Journal of Digital Art & Humanities

ISSN 2712-8148

Vol.2 Iss.2 **December 2021**

CONTENTS

Sentiment Evolution Analysis and Association Rule Mining for COVID-19 Tweets

Yassine Drias[1][0000-0002-8896-6170], Habiba Drias[2][0000-0001-7287-5170]

[1]University of Algiers, Algiers, Algeria
[2]USTHB – LRIA, Algiers, Algeria

https://doi.org/10.33847/2712-8148.2.2_1

Received 01.09.2021/ Revised 09.10.2021/Accepted 11.12.2021/Published 30.12.2021

Abstract. This article presents a data mining study carried out on social media users in the context of COVID-19 and offers four main contributions. The first one consists in the construction of a COVID-19 dataset composed of tweets posted by users during the first stages of the virus propagation. The second contribution offers a sample of the interactions between users on topics related to the pandemic. The third contribution is a sentiment analysis, which explores the evolution of emotions throughout time, while the fourth one is an association rule mining task. The indicators determined by statistics and the results obtained from sentiment analysis and association rule mining are eloquent. For instance, signs of an upcoming worldwide economic crisis were clearly detected at an early stage in this study. Overall results are promising and can be exploited in the prediction of the aftermath of COVID-19 and similar crisis in the future.

Keywords: COVID-19, Twitter Dataset, Tweets Analytics, Sentiment Analysis, Sentiment Evolution, Data Mining, Association Rule Mining, FP-growth.

1. Introduction

Nowadays, people are becoming more dependent on social media in their daily life. They use them to share their feelings and opinions about common subjects. The benefits of these platforms become more evident in situations such as social movements, natural disasters and pandemics. With the widespread of COVID-19 around the world, which forced the whole planet to adopt new drastic measures and behaviors, it becomes interesting to investigate the impact of this pandemic through social media.

The present work conducts a tweets analytics study on COVID-19, focusing on discovering social features and relations with the hope of achieving important insights. More precisely, we present a study that we performed by analyzing Twitter publications related to this disease, posted between the 27th February and the 25th March 2020. This is the same period during which COVID-19 was officially declared as a pandemic by the World Health Organization, which made it the hottest topics of public concern in the world at that time. We believe that our study is comprehensive as it deals with several and various aspects. For instance, we use an elegant approach to extract data from Twitter by covering a large number of hashtags. Furthermore, we use a substantial sentiment analysis repertoire containing ten different emotions. In addition to that, we perform tasks such as an analysis of tweets trends and distributions with descriptive statistics as well as mining frequent patterns and association rules.

The findings of our study could help dressing an inventory for the history of the outbreak and especially how it was apprehended by the world population. Technically, we first collect tweets related to COVID-19 and store them in a dataset that we make available for download. In a second step, we perform a preprocessing phase on the crawled data followed by a statistical analysis in order to extract knowledge and

facilitate its understanding. Thereafter, we propose a new algorithm using the inverted file lexicon-based approach to conduct a sentiment analysis task on a large number of tweets. We also study the sentiment evolution over a period of four weeks to come up with the variations related to the feelings of the considered twitter users. In addition to these contributions, we adapt the FP-Growth algorithm to make it able to efficiently extract the most frequent patterns directly form social media with the aim of grasping social features related to COVID-19. Some interesting association rules were then derived and analyzed.

This study has been significantly improved from a preliminary work published as a preprint published in [1]. In terms of the construction of the dataset, a new task consisting in bot detection was undertaken. We found out that around 4.1% of the extracted tweets were provided by software applications. All these tweets were eliminated from the dataset to avoid any kind of noisy or biased data. The bot elimination helped to enhance the results of the different tasks including the sentiment evolution analysis and the association rule mining. A more in-depth analysis of the achieved outcomes such as the prediction of important events like the economic crisis were included. Additionally, we considerably improved the presentation of the article and the visualization of the results with completely new and enhanced graphs. Finally, we reported and discussed some limitations of the study.

The paper is organized as follows. The next section discusses background and some related works. Section 3 presents the approach of the construction of the COVID-19 tweets dataset as well as the bot detection and some preprocessing tasks. Section 4 exposes three important datamining tasks applied on the dataset; an exploratory study, an analysis of sentiments evolution and an association rule mining process. The experimental results of these tasks are presented in Section 5. The main contributions of this work are highlighted in Section 6. Finally, Section 7 concludes the paper and discusses some limitations of the study.

2. Background

COVID-19 appeared in December 2019 in Wuhan China and within a couple of months, it strongly impacted the world, which led to the evocation of well-founded concerns about the future. The World Health Organization declared the outbreak a public health emergency of international concern on 30 January 2020, and a pandemic on 11 March 2020. Studies from all around the world have been performed within a short period of time as COVID-19 took its toll in almost all countries [2]. This fact translates the importance of studying this virus in order to speed up the mastery of the disease and the discovery of a remedy. On the other hand, an impressive number of articles covering COVID-19 has been widely published on the media on a daily basis, relating the disease spread, the population concerns and also helping to cope with cultural customs and urge for people isolation to limit the propagation [3].

This invisible virus is currently considered to be one of the most dangerous enemies of humanity and the metaphor of a war situation was adopted by certain scientists who advice the use of some military strategies to combat this disease [4].

Doctors, affected persons and their relatives, politicians, economists, celebrities as well as ordinary people are actively talking about the pandemic on social media, which have become an essential means of communication in our daily lives. The analysis of internet users' behavior has shown that they are mostly either seekers of information from the Web or from social media (Facebook, Twitter, microblog, etc.) [5, 6, 7]. A better understanding of this data flow would allow to make informed and focused decisions on how to meet certain goals. It can be therefore of great benefit to people and to industry, even knowing that such data may contain some biased elements.

Tweets analysis has recently witnessed an increasing number of efforts. Most of them are interested in determining features of social interactions between users [8, 6] and their behavior [5, 7].

On the other hand, there is a recent and rich literature on sentiment analysis (SA) and its applications to various fields. SA has been also investigated for social networking data and is generally used by companies for analyzing the opinion and feelings of the customers about products, services and company strategies [9, 10, 11, 12, 13, 14].

Besides, developments on data mining have yielded advanced tools such as classification, clustering and association rule mining (ARM) [15; 16; 17]. ARM have also known an impressive development for use on huge volume and complex data [18, 19, 20].

