, negative or neutral opinions about the Covid-19 pandemic in smart cities. Our application provides different views through

imported data. It allows for proper integration of data for analysis.

The work presented in this paper is guided by the many concerns people have about the Covid-19 health crisis and its impact on their future in order to better understand the reasons for these significant concerns and worries.

We would like to point out that the procedures we implemented resulted in the polarity of feelings expressed on Twitter, namely negative, positive and neutral, and the percentage value of each polarity. Despite the fact that this is a topical subject, and that the programs used are effective and useful in other studies, this work is open to improvement, given that it remains incomplete. In particular, it has limitations linked to:

-Sentiment analysis does not recognize facial expressions (uses sensors for facial recognition).

-Sentiment analysis does not recognize the emoji used.

-Sentiment analysis is not up to date as the application needs to be refreshed every time the dataset is updated.

Perspectives

We have pushed the project as far as we can, but there are still many steps that need to be added to improve it. We considered the following main points:

- Improve the system and enable real-time analytics using the Twitter API.

- Allow users to input their own datasets for analysis.

- Analyze negative words and phrases to better understand what users are worried about.

- Add functionality to analyze and categorize emoji.

References

[1] S. Maurel, P. Curtoni, and L. Dini, “L’analyse des sentiments dans les forums,” Atelier Fouille des Données d’Opinion, 2008.

[2] https://www.tibco.com/, “what is sentiment analysis,” 2022.

[3] V. Albino, U. Berardi, and R. M. Dangelico, “Smart cities: Definitions, dimensions, performance, and initiatives,” Journal of urban technology, vol. 22, no. 1, pp. 3–21, 2015.

[4] https://www.twi global.com, “/what-is-a-smart-city,” consulted on December 2022.

[5] S. Selvakanmani, “Smart city–the urban intelligence of india,” International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 3, no. 6, pp. 302–307, 2015.

[6] Z. Jalil, A. Abbasi, A. R. Javed, M. Badruddin Khan, M. H. Abul Hasanat, K. M. Malik, and A. K. J. Saudagar, “Covid-19 related sentiment analysis using state-of-the-art machine learning and deep learning techniques,” Frontiers in Public Health, vol. 9, p. 2276, 2022.

[7] N. Yeasmin, N. Mahbub, M. Baowaly, B. Singh, Z. Alom, Z. Aung, and M. Azim, "Analysis and prediction of user sentiment on covid-19 pandemic using tweets. big data cogn. comput. 2022, 6, 65," 2022.

[8] R. Chandra and A. Krishna, "Covid-19 sentiment analysis via deep learning during the rise of novel cases," PloS one, vol. 16, no. 8, p. e0255615, 2021.

[9] M. Habibi, A. Priadana, and M. R. Ma'arif, "Sentiment analysis and topic modeling of indonesian public conversation about covid-19 epidemics on twitter," IJID (International Journal on Informatics for Development), vol. 10, no. 1, pp. 23–30, 2021.

[10] "Kaur, chhinder et sharma, anand , "analyse sentimentale covid-19 à l'aide de techniques d'apprentissage automatique," in Progress in Advanced Computing and Intelligent Engineering : Actes de l'ICACIE 2020, pp. 153–162.

[11] K. P. Iyer and S. Kumaresh, "Twitter sentiment analysis on coronavirus outbreak using machine learning algorithms," Eur. J. Mol. Clin. Med, vol. 7, no. 3, pp. 2663–2676, 2020.

[12] M. A. et Sumayh S. Aljameel et Irfan Ullah Khan et Nida Aslam et Sara Mhd. Bachar Charouf et Norah Alzahrani, "Modèle d'apprentissage automatique pour l'analyse des sentiments des tweets covid-19," Revue internationale sur les sciences avancées, l'ingénierie et les technologies de l'information, vol. 12, pp. 1206–1214.

[13] B. P. Pokharel, "Twitter sentiment analysis during covid-19 outbreak in nepal," Available at SSRN 3624719, 2020.

[14] A. H. Alamoodi, B. B. Zaidan, A. A. Zaidan, O. S. Albahri, K. I. Mohammed, R. Q. Malik, E. M. Almahdi, M. A. Chyad, Z. Tareq, A. S. Albahri, et al., "Sentiment analysis and its applications in fighting covid-19 and infectious diseases: A systematic review," Expert systems with applications, vol. 167, p. 114155, 2021.

[15] D. Dangi, D. K. Dixit, and A. Bhagat, "Sentiment analysis of covid-19 social media data through machine learning," Multimedia Tools and Applications, vol. 81, no. 29, pp. 42261–42283, 2022.

[16] J. Claussen and C. Peukert, "Obtaining data from the internet: A guide to data crawling in management research," Available at SSRN 3403799, 2019.

[17] A. MIGLANI, "Coronavirus tweets pnl - classification de texte." https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification/code, 2020.

