

Ministry of Higher Education and Scientific Research

University of Kasdi Merbah -Ouargla-

Faculty of Modern Information and Communication Technology

Computer science and information technology department



Academic Master Thesis

Field: Computer science and information technology

Branch: Computer science

Specialty: Fundamental computing

Theme

Sentiment Analysis in Arabic Tweets

Supervised by:

Mrs. W. SAADI

Presented by:

BERRIM Israr

2019/2020

Thanks

First, we thank Allah for giving us the knowledge, patience and health to be able to carry out this graduation project.

I would like to express my warm thanks to my promoter Mrs. W. SAADI for her entire disposition, and her judicious advice, and for having directed me all the semester. I would also like to thank the members of the jury for their precious time devoted to study our work.

My thanks and gratitude to our teachers who throughout the years of study have passed on their knowledge to us.

My thanks also go to all who have contributed directly or indirectly to the development of this work.

I also thank my family and friend for their support.

Dedication

This modest work is dedicated to

My parents

Whom affection, love, encouragement, and prayer of day and night
make me able to get such success and honor.

And to my dear, HARROUZ Mouad.

Who has been a constant source of support and encouragement
during the challenges of graduation

Whom without I would never reach this achievement

I am truly thankful for having you in my life

Table of contents

Table of contents	I
List of figures	IV
List of tables	V
General Introduction	1

Chapter 1 Social Big Data

I.1. introduction	3
I.2. Big data.....	3
I.3. Big Data characteristics	4
I.3.1. Volume.....	4
I.3.2. Velocity.....	5
I.3.3. Variety	5
I.4. Big data analytics.....	7
I.5. Big data frameworks.....	9
I.6. Application field of big data.....	11
I.7. Social big data	13
I.7.1. Social Media.....	14
I.7.2. Most popular social media.....	14
I.7.3. Challenges of using social big data	17
I.8. Conclusion.....	17

Chapter 2 Machine Learning

II.1. Introduction.....	18
II.2. What is machine learning?.....	18
II.3. Why machine learning ?	19
II.4. Machine learning methods	21
II.4.1. Supervised learning.....	22
II.4.1.1. Classification	23
II.4.1.2. Regression	27
II.4.2. Unsupervised Learning	28
II.4.3. Reinforcement Learning.....	29

II.4.4. Deep Learning.....	31
II.5. Conclusion	32

Chapter3 Sentiment Analysis in Social Media

III.1. Introduction	33
III.2. Sentiment analysis	33
III.3. Sentiment analysis levels	34
III.4. Sentiment analysis applications	35
III.5. Sentiment analysis approaches.....	37
III.5.1. Lexicon-Based approach.....	37
III.5.2. Machine learning approach.....	37
III.5.3. Hybrid approach	38
III.6. Sentiment analysis challenges.....	38
III.6.1. Natural Language Processing	39
III.6.2. Arabic Natural Language Processing.....	40
III.7. Social media Sentiment analysis.....	42
III.8. Related works.....	43
III.9. Conclusion.....	49

Chapter 4: Conception and Implementation

IV.1. Introduction.....	50
IV.2. Working environment	50
IV.2.1. Hardware environment.....	50
IV.2.2. Software environment and library.....	50
IV.2.3. Learning preprocessing.....	53
IV.3. Global Architecture.....	55
IV.4. Data Collection	56
IV.5. Data annotation.....	57
IV.6. Data preprocessing for training.....	57
IV.6.1. Arabic language preprocessing.....	57
IV.6.1.1. Text Normalization.....	57
IV.6.1.2. Text Tokenization	58

IV.6.1.3. Text Stemming	58
IV.6.1.4. Diacritics.....	59
IV.6.2. Preprocessing validation.....	62
IV.6.2.1. Creating the dataset.....	62
IV.6.2.2. Managing the lack values	63
IV.6.2.3. Database generation.....	63
IV.7. Analysis and Results.....	64
IV.7.1. Classification models.....	65
IV.7.1.1. Neural Network (NN).....	65
IV.7.1.2. Random Forest (RF)	66
IV.7.2. Presentation of our platform features.....	68
IV.7.3. Results.....	71
IV.7.3.1. Testing database parameters.....	71
IV.7.3.2. Classification.....	73
IV.7.4. Evaluation of our method.....	76
IV.8. Conclusion	77
General Conclusion	78
Bibliograph	79
Webography	88

List of figures

Figure I.1: The 3Vs of big data	6
Figure I.2: The structure of big data framework	9
Figure I.3: The conceptual map of social big data	16
Figure II.1: Machine learning methods classification	22
Figure II.2: Decision Tree Classification example.....	24
Figure II.3: Random Forest Classifier explication.....	25
Figure II.4: Artificial Neural Network layers.....	27
Figure II.5: Reinforcement learning, agent and environment interactions.....	30
Figure II.6: Deep neural network architecture.....	31
Figure IV.1: Global proposed architecture.....	55
Figure IV.2: The platform home page.....	69
Figure IV.3: Adding new sentences and annotations.....	69
Figure IV.4: The sentiment analysis page.....	70
Figure IV.5: Generate dataset page.....	70
Figure IV.6: Pretreatment page.....	71
Figure IV.7: Database classification depending on the number of sentence per annotation....	72
Figure IV.8: Database classification according to the number of words per sentence.....	73
Figure IV.9: Precision of NN Classifier.....	74
Figure IV.10: Precision of RF Classifier.....	75
Figure IV.11: Comparison between NN and RF classifiers.....	76

List of tables

Table III.1: Classification of the applied sentiment analysis approaches.....	48
Table IV.1: Evaluation queries and results.....	77

General Introduction

Big Data is the art of managing and exploiting large volumes of data, the term is used to describe the huge amount of data that is generated by organizations. The bulk of big data generated comes from many resources; one of the primary resources is social media.

The exponential growth of social media could facilitate interactions between users and create many forms of expression. Social media platforms like Facebook and Twitter are becoming one of the most popular tools for people to express their thoughts, emotions, reviews and feedback. Moreover, social sites and virtual communities platforms donate a large and huge data, which is called Social Big Data. It can be useful in many different and important aspects in high and accurate professional domains in the manner of researches.

Sentiment analysis (SA) or opinions mining (OM) is an attractive research subject because knowing a person's positive or negative feelings from the text he/she writes is an important phase in information-gathering and decision making. Sentiment analysis identifies the emotional tone behind a set of words or a body of text. It is an approach to natural language processing, which involves the use of data analyzing, machine learning (ML) and artificial intelligence (AI) to analyze texts for sentiment and subjective information. This field importance is present in various domains and applications.

Several works have been proposed in order to overcome the many challenges faced by sentiment analysis. This task is further complicated when it is applied to social media data that is known to be highly informal and noisy. In the other hand neither the different human languages processing is easy.

Arabic language is one of the most hard and complex languages around the world, due to its nature and characteristics it represent big challenge for sentiment analysis field. One of the most important challenges for sentiment analysis on Arabic is the different types of the Arabic language. It can be classified with respect to its morphology, syntax and lexical combinations into three different categories: Classical Arabic (CA), Modern Standard Arabic (MSA), and informal Arabic or Dialectal Arabic (DA).

This last is the most used on social media. Each Arab country has its own dialect, every specific dialect makes changes and variations whether in lexical, phonology and morphology, which make the task of sentiment analysis becomes more sophisticated. This motivates us to explore

the challenges to analyze tweets written in both modern standard Arabic (MSA) and Algerian Dialectal Arabic.

The Algerian dialect or *Darja* is very complex mainly due to its morphological complexity and the limited availability of software compatible with the Arabic language. It is originated from the classical Arabic and it has been mixed with Berber and French. An accent of Darja differs in almost all Algeria regions; Darja is used throughout the country with some differences according to the region. Moreover, the Algerian social media users post, comments or tweets mostly in Darja.

The objective of this work is to create a platform that includes a community system, which it can make it possible to make a classification of the polarity (Positive/Negative) of the Algerian Dialectal Arabic tweets. In addition, this project aim to add a new powerful learning and classification model in the sentiment analysis field, which is *Random Forest Classifier*, also to improve the analyzing of live tweets and to reach the ability of making live learning.

This work is organized as follows:

- Chapter1: We introduce the big social data by defining big data with its main characteristics and by explaining the correlation between social media and big data.
- Chapter2: We present the machine learning and mention its importance. Beside, to explain some of the most popular and useful machine learning algorithms.
- Chapter3: We institute a definition of sentiment analysis on social media data with its mains different approaches and challenges in natural language processing.
- Chapter4: We present the architecture of our database with introducing our classifiers. In addition to an illustrating the choice of methods by running the tests and discussing the results.
- Finely a conclusion is made in the context of this work.

Chapter 1

Social Big Data

I.1. introduction

Around the world, there are 3.5 billion social media users which equates to about 45% of the population according to a statistics elaborated by Emarsys,2019 WEB [1]. This wide number of users interact between one another in different ways such as, likes, comments, tweets, views, favorites, sharing digital pictures and videos and everything else can user do and interact with in any social media platform. Those interactions represent unstructured/semi-structured social data generated by a large number of internet applications and websites, the sheer volume and semantic richness of such data open enormous possibilities for utilizing and analyzing it for personal, commercial, researches, as well as societal purposes.

On account of the huge increase of using social media in recent days, social media has become one of the most representative and relevant data sources for big data.

Social big data A [1], comes from joining the two domains: social media and big data, where social big data will be based on the analysis of vast data that could come from multiple distributed sources but with a strong focus on social media. Social big data is mainly utilized to extract ideas from social media data and online social interactions for descriptive and predictive purposes to influence human decision making in various application domains.

In this chapter, we will focus on the representations and the definitions of different concepts related to big data and social big data.

I.2. Big data

Big data is a term for any large and complex collection of data, which becomes difficult to process using traditional data processing applications. In fact, big data is often associated with the processing of very large volumes of data, it is not only the quantity that is decisive. However, the association in the same analysis of varied data in order to deduce information, which is impossible to highlight with the classical analyzes of structured data.

The analyzing of millions of records stored in a traditional data warehouse is not big data. It is a voluminous analysis. Otherwise, combining customer consumption information stored in a data warehouse, with web browsing logs and call center discussions records, to try to anticipate a customer's departure, regardless of the size of the information, sources can be called big data B [1].

For many years, the concept of big data took hold: the collection of data, which has been focused on the 3V's of big data that are used to classify data as big data. This 3V model has been defined in 2001, by Doug Laney A [2].

The combination of data collected from various different sources some of the potential ones are : Open Science Data Cloud, Web Services, public dataset on Amazon, machine data, healthcare data, transaction data, social media and many others, then processing this data then using the result obtained is called Big Data analytics.

The aim of this popular phenomenon is to provide an alternative to traditional solutions based on databases and data analysis. It is not just about storage or access to data, but it is for analyzing data in order to make sense of them and explore their values. This analysis spans areas such as big data computing, data mining, statistics, and natural language processing B [2], machine-learning B [3], and numerous others. It has different application domains like healthcare, manufacturing, media and entertainment, government, finance, e-commerce and many others.

I.3. Big Data characteristics

An interesting definition of Big Data has been given by Doug Laney in 2001, he presented the theory of 3V as follows: *“high-volume, high-velocity and high-variety information assets that demand cost-effective, innovation forms of information processing for enhanced insight and decision making”* A [2]. Then more recently in 2012 the definition was updated by Gartner as: *”big data is high volume high velocity and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization”* A[3]. Both definitions refer to the three basic features of big data Volume, Velocity and Variety.

I.3.1. Volume

The term volume refers to large amounts of any kind of data generated through any different sources, it defines the data infrastructure capability of an organization's storage, management and delivery of data to end users and applications. Volume focuses on planning current and future storage capacity but also in reaping the optimal benefits of effectively utilizing a current storage infrastructure.

The volume of data is increasing at a phenomenal rate, data growth is passing our ability to decipher it. Digital Universe IDC A[4] conducted a study that revealed that data around the world doubles in size every two years. What's more important is that 3% of the data is organized, with only 0.5% ready for analysis.

The majority of big data is unstructured, and its volume is so large that processing it using traditional databases and software techniques is difficult, if not impossible. One of the technologies commonly and recently used for facing this problem is the framework Hadoop B [4]. We will get to mention the most used big data frameworks in the next sections.

The benefit from gathering, processing and analyzing these large amounts of data generate a number of challenges in obtaining valuable knowledge for people and companies.

I.3.2. Velocity

The velocity or the speed of data flow has experienced a similar evolution to that of the volume.

A big data system is a system in which data flows quickly between tools, databases, applications and sources, data arrives in the system from multiple sources and is often processed in real time to generate insights and update the system. In big data, "batch" is the oriented approach, which is gradually tending to give way to real-time data streaming. Increasingly, data is added, treated, processed and analyzed in real time.

For the user service, the velocity is decisive in situations where it is a question of taking into account in real time: the wishes expressed by the user and the state of availability in stock, to provide the best possible services. From this point of view, new algorithms and methods were needed to adequately process and analyze the online streaming data, artificial intelligence B[5] and machine learning have the right high potential.

I.3.3. Variety

Big data is also characterized by the huge variety of data processed, relational databases B[6] deal with structured data of the same type. In big data, the majority of data is unstructured or semi-structured. Moreover, for this reason, they must be worked and long prepared.

Variety refers to the different types of structured or unstructured collected data via sensors, organizations or social networks, such as text and numeric, videos, images, audio, log files, and

so on. Each data form has its own type of uniqueness in terms of how it is classified and stored on a cloud. What makes this unique format is how we can analyze them to create valuable solutions.

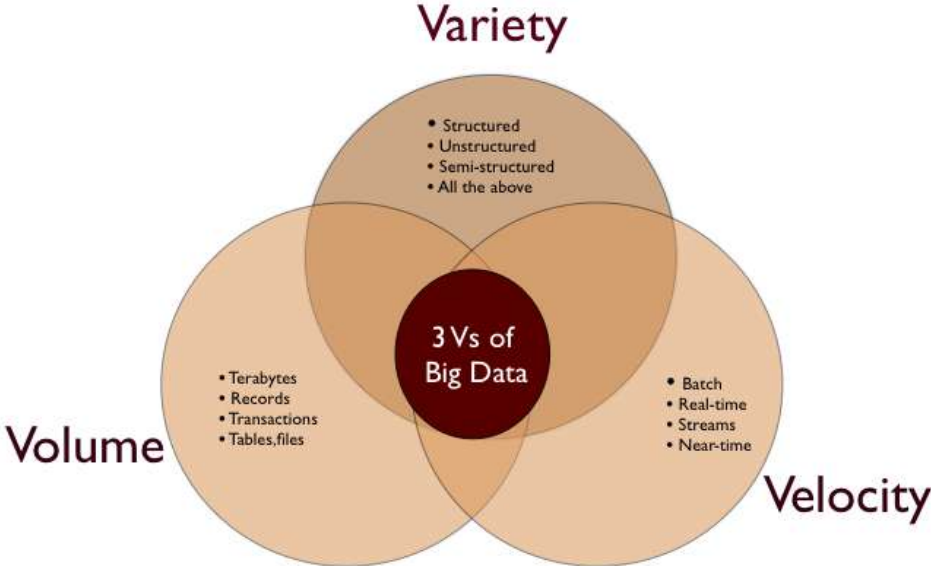


Figure I.1: The 3Vs of big data B[7].

Big data is the art of managing large volumes of data, complex and heterogeneous, for the most part unstructured, which circulate quickly in a given system that is not within the reach of conventional database management software.

In addition to the three proposed by Gartner, other organizations and software wanted to add other "V" to highlight other challenges posed by big data, they have extended this 3V model to 4V model by including a new "V", then extending this last to 5V model then to 6V.

The new 3 including "Vs" may take:

- **Value:** is the most important characteristic of any big-data-based application, because it allows us to generate useful information A [1]. It refers to the process of extracting valuable information from large sets of data, because the ultimate challenge of big data is to create value.
- **Veracity:** aiming at data integrity and the ability of the organization to use confidently the data, the variety of sources and the complexity of processing can pose problems with

regard to the evaluation of data quality, the data quality issue is structuring in any big data project.

- **Variability:** data arrives constantly from different sources and how efficiently it differentiates between noisy data or important data. The variability of the data leads to a variation in their quality.

Ripon and Arif in A [5], proposed a conjugation for the previous 6V model of big data with a 1C, this 1C is the complexity. In another study of big data characteristics, Gayatri, Alka and Raees in A [6] had largely mentioned most of all the proposed Vs and C of big data models, with a brief description for each one of them.

I.4. Big data analytics

Data Analytics B [8], is a science of examining data through a process of cleansing, transforming and modeling data with the aim of discovering useful information and unknown patterns. Where this information will be used to suggest conclusions and support decision-making. A powerful analytics tool can be used to achieve the steps of this process involving the application of specific algorithms to extract models from data. Data analytics purposes split to predictive, descriptive, prescriptive and diagnostic aims.

Big Data Analytics is where advanced analytic techniques operate on big data. It is a process where extracting the knowledge and the useful information is from a large amount of different kinds of data (structured, semi-structured, unstructured). This process is complex, it involves many special steps and phases including data acquisition and recording, information extraction and cleaning, data integration aggregation and representation, query processing, data modeling and analysis and interpretation. Each phase separately represents a collection of different techniques and methods, which aim to improve the analysis results. Depending on the provided results of each phase, some phases could be changed, adjusted and repeated B [9].

Big data analytics is one of the most vital aspects that is driving some of the biggest and best companies forward today. It is used in many industries to enable businesses and organizations to make better decisions. In the scientific and technical field, scientists and engineers face big data notably generated automatically by specific programs users or sensors or measuring

instruments or many other, thus, big data analytics is used to verify theories or to refute existing models. Therefore, the importance of big data analytics is represented by offering various business and research benefits.

The nature of big data analytics depends on the nature and the structure of the data, which implement various algorithms related to data mining B [10], machine learning, decision support, even visualization.

In the context of the use of machine learning for big data analysis, machine learning consists of automatically determining a formal model, describing the available data and allowing a certain level of generalization on new data. Where the objective of a learning method is to determine the model which minimizes generalization errors, i.e. which allows to obtain the most exact classification possible of new data. In general, machine learning is a compelling tool that is emerging as a solution for managing large amounts of data, especially for making predictions and providing suggestions based on the data sets.

Machine learning supports big data analytics implementation. If without machine learning to mine ever-growing massive data, big data analytics would be impossible. All other components within a framework of big data aim to support the machine learning process. Consequently, machine learning is the centerpiece of any big data analytics A [7].

As we mentioned, the data concerned about big data is very diverse, due to their nature and / or their level of structuring, their analysis will also be different. Big data analytics divide the analysis of stored big data and the analysis of exchanged big data, which is not possible to store due to their volume. This last relates to data exchanged and transmitted continuously, like streaming data from online media, data from sensors, or physics experience reports. When it comes to requesting a continuous, fast and endless data flow, it is not possible to query the entire flow, which could have the consequence of stopping the flow. The most used techniques in the search of data flow, do not explore the entire stream, but query selected data in the stream. These techniques work on a set restricted flow to extract the various patterns or items needed.

The storage and the operation on big data requires data partitioning but also a distribution of the treatments, the algorithms, the necessary access and the management of this data. The analysis of large stored data requires the implementation of processes and algorithms, In particular machine learning and data exploration, which must also be distributed to be efficient.

Therefore, several standard machine learning and data mining algorithms, such as matrix factorization algorithms, classification algorithms or even categorization algorithms, had distributed versions through the many recently developed big data platforms and frameworks.

I.5. Big data frameworks

Massive data arrays must be reviewed, structured, and processed to provide the required bandwidth. Nowadays, there’s probably no single big data software that wouldn’t be able to process enormous volumes of data, therefore, data processing engines are getting a lot of use in tech stacks for mobile applications, and many more.

Under the previous title we will get to mention the benefit of big data frameworks, their structure and briefly demonstrate some of them. Big data frameworks have been created to implement and support the functionality of such software; they help rapidly process and structure huge chunks of real-time data.

A big data framework identifies core and measurable capabilities in each of its six domains represent the structure of the big data framework, so that the organization can develop over time, which are depicted in the figure below.

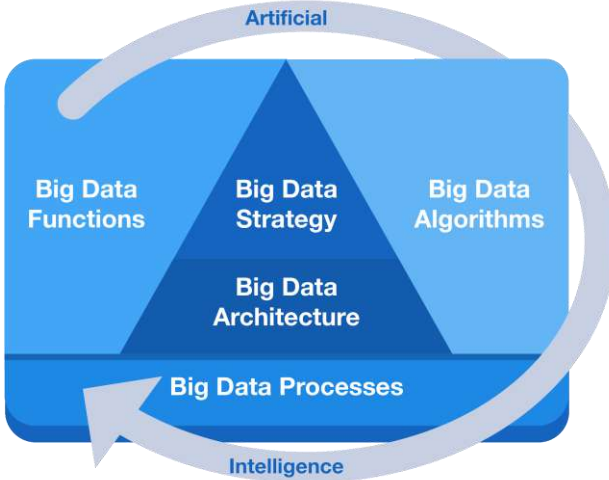


Figure I.2: The structure of big data framework WEB[2]

- Big data strategy: it's about how return on investment be realised, and where to focus effort in big data analysis.
- Big data architecture: it discusses the various roles that are present within Big Data Architecture and looks at the best practices for design.
- Big data algorithm: it aims to build a solid foundation that includes basic statistical operations and introduction to different classes of algorithms.
- Big data process: it brings structure, measurable steps and can be effectively managed on a basis; processes embed big data expertise within the organization by following similar procedures and steps.
- Big data function: it addresses critical success factors for starting a Big Data project in the organization.

There are many great Big Data frameworks on the market right now, most popular like Hadoop, Storm, Hive and Spark, most promising like Flink and Heron, most useful like MapReduce and Presto, also most underrated like Samza and Kudu.

A brief description of five most powerful and popular Apache big data frameworks in follow:

- **Apache Hadoop** WEB[3]: The Apache Hadoop software library is a framework that allows the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each one offering local computation and storage.
- **Apache Storm** WEB [4] : Apache Storm is a free and open source distributed realtime computation system. Apache Storm makes it easy to reliably process unbounded streams of data, doing for real time processing what Hadoop did for batch processing. Apache Storm is simple, can be used with any programming language.

- **Apache Samza** WEB[5]: Samza allows to build stateful applications that process data in real-time from multiple sources including Apache Kafka, while Kafka can be used by many stream processing systems, Samza is designed specifically to take advantage of Kafka's unique architecture and guarantees. It uses Kafka to provide fault tolerance, buffering and state storage.
- **Apache Spark** WEB [6]: Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine. It provides high-level APIs like Java, Scala, Python and R. It is 100 times faster than Big Data Hadoop and ten times faster than accessing data from the disk.
- **Apache Flink**: Apache Flink is an open source platform. It is a streaming data flow engine that provides communication, fault tolerance and data distribution for distributed computations over data streams. Flink is a scalable data analytics framework that is fully compatible with Hadoop. It can execute both stream processing and batch processing easily.

