PEOPLE'S DEMOCRATIC REPUBLICC OF ALGERIA

Ministery of Higher Education and Scientific Research

University of abou Baker BelkaidTlemcen

Faculty of Technology

Biomedical engineering department

Thesis project

To obtain :

A MASTER degree in Biomedical Engineering

Specialty :

Biomedical informatics

AUTOMATIC AND INTELLIGENT SENTIMENT ANALYSIS AND OPINION MINING ON SOCIAL MEDIA AND THE WEB IN THE MEDICAL FIELD

Presented by :

- Ikram HAMDI

Presenting on 10th of September , in front of the juries :

- Prof. Chikh Mohammed Amine	(President)
- Dr. Abderrahim Mohammed El Amine	(supervisor)
- Dr. Bechar Hassane	(Examiner)

University year 2019-2020

ACKNOWLEDGEMENT

This work would never have been possible without the support of a group of people whom I would like to thank here warmly.

I would acknowledge my supervisor Dr. ABDERRAHIM MOHAMED AMINE for his guidance, help and support, not just during my research, but also in my studies. His pieces of advice were really of a tremendous help to refine my work

I would also like to thank all the members of the jury, Dr. BECHAR HASENE and Prof. CHIKH MOHAMMED EL AMINE for having done me the honor of accepting to judge this master's thesis.

Of course, words cannot describe the immense gratitude to all my teachers in the Department of Biomedical Engineering, who provided me with a quality education during all my years of study at Tlemcen University.

Dedication

Every challenging work needs self-efforts as guidance of olders especially those who were very close to our heart .

My humble effort I dedicate

To

My sweet and loving

Mother & Father

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

What a good feeling? Ihave when, Iseethat it is time to dedicate my studies to my dearest husband “Mohamed” Who is shiming my way by his care,love and warms encouragement.

My soulmates IMANE , INSAF , HANONE

Thank you for being next to me in the hard moments

The index

List of Figures

Tables list

General Introduction

Social media analytics is a research axis focused on extracting useful insights from social media data, with the aim of helping individuals and organizations take the most optimum decisions regarding several disciplines of life (business, marketing, politics, health, etc

Sentiment analysis has become quite popular among different brands. So many organizations are now incorporating it into their business to gain useful insight and automate processes.

Some of the benefits of sentiment analysis include:

- **Optimize marketing Strategy**: Although so many organizations use social media to promote their brands; carrying out sentiment analysis enables organizations to optimize their marketing strategy.

Through sentiment analysis, organizations can know if their social media marketing strategy is right and the necessary changes they must put in place to make them stand out.

- **Evaluate ROI on marketing campaign**: The success of every marketing campaign does not only lead to an increase in likes, comments, and followers on the companies social media pages.

It also leads to an increase in the number of positive discussions about the brand among customers which helps the firm to measure its ROI on marketing campaigns.

- **Increase product quality**: Knowing customers opinion about your products and services enables organizations to discover key areas that require improvement in order to meet the customers need. Remember that products are services are not judged only by how well they performed but also how they are presented.

You can only get ideas on how to improve your products from your customers through surveys or casual discussions.

• **Improve customer service**: Sentiment analysis helps organizations manage customers' complaints and avoid leaving them feeling ignored.

It picks out negative discussion and alerts the customer service team members so

that you can respond to them when a customer complains about your products, the earlier you react, the better they get to forget about it, and this helps you have a satisfied customer.

• **Lead generation**: Adjusting your marketing campaign, having excellent customer support, and improving product quality to meet the customers need can increase sales and help you gain more customers.

Sentiment analysis also helps you discover what your customers need and help you create products that attract new and loyal customers. • Sales Revenue: One of the most significant benefits of sentiment analysis is to boost sales revenue.

When a company has a good customer service and improved product quality which can be achieved through sentiment analysis the sales revenue will automatically scale up.

The fact remains that when there are more positive discussions about your product on social media, your sales revenue will increase and when the discussions are negative reverse will be the case.

• **Crisis managemen**t: Sentiment analysis also plays a huge role in helping organizations manage the crisis. This is because constant monitoring of social media conversation can help mitigate against damage which could be caused by bad product quality, unacceptable customer service or other environmental harms in an emerging market. Once not managed on time, such negative feedbacks might go viral and cause a huge crisis to the image of the organization.

This thesis is organized into 03 chapters, preceded by a general introduction and followed by a general conclusion:

- The first chapter presents the social media, social network, the blogs, themicroblog and in the end we have different information about tweeter .
- The second chapter we speak about sentiment analysis , its types , how sentiment analysis work , and finally, its approach .
- The third chapter presents the steps followed in our work

CHAPTER 1:

SOCIAL MEDIA

1. Introduction

The definition of social media seems like an oversimplification. In the last several years, technology has brought us very far from where we started and social media almost seems like it is an entirely different animal. Social media and social networking have been instrumental in many major events around the world. It is fair to say that social networking is a subcategory of social media. Many people think that social media and social networking are one and the same and therefore can be used interchangeably. That is a misconception. It is a good idea to look at the differences between the two here and walk away with a clear understanding of the differences. (COHN, social media, 2011)

2. Developing a Social Media Strategy

Developing a social media strategy is an important part of making sure that each of your campaigns is helping we get closer to meeting our marketing goals. After we have decided which channels we plan to use, we all need to consider how often we plan to post and what types of content we will post. Start by considering what we know about our audience.Develop content that's engaging but also speaks to their needs and challenges. If we need help getting started, check out these social media marketing ideas .We should also consider how we plan to make engagement and response a consistent part of our social media management. (lyfemarketing, 2020)

3. Mobile Social Media

Mobile social media refer to the use of social media on mobile devices such as smartphones and tablet computers . Mobile social media are a useful application of mobile marketing because the creation, exchange, and circulation of user-generated content can assist companies with marketing research, communication, and relationship development. Mobile social media differ from others because they incorporate the current location of the user (location-sensitivity) or the time delay between sending and receiving messages (time-sensitivity). According to Andreas Kaplan, mobile social media applications can be differentiated among four types:

1. *Space-timers* (location and time sensitive): Exchange of messages with relevance mostly for one specific location at one specific point in time (e.g. Facebook Places WhatsApp).

2. *Space-locators* (only location sensitive): Exchange of messages, with relevance for one specific location, which is tagged to a certain place and read later by others (e.g. Yelp; Qype ,Tumbir ,Fishbrain)
3. *Quick-timers* (only time sensitive): Transfer of traditional social media mobile apps to increase immediacy (e.g. posting Twitter messages or Facebook status updates)
4. *Slow-timers* (neither location nor time sensitive): Transfer of traditional social media applications to mobile devices (e.g. watching a YouTube video or reading/editing a Wikipedia article) (wikipedia, Mobile social media, 2020)

4. Social Network

4.1 Definition

Social network is a website that brings people together to talk, share ideas and interests, or make new friends. This type of collaboration and sharing is known as social media. Unlike traditional media that is created by no more than ten people, social media sites contain content created by hundreds or even millions of different people. Below is a small list of some of the biggest social networks used today. (Hope, 2020)

4.2 Definition Of Social Media

If we consider what we understood media to be before the Internet existed, it was about television, newspapers, magazines, etc. Once media became available through the World Wide Web, the media was no longer static. Tremendous interactivity capabilities became available to everyone and it felt much more like a personal, one-on-one relationship than anything else. At the heart of social media are relationships, which is in common with social networking. Social media is a very broad term and really encompasses several different types of media, such as videos, blogs, etc. Social media is a place where we can transmit information to other people.

