POLITECNICO DI TORINO

DEPARTMENT OF CONTROL AND COMPUTER ENGINEERING

Master of Science in Computer Engineering

Master Degree Thesis

Applying Natural Language Processing techniques to analyze HIV-related discussions on Social Media

Supervisors
Prof. Paolo Garza
Prof. Risto Sarvas
Post. Doc. Aqdas Malik

Candidate
Antonino Angi

Academic Year 2019-2020

To my family:
mum, dad and
Katia

Abstract

Nowadays social media are being used to monitor the progress of viruses and share important prevention and treatment information. This has also allowed the creation of a community of people united by the same disease, to give themselves strength, comfort and advice.
The objective of this work is to extract and understand discussions about HIV on a popular social media platform: Twitter, a micro-blogging application.
Tweets with the hashtag *#HIV* were collected in the date range of one year, starting from November 12th 2018 to November 12th 2019. They were then filtered and cleaned using NLP techniques, which allowed the removal of duplicates, non-english texts and useless information, such as tweets only containing urls, mentions or hashtags. After the cleaning phase, the main analyzes carried out were sentiment analysis and content analysis which, using data mining and text mining algorithms were able to reveal their emotions and the most influential topics written about HIV.
This study illustrates the potential of using social media to analyse the spread of viruses and health conditions using two types of analyses for the same topic and dataset: sentimental analysis and content analysis. HIV-related messages were used by organisations and credible sources to disseminate information about treatment and prevention, but also by individual users to share their thoughts, emotions and experiences of living with HIV. Twitter is also used by celebrities and health authorities to respond to public concerns.
This work shows that many tweets are written for the purpose of giving information and emotional support with assistance from the online

community and also for health care professionals who support individuals living with HIV/AIDS. The algorithms and notions covered in this work can subsequently be used by the public health community or data scientists to analyze tweets regarding other viruses or diseases, showing how social media can be used to identify, detect and study outbreaks in a specific geographical area and in a specific period of time.

Acknowledgements

I would like to express my deep gratitude to Post. Doc. Aqdas Malik, for his useful advices, his patience and support in the revision of this work.
I would also like to express my gratitude to the supervisors of this work, Prof. Paolo Garza and Prof. Risto Sarvas, for the useful comments and revisions.
I also want to thank my friends, with a particular mention to Marco, Federico, Sara and Beatrice.
This work is dedicated to my family, without whose support I would not have been able to face the challenges that the life posed me in the latest years.

Thank you so much.

Contents

Chapter 1

Introduction

Social media have become a fundamental tool of the society making deep changes and influencing its culture. The sharing of contents - such as thoughts, photos, videos, emotions, their geographic location, their *likes* - by users with different backgrounds, race and ethnicity makes social media one of the biggest data container ever, specially with their heterogeneity, their number and their production in real time: every minute of the 2018, 360.000 tweets are created on Twitter, 293.000 status are updated on Facebook and 65.000 new photos are added on Instagram, just in the US and UK [1]. These different ways of interaction have changed how new information is produced and distributed because it is not only used by individuals, but also by organizations and journals to spread their news.

In times of political elections, social media are increasingly used making a huge impact. For example, according to their survey, "44% of the U.S. adults got information about the 2016 presidential election from social media. And 24% got news and information from social media posts by Donald Trump and Hillary Clinton. Trump had almost 10 million Twitter followers to Clinton's seven million, and his nine million Facebook followers were about double her number." [2] These numbers also emphasize the great use of social media by supporters of the respective parties.

The widespread public engagement with social media creates a ready

platform for its application also in the health field. Indeed, at the end of 2008, 74% of U.S. adults went online, many of those were also searching for health information (80% in 2010). In fact, searching for health information, and e-mail checking, is very popular among adults. [3]

Application of social media features are many, some of them can be found in micro-blogging services such as Twitter or in Social Networking sites such as LinkedIn or Facebook, a virtual place in which it is possible to create bonds with other people, share emotions with them, united by the same social and cultural factors.

1.1 Twitter

Twitter is a free micro-blogging platform characterized by text messages (tweets) with a length of maximum 280 characters. Every user has a personal page (or profile) with people that he follows (friends) and people that follows him (followers), making it an asymmetric *friend relationship* than Facebook. The heterogeneous network of Twitter is shown in *Figure 1.1*.
A person that signs up to Twitter has his own profile with the description of his activities, relationships and preferences which could help identify himself to find new friends or new events nearby. Twitter offers two types of personal profiles:

- **public**: the profile is visible by everyone, even by not logged-in users.

- **private**: the profile is visible only by a logged-in user that has a *friend* relationship with it.

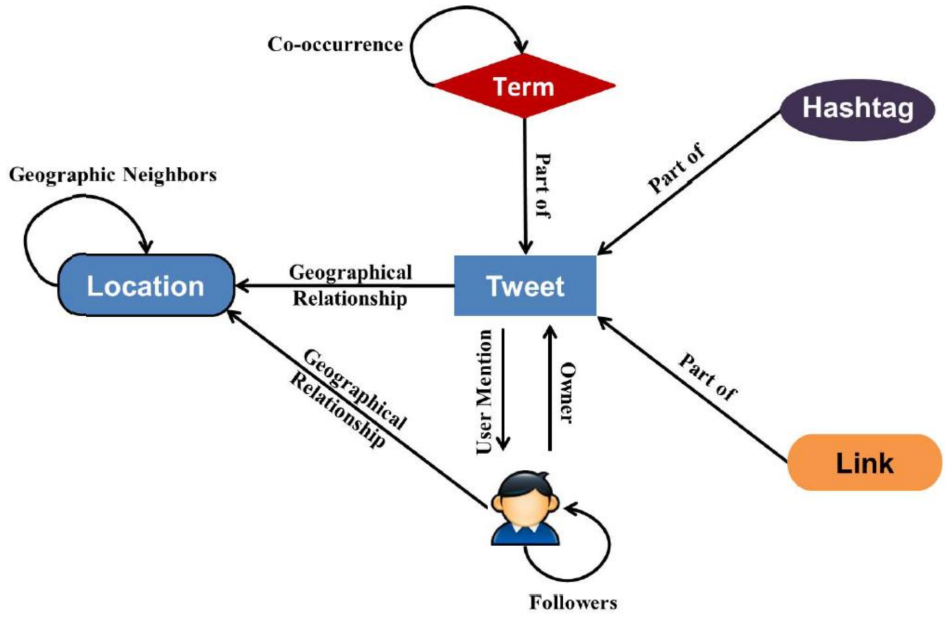

Figure 1.1: Twitter heterogeneous network [4]

1.2 HIV

"The human immunodeficiency virus (HIV) is a virus that attacks cells that help the body fight infection, making a person more vulnerable to other infections and diseases. An HIV-negative person gets the virus by coming into direct contact with certain body fluids from a person with HIV who has a detectable viral load. These fluids are:

- Blood
- Semen and pre-seminal fluid
- Rectal fluids
- Vaginal fluids
- Breast milk

People who are HIV-negative can prevent getting HIV by using PrEP (pre-exposure prophylaxis). Post-exposure prophylaxis (PEP) is a way to prevent HIV infection after a recent possible exposure to the virus. There are other ways to prevent getting or transmitting HIV through

injection drug use and sexual activity. If left untreated, HIV can lead to the disease AIDS (acquired immunodeficiency syndrome). AIDS is the late stage of HIV infection that occurs when the body's immune system is badly damaged because of the virus." [5]

1.3 Research Questions

The goal of this work is to visualize the predominant emotions of people and the most discussed topics, analyzing their tweets written from November 12th 2018 to November 12th 2019 on Twitter. This was possible through the use of social media mining algorithms for data collection and, subsequently, the use of text mining (a branch of Natural Language Processing) for the analysis of textual form tweets and the application of data mining algorithms for understanding the results.

Thus, the main research questions are:
RQ1: *Are Twitter individual users willing to share their personal experience of living with HIV?*
RQ2: *How are interpreted and classified the emotions among HIV-related tweets?*
RQ3: *What are the most discussed topics among twitter users who write about HIV?*

1.4 Thesis Structure

The thesis is structured in 6 chapters described as follows:

- The second chapter gives an overview of the current literature situation regarding the studies of diseases or viruses for healthcare purposes in social media platforms, giving also a quick look at pandemic cases.
- The third chapter focuses on the methodology and data science processes used in this work to perform the analysis presented in

the next chapters, including application of data filtering and data transformation.

- The fourth chapter presents all the obtained results from the analysis that have been carried out.
- The fifth chapter gives an opinion of the analysis performed and results obtained comparing it with the results of others author's work. It also answers to the research question presented in the first chapter.
- The sixth chapter gives a general comment of the work, presents some limitation of the algorithms used and shows future developments to improve the overall quality of the study.

Chapter 2

Literature Review

Social media platforms are being more and more used for healthcare purposes. Studies have been carried out to analyze the spread of diseases and viruses in real time and to detect and track infectious disease outbreaks. These platforms are also used to share personal knowledge on various arguments, such as personal health issues, treatments, but also side-effects.[6][7]
Social media studies and analyses have also been proven of being useful to create healthcare awareness, to ease social interaction between care provider-patient and patient-patient and to keep them motivated when fighting with their health problems and when adopting a healthy lifestyle.[8]

Many studies have been carried out to check and control the spread of diseases. For example, in the US, an algorithm has been used to detect flu infections through the analysis of tweets in a specific area with an accuracy of the 85% if compared to the last years' public data.[9]

2.1 Pandemic cases

During pandemic cases, social media were also used as a source of information. This is demonstrated by the case of H1N1, better known as swine flu, from 2009. A study carried out on over 5,000 tweets analysed the most popular topics during the pandemic showing that news and information were the most commonly tweeted H1N1-related material

and thoughts or jokes decreased over time. [10]

Social media has also been used during the last known pandemic case: the spread of SARS-CoV-2 or nCoV-2019 virus. The Chinese Government provided daily updates about surveillance and active cases on websites, but also on social media. Many experts, such as psychologists and psychiatrists were also called to provide support through social media by sharing strategies for dealing with stress that may be caused by social isolation or quarantine. [11]

2.2 HIV in social media

Social media platforms have also been studied in order to find other strategies for prevention and to stop the spread of HIV. In particular, many studies have in fact reported the importance of anonymity which helps to communicate more about HIV due to the stigma associated with it. Additionally, the increased social support provided by social media improved treatment adherence and eased the access to HIV testing and prevention services. HIV-positive people use social media in order to create social ties and build a sense of community, such as also to access health information and obtain emotional support.[12]
The ease with which health information, thoughts, doubts and experiences can be shared, the anonymity, the sense of collective support that transcends geographical barriers, the shame and fear of being stigmatized or being ashamed when talking about condoms make social media a very powerful tool when talking about HIV.
One study [13] described how teens use their phones to find medical information and share information about HIV and other sexually transmitted infections on social media. Anonymity on social media platforms helped to decrease stigma, but also fear and discrimination around HIV and allowed people to tell personal stories about their sexual orientation and HIV status in a way they would not with other people offline. [12]
A study has also reported the most common disadvantages when writing on social media to talk about HIV. Those cons were related to technology barriers, caused also by the cost of devices, the lack of physical

interaction or privacy. [12]

Chapter 3

Methodology and proposed data science process

This chapter presents the methods used to retrieve and analyze the data collected from the social media platform. Particularly, it focuses on giving a theoretical point of view for the social media mining, data mining and text mining algorithms and techniques that have been used. The third section of the chapter explains the methods used to preprocess and transform the dataset after the collection.

The first attempt to recover the tweets is using the official Twitter API, *Tweepy* using Python but it has some limitations: it just allows to collect tweets not older than a week. For such a short period there would have been few results that would not have allowed to obtain the desired results with the analyzes.