I.6. Application field of big data

Big data applications have introduced cutting-edge possibilities in every aspect of our daily life. We are living in a world of tremendous competition, as there are many various fields of big data application, they represent the picture of the impact of data in our lives.

Likewise big data applications have made our life better and smooth as well, further, we are clearly using big data for increasing our efficiency and productivity. Just bellow, we will bring up the benefits and competitive advantages provided by big data applications on some of the various fields in nowadays:

- **Big data applications in Healthcare**: the role of big data in medical was not mentionable before, but data science is dominating to improve healthcare nowadays. The clinical field could benefit from the large amounts of biological and clinical patient data, which could be generated and collected for making fast and intelligent decisions. Also big data has a great impact on reducing waste of money and time. A Literature review had been made by Jake, Deepika and Yiqing A[8] on big data application in biomedical research and healthcare.

- **Big data application in Ecommerce:** Ecommerce faces many challenges to achieve the business objectives. Ecommerce businesses can benefit the most from using big data, because of all the information they collect during day-to-day operations. This information is a large collection of data that organizations can use to make better decisions. Big data in e-commerce can provide competitive advantages by providing insights and analytical reports.
- **Big data in marketing:** Big data has made the marketing powerful, where it has become an essential part of any business. In marketing, big data is providing insights into which content is the most effective at each stage of a sales cycle and strategies for increasing conversion rates, in addition to prospect engagement and customer lifetime value. Alexander, Paulo, Paulo and Sergio A [9] had given the interest of research on big data in marketing in a text mining and a subject modeling literature analysis.
- **Big Data application in the government sector:** the application of big data can leave an enormous impact on this sector by collecting all the information about millions of people that helps to take any decision considering locals. Gang, Silvana and Ji-hyong in A[10] explained how the same way businesses use big data to pursue profits, governments use it to promote the public good.
- **Big data application in education:** a large amount of data is generated in schools and higher education, it can be used for understanding performance and behavior patterns of students, developed countries are recently acquiring new techniques as well, the work A[11] could briefly mention what big data application in education could be used for.
- **Big data application in the social media sector:** The platform of social media marketing completely depends upon the application of big data, because by the analysis of big data, marketers can identify the latest trends in the social media sector and take better and quicker decisions accordingly. Big data allows the analysis of the behavior of users, the more information obtained about consumers, the better way to target them through social media campaigns.

The list of big data applications in our recent days is more various, we can also mention the big data application in banking sector, in tourism, in media and entertainment, in disaster management, to ensure national security, in agriculture, in cloud computing, to provide customer oriented services, in telecommunication, and much more. In addition, the most popular achievement made by big data application in different field in short are A[11] :

- Bank of America Merrill Lynch: creates practical and effective solutions for clients based on a more comprehensive and holistic understanding of their requirements.
- British Airways' Know Me Program: uses the data collected to get a better insight into personal preferences and buying patterns of its frequent flier.
- Google: tunes algorithms in language processing to be culturally relevant (for instance differentiating between American and British idioms) and improving its speech recognition capabilities.
- Apple: granted a patent to collect data on body temperature and heart rate through audio buds.
- Facebook: recently started to decode the content of photographs (identifying faces and objects) and video.
- IBM's Deep Thunder weather analytics package: helps farmers know when to irrigate their crop.

I.7. Social big data

Social media and big data are two phenomena that cannot be more current. Social media generates massive data that forms the big data. This is called Social Big Data. The work A[1] considered social big data as a combination of big data and social media. According to the authors, social big data is needed for analysis of large amounts of data from diverse social media

sources. They defined the concept as follows: *“Those processes and methods that are designed to provide sensitive and relevant knowledge to any user or company from social media data sources when data sources can be characterized by their different formats and contents, their very large size, and the online or streamed generation of information”*.

According to the definition and the characteristics of big data, social big data is a high-volume, high-velocity, high-variety data that is generated from social media users interactions and actions through social media platforms. This data can be collected and analyzed to model social interactions and behavior in wide applications such as: trend discovery, social media analytics, sentiment analysis, and opinion mining.

I.7.1. Social Media

The Internet is a real medium for exchange and sharing, people are now exchanging information through platforms known as social media. Social media is not to be confused with a social network, the social network designates a platform for exchanges between Internet users in the form of a network, while social media brings together all of the exchange platforms including participative sites, such as wikis, forums and blogs. Social media includes all social networks. It therefore defines the websites allowing connecting internet users who will communicate exchange and form communities to socialize, or to achieve a common goal or interest, via different tools.

Over the past two decades, internet users connections have become increasingly subjoined by online media, beginning with message boards, progressing to email and instant messaging, and now blooming into online social media sites and apps. Social media has four major potential strengths: collaboration, participation, empowerment, and time, its technologies allow users to immediately publish information in near-real time A[12].

I.7.2. Most popular social media

At this age of digitization, and with the advent of the numerous social media platforms and apps, now it is possible for people to connect and network with each other, by being socially active on the internet through one of the many different tools. Besides the increasing number of social media users, and the growing popularity of mobile social media users, social media platforms and apps are going to grow even bigger as people adopt them into their everyday lives, as follow by the order of the most popular and used social media:

- **Facebook** : The largest social networking site in the world and one of the most widely used, apart from the ability to network with friends and relatives, you can also access different Facebook apps to sell online and you can even market or promote your business, brand and products by using paid Facebook ads. Facebook empowers more than 2 billion people around the world to share ideas, offer support and make a difference WEB [7].

Sharing opinion by posting texts or pictures, videos and stories, reacting on those posts through adding comments or choosing one of the emoji faces, connecting when sending messages or voice memos. All this represents rich and large big data, which can be used for analysis and research in various domains.

- **Instagram** : Instagram is a photo and video sharing social media app. It allows you to share a wide range of content such as photos, videos, stories and live videos. It has also recently launched IGTV (Instagram Television) for longer-form videos, Instagram is now part of the Facebook empire, more than 1 billion users, as a consequence large amounts of big data is there for analyzing and studying.

- **Twitter**: Twitter is a social media site for news, entertainment, sports, politics, and more, this social networking site enables you to post short text messages (called tweets).

What makes Twitter different from most other social media sites is that it only allows 280 characters in a tweet (140 for Japanese, Korean, and Chinese), unlike most social media sites that have a much higher limit, more than that Twitter has a strong emphasis on real-time information, things that are happening right now.

- **YouTube**: YouTube is a free video sharing website that makes it easy to watch online videos, it allows users to create and upload their own videos to share with others. YouTube's mission is to give everyone a voice and show users the world WEB [8]. Anyone with access to a computer or mobile device and an internet connection can watch YouTube content and share their own.

YouTube was created in 2005, it is now one of the most popular sites on the Web, with visitors watching around 6 million hours of video every month.

Signify that data analysis is the set of algorithms and methods used to extract and analyze knowledge from the social big data, and with the previous definition of social big data, which explain how social media is the main source of big data and the massive parallel processing paradigm of social big data. **Figure I.3** summarize the relation between the previous concepts as three basic social big data areas A [1]:

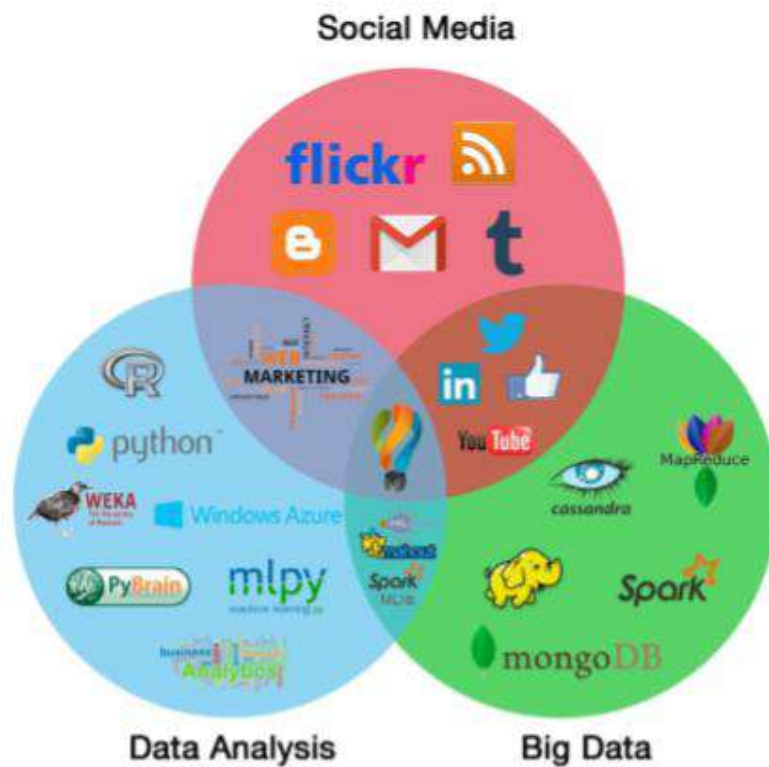


Figure I.3: The conceptual map of social big data A[1].

Using social big data in research is trendy, because of the richness and variety of this data, which refer to many difficulties and challenges of collecting and using it. The following title presents a brief explanation of those different difficulties and challenges of using social big data.

I.7.3. Challenges of using social big data

Social big data is large and various, as we mentioned previously, it can be texts, images, videos, and much more. Usually this data is unstructured or semi-structured, it is noisy and sometimes uncompleted, many other problems researchers face while using social media data.

We can briefly allude to: dispassionate data (noisy texts) colloquial and personal elements in users language, the uncompleted pictures or images, low data relevance and quality, moreover exploring who the social media users are, geographic and biases, the location information for many posts may be locally incorrect or completely missing, self-selecting users and ethical concerns. In-addition filtering out spatially irrelevant or biased observations, expensiveness of data acquisition because social media data is not always available and not all social media platforms allow to use any of their data, also challenges in preparing, processing and cleaning the data. Finally, there are many ethical issues, such as privacy, free speech, data leakage, and revealing users' identities

I.8. Conclusion

Research and business in social big data is increasing, because of the popularity of social media, which allow users to share and exchange every detail of their lives in almost real time, in various ways and different platforms and applications, therefore, this data is large amount (volume), fast and speedy (velocity) and various (variety).

In the previous chapter we defined big data with its main characteristics and big data analytics, also we had briefly referred to the most common big data frameworks. Social Big data as well was briefly presented.

Social big data analytics has faced many challenges and difficulties. Several techniques allow using social big data in various fields, but because of the automation, wide applications and the many other advantages of machine learning, social big data analytics A[13] has become much simplified and much easier. The next chapter contains more details about the different used techniques and algorithms of machine learning.

Chapter 2

Machine Learning

II.1. Introduction

The manifestations of intelligence are generally in the understanding of a situation and problem solving in particular cases, when the ability of solving problems is sufficient to offer various solutions to the considering problem. Intelligence also appears in the decision, which is the selection of the most appropriate solution for the current problem.

The machine could solve problems, but *Could the machine learn how to understand a particular problem and take the right decision for solving it ?* and *How close could machine capabilities come to those of human beings ?* A conference at Dartmouth University in 1956 explored these questions and led to the appearance of the term ‘AI’ A [14] , the abbreviation of Artificial Intelligence .

At the age of “Big Data”, Artificial intelligence is also on the doorstep of the technology revolution, as technology helped us for understanding how our minds work, our concept of what constitutes AI has changed, more than increasingly complex calculations, the work in AI concentrated on human decision making processes and carrying out tasks in ever more human ways.

Artificial intelligence is concerned with intelligent behavior in artifact, which involve perception, reasoning, learning, communicating and acting in complex environment, therefore AI has two principal goals which are : the development of machine that can do these things as well as human can, or even better, and understanding this behavior whether it occurs in machines or in humans B[11].

In addition to these goals, the realization, that rather than teaching computers everything they need to know how to carry out tasks. It might be possible to teach them to learn for themselves, and recently the emergence of internet and big data, was the two important breakthrough that led to the rising of Machine Learning, which is driving AI development forward with the speed it currently has. This chapter will define machine learning with its most important benefits and methods.

II.2. What is machine learning?

Generally, Machine Learning B [12] is a recent application of AI that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. The idea is to give machine access to data and let them learn for themselves, it is

based on the understanding of the machine to the structure of data and turn that data into models that can be understood and utilized by people.

Machine learning is a modern science (field within computer science) differs from traditional computational approaches; it is for discovering patterns and making predictions from data based on statistics, data fodder, pattern recognition and predictive analytics. Machine learning algorithms authorize computers to train on data inputs and use statistical analysis in sequence to output values that fall within a specific range. As a result of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs. It is very effective in situations where insights need to be discovered from large, diverse and changing data sets.

The iterative aspect of Machine Learning is essential, as it allows models to adapt independently when exposed to new data, they learn from previous calculations to create reliable and repeatable decisions and results. Arthur Samuel (1959) A[15] defined it as follows: “*Machine Learning is a field of study that gives computers the ability to learn without explicitly being programmed.*”. He had made a Checkers playing program which could learn over time, at first it could be easily won, But over time, it learnt all the board position, thus became one of the best chess players.

There are typically three phases of machine learning. The training phase where the training data is used for training the model by matching the given input and the expected output, then testing phase here the learning model is measured by the quality and the estimation of the properties as error, recall and precision. Finally the application phase, where the model is exposed to real world data for which results need to be derived.

Under the next title, we will answer the question why every technology user today has benefitted from machine learning, by setting the most important benefits and advantages of machine learning with brief explanation.

II.3. Why machine learning ?

Machine learning A [16] extracts expressive information from raw data and provides specific results, by processing this data and by learning from it, this information helps in solving complex and data-rich problems. Besides, machine-learning algorithms use this technique to find different insights and data features without being programmed to do so.

Recently, many top-ranked companies like Google, Amazon, and Microsoft have adopted machine learning in their business, machine learning and Artificial Intelligence have created a lot of buzz not only in the business sector, hence, one of most important advantages of machine learning is :

- **Wide application of Machine Learning:** Machine-learning applications go far beyond computer science; many other industries stand to benefit from it. Learning algorithms are recently adopted in almost all science and development fields, they are used in a wide variety of applications such as: banking and financial sector, healthcare, retail, publishing, and business where Google and Facebook are using machine learning to push relevant advertisements , which are based on users past search behavior B [13]
- **Pattern and trends recognition B[14]:** machine learning can investigate large volumes of data and recognize specific trends and patterns that would not be apparent to humans, it indicates the use of powerful algorithms for identifying the regularities in the given data, which is used in the new age technical domains like computer vision B[15], speech recognition and face recognition
- **Automation of Machine Learning:** since machine-learning means giving machines the ability to learn, it lets them make predictions and enhance the algorithms on their own. After, feeding the algorithm with whether known or unknown data, this last will learn about this data and extract specific results, therefore with machine learning human intervention is not always needed.
- **Permanent Improvement:** by time machine-learning algorithms will make better decisions, which happen by gaining the experience from the continuous improvement in precision and efficiency. In another word, the learning is continual, the machine can update the prediction model, but it still can also reuse and retrain the useful knowledge skills during time.

- Multi-dimensional and Multi-variety data: machine-learning algorithms may generate unexpected population groupings, they are used to handle multi-dimensional and multi-variety data in a dynamic environment.
- Machine learning is useful where large-scale data is available, the large-scale deployment of machine learning beneficial in improving speed and accuracy.
- Understanding no-linearity in the data and generating a function mapping input to output, therefore, machine learning is recommended for solving classification and regression problems, which are the two approaches of supervised learning.

II.4. Machine learning methods

Depending on how learning is received or how feedback on the learning is given to the system developed, machine-learning methods are divided into categories, in general: Supervised Learning and Unsupervised Learning, which are further classified into methods. Classification and Regression for the supervised learning, clustering for the unsupervised learning. Additionally machine-learning methods include Reinforcement Learning and Deep Learning, each method gives a particular result, and utilizes various forms of data, also has a specific purpose and action. Machine learning methods are classified as the figure below clarify:

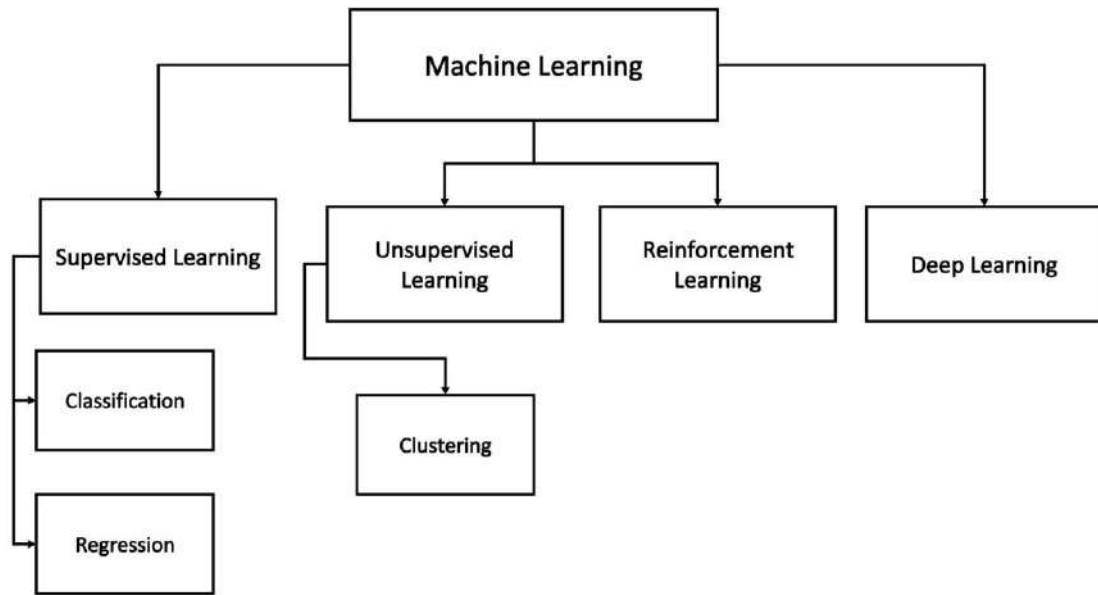


Figure II.1: Machine Learning methods classification

II.4.1. Supervised learning

Supervised learning is where the algorithms training is based on example input and output data that is labeled by humans, which means the computer will be provided with examples input labelled with the desired output, for the aim of making the algorithm able to learn. The input data goes through the supervised algorithm, where it is used to train the model. Once the model is trained based on the known data, the unknown data can be used into the model and get a new response. The algorithm will learn by comparing its results with the giving output, so it can find errors, and modify the model accordingly.

Supervised learning is used whenever the aim is to predict a certain output from a given input, with taught examples of input/output pairs; the goal is to make accurate predictions for new, never-before-seen data. Supervised learning often requires human effort to build the training set, but afterward automates and often speeds up an otherwise laborious or infeasible task B [16].

Some of the popular supervised learning applications are Natural Language Processing, Image classification, predictive analysis, pattern recognition, spam detection, Speech/Sequence processing and Sentiment analysis (On social media).

Supervised learning uses classification algorithms and regression techniques to develop it different models and algorithms, therefore it is further divided into Classification and Regression:

II.4.1.1. Classification

Classification is when the computer program learns from the input data given to it as supervised learning, and then uses this learning to classify new observations, where the process is to find the features that will help to separate data into different classes.

The goal is to predict a class label, depending on the training feature, the class label will be represented as a choice from a predefined list of possibilities in the multiclass classification, such as in speech recognition, handwriting recognition, biometric identification, document classification.

Or between exactly two classes in binary classification, like in whether the mail is spam or non-spam, it's like trying to answer the yes/no question, or in identifying if the person is male or female, also in sentiment analysis by detecting if the text is positive or negative.

The classification algorithms are in types we are mentioning: Linear Classifiers, Naive Bayes Classifier, Nearest Neighbor, Support Vector Machines B [17], Decision Trees, Boosted Trees, Random Forest and Neural Networks, in this section we will clarify these two last algorithms and explain how they work.

- **Random Forest:** Random Forest Classifier is a supervised classification algorithm B [18], which can be used for regression and classification problems. This classifier is considered as an overall classifier, that it combines more than one classification algorithm of the same or different type to classify objects.

As the name suggests, it implicate creating a set of decision trees, the higher the number of trees in the forest, the more precise the results obtained, each decision tree is a single classifier and the target prediction is based on the majority voting method, therefore, we will first examine how basic decision trees work :

Decision tree A [17] builds models in the form of a tree structure, it decomposes the data set into smaller and smaller subsets while concurrently an associated decision tree is incrementally developed. The result is a tree with decision nodes and sheet nodes, a

decision node has two or more branches and a sheet node represents a classification or decision. **Figure II.2** represents a simple example of decision tree classification of whether the patient needs to be sent to the hospital or to stay home, considering breathing difficulties and bleeding as features.

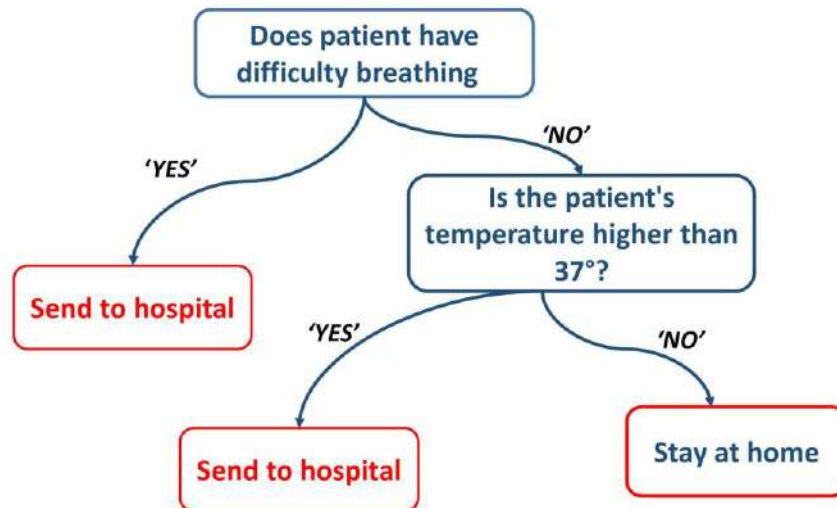


Figure II.2: Decision Tree Classification example.

Obviously the data is not always clean and easy to classify, but the logic which the decision tree works with is the same, the main purpose here is to find the feature which will allow to divide the observation at each node, thus that the resulting groups are as different from each other as possible.