3. Crawling tweets and building the dataset

In this section, we present the different steps performed to build a dataset of tweets on COVID-19 of a high quality. We provide metadata for the dataset, which merely describes its specifications such as the content of the tweet, the terms, the hashtags, the mentions, the links, the date of publication and the author.

3.1. Extraction of raw data

We undertook a task of tweets crawling during a consecutive period of 28 days starting on the 27th of February and ending on the 25th of March 2020. The extracted tweets are related to the Coronavirus topic and highlighted by the following hashtags: #COVID-19, #COVID19, #COVID, #Coronavirus, #NCoV and #Corona. During this phase, the world has known tremendous changes affecting its whole organization. The huge number of infected and dead people provoked public panic and fear, which raised supply shortages not only in pharmaceuticals and PPE (Personal Protected Equipment) but also in food. Several countries have known a bad public health management and a rapid scalability in world economic crisis has been observed during a very short lapse of time. Rushes on public markets have been seen before certain governments decreed isolation and quarantine for inhabitants and especially for returning travelers. All these upheavals have considerably impacted the usual behavior of people and were visible in their interactions and exchanges on social media. The aim of this study is to shed light on all these aspects through data mining tasks on the extracted collection of tweets. The latter contains more than half a million tweets written in several languages and sent from over 130 countries.

3.2. Bot detection

Following the data extraction phase, we cleaned up the obtained data by removing any noisy parts. We paid particular attention to the provenance of the tweets and wanted to make sure they were generated by real persons rather than by bots. A bot is a software application that performs automated tasks over the Internet and is generally used to promote certain ideas and spread disinformation at a large scale. This represents a major concern on the credibility and the authenticity of the information provided on social media [21]. In recent years, the majority of social media platforms have dedicated efficient tools and algorithms focused on detecting bots and deleting their content. Twitter for example suspended more than 70 million accounts between May and June 2018 and reached a suspending rate of more than 1 million per day in recent months [22].

In order to evaluate the presence of bots in our dataset, we implemented a bot detection program based on the open-source tool *Tweetbotornot* [23] using R

programming language. We rewrote some parts of the code and rebuilt the package to make it work. We adopted a user-level model based on features such as the user's biography, location, number of followers, profile picture, etc. Using additional features based on the user's tweets would improve the accuracy. However, due to Twitter rate limits, such method can be very slow, and its implementation would be quite impossible with voluminous datasets like ours.

3.3. Preprocessing the tweets

A textual preprocessing phase was first held to eliminate hyperlinks, mentions and punctuation. Stop words such as articles, prepositions of time and place, conjunctions and superfluous and unnecessary words, were removed as they do not have impact on the text semantics. We then determined the set of terms to be considered in the rest of the study by applying the *Porter stemming algorithm*, which consists in reducing each word to its stem. Note that some words were misspelled, an example is the word *carona* which was tweeted by *President Trump* and reused 8013 times by other users. This kind of typos was simply eliminated.

4. Used Data Mining techniques

In this section, we describe three data mining techniques we explored for the dataset we built. First, tweets analytics were investigated in order to get preliminary insights from the considered users communication. Then a sentiment analysis was handled to discover the users' sentiments during the studied period as well as their evolution throughout the weeks. Association rule mining has also been considered in order to enrich our findings on COVID-19 insights.

4.1. Tweets analytical insight for COVID-19

As a second contribution following the dataset construction, we performed an analytical study in order to mine knowledge allowing an understanding and a mastery of the considered users insights. This analysis aims to yield data trends and distributions, descriptive statistics as well as data groupings. It also helps formulating hypotheses that can contribute in the interpretation of certain events.

4.2. Tweets sentiment evolution analysis

In addition to the previous task, a sentiment analysis was investigated aiming at capturing the tone of the tweets. We know that during the COVID-19 period, people changed their behaviors and experienced different feelings and emotions compared to those they have known before. This has considerably impacted their lifestyle, which has shifted to something completely new. The algorithm described in the next subsection was adopted and implemented to shed light on tweeters sentiments.

4.2.1. Sentiment analysis algorithm

The sentiment analysis algorithm we propose relates on a lexicon-based approach. As input, it takes the set of preprocessed words of the dataset and a sentiment lexicon. There are several known emotions lexicons that can be used for sentiment analysis. As examples we cite AFINN, Bing and NRC [24, 14]. AFINN computes a score from the range (-5,5) to each word and deducts the positive sentiment if the score is positive and the negative sentiment otherwise. Bing assigns straightly a positive or a negative emotion to each word. NRC considers ten categories of sentiments, which are *positive*, *negative*, *anger*, *anticipation*, *disgust*, *fear*, *joy*,

sadness, *surprise*, and *trust* and assigns to each word at least one of these categories. We adopt the NRC lexicon as it provides specific sentiments other than positive and negative. This lexicon contains a total of 6468 words distributed over the ten categories. The number of words in each category is shown in Table 1.

Table 1. Sentiment categories and their respective cardinalities in NRC

Category	Anger	Anticipation	Disgust	Fear	Joy	Negative	Positive	Sadness	Surprise	Trust
Cardinality	1247	839	1058	1476	689	3324	2312	1191	534	1231

The framework of the algorithm is shown in Fig. 1. After the preprocessing step, the list of words contained in the dataset is determined. Meanwhile, the NRC inverted file is built and converted to a hash table. Initially each category is assigned a counter equal to 0, which represents the number of words that are associated to that category. Then, for each word in the dataset, the algorithm calculates its hash key to access the index containing its address in the NRC lexicon structure. Next, using that address, the algorithm retrieves the categories to which the word belongs from the NRC lexicon. Each category counter is then incremented accordingly. When the process is completed, the score of each category is computed as the ratio of its corresponding count over the total number of words in the dataset.