Python

Python is an interpreted high-level object-oriented and structured programming language that is implemented in C. Python does not convert its code into machine code but into something called byte code. Variables, functions and methods are not type-declared in source code and

the code is checked at run-time which allows to save compiling-time compared to other programming languages that need the compilers. It uses virtual environments which is used to keep dependencies required by different projects separate. Methods and variables are created in Stack memory, whose frame are then destroyed automatically whenever methods return. The heap memory stores instead the objects and the instance variables and when those variables return, a process called Garbage Collection will automatically frees those blocks of memory that are no longer used.

Tweepy

Tweepy is a library used in Python for accessing the Twitter API (Application Programming Interface). An API is a a set of HTTP requests that return a structure in JSON (JavaScript Object Notation) format, as described in *Section 3.1.1*. Tweepy, in particular, allows to stream tweets in real-time directly from the micro-blogging application. The first thing necessary to use Tweepy is to request an account developer and specify the reasons and purposes for which to use the API. After registering for a developer account and creating a new app, as shown in *Figure 3.1*, the new keys and access tokens are generated, those are necessary to write the Python code.

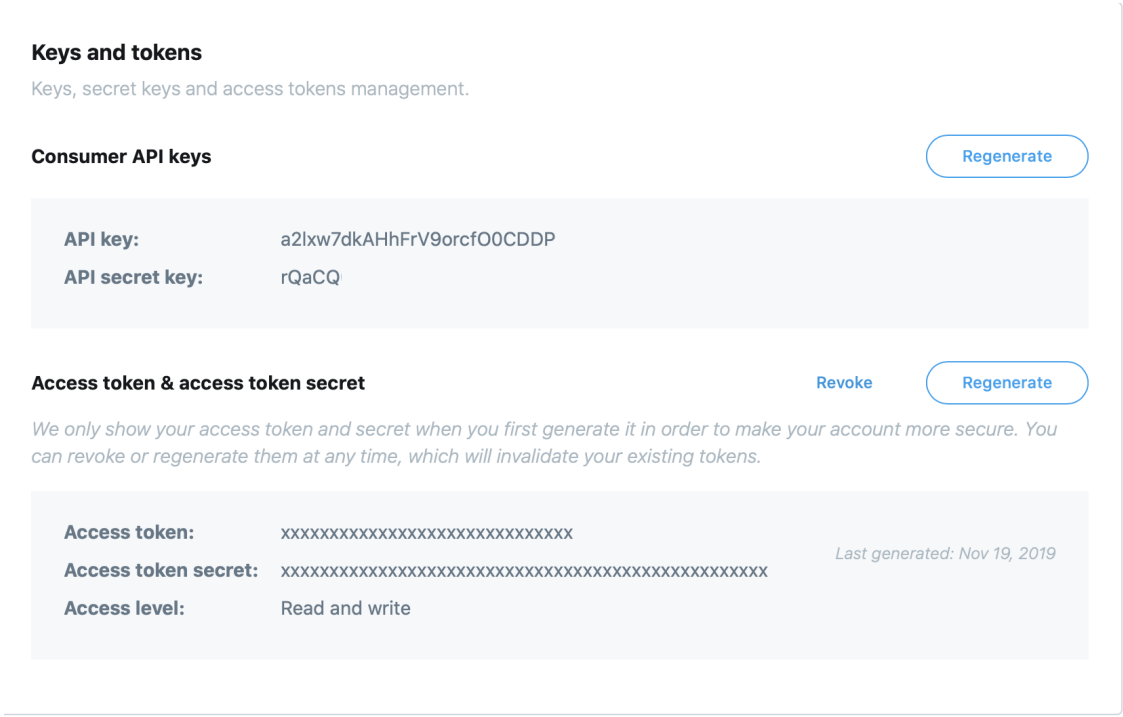

Figure 3.1: Example of keys and access tokens for Tweepy

```
1  import tweepy
2
3  authorization = tweepy.OAuthHandler(consumer_key_code,
   ↪  consumer_secret_code)
4  authorization.set_access_token(access_token_number,
   ↪  access_token_secret_number)
5
6  api = tweepy.API(authorization)
7
8  public_tweets_collection = api.home_timeline()
9  for tweet in public_tweets_collection:
10     print(tweet.text)
```

Algorithm 3.1: Snippet of Code for Tweepy

The example shown in *Algorithm 3.1* downloads the user's home timeline tweets and print each one of their texts to the console. Twitter

requires all requests to use OAuth for authentication.[14]
OAuth is a network protocol designed for authorization: a user authorizes a web or mobile app to access their data in online services, this is better than sharing credentials with third-party services.

3.1 Social Media Mining

Social Media Mining is the process of extracting information using data mining algorithms from social media. This information is then analysed to retrieve what is of interest for marketing and beyond, such as advertising matching and targeted support. Social Media Mining is also called "Social Media Analysis", this analysis focuses on the use of machine learning, data mining and statistics algorithms as well as sociology and media psychology.

When performing Social Media Mining it is important to take care of the each user's privacy because social media may contain sensitive data, such as medicals' or payments', that if analysed without paying attention, could bring to some privacy problems. It is therefore important to analyze the data and have results without being able to uniquely identify a particular user. Indeed, through the use of algorithms known as *Data matching* it is possible to cross data from different social media obtaining common information about a user or his behaviour.

3.1.1 Data collection

Data collection phase is the first step of analyzing social media. The main difficulties to face with are about the huge amount of data that have to be filtered and then preprocessed in order to avoid noise and outliers.
Data collection can occur in three different ways:

- **Manual**: the analyst manually selects data which are highly accurate and reliable. Duplicates and noise are generally avoided. The main disadvantage is that it may take much time especially if the date range is wide and there are many data to collect

- **Automatic**: the analyst uses API platforms distributed by the social media or some 3rd part software to extract information (like web crawlers). This is the fastest way to collect data, but the analyst should be aware of the software and hardware requirements and he should be prepared to delete duplicates and useless information in a filtering phase.
- **Mixed**: it uses both the Manual and Automatic method: the first one to collect more sensitive data and the second for the remaining part

Another important aspect when collecting data is the range of dates which could be:

- **Limited**: data are collected in a fixed and limited range of dates
- **Tracking**: data are initially collected with the limited method and then compared to another range of dates to make a benchmark report
- **On-going**: it is a real-time data collection in which data are collected and analysed withing strict deadlines (hourly or daily). This method is most used to analyze trends

For this work, HIV-related tweets were collected from Twitter using a code under MIT license [15] and the terminal as shown in *Figure 3.2*. The range of dates used for data collection was of one year, from November 12th 2018 to November 12th 2019. The Python code produced a *.csv* file with 160,658 tweets organised in *username, date, #retweets, #likes, text, mentions, hashtags*:

- **username**: the tweet author's name on Twitter
- **date**: when the tweet was written
- **#retweets**: how many *retweets* (re-posting) does the tweet have
- **#liked**: how many *likes* does the tweet have
- **text**: the text of the tweet
- **mentions**: who does the tweet *mention* - a mention is when someone uses the @ sign immediately followed by a username.

- **hashtags**: what *hashtags* does the tweet have - an hashtag is a type of tag used in social medias by users to dynamically find messages with a specific content.

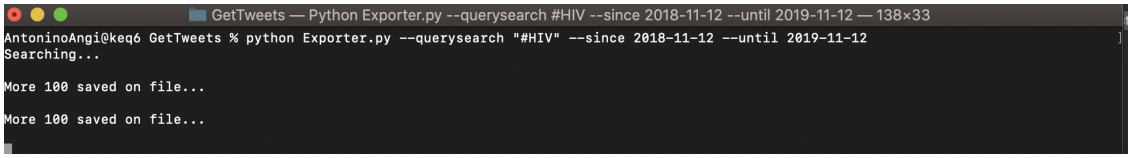

Figure 3.2: Terminal input to retrieve the tweets

When a user enters on Twitter, a page scroll loader starts and shows more tweets the more the user scrolls down.
All of this is done through calls to a JSON provider[15].

JSON

JavaScript Object Notation (JSON) is a format used to represent general data based on JavaScript object syntax. The main features of a JSON file are:

- It is composed by name/value pairs
- It is composed by an ordered list of values
- A JSON object begins with '{' and ends with '}'
- It is easily understandable by humans
- It is easily parsed by computers

A JSON provider is a software tool to convert JSON String to JSON objects.

Examples of code usage

Other possible uses of the Python code could be:

- Get tweets by username, as shown in *Algorithm 3.2*

```
1  tweetCrit_show =
   ↪  got.manager.tweetCrit().setUsername('mariored')
2  tweet =
   ↪  got.manager.TweetManager.getTweets(tweetCrit_show)[0]
3
4  print tweet.text
```

Algorithm 3.2: Example on how to get tweets by username [15]

- Get tweets by query search and bound dates, as shown in *Algorithm 3.3*

```
1  tweetCrit_show =
   ↪  got.manager.tweetCrit().setQuerySearch('america').
   ↪  setSince("2016-04-22").setUntil("2016-11-05")
2  tweet =
   ↪  got.manager.TweetManager.getTweets(tweetCrit_show)[0]
3
4  print tweet.text
```

Algorithm 3.3: Example on how to get tweets by query search and bound dates [15]

- Get the last 10 top tweets by username, as shown in *Algorithm 3.4*

```
1  tweetCrit_show =
   ↪  got.manager.tweetCrit().setUsername("mariored").
   ↪  setTopTweets(True).setMaxTweets(25)
2  tweet =
   ↪  got.manager.TweetManager.getTweets(tweetCrit_show)[0]
3
4  print tweet.text
```

Algorithm 3.4: Example on how to get the last 10 top tweets by username [15]

Author of tweet's information

Other user features, such as *location, #followers, #following, profile verified*, were also added to give a complete overview of the author of the tweet. This was possible thanks to a parsing tool library, called *BeautifulSoup*, which is a Python library that allows extracting data from HTML files. This data can then be converted to JSON for an easier manipulation.

- **location**: the geographical location the user added in his profile
- **#followers**: how many people the user follows
- **#followers**: from how many people the user is followed
- **#profile verified**: if a profile is verified or not - a verified profile lets other users know that the profile is authentic. Typically this includes accounts maintained by users in different artistics or public environment, such as fashion, music, art, government. [16]

At the end, the final *.csv* file is organised in *username, date, #retweets, #likes, text, mentions, hashtags, location, #followers, #following, profile verified*.

3.2 Data mining

Data mining is a process of automatic extraction unknown and potentially useful information from datasets using patterns. The flow of this process is known as *Knowledge Discovery from Data*, a representation is shown in *Figure 3.3* [17].

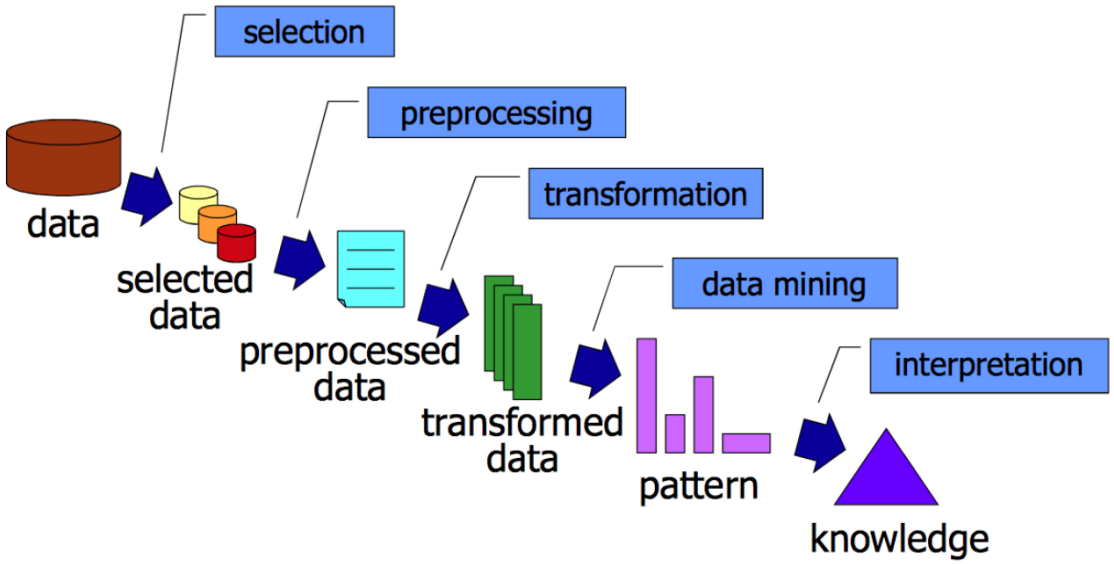

Figure 3.3: Knowledge Discovery from Data

It all starts from the **data selection**: starting from a big dataset, only a part of it will be used.