Social media is a vehicle for communication. Social media lets everyone share content that other people can share, in turn, with their online connections. We create the buzz through social media. (COHN, Social media, 2011)

Figure 1. Social Network

4.3 Social media and social networking

One thing that social media and social networking have in common is that they both depend on viral marketing to become truly successful. If the content goes viral, more and more people will be paying attention and the more online traffic we have, the better our chances are of increasing our business. A simple way to look at the basic difference between social media and social networking is that social media helps people to make the connection and social networking enhances that connection. People get together because they have common interests, passions, and causes and they continue to strengthen their relationships as they get to know each other through interaction over time. (COHN, social media and social networking have in common, 2011)

4.4 Examples Of Social Networks

- Classmates https://www.classmates.com/
One of the largest and most used websites for connecting people who graduated from a high school and allows we to keep in touch with them and any future reunions.

- **DeviantArt**https://www.deviantart.com/
A social media platform for sharing original artwork.

- **Facebook** https://www.facebook.com/
The most popular social networking websites on the Internet. Facebook is a popular destination for users to set up personal space and connect with friends, share pictures, share movies, talk about what we are doing, etc.

- **Google+** https://plus.google.com/
The latest social networking service from Google.

- **Instagram** https://www.instagram.com/
A mobile photo sharing service and application available for the iPhone, Android, and Windows Phone platforms.

- **LinkedIn** https://www.linkedin.com/feed/
One of the best if not the best locations to connect with current and past coworkers and potentially future employers.

- **Mastodon**https://joinmastodon.org/
A free, federated, social microblogging service with over two million users. Any Mastodon user can operate a node (social subdomain) with its own theme and set of rules.

- **Mix** https://mix.com/
Another very popular community of Internet users who vote for web pages they like and dislike. Mix also allows users to create personal pages of interesting sites they come across.

- **MySpace** https://myspace.com/

Once one of the most popular social networks and viewed website on the Internet.

- **Pinterest**https://www.pinterest.com/
A popular picture and sharing service that allows anyone to share pictures, create collections, and more.

- **Reddit** https://www.reddit.com/
Community of registered users (redditors) submits content that is upvoted by the community. Reddit has a subreddit (board) for almost every category.

- **Twitter** https://twitter.com/
Another fantastic service that allows users to post 140 character long posts from their phones and on the Internet. A fantastic way to get the pulse of what's going on around the world.

- YouTube https://www.youtube.com/
An excellent network of users posting video blogs or vlogs and other fun and exciting videos. (computerhope, Examples of social networks, 2020)

4.5 The first social media website

The Bolt.com social networking website was created in 1996, by Jane Mount and Dan Pelson. Although it is not considered the first true social media website, it technically was the first to be created. It was officially shut down in October 2008.
The first true social media website is considered to be SixDegrees.com, created in 1997 by Andrew Weinreich. SixDegrees.com is still in operation today, and is available at http://sixdegrees.com. (computerhope, the first social media website, 2020).

5. Blog

A blog (shortening of "weblog") is an online journal or informational website displaying information in the reverse chronological order, with the latest posts appearing first. It is a platform where a writer or even a group of writers share their views on an individual subject . (Skrba, 2018)

6. Microblog

A microblog is a type of blog in which users can post small pieces of digital content like pictures, video or audio on the Internet. These posts, called microposts, are immediately available to a small community or public. It differs from a blog due to its smaller content.
Microblogging is highly popular among users due to its portability and immediacy. (Techopedia, 2016)

7. Twitter

7.1 Introduction

Figure 2 Twitter Logo

Twitter is an American microblogging and social networking service on which users post and interact with messages known as "tweets". Registered users can post, like, and retweet tweets, but unregistered users can only read them. Users access Twitter through its website interface,

through Short Message Service (SMS) or its mobile-device application software .("app").

Twitter, Inc. is based in San Francisco, California, and has more than 25 offices around the world. Tweets were originally restricted to 140 characters, but was doubled to 280 for non-Asian languages in November 2017.

Twitter is an online social networking service that enables users to send and read short 140- character messages called "tweets".

Registered users can read and post tweets from their accounts.

Friends, family, coworkers can communicate and stay connected through the exchange of quick, frequent messages. Tweets can contain words, photos, videos and links. (Wikipedia, 2020)

7.2 Twitter Terminology

- Retweet: a tweet that we forward to our followers that always retains original attribution.

- @: the @ sign is used to call out usernames in tweets.

- hachtag: any word or phrase immediately preceded by the (hachtag) symbol. By clicking hachtag you can see other tweets containing the same keyword or topic.

- Follower: another person who receives our tweets on their Home stream.

- Direct Messages: private messages sent from one Twitter user to another. It can also be used in groups.

Home: timeline displaying a stream of tweets from accounts we have chosen to follow. (Sengottaiyan, 2009)

7.3 Twitter Is So Popular

In addition to its relative novelty, Twitter's big appeal is how scan-friendly it is: we can track hundreds of interesting Twitter users and read their content with a glance. This is ideal for our modern attention-deficit world.

Twitter employs a purposeful message size restriction to keep things scan-friendly: every microblog tweet entry is limited to 280 characters or less. This size cap promotes the focused and clever use of language, which makes tweets easy to scan, and challenging to write. This size restriction made Twitter a popular social tool. (Dharmendra, 2020)

CHAPTER 2: SENTIMENT ANALYSIS

1. introduction

The use of social media has become an integral part of daily routine in modern society. Social media portals offer powerful public platforms where people can freely share their opinions and feelings about various topics with large crowds. (Öztürk, 2018)
Sentiment analysis and opinion mining has become a major tool for collecting information from customer reviews on user sentiments and emotions, especially for online video streaming services and social networks. The increasing use of smartphones has popularized subscription to various streaming services that provide streaming media and video-on-demand. (H. Sankar and V. Subramaniyaswamy, 2019)

The use of sentiment analysis

- Business: In marketing field companies use it to develop their strategies, to understand customers' feelings towards products or brand, how people respond to their campaigns or product launches and why consumers don't buy some products.
- . Politics: In political field, it is used to keep track of political view, to detect consistency and inconsistency between statements and actions at the government level.
- Public Actions: Sentiment analysis also is used to monitor and analyse social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

2. Sentiment Analysis

Sentiment Analysis is a term that you must have heard if you have been in the Tech field long enough. It is the process of predicting whether a piece of information (i.e. text, most commonly) indicates a positive, negative or neutral sentiment on the topic (Al-Masri, 2019).
- Sentiment analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. Sentiment analysis allows businesses to identify customer sentiment toward products, brands or services in online conversations and feedback. (monkeylearn, Sentiment Analysis, 2020)

3. Types of Sentiment Analysis

Sentiment analysis assumes various forms, from models that focus on polarity (positive,negative, neutral) to those that detect feelings and emotions (angry, happy, sad, etc), or even models that identify intentions (e.g. interested v. not interested).
Here are some of the most popular types of sentiment analysis: (monkeylearn, Types of Sentiment Analysis, 2020)

3.1. Fine-grained Sentiment Analysis

If polarity precision is important to our business, we might consider expanding our polarity categories to include:

- Very positive
- Positive
- Neutral
- Negative
- Very negative: this is usually referred to as fine-grained sentiment analysis, and could be used to interpret 5-star ratings in a review, for example:
- Very Positive = 5 stars
- VeryNegative = 1 star
  (monkeylearn, Fine-grained Sentiment Analysis, 2020)

3.2. Emotion detection :

Emotion detection aims at detecting emotions, like happiness, frustration, anger, sadness, and so on. Many emotion detection systems use lexicons (i.e. lists of words and the emotions they convey) or complex machine learning algorithms.
One of the downsides of using lexicons is that people express emotions in different ways. Some words that typically express anger, like bad or kill (e.g. your product is so bad or your customer support is killing me) might also express happiness (e.g. this is bad ass or you are killing it). (monkeylearn, Emotion detection, 2020)

4. Sentiment analysis work

Sentiment analysis uses various Natural Language Processing (NLP) methods and algorithms, which we'll go over in more detail in this section. and the main types of algorithms used include:

4.1. Rule-based Approaches

Usually, a rule-based system uses a set of human-crafted rules to help identify subjectivity, polarity, or the subject of an opinion.
These rules may include various techniques developed in computational linguistics, such as:

- **Stemming**, tokenization, part-of-speech tagging and parsing.
- **Lexicons** (i.e. lists of words and expressions).