[18] N. Abd Rahim and S. Rafie, "Sentiment analysis of social media data in vaccination," Int J, vol. 8, no. 9, 2020.

[19] S. Balech and C. Benavent, "Les techniques du nlp pour la recherche en sciences de gestion," hal-02400308ff, 2019.

[20] A. Bayhaqy, S. Sfenrianto, K. Nainggolan, and E. R. Kaburuan, "Sentiment analysis about e-commerce from tweets

using decision tree, k-nearest neighbor, and naïve bayes," in 2018 international conference on orange technologies (ICOT), pp. 1–6, IEEE, 2018.

[21] https://openclassrooms.com, "Nettoyez et normalisez les données," consulted on 05/07/2023.

[22] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," A practical guide to sentiment analysis, pp. 1–10, 2017.

[23] H. Tran-Ngoc, S. Khatir, T. Le-Xuan, G. De Roeck, T. Bui-Tien, and M. A. Wahab, "A novel machine-learning based on the global search techniques using vectorized data for damage detection in structures," International Journal of Engineering Science, vol. 157, p. 103376, 2020.

[24] https ://towardsdatascience.com, "data science," consulted on 05/08/2023.

[25] S. e. C. L. d. S. Wang, Gai-Ge et Deb, "Optimisation de l'élevage d'éléphants," in 2015 3e symposium international sur l'informatique et l'intelligence d'affaires (ISCBI), pp. 1–5.

[26] S. M. Baneamoon, "Combining deep learning and elephant herding optimization for pedestrian detection from a drone-based images,"

[27] D. B. Avval, P. O. Heris, N. J. Navimipour, B. Mohammadi, and S. Yalcin, "A new qos-aware method for produc-

tion scheduling in the industrial internet of things using elephant herding optimization algorithm," Cluster Computing, pp. 1–16, 2022.

[28] J. Li, H. Lei, A. H. Alavi, and G.-G. Wang, "Elephant herding optimization: variants, hybrids, and applications," Mathematics, vol. 8, no. 9, p. 1415, 2020.

[29] O. Sambariya, DK et Nagar, "Application du contrôleur fopid pour lfc utilisant la technique d'optimisation de l'élevage d'éléphants," in 2018 3e Conférence internationale IEEE sur les tendances récentes en électronique, technologies de l'information et de la communication (RTEICT), pp. 833–837.

[30] A. E. Hassanien, M. Kilany, and E. H. Houssein, "Combining support vector machine and elephant herding optimization for cardiac arrhythmias," arXiv preprint arXiv:1806.08242, 2018.

[31] V. Kévan, "www.blog businessdecision.com -tutoriel | machine learning : comprendre ce qu'est un réseau de neurones et en créer un !," blog, vol. 13, 10 novembre 2020.

[32] C. Delmegani, "Dark side of neural networks explained," AI Multiple Logo, 2022.

[33] J. F. Torres, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso, "Deep learning for time series forecasting: a survey," vol. 9, no. 1, pp. 3–21, 2021.

[34] M. Bouaziz, “Reseaux de neurones recurrents pour la classification de sequences dans des flux audiovisuels parall eles –université d’avignon,” 2017.

[35] F. R. V. Alves and R. P. M. Vieira, “The newton fractal’s leonardo sequence study with the google colab,” vol. 15, no. 2, p. em0575, 2019.

[36] B. M. Randles, I. V. Pasquetto, M. S. Golshan, and C. L. Borgman, “Using the jupyter notebook as a tool for open science: An empirical study,” pp. 1–2, IEEE, 2017.

[37] L.Bastien, “https:/www.lebigdata.fr/python-langage-definition,” consulted on May 25,2023.

[38] https/numpy.org/doc/stable/user/whatisnumpy.html, “What is numpy?,” consulted on June 1th, 2023.

[39] http ://www.python-simple.com/python-pandas/panda intro.php, “Introduction à pandas,” on June11th,2023.

[40] https ://pandas.pydata.org/docs, “pandas documentation,” consulted on June11th,2023.

[41] JACOB, “https:/streamhacker.com/2010/ 05/17/text- classification de texte pour l’analyse des sentiments prÉcision et rappel,” consulted on June 06 2023.

[42] J. JVC, “www.data-transitionnumerique.com/machine-learning-python,” 7 JUILLET 2021.

Abstract

Machine learning becomes necessary. consists in creating systems that learn or improve performance according to the data they process. It is a decision-making tool thanks to its predictive power.

In our project, we will be focusing on the analysis of of sentiment in social networks, more specifically on the Twitter platform, in the context of the coronavirus pandemic. Our main objective will be to determine the emotional tone of users' discourse by classifying their messages into three main categories: positive, neutral and negative .

We will use machine learning and natural language processing techniques to classify tweets. We will combine Long Short-Term Memory (LSTM) model with Elephant Herding Optimization algorithm (EHO).