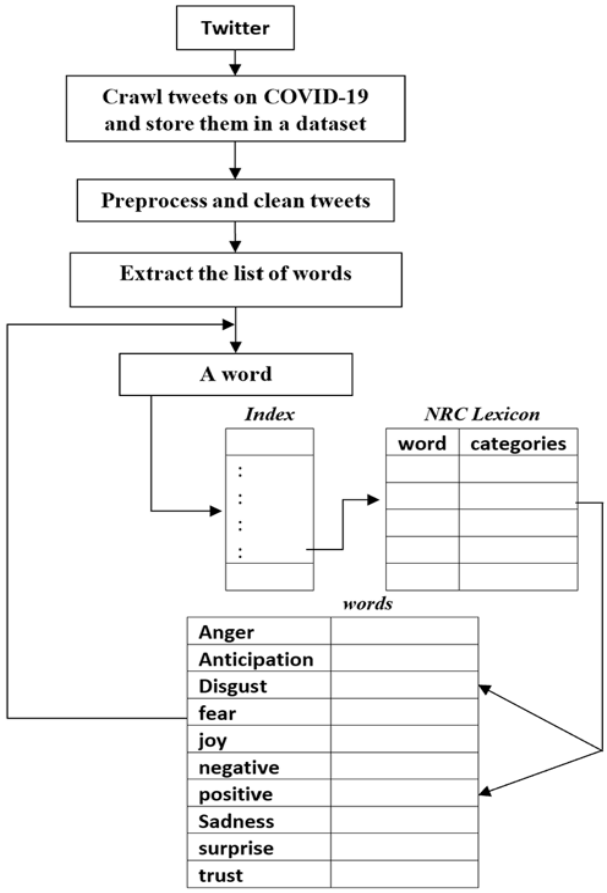

Fig. 1. Sentiment Analysis Flowchart.

4.3. Tweets Association Rule Mining

Data mining technology proposes a large spectrum of tools to extract interesting and potentially useful patterns from huge and complex volumes of data. Association Rule Mining (ARM) is one of those techniques that can be adapted to various domains such as health, trade and industry. The generation of ARM is achieved within two steps:

- Discovering Frequent Patterns for a *minimum support*.
- Filtering association rules from the extracted frequent patterns with respect to a *minimum confidence*.

The patterns correspond to the most frequent words used in the tweets and hashtags, which represent all insights on this social media for COVID-19. Since the algorithms for Frequent Pattern Mining (FPM) are computationally expensive, a lot of research has been carried out for further improving their effectiveness and efficiency. FP-Growth algorithm [19] was developed in order to cope with the main drawbacks of previous FPM algorithms, especially Apriori and ECLAT. Therefore, we use FP-Growth to discover frequent patterns and association rules in our tweets dataset.

FP-Growth on COVID-19 Tweets

The FP-Growth algorithm draws its strength from the fact it uses a sophisticated and optimal data structure called FP-tree that condenses only relevant data in a vertical format. By scanning the tree, the frequent words respecting the minimum support are determined. The support of a word is calculated as the number of its occurrences in the tweets and a minimum support is introduced for the algorithm as an input. Our adaptation of the FP-Growth algorithm for mining frequent tweets keywords is outlined in Algorithm 1.

Algorithm 1. FP-Growth on COVID-19 tweets

Algorithm: FP-Growth-Tweets

Input: *D* the tweets dataset;
minsup, the minimum support count threshold.

Output: subsets of frequent keywords of length 1, 2, ...

Begin

1. Scan the dataset *D* once and find frequent words (single word pattern);
2. Sort frequent words in frequency descending order to constitute the *f-list*;
3. Construct FP-tree by scanning once more *D*;
4. **For each** path of the tree do
    a. *l* := 2 ;
    b. output all sub-paths (non-necessary consecutive) of length equal to *l* with *minsup*;
    c. increment *l*;
5. **End for**

End

The advantage of the *f-list* is to eliminate at the beginning the words that do not respect the minimum support. Then the construction of the *FP-tree* data structure considers only the *f-list* words, which makes it reduced to only relevant information. It starts with an empty node, then it creates nodes taken sequentially from the *f-list* and links them if they belong to the same tweet, while scanning the collection of tweets and incrementing their frequency.

4.3.2. Association Rule Mining for COVID-19 Tweets

Association rules are derived from the frequent patterns calculated by the adapted FP-Growth algorithm. Suppose w_1, w_2 and w_3 are words appearing frequently in the collection of tweets with a minimum support equal to *minsup*. Then, from this subset of words (called itemset in the traditional frequent patterns mining algorithms), we can generate four association rules that are:

$\rightarrow w_1, w_2, w_3$ (*minsup*, *minconf*)
$w_1 \rightarrow w_2, w_3$ (*minsup*, *minconf*)
$w_1, w_2 \rightarrow w_3$ (*minsup*, *minconf*)
$w_1, w_2, w_3 \rightarrow$ (*minsup*, *minconf*)

The first rule means that in all tweets, the three words exist with the indicated minimum support. In the second one, the rule is interpreted as: whenever w_1 exists in a tweet, w_2 and w_3 appears in the same tweet. The other rules follow the same principle. Note that we need to specify another measure for the rule besides the support, which is the confidence. The confidence of a rule is defined to be the probability whenever an antecedent of the rule is in a tweet, the consequent is also in the same tweet. It is calculated as follows:

$$Confidence(rule) = \frac{\big(support(antecedent, consequent)\big)}{\big(support(antecedent)\big)}$$

5. Results

All the experiments showcased in this section were held on a laptop running Windows 10 with an Intel Core i5-7300U CPU at 2.60 GHz and 8GB of RAM. *Java* and *R* programming languages were used to implement the different techniques described in the previous section.

5.1. Dataset Construction

We built a dataset composed of 653 996 tweets posted by 390 458 users, using *NodeXL* and *Java*. The dataset is available online and can be downloaded from the *Zenodo* platform [25]. The specifications of the dataset as well as the results of the bot elimination are shown in the following subsections.

5.1.1. Descriptive dataset and metadata

The created dataset includes the features shown in Fig. 2 with an example for each one of them. The Attributes are all of string type except for the date which follows the *month/day/year* format. The dataset contains 653 996 tweets with no missing values. Table 2 gives an overview of a portion of the tweets included in the dataset.

Author	jmcrhealthdept
Recipient	tndeptofhealth
Hashtags	coronavirus
Language	en
Relationship	Mentions
Location	United States
Date	2/27/2020
Source	Twitter for iPhone

Fig. 2. Example of extracted features.

Table 2. A sample of tweets extracted from the dataset

N°	Tweet
1	Concerned about #coronavirus? We are closely monitoring the situation along with @TNDeptofHealth and @CDCgov. For the general American public, the immediate health risk from COVID-19 is believed to be low at this time. Find current and accurate info at https://t.co/aQqARorfCM. https://t.co/Hjc3OnMSLY
2	Australia now has more than 1000 confirmed cases of coronavirus .. case numbers rising steeply over last 2 days. ABC is compiling the latest statistics from states and territories .. This graph is up to 7.30 tonight https://t.co/Zz8B3oVSHM https://t.co/Wb7fss0Ivi
3	Es imposible no estresarse viendo las líneas de carros afuera de los supermercados, la siguiente semana no tendrá clases y el resto del semestre lo haré virtual como medidas preventivas #COVID2019.
4	#PlaceCoronavirusDansUnFilm Ququ'on a fait au bon Caronavirus ?