Here's a basic example of how a rule-based system works:
1 - Defines two lists of polarized words (e.g. negative words such as bad, worst, ugly, etc and positive words such as good, best, beautiful, etc).
2 - Counts the number of positive and negative words that appear in a given text.
3 - If the number of positive word appearances is greater than the number of negative word appearances, the system returns a positive sentiment, and vice versa. If the numbers are even, the system will return a neutral sentiment.
Rule-based systems are very naive since they don't take into account how words are combined in a sequence. Of course, more advanced processing techniques can be used, and new rules added to support new expressions and vocabulary. However, adding new rules may affect previous results, and the whole system can get very complex.
Since rule-based systems often require fine-tuning and maintenance, they'll also need regular investments.

4.2. Automatic Approaches

Automatic methods, contrary to rule-based systems, don't rely on manually crafted rules, but on machine learning techniques. A sentiment analysis task is usually modeled as a classification problem, whereby a
classifier is fed a text and returns a category, e.g. positive, negative, or neutral.

Here's how a machine learning classifier can be implemented:

The Training and Prediction Processes

In the training process (a), our model learns to associate a particular input (i.e. a text) to the corresponding output (tag) based on the test samples used for training. The feature extractor transfers the text input into a feature vector. Pairs of feature vectors and tags (e.g. positive, negative, or neutral) are fed into the machine learning algorithm to generate a model.
In the prediction process (b), the feature extractor is used to transform unseen text inputs into feature vectors. These feature vectors are then fed into the model, which generates predicted tags (again, positive, negative or neutral).

Feature Extraction from Text

The first step in a machine learning text classifier is to transform the text extraction or text vectorization, and the classical approach has been bag-of-words or bag-of-ngrams with their frequency.
More recently, new feature extraction techniques have been applied based on word embeddings (also known as word vectors). This kind of representations makes it possible for words with similar meaning to have a similar representation, which can improve the performance of classifiers.

Classification Algorithms

The classification step usually involves a statistical model like Naïve Bayes, Logistic Regression, Support Vector
Machines, or Neural Networks:

- Naïve Bayes: a family of probabilistic algorithms that uses Bayes's Theorem to predict the category of a text.
- Linear Regression: a very well-known algorithm in statistics used to predict some value (Y) given a set of features (X).
- Support Vector Machines: a non-probabilistic model which uses a representation of text examples as points in a multidimensional space. Examples of different categories (sentiments) are mapped to distinct regions within that space. Then, new texts are assigned a category based on similarities with existing text sand the regions they're mapped to.
- Deep Learning: a diverse set of algorithms that attempt to mimic the human brain, by employing artificial neural networks to process data.

4.3. Hybrid Approaches

Hybrid systems combine the desirable elements of rule-based and automatic techniques into one system. One huge benefit of these systems is that results are often more accurate.
(monkeylearn, Sentiment Analysis Algorithms , 2020)

5. Sentiment Analysis Approach

Sentiment **C**lassification techniques can be roughly divided into machine learning approach, lexicon based approach and hybrid approach . The **M**achine **L**earning Approach (**ML**) applies the famous ML algorithm and uses linguistic features.

The **L**exicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The hybrid Approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods. The various approaches and the most popular algorithms of **SC** are illustrated as mentioned.

The text classification methods using **ML** approach can be roughly divided into supervised and unsupervised learning methods. The supervised methods make use of a large number of labeled training documents. The unsupervised methods are used when it is difficult to find these labeled training documents.

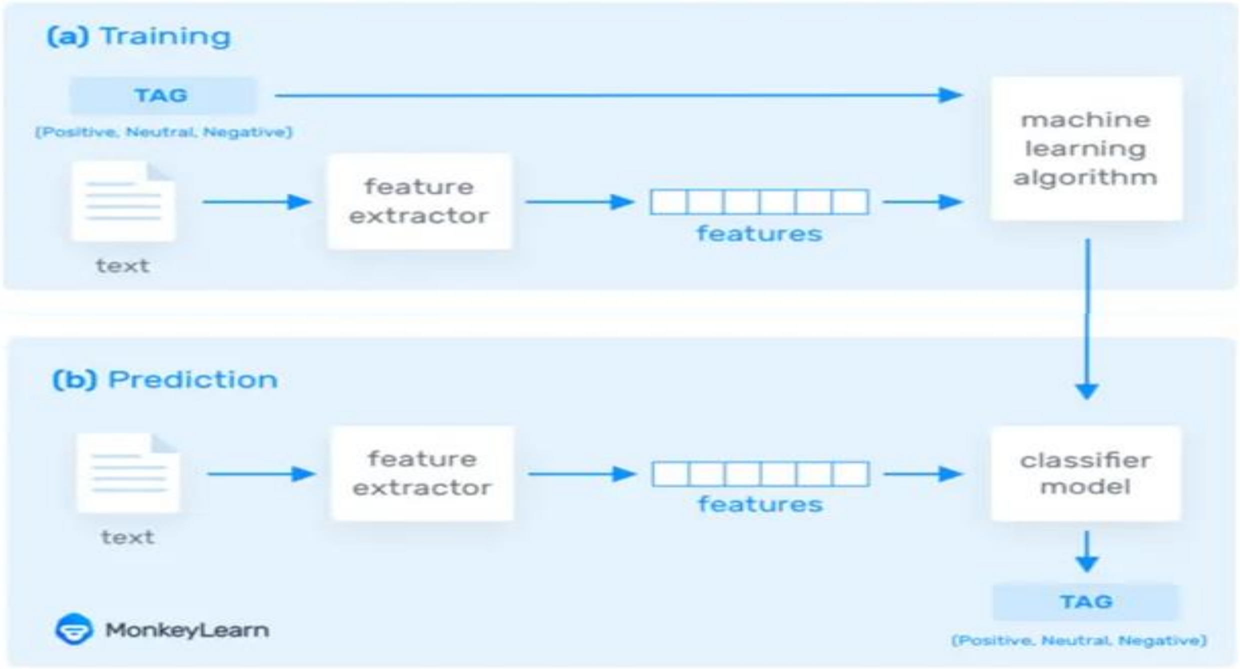


Figure 3. Sentiment Analysis Work

The lexicon-based approach depends on finding the opinion lexicon which is used to analyze the text. There are two methods in this approach. The dictionary-based approach which depends on finding opinion seed words, and then searches the dictionary of their synonyms and antonyms. The corpus based approach begins with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations. This could be done by using statistical or semantic methods.