Moreover, random forest classifier consists of a large number of individual decision trees that works, as an ensemble with random sampling, is a combination of Breiman's idea and random selection of features "bagging" A [18]. Where the idea was to make the prediction precise by taking average or mode of the output of multiple decision trees, the more the number of decision trees is considered the more precise output will be.

Breiman also give another definition in A [19] as “A *random forest* is a classifier consisting of a collection of tree structured classifiers $\{h(x, \Theta_k), k=1, \dots\}$. Where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x ”. This definition implicate that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The error for forests converges as to a limit as the number of trees in the forest becomes large. The reason for this great effect is that the trees protect each other from their individual errors, while some trees may take the wrong decisions; many other trees will take the right ones, so as a group of trees the forest is able to move in the correct direction.

Figure II.3 clarifies the distribution of the input data to the multiple trees of the random forest where they work individually, for getting a primary classification, then obtaining the final class depending on the majority.

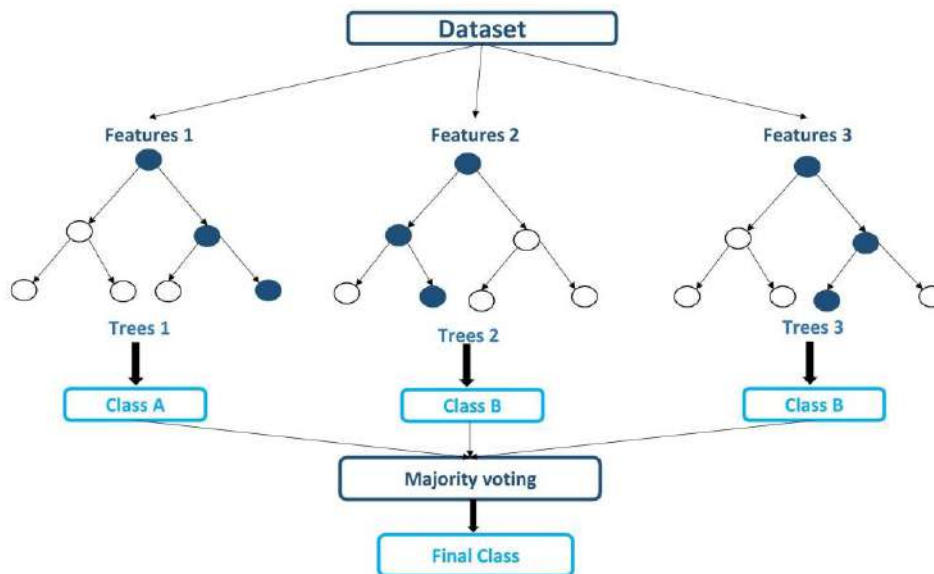


Figure II.3: Random Forest Classifier explication.

The application of the Random Forest Classifier is common in different domains; first as Weizhong A [20] elaborate on the objective of evaluating effectiveness of random forest classifiers on aircraft engine fault diagnosis, by designing a real-world aircraft engine fault diagnostic system.

Besides Grant A [21], the work was to treat the particular of how to successfully apply the random forest algorithm in a proteomics profiling study for construct classifiers and discover peak intensities most likely responsible for the separation between the classes.

- **Neural Network:** Neural Network learning algorithm, or an artificial neural network B[19] is a computational learning system that uses a network of functions to understand input data and interpret it into desired output, it consists of units or neurons, which are inspired by the biological neural in the brain of the human nervous system. The preliminary theoretical base for contemporary neural networks was independently proposed by Alexander Bain(1873) and William James (1890) B[20], In their work, both pansies and body activity resulted from interactions between neurons within the brain, realizing the potential for man-made system based on neural models.

The artificial neurons are arranged in layers, which convert an input vector into some output, each neuron takes an input, applies a function to it and then passes the output on to the next layer; each node/neuron is associated with weight (w). This weight is given under the relative importance of that particular neuron or node. The node function or activation function, purpose is to introduce non-linearity of data to the neurons, in order to learn this representation. The function f will provide output as: $y = f(wI. xI + b)$, where y is the output of the neuron with $w1$ as the associated weight and $x1$ as a numerical input, and bias b .

Generally, the networks are defined to be feed-forward with the input, it implies that a neuron feeds its output to the neurons on the next layer, but there is no feedback to the previous layer. **Figure II.4** captures the interaction between layers, knowing that there are many variations of the relationships between nodes, or artificial neurons:

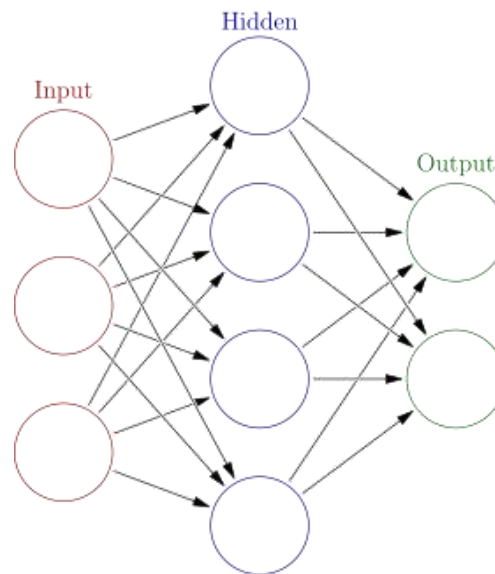


Figure II.4: Artificial neural network layers.

The neural net learning algorithm learns from processing many labeled data, which were provided during training, and use them to learn and to define what characteristics of the input are needed to construct the correct output, once a sufficient number of examples have been processed, the neural network can begin to process new, unseen inputs and successfully return accurate results.

The more examples and variety of inputs the algorithm trains, the more correct the results typically become because the program learns with experience.

Artificial Neural Networks can be applied to a variety of problems and can assess many different types of input, including images, videos, files, databases, and more.

II.4.1.2. Regression

Regression is a supervised learning technique predict continuous responses B[21]. Regression models are used to predict a continuous value such as salary, prices and weights. It is one of the most important and broadly used machine learning and statistics tools; it allows making predictions from data by learning the relationship between features of the data and observed continuous-valued output. Many different models can be used, such as, Polynomial Regression,

Support Vector Regression B [22], Decision Tree Regression and Random Forest Regression., the simplest is the Simple Linear Regression B [23].

Regression is used in a massive number of applications ranging from predicting stock prices or person age to understanding gene regulatory networks, in the same manner for business it is a very powerful statistical technique and can be used to generate insights on consumer behavior, understanding business and factors influencing profitability.

As an example, Dean A [22] suggested solving a global environmental problem, by using Simple Linear Regression to assess the success of the Montreal protocol in reducing Atmospheric Chlorofluorocarbons. He modeled the rate of change during periods that correspond to the times prior to and subsequent to implementation of the Montreal Protocol, with a simple linear equation, for predicting the atmospheric concentrations.

II.4.2. Unsupervised Learning

In unsupervised learning, the training data is unknown and unlabeled, so the learning algorithm is left to find commonalities between its inputs data. Without the aspect of known data, the input cannot be guided to the algorithm, which is where the unsupervised term originates. Unsupervised learning A [23] goal is discovering and modelling the hidden patterns, or the underlying structure in the given input data, so as to learn about the data, which is fed to the learning algorithm and used to train the model, that help to find features which can be useful for categorization. The trained model tries to search for a pattern and give the desired response.

As unlabeled data are more numerous and obtainable than labeled data, machine learning methods that facilitate unsupervised learning are particularly valuable.

Otherwise, we cannot get precise information, and the results are less accurate, because the input data is not known and unlabeled by humans, the machine requires doing this itself. Unsupervised learning is commonly used for transactional and preprocessing data, during exploratory analysis, or to pre-train supervised learning algorithms.

One of the most popular unsupervised learning techniques is clustering.

- **Clustering** : Clustering is an important technique in the unsupervised learning, It basically deals with finding a structure or pattern in a collection of uncategorized data, which signify that clustering is a process of grouping similar entities together, which

will help profiling the attributes of those different groups, by giving insights into underlying pattern of these groups, those groups name cluster. The goal of this unsupervised machine learning technique is to find similarities in the data point and group similar data points together.

Clustering algorithms will process the data, and define the existence of the natural clusters or the groups in the data. We can also control the cluster numbers that the algorithms should identify, allowing us to adjust the granularity of these groups. There are many algorithms developed to implement this technique, the most popular and widely used algorithms in machine learning are, K-means Clustering Algorithm [24], Mean-Shift Clustering Algorithm, Hierarchical Clustering, K-NN (k nearest neighbors), and many others.

II.4.3. Reinforcement Learning

Richard in B [25], defined reinforcement learning as the learning of a mapping from situations to actions, to maximize a scalar reward or reinforcement signal. The learner is not told which action to take, as in the most forms of machine learning, but instead must discover which action produces the highest reward by trying them.

Reinforcement Learning algorithms is a zone of Machine Learning inspired by behaviorist psychology and biological learning systems B [25]. It deals with learning via interaction and feedback; it solves tasks with trial and error by acting in an environment and receiving rewards for it.

Reinforcement learning contrasts with other machine learning approaches in that the algorithm is not explicitly told how to perform a task, but works through the problem on its own. It is about making decisions sequentially, where the output depends on the state of the current input, and the next input depends on the output of the previous input. Therefore, we do not give labels to each decision but we give labels to sequence of dependent decisions.

Three major components make up reinforcement learning, **Agent**: the learner and the decision maker. **Environment**: where the agent learns and decides what actions to perform. **Action**: a set of actions, which the agent can perform. The agent learns by interacting with its environment, that it receives rewards by performing correctly and penalties for performing incorrectly. The agent learns without intervention from a human by maximizing its reward and minimizing its penalty. Reinforcement learning happens when the agent chooses actions that

maximize the expected reward over a given time. This is easiest to achieve when the agent is working within a sound policy framework.

The next figure (**Figure II.5**) represents the interactions between the reinforcement learning components, with indicating the agent state in the environment.

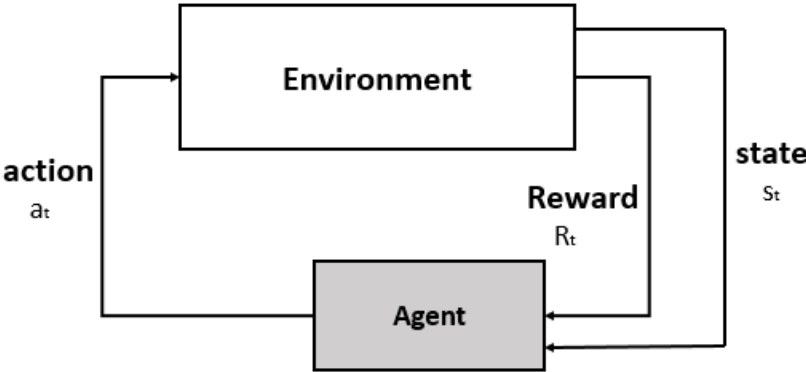


Figure II.5: Reinforcement learning, agent and environment interactions A [24]

There are two types of reinforcement learning: positive and negative. In the case of positive reinforcement, an event, which occurs following a specific behavior, reinforces the frequency of this behavior. The event therefore has a "positive" effect on the behavior of the model. Its advantages are: it supports change for a long period and maximizes the performance. Otherwise too much reinforcement can lead to surcharges of states which can diminish the results.

In the case of the negative reinforcement, behavior is reinforced because negative conditions are prevented. This increases the frequency of appropriate behavior, but only achieves a minimal result. This type is good for the Increase of the behavior and that it provides defiance to the minimum standard of performance, but it only provides enough to meet up the minimum behavior.

Reinforcement learning has various practical applications, it can be used in data processing, industrial automation, text mining, trade execution, healthcare and robotics such as Athanasios and Lazaros A [25] proposed a survey of Model-Based Reinforcement learning applications.

II.4.4. Deep Learning

Deep learning B [26] is a type of artificial intelligence derived from machine learning where the machine is able to learn by itself, instead of teaching computers to process and learn from data, which is how machine-learning works, with deep learning, the computer trains itself to process and learn from data. It relies on a network of artificial neurons inspired by the human brain. This network is made up of tens or even hundreds of layers of neurons, each receiving and interpreting the information from the previous layer.

The difference between deep learning and neural networks is in the depth of the model, deep learning is an expression used for complex neural networks. It is related to transformation and extraction of features, which aim to develop a relationship between stimuli and associated neural responses present in the brain. The complexity is awarded by elaborate patterns of how information can pass throughout the model. The figure (Figure II.6) below represents an example of a deep neural network and shows how the architecture has become more complex.

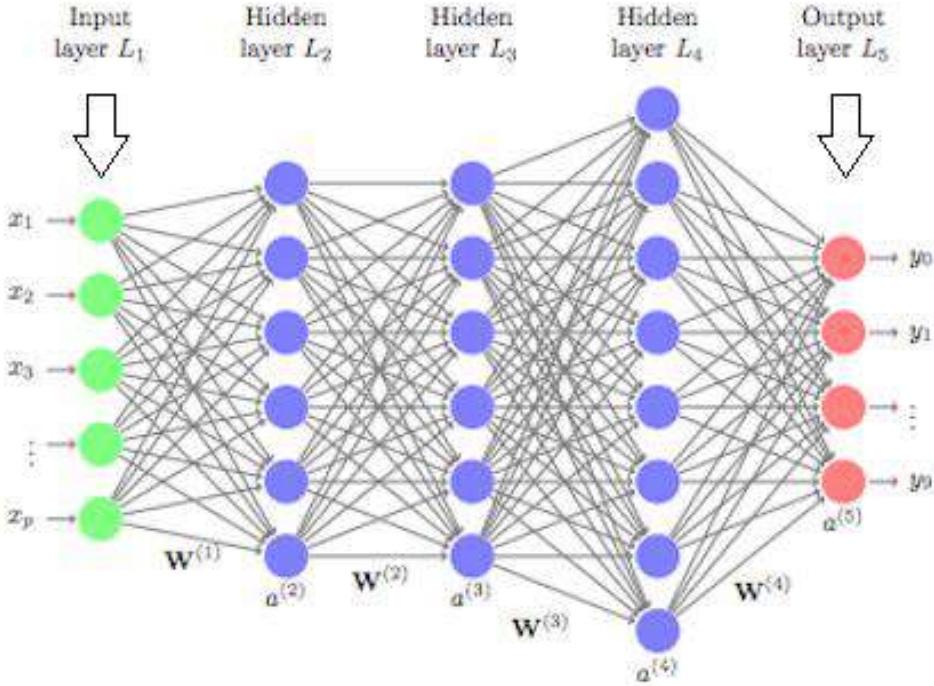


Figure II.6: Deep neural network architecture A [26]

Deep learning uses no linear processing units multiple layers for transformation and extraction features of the data, it also represents concepts in multiple hierarchical fashion, which

correspond to various levels of abstraction. In deep learning, algorithms can be either supervised and serve to classify data, or unsupervised and perform pattern analysis.

The most popular deep learning algorithms are: Convolutional Neural Network (CNN), Recurrent Neural Networks (RNNs) [27], Long Short-Term Memory Networks (LSTMs) [28], Deep Belief Networks (DBN) and many others.

II.5. Conclusion

Machine learning is an artificial intelligence technology and a very powerful tool that allows computers to learn and develop, in the previous chapter we introduced the machine learning and mentioned its importance, we defined supervised learning whether classification and regression, and unsupervised learning with clustering, we also briefly explained the reinforcement learning and the deep learning.

Beside, we had explained some of the most popular and useful machine learning algorithms, and mentioned the many various applications of each one. One of the most common application fields of machine learning in social big data is sentiment analysis. The next chapter explain the application of machine learning algorithms in sentiment analysis of social big data and accommodate its different approaches and challenges.

Chapter 3

Sentiment Analysis in Social Media

III.1. Introduction

In the previous chapters, we mentioned the explosive growth of social media, and the large increase of social media users, which are allowed to conduct many activities such as interacting, sharing, posting and manipulating contents. The resulting unstructured data generated by users imposes new computational techniques from social media mining, it offers an opportunity to study and understand individuals at different and unprecedented situations and events.

Sentiment Analysis also known as *Opinion Mining* is one class of computational techniques, which automatically extracts and recapitulates the opinions of such immense volumes of data that is unable to be processed by humans. Sentiment Analysis in Social Media begins with an overview of the latest research trends in the field; it discusses the sociological and psychological processes underlying social network interactions. The author in B[29] shows how social network streams pose numerous challenges for sentiment analysis, due to their large-scale, short, noisy, context- dependent and dynamic nature, further, it shows how to apply sentiment analysis tools for a particular application and domain, and how to get the best results for understanding the consequences.

Sentiment Analysis utilizes methodologies, theories, and techniques, from a number of scientific and computing domains, including psychology and sociology, natural language processing, machine learning, big data, and statistical methodologies. In this chapter, we present a definition for Sentiment Analysis and its different application fields; we will cite the Natural Language Processing and the challenges of the different languages. In addition, we will discuss the sentiment analysis approaches and techniques, and will classify previous works depending on the used techniques.

III.2. Sentiment analysis

Sentiment Analysis or Opinion Mining B [30] is the process of determining the emotional tone behind a series of words, it consists of building automatic tools capable of extracting subjective information from texts in natural language, so as to create structured and exploitable knowledge that can be used by a decision support system or a decision maker. This analysis is used to better understand the perception, opinions, emotions, attitudes and feelings expressed in an online report.

First, we must focus on what we mean about sentiment. Simplistically, sentiment describes the feeling that comes from within a comment or review it answers the questions: is someone for or against a specific subject? And did he like or dislike something?

Therefore, sentiment analysis is the process of examining, pre-processing, transforming and classifying a collected dataset of natural language texts, in order to extract the subjective information of this text and decide its sentiment.

When a sentence in a text is objective as in “it is raining”, no other basic task is required. Otherwise, when a sentence is subjective as in “I love the rain!” the sentiment is usually described as a binary opposition, but it is often more complex. There are comments and reviews that offer neither a good or bad opinion, which is known as a neutral opinion. Hence classifying whether the sentence is expressing objective information or subjective views and opinions is the classification of subjectivity task and, the classification of polarity is the task that determines the sentences, which express positive, negative or neutral polarities. Sentiment analysis includes many other tasks such as: Opinion retrieval, Opinion summarization, Opinion holder identification, Topic/sentiment dynamics tracking, Opinion spam detection and many others.

III.3. Sentiment analysis levels

To apply sentiment analysis we should define the text that will be analyzed in the case of a study considered. Sentiment analysis can occur at generally three levels:

- **Document level:** it determines the polarity of an entire text (classify a review as positive, negative, or neutral). It works best when a single person writes the document and expresses only one view on a single entity.
- **Sentence level:** it determines the polarity of each sentence contained in a text; it usually involves subjectivity classification of the sentence (whether the sentence is objective or subjective) and sentiment classification of subjective sentences (positive or negative).

- **Aspects/ Features level:** it performs a finer analysis than the other levels. It is based on the idea that an opinion consists of a feeling and a target, it identifies and extracts object features that have been commented on by the opinion holder and determines whether the opinion is positive, negative, or neutral. This level of analysis makes it possible to differentiate the aspects which are liked or not by the authors of the texts and thus makes it easier to determine possible treatment.

Sentiment analysis is extremely useful in social media monitoring because it provides an overview of the public's opinion on certain topics. The ability to extract insights from social web data is a practice that is widely adopted by companies around the world; therefore, the use of sentiment analysis is both broad and powerful.

III.4. Sentiment analysis applications

The importance of sentiment analysis is present in various domains, such as politics, medical field, emergencies, economy, security and sociology, so several applications have emerged in this context. In the following section, we will mention some of the works in the application of sentiment analysis in various fields.

- **Medical:** in the medical domain **SINAI WEB** [09] Research group A [27], generated a corpus by crawling the website Masquemedicos with Spanish opinions about medical entities written by patients. They presented a new resource called **COPOS WEB** [10] (**Corpus Of Patient Opinions in Spanish**), in order to demonstrate the validity of this corpus, they carried out different experiments with the main methodologies applied in polarity classification. Lexicon-based method A[28] for the semantic orientation and the data mining system Rapidx as a tool for classifying the polarity in the corpus.
- **Emergencies and natural disasters:** Ghazaleh, Xia, Ross and Huan in A[29] explored applications of sentiment analysis and demonstrated how sentiment mining in social media can be exploited to determine how local crowds react during a disaster, and how such information can be used to improve disaster management. They discussed the relationship among social media, disaster relief and situational awareness and explained how sentiment analysis in social media can be used in these contexts, and how such information can be used to help assess the extent of the devastation and find people who

are in specific need during an emergency. Further, in order to enable quick analysis of real-time geo-distributed data, they detailed the applications of visual analytics with an emphasis on sentiment visualization.

- **Politics:** in politic sector Andrea, Luigi, Stefano and Giuseppe A [30], applied a method recently proposed by other social scientists to three different scenarios, by analyzing on one side the online popularity of Italian political leaders throughout 2011, and on the other the voting intention of French Internet users in both the 2012 presidential ballot and the subsequent legislative election. Their analysis shows a remarkable ability for social media to forecast electoral results, as well as a noteworthy correlation between social media and the results of traditional mass surveys. They also illustrate that the predictive ability of social media analysis strengthens as the number of citizens expressing their opinion online increases, if the citizens act consistently on these opinions.
- **Economy and business:** Jasmina, Miha, Nada, and Martin A[31], presented a stock market application of tweets sentiment analysis, they studied whether Twitter feeds, expressing public opinion concerning companies and their products, are a suitable data source for forecasting the movements in stock closing prices. They used sentiment analysis to predict the changes in the phenomenon of interest, and the term predictive sentiment analysis to denote the approach. Besides, they adapted the Support Vector Machine classification mechanism to categorize tweets into three sentiment categories (positive, negative and neutral), and employed the Granger causality test and show that sentiment polarity can indicate stock price movements a few days in advance.
- **Security:** a work by Daniel, Bogdan and Alexander A [32] combined between security and emotion; this work gauged the presence and atmosphere surrounding security-related discussions on **GitHub** WEB [11], as mined from discussions around commits and pull requests. They used **NLTK** WEB [12] tool for the natural language processing to perform sentiment analysis. Further, they found that security-related discussions account for approximately 10% of all discussions on GitHub. And they encompass more negative emotions than other discussions, which confirm the anecdotal evidence that implementing application security can often lead to frustration and anger among developers, and is a source of tension to the overall project atmosphere.

III.5. Sentiment analysis approaches

Sentiment analysis field has been well studied by researchers in the past few years. Many different methods and techniques have been developed and tested through different tasks and at different levels. However, a lot of work is yet to be done. Sentiment analysis is the contrary to simple text classification due to the many challenges of the field; three types of techniques had been used to classify opinion, which are: machine learning based approach, NLP and lexical resources based approach and hybrid approach.