Fig. 3 shows the WordCloud of the raw dataset before proceeding to the text preprocessing and the bot detection phases. Only an English stop words list was applied to the text.

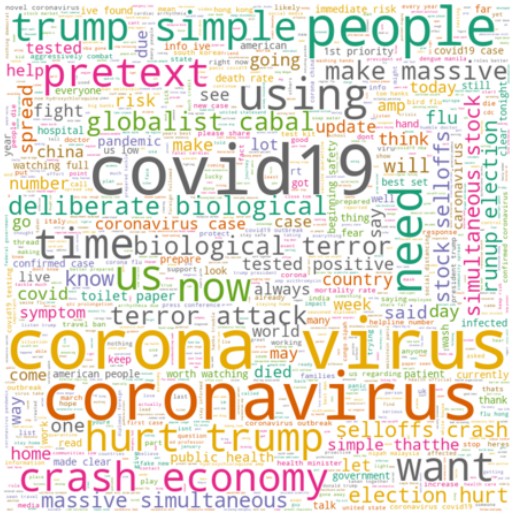

Fig. 3. TweetsWorldCloud

5.1.2. Bot detection

Fig. 4 shows the results of the bot detection task. We found out that among a total of 390 458 users present in our dataset, 26 874 (6.88%) were determined as bots accounts while 363 584 were tagged as real users. This means that around 4.1% of the tweets were provided by software applications. All these tweets were eliminated from the dataset to avoid any kind of noisy or biased data.

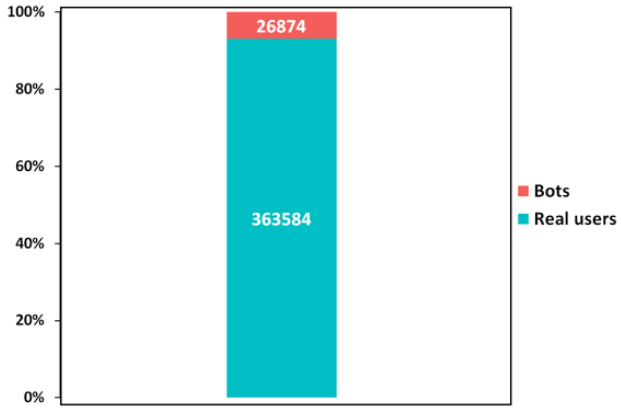

Fig. 4. Number of bots versus Number of real users in the dataset

5.2. Tweets Analytics

The experiments of the developed program for analytics were performed on the tweets after the text preprocessing and the bot elimination phases like explained in Section 3. The different outcomes are exhibited in the following subsections.

5.2.1. Top hashtags

The hashtags #COVID-19, #COVID19, #COVID, #Coronavirus, #NCoV and #Corona that served for the construction of the dataset are not considered since each tweet contains at least one of them. Fig. 5 shows the 25 most used hashtags, which report some events and situations people are dealing with, such as: pandemic, update, outbreak, stay at home, curfew, confinement and quarantine. On the other hand, countries and regions such as China, Iran, Wuhan, India and Italy, that were the most affected by the virus during that period are cited. Note that the wide virus spread wave in India was just starting at the end of the studied period, we assume it is evoked among these countries because of its large population.

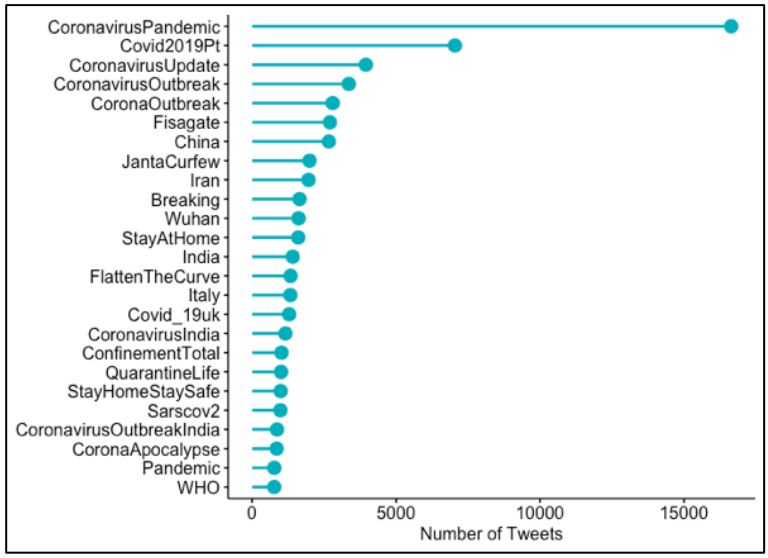

Figure 5. Top 25 most used hashtags

5.2.2. Top keywords

The 25 top words are shown in Fig. 6 with their respective frequencies. These terms are eloquent as they are similar to the words expressed by the majority of people in real life. We observe four sets of words with approximately the same frequency, cited in a decreasing order of their occurrence as follows:

- virus, people
- Flu, Cases, Trump, test
- Spread, Health
- Work, outbreak, death, update, home, China, hands, report, world, country, call, confirmed, infected, week, risk, panic, pandemic

We can observe that most these words belong to the new vocabulary utilized by people and the media in real life during this period.

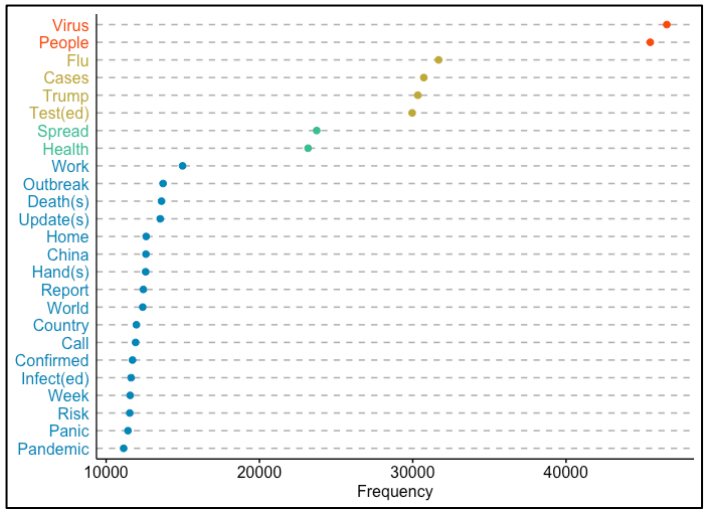

Fig. 6. Top 25most used terms.