CHAPTRE 3 : APPLICATION

I. CORPUS

1. Introduction

The word "corpus", derived from the Latin word meaning "body", may be used to refer to any text in written or spoken form. However, in modern Linguistics this term is used to refer to large collections of texts which represent a sample of a particular variety or use of language(s) that are presented in machine readable form . (Azeredo, Corpus, 2008)

2. Types of corpora

There are many different kinds of corpora. They can contain written or spoken (transcribed)language, modern or old texts, texts from one language or several languages. The texts can be whole books, newspapers, journals, speeches etc, or consist of extracts of varying length. The kind of texts included and the combination of different texts vary between different corpora and corpus types. (Azeredo, Types of corpora, 2008)

3. Types Of Text Corpora

A text corpus can be classified into various categories by the source of the content, metadata, the presence of multimedia or its relation to other corpora. The same corpus can fall into more than one category if it fulfills the criteria for more categories. (sketchengine, Types of text corpora)

3.1. Monolingual Corpus

Monolingual corpus is the most frequent type of corpus. It contains texts in one language only. The corpus is usually tagged for parts of speech and is used by a wide range of users for various tasks from highly practical ones, e.g. checking the correct usage of a word or looking up the most natural word combinations, to scientific use, e.g. identifying frequent patterns or new trends in language. Sketch Engine contains hundreds of monolingual corpora in dozens of languages. (sketchengine, Monolingual corpus)

3.2. Parallel Corpus

A parallel corpus consists of two monolingual corpora. One corpus is the translation of the other. For example, a novel and its translation or a translation memory of a CAT tool (A CAT tool is a computer assisted translation tool, a software that helps translators maintain consistency in terminology across their translation jobs and also aids the translation process by suggesting (or translating automatically) passages which the translator already translated in the past.)could be used to build a parallel corpus. Both languages need to be aligned, i.e. corresponding segments, usually sentences or paragraphs, need to be matched. The user can then search for all examples of a word or phrase in one language and the results will be displayed together with the corresponding sentences in the other language. The user can then observe how the search word or phrase is translated. (sketchengine, Parallel corpus)

3.3. Multilingual Corpus

A multilingual corpus is very similar to a parallel corpus. The two terms are often used interchangeably. A multilingual corpus contains texts in several languages which are all translations of the same text and are aligned in the same way as parallel corpora. When only two languages are selected, a multilingual corpus behaves as a parallel corpus. The user can also decide to work with one language to use it as a monolingual corpus. (sketchengine, Multilingual corpus, 2020)

3.4. Comparable Corpus

A comparable corpus is a set of two or more monolingual corpora whose texts relate to the same topic., however, they are not translations of each other, and therefore, there are not aligned. When users search these corpora they can use the fact, that the corpora also have the same metadata. (sketchengine, Comparable corpus, 2020)

3.5. Learner Corpus

A learner corpus is a corpus of texts produced by learners of a language. The corpus is used to study the mistakes and problems learners have when learning a foreign language. (sketchengine, Learner corpus, 2020)

3.6. Diachronic Corpus

A diachronic corpus is a corpus containing texts from different periods and is used to study the development or change in language.
In addition, there is a specialized diachronic feature called Trends, which identifies words whose usage changes the most of the selected period of time. (sketchengine, Diachronic corpus, 2020)

3.7. Specialized Corpus

A specialized corpus contains texts limited to one or more subject areas, domains, topics ,etc.
Such corpus is used to study how the specialized language is used. (sketchengine, Specialized, 2020)

3.8. Multimedia Corpus

A multimedia corpus contains texts which are enhanced with audio or visual materials or other type of multimedia content. (sketchengine, Multimedia, 2020)

II. The Tools Used

1. Python

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python’s elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms.

The Python interpreter and the extensive standard library are freely available in source or binary form for all major platforms from the Python Web site, https://www.python.org/, and may be freely distributed. The same site also contains distributions of and pointers to many free third party Python modules, programs and tools, and additional documentation.

The Python interpreter is easily extended with new functions and data types implemented in C or C++ (or other languages callable from C). Python is also suitable as an extension language for customizable applications.

This tutorial introduces the reader informally to the basic concepts and features of the Python language and system. It helps to have a Python interpreter handy for hands-on experience, but all examples are self-contained, so the tutorial can be read off-line as well.

For a description of standard objects and modules, see The Python Standard Library. The Python Language Reference gives a more formal definition of the language. To write extensions in C or C++, read Extending and Embedding the Python Interpreter and Python/C API Reference Manual. There are also several books covering Python in depth.

This tutorial does not attempt to be comprehensive and cover every single feature, or even every commonly used feature. Instead, it introduces many of Python's most noteworthy features, and will give you a good idea of the language's flavor and style. After reading it, you will be able to read and write Python modules and programs, and you will be ready to learn more about the various Python library modules described in The Python Standard Library. (docs.python, python, 2020)

2. Anaconda

Anaconda is a free and open-source distribution of the Python and programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and mac OS. It is developed and maintained by Anaconda, Inc., which was founded by Peter Wang and Travis Oliphant in 2012. As an Anaconda, Inc. product, it is also known as Anaconda Distribution or Anaconda Individual Edition, while other products from the company are Anaconda Team Edition and Anaconda Enterprise Edition, which are both not free.

Package versions in Anaconda are managed by the package management system*conda*. This package manager was spun out as a separate open-source package as it ended up being useful on its own and for other things than Python. There is also a small, bootstrap version of Anaconda called Miniconda, which includes only conda, Python, the packages they depend on, and a small number of other packages. (wikipedia, Anaconda, 2020)

3. Twitter API

Twitter's API allows us to access certain points of a public profile. As a basic use of the API, we could write a program where you can search for someone's username and it'll return the profile page. Instead of walking up to the Twitter office every time you have a request, the API gives access to the program to return the profile page. (Chen, 2019)

4. Libraries:

- **Comma Separated Values(CSC)**: The CSV module implements classes to read and write tabular data in CSVformat. It allows programmers to say, "write this data in the format preferred by Excel," or "read data from this file which was generated by Excel," without knowing the precise details of the CSV format used by Excel. Programmers can also describe the CSV formats understood by other applications or define their own special-purpose CSV formats. (docs.python, docs.python, 2020)

- **Natural Language Toolkit NLTK:** is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum. (nltk, 2020)

- **Tweepy** is an open source Python package that gives we a very convenient way to access the Twitter API with Python. Tweepy includes a set of classes and methods that represent Twitter's models and API endpoints, and it transparently handles various implementation details, such as (Garcia, 2018):

    - Data encoding and decoding
    - HTTP requests
    - Results pagination
    - OAuthauthentication
    - Rate limits
    - Streams

- **pandas** is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.(wikipedia, Pandas, 2020)

- **Regular expression** (re) **:**A RegEx, or Regular Expression, is a sequence of characters that forms a search pattern. RegEx can be used to check if a string contains the specified search pattern. (3schools)

- **Langdetect**: Langdetect is a library that detects the language of specific text, it supports more than 55 languages.

- **spaCy**is a free and open-source library for **Natural Language Processing** (NLP) in Python with a lot of in-built capabilities. It’s becoming increasingly popular for processing and analyzing data in NLP. Unstructured textual data is produced at a large scale, and it’s important to process and derive insights from unstructured data. To do that, you need to represent the data in a format that can be understood by computers. NLP can help you do that.(Singh, 2019)

- **Demoji**: Demoji is a library that accurately detects and removes emojis in text.