Data Preprocessing

The preprocessing phase is needed for data cleaning and data reduction to generate data ready to be transformed. It allows to generate data with a better quality by eliminating any noise, such as outliers, unnecessary to the analysis, removing any conflicts coming from the merging with different sources and consequentially save time for all the operation that will be working on the dataset, such as algorithms and results analysis.

Data can be characterized by being discrete or continuous or, in particular, by different attributes which could be:

- **Nominal**: those with a value that represents a characteristic, such as colors, ID and they have no ranking positions
- **Ordinal**: those with values that can be ranked, so they are ordered between them

- **Numeric**: they have a real integer value, such as temperature
- **Binary**: characterized by only two values, such as pass or fail

Preprocessing operations

The main goal of the preprocessing operations is to have high quality data in a reduced size. This can be achieved with operations such as **data cleaning** in which noise, outliers and useless information is deleted from the dataset; **aggregation** of two or more attributes into a single one to have reduce the size of the dataset; **analyzing missing values** to delete or replace them with a value understandable by the computer; and **sampling**, which focuses on reducing the time of processing the entire dataset by taking only a piece of it (sample) because processing the entire dataset could be time consuming. This can be done in different types such as with or without replacement.

Data Transformation

In the data transformation, data are transformed in a suitable form and to subsequently apply mining algorithms.

- **Normalization**: assigning a value to the data so that each entry falls into a range between 0 and 1. This could be done using the Z-score, whose function is shown in *Formula 3.1* and where X is a continuous variable that represents the data raw score, σ represents the data standard deviation and μ is the mean of the data
- **Conversion**: data are converted in a more useful, organized and structured way for the computer. In case of pictures, audios and videos, those are converted in text form so that they can be analyzed with the standard algorithms
- **Discretization**: it separates the domain of a continuous attribute in a set of intervals to minimize the cardinality of the domain [18]
- **Attribute Transformation**: each value for an attribute is mapped into a new set of values. This can be done with normalization, similarity and dissimilarity values or distances calculation (Euclidean or Minkowski)

$$Z = \frac{X - \mu}{\sigma} \tag{3.1}$$

Formula 1: Z-Score

3.2.1 Classification process

In data mining, classification is a process that assigns class labels to items not yet labeled. The goal is to generate a classification model that assigns the right class label to the right data. This is done by taking as input a training data (or training set) that contains data correctly labeled, this generates the model that will be then used to classify data of a test set, which contains items not yet labeled.
This classification process could be of two main types: supervised and unsupervised.
Supervised classification is made using labels already known, so that, during the classification process, the test set is classified using those class labels. The objective is to find a relation or function that produces the most reliable output label, which highly depends on the training set given as input that represents the so called *ground truth*. Decision trees and Bayes classifiers are some of the supervised classification methods. Classification methods are generally accurate for the outcome quality of the prediction, efficient when it comes to build the model and robust against noise and outliers.
In the **unsupervised** classification, labels are not known so items can be labeled only after the generation of the classification model. An unsupervised classification technique is clustering.

Decision tree

A decision tree is a type of supervised classification technique. It is structured as a flowchart where each internal node tests an attribute of the dataset and the leaf represent the class label, an example is shown

in *Figure 3.4*.

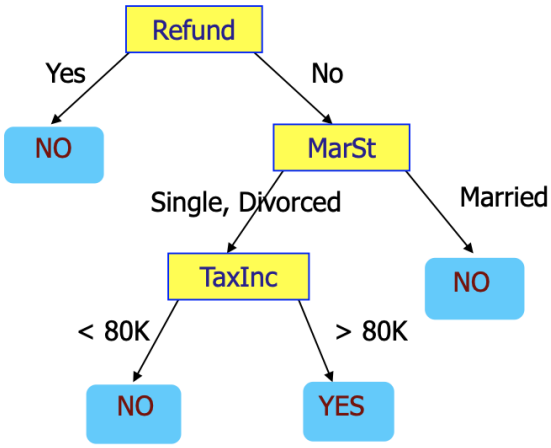

Figure 3.4: Decision tree [19]

When using a decision tree, the analysis starts from a root node and, with a top-down approach, it confronts the dataset items' attributes with each node and then split them in order to have the label that best classifies each item.
Most known and widely used splitting parameters are GINI index, shown in *Formula 3.2*, and Entropy, shown in *Formula 3.3*.

$$GINI = 1 - \sum_{j=1}^{m} f(i,j)^2 \tag{3.2}$$

Formula 2: GINI index

$$Entropy = -\sum_{j=1}^{m} f(i,j) log f(i,j) \tag{3.3}$$

Formula 3: Entropy

A decision tree is generally simple to understand, interpret and visualize, but it may incur in underfitting and overfitting problems in which

the decision tree may be not very precise (underfitting) or may be too specific with a high number of nodes (overfitting). These problems, however, can be partly solved with some halting and pruning solutions.

Bayesian classifier

Bayesian classifier is a family of supervised probabilistic classification techniques based on the Bayes theorem, shown in *Formula 3.5*, which starts from the conditional probability described in *Formula 3.4*. Bayesian theorem supposes that *P(A)* is the "A Priori" or marginal probability: they are calculated independently of each other. The only problem is how to calculate the conditional probability $P(B \mid A)$, this was solved using the **Naive hypothesis**: all attributes are statistical independent, but since it is not always true, it may affect the model quality generating lower accuracy results.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{3.4}$$

Formula 4: Conditional Probability

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{3.5}$$

Formula 5: Bayes Theorem

Clustering

Clustering is a data-mining unsupervised technique in which data are divided into groups, this assures that the homogeneity between items in each group and the heterogeneity between groups are maximum. There are two main type of clustering:

- **Partitional**: as shown in *Figure 3.5*, data items are divided into subsets (clusters) that do not overlap so that each item is inside one specific subset

- **Hierarchical**: as shown in *Figure 3.6* data are organized in a set of nested clusters as a hierarchical tree

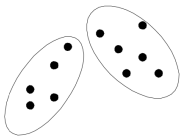

Figure 3.5: Partitional Cluster [20]

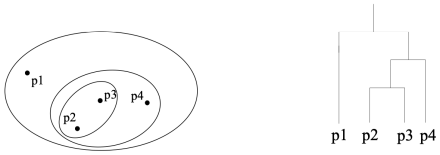

Figure 3.6: Hierarchical Cluster in traditional and dendogram form [20]

The most used algorithms for clustering are K-means, DBSCAN and hierarchical clustering. The results from the these algorithms may then be used for studies such as topic or context analysis and sentiment analysis.
Nowadays data mining techniques are widely used for many different purposes, such as social media mining and text mining.

3.2.2 Text Mining

Text mining is a branch of Natural Language Processing (NLP) that deals with the extraction of unstructured data from textual inputs and its conversion into a usable and meaningful information. This information is then analyzed and classified so that at the end there are documents grouped with the same content.

Data mining mainly deals with unknown and potentially useful information from datasets which are not understandable by humans. Information extracted through text mining is instead easily readable by humans, but not structured in such a way that algorithms can work on it.

Natural Language Processing

Natural Language Processing (NLP) is a branch of artificial intelligence and computer science defined as the process of extracting and analyses information from the human language. This information is then manipulated in order to analyze the meaning of some documents with applications such as sentiment analysis, chatbot, speech recognition, machine translation and advertisement matching. NLP is characterised by two main components: Natural Language Understanding (NLU) and Natural Language Generation (NLG).

Natural Language Understanding (NLU) maps the given input into an analyzable and structured representation. It mainly focuses on the understanding the meaning of the human language, not the processing of it. Therefore, one of the hardest challenge for NLU is the possibility of having texts with different meaning, containing sarcasm or irony which could bring ambiguity and misinterpretation.

Natural Language Generation (NLG) is the process of generating meaningful sentences from an internal representation: the structured data from NLP and NLU are then converted into text written in a human language form.

Text Mining in Social Media

In social media, ideas, thoughts and emotions are often written verbatim and rarely numerically. With this perspective, text mining is useful to perform all the analysis of documents required because it can give a deep meaning of the text extracting its content. Indeed, there are many application of text mining in social media data:

- **Context analysis**: it is used to understand the topic of some documents and group them by their main content
- **Sentiment analysis and Customer Care service**: it is used to understand the feeling of a person. In marketing, it is also used to understand the emotion of a client towards a company and consequentially improve the service

- **Cybercrime prevention**: text mining is also used to identify possible crimes over the internet which are then approved or declined by other entities
- **Language detection**: it is used to understand what language is used in different topics

A tool with which it is possible to use NLP techniques is the Python NLTK library.

NLTK Library

Natural Language Toolkit (NLTK) is a library used in Python for natural language processing techniques as well as all the methods for text processing, such as classification, tokenization and stemming. NLTK improves already existing Python's text analysis and treatment functions, so that it is easier to manipulate and execute complex linguistic analysis.

3.3 Preprocessing phase and data cleaning

After the collection phase, a data cleaning phase is performed in order to delete all the unnecessary data which would have brought noise to the data set. In particular, the removed tweets were:

- Those only containing retweets
- Those that were not written in English which would have caused problems in executing NLP algorithms
- Those duplicated
- Those only containing hashtags, urls or mentions which would have not brought any relevant information

Since these tweets are written by humans, they need a preprocessing phase in order to be more organised, structured and usable by computers. Without this phase, the analysis performed may have brought to wrong results.

3.3.1 Filtering phase

Retweets filter

To remove all the retweets (shared tweets), a filter function has been used that, by scanning the tweet's text, it checked if there was any *#RT* or *@RT* inside the text. A snippet of the code is shown in *Algorithm 3.5*

```
1 if "#RT" and "@RT" not in text
```

Algorithm 3.5: Snippet of code to remove Retweets

Non-English tweets filter

The presence of non-English tweets can lead to different and wrong result when applying NLP algorithms. The detection and removal of those tweets was performed using the *langdetect* library in Python, as shown in *Algorithm 3.6*.

```
1 language = detect(text)
2 if language == 'en' #checks if the text is written in English
```

Algorithm 3.6: Snippet of code to remove non-English tweets

Langdetect is a Python library used for language detection. It currently supports and detect 55 languages with a precision of 99% for 53 of them.[21]
For example, in *Algorithm 3.7*, the Python console returns a result of [**en**: 0.7142848939480231, **it**: 0.2857138245468812 7].

```
1 from langdetect import detect_langs
2 detect_langs("Hello, my name is Antonino and I am from Italy. I
  ↪  like pizza.")
```

Algorithm 3.7: Snippet of example to detect languages

Duplicates filter

To check and filter for duplicated tweets, a confidence function was applied: two tweets were considered duplicated if their similarity score was above a confidence value, if so, they would have been removed from the dataset. This was possible using the class *SequenceMatcher* and, in particular, the method *ratio*.

SequenceMatcher is a class from the Python library *difflib* which compares pairs of sequences of any type, the initialization is shown in *Algorithm 3.8*.

```
__init__(isjunk=None, a='', b='')
```

Algorithm 3.8: Initialization code for SequenceMatcher class

Junks are elements or parts of text the algorithm will not match on, like spaces or blank lines.
The method *ratio* of the SequenceMatcher class returns a float value in [0,1] representing the similarity score between input strings. It sums the sizes of all matched sequences and calculates the ratio as:

$$ratio = 2.0 * \frac{M}{T}$$

where M = number of matches , T = total number of elements in both sequences. If the result is 1.0, then the sequences are identical; if the result is 0.0, then they are totally different and have nothing in common. An example of use is shown in *Algorithm 3.9*.

```
from difflib import SequenceMatcher
if (SequenceMatcher(None, text1, text2).ratio() < confidence)
```

Algorithm 3.9: Sample code to check similarity among two texts

Hashtag, Url and Mentions only filter

Regular expressions were used in order to delete meaningless tweets: those that are composed of only Urls or Hashtags or Mentions or the only combinations of those three which would not have brought any relevant information.