III.5.1. Lexicon-Based approach

The lexicon-based approach depends on finding the opinion word lexicon which is used to analyze the text, It identifies the polarity of a text using two sets of words divided into, positive represent the desired expression and negative represent the undesired expression, it require the sentiment lexicon to generate it either manually or semi-automatically.

The model counts in the text the number of positive words and the number of negative words, the sum gives an overall evaluation of the feeling of the text. The input text will be converted to tokens by the Tokenizer of NLP system; every token will be compared with the lexicon in the dictionary or in the corpus. If there is a positive mark, the result will be added to the total pool of score for the input text, and then the total score of the text is incremented. Else, the score is decremented or the word is tagged as negative, the text is possibly neutral if the numbers are equal.

This technique is governed by the use of two methods: the dictionary-based approach A [33] and corpus-based approach A [34]. Both approaches could be done by using statistical or semantic methods. The dictionary-based approach depends on finding opinion root or seed words, and then searches the dictionary of their synonyms and antonyms. The corpus-based approach starts with a list of seed opinion words, and then finds other opinion words in a large corpus to find opinion words with context specific orientations.

III.5.2. Machine learning approach

Sentiment classification and categorizing text into positive, negative or neutral categories require more practical techniques, hence the use of the machine learning technique with their fully automatic implementation and an ability to handle large collections of data. Machine

learning techniques are most useful techniques for sentiment analysis, because the steps of training a dataset by learning the documents then testing it to validate the performance, are very powerful and accurate. There are a number of machine learning algorithms used to classify texts.

Machine Learning-Based Sentiment classification methods can be categorized into: supervised and unsupervised. In sentiment analysis the most used is supervised learning, it generally comprises of four steps: Data collection, Pre-processing, Training data, Classification and writing results.

It begins with the collection of training data as tagged corpus, and then the classifier will be trained on this data and present a series of feature vectors from it. Therefore, a model will be created based on the training dataset, which will be used over the new text for classification purposes. Results will be written based on the type of representation selected; the performance tuning and the execution precision are done before the release of the algorithm.

III.5.3. Hybrid approach

Hybrid approach combines the strengths of the two previous approaches; it could collectively expose the accuracy of a machine learning approach and the speed of lexical approach. This approach takes into account all the processing linguistics of the lexical approach before starting the learning process as in statistical approaches. The hybrid approach gives the high accuracy from the machine learning and the stability from the lexicon-based approach.

The lexicon based approach tools and techniques use dictionaries and lexicons as the major source of search for sentiment classification. These lexicons have seeded semantic orientations that are later compared with the input data set for classification. Machine learning based approaches instead follow the learning algorithms to create the training data set. Then based on the trained dataset the inputs are compared and classified as either positive, negative or any other sentiment.

III.6. Sentiment analysis challenges

Sentiment analysis is the operation of determining the feeling polarity behind a series of words or a text, therefore the challenges here are mostly about the nature of the text and the language. Sentiment analysis challenges are wide and many, they present the obstacles in analyzing the

accurate meaning of sentiments and detecting the suitable sentiment polarity. Some of the challenges that sentiment analysis usually face are: subjectivity and tone, context and polarity, irony and sarcasm, comparisons, emoji, defining neutral and many others.

Since Sentiment analysis is the application of natural language processing methods and algorithms and text analysis techniques to identify and extract subjective information from text, the most considered challenges are related to the NLP domain. Under the next title, we identify the natural language processing and explain the practice of its most useful algorithms.

III.6.1. Natural Language Processing

NLP is used today for a wide variety of use cases. It become the driving force behind many common applications as Language Translation, Word Processors and Personal assistant. Another primary use case for NLP is Sentiment analysis; it is one of the most important fields of NLP, because it is the process of unearthing or mining meaningful patterns from text data.

The process of understanding and manipulating language is complex; NLP uses different techniques to handle different challenges of the natural language. The two main techniques used for natural language processing are syntactic analysis and semantic analysis.

- **Syntactic analysis** consists of identifying the grammatical rules in a sentence in order to decipher their meaning, by applying these grammatical rules to a group of words and derive meaning from them. It shows how the natural language aligns with the grammatical rules.
- **Semantic analysis** conducted based on analyzing the grammar of a sentence. Word segmentation consists of dividing a text into units, while morphological segmentation divides words into groups; it implies applying computer algorithms to understand the meaning and interpretation of words and the structure of sentences. Semantic analysis is one of the difficult aspects of NLP that has not been fully resolved yet.

Programming languages like Java and Python are used to implement the previous techniques. Python is a simple and powerful programming language with excellent functionality for processing linguistic data B [31]. The Natural Language Toolkit B[31] (NLTK) is a platform

in Python, defines an infrastructure that can be used to build NLP programs that work with human language data, it provide basic classes to represent these data, and contains a text processing libraries for applying the many algorithms in NLP.

Otherwise, natural language processing is considered a difficult problem and not an easy task in computer science. Human language is inherently complex and its different rules are hard for a computer to understand, these rules adduced many difficulties and challenges for NLP, including the abstract use of language and the sarcasm detection, also the fact that a sentence can change meaning depending on the words which the speaker puts stress on. The different human languages is another challenge for NLP, the ambiguity and precise characteristics of the different languages (Spanish, Chinese, Arabic,...) make the task much more complicated for machines to mastery these last.

III.6.2. Arabic Natural Language Processing

Arabic is the official language of 22 countries and recognized as the fourth most used language of the Internet. An introduction to Arabic Natural Language Processing by Nizar Y. Habash B[32], mentioned how Arabic language split into three varieties among Classical Arabic (CA), which is the form of the Arabic language used in previous centuries literary texts, it is essentially the form of the language found in the Quran. With some necessary modifications for the use of classical Arabic in modern times the Modern Standard Arabic (MSA) has appeared. MSA is the official language of the Arab world, it is the primary language of the media and education, it is mainly written not spoken.

Arabic Dialects (AD) are generally restricted and used for informal daily communication, they were primarily spoken not written, however this changed with the Arab access to the electronic social media communication, AD can be vary or divided depending on dimensions, among geography or social class. Both MSA and AD could be written either in Arabic or in Roman script (Arabizi). It corresponds to Arabic written with Latin letters, numerals and punctuation.

A set of challenges and difficulties of Arabic NLP submitted by Ali and Khaled in A [35] and by Khaled, Sanjeera, Manar and Azza in A[36], in the following section we will mention the most important difficulties according to these two previous works:

- **Arabic orthography:** first, Arabic script written in consonantal alphabet from right to left, where the letters have contextual variants. Arabic orthography relate to many problems and challenges for ANLP. Such as the optional diacritics, symbols that carry the intended pronunciation of words, to helps clarify the sense and meaning of the word, like in “مَدْرَسَةٌ” (school, madrasa) and “مُدْرَسَةٌ”, which it mean (teacher/female, almударisa). And the long vowels (“ا”, Alef), (“و”, Waw), and (“ي”, Ya'a) can appear at the beginning, in the middle, or at the end of a word, and it has many forms of pronunciation or no pronunciation at all, as in “ذهيوا” no pronunciation for the letter Alef at the end. Also the fact that the shape of a letter can be changed depending on whether it is connected with a former and subsequent letter, or just connected with a past letter.

Further, Arabic is morphologically rich, where a core word has many inflected forms, it is known by the template morphology where words involve roots and illustrations in the form of patterns, and fastened with affixes. Arabic verbs have many different forms B[32], based on gender, number, person, aspect, mood, voice, conjunction and many others, these verbs are basically three or four characters root verb, which can represent a descriptors by adding patterns to them, as we can mention the verb (“فتح”, he opened) and (“مفتوح”, open/ it's open).

Furthermore, these can be more complicated when it comes to words with prefixes and suffixes, the prefixes can be relational words or conjunctions alphabet, and the suffixes are largely protests or individual/ possessive. Both are permitted to be mixed, therefore a word can have zero or more affixes. The example “وسيحضرونها” (and they will bring it), where the Lemma is “حضر” accepts three prefixes: “س”, “و”, and “ي”, and two suffixes: “ن” and “ها”, shows how Arabic morphology is complex and represent a challenge for ANLP in tokenization, stemming and lemmatization.

Another morphologic challenge in Arabic language is that we can compose a word to another by a conjunction of two words. as in “مادام” (as long as) comes from the compound of a particle “ما” and the verb “دام”, and the word “كيفما” (however) comes from “كيف” and “ما”. These compound words are important for understanding the Arabic text, which is a challenge to POS tagging and applications that require semantic processing. These morphology and sentence structure give the ability to incorporate a broad number of adds to each word, which makes the combinatorial expansion of possible words.

- **Arabic grammar:** the Arabic grammar differentiate between two types of sentences: verbal and nominal. Verbal sentences usually begin with a verb; nominal sentences begin with a noun or a pronoun. These sentences can be syntactically-flexible which gives a wider range of syntactic variability and types of variations possible, or may contain multi word expressions as in the medical terminology “الدم فقر” (Anemia) that consists of two words that has the literal meaning “فقر” (poor) and “دم” (blood). Also may contain arguments, adjectives and pronominal anaphora which is ambiguous whether the third personal pronoun, called “ضمير الغائب” in Arabic, is present in this case the difficulty is to identify the correct antecedent indicated by the third personal pronoun (her/hers/it/its), (him/his/it/its), (them/their) or hidden which causes grammatical mistakes in automated NLP system. These will make the sentence more complex to be understood, that is why Arabic syntax is intricate and recognized as a very difficult problem in NLP.
- **Semantic ambiguity:** Arabic have constituent boundary ambiguity like in: “صاحب المنزل الجديد”, it could mean (the new house owner) or (the new owner of the house), this depends on the boundary of the adjective phrase within this noun construct. Beside semantic ambiguity in Arabic is further complex and vague, where sentences and phrases may be interpreted in different ways, as “يحب علي أحمد أكثر من إبراهيم”, (Ali likes Ahmed more than Ibrahim.), this could mean that Ali likes Ahmed more than Ali likes Ibrahim, or Ali and Ibrahim like Ahmed, but Ali likes Ahmed more than Ibrahim likes Ahmed.

Finally, in addition to all these challenges and problems in ANLP, many other difficulties will increase when it comes to dialect especially that in social media the dialect is the most used. Every specific dialect makes changes and variations whether in lexical, phonology and morphology. This will make the sentiment analysis in social media more confrontational and challenging.

III.7. Social media Sentiment analysis

Social media is a huge virtual space where to express and share individual opinions, influencing any aspect of life. Social media sentiment analysis is applying the sentiment analysis process to analyze online mentions or messages and determine the feeling behind them, it discovers

whether the user is reacting positively, negatively or neutral to a particular topic, product, service or message.

The new sentiment analysis technologies enable the automatic analysis of the information distributed through social media to identify the polarity of posted opinions, these opinions and reviews are increasing their importance in the evaluation of products and services by potential customers. Therefore, Social media sentiment analysis is essential to run a successful social media campaign, to boost brands and to provide the essential knowledge about social media marketing performance.

A sentiment analysis tool is a software that analyzes text or conversations and evaluates the tone, intent, and emotion behind each message or each word. Within the importance of sentiment analysis in social media, and with the use of AI, many different sentiment analysis tools has been developed such as: Social Mention WEB [13], RapidMiner WEB[14], Quick Search, Brandwatch WEB[15], Repustate, and many others.

III.8. Related works

The existing related works on sentiment analysis can be classified from different points of views: used technique, view of the text (language), level of detail of text analysis, source of data, rating level and many others. In this section we present classification based on the previous approaches which we identified (lexicon approach, machine learning approach, hybrid approach), social media as source of data and English also Arabic (MSA and DA) as text language.

- **An Analysis of Online Twitter Sentiment Surrounding the European Refugee Crisis:** Linguistic Inquiry and Word Count (LIWC) WEB [16] is a natural language processing tool, uses a lexical approach to perform sentiment analysis and the large LIWC internal dictionaries. It has been chosen by David and Josephine in their lexicon-based sentiment analysis study on tweets A [37].

LIWC contains a number of sentiment categories, four LIWC categories were selected for the sentiment analysis results presented in this work: : positive emotion, negative emotion, anger and anxiety, with a list of words and word steam associated to each category. LIWC also contains a mean score for each category list for both English and German languages. They used these means as a baseline to compare the sentiment scores

of each experiment, and they calculated their own mean score based only on their dataset.

The tweet text contained within each tweet object for each day and each language. Was extracted and stored in a single file, each file passed through the LIWC tool. LIWC generates the score, which is calculated per day and per language based on the total number of words present within the tweets that match words, word stems, emoticons and expressions categorized into the specified categories of the English and German internal LIWC dictionaries.

- **Lexicon-based approach for sentiment analysis of Arabic tweets:** by Mahmoud Al, Safa and Izzat A[38], they built a sentiment lexicon of about 120,000 Arabic terms, each sentiment term values range between 0%–100% with the positive, neutral and negative words taking values in the ranges (60%–100%), (40%–60%) and (0%–40%), respectively. The process of building the sentiment lexicon was divided into: collect Arabic stems, translate them into English and use online English sentiment lexicons to determine the sentiment value of each word.

They also built a SA tool based on predicate calculus. After, applying the pre-processing and stemming phases on the tweets, sentences were mapped into sentiment vector that combines the sentiment values of individual words to compute the sentiment orientation of the sentences. In addition, they chose to formulate words and sentences using a variant of predicate logic and use the corresponding predicate calculus to compute the overall sentiment orientation of a tweet.

In order to perform their experiments, they manually labelled a dataset of tweets, and tested it by balancing the number of tweets in each of the considered three classes (positive, negative and neutral). Moreover, they selected the tweets to be of similar length in terms of the numbers of words and characters.

- **Sentiment lexicon for sentiment analysis of Saudi dialect tweets:** SauDiSenti is a sentiment lexicon for sentiment analysis of Saudi dialect tweets; it was built by Abdul Mohsen, Qubayl and Abdulaziz A [39]. Their work presented two resources: the Saudi

dialect sentiment lexicon (SauDiSenti) comprises 4431 words and phrases from modern standard Arabic (MSA) and Saudi dialects manually extracted from a previously labelled dataset of tweets obtained from trending hashtags in Saudi Arabia. Additionally, a testing dataset comprising 1500 tweets evenly distributed over three classes: positive, negative, and neutral.

To build SauDiSenti lexicon, they used a dataset of tweets previously labelled as the Saudi dialect twitter corpus (SDTC), comprising 5,400 tweets containing Saudi dialect and MSA. By the help of two annotators, they added all the negative words and phrases provided by each annotator and removed the duplicated words, then gave them -1 as the score. The same procedure was applied to the positive words and phrases, with +1 as the score.

They calculated the performance of SauDiSenti lexicon based on four threshold values. The tweet is classified as a positive tweet if the tweet score was greater than or equal to 0, totally greater than 0, greater than or equal to 1, or even totally greater than 1. Otherwise, the tweet is classified as negative tweet

- **Sentiment Analysis of Arabic Tweets in e-Learning:** using the two machine learning algorithms, Support Vector Machine (SVM) and Naive Bayes (NB), by Hamed, Renxi, Khalid and Dayou in A [40]. The aim of the study is to develop a framework to analyze Twitter “tweets” as having negative, positive or neutral sentiments in education or, in other words, to illustrate the relationship between the sentiments conveyed in Arabic tweets and the students’ learning experiences at universities.

They grabbed the tweets by an application was developed in C# and used Twitter's official Developers API to download them. The tweets were preserved in a database, then manually filtered labelled as negative (-1), positive (1) or neutral (0). They used Rapidminer to pre-process their data with Tokenization, Stop-word process, Light stem, Filter token by length. NB and SVM were used to build the classification models used to classify tweets as negative, positive and neutral. Finally, precision and recall methods were used to evaluate the classification results.

- **Sentiment Analysis of Facebook comments published in standard Arabic or Moroccan dialect using a machine learning approach:** by Abdeljalil, Mohcine, Hafdalla and Fatima-Zahra A[41]. The process of this work begins with the collection of comments and their annotation using crowdsourcing and a group of volunteers to define the polarity of comments, positive or negative, this task was followed by a text preprocessing phase in order to extract Arabic words reduced to their roots. These words are used for the construction of input variables, which were automatically retrieved from the formed composite from preprocessed comments.

To classify Facebook comments, they applied three supervised classification algorithms (implemented on R software): Naïve Bayes (NB), Random Forests (FA) and Support Vector Machines (SVM).

- **Arabic Sentiment Analysis using Supervised Classification** by Rehab and Islam A [42], where they applied the Naïve Bayes, SVM and K-Nearest Neighbor classifiers on an in-house developed dataset of tweets/comments.

They generated their dataset by collecting tweets and Facebook comments using the crawler and an annotation tool, which allow user to choose a label from Positive(1), Negative(-1) Neutral(0) and Other (the tweet/comment is deleted from the dataset) for each tweet or comment. At least three different users must rate every tweet/comment and majority voting is used to assign the final label.

Further, Rapidminer was used for the Tokenize, Stem (Arabic), Filter Stopwords(Arabic), and Terms operators on the collected data. Two folders were fed to The Process Documents from Files Operator in Rapidminer, one that contains the positive reviews and the other one with negative.

10-fold cross validation was employed to split the data into training and testing sets, and X-Validation operator was used, which is a nested one that consists of an operator for the classifier and another operator for calculating the performance of the classifier. In a comparison between The Naïve Bayes, SVM and KNN classifiers for detecting the polarity of a given review, the best precision was achieved by SVM.

- **Classification Approach for Sentiment Analysis:** Combination approach of feature extraction and classification techniques in A[43] by Sumita and Mamta, where two types of datasets were generated manually, one for training and another for testing, X:Y was the relation present within the training set. X represented the score of the probable opinion word and the representation whether the score is positive or Y. did negative.

The testing data will pass through: data preprocessing, lexical analysis of sentences, extraction of features by POS tagging (Part-Of-Speech tag) label tool WEB [17]. Defining Positive, Negative and Neutral words was done with the help of Stanford parser WEB[18] open source, while the opinion mining application Sentiwordnet WEB[19] was generated with three relevant polarities present for each word which are positivity, negativity and subjectivity.

The N-gram algorithm was applied for the feature extraction and K-Nearest Neighbor Classifier was selected, with the score of the opinion word related to a feature within the review as input to this classifier to classify input data into positive, negative and neural classes.

- **An hybrid scheme for Arabic Tweets Sentiment Analysis** in A [44] by Haifa and Aqil, which combines semantic orientation and machine learning techniques. Through this approach, the lexical-based classifier labeled the training data, a time-consuming task often prepared manually. The output of the lexical classifier was used as training data for the SVM machine learning classifier.

After preprocessing the data, SentiWordNet was used to extract some sentiment words after translating it into Arabic. This was followed by adding their own list of essential sentiment words. Thereby their sentiment lexicon consists of 1500 sentiment words. The overall tweets polarity was determined according to the cumulative score of the positive degree of all the sentiment words in that tweet. The n-gram models were used for extracting features.

SVM classifier was used to predict the polarity class of the unclassified tweets, those that failed the lexical-based classifier. It does this by building a model from the tweets that were classified by the lexical classifier.

The table below classifies the previous works depending on the used approach of each work, the treated language and the social media kind as data source. With mentioning the including tools of each one and the accuracy if it was stated.

Applied Approach	Article	Social media kind	Language	Main Tools	Precision
Lexicon-based Approach	[37]	Twitter	English and German	LIWC	**
	[89]	Twitter	Arabic (MSA)	Twitter's APIs	86.89%
	[39]	Twitter	Arabic (MSA and Saudi dialects)	**	**
Machine Learning approach	[40]	Twitter	Arabic (MSA)	Twitter's APIs, Rapidminer, TF / TF-IDF, n-gram	(SVM) 73.15%
	[41]	Facebook	Arabic (MSA and Moroccan dialects)	TF / TF-IDF, n-gram	(SVM) ~78%
	[42]	Twitter and Facebook	Arabic (MSA)	Rapidminer	**
Hybrid Approach	[43]	Twitter	English	POS tagging, SentiWordNet, n-gram, KNN	86%
	[44]	Twitter	Arabic (MSA and Saudi dialects)	Twitter's API SentiWordNet, n-gram, SVM,	84.01%

Table III.1: Classification of the previous related works.

III.9. Conclusion

Sentiment analysis is a process during which the polarity is positive, negative or neutral of a given text, it is usually applied on social media platforms, specifically on the user's posts, comment, tweets or even messages. In this chapter, we introduced a definition of sentiment analysis especially on social media platforms. Also with explaining and clarifying the wide application fields of sentiment analysis on social media and its effect on our daily lives, we also defined the different approaches of sentiment analysis.

While sentiment analysis is a process applied in texts written by humans, Natural Language processing (NLP) was an important title in our chapter, where we explained how the NLP system works, with its different techniques. Additionally we brought up some of the wide challenges faced by NLP in particularly Arabic languages and its different Dialectes. Finally, we classified some related works according to the used approach. Next chapter contains the entire description of used technique in our work based on a description of our practice of the Arabic sentiment analysis in social media and more explanations about the used techniques, algorithms and tools.

Chapter 4

Conception and Implementation

IV.1. Introduction

Sentiment analysis on social media data is important for many different fields, but it is not an easy task. In the previous chapters, we cited some of the many challenges faced by sentiment analysis, especially the difficulties of using social media data. We also defined the main connection between each big social data and the sentiment analysis task. Moreover, we introduced the different approaches for sentiment analysis, which acquired us to reach the machine learning algorithms.

After recapitulating the different approaches and methods used for sentiment analysis, we will now present our proposed method and give more details of our work by presenting the classifiers, the architecture of our database and the datasets for the classification. We will explain the preprocessing tasks and finally demonstrate the choice of methods by running the tests and discussing the results.

IV.2. Working environment

IV.2.1. Hardware environment

To accomplish our work we used the following hardware configuration: AMD E2-1800 processor Windows 8 Memory: 4 GB Hard drive: 750 GB Graphics: AMD Radeon HD 7340.

IV.2.2. Software environment and library

- **Python:** AI as a subfield of computer science, focuses on designing computer programs and machines capable of performing tasks that humans are naturally good at, including natural language understanding, speech comprehension, and image recognition. Vast ranges of different programming languages and environments have been used to enable AI and machine learning research and application development. Among them, Python programming language, it gains huge growth of popularity with the scientific computing community over the last years, where the most recent machine learning and deep learning libraries are now Python-based.

Python WEB [20] is an interpreted programming language; it runs on an interpreter system, where the code can be executed as soon as it is written, which offers rapidity in prototyping. Python has a simple, easy to learn syntax similar to the English language, this syntax allows developers to write programs with fewer lines than some other programming languages, and this syntax also emphasizes readability and therefore reduces the cost of program maintenance.