- **Http.client:** This module defines classes which implement the client side of the HTTP and HTTPS protocols. (docs.python, http.client , 2020)

- **Farasa**: Farasa consists of the segmentation/tokenization module, POS tagger, Arabic text Diacritizer, and Dependency Parser. We measure the performance of the segmenter in terms of accuracy and efficiency, in two NLP tasks, namely Machine Translation (MT) and Information Retrieval (IR). Farasa outperforms or equalizes state-of-the-art Arabic segmenters(Stanford and MADAMIRA), while being more than one order of magnitude faster. (Ahmed Abdelali, 2016)

- **Scikit-learn** :Scikit-learn is a free Python Machine Learning Library. It is a very useful tool for data mining and data processing and can be used for both personal and commercial purposes. Scikit-learn is potentially the most useful library for machine learning in Python. The sklearn library provides a range of powerful methods for machine learning and statistical analysis, including classification, regression, clustering and dimensional reduction.
  Python Scikit-learn allows users to perform various Machine Learning tasks and provides a means to incorporate Machine Learning in Python. It needs to work with Python ‘s scientific and numerical libraries, namely Python SciPy and Python NumPy, respectively. It’s basically a SciPy toolkit that features a variety of Machine Learning algorithms. (prashanth, 2020)

- ***TextBlob*: *TextBlob* is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. (textblob)

III. Collecting tweets

Step 1: Getting Twitter API keys

We create a twitter account to get tweeter API keys

Step 2: Connecting to Twitter Streaming API and downloading data

We will be using a Python library called Tweepy to connect to Twitter Streaming API and downloading the data.

We choose to collect tweets in the medical field.
the collection is made based on the keywords and hachtagexp (Algeria health,

corona virus, # covid-19, #stayhome, ministry of health,#كورونا , medicament, santé ,médecins , الجيش الابيض , أطباء الجزائر, عدد الاصابات , #الحجر_الصحي ,….)

We collected 2075 tweets

number of tweetscollected	2075
number of tweets after deletion of duplicates	1632
number of words in the corpus	41544
word / tweet	25

Table 1. Collected Tweet

IV. Pre-processing

1. Automatic detection of the language of tweets:

Before starting our data cleaning procedure we collected 2075 tweets they are sorted according to their language using the "detect" library, we made up four categories: Arabic, French, English, mixed We obtained the following results: 796 Arabic tweets, 505 French tweets, 116 English tweets, 215 mixed tweets .

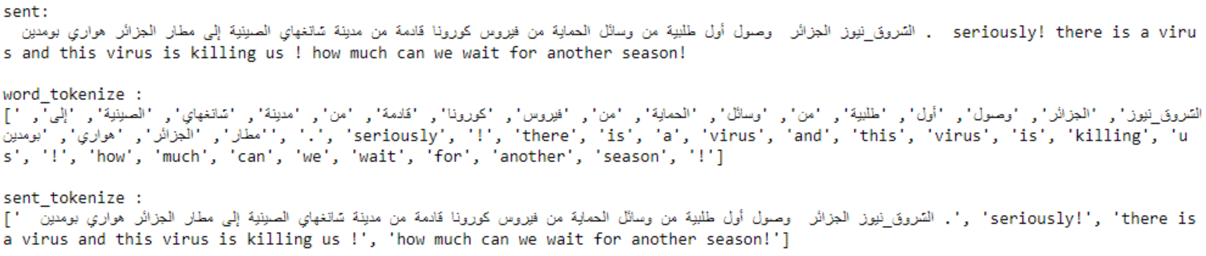
language	number of tweets	Percentage
Arabic	796	48.77%
French	505	30.94%
English	116	7.10%
Mix	215	13.17%

Table 2. Detection Result

2. Tokenization

Tokenization is a common task in Natural Language Processing (NLP). It’s a fundamental step in both traditional NLP methods like Count Vectorizer and Advanced Deep Learning-based architectures like Transformers. Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or sub words. Hence, tokenization can be broadly classified into 3 types – word, character, and sub word (n-gram characters) tokenization. (PAI, 2020)

Exp :



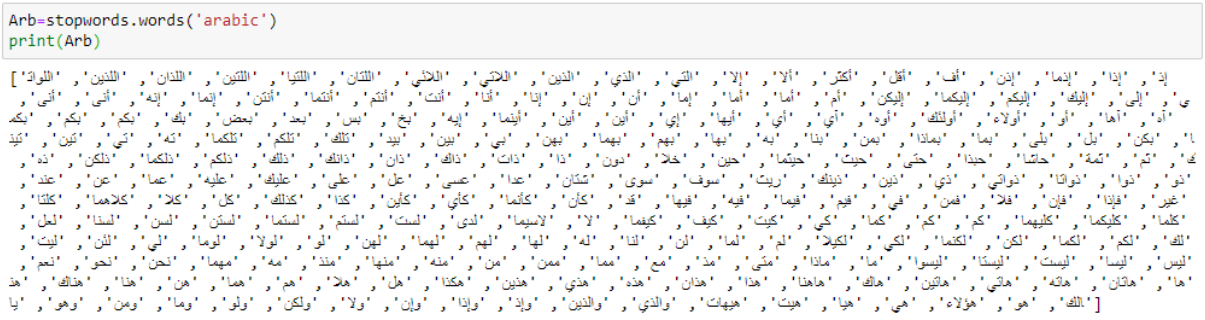
```
sent:
 الشروق_نيوز الجزائر  وصول أول طلبية من وسائل الحماية من فيروس كورونا قادمة من مدينة شانغهاي الصينية إلى مطار الجزائر هواري بومدين  .  seriously! there is a viru
s and this virus is killing us ! how much can we wait for another season!

word_tokenize :
['الشروق_نيوز', 'الجزائر', 'وصول', 'أول', 'طلبية', 'من', 'وسائل', 'الحماية', 'من', 'فيروس', 'كورونا', 'قادمة', 'من', 'مدينة', 'شانغهاي', 'الصينية', 'إلى', '
مطار', 'الجزائر', 'هواري', 'بومدين', '.', 'seriously', '!', 'there', 'is', 'a', 'virus', 'and', 'this', 'virus', 'is', 'killing', 'u
s', '!', 'how', 'much', 'can', 'we', 'wait', 'for', 'another', 'season', '!']

sent_tokenize :
['  الشروق_نيوز الجزائر  وصول أول طلبية من وسائل الحماية من فيروس كورونا قادمة من مدينة شانغهاي الصينية إلى مطار الجزائر هواري بومدين .', 'seriously!', 'there is
a virus and this virus is killing us !', 'how much can we wait for another season!']
```

Figure 4. Tokenization Example

3. Stop Words

Stopwords are the words which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. For example, the words like (the, he, have , le , la , les , من ,و ,على,ال ,etc.)