A regular expression (or *regex*) is a sequence of symbols that identifies a group of strings. It is used as search pattern to scan texts and find and manipulate strings.
As shown in *Algorithm 3.10*, in Python the library *re* allows the manipulation of regular expressions.

```
import re
only_url_string =
 ↪  re.split('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+] |[!*\(\),
 ↪  ]|(?:%[0-9a-fA-F][0-9a-fA-F]))+', text)
only_hashtags_string = re.split('^\s*(?:#\w+\s*)+$', text)
only_mentions_string = re.split('^\s*(?:@\w+\s*)+$', text)
```

Algorithm 3.10: Snippet of regex to check if a text only contains urls, hashtags or mentions

3.3.2 Tokenization

Tokenization is the process of forming tokens (sequence of characters) from input stream and it can be defined as the first step in NLP. Tokens can be identifiers, operators, keywords, symbols and constants. For this purpose, the *word_tokenize* method from the NLTK Python library is used which returns a list of the words (punctuation and numbers included) of the text. An example is shown in *Algorithm 3.11*, which gives as result:
['Hello', ',', 'my', 'name', 'is', 'Antonino', 'and', 'I', 'am', 'from', 'Italy'].

```
import nltk
print (nltk.word_tokenize("Hello, my name is Antonino and I am
 ↪  from Italy"))
```

Algorithm 3.11: Sample code for tokenizing a text

3.3.3 Stop words removal

Stop words are commonly used words that do not bring any benefit for the NLP analysis, indeed, search engines automatically ignore those words when inserted.
For this purpose, the *stopwords* from the NLTK Python library is used. An example is shown in *Algorithm 3.12* which prints:
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', ...]

```
1 from nltk.corpus import stopwords
2 print(stopwords.words('english'))
```

Algorithm 3.12: Sample code that shows the stopwords for the English language

3.3.4 Stemming

Stemming is the process of normalizing words into their root form. This process is used to delete unnecessary inflections of words, such as singular-plural or different verb tense, as shown in *Table 3.1*.
The result is less understandable by humans, but more easily comparable for machines and brings to a higher level of reliability when performing the analysis.

There are many algorithms for stemming, the most used ones are:

- **Lookup Table**: it uses a table with an entry for each inflected words and the corresponding root form. There are two main problems: the first is that the dimension of the table in the memory is very big and may cause overstemming; the second is that the unknown words are not handled

- **Suffix-stripping algorithms**: they are based on rules and patterns used to remove suffixes. The main problem is that they do not handle specific or uncommon words
- **Lemmatisation algorithms**: they use different normalization rules based on the context of the word in the sentence. The main problem is that it may be difficult sometimes to identify the right context for that specific word
- **Stochastic algorithms**: they use probability to find the root form of a word given the context of the sentence. The main problem is that sometimes the context may not be accurate, specially when related to irony or sarcasm sentences
- **Porter algorithm**: it is one of the most used, considered also the standard algorithm used for English stemming [22]
- **Hybrid approaches**: they combined two or more of the above algorithms

For this purpose, *PorterStemmer* from the NLTK Python library is used.

Original Word	**Stemmed Word**
Like	Like
Likes	Like
Book's	Book
Booked	Book
Connection	Connect
Connected	Connect

Table 3.1: Some examples of the stemming process

3.3.5 Privacy protection

Privacy protection is important to avoid any sort of tracking. Through the analysis of a user's data, for example the likes that he puts, the pages that he follows, the conversations in which he participates, it is possible to find private information that violate his privacy. For this reason, some words of the tweets were replaced with their synonyms so

that the tweet is no longer traceable to the author, this was possible thanks to the *wordnet* method from the NLTK Python library, then, for those authors, a check if they were an organization, celebrity, hospital or normal people was carried out.

Wordnet

Wordnet is a lexical database for the English language, it is used to get the meaning of words, their synonyms, antonyms and more. An example of use is shown in *Algorithm 3.13*, which returns ['correct', 'proper', 'correctly', 'justly', 'properly'].

```
from nltk.corpus import wordnet
synonym = wordnet.synsets("right")
```

Algorithm 3.13: Sample code for showing synonym substitution

After these steps, the dataset is organised in *username, date, #retweets, #likes, text, mentions, hashtags, location, #followers, #following, profile verified (true or false), processed.*

3.3.6 Summary of the cleaning phase

After the entire cleaning phase, the number of tweets has been decreased of 23.56%, bringing the dataset to be composed of 122,807 tweets. This is summarized in *Table 3.2*.

Social Media	**Tweets before cleaning**	**Tweets after cleaning**	**%Tweets reduced**	**Start tweets date**	**End tweets date**
Twitter	160,658	122,807	23.56%	12/11/2018	12/11/2019

Table 3.2: Summary of cleaning phase

Chapter 4

Results

This chapter covers all the analysis that have been carried out. The aim of the analysis work was to use the most appropriate techniques for establishing what tweets were mostly visualized, liked or retweeted, what months were most popular for writing new tweets and what hashtags were used. A sentiment analysis and a content analysis were then performed to see what reaction did the user have and what were the most written arguments.
In the following sections, an individual user is called *celebrity* if he has a verified profile and has more than 20,000 followers.

4.1 Most retweeted tweets

To analyze the most visualized and reacted tweets, a table was created with the 20 tweets' information that have been more retweeted.
Table 4.1 shows that the two most retweeted tweets were the ones written by a *celebrity* located in Washington, DC.
The other most retweeted tweet was from an *individual* user - person with less than 20.000 followers and without a verified profile - located in London which received 3771 retweets.

#Retweets	Location	Type Of User
7923	Washington, DC	celebrity
3969	Washington, DC	celebrity
3771	London, Europe	individual
1914		individual
1619	London	individual
1576	London	celebrity
1459	Geneva, Switzerland	organization
1372	Corvallis, OR	individual
1162		individual
1141	Johannesburg, South Africa	celebrity
897	London, UK	organization
868	Geneva, Switzerland	organization
624	Baltimore, MD	celebrity
603	General Santos, Philippines	individual
564		individual
556	London, Europe	individual
536	London, England	celebrity
534	Zurich, Switzerland	individual
527	Pakistan	individual
494	Cape Town South Africa	celebrity

Table 4.1: Most Retweeted Tweets Information

4.2 Most liked tweets

Table 4.2 shows the most 20 tweets' information that have been more liked. As also shown in *Table 4.1*, a celebrity user received more likes and the two most liked tweets were the ones that also got many retweets already described in *Section 4.1*.
This shows that a tweet that receives many likes is conducive to being retweeted many times, and vice versa.

#Liked	Location	Type of User
23758	Washington, DC	celebrity
16986	London, Europe	individual
14847	Washington, DC	celebrity
6055	London	celebrity
4244	London	individual
4027		individual
3108		individual
2433	London, Europe	individual
1993	Baltimore, MD	celebrity
1982	London, England	celebrity
1849	London, UK	organization
1803	London	celebrity
1774	London, Europe	individual
1588	Zurich, Switzerland	individual
1470	Manhattan, NY	celebrity
1430	Geneva, Switzerland	organization
1413		individual
1377	London	celebrity
1364	Johannesburg, South Africa	celebrity
1340	Glasgow, Scotland	individual

Table 4.2: Most Liked Tweets Information

4.3 Temporal Analysis

A temporal analysis was performed: an histogram was created by putting on the abscissa axis the year and the month, and on the ordinate axis the number of tweets, it allowed to notice in which month the #HIV hashtag was mostly used. *Figure 4.1* shows the three months with more tweets were December, March and May:

- **December 2018**: the 1st of December is the World AIDS Day, whose hashtag was *#WAD2018* and it was used 1207 times, as also shown in *Figure 4.3*

- **March 2019**: the 10th and the 20th of March there are two awareness days, respectively, National Women and Girls HIV/AIDS and National Native HIV/AIDS
- **May 2019**: the 18th and the 19th of May there are two awareness days, respectively, HIV Vaccine Awareness Day and National Asian & Pacific Islander HIV/AIDS

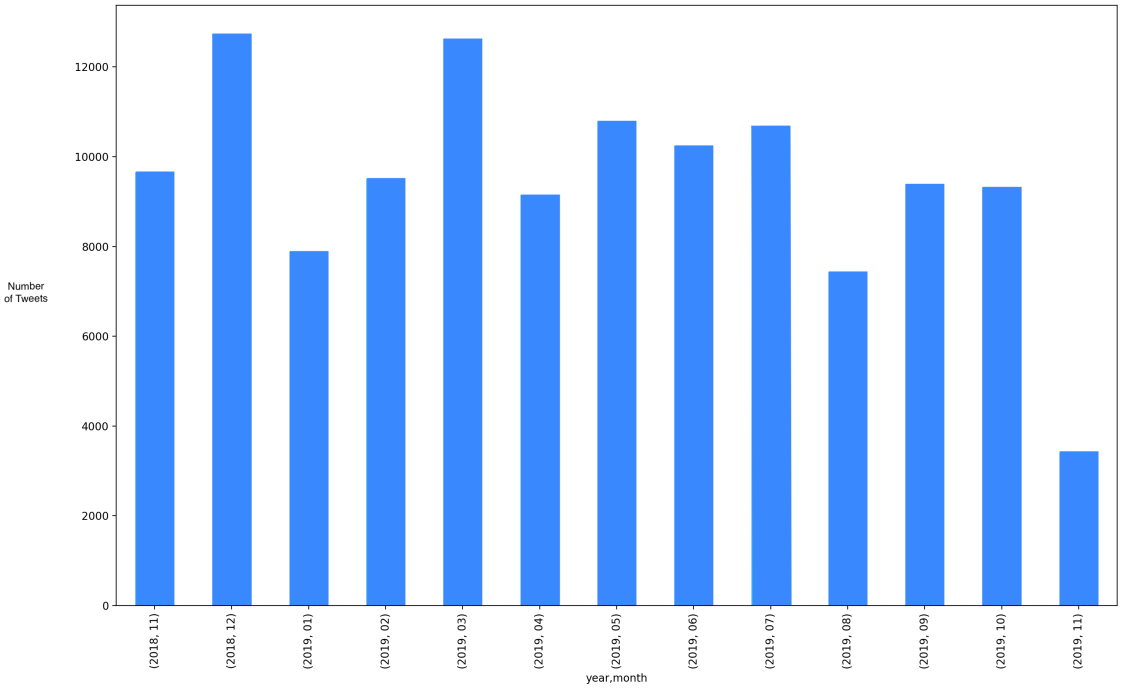

Figure 4.1: Temporal Analysis Histogram - Months view

Figure 4.2 shows the days when many tweets were written. Some of those days are also included in the *Figure 4.1*, representing the peak also for the respective months; some of them are not.
As visible in both graphs, the most tweeted day refers to the month of December, in fact, the 1st of December corresponds to the World AIDS Day and collected 3390 tweets. The other most tweeted day is the 30th and 29th of November which respectively collected 1633 and 1126 tweets, those tweets also referred to the World AIDS Day.
Another most tweeted day is the 5th of March, those tweets both referred to the National Women and Girls HIV/AIDS day and to another

event. In fact, in that day there was a Conference on Retroviruses and Opportunistic Infections (*CROI*) which announced a new possible HIV cure based on stem cells [23] and also talked about HIV prevention.
Those day could have been characterized by more tweets, but they are national events and may not be known outside US.
The 6th of February was another day with many tweets. In fact, the Joint United Nations Programme on HIV/AIDS (*UNAIDS*) listened and accepted what has been said by the President of the US the day before: to stop HIV transmission in the country by 2030. [24]
This 5 most tweeted days are summarized in *Table 4.3*.