5.2.3. Number of tweets per country

The first ascertainment about the origin of the tweets is that a significant number of users do not give access to their geolocation information and therefore do not indicate their country. That being the case, the results hereafter are valid only for tweets whose authors provide geolocation information. Fig. 7 shows the worldwide tweets distribution based on the number of tweets in each country. The color palette indicates the density of the tweets, starting from blue for countries that count a low number of tweets and going up to red for countries containing a large number. Note that the count of tweets for the countries colored in gray is equal to 0. We observe that the majority of the published tweets emanate from the USA. The second remark is that users belong to several countries from all the continents. We also notice that the number of tweets is for some countries as important as their respective demography. For instance, India is ranked in 3rd place while it has the second largest population in the world. The same observation can be done for Nigeria and South Africa that count respectively the largest populations in Africa. Also, the countries the most affected by the virus such as Spain, France and Italy show a high density of tweets.

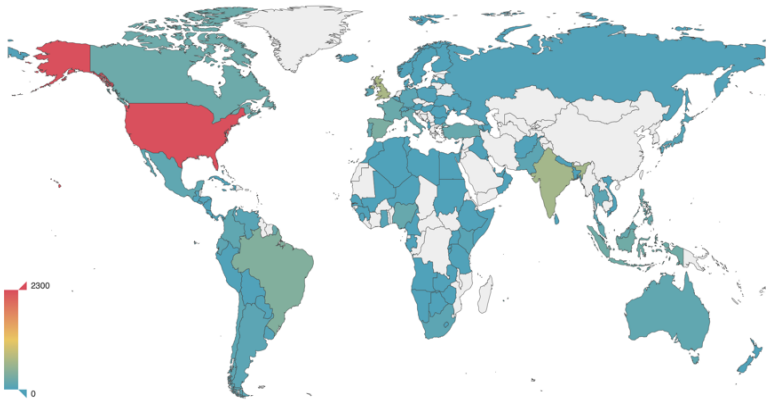

Fig. 7. Tweets distribution over countries.

5.2.4. Tweets' languages

The languages of tweets were also explored, Fig. 8 gives an insight about this feature. We note that English prevails over the other languages. This can be explained by the fact that according to the previous statistics, the three top countries in terms of the number of tweets are the USA, United Kingdom and India, where the communication language is English. Also note that English is scientifically and technologically universal and is written by a large population in other countries. Spanish is ranked in the second position in this graph because it is spoken in several countries in Latin America like Colombia and Mexico that appear in the previous figure. Also, all the languages spoken in highly affected countries during this period exist in this graph.

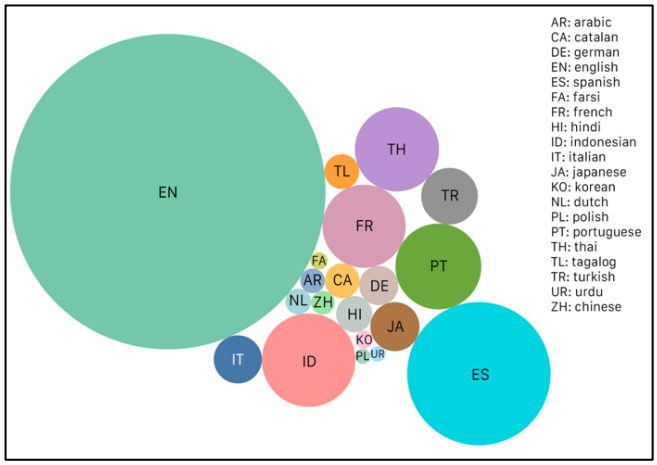

Fig. 8. Most used languages

5.2.5. Correlations

We can think about several possible correlations between the parameters identified previously. We investigated the correlation between the number of new infected cases and the number of tweets published each day. The intuition is that the evolution of the number of infected cases urges people to communicate more about the virus via social media. Figure 9 depicts the case of the UK, where the *x* axis represents the number of tweets per day and the *y* axis the number of new cases tested positive to the coronavirus per day. We observe a positive trend for the correlation confirmed by a Pearson coefficient equal to 0.55.

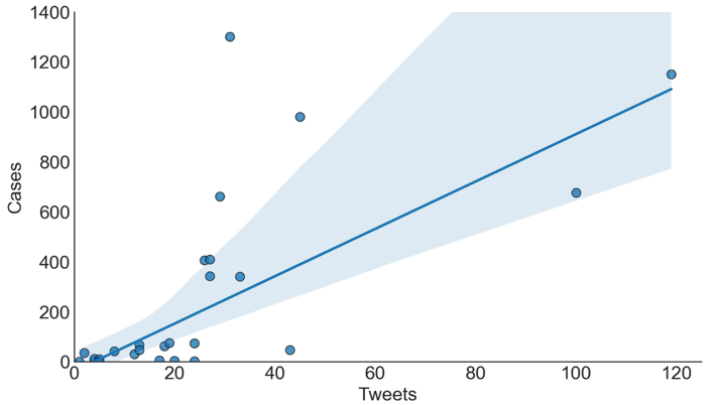

Fig. 9. Correlation between the number of new cases and the number of tweets in the UK.

5.3. Sentiment analysis

5.3.1. Tone of the tweets

Fig. 10 shows the results yielded by the sentiment analysis algorithm. We observe that the *negative* sentiment has a score slightly greater than that of the *positive*. *Fear* has an important score whereas *joy* and *disgust* have the lowest. Despite this difficult period and sad situation, the users seem to have trust in gaining the battle against the virus.

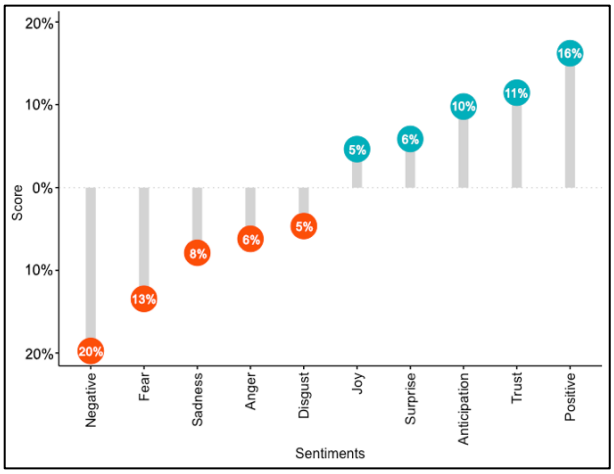

Fig. 10. The tone of tweets captured by the sentiment analysis.