Python is powerful yet very accessible, some of the most important reasons, which qualified python to be one of the most suitable language for writing machine-learning code, are:

- **Inbuilt libraries for AI projects**: python has libraries for almost all kinds of AI and machine learning projects such as **Scikit-learn, Pandas, NLTK, NumPy, TonsorFlow, Matplodlib** and many others.
 - **Flexibility**: Python offers to choose using OOPs or scripting with no need to recompile the source code, developers can implement any changes and quickly see the results. Programmers can combine Python and other languages to reach their goals.
 - **Platform independence**: Python can run on a wide variety of hardware platforms with the same interface on all platforms, python for machine learning development can run on any platform including Windows, MacOS, Linux, UNIX and others.
 - **Readability**: Python is very easy to read, which leads to more efficient exchange of algorithms, ideas, and tools between AI and ML professionals.
- **Django**: is a high-level Python web application framework used for rapid development, pragmatic, maintainable, clean design, and secure websites; it uses Python programming language, which is popular and easy to use. One of Django's main goals is to simplify work for developers, it allows developers to focus on components of the application that are new instead of spending time on already developed components. Django has excellent documentation for real-world applications. Django is also exceedingly scalable and reassuringly secure.

All the functionality comes in the Django framework in the form of web applications, which need to be imported according to the need. Django has solved some major issues for web-developers that were solved at the expense of time and money before its existence. Moreover, since Django is an open source release, companies and organizations around the world as the Washington post, Google, Curse gaming and many others use Django in both large and small projects B [33].

- **Pandas**: is a software library written for the Python programming language, it is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most

powerful and flexible open source data analysis / manipulation tool available in any language. Pandas is well suited for many different kinds of data. In particular, it offers data structures and operations for manipulating numerical tables and time series.

The Library particularized and highlights many important points such as a fast and efficient DataFrame object for data manipulation with integrated indexing. Tools for reading and writing data between in-memory data structures and different formats, intelligent data alignment and integrated handling of missing data. Flexible reshaping and pivoting of data sets, intelligent label-based slicing, fancy indexing, and subsetting of large data sets, high performance merging and joining of data sets, highly optimized for performance and others WEB[21]. Pandas can be combined with other powerful libraries and python toolkits. This combination of environments will support doing data analysis, it excels productivity and performance.

- **Bootstrap (html / CSS / JavaScript / jQuery):** is a framework, used to create the graphical interface (the model and static site).
- **JavaScript:** JavaScript is a programming language for the web. It is mainly used to improve web pages to provide a more user-friendly experience.
- **NLTK (Natural Language Toolkit):** NLTK is an open-source platform and a suite of Python modules (libraries and programmes) for natural language processing. It is a leading platform for building Python programs to work with human language data. NLTK provides easy-to-use interfaces to over 50 corpora and lexical resources, along with a suite of text processing libraries for classification, tokenization, stemming, and tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum. WEB [12].

Arabic is the largest member of the Semitic language family, it has received comparatively little attention in modern computational linguistics. NLTK tool delivers state-of-the-art performance in a variety of Arabic language processing tasks; it transfers texts from human language into machine-readable format by passing through processing steps.

IV.2.3. Learning preprocessing

- **Scikit-Learn:** is one of the most useful libraries for machine learning in Python. It contains many effective tools for machine learning and statistical modelling including classification, regression and clustering and dimensionality reduction WEB [22].

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python, it comes loaded with a lot of features including: Cross Validation, Feature extraction, Data pre-processing, Feature selection, Various toy datasets, Supervised learning algorithms, Unsupervised learning algorithms and many others. The Scikit-learn library is focused on modelling data. It is not focused on loading, manipulating and summarizing data. For these features, refer to NumPy and Pandas.

- **TensorFlow:** is an end-to-end open source platform for machine learning developed by google. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in machine learning and developers easily build and deploy machine learning powered applications. TensorFlow ecosystem provides a collection of workflows to develop and train models using Python, JavaScript, or Swift, and to deploy in the cloud, on-prem, in the browser, or on-device no matter what language used WEB [23]. TensorFlow is one of the most widely used AI tools in machine learning, it is used to develop and run machine learning and deep learning applications. It offers a very high level and abstract approach to organizing low-level numerical programming. Moreover, supporting libraries that can allow software to run without changes on regular CPU. Its supported platforms include **Linux, MacOS, Windows, and Android.**

TensorFlow has better computational graph visualizations; it helps to execute subparts of a graph that gives it an upper hand as it allows introducing and retrieving discrete data. The libraries in TensorFlow are deployed on a hardware machine. Where Google backs the library management. In addition, has the advantages of seamless performance, quick updates, and frequent new releases with new features. Besides, TensorFlow is designed to use various backend software with also highly parallelism. It has a unique approach that allows monitoring the training progress of the models and tracking several metrics.

- **Google Colab** : Google Colab or Colaboratory is an online and free cloud service, created by Google, it is a Jupyter notebook WEB[24] environment, which runs entirely in the cloud. It is intended for education and research in machine learning. This platform allows training machine learning and deep learning models directly in the cloud without any setup needed, except a Google account and a browser.

Google Colab offers to write and execute code in Python, to document code that supports mathematical equations, to create/Upload/Share notebooks, to import/Save notebooks from/to Google Drive. In addition, it allows to import/Publish notebooks from GitHub, to import external datasets e.g. from Kaggle, to integrate PyTorch, TensorFlow, Keras and others. The most important feature that Google Colab offers is the use of GPU (Graphics Processing Unit). Colab supports GPU and it is completely free. Moreover, the ability of choosing different types of runtime in Colab is what makes it so popular and powerful.

In our project, we use Google Colab due to the inability and the deficiency of our personal computers. Colab is available for direct use in WEB [25], getting started needs only signing up on the Google account.

IV.3. Global Architecture

The following figure (Figure IV.1) illustrates the architecture of our project by setting the applicable steps and the necessary tasks by order. Moreover, it summarizes the conception of our project.

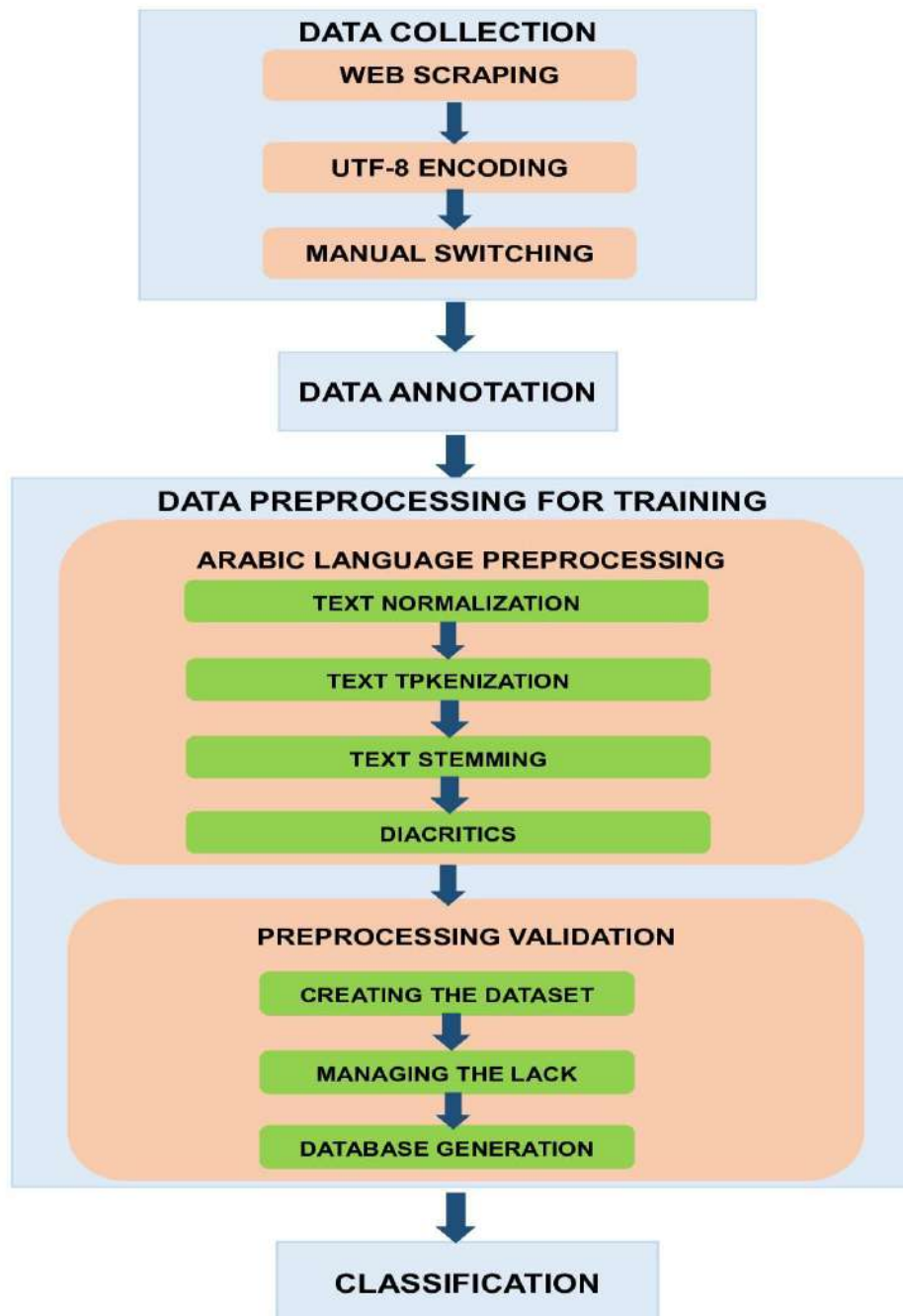


Figure IV.1: Global proposed architecture.

IV.4. Data Collection

- **Web scraping:** Web scraping refers to the extraction of data from a website. It is a term used to describe the use of a program or algorithm to extract and process large amounts of data from the web. The web scraper will be given one or more URLs to load before scraping. The scraper then loads the entire HTML code for the page. Moreover, it will extract all the selected data. Most web scrapers will output data to a CSV or Excel spreadsheet. Many useful packages for web scraping are available in Python; to get data from twitter we used the tow following libraries:
 - **Beautiful Soup:** is a Python library designed for quick turnaround projects like screen scraping, and it is used for pulling data out of HTML and XML files.
 - **Get-Old-Tweets-3:** is a Python 3 library and a corresponding command line utility that allows scraping data from twitter; it also allows scraping historical and old tweets. Get-Old-Tweets-3 allows scraping tweets using a variety of search parameters such as start/ end dates, username(s), text query search, and reference location area. Additionally, we can specify which tweet attributes are included like: username, tweet text, date, retweets and hashtags.
- **UTF-8 encoding:** is a variable-width character encoding capable of encoding all 1,112,064 valid character code points in Unicode using one to four one-byte (8-bit) code units.

The basic Arabic range encodes the standard Arabic letters and diacritics, all Arabic characters can be encoded using a single UTF-16 code unit (2 bytes), but they may take either two or three UTF-8 code units (1 byte each).

To retrieve the data, we use (**Get-Old-Tweets-3**) library, it helps us with **beautiful soup** to get all the not annotated Arabic and Algerian Dialectal tweets. Moreover, with manual work we were able to annotate almost all the retrieved data. In addition, to have help in annotation we generated a mini script, which aims to gather hashtags of each tweet and decide its polarity depending on the polarity of the words on hashtags. Finally, our final corpus is a fusion between available corpus on the internet and the tweets, which we retrieved and annotated. This corpus is in the following form: *Sentence --> annotation.*

- **Manual switching:** After getting the data, we found that inconsistency errors still exist. These errors are caused by the formation of sentences in Twitter. As a result, we establish the obligation of making a manual passage, which allows verifying the sentences.

We consider a sentence incorrect when an inconsistency is detected, the object polarity that does not match with the verb polarity as in: “راني ميت بالضحك” where “ميت” polarity is the opposite of “ضحك” polarity. Another case, where a sentence is incorrect, when a subject takes the place of the object, for example: “خويا طير مالك”.

Then we make the annotation at the time of correction.

IV.5. Data annotation

The proposed structure for an annotation is an object or a record, the record attributes are put in lists, and these lists are for objects, verbs, subjects and all possible words.

Annotation {Words = [Objects = [O1, O2, O3, On], Verbs = [A1, A2, A3, An], Subjects = [AC1, AC2, AC3, ACn]]}, Oi: object, Ai: verb, ACi: subject, Mi: Words.

Example:

Negative {Words = [فرح . انا . كان . كل . صباح]}

IV.6. Data preprocessing for training

IV.6.1. Arabic language preprocessing

First step in the Arabic language processing is deleting the Latin words and the emoji from our sentences. In cases where the sentence is composed with only Latin words plus emoji, it will all be deleted.

IV.6.1.1. Text Normalization

It is a process that converts a list of words to a uniform sequence by transforming the words to a standard format. It includes converting numbers into words or removing numbers, removing

punctuations, converting all characters to lower or uppercase, remove diacritics such as accents, umlauts or other, reduce or decompose characters to normal forms and many other tasks. In addition, the normalization process might also compromise an NLP task such as removing stop words, sparse terms, and particular words.

In Arabic, normalization generally contain strip diacritics (‘لِمُخْتَرَعَاتٍ’ into ‘لمخترعات’), strip elongation (‘العربية’ into ‘العربية’). Normalize Hamza (the ‘أ’ into ‘ا’ by removing the Hamza ‘ء’), normalize LamAlef (LamAlef ligatures ‘لا’ into two letters (LAM and ALEF) ‘ل’ and ‘ا’), normalize spell errors (TEH_MARBUTA ‘ة’ into HEH ‘ه’, ALEF_MAKSURA ‘ى’ into YEH ‘ي’), and many other tasks.

IV.6.1.2. Text Tokenization

It is the process of splitting the given text into smaller pieces called tokens, where the text can be divided into words with the method `word_tokenize()`, for example: ‘سأبدأ يومي ببعض الآيات’ will gives: ‘سأبدأ’, ‘يومي’, ‘ببعض’, ‘الآيات’, ‘القرآنية’, ‘نحن’, ‘بحاجة’, ‘الى’, ‘’, ‘’, ‘القرآنية نحن بحاجة الى السكينة’. Or into small phrases with the method `sent_tokenize()`, for the same example the result here is: ‘سأبدأ يومي ببعض الآيات القرآنية’, ‘نحن بحاجة الى السكينة’. Tokenization and normalization reduce sparsity and perplexity and decrease the number of out-of-vocabulary words.

NLTK supports stop word removal, it contain the list of Arabic stop words in the corpus module such as the prepositions ‘حروف الجر’, the conjunctions ‘حروف العطف’, the demonstrative nouns ‘أسماء الإشارة’ and the adverbs of space and time ‘ظروف المكان والزمان’ and many others. To remove stop words from a sentence, we can divide the text into words and then remove the word if it exists in the list of stop words provided by NLTK.

IV.6.1.3. Text Stemming

Word stemming in Arabic is the process of removing all of a word's prefixes and suffixes, or the conversion of plural to singular, or the derivation of a verb from the gerund form to produce the stem or root. It goes after finding the origin (root) of words in the natural Arabic language by getting rid of any additions in words, because Arabic words may have more complicated forms than any other language with such additions. The root has a general, basic meaning that forms the basis of many related meanings, and variations of the root determine the actual

meaning of the word. Removing the additions will make changes in the forms of words, which may sometimes make as well changes in the meaning of words. The simple following example explain the general idea of stemming: the words in ‘القلوب لا تأكل بل تحب’ will be stemmed to ‘قلب’, ‘لا اكل’, ‘حب’. The stemming process makes the classification operations less dependent on particular forms of words and reduces the potential size of vocabularies, which might otherwise have to contain all possible forms.

IV.6.1.4. Diacritics

In Arabic, diacritics are added to the characters of a word (as short vowels) in order to convey certain information about the meaning of the word as a whole and its place within the sentence. Arabic Text Diacritization (ATD) presents an important problem and challenge in Arabic natural language processing. Diacritics are also encoded similarly to the standard Arabic letters, so the NLTK can use them to improve the meaning of the words.

The following algorithm (**Algorithm 1**) illustrate cleaning and preprocessing data steps.

Algorithm 1: Cleaning and preprocessing data.

```

import re

from unidecode import unidecode

def tokenization(text):
    return set(word_tokenize(text))

def is_arabizi(text):
    html = len(re.findall(r'[A-Za-z0-9]+', text))

    if(html>0):
        text = unidecode(text)
        return text.replace(text, f' ')
    else:
        return text

def remove_fr(texte):
    word= tokenization(texte)

    x=str(texte)

    for val in word:
        tt=is_arabizi(val)
        x = x.replace(val, tt)

    return x

arabic_diacritics = re.compile("""
    ^          |# Tashdid
    ^          |# Fatha
    ^          |# Tanwin Fath
    ^          |# Damma
    ^          |# Tanwin Damm
    ^          |# Kasra
    ^          |# Tanwin Kasr
    ^          |# Sukun
    -         # Tatwil/Kashida
""")

```



```

        """, re.VERBOSE)

def remove_diacritics(text):
    text = re.sub(arabic_diacritics, '', text)
    return text

def normalize_arabic(text):
    text = re.sub("[|" , "[|]", text)
    text = re.sub("°" , "°", text)
    text = re.sub("ك" , "گ", text)
    return text

def remove_stop_words(texte):
    word= tokenization(texte)
    x=str(texte)
    ma_lis=[]
    for val in word:
        for text in stopwords:
            if(val == text):
                x = x.replace(val, '')
    return x

def remove_short(text):
    lower_word=text.lower()
    liste=tokenization(lower_word)
    sentence=''
    for word in liste:
        if(len(word) == 1):
            new=''
            sentence=sentence+''+new
        else:
            if(sentence == ''):
                new=word

```

```
        sentence=sentence+' '+new
    else:
        new=word
        sentence=sentence+' '+new
    return sentence
```

IV.6.2. Preprocessing validation

IV.6.2.1. Creating the dataset

For training our classification system, we need to have a dataset. This dataset is composed with a set of sentences associated with their annotation (positive, negative).

The dataset is divided into two parts. The first part contains the sentences with 100% precision, i.e. all the components of sentences and words are part of the same annotation. For this, we use the existing twitter database. Knowing that, the sentences of these two databases are composed with a set of words (verbs, subjects and objects) and each sentence is assigned to an annotation. The second part contains sentences with less than 100% precision. In this part, we create several datasets with different precision, and then we choose the best one for training.

Finally, we combine the two parts into a single dataset, which contains sentences with 100% precision and others with less than 100% precision. The goal of creating a dataset with different precisions here, is building a powerful learning model.

IV.6.2.2. Managing the lack values

The classification system has a static number of inputs, and there are some phrases in the database, which have fewer words than the number of the classifier inputs. This lack of data will create a problem for us. To solve it, we propose to fill the missing data of the sentence or of the set of words with significant values of the same annotation. At this point, we take the values from our database, and we choose the ones that frequent the most.

Otherwise, some sentences in our data lack verbs, or the pretreatment tasks may delete the sentences verbs, which will make the sentences more ambiguous. To solve this problem we used **LSTM B** [34], which can predict the right missing verb for each sentence and correct their meaning. For example passing “انا فرحان” to the **LSTM** neural network will transform it to “انا راني فرحان”.

After creating the dataset, we divided it into two parts. A part, which represents 70% of the dataset, intended for training, and the rest is used as test data.

IV.6.2.3. Database generation

The proposed structure for the database generation is a list of lists of objects or records. The parent list contains all the elements of the database, therefore all the annotations. Each node of the annotations contains a list of objects, subjects and verbs belonging to the annotation.

Databases = [DB1, DB2, ..., DBn];

DBi = [S1, S2, ..., Si];

Si = [Sentence 1, Sentence 2 ..., Sentence i].

The following algorithm illustrates the steps in generating datasets.

Algorithm 2: Generation of different databases

Input :

N : subject number

M : verb number

L: object number

K : sentences number per annotation

DB : our database

ML : existing database

P : precision

S : generated database number

Output :

The database DBG with N,M,P precision

Beginning :

For i=0 to S do

 For each annotation in DB do

 for i=0 to K do

 · Add extracted annotation sentence from ML to
 annotation_table

 · Complete sentence = create sentence (N,M,L,P)

 · Add Sentence to sentences_table

 * Add sentences_tables to sentence/annotation list (SA)

 • Add (SA) to DBG(i)

 do

 do

do

End

IV.7. Analysis and Results

In this section, we will demonstrate the choice of methods by running the tests and discussing the results.

IV.7.1. Classification models

First, we have chosen the classifier "**Neural Network Classifier**" to classify the dataset; it is based on the linear classification of data in a space of n dimensions. Moreover, we used another classifier "**Random Forest Classifier**". It has the ability to perform both regression and classification tasks at the same time.

IV.7.1.1. Neural Network (NN)

In the previous part, we mentioned that we created a dataset for training with the annotation precision parameter. Now, we create another dataset using a new parameter, which depends on the number of data inputs for the NN classification algorithm.

The NN classifier accepts a static number of data inputs, so all sentences (tweets) must have a static number of components. Therefore, we normalize and encode the sentence hence it is in the following form:

Sentence/Tweet = {subject1, subject2, subject3, subject4, verb1, verb2, verb3, verb4, object1, object2, object3}.

We suppose that a subject can make only one verb, so the number of verbs is equal to the number of subjects. The parameter used in the creation of the datasets, is n and m , such that n is the number of objects; m is the number of verbs and subjects at the same time. We apply the classification by the classifier NN on the datasets with the following parameters: $-n$ varied in the interval $[1, 3]$ and $-m$ varied in the interval $[1, 3]$.

In conclusion, after making tests and comparing the results (See result in **Figure IV.9**), we noted that the best parameters for a good classification model are:

$n=1, m=1$ or $n=1, m=2$, learning ration =0.005 sigmoid activation function, no_of_in_nodes = 3, no_of_out_nodes = 2, no_of_hidden_nodes = 4.

However, we have realized that there is an overfitting. For this, we have optimization for another classifier, which is "Random Forest Classifier". The following algorithm illustrates how we applied the NN classifier on our dataset.

Algorithm 3: NN classifier application on the dataset

Input :

learning ration =[0.0005,0.005,0.05,0.5]

activation function =[*sigmoid,Tanh,PRelu,Elu*]]

N: varied object number [1,3]

M: verbs and subjects number [1,3]

Output :

The best parameters of NN.

Beginning

- G=NM (generate all the possible combination) [n=2,m=1 for example]
- DB = generate set of datasets using G
- For each dataset in DB do

- Train_dataset = 70% of dataset

- Test_dataset = 30 % of dataset

- Parameters = *learning ration* **activation function*

- Best_parameter =

- Best_train=0

- For each parameter in Parameters :

- TC = Train_NN (Train_set,Best_parameter)

- if (Best_train < TC(Test_dataset)) so
Best_parameter =parameter

- Add Best_train , Best_parameter to final result table

table_train

- BEST= Retrieve the parameters of the first stagnation in
table table_train

do

do

END.