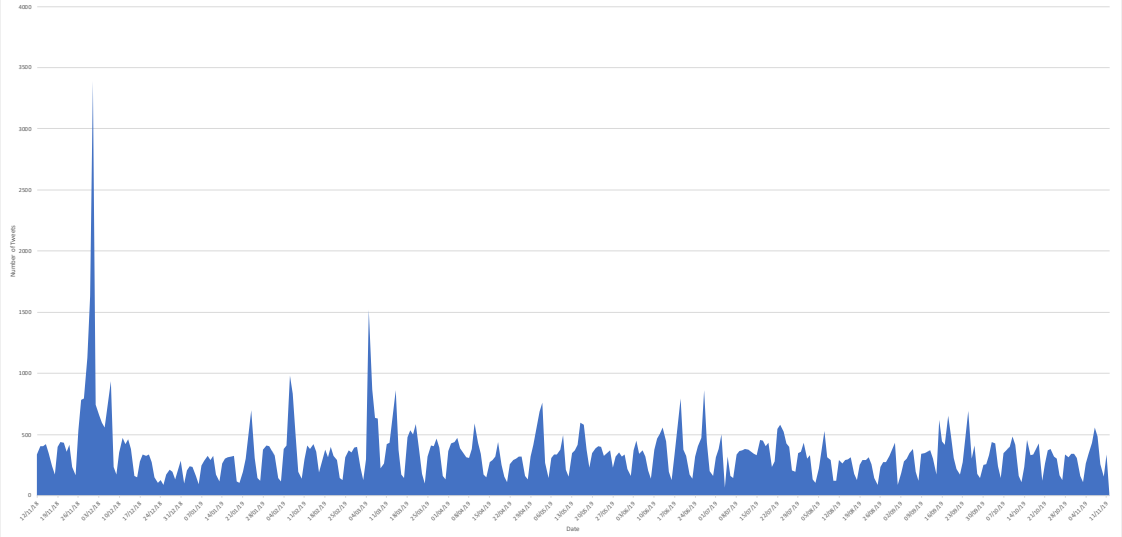

Figure 4.2: Temporal Analysis Histogram - Days view

#Tweets	Date	Event
3390	01/12/2018	World AIDS Day
1633	30/11/2018	a day before the World AIDS Day
1516	05/03/2019	CROI 2019
1126	29/11/2018	two days before the World AIDS Day
984	06/02/2018	US President speech

Table 4.3: 5 most tweeted days

4.4 Hashtag Analysis

All the hashtags in the tweets were analysed to create a table with the 20 most used ones, the number of times they were mentioned and the meaning of their acronym, as shown in *Figure 4.3*.

- **AIDS**: the Acquired Immune Deficiency Syndrome is a disease of the immune system caused when an HIV-positive person is not treated with the antiretroviral therapy (ART) which helps to reduce the viral load in the bloodstream
- **PrEP**: Pre-exposure prophylaxis is a way to prevent HIV in HIV-negative people
- **WorldAIDSDay** or **WAD2018**: Awareness day, 1st of December, as discussed in *Section 4.3*
- **UequalsU**: "Undetectable Equals Untransmittable, it is a campaign which focuses on the fact that if the viral load of a HIV-positive person is consistently undetectable for 6 months, the HIV medications are continuously taken and he gets tested and treated for STIs as needed, HIV is not passed to the sexual partners" [25]
- **STI** or **STD**: Sexually Transmitted Infection or Sexually Transmitted Disease identify all the infections or diseases that can be passed to the sexual partners, such as Chlamydia, Syphilis, etc...
- **TB**: "Tubercholosis is an opportunistic infection (OI). OIs are infections that occur more often or are more severe in people with weakened immune systems than in people with healthy immune systems. HIV weakens the immune system, increasing the risk of TB in people with HIV. Infection with both HIV and TB is called HIV/TB coinfection. Latent TB is more likely to advance to TB disease in people with HIV than in people without HIV. TB disease may also cause HIV to worsen." [26]
- **KnowYourStatus**: incites people to get tested so that they *know their status*.
- **LGBTQ**: Lesbian, Gay, Bisexual, Transgender, Queer, it is a term that identifies the LGBTQ community

- **Stigma**: it is related to an attitude or beliefs about HIV-positive people. It can bring to discrimination when HIV-positive people are treated differently than other people just because of their HIV status.

- **SRHR**: "Sexual and Reproductive Health and Rights, it is the concept of human rights applied to sexuality and reproduction. It is a combination of four fields that in some contexts are more or less distinct from each other, but less so or not at all in other contexts. These four fields are sexual health, sexual rights, reproductive health and reproductive rights." [27]

- **HCV**: "Because both HIV and HCV can spread in blood, a major risk factor for both HIV and HCV infection is injection drug use. Sharing needles or other drug injection equipment increases the risk of contact with HIV- or HCV-infected blood. According to the Centers for Disease Control and Prevention (CDC), approximately 25% of people with HIV in the United States also have HCV. Infection with both HIV and HCV is called HIV/HCV coinfection. In people with HIV/HCV coinfection, HIV may cause chronic HCV to advance faster" [28]

- **IAS2019**: Conference on HIV Science organised in Mexico City, Mexico

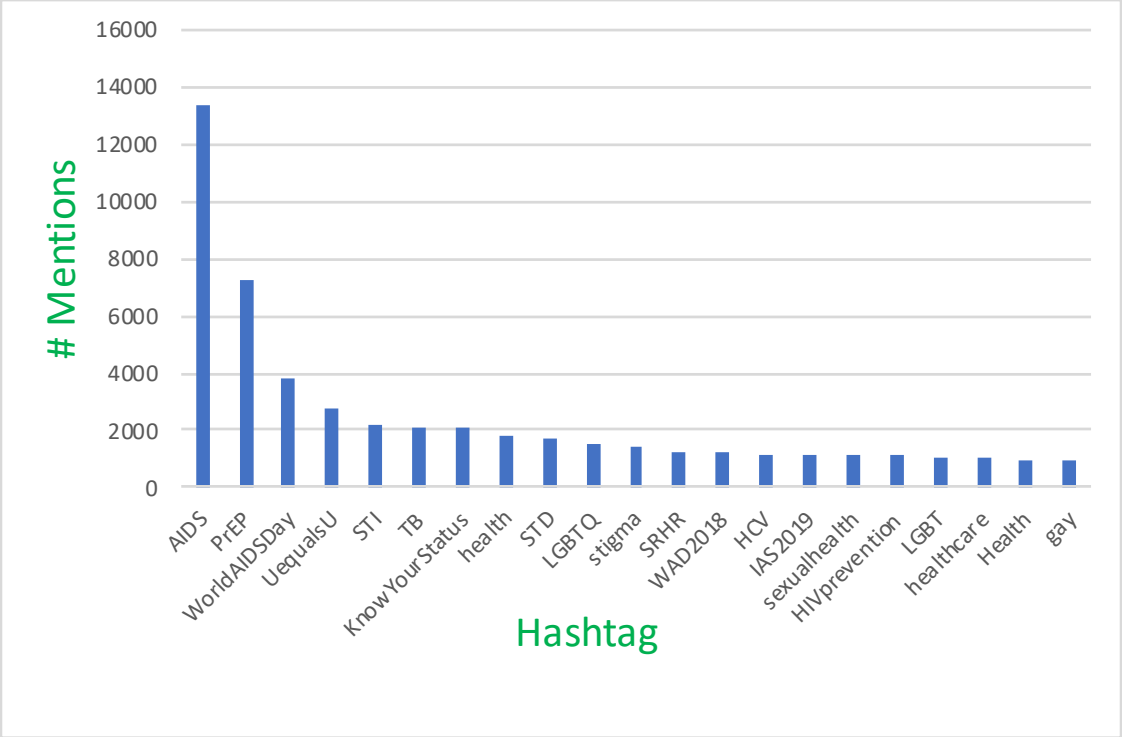

Figure 4.3: Hashtag Analysis Result

4.5 Sentiment Analysis

Sentiment analysis is an application of text mining techniques that measures people's opinions through natural language processing (NLP). These opinion are generally collected from social media and review sites. When data are analyzed, a polarity score shows the inclination of people's texts, that inclination could be expressed with *Positive*, *Neutral* or *Negative* result.
These applications are always evolving, but still present many problems when analyzing a text, specially those containing sarcasm and irony which could bring to the misinterpretation of the text.

The sentiment analysis performed using an unsupervised technique referred to the *TextBlob* Python library and involves using a rule-based approach to analyze the text.
As shown in *Figure 4.4*, the sentiment analysis returned tweets classified

as mostly Positive and Neutral, which only the 11.45% was Negative. Those that have been classified as negative have words such as "shame" or "stigma" or "afraid" which were considered negative by the method.

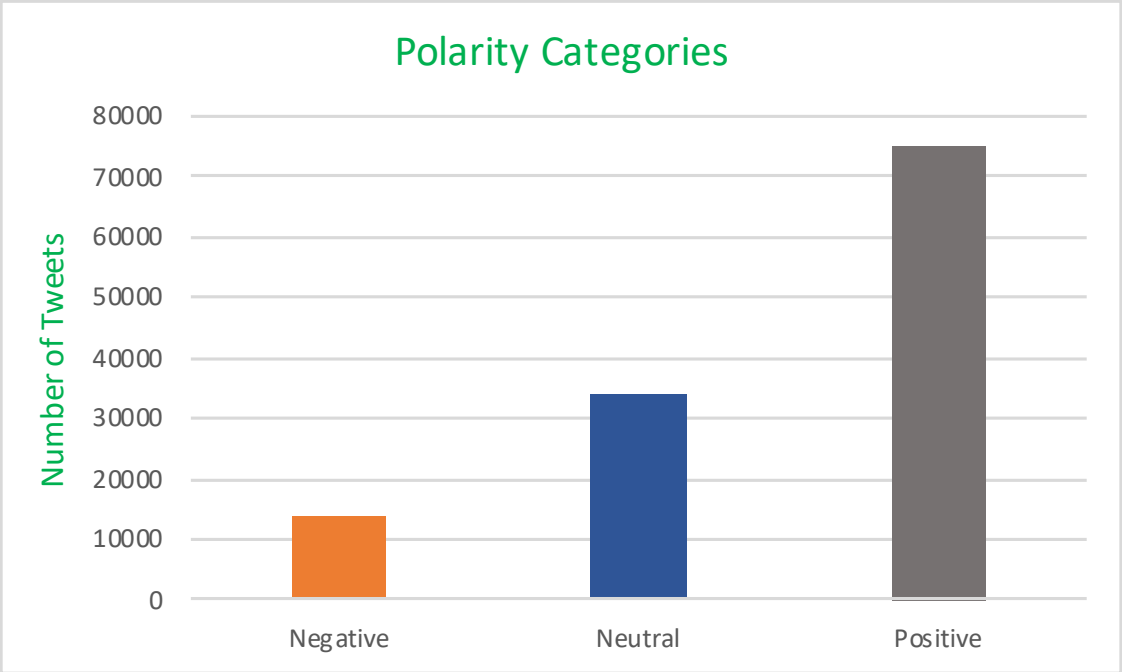

Figure 4.4: Sentiment Analysis Result

TextBlob

TextBlob is a Python library, based on NLTK, that allows the manipulation of words and sentences. To make a sentiment analysis, it is necessary to use the *sentiment* method whose *polarity* field returns a value in [-1, 1]: if the value is greater than 0, then the text is classified as positive; if the value is lower than 0, then the text is classified as negative; otherwise it is classified as neutral. An example of use of the library is shown in *Algorithm 4.1*, which gives a result of 0.5. The output got from TextBlob shows also a subjectivity score which could be in [0, 1] and quantifies the amount of personal opinion contained in the text.
The implementation of the sentiment analysis follows a Naive Bayes algorithm trained on a dataset of movie reviews.