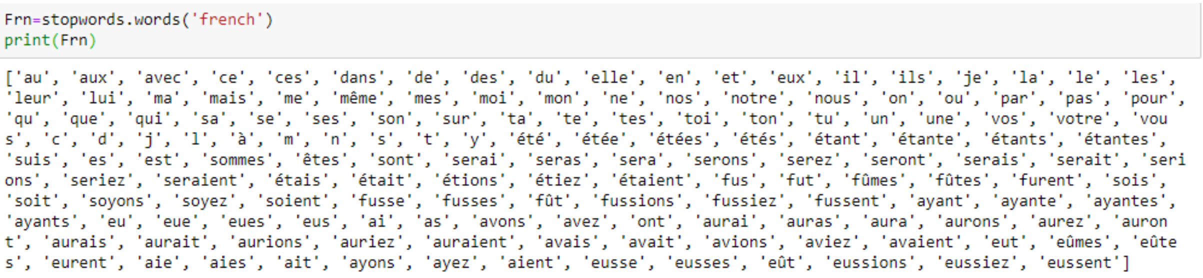
Such words are already captured :



```
Arb=stopwords.words('arabic')
print(Arb)
```

```
['إذ', 'إذا', 'إذما', 'إذن', 'أف', 'أقل', 'أكثر', 'ألا', 'إلا', 'التي', 'الذي', 'الذين', 'اللاتي', 'اللائي', 'اللتان', 'اللتيا', 'اللتين', 'اللذان', 'اللذين', 'اللوات
ي', 'إلى', 'إليك', 'إليكم', 'إليكما', 'إليكن', 'أم', 'أما', 'أما', 'إما', 'أن', 'إن', 'إنا', 'أنا', 'أنت', 'أنتم', 'أنتما', 'أنتن', 'إنما', 'إنه', 'أنى', 'أنى', 
'آه', 'آها', 'أو', 'أولاء', 'أولئك', 'أوه', 'آي', 'أي', 'أيها', 'إي', 'أين', 'أين', 'أينما', 'إيه', 'بخ', 'بس', 'بعد', 'بعض', 'بك', 'بكم', 'بكم', 'بكم
ا', 'بكن', 'بل', 'بلى', 'بما', 'بماذا', 'بمن', 'بنا', 'به', 'بها', 'بهم', 'بهما', 'بهن', 'بي', 'بين', 'بيد', 'تلك', 'تلكم', 'تلكما', 'ته', 'تي', 'تين', 'تين
ك', 'ثم', 'ثمة', 'حاشا', 'حبذا', 'حتى', 'حيث', 'حيثما', 'حين', 'خلا', 'دون', 'ذا', 'ذات', 'ذاك', 'ذان', 'ذانك', 'ذلك', 'ذلكم', 'ذلكما', 'ذلكن', 'ذه', 
'ذو', 'ذوا', 'ذواتا', 'ذواتي', 'ذي', 'ذين', 'ذينك', 'ريث', 'سوف', 'سوى', 'شتان', 'عدا', 'عسى', 'عل', 'على', 'عليك', 'عليه', 'عما', 'عن', 'عند', 
'غير', 'فإذا', 'فإن', 'فلا', 'فمن', 'في', 'فيم', 'فيما', 'فيه', 'فيها', 'قد', 'كأن', 'كأنما', 'كأي', 'كأين', 'كذا', 'كذلك', 'كل', 'كلا', 'كلاهما', 'كلتا', 
'كلما', 'كليكما', 'كليهما', 'كم', 'كم', 'كما', 'كي', 'كيت', 'كيف', 'كيفما', 'لا', 'لاسيما', 'لدى', 'لست', 'لستم', 'لستما', 'لستن', 'لسن', 'لسنا', 'لعل', 
'لك', 'لكم', 'لكما', 'لكن', 'لكنما', 'لكي', 'لكيلا', 'لم', 'لما', 'لن', 'لنا', 'له', 'لها', 'لهم', 'لهما', 'لهن', 'لو', 'لولا', 'لوما', 'لي', 'لئن', 'ليت', 
'ليس', 'ليسا', 'ليست', 'ليستا', 'ليسوا', 'ما', 'ماذا', 'متى', 'مذ', 'مع', 'مما', 'ممن', 'من', 'منه', 'منها', 'منذ', 'مه', 'مهما', 'نحن', 'نحو', 'نعم', 
'ها', 'هاتان', 'هاته', 'هاتي', 'هاتين', 'هاك', 'هاهنا', 'هذا', 'هذان', 'هذه', 'هذي', 'هذين', 'هكذا', 'هل', 'هلا', 'هم', 'هما', 'هن', 'هنا', 'هناك', 'هن
الك', 'هو', 'هؤلاء', 'هي', 'هيا', 'هيت', 'هيهات', 'والذي', 'والذين', 'وإذ', 'وإذا', 'وإن', 'ولا', 'ولكن', 'ولو', 'وما', 'ومن', 'وهو', 'يا']
```

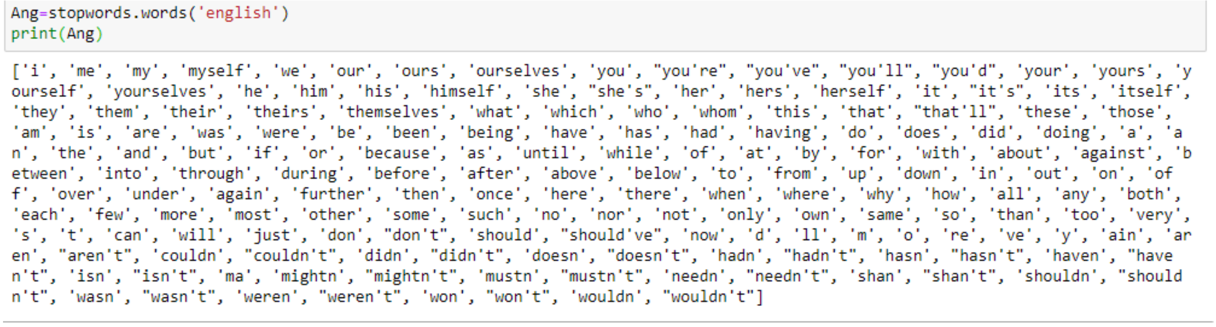
Figure 5 Arabic Stopwords



```
Frn=stopwords.words('french')
print(Frn)
```

```
['au', 'aux', 'avec', 'ce', 'ces', 'dans', 'de', 'des', 'du', 'elle', 'en', 'et', 'eux', 'il', 'ils', 'je', 'la', 'le', 'les',
'leur', 'lui', 'ma', 'mais', 'me', 'même', 'mes', 'moi', 'mon', 'ne', 'nos', 'notre', 'nous', 'on', 'ou', 'par', 'pas', 'pour',
'qu', 'que', 'qui', 'sa', 'se', 'ses', 'son', 'sur', 'ta', 'te', 'tes', 'toi', 'ton', 'tu', 'un', 'une', 'vos', 'votre', 'vou
s', 'c', 'd', 'j', 'l', 'à', 'm', 'n', 's', 't', 'y', 'été', 'étée', 'étées', 'étés', 'étant', 'étante', 'étants', 'étantes',
'suis', 'es', 'est', 'sommes', 'êtes', 'sont', 'serai', 'seras', 'sera', 'serons', 'serez', 'seront', 'serais', 'serait', 'seri
ons', 'seriez', 'seraient', 'étais', 'était', 'étions', 'étiez', 'étaient', 'fus', 'fut', 'fûmes', 'fûtes', 'furent', 'sois',
'soit', 'soyons', 'soyez', 'soient', 'fusse', 'fusses', 'fût', 'fussions', 'fussiez', 'fussent', 'ayant', 'ayante', 'ayantes',
'ayants', 'eu', 'eue', 'eues', 'eus', 'ai', 'as', 'avons', 'avez', 'ont', 'aurai', 'auras', 'aura', 'aurons', 'aurez', 'auron
t', 'aurais', 'aurait', 'aurions', 'auriez', 'auraient', 'avais', 'avait', 'avions', 'aviez', 'avaient', 'eut', 'eûmes', 'eûte
s', 'eurent', 'aie', 'aies', 'ait', 'ayons', 'ayez', 'aient', 'eusse', 'eusses', 'eût', 'eussions', 'eussiez', 'eussent']
```

