

PROPAGANDA IN DIGITAL COMMUNICATION, TWITTER PREDICTIVE
MODULE BASED ON USER

A Thesis

presented to

the Faculty of Natural and Applied Sciences

at Notre Dame University-Louaize

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

MIKEL SFEIR

JUNE 2021

Propaganda in digital communication, twitter predictive module based on user behavior.

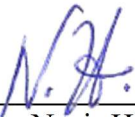
By

Mikel Sfeir

Approved by:



Dr. Hicham Hage: Associate Professor of Computer Science
Thesis Advisor.



Dr. Nazir Hawi: Associate Professor of Computer Science
Member of Committee.

June 4, 2021

Date of Thesis Defense

Abstract

“In the truth there is no news and in the news there is no truth”, a soviet-era political joke, still applies to the present times. This study will highlight all the communication forms and how they are used nowadays as a tool for propaganda especially in the digital world. When website traffic, clicks, and reach are at stake, information disorder in the form of misleading/fake information would be used as a tool to enhance data metrics across the digital world.

This study tackles information disorder during all its phases: creation, production, communication, and the user experience through the process.

First, the disruptive user behavior is modeled, and as a classifier, that categorizes Twitter users based on their behaviors, is built. Second, the propaganda detection module is built, which given a trending keywords on Twitter, determines whether or not it is part of a propaganda campaign. This approach takes a hybrid form since the analysis is based on the combination of all four phases of information disorder.”

Keywords: Propaganda, digital communication, information disorder, user behavior

Table of Contents

Abstract	iii
Table of Contents	iv
List of Figures	vi
Acknowledgments	viii

Chapter 1: Introduction and Problem Definition 1

1.1 Approach and Main Results	2
1.2 Thesis Organization.....	3

Chapter 2: Background and Motivation..... 4

2.1 Propaganda in digital communication.....	4
2.1.1 Propaganda	4
2.1.2 An evolution of information disorder.....	5
2.1.3 Forms in the digital world	5
2.1.4 Knowledge, power and control.....	6
2.2 Information disorder.....	7
2.2.1 Types	7
2.2.2 Phases	8
2.2.3 Elements	10
2.3 User Experience	13
2.3.1 Available user data	13
2.3.2 Cognitive biases.....	15
2.3.3 Visual Cues.....	16
2.3.4 Social behavior	17
2.3.5 User behavior in Lebanon – Survey	18
2.4 Social media used for information disorder	21
2.4.1 Forms.....	22
2.4.2 Applications.....	22
2.4.3 User Targeting.....	22

Chapter 3: State of the art of deception detection in social media..... 24

3.1 Creation (Agent).....	25
---------------------------	----

3.2 Production (Message).....	26
3.3 Communication (Platforms).....	27
3.4 User Experience (Receiver)	27

Chapter 4: Twitter propaganda predictive module based on user behavior30

4.1 Keywords	32
4.1.1 Key words extraction.....	32
4.1.2 Tweets retrieval	33
4.1.3 Users extraction	34
4.1.4 Keyword test.....	35
4.2 User categorization.....	35
4.2.1 User activity and background.....	36
4.2.2 Analytical layer	38
4.3 Propaganda detection	48
4.4 Testing and validation	51
4.4.1 Test sample.....	51
4.4.2 Module Adaptation.....	52
4.5 Finding and discussion	54