TextBlob could also be used for other NLP-purposes such as Part-of-speech tagging, a Naive Bayes or Decision Tree classification and spelling correction.

```
from textblob import TextBlob
text = TextBlob("Hello, I love pizza.")
result = text.sentiment.polarity
```

Algorithm 4.1: Sample code for TextBlob

4.6 Content Analysis

Content analysis is a quantitative research technique used to compress large amounts of text, expressed as qualitative data, into categories. This is done with a process of identifying the meaning of the messages in general texts or images and then assign them labels so that they can be grouped by their content or topic. Those contents are then analyzed for their meaning and the relationships between them.

4.6.1 Latent Dirichlet Allocation (LDA)

The analysis uses the **Latent Dirichlet Allocation (LDA)**, an unsupervised generative modelling technique to train the model. A generative modelling technique is a machine learning model that generates an output considering the distribution of some objects. The LDA technique makes two assumption:

- The texts are produced from a mixture of topics. Each text belongs to a topic with a certain degree
- Each topic depends on the words' probabilities: words frequently occurring together will have more probability, as shown in *Table 4.4* and in *Table 4.5*, where the element $P(\mathrm{A}|\mathrm{B})$ represents the conditional probability whose definition has been described in *Section 3.2.1*

	Topic 1	**Topic 2**	**...**	**Topic K**
Document 1	$P(\mathrm{t}_1\|\mathrm{d}_1)$	$P(\mathrm{t}_2\|\mathrm{d}_1)$	...	$P(\mathrm{t}_\mathrm{K}\|\mathrm{d}_1)$
Document 2	$P(\mathrm{t}_1\|\mathrm{d}_2)$	$P(\mathrm{t}_2\|\mathrm{d}_2)$	...	$P(\mathrm{t}_\mathrm{K}\|\mathrm{d}_2)$
...	...	...	...	...
Document M	$P(\mathrm{t}_1\|\mathrm{d}_\mathrm{M})$	$P(\mathrm{t}_2\|\mathrm{d}_\mathrm{M})$	...	$P(\mathrm{t}_\mathrm{K}\|\mathrm{d}_\mathrm{M})$

Table 4.4: Document - Topic Distribution

	Word 1	**Word 2**	**...**	**Word N**
Topic 1	$P(\mathrm{w}_1\|\mathrm{t}_1)$	$P(\mathrm{w}_2\|\mathrm{t}_1)$	...	$P(\mathrm{w}_\mathrm{N}\|\mathrm{t}_1)$
Topic 2	$P(\mathrm{w}_1\|\mathrm{t}_2)$	$P(\mathrm{w}_2\|\mathrm{t}_2)$	...	$P(\mathrm{w}_\mathrm{N}\|\mathrm{t}_2)$
...	...	...	...	...
Topic K	$P(\mathrm{w}_1\|\ \mathrm{t}_\mathrm{K})$	$P(\mathrm{w}_2\|\mathrm{t}_\mathrm{M})$	...	$P(\mathrm{w}_\mathrm{N}\|\mathrm{t}_\mathrm{K})$

Table 4.5: Topic - Word Distribution

When using an LDA model, it is important to specify the number of topics of interests for which the model is going to be trained to find in the texts. This is done by doing many tests on multiple LDA models to find the one the has the highest **coherence score**, but at the same time the lowest number of topics. The coherence score is, in fact, a measure of how good the produced model is. Each text will then be labeled by the main topic which also contains other texts within the same topic. This topics could then be considered as clusters in which the coherence score is the distance between those clusters. A representation of the coherence score value is shown in *Formula 6*.

$$Coherence = \sum_{i<j} score(w_i, w_j) \tag{4.1}$$

Formula 6: Coherence score value

The higher the coherence score, the less the overlapping topics will be. A graphic representation of the LDA model is shown in *Figure 4.5*.

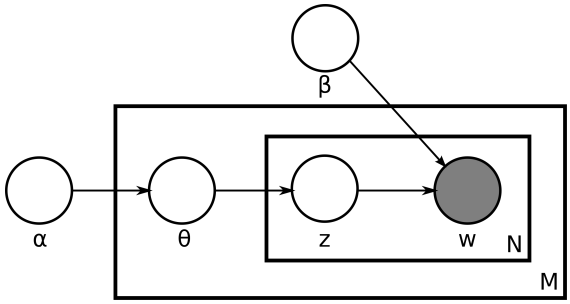

Figure 4.5: Graphic representation of LDA model[29]

4.6.2 Topics of interest

When doing a content analysis, the number of topics of interest is relevant. There are two main LDA algorithms to analyze different documents and retrieve their main topics: Mallet (**MA**chine **L**earning for **L**anguag**E** **T**oolkit) and Gensim. The general difference between Mallet and Gensim's LDA is that Gensim uses a type of Bayes sampling, called Variational Bayes sampling, which is faster but, at the same time, less precise that Mallet's.[30]

In order to find the optimal number of topics, many different values were tried. A value of number of topics can be considered optimal if it gives the highest coherence value score, a measure of how good a given topic model is. As shown in *Figure 4.6*, multiple LDA models were trained to find the one that has the highest coherence score, but, at the same time, the lowest number of topics because choosing a high-value can result in having more granular sub-topics. This plot is called *elbow* plot because when the coherence score keeps increasing, the first model that gives the highest score before flattening out is picked. As shown in *Table 4.6*, the number of topics that guarantees a good balance between high coherence score and low number of topics is 8.
The analysis with 8 topics is shown in *Section 4.6.2*.

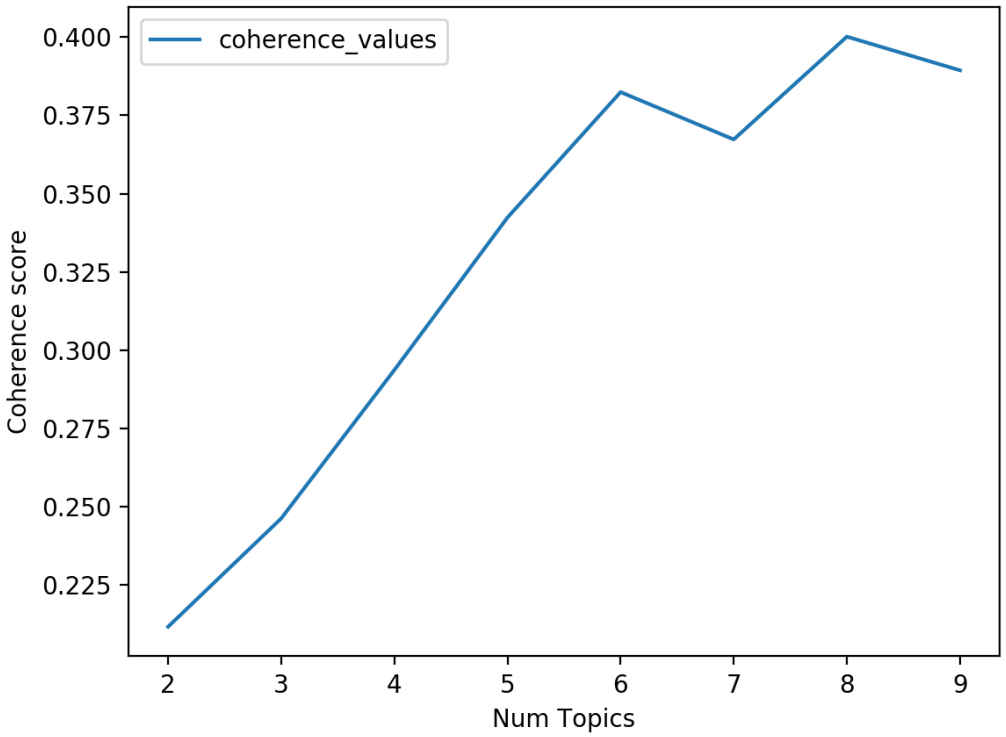

Figure 4.6: Optimal topic model graph with coherence scores

#Topics	Coherence scores
2	0.2116
3	0.2462
4	0.2936
5	0.3424
6	0.3824
7	0.3673
8	0.4002
9	0.3894

Table 4.6: Coherence scores for different topic models

As shown in *Algorithm 4.2*, the LDA model was provided by the

gensim library and it was built with *num_topics = 8*.

```
lda_model_scheme =
↪  gensim.models.ldamodel.LdaModel(corpus=corpus_chosen,
↪  id2word=dictionary_chosen, num_topics=8, passes=10)
```

Algorithm 4.2: LDA model training

The number of passes indicates how many times the model passes through the corpus during training, while *corpus* and *id2word* indicates respectively the words collected as vector and the vocabulary size used for topic printing.
To represent and visualize the topic modeling, the Python library *pyLDAvis* was used, whose characteristics are:

- The saliency value of a word measures of how much that specific word tells about the topic
- The λ value is set to 1 and it represents the relevance metric
- The size of the bubbles measure the importance of the topics related to the data which is measured as a percentage of tokens
- The distance between two bubbles (topics) represents their semantic relationship and their inter-topic distance
- When a topic is selected, it shows the frequency of each word given that topic, in descending order. An example is shown in *Figure 4.8* in the Section 4.6.2, related to a model with 5 topics
- By default, the x and y axes are labeled PC1 and PC2, which stands for Principal Components 1 and Principal Components 2

An example of a number of topics different than the recommended one is shown in *Section 4.6.2*. Indeed, it shows different size bubbles, proving that the LDA model trained with *num_topic = 5* gives an insufficient coherence score.

Number of topics = 5

Figure 4.7 shows the most 30 salient terms collected between the 5 topics. It is visible that the most salient words are "people", "health" and

"living" showing that users usually write about HIV referring to people, their health condition, inviting them to get tested to prevent the spread of the virus and being aware of their status.
Figure 4.8 shows the most influential topic of the data whose words could be summarised into "testing" and "awareness" main keyword because the lack of awareness can facilitate the spread of the disease. Indeed, it has been shown that HIV-positive people partially reduce the risk of the infection once they know about their condition [31].