IV.7.1.2. Random Forest (RF)

Random Forest Classifier is a supervised classification algorithm, which can be used for regression and classification problems. This classifier is considered a set classifier, i.e. it

combines more than one classification algorithms of the same or different type to classify objects.

With the Random Forest Classifier, as the name suggests, it involves creating a set of decision trees. The higher the number of trees in the forest, the more precise the results obtained. Each decision tree is a single classifier and the target prediction is based on the majority voting method. Random Forest Classifier has two parameters, '**n-estimate**' which represents the number of decision trees in the classifier. The second parameter is '**maxdepth**', which means the maximum depth the tree can have, the value of this parameter is the length of the input data.

For **RF** classify, we use the same datasets that we used in the “**NN**”, that is to say with the same dataset parameters, **n** and **m** and precision.

As shown in **Figure IV.10** The RF classifier algorithm divides the dataset on the number of decision trees randomly. Then, each tree applies a training on its dataset. The advantage of this algorithm is that the learning time is reduced since the learning of the decision trees is executed in parallel.

In the class prediction phase, the classifier takes the query annotation and runs it on all the trees of the RF classifier. Then, a voting system is applied, the class most predicted by the classifiers is chosen as the final class. Algorithm 4 illustrates how we applied the “Random Forest” classifier on our dataset.

Algorithm 4: RF classifier application on the dataset

Input:

N: varied subject number [1,3]

M:verbs and objects number [1,3]

Nbr_tree : decision trees number [10,20]

Output :

The best parameter of RF.

Beginning

- G=NM (generate all the possible combinations) [n=2,m=1 for example]
- DB = generate set of datasets using G
- For each dataset in DB do

```

- Train_dataset = 70% of dataset
- Test_dataset = 30 % of dataset
- Best_parameter =
- Best_train=0
- For each nbr in Nbr_tree do
      - TR = Random_forest (Train_set,nbr)
      - if (Best_train < TR(Test_dataset) ) so
Best_parameter =Parameter
      - Add best_train , nbr final table result table_train
      - BEST= Retrieve the parameters of the first stagnation in
table table_train
End

```

The following example illustrates query with its result, which is an annotation:

Query: “راني فرحان”

Request structure: “أنا.” “فرح”

Query Result: Positive.

IV.7.2. Presentation of our platform features

Our Platform starts with a home page that contains icons buttons for different tasks and shows the number of the platform users also sentences and words number. This home page allows the user to authenticate as a member, if he already has an account, he can identify himself to access the features of the site.

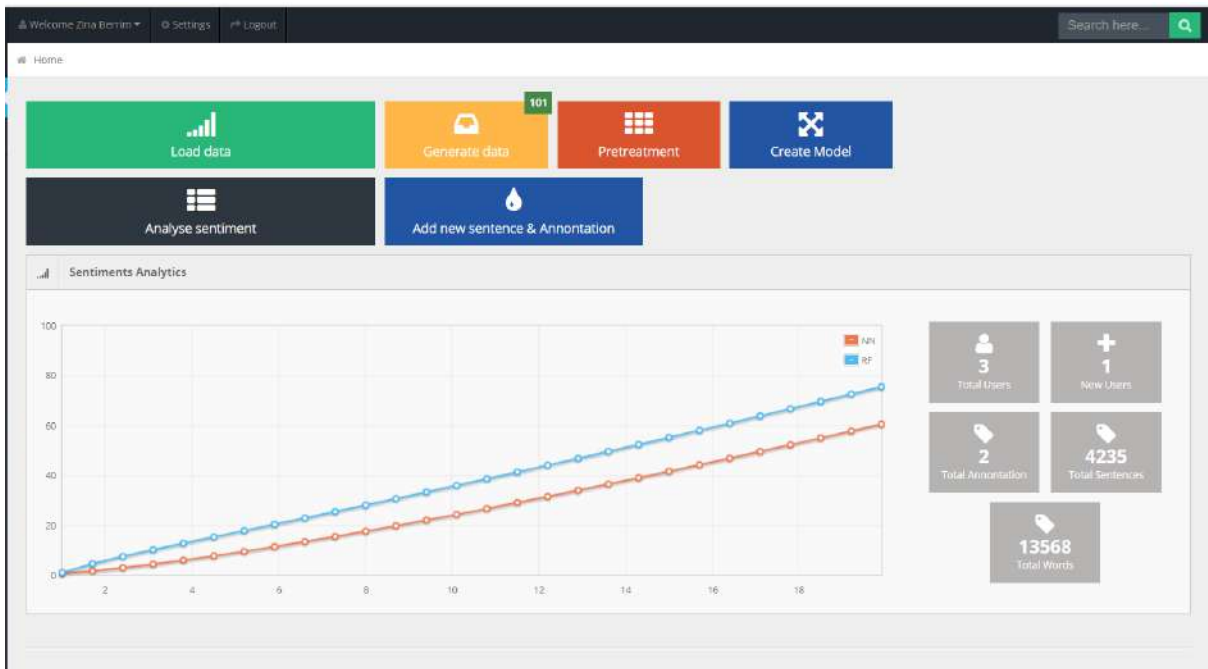


Figure IV.2: the platform home page

The user can add new sentences with their annotations to the database, he just have to write the sentence and the corresponding annotation

The screenshot shows the 'Add new sentence & Annotation' form. The page title is 'Add new sentence & Annotation'. The form contains two input fields: 'Sentece' (sic) with the value 'إلى بيت بالقرحة' and 'Annotation' with the value 'Positive'. A green 'Save' button is located below the input fields.

Figure IV.3: Adding new sentences and annotations

To test the annotation of a sentence, all you have to do is choose Dashboard then Sentiment Analysis, then give the request. Once the user clicks on ‘Query’ classification step will be launched to determine the annotation corresponding to the sentence.

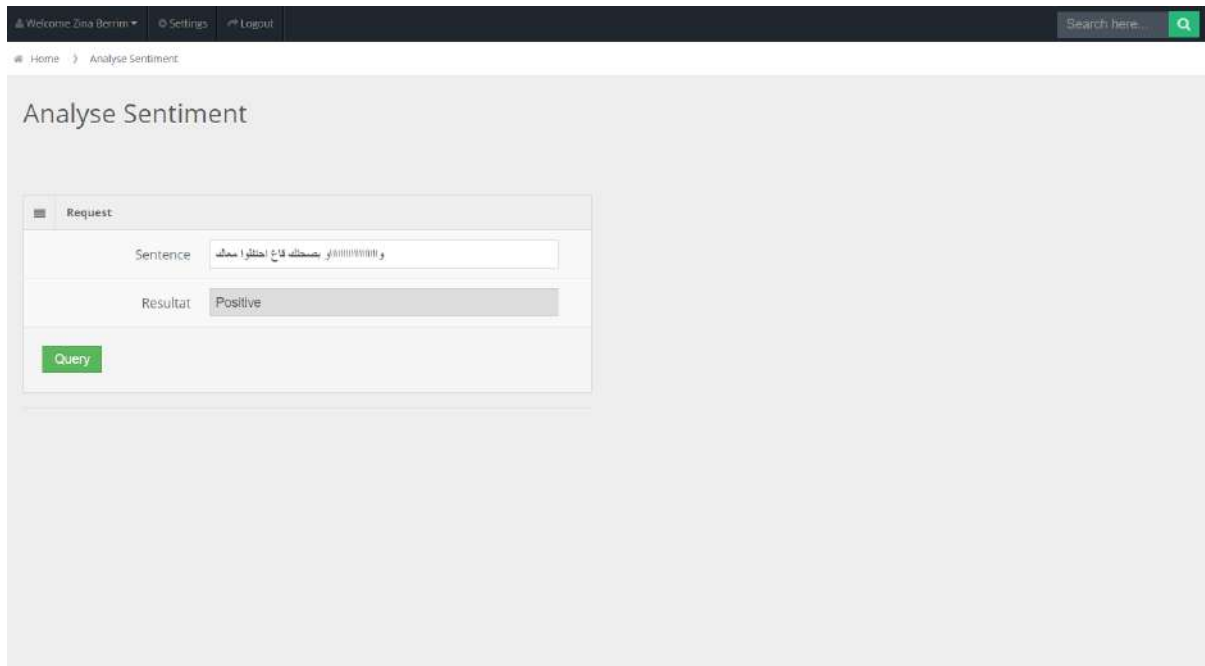


Figure IV.4: The sentiment analysis page.

The user can also generate different datasets to have done all the possible tests; you just have to fill in the fields of number of objects and verbs and the precision of the database.

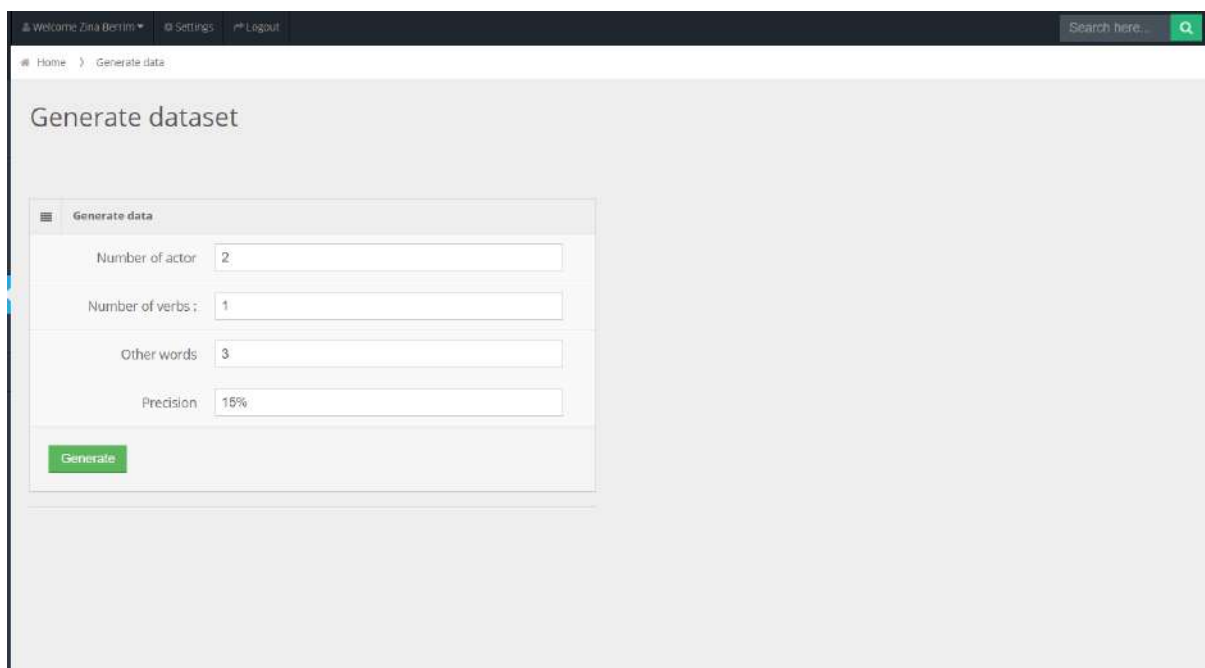


Figure IV.5: Generate dataset page.

Finally, Platform gives us the hand to do the preprocessing on sentences so that we can pass the learning task and correct the false words & false annotation.

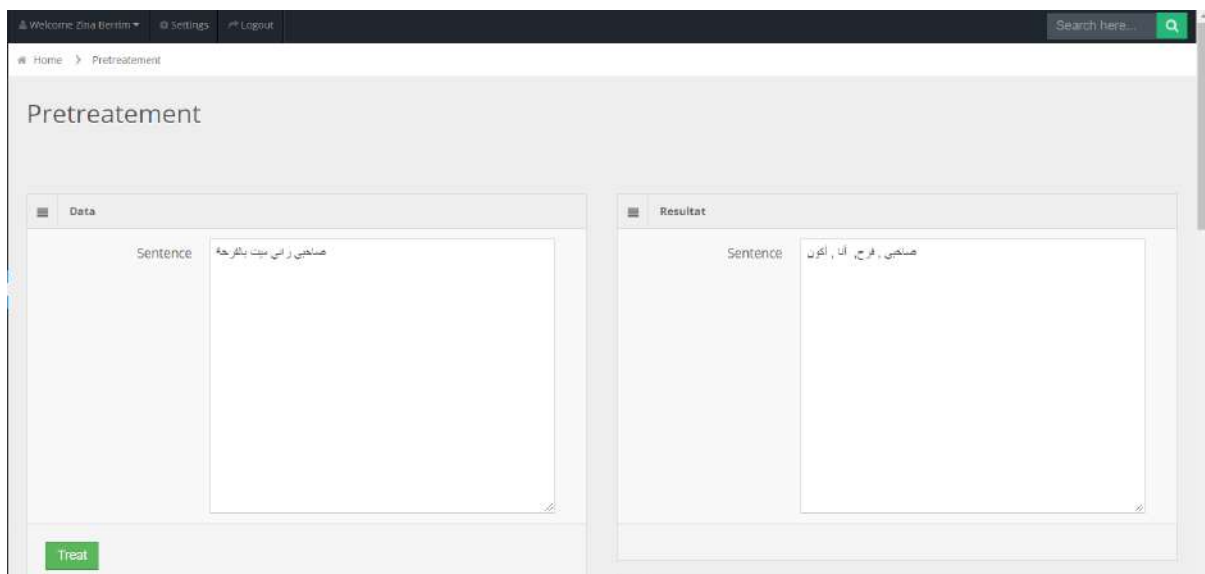


Figure IV.6: Pretreatment page.

IV.7.3. Results

In this section, we introduce the choice of parameters in order to generate the database, which will be used for the classification. Then, we compare the generated database with the existing database.

IV.7.3.1. Testing database parameters

In order to generate the database, we dispose of 2 parameters: the number of sentences per annotation and the number of words per sentence.

- **The number of sentences per annotation:** At this point, we show the generation result of the database by changing the number of sentences per annotation. In addition, we measure the precision of the classification. The following figure (**Figure IV.7**) shows the classification of the database depending on the number of sentences per annotation.

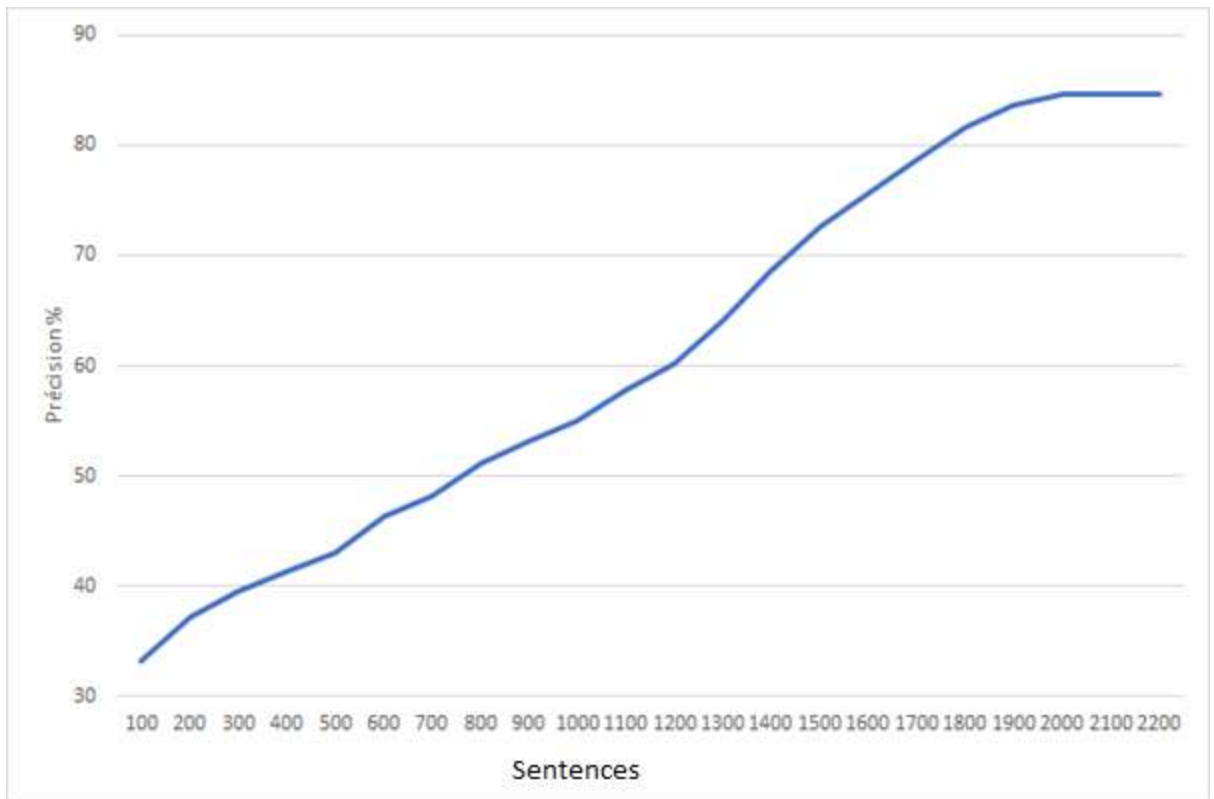


Figure IV.7: Database classification depending on the number of sentence per annotation

- **Discussion:** In these tests, we executed the Random Forest Classifier on the database generated by changing the number of sentences per annotation. After determining the number of annotations in 2, We determine the number of words per sentence at 3 up to 6 entities.

The result clarified that by increasing the number of sentences per annotation, the precision of the classification increases. The explanation of this result is that with a wide variety of words results in a reliable classification. On the other hand, the classifier stagnates with 2000 sentences, as long as we have a rich database. After these results, we will take 2000 sentences.

- **The number of words per sentence:** After having the number of sentences per annotation. We put the database generation tests by changing the number of words per sentence. The goal here is to know the best number of words we choose for the generation. The following figure (**Figure IV.8**) shows the classification of the database according to the number of words per sentence.

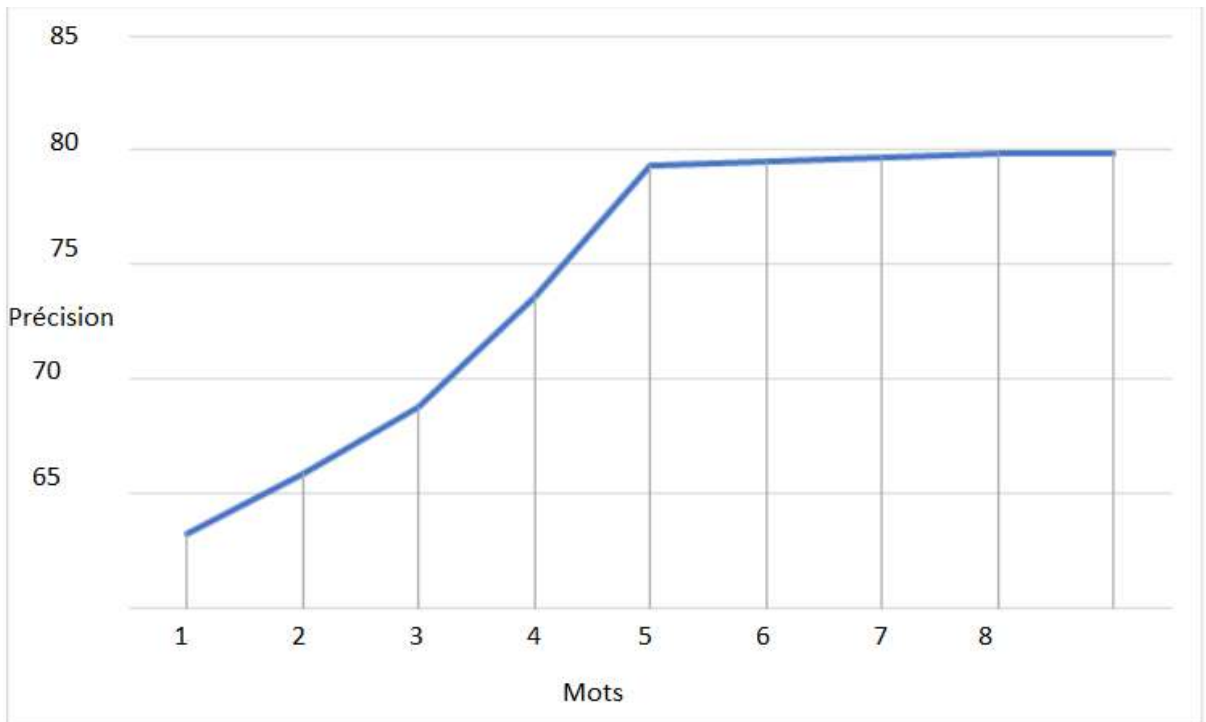


Figure IV.8: Database classification according to the number of words per sentence

- **Discussion:** In these tests, we also executed the Random Forest Classifier on the database generated by changing the number of words per sentence. After having the sentences number in 2000 sentences according to the previous tests.

The result shows that by increasing the number of words per sentence, the precision of the classification increases. This is caused by the words variety, which helps to better classify the sentences. Besides, since the stagnation of the classifier is when the number of words is 5. Therefore, we take 5 words per sentence in the generation of the database.

IV.7.3.2. Classification

- **Test the NN classification by varying the degree of the precision parameters of the database:** Regarding the NN test, we run the classifier on the database by varying the precision of the database. That is for the NN classifier power. **Figure IV.9** shows the graph, which illustrates the precision of the NN classification depending on the precision of the database (validity and cleanliness of database).