Chapter 5: Conclusion and future works58

5.1 Main Contributions and results of the Thesis.....	59
5.2 Possible Extensions and Future Work.....	60

Bibliography	61
Appendix A: Questionnaire	64

List of Figures

Figure 1: Sample of ads that were targeting US voters based on their classification.....	7
Figure 2: Information Disorder.....	8
Figure 3: Characteristics for each element of an example of information disorder	13
Figure 4: Survey respondent's demographics.....	19
Figure 5: Offline and Online platforms usage	20
Figure 6: Participants answers analysis between checking and sharing of news	21
Figure 7: Information Disorder Main Phases	24
Figure 8: Twitter infographic.....	29
Figure 9: Module general flow chart	31
Figure 10: Keyword phase flow chart.....	32
Figure 11: Trending keywords from twitter sample	33
Figure 12: Sample of retrieved top users	35
Figure 13: Categorization phase flow chart.....	36
Figure 14: Twitter get user background and statuses processes	37
Figure 15: Data base join.....	39
Figure 16: Activity count per day for users on February 2020.....	40
Figure 17: User Activity April to July 2020 #الدرون_بالدرون.....	40
Figure 18: Correlation Analysis #الدرون_بالدرون activity from April to July 2020	41
Figure 19: Sample of the extracted variables via QlikView.....	43
Figure 20: Variable results analysis.....	46
Figure 21: Categories conditions for the variables	46
Figure 22: Leader variables grading results after the application of the conditions.....	47
Figure 23: Grading of categories and the classification process	47
Figure 24: Final phase, propaganda detection	49
Figure 25: Normal campaign categorized users keyword #fnflebanon	49
Figure 26: Users retrieval based on the keyword #مقبره_الميركافا.....	50
Figure 27: User categorization on the keyword Feyrouz.....	52
Figure 28: Sample keywords with predictions	52

Figure 29: Adaptation table for France and UAE.....	53
Figure 30: Questionnaire answers of Knowledge/ Twitter detected fake.....	55
Figure 31: User Categorization.....	56
Figure 32: Keywords Summary.....	56

Acknowledgments

I would like to start by thanking, the faculty of computer sciences at NDU and all the professors whom I took courses with, especially my advisor, a true and honest person to look up to, Dr. Hicham Hage. Achieving this thesis would not have been possible without his guidance and support throughout the whole research and development process, especially in the hard times we're living in.

I would like to thank my friend Ramzi Awad who with our never-ending discussions around the topic have helped to develop the political aspect of the research that happened to coincide with his research interests.

Also, I would like to thank Dr. Sarah Abdallah a close friend who supported me in my research and helped me achieve my goals in the technical aspects of the research.

Finally, I would like to extend my sincere thanks to Jean Louis Cardahi, who pushed me back to university and supported my growth professionally and academically without any restraints or conditions.

Chapter 1: Introduction and Problem Definition

“In Pravda there is no news, and in Izvestiya there is no truth” [22]. This is a well-known political joke from the soviet-era in which Pravda means truth and Izvestiya means news. In other words, not all the news is true because truth does not attract readers, nor does it interest people. Information disorder and propaganda has been present since the roman era [7], when truth was deflected by the news to make people change their beliefs and actions.

Indeed, we live in a world that is boosted by the digital era which is growing exponentially which only means that internet users are facing an exponential growth of information disorder as well. Users are being targeted with information to deceive and distort their beliefs and thoughts. The danger of information disorder is that the communicated information might be misleading or incorrect. Moreover, digital propaganda depends on the technological advancement of the platforms it is using, allowing it to enhance its ability to target users. Campaigns are organized, computerized, and automated to reach a certain result in the fastest possible way. The difficulty in attempting to identify and control information disorder is that it comes in different types, has many phases and includes different elements [13].

With the technological advancements, campaigning is easier and faster and could be misleading with few or no legitimacy in its content. As a result, propaganda based on information disorder reaches the user, who is being targeted for his/her social behaviors, cognitive biases, and digital footprints, in many forms and different content. User data is retrieved from the platform they are using.

Our world has recently faced many forms of propaganda, the most notorious of which was managed by the company Cambridge Analytica [11], who handled the United States presidential election campaign that led to the election of Donald Trump [15] and that which supported the Brexit law [4].

Propaganda is being applied on the digital platforms through organized and computerized campaigning of information disorder communications. Bots play the most efficient role in the computational forms of information disorder. Moreover, users are targeted based on their backgrounds, beliefs, and any shared personal content that could be available online [1].

The purpose of the research is to establish a propaganda predictive module, that should detect and flag propaganda campaigns based on the behaviors of the users related to those specific campaigns.

1.1 Approach and Main Results

For many researchers, propaganda prediction was tackled from one angle which means that each focused on one of the four phases of information disorder. The approaches are discussed and classified based on the phases of information disorder which researchers worked on for their prediction module.