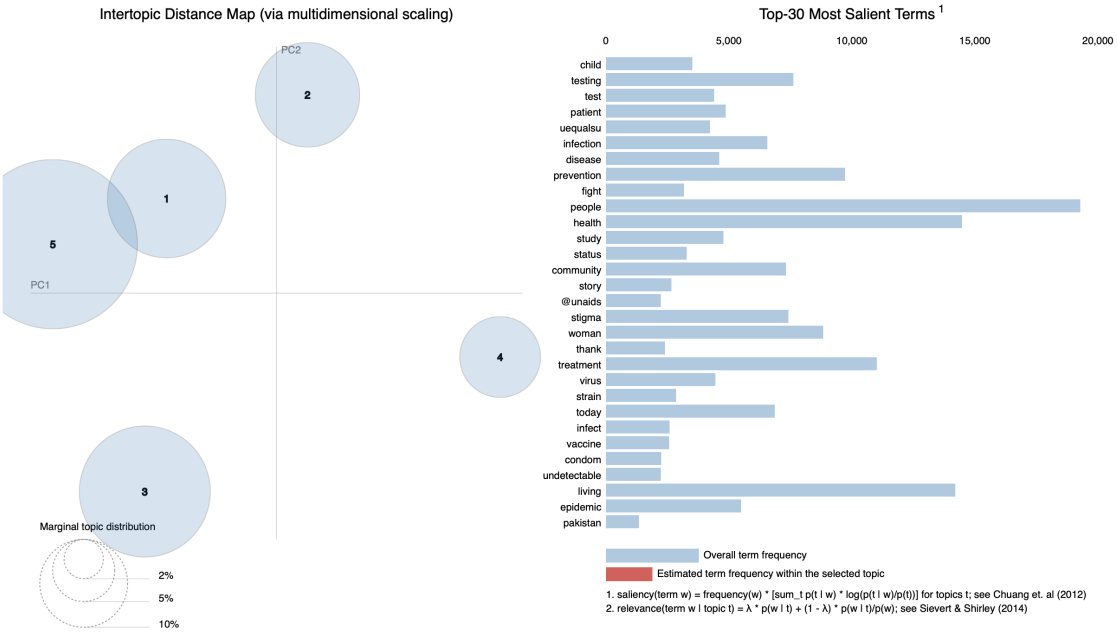

Figure 4.7: Model visualization with 5 Topics

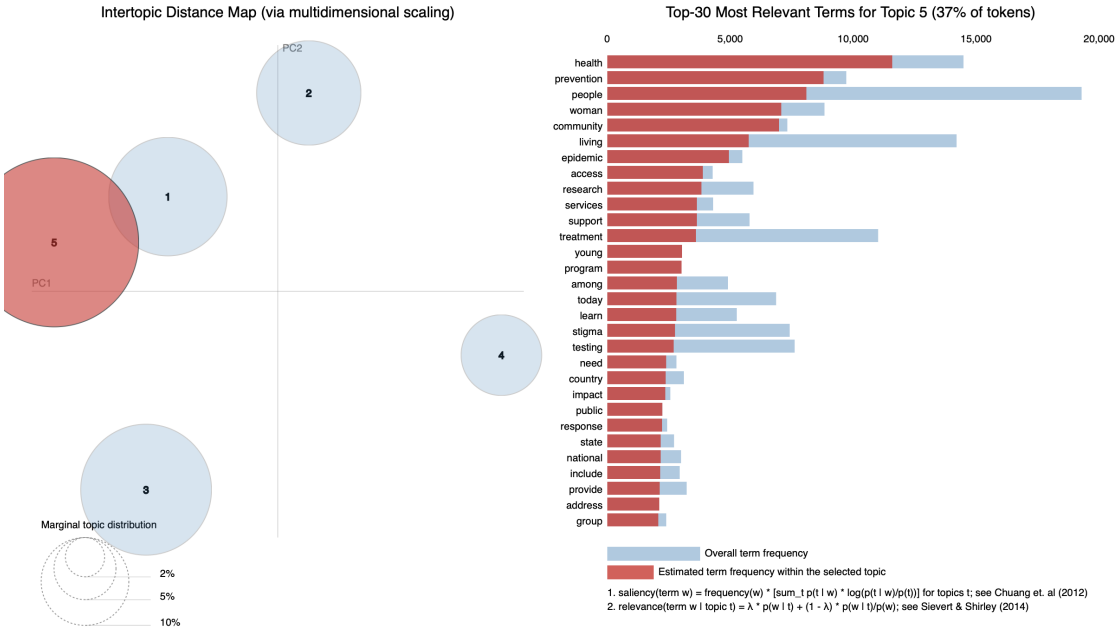

Figure 4.8: Model visualization with 5 Topics, one topic selected

Number of topics = 8

Figure 4.9 shows the most 30 salient terms collected between the 8 topics. The all equal sized bubbles on the left of the figure prove that the chosen number is the best one and represents the topics of the tweets, returning the highest coherence value. It is visible that one of the most salient term is "aid", "live", "test" and "treatment", this shows that people encourage others to get tested and live a normal life using the necessary treatments.

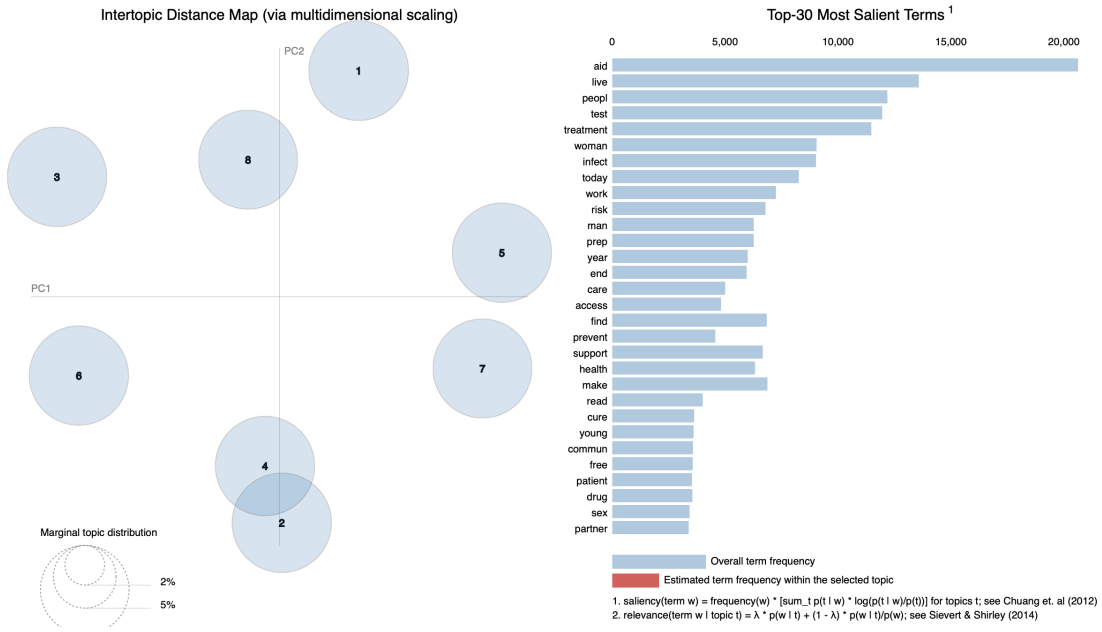

Figure 4.9: Model visualization with 8 Topics

Chapter 5

Discussion

This work performs different types of analysis applied to the same dataset, this gives detailed information on the main topics covered and emotions transmitted on the tweets. It also gives an overview of the type of users and their ability to express themselves and open up by sharing their experiences to help those in the same situations.

Analyzing the most retweeted and liked tweets, shown in *Section 4.1* and *Section 4.2* the two most retweeted tweets were the ones written by a *celebrity*, this can be seen as if users are likely to trust and share what celebrities say. In particular, that *celebrity* wrote that a donation would have been made in the US that would have enabled the purchase of HIV medicines and preventive treatments for over 100,000 people, that could easily explain the number of retweets. Another tweet that got many likes was from a user that spoke about his personal experience of living with HIV, showing how he is now living a normal life, recommending other people to get tested and to be treated if necessary. As shown in those two sections, engagement levels of tweets by relevant organizations are not very high. They should promote HIV testing via Twitter which can subsequently be used as a good option in proactive support, such as encourage people to use condoms or other contraceptive methods. Indeed, the possibility of using HIV conversation on social media could be a method to detect HIV outcomes encouraging users to get tested. Most of the tweets that got more likes and retweets were from individuals and celebrities and not from organizations, this

could be explained because some organizations may have contacted celebrities to speak about HIV because they might be more influential and have more public response than organizations.

Professionals should therefore promote the use of condoms and they should try to convince people not to be embarrassed about obtaining condoms or other contraceptive methods. A study shows that the embarrassment associated with purchasing condoms exceeds that of using condoms and people generally use different strategies to purchase condoms because they feel embarrassed, this attitude generally decreases with age and experience. [32]
Fortunately today it is possible to easily buy condoms online even if their use and the taboo that revolves around it is always relevant, in heterosexual and homosexual relationships.

In an interesting study, participants described how not using the condom established trust, making the relationship to grow to a first level. Some of them preferred unprotected intercourses while knowing the risk of STDs, because that created a sort of intimacy placing the emotions they proved over concerns about their health. Other people used a sort of "negotiation", like using condoms until their HIV-status was defined after tests, or even using condoms until their relationship was a monogamy. [33]

The main topic of the dataset is about testing and prevention. This is particularly noticeable in the peaks of the days with the greatest number of tweets that concerned World AIDS Day, in *Section 4.3*, but also in the content analysis, in *Section 4.6*.
In particular, as visible in *Table 4.3*, there are days that have generated a large number of tweet even if they are not mainly focused on HIV. These results may contrast with the monthly view for the highest number of tweets showed in *Figure 4.1*, which shows that a large number of tweets had occurred during the month of May, probably justified for National Asian & Pacific Islander HIV/AIDS, yet in those days there are no peaks of tweets. This could be explained since the event is national or generally restricted to a part of the globe and they event is

not known worldwide.
Many tweets have also been written about donations that have been made or promises of hope that have been said. This is noticeable from the peak of the tweets for the 6th of February because the day before the President of the US, Donald Trump, made a speech in which he said that by 2030 there would not be HIV transmission anymore in the country. [24]

Social media are therefore used by users to create a sense of community, breaking down geographical barriers and defeating stigma. The word "stigma" is also visible between the most used hashtags, in *Section 4.4* and represents the discrimination regarding HIV-positive people just because of their HIV status. A study compared stigma among people with HIV and stigma among people with cancer.
HIV-positive people suffered more of stigma than the individuals with cancer. The effect of these two conditions are mostly felt through the mechanism of stigma, because the type of illness does not appear to directly discourage people's self-perception. [34]

"Stigma" is also used in many tweets which have been classified as negative when performing the sentiment analysis in *Section 4.5*. This can be considered a limit of the technique used because those tweets spoke in particular of eliminating stigma and breaking down the fear of one's status, but they were misinterpreted. Tweets classified as neutral were the ones that contained some statistics about a country or some specific term related to HIV, an example is "pre-exposure prophylaxis" or "PrEP".

Different algorithms were also tried for the sentiment analysis, but most of them needed supervised learning methods which would have brought to manually put labels (i.e. negative: -1; neutral: 0; positive: 1) to set up a training set. This could have taken too much time because of the dimension of the dataset (122,807 tweets). At the end, an algorithm was tried with R programming language, the *sentiment* package and, in particular, the *classify_emotion* method that classifies the emotion of a set of texts using a naive Bayer classifier trained on Strapparaca

and Valitutti's emotions lexicon [35]. However, this algorithm did not produce any useful result because the algorithm just focused word by word to understand the polarity and put the label and not by the general context of the sentence.

The content analysis showed the most used topics which, as shown in *Section 4.6*, were mostly referred to "aid", "test" and "treatment". Users therefore encourage people to ask for help if someone thinks of being HIV-positive, because of some unprotected sex, thus, increasing the possibility of get tested, being treated and then leading to a normal life.

A topic that could be considered missing in the result of the content analysis is the one about pregnancy by HIV-infected women. Indeed, women living in poor countries often need support and also services to understand the risk that they and their baby might have.
Indeed, a study shows that many women think that their baby would not be HIV-positive if they became pregnant even if they had HIV, this can be true because the risk of HIV transmission from mother to child may be reduced by treatment with HIV antiretroviral medicines that could be initiated in women before, during, and after pregnancy. [36] [37]

5.1 Answers to the Research Questions

This section shows the answers to the research questions of the work presented in the Introduction, *Section 1.3*.

5.1.1 RQ1: individual's experience of living with HIV

The first research question was *Are Twitter individual users willing to share their personal experience of living with HIV?*.
Many tweets were read and analyzed, only few of them written by organizations that invite people to get tested. Nevertheless, many of them were written by individual and celebrity users who demonstrate

how, despite being HIV-positive, it is possible to live a normal life if all the conditions to treat the pathology are adopted and followed, such as take the necessary care and get advice from doctors. Those tweets got many likes and retweets, as showed in *Section 4.1* and in *Section 4.2* and they were written by individuals and celebrities who had more feedback (visible by the number of retweets and likes) from other users than tweets written by organizations which did not have a high level of engagement. This could mean that people admire their lifestyle and the courage of writing publicly on a social platform creating bonds between users so that those with a HIV-positive status do not feel alone and see that their life is not over.

5.1.2 RQ2: emotion expressed by the users

The second research question was *How are interpreted and classified the emotions among HIV-related tweets?*.
Many tweets encourage other users to break down the stigma that revolves around HIV, but since the method sees the word "stigma" or "afraid", they have been classified as negative. The positive tweets are from users that share their thoughts and experiences giving courage to other users, this creates a community of people united by the same disease, to give themselves strength, comfort and advice. Most of the neutral tweets were from organizations that write about donations that have been made or about some statistical results from treatment and medications.