5.3.2. Sentiment evolution

To track the sentiments evolution during the studied period, we divided the 28 days into four intervals each constituting a week. The first week starts on the 27th of February 2020, the second on the 5th of March 2020, the third on the 12th of March 2020 and the fourth on the 19th of March 2020. We used these dates as point-in-time snapshots in order to capture the score of each sentiment and then compare its variation over time and analyze its evolution.

Fig. 11 displays the sentiment density variation over the sentiment rate during the 4 weeks period. We can distinguish three intervals of scores:

The first interval [0.13, 0.27] incorporates the highest sentiments rates and includes two sentiments "positive" and "negative". This reflects that both sentiments are the most prevalent among the rest during the studied period. The low density of these sentiments is due to fact that they are distributed over a relatively large interval, which results in a prominent evolution over time.

The second interval [0.08, 0.15] includes two other sentiments "fear" and "trust" with medium density.

The third interval [0.02, 0.10] contains the six remaining sentiments, namely "anger", "anticipation", "disgust", "joy", "sadness" and "surprise". These sentiments have a high density because of the reduced range of values they occupy. This means that their evolution over time is less prominent compared to the sentiments of the first and second interval.

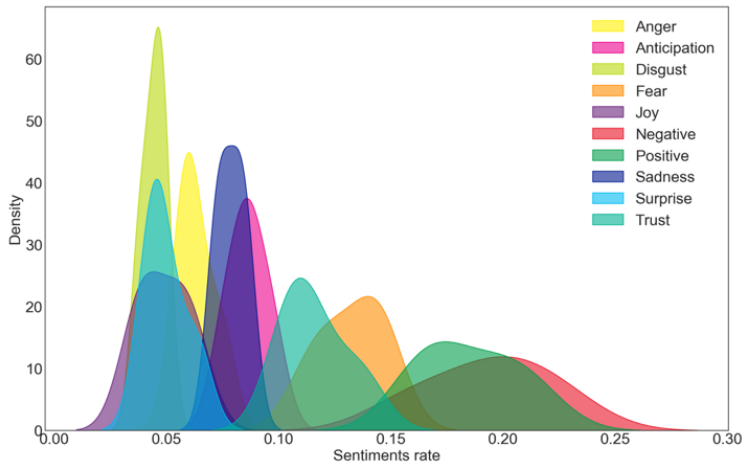

Fig. 11. Sentiment density variation.

Fig. 12 details the evolution of the sentiments during the four weeks period. We notice that the *negative* sentiment was high in the beginning of the period then it knew a little increase in the second week and since then it has considerably decreased. On the contrary, the *positive* sentiment follows the opposite behavior. *Fear* has almost the same evolution as the *negative* sentiment but with less magnitude whereas *trust* behaves as the *positive* sentiment with a lower rate. On the other hand, *Joy* is at a constant slow increase throughout the studied period and has one of the lowest scores. *Anticipation* has an overall constant curve whereas *anger*, *disgust*, *sadness* and *surprise* vary with very low rate from week to week.

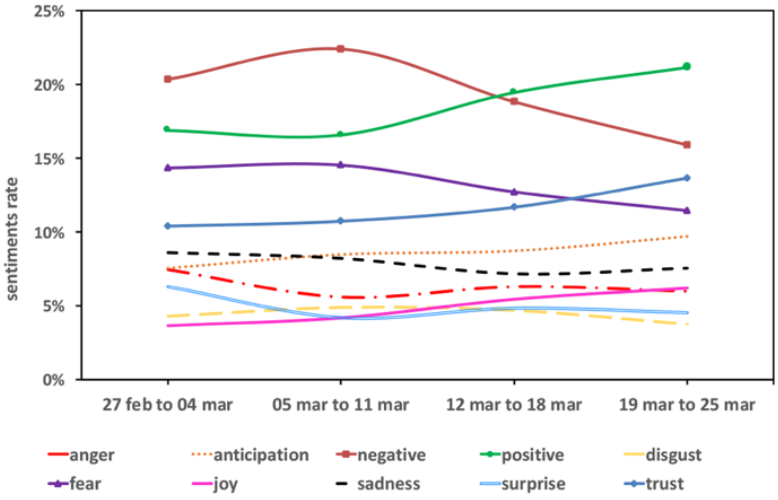

Fig. 12. Sentiment evolution over time

As tweets can be distributed by country as shown in Fig. 7, the sentiment analysis can be carried out for each country and each region by merging the tweets of the countries of the region in one dataset. Examples were performed for the USA and Canada separately and for the region of North America including both countries.

India has also undergone such treatment. Note that the sentiments evolution can also be drawn for countries and regions.

5.4. ARM experimental results

The adapted FP-growth algorithm was implemented in Java. The most frequent words in the tweets published in English are reported in Table 3 in descending order of their minimum support. We can confirm that these words constitute an important part of the daily vocabulary used in real life by people during the pandemic.

Table 3. The most frequent patterns with their minimum support

Word	MinSup
people	26214
virus	22830
cases	19118
flu	18363
tests	17163
spread	14650
health	13574
trump	11149
time	10104
work	9627
covid	9511
make	8731
outbreak	8629
pandemic	8319
day	8289
deaths	8265
china	8103
prepare	8030
update	7954

Similarly, the pairs of the most frequent words appear in Table 4, which shows examples of association rules ordered with respect to their minimum support on a total number of patterns equal to 19560. These associations are impressively similar to the ones people daily employ during the pandemic. For instance, the rule *wash* → *hand* has the highest confidence, which corresponds exactly to what we hear many times a day through the media.

On the other hand, some rules predicted the impact of COVID-19 and its imposed restrictions such as the quarantine measures. An example is the rule "*crash* → *economy"*, which has a high minimum confidence score. In fact, following the period of the study, several countries have known economic difficulties and crisis and some of them have publicly discussed the possibility of proceeding to an early lift of the lockdown in order to reduce the negative economic impact of the pandemic. The most significant event related to this situation was the holding of an economic summit that brought together the EU leaders for 4 days of talks and which was crowned on the 21st of July 2020 with an agreement on a post-coronavirus economic recovery plan.