The research includes a questionnaire targeting information disorder in Lebanon and how people receive and handle this information. The questionnaire helps to understand the platforms that share such information, and its target is to understand whether or not users are aware of propaganda.

The developed model delves into the analysis of information disorder in its different phases: it starts with the communication phase by grabbing the trending keywords from Twitter then moves to the creation phase where it analyzes the agents based on their content and behaviors. As for the production phase, the code developed for the module returns if the keyword is inorganically pushed, hence propaganda. It classifies the users into four main categories: normal users, leaders, media agents, and propaganda agents. Agents that are engaging the most in a certain keyword are directly identified, and the keyword is classified as propaganda.

As a result, throughout the research, networks of users are identified. In some of those circles, agents are very well organized: their correlation average is 0,98, which proves the organizational and inorganic approach of the campaign. On the other hand, around 50% of

the users that are analyzed are propaganda agents, and their behavior is boosted or correlated. Mainly, out of all the keywords analyzed, more than 90% are not organic keywords; 60% are propaganda being backed up by propaganda agents.

1.2 Thesis Organization

The paper is organized as follows: Chapter 2 provides the background and motivation of information disorder, hence the discussion of propaganda and its involvement in the digital world. Information disorder is discussed in its specifics, starting with its forms, then phases, and finally its elements, emphasizing the role of the users in such circumstances, striving to understand the users' experience and behaviors.

Chapter 3 provides an overview of the previous work on information disorder prediction. The approaches to solve the problem are filtered based on the phases of information disorder and are discussed.

Chapter 4 develops the module for propaganda prediction over Twitter. It is divided into data retrieval and gathering, retrieval and categorization of users, and propaganda prediction module.

Chapter 5 sums up the work and discusses the results and conclusions. It also elaborates on the contribution of the research.

Chapter 2: Background and Motivation

Propaganda is evolving at an exponential rate with the technological advancement and the digital transformation. Propaganda was used in many forms, and digital propaganda has recently taken its course in the political arena, where it has affected important decisions and elections results. Propaganda is not new to the world, but the digitalization has made it spread faster and farther, at a more successful rate. It provides its creators with knowledge, power, and control in order to gain domination over a certain subject or to get certain aspired results.

Since users play the most important part in influencing the results of any specific event, such as political/election campaigns, they are targeted based on the personal data they share online. Their behaviors, interests, and digital footprints are heavily used in the creation of propaganda campaigns that result in the deflection of people's beliefs.

2.1 Propaganda in digital communication

The use of the digital world to push propaganda has become a main factor of cyber warfare. The most notable example is the 2016 US presidential election [14] whose results were directly impacted by one digital marketing campaign as part of a major propaganda campaign that led to the election of Donald Trump. Propaganda in the digital world is highly guided by automated accounts, users, or bots that push information in a targeted manner to reach regular users and target their behaviors from every angle, be it political, sociological, or business. Alternatively, propaganda could be the result of organic content shared or generated by groups of users collaborating deliberately. Information disorder in all its forms, phases, and elements creates a build-up that leads to propaganda.

2.1.1 Propaganda

Propaganda is the dissemination of information, facts or false facts, to manipulate public opinion. Modern propaganda operates with all types of information disorder. Modern

propaganda is based on scientific analysis of psychology and sociology [30]. Individuals are no longer viewed as unique beings; they are rather viewed as a set of characteristics, such as motivation, feelings, etc., they have in common with each other. Thus, their behavior is paramount to and at the core of propaganda campaigns. Moreover, they are considered as part of the mass and are categorized because in that way their psychic defenses would be weakened and reactions would be easier to provoke; thus, those behind propaganda profit from the process of emotions, reactions, and behaviors diffusion through the mass [30].

Propaganda has the means to prevent messages from being considered oppressive and to be adopted by people of their own accord. Therefore, people end up being manipulated into following certain dogmas enthusiastically and doing what they are targeted to do without being consciously aware of it. In general, a well-organized propaganda uses every entity of the offline and online media world to target people and turn them to supporters: the deflection starts with small groups surrounding an individual to make him/her lose all his/her defenses, equilibrium, and resistance toward them; consequently, the action of propaganda becomes possible [30].