Figure 6. French Stopwords



```
Ang=stopwords.words('english')
print(Ang)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'y
ourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those',
'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'a
n', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'b
etween', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'of
f', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both',
'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'ar
en', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "have
n't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "should
n't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

Figure 7. English Stopwords

So , we created our **own** dictionary which contains the stopwords of the three languages (Arabic , French , English) to eliminate them from our corpus .

4. Lemmatization

Lemmatisation (or lemmatization) in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lrmma, or dictionary form.

In computational linguistics, lemmatisation is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatisation depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document. As a result, developing efficient lemmatisation algorithms is an open area of research . (wikipedia, Lemmatisation, 2020)

4.1. Lemmatization of English tweets

We used in the analysis of English texts the "WordNetLemmatizer" library for lemmatization and the "pos_tag" process of the NLTK package for the classification of words according to their placement in the sentence and their labeling.

4.2. Lemmatization of French tweets :

We used in the analysis of French texts the library "fr_core_news_md" for the lemmatization of the Spacy package. We note that the lemmatization and classification of words according to their location in the sentence is carried out by the library itself.

4.3. Lemmatization of Arabic tweets

Farasa (which means “insight” in Arabic), is a fast and accurate text processing toolkit for Arabic text. Farasa can do segmentation, lemmatization, POS tagging, Arabic diacritization, dependency parsing, constituency parsing, named-entity recognition, and spell-checking

V. Annotation of the corpus

We manually annotated each tweet in our corpus using one of the three annotations: positive, negative or neutral.

We have affected the value '1' for the positive tweets, the value '-1' for the negative tweets and the value '0' for the neutral tweets .

We got the following results :

positive tweets	Negative tweets	Neutral tweets
24,10%	34.47%	41.41%

Table 3. Annotation

Before starting the binary classification of tweets, we have removed neutral tweets.

VI. Classification of tweets

We used two classifiers widely used in the literature: SVM (Support Vector Machine) and Naive Bayes.

For the evaluation, we used recall, precision and f-score to measure the performance of the classifiers used.

We obtained the following results :

Classifier	Precision	Recall	f-score
SVM	70%	68%	69%
Naive Bayes	74%	69%	71%

Table 4. Classification result

The results obtained show that classification using Naive Bays gives better results compared to SVM.

Conclusion

During this chapter we have presented the phase of collecting and annotating tweets from our corpus. We also described the preprocessing doing on our annotated corpus then we gave the results in terms of precision and recall for both classifiers chosen to detect feelings in tweets. We have noticed that Naive Bayes gives the best performance. This work allowed us to openseveral research perspectives in particular are testing other types of classifiers. It is also important to underline that the annotation of our corpus was done by one person and it is better to do it by two people (and use a third person if the two annotations are not the same) to avoid possible errors in the annotation phase and thus obtain the best classifiers.

General conclusion

In this study , we were intersecting on the automatic and intelligent sentiment analysis and opinion mining on social media and the web in the medical field.

In the first time we do some research about social media and social network , also , about tweeter .

Then , we spoke about the state of art in the methods used in sentiment analysis .

In the end , we details the steps following in our study. We start by the collect of tweets , then , the pre-processing which contain (Automatic detection of the language of tweets , Tokenization ,the delete of stop word , the lemmatization), after , the annotation of the corpus , and finally , the classification where we have noticed that Naive Bayes gives the best performance.

References

1. **COHN, MICHAEL.** social media. *compukol.* [Online] 10 10, 2011. https://www.compukol.com/social-media-vs-social-networking/?fbclid=IwAR3Kkt5p47cgU7mm1x1yCbgpuoP2ygtT6MIZE2HshiOg3-l5GGzWx0TwgT8.

2. **lyfemarketing.** Developing a Social Media Strategy. *lyfemarketing.* [Online] 07 23, 2020. https://www.lyfemarketing.com/blog/what-is-social-media-management/.

3. **wikipedia.** Mobile social media. *wikipedia.* [Online] 08 11, 2020. https://en.wikipedia.org/wiki/Social_media#Mobile_social_media.

4. **Hope, Computer.** Social network. *computerhope.* [Online] 02 08, 2020. https://www.computerhope.com/jargon/s/socinetw.htm?fbclid=IwAR3czrQyvl8mN082yhfDgWAW_Eg4YoOvIPkPUcS3ip-z5bbBu8eUR9d6aqc.

5. **COHN, MICHAEL.** Social media. *compukol.* [Online] 10 10, 2011. https://www.compukol.com/social-media-vs-social-networking/?fbclid=IwAR3Kkt5p47cgU7mm1x1yCbgpuoP2ygtT6MIZE2HshiOg3-l5GGzWx0TwgT8.

6. —. social media and social networking have in common. *compukol.* [Online] 10 10, 2011.

7. **computerhope.** Examples of social networks. *computerhope.* [Online] 02 08, 2020. https://www.computerhope.com/jargon/s/socinetw.htm?fbclid=IwAR30Ay5rByQ6Vug1R7FeI_NIexovX_vgLpIy2JYml2iMAhnXvCstihNcXio.

8. —. the first social media website. *computerhope.* [Online] 02 08, 2020. https://www.computerhope.com/jargon/s/socinetw.htm?fbclid=IwAR239GIXhyHqMRUIaTFQ9VTuFvWTye9cT83JmmyXuTpWXvOHsHeUp7bNo9I.

9. **Skrba, Anya.** Blog. *firstsiteguide.* [Online] 04 08, 2018. https://firstsiteguide.com/what-is-blog/?fbclid=IwAR2tJ21ayb24ymQXVmZNzUt2xgTcky_t6UNgOOx_QWmeH5FIEzQIRL1pYik.

10. **Techopedia.** Microblog . *Techopedia.* [Online] 12 05, 2016. 11. **Wikipedia.** Twitter. *Wikipedia.* [Online] 08 04, 2020. https://en.wikipedia.org/wiki/Twitter?fbclid=IwAR1lsbOcUq63kJBb5YoFu4e_bxz_29GVr9ux-jCRhIwF6dw_VKc-Qesua5U.

12. **Sengottaiyan, Jeyakumar.** Twitter PPT. *slideshare.* [Online] 09 15, 2009. https://www.slideshare.net/rsjeyakumar/twitter-ppt?next_slideshow=3.

13. **Dharmendra.** Twitter is so popular. *netkiduniya.* [Online] 08 10, 2020. https://netkiduniya.com/2019/07/what-is-twitter/?fbclid=IwAR0D3H2Y2wbY5fS9dEAhPXJUdJa3NCjLb0SyqICFO-fj4S9TCbRqG2HJyX4.

14. **Öztürk, Nazan.** Sentiment analysis. *sciencedirect.* [Online] 04 01, 2018. https://www.sciencedirect.com/science/article/abs/pii/S0736585317304999.

15. **H. Sankar and V. Subramaniyaswamy, V. Vijayakumar.** Intelligent sentiment analysis approach using edge computing-based deep learning technique. *Wiley Online Library.* [Online] 03 01, 2019. https://onlinelibrary.wiley.com/doi/pdf/10.1002/spe.2687.

16. **Al-Masri, Anas.** Creating The Twitter Sentiment Analysis Program in Python with Naive Bayes Classification. *towardsdatascience.* [Online] 02 16, 2019. https://towardsdatascience.com/creating-the-twitter-sentiment-analysis-program-in-python-with-naive-bayes-classification-672e5589a7ed.

17. **monkeylearn.** Sentiment Analysis. *monkeylearn.* [Online] 01 02, 2020. https://monkeylearn.com/sentiment-analysis/.