Table 4. Examples of generated association rules with 2 frequent patterns

Association Rules	MinSup	MinConf
flu → virus	16068	0.5
found → virus	4712	0.68
confirmed → cases	9152	0.78
positive → tested	7689	0.72
wash → hand	7024	0.94
attack → biological	7436	0.83
low → preparation	3913	0.61

Association Rules	MinSup	MinConf
crash → economy	7486	0.92
massive → hurt	7416	0.85
rate → flu	4337	0.51

Table 5 and Table 6 show the association rules identified by the algorithm for respectively the 3 and the 4 most frequent patterns with a low support but a significant confidence. These rules are also expressive relatively to what people were living during these days. An eloquent example that confirms people panic, which was also reported by the media is the rule "*toilet → panic paper"*. Rules predicting an emanant economic crises such as "2020 → economy crash" and "economic → stock crash" can also be found in Table 5.

Table 5. Examples of generated association rules with 3 frequent patterns

Association Rules	MinSup	MinConf
low → prepare risk	3905	0.61
selloff → economy simultaneously	7416	0.99
terror → stock hurt	7416	0.98
economic → stock crash	7427	0.75
bird → virus flu	1918	0.83
2020 → economy crash	7420	0.5
priority → prepare american	2417	0.69
aggressive → risk people	2418	0.84
soap → home hand	1940	0.59

Table 6. Examples of generated association rules with 4 frequent patterns

Association Rules	MinSup	MinConf
low → risk made american	2414	0.41
minister → prepare year india	1480	0.43
season → health doctor role	1480	0.41
men → season journey role	1480	0.78
secretary → update confirmed units	1463	1
communicating → update confirmed statesamerican	1463	0.29
communicating → updates confirmed families	1463	0.29
update → cases units low	1463	0.18

6. Discussion

Within the framework of this study, we were interested in investigating people communication via social media during the starting period of the rapid spread of COVID-19 all around the world. Twitter was selected to undertake the study as it is one of the most used information sharing networks. Also, it offers different APIs for crawling tweets. As a first contribution, a dataset of 653 996 tweets was created and preprocessed in order to highlight useful insights.

The second contribution is an exploratory study on the collection of tweets, which was carried out to yield useful information and descriptive statistics. An important number of hashtags and topics were shared by users on Twitter and the most popular ones are exhibited in Fig. 5 and Fig. 6 respectively. Features such as the number of tweets posted per country (Fig. 7) along with the most used languages (Fig. 8) were extracted. An example of a correlation between the number of tweets posted per day and the number of infected cases per day was determined for the UK (Fig. 9).

A sentiment analysis followed, and the results showcased the tone of the tweets relatively to ten emotions, which are *anger*, *anticipation*, *disgust*, *fear*, *joy*, *negative*, *positive*, *sadness*, *surprise* and *trust*. The rate of words related to the negative

sentiment was close to 20% whereas that of the positive sentiment was a little bit greater than 16%. The rates corresponding to the other emotions are presented in Fig. 10. The evolution of the sentiments over a period of 4 weeks is illustrated in Fig. 11 and Fig. 12.

Data Mining technologies were afterwards exploited, and the FP-Growth algorithm was especially adapted to the tweets dataset in order to discover the most frequent patterns in an effective and efficient way. The derived association rules shed light on our understanding related to people interaction on COVID-19. The study highlights four major topics on COVID-19 that are sanitary measures, economy crisis, the origin and current state of the disease as well as social behaviors. An example of each one of them is respectively "*soap → home hand"*, "*bird → virus flu"*, "*2020 → crash economy"* and "*toilet → panic paper"*. For more details, Tables 4 to 6 show the main generated association rules.

7. Conclusion

This work allowed us to find out the state of mind of Twitter users during the first stage of the widespread of COVID-19. Several tasks were carried out to lead to an in-depth study. The beginning was the construction of a large dataset containing tweets posted between the 27th of February 2020 and the 25th of March 2020. Tweets coming from bots were eliminated in order to restrain the analysis we conducted to only real users, as our prime concern was to identify insights expressed by Twitter users on COVID-19. Then three main data mining processes were investigated on descriptive statistics, tweets sentiments analysis and frequent pattern and association rule mining respectively. All the findings of the different analyses are exhibited in the Results section using graphical formats and comments.

First, let us observe that the discussions and exchanges on COVID-19 were massive between the Twitter users. This proves that people's daily life was impacted by the pandemic phenomenon. Several aggregates and insights were determined thanks to the tweets analytics process.

The Twitter users' sentiments were also revealed, we know especially that in the beginning of the studied period, this population was afraid and that panic gradually faded until it was overcome by the end of the period.

Frequent used patterns such as "wash hand" were also shared among the people in real life. The confidence of such association was relatively high, which translates a good representativeness of people through the dataset. Also, predictions such as economic crisis were mined by the rules "2020 → crash economy" and "crash 2020 → economy". In fact, the predictions turned out to be correct as many countries have known economic difficulties and crises following the period of our study. These achieved findings can contribute in building the history of the COVID-19 pandemic.

At last, it is worth noting that this study has some limitations. First, the number of tweets extracted to undertake the study was limited to the machine capacity. Increasing the size of the dataset can improve the outcomes quality. Second the suppression of the tweets coming from organizations [26] was not treated due to the important size of our dataset (390,458 users), Twitter rate limits and time constraint. Addressing the individuals/organizations separation aspect can be very slow and its implementation would require a long period of time. Third, only tweets posted in English were considered in the sentiment analysis. A Multi-language analysis would potentially disclose more relevant insights. Fourth, certain tweets did not contain geolocation information, which could have affected the result of the distribution of the tweets over countries.