2.1.2 An evolution of information disorder

Information disorder in all its forms means that a person has digitally received information that could be wrong, misleading, or even correct but not in the correct timeframe. This information is pushed based on this person's digital footprint that by itself is a reflection of his/her physical behavior. Whether it is misinformation or disinformation, an agent is involved in creating the content which is then produced and communicated. The content can be positive or negative and produced by one or many agents, but its target is one. In other words, information disorder in all its forms is used to create and diffuse propaganda.

2.1.3 Forms in the digital world

Computational propaganda is the use of algorithms, automation, and human curiosity to purposefully distribute misleading information over social media networks [23]. Nowadays social media are top platforms for political engagement and important channels for disseminating news. Social media platforms are the primary media over which young people develop their political identities [23].

- Social media are used to create monopoly platforms for public life.
- The majority of young voters use social media to share information on political news.
- Social media are used as tools for public opinion manipulation by targeting particular segments of the public (categorization targeting) via ads.

Propaganda takes many forms in the digital world and has multiple targets.

- Computational propaganda and social media bots have been more broadly used to manipulate online discussion.
- Political contexts are controlled and dominated by organized misinformation campaigns and governments.
- Individual users operate and design fake and highly automated social media accounts.

2.1.4 Knowledge, power and control

Cambridge Analytica is the name of the company that managed many digital propaganda campaigns before it was shut down. Most known campaigns were the 2016 US presidential election and the Brexit campaigns. In brief, the most important factors that were used in Trump's election campaign and that made all the difference are stated as follow:

- Social Media: they were used to gather information about every US voter; different platforms were used, the main one being Facebook which gave away the data from 86 million user profiles.
- User Behavior: most US voters were categorized using computational tools [8] based on their different behaviors, religions, backgrounds, and any other available pattern.
- Information disorder: social media algorithms and automation run ads based on the categorization established to mislead voters, change their mindsets, and push them to vote [20]. An average of 1million USD was spent daily on Facebook ads alone.
- Propaganda: Donald Trump's ranking improved and went up from the lowest to the highest position in the US presidential race [26]. He won the election making this case one of the most successful computational propaganda ever run.

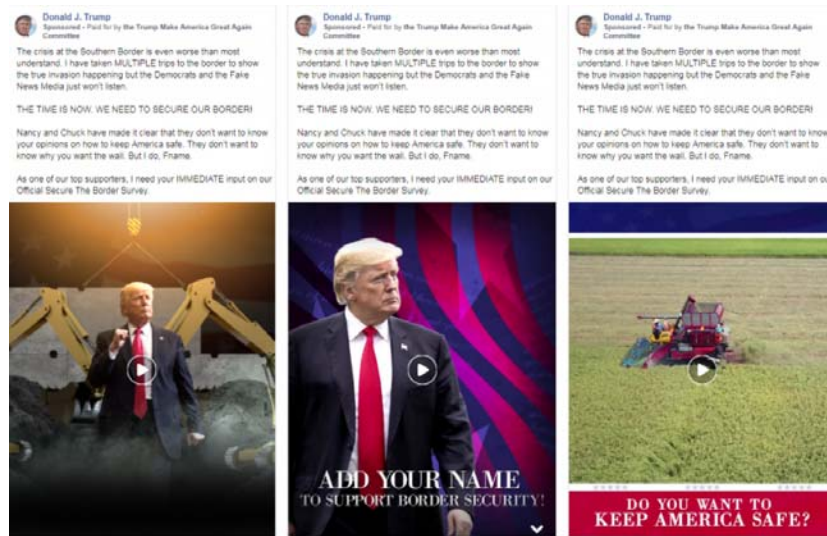


Figure 1: Sample of ads that were targeting US voters based on their classification

2.2 Information disorder

Information disorder is one of the most discussed topics today. The manipulation of information, whether in the form of fake news or any other form, to create misleading content has damaged a lot of people, businesses, and even countries [8]. Information disorder use traces back to the Romans, who used it in different forms but with the same purpose, intentionally manipulating people's beliefs and opinions [16]. With the technological advances nowadays, information disorder has reached different levels, where targeting and retargeting each individual based on their behaviors can be based on precise backgrounds, behaviors, or social circles. Social media have boosted this process by making it faster and easier for the agents to create, communicate, and share their content. Technological advances have even helped in the targeting process of the users based on their interaction and behavioral patterns.