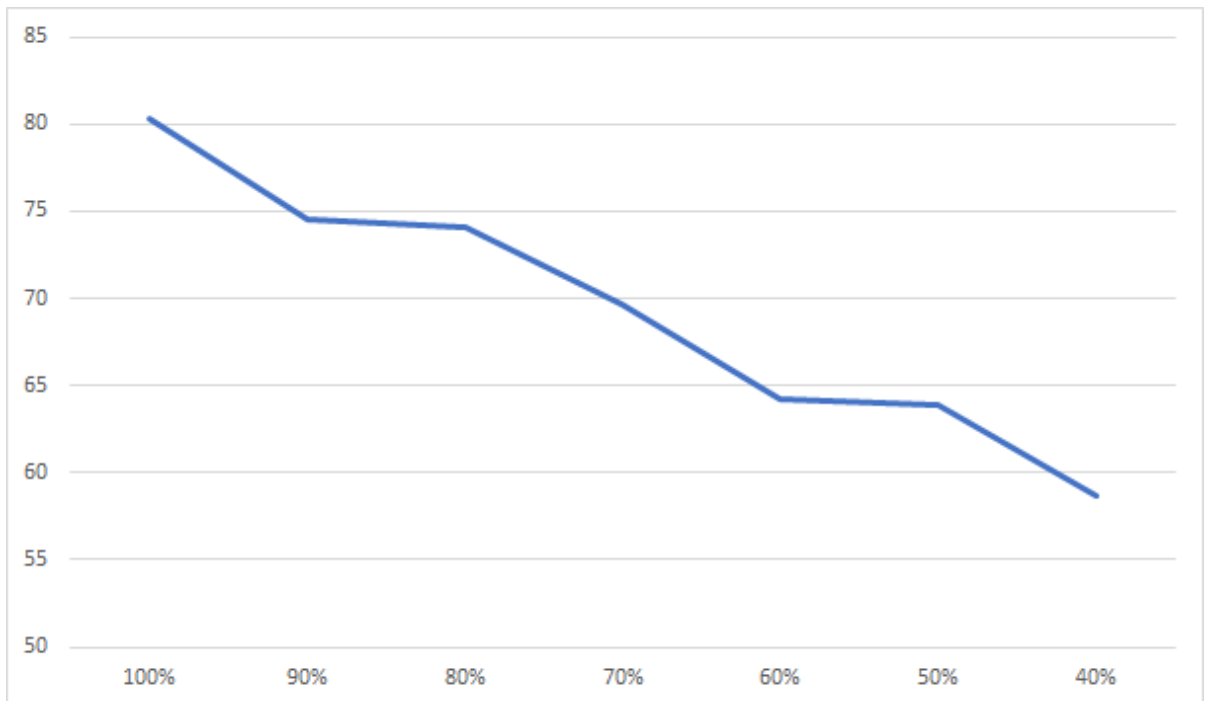


Figure IV.9: Precision of NN classifier

- **Discussion :** After making the tests, we saw that the NN classifier gives satisfactory results, the classification precision of which is greater than 70% with the database precision greater than 80%.
- **Test the RF classification by varying the degree of precision parameters of the database:** Regarding the RF test, we run the classifier on the database by varying the precision of the database. That is for the RF classifier power. **Figure IV.10** shows the graph, which illustrates the precision of the RF classification depending on the precision of the database.

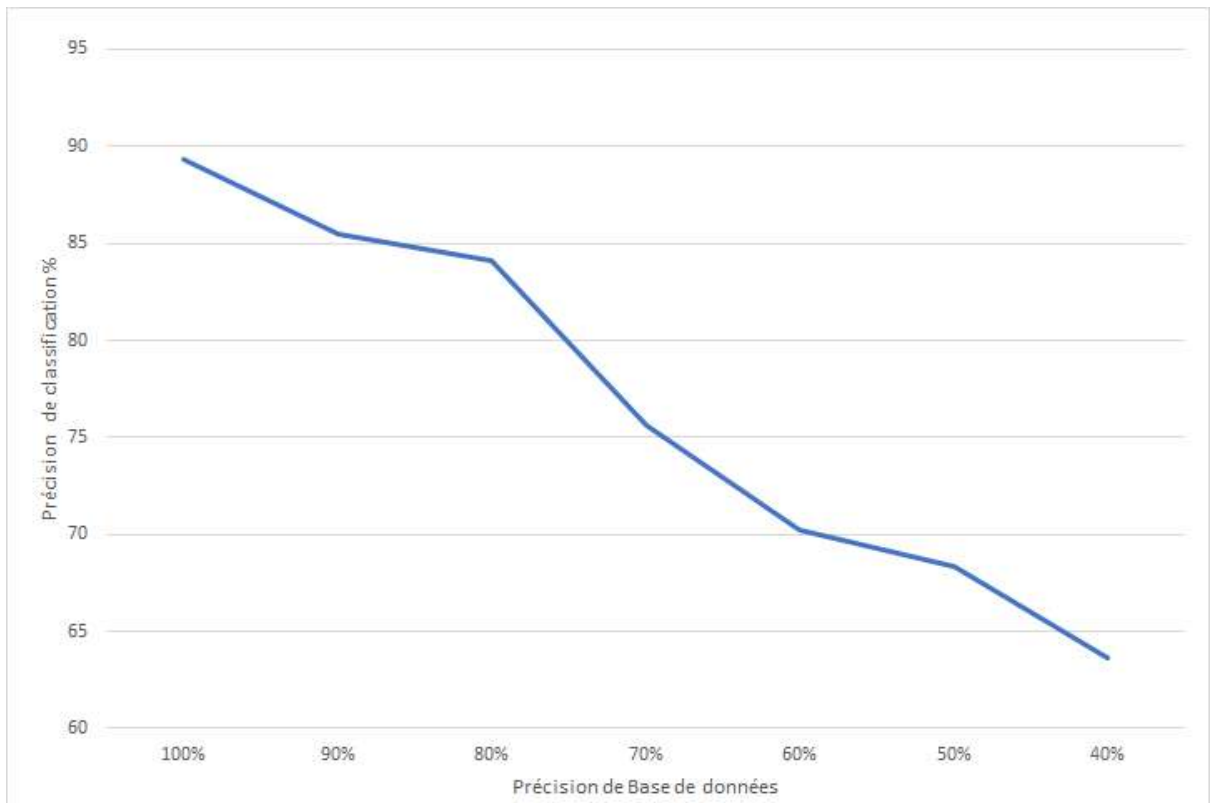


Figure IV.10: Precision of RF classifier

- **Discussion:** After making the tests, we saw that the RF classifier gives satisfactory results, the classification precision of which is greater than 85% with the database precision greater than 70%.
- **Comparing of the classification between RF and NN:** After testing the NN and RF classifiers, we make a comparing classification in order to choose the best classifier for the platform. **Figure IV.11** shows the graph, which illustrates a comparison between the two classifiers.

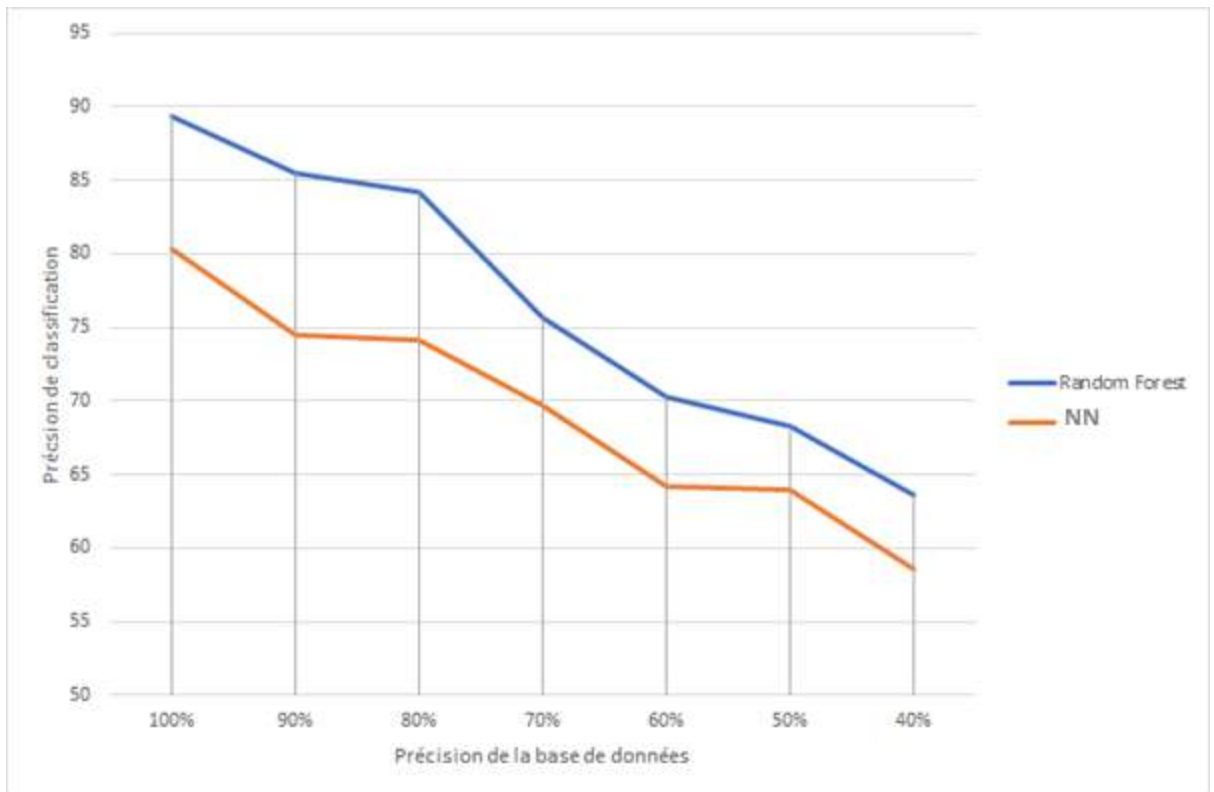


Figure IV.11: Comparison between NN and RF classifiers

- Discussion:** Depending on the obtained result, we notice that the curve of the RF classifier is above the curve of the NN classifier. The RF classifier gives 86% precision with 85% database precision, whereas the NN classifier gives 79% precision. After this comparison, we choose the Random Forest classifier.

IV.7.4. Evaluation of our method

In this section, we introduce the evaluation of our proposed approach for sentiment analysis of Arabic Algerian Dialectal tweets. The evaluation is made with the use of random Algerian Dialect sentences. We make tests for each sentence and we obtain the previous results. We notice that most of the entire results were right except two: query [1] and query [3], which is a good evaluation considering the lack of using powerful resources and machines. The following table presents the queries and results of evaluation.

General Conclusion

Sentiment analysis is a trending topic that based on analyzing and classifying texts written in natural human language. It explored the challenge of understanding and treating human language by machines. Recently, social media has been a great source of data for sentiment analysis field just as well as a big challenge due to its nature and noisiness.

Moreover, while sentiment analysis treat natural human language, this field cannot be separated to the natural language processing tasks. These last differentiate from language to another, Arabic is one of the most challenging and demanding language in sentiment analysis and natural language processing considering its nature and characteristics.

In the state of the art of this paper we managed making a useful introduction to the field of sentiment analysis on social media, by clearing and explaining the connection between each: social big data, artificial intelligence and machine learning algorithms and sentiment analysis. We defined each field and concluded the fact that making a powerful sentiment analysis project is depended on using social big data and machine learning techniques.

In this work, we presented our platform that allows generating a database, which contains tweets written in both Arabic and Algerian Dialectal Arabic with their polarity classes. The main functionality in the platform is to classify a given Arabic tweet or Algerian Arabic Dialectal tweet to positive or negative polarity. This work succeeded in adding a new powerful learning and classification model in the sentiment analysis field.

Otherwise, in this work we were limited to have better results owing to the lack of powerful machines and resources. In our future perspectives, we aim to work on creating big dataset for this project, and adding N-gram notion in the development of this work. In addition, we will work on using emoji in tweets to detect polarity.








Bibliography:









- 📖 A [1]: GEMA Bello-Orgaza, JASON J. Jungb, DAVID Camachoa, Social big data: Recent achievements and new challenges. ELSEVIER [online]. 2015. Available on : <https://www.sciencedirect.com/science/article/pii/S1566253515000780> (accessed 23/02/2020).
- 📖 A [2]: DOUG Laney, 3D Data Management: Controlling Data Volume, Velocity, and Variety, Technical report. application delivery strategies meta group [online]. 2001. 949. Available on : <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf> (accessed 23/02/2020).
- 📖 A [3]: Mark Beyer and Douglas Laney, The Importance of 'Big Data': A Definition, Gartner, Stamford, CT 2012.
- 📖 A [4]: EMC Digital Universe with Research & Analysis by IDC (International Data Collaboration). The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. April 2014. Sponsored by: EMC². 2014. 17.
- 📖 A [5]: ARIF Ahmed and RIPON Patgiri, Big Data: The V's of the Game Changer Paradigm. IEEE (Institute of Electrical and Electronics Engineers) [online]. 2016. Available on : https://www.researchgate.net/publication/311642627_Big_Data_The_V's_of_the_Game_Changer_Paradigm (accessed 25/02/2020).
- 📖 A [6]: GAYATRI Kapil, ALKA Agrawal, and R. A. Khan, A Study of Big Data Characteristics. SIST-Department of Information Technology, Babasaheb Bhimrao Ambedkar University [online]. 2016. Available on: https://www.researchgate.net/publication/315867458_A_study_of_big_data_characteristics (accessed 25/02/2020).
- 📖 A [7]: Caesar Wu, Rajkumar Buyya, Kotagiri Ramamohanarao, Big Data Analytics = Machine Learning + Cloud Computing [online]. 2016. Available on : <https://arxiv.org/ftp/arxiv/papers/1601/1601.03115.pdf> (accessed 10/06/2020).
- 📖 A [8]: JAKE luo, MIN Wu, DEEPIKA Gopukumar et al. Big Data Application in Biomedical Research and Health Care: A Literature Review. Biomedical Informatics Insights [online]. 2016:8 1–10 doi: 10.4137/Bii.s31559. Available on : https://www.researchgate.net/publication/291387301_Big_Data_Application_in_Biomedical_Research_and_Health_Care_A_Literature_Review (accessed 27/02/2020).
- 📖 A [9]: ALEXANDRA Amado, PAULO Cortez, PAULO Rita et al. Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. european research on management and business economics [online]. 2017. Available on : <https://www.sciencedirect.com/science/article/pii/S2444883417300268> (accessed 27/02/2020).
- 📖 A [10]: GANG-HOON KIM, SILVANA TRIMI, AND JI-HYONG CHUNG. Big-Data Applications in the Government Sector. Communications of the ACM (Association for Computing Machinery) [online]. 2014. VOL. 57. NO. 3. Available on : https://www.researchgate.net/publication/260865566_Big_Data_Applications_in_the_Government_Sector_A_Comparative_Analysis_among_Leading_Countries (accessed 27/02/2020).

- 📖 **A [11]:** VIKAS Dhawan and NADIR Zanini. Big data and social media analytics. m Research Matters: A Cambridge Assessment publication **[online]**. 2014. issue 18. Available on :<https://www.semanticscholar.org/paper/Big-data-and-social-media-analytics-Dhawan-Zanini/9a9eaba040ab9a15a06686b23cbd2b39d2e4cf75> (accessed 27/02/2020).
- 📖 **A [12]:** MICHAEL J. Magro. A Review of Social Media Use in E-Government. ADM SCI administrative science. **[online]**. 2012. 2, 148-161; doi:10.3390/admsci2020148. Available on: https://www.researchgate.net/publication/227439181_A_Review_of_Social_Media_Use_in_E-Government (accessed 27/02/2020).
- 📖 **A [13]:** Norjihani Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem et al. Big Social Media Data Analytics: A Survey. Computers in Human Behavior. 2018. doi: 10.1016/j.chb.2018.08.039.
- 📖 **A [14]:** JAMES Mor. Dartmouth Artificial Intelligence conference. The next fifty years. 1956. Dartmouth College, Hanover, New Hampshire. AI Magazine Volume 27 Number 4 (2006).
- 📖 **A [15]:** SAMUAL Arthur, Some Studies in Machine Learning: Using the game of checkers. IBM JOURNAL **[online]**. NOVEMBER 1967, 601-617. Available on: <https://www.cs.virginia.edu/~evans/greatworks/samuel.pdf> (accessed 18/06/2020).
- 📖 **A [16]:** M. I. Jordan and T. M. Mitchell. Machine Learning: Trends, perspectives, and prospects. SCIENCE sciencemag.org **[online]**. 2015, VOL 349, ISSUE 6245, 255-260. Available on: <https://science.sciencemag.org/content/349/6245/255> (accessed 10/03/2020).
- 📖 **A [17]:** J.R. Quinlan. Simplifying Decision Trees. International Journal of Man-Machine Studies **[online]**. 1987, Number 3, Vol. 27, 221-234. Available on: <https://dspace.mit.edu/bitstream/handle/1721.1/6453/aim-930.pdf?sequence=2> (accessed 11/03/2020).
- 📖 **A [18]:** LEO Breiman. Bagging Predictors. Technical Report No. 421 **[online]**. Berkeley, California 94720: University of California, 1994. Available on: <https://www.stat.berkeley.edu/~breiman/bagging.pdf> (accessed 11/03/2020).
- 📖 **A [19]:** LEO Breiman. RANDOM FORESTS--RANDOM FEATURES. Technical Report 567 **[online]**. Berkeley, California CA 94720: University of California, 1999. Available on: <https://www.stat.berkeley.edu/~breiman/random-forests.pdf> (accessed 11/03/2929).
- 📖 **A [20]:** WEIZHONG Yan. Application of Random Forest to Aircraft Engine Fault Diagnosis. Conference: Computational Engineering in Systems Applications, IMACS Multi-conference. 2006, Volume: 1, DOI: 10.1109/CESA.2006.4281698.
- 📖 **A [21]:** GRANT Izmiran. Application of the Random Forest Classification Algorithm to a SELDI-TOF Proteomics Study in the Setting of Cancer Prevention Trial. Annals of the New York Academy of Sciences **[online]**. 2004, 1020(1):154-74. Available on:

https://www.researchgate.net/publication/8500332_Application_of_the_Random_Forest_Classification_Algorithm_to_a_SELDI-TOF_Proteomics_Study_in_the_Setting_of_a_Cancer_Prevention_Trial (accessed 11/03/2020).

- 📖 A [22]: DEAN Nelson. Using Simple Linear Regression to Assess the Success of the Montreal Protocol in Reducing Atmospheric Chlorofluorocarbons. *Journal of Statistics Education* [online]. 2009, Volume 17, Number 2. Available on: <https://www.tandfonline.com/doi/full/10.1080/10691898.2009.11889520> (accessed 12/03/2020).
- 📖 A [23]: ZOUBIN Ghahramani. Unsupervised learning [online]. Gatsby Computational Neuroscience Unit: University College London, UK, 2004, 32. Available on: https://link.springer.com/chapter/10.1007/978-3-540-28650-9_5 (accessed 12/03/2020).
- 📖 A [24]: ROLHOLLAH Amiri, HANI Mehrpouyan, LEX Fridman, Ranjan et al. A Machine Learning Approach for Power Allocation in HetNets Considering QoS. Conference: 2018 IEEE International Conference on Communications (ICC 2018) [online]. 2018, DOI: 10.1109/ICC.2018.8422864. Available on: https://www.researchgate.net/publication/323867253_A_Machine_Learning_Approach_for_Power_Allocation_in_HetNets_Considering_QoS (accessed 12/03/2020).
- 📖 A [25]: ATHANASIOS S. Polydoros and LAZAROS Nalpantidis. Survey of Model-Based Reinforcement Learning: Applications on Robotics. *Journal of Intelligent and Robotic Systems* [online]. 2017, 86(2):153—173. Available on: https://www.researchgate.net/publication/312921419_Survey_of_Model-Based_Reinforcement_Learning_Applications_on_Robotics (accessed 14/03/3030).
- 📖 A [26]: ANDEREAS Holzinger, BERND Malle, PETER Kieseberg et al. Machine Learning and Knowledge Extraction in Digital Pathology Needs an Integrative Approach. *Towards Integrative Machine Learning and Knowledge Extraction* [online]. 2017, DOI: 10.1007/978-3-319-69775-8_2 (pp.13-50). Available on: https://www.researchgate.net/publication/320687279_Machine_Learning_and_Knowledge_Extraction_in_Digital_Pathology_Needs_an_Integrative_Approach (accessed 14/03/2020).
- 📖 A [27]: FLOR Miriam Plaza-del-Arco, M. TERESA Martín-Valdivia, SALUD Maria Jimenez-Zafra et al. COPOS: Corpus Of Patient Opinions in Spanish. Application of Sentiment Analysis Techniques. *Natural Language Processing Magazine* [online]. 2016, N°57, 83-90p. Available on: https://www.researchgate.net/publication/308368140_COPOS_Corpus_of_patient_opinions_in_Spanish_Application_of_sentiment_analysis_techniques (accessed 05/04/2020).
- 📖 A [28]: MAITE Taboada, JULIAN Brooke, MILAN Tofiloski et al. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics* [online]. 2011, Volume 37, Number 2, p.267-307 Available on: https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00049 (accessed 05/04/2020).

-  **A [29]:** GHAZALEH Beigi, XIA Hu, ROSS Maciejewski et al. An Overview of Sentiment Analysis in Social Media and its Applications in Disaster Relief. Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence [**online**]. 2015, DOI: 10.1007/978-3-319-30319-2_13, pp 313-340 .Available on: https://www.researchgate.net/publication/288516377_An_Overview_of_Sentiment_Analysis_in_Social_Media_and_Its_Applications_in_Disaster_Relief (accessed 06/04/2020).
-  **A [30]:** ANDREA Ceron, LUIGI Curini, STEFANO M Iacus et al. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media & Society [**online**]. 2014, Volume: 16 issue: 2, page(s): 340-358. Available on : https://www.researchgate.net/publication/277718038_Every_Tweet_Counts_How_Sentiment_Analysis_of_Social_Media_Can_Improve_Our_Knowledge_of_Citizens'_Political_Preferences_with_an_Application_to_Italy_and_France (accessed 06/04/2020).
-  **A [31]:** JASMINA Smailović, MIHA Grčar, NADA Lavrač et al. Predictive Sentiment Analysis of Tweets: A Stock Market Application. From book Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data: Third International Workshop, HCI-KDD 2013, Held at SouthCHI 2013, Maribor, Slovenia, July 1-3, 2013. Proceedings. 2013, (pp.77-88). Available on: <http://first.ijs.si/FirstShowcase/Content/pub/HCI-KDD-2013.pdf> (accessed 06/04/2020).
-  **A [32]:** DANIEL Pletea, BOGDAN Vasilescu, ALEXANDER Serebrenik. Security and Emotion: Sentiment Analysis of Security Discussions on GitHub. Proceedings of the 11th Working Conference on Mining Software Repositories [**online**]. 2014, Pages 348–351. Available on: https://www.researchgate.net/publication/264799488_Security_and_emotion_Sentiment_analysis_of_security_discussions_on_GitHub (accessed 06/04/2020).
-  **A [33]:** TANVI Hardeniya and DILIPKUMAR A. Borikar. Dictionary Based Approach to Sentiment Analysis - A Review. International Journal of Advanced Engineering, Management and Science (IJAEMS) [**online**]. 2016, Vol-2, Issue-5, 317-322p. Available on: https://www.academia.edu/26501071/Dictionary_Based_Approach_to_Sentiment_Analysis_A_Review (accessed 06/04/2020).
-  **A [34]:** DARWICH Mohammad, SHAHRUL Azman Mohd Noah, NAZLIA Omar et al. Corpus-Based Techniques for Sentiment Lexicon Generation: A Review. Journal of Digital Information Management [**online**]. 2019, Volume 17, Number 5, 296-305p. Available on: https://www.researchgate.net/publication/337021440_Corpus-Based_Techniques_for_Sentiment_Lexicon_Generation_A_Review (accessed 06/04/2020).
-  **A [35]:** ALI FARGHALY and KHALED SHAALAN, Arabic Natural Language Processing Challenges and Solutions. ACM Transactions on Asian Language Information Processing [**online**]. 2009, Vol.8, No.4 Article 14. Available on: https://www.researchgate.net/publication/206006010_Arabic_Natural_Language_Processing_Challenges_and_Solutions (accessed 07/04/2020).