5.1.3 RQ3: relevant discussed topics

The third research question was *What are the most discussed topics among twitter users who write about HIV?*.
The most discussed topics were about living and aid, this reinforces the ability of social media to act as a community. Another important topics were related to get tested or treated with medications.
An important aspect is how also the topic "woman" is relevant. This may be connected to the disbelief that women can not get HIV, but only men and in particular, only homosexuals. This is not true, in fact,

prevention with tests must be done by both genders who have recently had unprotected sexual intercourse.
Summarized, the most discussed topics were:

- living and aid
- get tested
- treatment medications
- woman

New methods for how to use social media and "big data" in healthcare field have been developed and they can be considered as the first step in establishing how these data can be used in prevention, detection and treatment of diseases.

Chapter 6

Conclusions

This work focused on the extracting, cleaning and then analyzing texts (tweets) written in a social media platform, Twitter. Tweets were written by individual users or organization and for this reason, the dataset had also to be filtered from retweets, non-english texts and tweets only containing hashtag, url and mentions.
The analysis was performed thanks to a combination of social media, data mining and text mining techniques to better understand the communication around HIV on social media, focusing on the feelings expressed and on the most discussed topics.

The result of this work could be useful for public health figures to understand people's needs and thoughts about HIV on Twitter and provide a summary of public opinion to win the fight against HIV/AIDS. It also illustrates the potential of using social media to analyse the spread of viruses or diseases using two types of analyses for the same topic and dataset: sentiment and content analysis. Indeed, the algorithms and techniques presented could therefore be useful to study and analyse not only hashtags related to viruses or diseases but for every topic of interest. Related to viruses and diseases, public health community or, in general, data scientists could use these techniques to analyze other tweets, showing how social media can be helpful to identify, detect and study outbreaks in a specific geographical area and in a specific period of time.

6.1 Limitation and future developments

A general limitation could be noticed from the dataset: the majority of the tweets were from developed countries and are therefore not representative of the entire population. This could be considered a sensor of how Twitter, and maybe other social media platforms and social network sites, are not popular outside developed countries.[38]

Twitter is sometimes used to express thoughts, most of the times in an informal way even with the use of irony and sarcasm, this could bring the algorithms to misinterpret the texts and give low accuracy results. This was also noted in the results which showed how some tweets were marked as negative, not because they were, but because they had words like "stigma" and "fear" even if used ironically, that led to a partially wrong result for the sentiment analysis. It could also happen that real negative tweets are promptly reported by the community and then subsequently removed, thus leading to an imbalance towards positive tweets. The length of the tweets is of significant importance for sentiment analysis because with short texts it is likely to incur in unreliable results. However, this has been improved by increasing the maximum number of characters that can be written in a tweet. A possible solution could be to compare the results obtained with sentiment analysis with a human judgment. However, this is highly dependent on the size of the available dataset.

Context analysis, just as the sentiment analysis, has the problem of discrediting the word from an ironic or sarcastic context. Certain words placed in different contexts can in fact express diametrically opposite meanings. It is also highly time consuming [39] and not so easy to computerize because the results obtained must then be reviewed by the human eye to try to understand the topic in common between certain words.

Although the data, algorithms and results have these limitations, we can conclude that the results obtained are in line with what had been

expected from the beginning and they highlighted different social aspects: how social media reflect the average population thoughts and how a sense of community breaks down geographical barriers, allowing people to confront and open up, expressing themselves.

As future developments, for the sentiment analysis it could be possible to use a vocabulary of lexicons used for the expression of feelings and a Like/Retweet analysis because a high number of likes and retweets could be an index of a positive message. Both methods, combined with a supervised technique could then allow to increase the accuracy of the analysis and give better results.
For the content analysis, it could be possible to use labels generated by the LDA method to create a training set. This training set could then be taken as input to generate a classification model (with techniques such as Bayesian or SVM or KNN) to label new texts. This would take as granted the possible errors made by LDA but, at the same time, could be used for real time tweets, speeding up the entire content classification and allowing health authorities to respond to public concerns.

Appendices

Appendix A

List of Algorithm

Appendix B

List of Tables

Appendix C

List of Figures

Bibliography

[1] Waddell, L. and Hernandez, S., "Online in 60 seconds [infographic] – a year later." https://blog.qmee.com/online-in-60-seconds-infographic-a-year-later, 2019. [Online; accessed 17-February-2020].

[2] C. B. Williams, "Introduction: Social media, political marketing and the 2016 us election," 2017.

[3] H. Korda and Z. Itani, "Harnessing social media for health promotion and behavior change," *Health promotion practice*, vol. 14, no. 1, pp. 15–23, 2013.

[4] F. Chen and D. B. Neill, "Non-parametric scan statistics for disease outbreak detection on twitter.," *Online journal of public health informatics*, vol. 6, no. 1, 2014.

[5] HIV.gov, "About hiv & aids," 2019, (accessed 24-January-2020). https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids.

[6] A. Kotov, "Social media analytics for healthcare.," *Healthcare data analytics*, vol. 1, pp. 309–340, 2015.

[7] K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using twitter data: demonstration on flu and cancer," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1474–1477, 2013.

[8] A. A. Khan and S. Krishna, "Using social media in primary healthcare settings," 2013.

[9] Broniatowski, David A and Paul, Michael J and Dredze, Mark, "National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic," *PloS one*, vol. 8, no. 12, 2013.

[10] C. Chew and G. Eysenbach, "Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak," *PloS one*, vol. 5, no. 11, 2010.

[11] Y. Bao, Y. Sun, S. Meng, J. Shi, and L. Lu, "2019-ncov epidemic: address mental health care to empower society," *The Lancet*, 2020.

[12] Taggart, Tamara and Grewe, Mary Elisabeth and Conserve, Donaldson F and Gliwa, Catherine and Isler, Malika Roman, "Social media and hiv: a systematic review of uses of social media in hiv communication," *Journal of medical Internet research*, vol. 17, no. 11, p. e248, 2015.

[13] Z. Divecha, A. Divney, J. Ickovics, and T. Kershaw, "Tweeting about testing: Do low-income, parenting adolescents and young adults use new media technologies to communicate about sexual health?," *Perspectives on sexual and reproductive health*, vol. 44, no. 3, pp. 176–183, 2012.

[14] Joshua Roesslein, Harmon758, Aaron Hill, "Getoldtweets-python," 2016, (accessed 25-January-2020). https://github.com/tweepy/tweepy/blob/master/.

[15] Jefferson-Henrique, "Get old tweets programatically," 2016, (accessed 25-January-2020). https://github.com/Jefferson-Henrique/GetOldTweets-python/blob/master.

[16] Twitter Help Center, "About verified accounts," 2019, (accessed 26-January-2020). https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts.

[17] Elena Baralis, "The data mining process." https://dbdmg.polito.it/wordpress/wp-content/uploads/2018/10/5-DMProcess.pdf. [Online; accessed 25-February-2020].

[18] Elena Baralis, "Data preprocessing." https://dbdmg.polito.it/wordpress/wp-content/uploads/2018/10/6-DMPreProc.pdf. [Online; accessed 25-February-2020].

[19] Elena Baralis, "Classification fundamentals." https://dbdmg.polito.it/wordpress/wp-content/uploads/2018/10/8-DMClassification.pdf. [Online; accessed 30-May-2020].

[20] Elena Baralis, Tania Cerquitelli, "Clustering fundamentals." https://dbdmg.polito.it/wordpress/wp-content/uploads/

2018/10/9-DMClustering.pdf. [Online; accessed 24-February-2020].

[21] Software in the Public Interest (SPI), "Package: python3-langdetect (1.0.7-4)," 2019, (accessed 26-January-2020). https://packages.debian.org/en/sid/python3-langdetect.

[22] Wikipedia contributors, "Stemming — Wikipedia, the free encyclopedia." https://en.wikipedia.org/w/index.php?title=Stemming&oldid=927284897, 2019. [Online; accessed 23-February-2020].

[23] K. Allers, G. Hütter, J. Hofmann, C. Loddenkemper, K. Rieger, E. Thiel, and T. Schneider, "Evidence for the cure of hiv infection by ccr5δ32/δ32 stem cell transplantation," *Blood, The Journal of the American Society of Hematology*, vol. 117, no. 10, pp. 2791–2799, 2011.

[24] S. Barton-Knott, "Unaids welcomes pledge by the president of the united states of america to stop hiv transmission in the country by 2030." https://www.unaids.org/en/resources/presscentre/pressreleaseandstatementarchive/2019/february/20190206_usa, 2019. [Online; accessed 16-March-2020].

[25] Sarah Chown, "Undetectable = untransmittable." http://www.youthco.org/uequalsu, 2017. [Online; accessed 28-January-2020].

[26] aidsinfo.nih.gov, "Hiv and tuberculosis (tb)." https://aidsinfo.nih.gov/understanding-hiv-aids/fact-sheets/26/90/hiv-and-tuberculosis--tb-, 2019. [Online; accessed 28-January-2020].

[27] Wikipedia contributors, "Sexual and reproductive health and rights — Wikipedia, the free encyclopedia." https://en.wikipedia.org/w/index.php?title=Sexual_and_reproductive_health_and_rights&oldid=936235665, 2020. [Online; accessed 28-January-2020].

[28] aidsinfo.nih.gov, "Hiv and hepatitis c." https://aidsinfo.nih.gov/understanding-hiv-aids/fact-sheets/26/88/hiv-and-hepatitis-c, 2019. [Online; accessed 28-January-2020].

[29] Wikipedia contributors, "Latent dirichlet allocation — Wikipedia, the free encyclopedia." https://en.wikipedia.org/w/index.php?title=Latent_Dirichlet_allocation&oldid=953599124, 2020. [Online; accessed 22-June-2020].

[30] Greg Rafferty, "Lda on the texts of harry potter - topic modeling with latent dirichlet allocation." shorturl.at/qtGZ3, 2018. [Online; accessed 20-February-2020].

[31] Brescia, Valerio and Myriam, Caratù and Scaioli, Giacomo, "A community-based social marketing strategy to prevent hiv and fight stigma," *International Journal of Business and Management; Vol. 14, No. 10*, 2019.

[32] S. G. Moore, D. W. Dahl, G. J. Gorn, C. B. Weinberg, J. Park, and Y. Jiang, "Condom embarrassment: coping and consequences for condom use in three countries," *AIDS care*, vol. 20, no. 5, pp. 553–559, 2008.

[33] A. M. Corbett, J. Dickson-Gómez, H. Hilario, and M. R. Weeks, "A little thing called love: Condom use in high-risk primary heterosexual relationships," *Perspectives on sexual and reproductive health*, vol. 41, no. 4, pp. 218–224, 2009.

[34] B. L. Fife and E. R. Wright, "The dimensionality of stigma: A comparison of its impact on the self of persons with hiv/aids and cancer," *Journal of health and social behavior*, pp. 50–67, 2000.

[35] Timothy P. Jurka, "classify_emotion: classifies the emotion (e.g. anger, disgust, fear, joy,...." https://rdrr.io/github/abhy/sentiment/man/classify_emotion.html, 2019. [Online; accessed 02-March-2020].

[36] Wikipedia contributors, "Hiv and pregnancy — Wikipedia, the free encyclopedia." https://en.wikipedia.org/w/index.php?title=HIV_and_pregnancy&oldid=951409686, 2020. [Online; accessed 20-April-2020].

[37] R. L. Sowell, K. D. Phillips, and T. R. Misener, "Hiv-infected women and motivation to add children to their families," *Journal of Family Nursing*, vol. 5, no. 3, pp. 316–331, 1999.

[38] H. Taubenböck, J. Staab, X. X. Zhu, C. Geiß, S. Dech, and M. Wurm, "Are the poor digitally left behind? indications of urban

divides based on remote sensing and twitter data," *ISPRS International Journal of Geo-Information*, vol. 7, no. 8, p. 304, 2018.

[39] A. Agrawal, W. Fu, and T. Menzies, "What is wrong with topic modeling?(and how to fix it using search-based se)," *arXiv preprint arXiv:1608.08176*, 2016.