2.2.1 Types

The three main types of information disorder can be narrowed to MIS-, DIS- and MAL information [8]. "Fake News" discourse joins the three types of information disorder, but it is important to distinguish the messages that are false from those that are true, as well as the messages that are produced, distributed, and created to do harm from those that are not.

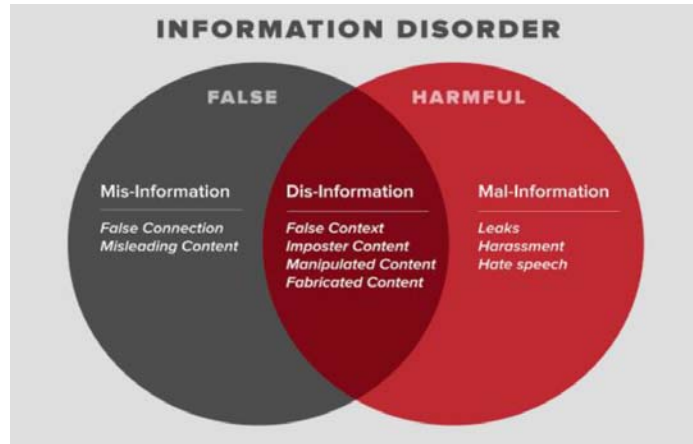


Figure 2: Information Disorder [8]

2.2.1.1 Misinformation

Misinformation is wrong information that is not deliberately created. One of the most important examples is the Champs Elysee attack in 2017 which inspired a great push of misinformation. Users on social media unwittingly published numbers of rumors like the killing of a second policeman [8,16].

2.2.1.2 Disinformation

Disinformation is wrong information that was purposefully created to harm a person, country, organization, or social group. For example, a sophisticated duplicate version of the Belgian newspaper Le Soir was created with false articles about the French president [8,16].

2.2.1.3 Mal-information

Mal-information is true information twisted to create harm. It could be used at the wrong time or the wrong place. An example of how mal-information could be used is the French president leaks which contained real emails and which happened only few hours before the election blackout period. Macron's presidential campaign has allegedly employed disinformation at later stages to diminish the impact of the leaks [8,16].

2.2.2 Phases

To dive deeper into information disorder in all its types, a thorough study of the phases is essential. It is very important to consider the different phases of a particular information disorder with its elements, because the agent who delivers/spread the content is often

different from the agent that creates it. For example, the mastermind behind the content of the Brexit Campaign is completely different from the low paid trolls who mass delivered it. Hence the focus is on the phases of each of these elements: how they shift between phases and how the whole mass information disorder makes it through to its audience and accomplishes its intended work.

2.2.2.1 Creation

Agents behind the creation of any message or content can have different backgrounds and this depends on their intentions. Whether they are working for the public or private sector, they can be seen or be even working in the shadows or for the unknown. The creation can be misleading with its content and can be transferred to the production phase without anyone comprehending its purpose. Propaganda initiation starts with the creation of the content which could be negative, positive, or even neutral about a certain topic, leading in the end to a targeted purpose.

A propaganda content creation can be led by one entity and later produced and communicated in different layers and forms. Alternatively, it can be organically created by many entities and later on produced and communicated horizontally.

2.2.2.2 Communication

The communication phase is the last phase of the information disorder process. It is when the message has been distributed or made public. There are many forms of distribution and platforms. The distribution could be on offline or online platforms. Offline is the traditional way to communicate messages or information: it can be through newspapers, billboards, or even television. Online platforms today have been growing exponentially in the communication platforms including any digital platform and mainly social media. This paper discusses social media communication platforms and how information disorder is communicated and processed to lead to propaganda.