18. —. Types of Sentiment Analysis. *monkeylearn.* [Online] 01 02, 2020. https://monkeylearn.com/sentiment-analysis/.

19. —. Fine-grained Sentiment Analysis. *monkeylearn.* [Online] 01 02, 2020https://monkeylearn.com/sentiment-analysis/.

20. —. Emotion detection. *monkeylearn.* [Online] 01 02, 2020. https://monkeylearn.com/sentiment-analysis/.

21. —. Sentiment Analysis Algorithms . *monkeylearn.* [Online] 01 02, 2020. https://monkeylearn.com/sentiment-analysis/.

22. **Azeredo, Rogerio.** Corpus. *scribd.* [Online] 10 2008. https://fr.scribd.com/presentation/20085715/Corpus-Linguistics.

23. —. Types of corpora. *Scribd.* [Online] 10 2008. https://fr.scribd.com/presentation/20085715/Corpus-Linguistics.

24. **sketchengine.** Types of text corpora. *sketchengine.* [Online] https://www.sketchengine.eu/corpora-and-languages/corpus-types/.

25. —. Monolingual corpus. *sketchengine.* [Online] https://www.sketchengine.eu/corpora-and-languages/corpus-types/.

26. —. Parallel corpus. *sketchengine.* [Online] https://www.sketchengine.eu/corpora-and-languages/corpus-types/.

27. —. Multilingual corpus. *sketchengine.* [Online] 03 2020. https://www.sketchengine.eu/corpora-and-languages/corpus-types/.

28. —. Comparable corpus. *sketchengine.* [Online] 03 2020. https://www.sketchengine.eu/corpora-and-languages/corpus-types/.

29. —. Learner corpus. *sketchengine.* [Online] 03 2020. https://www.sketchengine.eu/corpora-and-languages/corpus-types/.

30. —. Diachronic corpus. *sketchengine.* [Online] 03 2020. https://www.sketchengine.eu/corpora-and-languages/corpus-types/.

31. —. Specialized. *sketchengine.* [Online] 03 2020. https://www.sketchengine.eu/corpora-and-languages/corpus-types/.

32. —. Multimedia. *sketchengine.* [Online] 03 2020. https://www.sketchengine.eu/corpora-and-languages/corpus-types/.

33. **docs.python.** python. *docs.python.org.* [Online] 08 18, 2020. https://docs.python.org/3.8/tutorial/index.html.

34. **wikipedia.** Anaconda. *wikipedia.* [Online] 07 18, 2020. https://en.wikipedia.org/wiki/Anaconda_(Python_distribution).

35. **Chen, Jenn.** Twitter’s API. *sproutsocial.* [Online] 08 09, 2019. https://sproutsocial.com/insights/what-is-an-api/.

36. **docs.python.***docs.python.* [Online] 08 18, 2020. https://docs.python.org/3/library/csv.html.

37. **nltk.** natural language toolkit. *nltk.* [Online] 04 13, 2020. https://www.nltk.org/#natural-language-toolkit.

38. **Garcia, Miguel.** tweepy. *realpython.* [Online] 08 18, 2018. https://realpython.com/twitter-bot-python-tweepy/.

39. **wikipedia.** Pandas. *wikipedia.* [Online] 08 06, 2020. https://en.wikipedia.org/wiki/Pandas_(software).

40. **3schools.** Python regex. *w3schools.* [Online] https://www.w3schools.com/python/python_regex.asp.

41. **Singh, Taranjeet.** spaCy . *Realpython.* [Online] 09 02, 2019. https://realpython.com/natural-language-processing-spacy-python/#what-are-nlp-and-spacy.

42. **docs.python.** http.client . *docs.python.* [Online] 08 18, 2020. https://docs.python.org/3/library/http.client.html.

43. **Ahmed Abdelali, Kareem Darwish, Nadir Durrani, Hamdy Mubara.** Farasa . *qatsdemo.cloudapp.* [Online] 2016. http://qatsdemo.cloudapp.net/farasa/.

44. **prashanth, Mahi.** Scikit learn library in Python. *medium.* [Online] 07 06, 2020. https://medium.com/@mahiprashanth866/scikit-learn-library-in-python-6c6f592f2b52.

45. **textblob.** textblob. *textblob.readthedocs.* [Online] https://textblob.readthedocs.io/en/dev/.

46. **Walaa Medhat and Ahmed Hassan, Hoda Korashy.** Figure 1. Sentiment analysis process on product reviews. *sciencedirect.* [Online] 12 04, 2014. https://www.sciencedirect.com/science/article/pii/S2090447914000550.

47. —. Sentiment classification techniques. *sciencedirect.* [Online] 12 04, 2014. https://www.sciencedirect.com/science/article/pii/S2090447914000550.

48. —. Sentiment classification techniques. *sciencedirect.* [Online] 12 04, 2014. https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0045.

49. —. Machine learning approach. *sciencedirect.* [Online] 12 04, 2014. https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0045.

50. —. 4.1.1. Supervised learning. *sciencedirect.* [Online] 12 04, 2014. https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0045.

51. —. Probabilistic classifiers. *sciencedirect.* [Online] 12 04, 2014. https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0045.

52. —. Naïve Bayes Classifier (NB). *sciencedirect.* [Online] 12 03, 2014. https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0045.

53. —. Bayesian Network (BN). *sciencedirect.* [Online] 12 04, 2014. https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0045.

54. —. Maximum Entropy Classifier (ME). *sciencedirect.* [Online] 12 04, 2014. https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0045.

55. —. Linear classifiers. *sciencedirect.* [Online] 2014, 04 2014, 12. https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0045.

56. —. Support Vector Machines Classifiers (SVM). *sciencedirect.* [Online] 12 04, 2014. https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0045.

57. —. Support Vector Machines Classifiers (SVM). *sciencedirect.* [Online] 12 04, 2014. https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0045.

58. —. Neural Network (NN). *sciencedirect.* [Online] 12 04, 2014. https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0045.

59. —. Decision tree classifiers. *sciencedirect.* [Online] 12 04, 2014. https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0045.

60. **Chen, Jenn.** an API. *sproutsocial.* [Online] 08 09, 2019. https://sproutsocial.com/insights/what-is-an-api/.

Abstract

Analyzing data on social media is a very interesting area of research. In this study, we were intersecting on the automatic and intelligent sentiment analysis and opinion mining on social media and the web in the medical field. In the first time we do some research about social media and social network, also, about tweeter . Then, we spoke about the state of art in the methods used in sentiment analysis. In the end , we details the steps following in our study. We start by the collect of tweets , then , the pre-processing which contain (Automatic detection of the language of tweets , Tokenization ,the delete of stop word , the lemmatization), after , the annotation of the corpus , and finally , the classification where we have noticed that Naive Bayes gives the best performance. The results obtained for this work are very promising and open future lines of research .

Key words: social networks, tweets, corpus, preprocessing , sentiment analysis , classification

Resumé

L’analyse des données sur les réseaux sociaux est un domaine de recherche en plein ébullition. Dans ce cadre , nous sommes intéressés a l’analyse automatique et intelligente des sentiments et fouille d'opinions sur les réseaux sociaux et le Web dans le domaine médical. Premièrement , on a présente des informations sur les réseaux sociaux et tweeter . Après , on a parlé sur l’états de l’art des méthodes utilise dans l’analyse des sentiment . Et enfin , on a **détaillé** les **différents étapes suivie** dans notre projet .**Alors,** on a commencé par la **collecte** des **tweets** , puis, le prétraitement de données (qui contient la **détection** automatique des données , la tokenisation , l’élimination des mots vides**,** la **lemmatisation**) , ensuit , l’annotation de notre corpus , et f**inalement , la classification dont** on a remarqué que le Naive Bayes donne les meilleurs performance . **Les résultats obtenus pour cette tâche sont très prometteurs et ouvrent des pistes de recherche** .

Mots clés : réseaux sociaux, tweets, corpus, prétraitement, l’analyse des sentiments , classification .