References

1. Y. Drias, and H. Drias, "Mining Twitter Data on COVID-19 for Sentiment analysis and frequent patterns Discovery". medRxiv preprint, **(2020)**/ https://doi.org/10.1101/2020.05.08.20090464
2. Z. Wu and J.M. McGoogan, "Characteristics of and important lessons from the coronavirus disease (COVID-19) outbreak in China: Summary of a Report of 72314 cases from the Chinese center for disease control and prevention". JAMA. **(2020)**. https://doi.org/10.1001/jama.2020.2648
3. J. Li and X. Guo, "Global Deployment Mappings and Challenges of Contact-tracing Apps for COVID-19", SSRN Electronic Journal, **(2020)**. https://doi.org/10.2139/ssrn.3609516
4. D.N. Maxwell, T.M. Perl and J.B. Cutrell, "The art of war in the era of coronavirus disease (COVID-19)", Clinical Infectious Diseases, ciaa229, **(2019)**. https://doi.org/10.1093/cid/ciaa229
5. D.R. Bild, Y. Liu, R.P. Dick and Z. Morley Mao, "Aggregate characterization of user behavior in Twitter and analysis of the retweet graph", ACM Transactions on Internet Technology (TOIT), vol. 15, no 4, **(2015)**. https://doi.org/10.1145/2700060
6. S.S. Ercetin and N.B. Neyisci, "Social network analysis: A brief introduction to the theory", In: Ercetin S. (eds) Chaos, Complexity and Leadership, Springer Proceedings in Complexity, Springer, Cham, 167-171, **(2014)**. https://doi.org/10.1007/978-3-319-18693-1_16
7. Q. Yan, L. Wu and L. Zheng, "Social network based microblog user behavior analysis. Physica A: Statistical mechanics and its applications", 7(392), 1712-1723, **(2013)**. https://doi.org/10.1016/j.physa.2012.12.008
8. T.D. Baruah, "Effectiveness of social media as a tool of communication and its potential for technology enabled connections: A micro-level study". International Journal of Scientific and Research Publications, 2(5), pp: 1-10, **(2012)**.
9. F. A. Pozzi, E. Fersini, E. Messina and B. Liu, B, "Sentiment analysis in social media". Morgan Kaufmann, **(2016)**.
10. K.S. Houtan, T. Gagne, C.N. Jenkins and L. Joppa, "Sentiment analysis of conservation studies captures successes of species reintroductions". Patterns 1, 100005, **(2020)**. https://doi.org/10.1016/j.patter.2020.100005
11. M. Thelwall, K. Buckley and G. Paltoglou, "Sentiment strength detection for the social web". JASIST, 63(1); pp:163-173, **(2012)**. https://doi.org/10.1002/asi.21662
12. B. Liu, "Sentiment analysis and subjectivity". Handbook of Natural Language Processing, 2nd edition, **(2010)**.
13. X. Guo and J. Li, "A Novel Twitter Sentiment Analysis Model with Baseline Correlation for Financial Market Prediction with Improved Efficiency". International Conference on Social Networks Analysis, Management and Security, **(2019)**. https://doi.org/10.1109/SNAMS.2019.8931720
14. W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, 5(4), Elsevier, **(2014)**. https://doi.org/10.1016/j.asej.2014.04.011
15. J. Han, J. Pei and M. Kamber, "Data mining: concepts and techniques". Elsevier, **(2011)**. https://doi.org/10.1016/C2009-0-61819-5
16. X. Wu, X. Zhu and G. Wu, "Data mining with big data", IEEE transactions on knowledge and data engineering, 1(26); pp:97-107, **(2013)**. https://doi.org/10.1109/TKDE.2013.109
17. K. Heraguemi, N. Kamel and H. Drias, "Association Rule Mining Based on Bat Algorithm", Bio-Inspired Computing - Theories and Applications, Springer, **(2014)**. https://doi.org/10.1007/978-3-662-45049-9_29
18. C.C. Aggarwal, A.B. Mansurul and A.H. Mohammad, "Frequent pattern mining algorithms: A survey". Springer, Cham; pp:19-64, **(2014)**. https://doi.org/10.1007/978-3-319-07821-2_2
19. P. Fournier-Viger, J.C.W. Lin, B. Vo, T.T. Chi, J. Zhang and H.B. Le, "A Survey of itemset mining", WIREs data mining and knowledge discovery, Wiley, **(2017)**. https://doi.org/10.1002/widm.1207
20. H. Drias, C. Hireche and A. Douib, "Datamining techniques and swarm intelligence for problem solving: Application to SAT". World Congress on Nature and Biologically Inspired Computing, NaBIC, **(2013)**. https://doi.org/10.1109/NaBIC.2013.6617862
21. Y. Drias and G. Pasi, "Credible Information Foraging on Social Media", Trends and Innovations in Information Systems and Technologies, Advances in Intelligent Systems and Computing, vol 1159 Springer, **(2020)**. https://doi.org/10.1007/978-3-030-45688-7_43

22. C. Timberg and E. Dwoskin, "Twitter is sweeping out fake accounts like never before, putting user growth at risk", The Washington Post, July 6, 2018, **(2018)**. https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/
23. M. Kearney, "Tweetbotornot: Detecting Twitter bots". web app: https://mikewk.shinyapps.io/botornot/,(**2018)**.
https://doi.org/10.13140/RG.2.2.10732.82562
24. Neviarouskaya, H. Prendinger and M. Ishizuka, "Sentiful: A lexicon for sentiment analysis", IEEE Transactions on Affective Computing, 2; pp:22-36, **(2011)**. https://doi.org/10.1109/T-AFFC.2011.1
25. Y. Drias and H. Drias, "COVID-19 Tweets: A dataset contaning more than 600k tweets on the novel Coronavirus (Version 1.0) [Data set]", Zenodo, **(2020)**. http://doi.org/10.5281/zenodo.4024177
26. Z. Wood-Doughty, P. Mahajan, M. Dredze, and J. Hopkins, "Classifying Individuals versus Organizations on Twitter", Proceedings of the Second Workshop on Computational Modeling of People Opinions, Personality, and Emotions in Social Media, pages 56-61 New Orleans, Louisiana, **(2018)**. http://doi.org/10.18653/v1/W18-1108

Aims and Objectives

Published online by Institute of Certified Specialists twice a year, **Journal of Digital Art & Humanities (JDAH)** is an international peer-reviewed journal which **aims** at the latest ideas, innovations, trends, experiences and concerns in the field of the digital arts & humanities. JDAH bridges humanitarian, artistic, and scientific disciplines, allowing author(s) to express the views on the subjects studied using modern digital/information technology. It is a nexus for information exchange among academia and industry addressing theory, criticism, and practice. The effective dissemination of original ideas/results generated by the human brain and presented/reflected in articles created using modern information/digital technology is **the main objective of JDAH**.

Topics to be discussed in this journal include the following: Record Review of Social Media; Script of Digital Pedagogical Technology; Digital Pedagogical Technology; Excerpts from Digital Education; The Critical Thinking Initiative; Digital Artwork.

All articles are in open access distributed under Publisher decision.

Publisher

Institute of Certified Specialists (ICS)
95a Lunacharskogo str., Perm, Russian Federation

Journal URL: https://ics.events/journal-of-digital-art-humanities/

Email: conf@ics.events

The picture on JDAH cover was painted by Antonio Donizeti da Cruz.