-  **A [36]:** SHAALAN Khaled, SANJEERA Siddiqui, ALKHATIB Manar et al. Challenges in Arabic Natural Language Processing. In book: Computational Linguistics, Speech and Image Processing for Arabic Language **[online]**. 2018, 59-83p. Available on: https://www.researchgate.net/publication/327753798_Challenges_in_Arabic_Natural_Language_Processing (accessed 07/06/2020).
-  **A [37]:** POPE David and GRIFFITH Josephine. An Analysis of Online Twitter Sentiment Surrounding the European Refugee Crisis. In Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management **[online]**. 2016, Volume 1: KDIR, pages 299-306. Available on: <https://www.semanticscholar.org/paper/An-Analysis-of-Online-Twitter-Sentiment-Surrounding-Pope-Griffith/9f3b1d9a315981035302d8a293586052c97b2fe5> (accessed 08/04/2020).
-  **A [38]:** AL-AYYOUB Mahmoud, BANI ESSA Safa and ALSMADI Izzat. Lexicon-based sentiment analysis of Arabic tweets. International Journal of Social Network Mining **[online]**. 2015. Available on: https://www.researchgate.net/publication/279963345_Lexicon-Based_Sentiment_Analysis_of_Arabic_Tweets (accessed 08/04/2020).
-  **A [39]:** AL-THUBAITYA Abdulmohsen, ALQAHTANIB Qubayl and ALJANDALB Abdulaziz. Sentiment lexicon for sentiment analysis of Saudi dialect tweets. Procedia Computer Science **[online]**. 2018, 142, 301-307. Available on: <https://www.mendeley.com/catalogue/8afacf6d-49ed-39e4-b86e-f9b0fd73c2e4/> (accessed 08/04/2020).
-  **A [40]:** AL-RUBAIEE Hamed, QIU Renxi, ALOMAR Khalid et al. Sentiment Analysis of Arabic Tweets in E-Learning. Journal of Computer Science **[online]**. 2016, 12 (11): 553.56. Available on: https://www.researchgate.net/publication/316585452_Sentiment_Analysis_of_Arabic_Tweets_in_e-Learning (accessed 09/04/2020).
-  **A [41]:** ELOUARDIGHI Abdeljalil, MAGHFOUR Mohcine, HAMMIA Hafdalla et al. Analyse des sentiments à partir des commentaires Facebook publiés en Arabe standard ou dialectal marocain par une approche d'apprentissage automatique. Published in EGC Computer Science **[online]**. 2018. Available on: <https://www.semanticscholar.org/paper/Analyse-des-sentiments-%C3%A0-partir-des-commentaires-en-Elouardighi-Maghfour/6d7e9279aa804f5f8f41fedf8aa0f2b230cb2b37> (accessed 09/04/2020).
-  **A [42]:** M. DUWAIRI Rehab and QARQAZ Islam, Arabic Sentiment Analysis using Supervised Classification. 2014 International Conference on Future Internet of Things and Cloud. 27-29 Aug. 2014. Barcelona, Spain. Institute of Electrical and Electronics Engineers, 15 December 2014. Available on: <https://ieeexplore.ieee.org/document/6984256> (accessed 10/04/2020).
-  **A [43]:** SUMITA Sharma and Dr. MAMTA Bansal, Classification Approach for Sentiment Analysis. INTERNATIONAL JOURNAL OF RESEARCH IN ELECTRONICS AND COMPUTER ENGINEERING **[online]**. 2018, VOL 6, ISSUE 3. Available on:

https://www.researchgate.net/publication/337590789_Classification_Approach_for_Sentiment_Analysis (accessed 10/04/2020).

- 📖 **A [44]:** K. ALDAYEL Haifa and M. AZMI Aqil. Arabic tweets sentiment analysis – a hybrid scheme. Journal of Information Science [online]. 2016, Vol. 42(6) 782 –797. Available on: https://www.researchgate.net/publication/283664122_Arabic_tweets_sentiment_analysis_-_A_hybrid_scheme (10/04/2020).
- 📖 **B [1]:** JEAN-charles Cointot and YEVES Eychen. La Révolution Big data- Les données au cœur de la transformation d’entreprise. 5 rue Laromiguière, 75005 Paris: Dunod. 2014. 236p.
- 📖 **B [2]:** NITIN Indurkha and FRED J. Damerau. Handbook of Natural Language Processing [online]. Second edition. 6000 Broken Sound Parkway NW: CRC Press, 2010, 704p. Available on: https://books.google.dz/books?id=nK-QYHZ0-gC&hl=fr&source=gbs_similarbooks (accessed 11/06/2020).
- 📖 **B [3]:** PIRMIN Lemberger. Big Data et Machine Learning: les concepts et les outils de la data science. Second Edition. 11 rue paul bert 92240 malakoff: Dunod, 2016, 255p.
- 📖 **B [4]:** GARRY Turkington, TANMAY Deshpande and SANDEEP Karanth. Hadoop: Data Processing and Modelling. [online]. Birmingham B3 2PB, UK: Packt Publishing Ltd, 2016, 979p. Available on: https://books.google.dz/books?id=LWzWDQAAQBAJ&hl=fr&source=gbs_book_other_versions (accessed 11/06/2020).
- 📖 **B [5]:** BEN Coppin. Artificial Intelligence Illuminated. [online]. Sudbury Canada: Jones and Bartlett Publishers, 2004, 739p. Available on: <https://www.pdfdrive.com/artificial-intelligence-illuminated-e1083124.html> (accessed 11/06/2020).
- 📖 **B [6]:** LEVENE Mark, LOIZOU George. A Guided Tour of Relational Databases and Beyond [online]. Springer Science & Business Media, 1999 - 625 pages: Available on: https://books.google.dz/books/about/A_Guided_Tour_of_Relational_Databases_an.html?id=CkYpI7QsLIQC&redir_esc=y (accessed 11/06/2020).
- 📖 **B [7]:** PHILIP Russom. Big Data Analytics-FOURTH QUARTER. [online]. TDWI (The Data Warehousing InstituteTM) Research, 2011. Available on: ftp://ftp.software.ibm.com/software/tw/Defining_Big_Data_through_3V_v.pdf (accessed 11/06/2020).
- 📖 **B [8]:** EMC Education Services. Data Science and Big Data Analytics [online]. John Wiley & Sons, Inc., Indianapolis, Indiana. 2014 - 435 pages. Available on: <https://www.pdfdrive.com/data-science-and-big-data-analytics-e58447171.html> (accessed 11/06/2020).
- 📖 **B [9]:** SORAYA Sedkaoui. Data Analytics and Big Data [online]. United States: John Wiley & Sons, 2018 - 224 pages. Available on: <https://books.google.dz/books?id=0pFeDwAAQBAJ&printsec=frontcover&dq=Data+Analytics+and+Big+Data&hl=fr&sa=X&ved=2ahUKEwiR7rWbtMvrAhVDOBoKHTRPDr>

[cQuwUwAHoECAEQCQ#v=onepage&q=Data%20Analytics%20and%20Big%20Data&f=false](https://books.google.dz/books/about/Data_Mining_Concepts_and_Techniques.html?id=pQws07tdpjoC&redir_esc=y) (accessed 12/06/2020).

- 📖 **B [10]:** JIAWEI Han, JIAN Pei and MICHELINE Kamber. Data Mining: Concepts and Techniques. Wyman street, Waltham, MA 02451 USA: Elsevier, 2011 - 744 pages. Available on: https://books.google.dz/books/about/Data_Mining_Concepts_and_Techniques.html?id=pQws07tdpjoC&redir_esc=y (accessed 12/06/2020).
- 📖 **B [11]:** NILS J. Nilsson. Artificial Intelligence: A New Synthesis [online]. San Francisco, CA 94104-3205 USA: Morgan Kaufmann Publishers. 1998. 493p. Available on : <https://books.google.dz/books?id=GYOFSd6fETgC&printsec=frontcover&dq=Artificial+Intelligence:+A+New+Synthesis&hl=fr&sa=X&ved=2ahUKEwiWlp3MvsvrAhVkoUKHbKCBpoQuwUwAHoECAUQCg#v=onepage&q=Artificial%20Intelligence%3A%20A%20New%20Synthesis&f=false> (accessed 18/06/2020).
- 📖 **B [12]:** MASSIH Reza-AMINI, Apprentissage machine de la théorie à la pratique. 75240 Paris Cedex 05: Edition Eyrolles, 2015, 267p.
- 📖 **B [13]:** JAMES Seligman, Artificial intelligence + Machine learning in marketing management [online]. First Edition. Publisher: Lulu.com, 2018, 319p. Available on: https://www.researchgate.net/publication/327768607_ARTIFICIAL_INTELLIGENCE_MACHINE_LEARNING_in_MARKETING_MANAGEMENT (accessed 18/06/2020).
- 📖 **B [14]:** CHRISTOPHER M.Bishop. Pattern Recognition and Machine Learning [online]. University of California, Berkeley, CA 94720, USA: Springer Science+Business Media, LLC. 2006, 729p. Available on: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf> (accessed 18/06/2020).
- 📖 **B [15]:** N. SEBE, IRA COHEN,ASHUTOSH GARG et al. Machine Learning in Computer Vision [online]. Dordrecht, The Netherlands: published by Springer, 2005, 237p. Available on: https://books.google.dz/books/about/Machine_Learning_in_Computer_Vision.html?id=le mw2Rhr_PEC&redir_esc=y (accessed 19/06/2020).
- 📖 **B [16]:** ANDREAS C. Müller & SARAH Guido. Introduction to machine learning with python. First Edition. 1005 Gravenstein Highway North, Sebastopol, CA 95472: O'Reilly Media, 2016, 367p.
- 📖 **B [17]:** LIPO Wang, Support Vector Machine: Theory and Applications [online]. First Edition. Berlin Heidelberg: Springer-Publishing Company, 2005, 431p. Available on: https://books.google.dz/books/about/Support_Vector_Machines_Theory_and_Appli.html?id=uTzMPJjVjsMC&redir_esc=y (accessed 19/06/2020).
- 📖 **B [18]:** YU.L Pavlov. Random Forest [online]. First Edition. The Netherland: Ridderprint BV Ridderkerk, 2000, 122P Available on: https://books.google.dz/books/about/Random_Forests.html?id=07gpKU3npYUC&redir_esc=y (accessed 19/06/2020).

- 📖 **B [19]:** B. YEGNANARAYANA. Artificial Neural Networks **[online]**. First Edition. New-Delhi India: Prentice-Hall of India Pvt.Ltd, 2006, 476p. Available on: https://books.google.dz/books/about/ARTIFICIAL_NEURAL_NETWORKS.html?hl=fr&id=RTtvUVU_xL4C&redir_esc=y (accessed 19/06/2020).
- 📖 **B [20]:** KEVIN L. Priddy, PAUL E. Keller. Artificial Neural Networks: An Introduction **[online]**. United States of America: SPIE (Society of Photo-Optical Instrumentation Engineers), 2005, 165p. Available on: https://books.google.dz/books/about/Artificial_Neural_Networks.html?id=BrnHR7esWmkC&redir_esc=y (accessed 19/06/2020).
- 📖 **B [21]:** LUDWIG Fahrmeir, THOMAS Kneib, And STEFAN Lang et al. Regression: Models, Methods and Applications **[online]**. First Edition. Berlin Heidelberg: Springer-Publishing Company, 2013, 698p. Available on: https://books.google.dz/books/about/Regression.html?id=EQxU9iJtipAC&redir_esc=y (accessed 20/06/2020).
- 📖 **B [22]:** AWAD Mariette and KHANNA Rahul. Support Vector Machine **In:** Efficient Learning machines. First Edition. ApressOpen, 2015, 70-82p.
- 📖 **B [23]:** SANFORD Weisberg. Applied Linear Regression **[online]**. Third Edition. Hoboken-New Jersey: John Wiley & Sons, 2005, 352P. Available on: https://books.google.dz/books/about/Applied_Linear_Regression.html?id=xd0tNdFOOjcC&redir_esc=y (accessed 20/06/2020).
- 📖 **B [24]:** WU Junjie. Advances in K-means Clustering: A Data Mining Thinking **[online]**. First Edition. Berlin Heidelberg: Springer-Publishing Company, 2012, 180p. Available on: https://books.google.dz/books/about/Advances_in_K_means_Clustering.html?id=pI2_F8SqWcQC&redir_esc=y (accessed 20/06/2020).
- 📖 **B [25]:** G.Barto Andrew and S. Sutton Richard. Reinforcement Learning: An Introduction **[online]**. First Edition. London-England: The MIT press, 2014.2025, 352p. Available on: <https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf> (accessed 20/06/2020).
- 📖 **B [26]:** IAN Goodfellow, YOSHUA Bengio and AARON Courville. Deep learning **[online]**. London England: The MIT press, 2016, 775p. Available on: https://books.google.dz/books/about/Deep_Learning.html?id=Np9SDQAAQBAJ&redir_esc=y (accessed 20/06/2020).
- 📖 **B [27]:** LARRY Medsker, LAKHMI C. Jain. Recurrent Neural Networks: Design and Applications **[online]**. CRC-press (Chemical Rubber Company) 2001, 416p. Available on: <https://doc.lagout.org/science/Artificial%20Intelligence/Neural%20networks/Recurrent%20Neural%20Networks%20Design%20And%20Applications%20-%20L.R.%20Medsker.pdf> (accessed 20/06/2020).
- 📖 **B [28]:** JASON Brownlee. Long Short-Term Memory Networks with Python **[online]**. Edition 1v.5. Machine Learning Mastery, 2019, 246p. Available on: <https://books.google.dz/books?id=m7SoDwAAQBAJ&dq=Jason+Brownlee,+Long+Shor>

[t-Term+Memory+Networks+With+Python,+2019.&hl=fr&source=gb_s_navlinks_s](#)
(accessed 20/06/2020).

- 📖 **B [29]:** FEDERICO Alberto Pozzi, ELISABETTA Fersini and ENZA Messina et al. Sentiment Analysis in Social Networks. First Edition. Elsevier Science. 2016. 284p.

- 📖 **B [30]:** BING Liu. Sentiment analysis and opinion mining **[online]**. Morgan & Claypool Publishers, 2012 - 167 pages. Available on: https://books.google.dz/books/about/Sentiment_Analysis_and_Opinion_Mining.html?id=Gt8g72e6MuEC&redir_esc=y (accessed 04/04/2020).

- 📖 **B [31]:** STEVEN Bird, EWAN Klein and EDWARD Lope. Natural Language Processing with Python: Analyzing Text with the Natural **[online]**. First Edition. Gravenstein highway north, Sebastopol: O'Reilly Media, 2009 - 504 pages. Available on: <https://www.nltk.org/book/> (accessed 07/04/2020).

- 📖 **B [32]:** Y.HABASH Nizar. Arabic Natural Language Processing **[online]**. Morgan & Claypool Publishers, 2010 - 167 pages. Available on: https://books.google.dz/books/about/Introduction_to_Arabic_Natural_Language.html?id=kRIHCnC74BoC&redir_esc=y (accessed 07/04/2020).

- 📖 **B [33]:** Jeff Forcier, Paul Bissex and Wesley J Chun. Python Web Development with Django **[online]**. Addison-Wesley, 2009 - 377 pages. Available on: <https://books.google.dz/books?id=M2D5nnYlmZoC&printsec=frontcover&dq=Python+Web+Development+with+Django&hl=fr&sa=X&ved=2ahUKEwiCjr-4u9LrAhVx5eAKHRTDAQgQ6wEwAXoECAUQAQ#v=onepage&q=Python%20Web%20Development%20with%20Django&f=false> (accessed 12/08/2020).

- 📖 **B [34]:** [JASON Brownlee](#). Long Short-Term Memory Network with Python: Develop Sequence Prediction Models with Deep Learning **[online]**. Machine Learning Mastery. 2017 - 246 pages Available on: https://books.google.dz/books?id=m7SoDwAAQBAJ&hl=fr&source=gb_s_book_other_versions (accessed 20/08/2020).

Webography:

- 🌐 **WEB [1]:** EMERSAY. Top 5 Social Media Predictions for 2019 [online]. Available on: <https://emarsys.com/learn/> (accessed 22/02/2020).
- 🌐 **WEB [2]:** Enterprise Big Data Framework. An overview of the Big Data Framework [online]. 7 mai. 2019. Available on: <https://www.bigdataframework.org> (accessed 24/02/2020).
- 🌐 **WEB [3]:** [Apache Hadoop](https://hadoop.apache.org/). Apache Hadoop [online]. Available on : <https://hadoop.apache.org/> (accessed 24/02/2020).
- 🌐 **WEB [4]:** [Apache Storm](https://storm.apache.org/index.html). Apache Storm [online]. Available on: <https://storm.apache.org/index.html> (accessed 24/02/2020).
- 🌐 **WEB [5]:** [Apache Samza](http://samza.apache.org/). Apache Samza [online]. Available on: <http://samza.apache.org/> (accessed 25/02/2020).
- 🌐 **WEB [6]:** Apache Spark. [Apache Spark™ - Unified Analytics Engine for Big Data](https://spark.apache.org/) [online]. Available on: <https://spark.apache.org/> (accessed 25/02/2020).
- 🌐 **WEB [7]:** Facebook. [About Facebook](https://www.facebook.com/about). [online]. Available on: <https://www.facebook.com/about> (accessed 26/02/2020).
- 🌐 **WEB [8]:** YouTube. About YouTube [online]. Available on: <https://www.youtube.com/about/>. (accessed 11/06/2020).
- 🌐 **WEB [09]:** SINAI-Intelligent Access to Information Systems. Home [online]. Available on: <http://sinai.ujaen.es/> /// <http://150.214.174.171:8059/en> (accessed 28/07/2020).
- 🌐 **WEB [10]:** SINAI-Intelligent Access to Information Systems. Home [online]. Available on: <http://150.214.174.171:8059/en/research/resources/copos> (accessed 28/07/2020).
- 🌐 **WEB [11]:** GitHub [online]. Available on: <https://github.com/> (accessed 29/07/2020).
- 🌐 **WEB [12]:** NLTK-Natural Language Toolkit. NLTK 3.5 Documentation [online]. Available on: <https://www.nltk.org/> (accessed 29/07/2020).
- 🌐 **WEB [13]:** Socialmention [online]. Available on: <http://socialmention.com/> (accessed 29/07/2020).
- 🌐 **WEB [14]:** RapidMiner [online]. Available on: <https://rapidminer.com/> (accessed 29/07/2020).
- 🌐 **WEB [15]:** Brandwatch [online]. Available on: <https://www.brandwatch.com/fr/> (accessed 29/07/2020).
- 🌐 **WEB [16]:** LIWC-Linguistic Inquiry and Word Count. DISCOVER LIWC2015 [online]. Available on: <https://liwc.wpengine.com/> (accessed 08/08/2020).

- 🌐 **WEB [17]:** The Stanford Natural Language Processing Group. Stanford Log-linear Part-Of-Speech Tagger [**online**]. Available on: <https://nlp.stanford.edu/software/tagger.shtml> (accessed 08/08/2020).
- 🌐 **WEB [18]:** Stanford Parser [**online**]. Available on: <http://nlp.stanford.edu:8080/parser/> (accessed 09/08/2020).
- 🌐 **WEB [19]:** ESULI Andrea. GitHub [**online**]. Available on: <https://github.com/aesuli/SentiWordNet> (accessed 09/08/2020).
- 🌐 **WEB [20]:** Python. About [**online**]. Available on: <https://www.python.org/doc/essays/blurb/> (accessed 12/08/2020).
- 🌐 **WEB [21]:** Pandas. Getting started [**online**]. Available on: <https://pandas.pydata.org/about/> (accessed 14/08/2020).
- 🌐 **WEB [22]:** Scikit-learn machine learning in python. Getting started [**online**]. Available on: <https://scikit-learn.org/stable/> (accessed 12/08/2020).
- 🌐 **WEB [23]:** Tensorflow [**online**]. Available on: <https://www.tensorflow.org/> (accessed 14/08/2020).
- 🌐 **WEB [24]:** Jupyter. Install [**online**]. Available on: <https://jupyter.org/> (accessed 20/08/2020).
- 🌐 **WEB [25]:** Colaboratory. Bienvenue dans Colaboratory [**online**]. Available on: <https://colab.research.google.com/notebooks/intro.ipynb> (accessed 20/08/2020)

Abstract

Sentiment analysis is one of the most active research areas in natural language processing in recent years. It is the computational study of sentiments and emotions expressed in written human languages. While emotions are central to almost all human activities, especially in Social Media, social big data becomes one of the most important source of data for sentiment analysis. This work consists in developing a platform that presents our contribution in the field of classification of the polarity (Positive/Negative) by using machine learning techniques (classification model) and a generated database. This last is consisted of Twitter comments or “tweets” written in both Modern Standard Arabic (MSA) and Algerian Dialectal Arabic. For the favour of the sentiment analysis field, in our work we fortunately managed the use of *Random Forest* learning model.

Keywords: Social Big Data, Sentiment Analysis, Natural Language Processing, Machine Learning, Algerian Dialectal Arabic, Twitter, Random Forest.

Résumé

Récemment, *l'analyse des sentiments* est l'un des domaines de recherche les plus actifs dans le traitement du langage naturel. C'est l'étude computationnelle des sentiments et des émotions exprimés dans le langage naturel écrit. Tant que les émotions sont au cœur de presque toutes les activités humaines, en particulier dans les réseaux sociaux, les méga données sociaux devenus l'une des sources de données les plus importantes pour l'analyse des sentiments. Ce travail consiste à développer une plateforme qui présente notre contribution dans le domaine de la classification de la polarité (positive / négative) en utilisant techniques d'apprentissage automatique (un modèle de classification) et une base de données générée. Cette dernière est composée de commentaires Twitter ou «tweets» rédigés à la fois en Arabe Standard Moderne (ASM) et en Arabe Dialectal Algérienne. Dans notre travail nous avons pris l'initiative d'utilisé avec réussite *Random Forest* comme modèle d'apprentissage dans le contexte de l'analyse des sentiments.

Mots clé : Méga Données Sociaux, L'analyse des sentiments, Traitement du langage naturel, apprentissage automatique, Arabe Dialectal Algérienne, Twitter, Random Forest.