2.2.2.3 Production

The production phase is when the content is transformed into an appealing visual, video, paragraph or any form that is used in the communication phase. Depending on the purpose

and the communication platform, the production phase gives an end product in the form that is most appealing to its targeted audience.

Based on all the input from the creation phase, the production involves other layers and transforms the content to make it the most appealing for every targeted society or community. The language used could include framing, biases, metaphors, and other approaches [13].

Re-production is the repetition of the production after the message has been communicated. Reproducing the message delivers the same meaning in most of the cases, and it is done by external parties that receive the message, support it, and adapt it in their own language and terms to satisfy their own readers, communities, and targets.

2.2.3 Elements

The three elements that are involved in all the phases and information disorder types are the agent, message, and interpreter. The agent and the interpreter are people, and the message could be any form of content.

2.2.3.1 Agent

Agents are included in all three phases: creation, production, and distribution and have many targets in mind. The focus is on the characteristics of the agent which can vary from phase to phase.

Seven major characteristics can be portrayed for an agent [8]:

- 1) Type: an agent can be official, such as intelligence services, political parties, and news organizations. They can also be unofficial, like groups of people that have joined in and agreed on a certain idea or issue.
- 2) Organization: an agent can work individually, consistently, in tightly-organized organizations (e.g., lobbying groups or PR agencies) or in impromptu groups developed over common interests.
- 3) Motivation: the four motivating factors are financial, political, social and psychological.

- Financial: gaining direct or indirect financial profit by empowering or attacking any business or party;
 - Political: attempting to influence public opinion over a certain point or discrediting a political candidate;
 - Social: connecting with certain groups or communities online and offline.
 - Psychological: seeking reinforcement or prestige
- 4) Audience: Audiences differ with every agent. The audience can vary from an organization's internal mailing lists or consumers, to social groups targeting their socioeconomic characteristics, to an entire population of a country. Each agent gives different approaches for every audience.
- 5) Technology: Technology has eased up the process for the agent in that it has become much easier and cheaper. An agent working with technology is categorized by the platform he is using, and the time stamp he leaves behind.
- 6) Misleading: Some agents may or may not intend to deliberately mislead the target audience. An agent mainly follows the same patterns whether they intend to be misleading or not.
- 7) Harmful: Similarly to misleading, an agent may or may not intend to be harmful. Messages help identify the patterns in messages which can then be connected to different agents.

2.2.3.2 Message

The message takes form in the production phase and is communicated in the communication phase. It can be communicated, employing cues, in person (gossip, speeches etc.), in text (offline or online), or in audio/visual material (images, videos).

There are 5 main characteristics to be taken into consideration for a message:

- 1) Durability: some messages are created to stay relevant and impactful for a long time while others are designed to have a high impact for a short term.

- 2) Accuracy: the accuracy of a message is impacted by whether the message is directly or indirectly related to a certain topic or propaganda. Accurate messages can be captioned by a known source while inaccurate messages could lead to fake sources.
- 3) Legality: the legality of the messages is related to recognized hate speech, intellectual properties, and harassment or privacy infringements. Messages legality differs depending on jurisdiction.
- 4) Imposter content: the message can use official branding (logo ...) unofficially, or it may steal an individual's name or image in order to give the appearance of credibility. As such, the question here is whether the message imposes official sources or not.
- 5) Target: the agent always has an intended audience in mind and translates the message in different forms to reach the right audience. These forms could be linguistic, visual, or even structural, based on the cultural background of the targeted audience.

2.2.3.3 Interpreter

Audiences are rarely passive recipients of content. An audience is made up of a group of individuals, each of whom understands and interprets the content according to his/her own personal experience and socio-cultural and political positions [8].

Understanding the ritualistic aspect of the communication process is critical for understanding how and why an audience reacts to content in different ways. The types of information everyone consumes and the way they make sense of it are directly impacted by their self-identity and the communities/cultures they associate with.



Figure 3: Characteristics for each element of an example of information disorder [8]

2.3 User Experience

It all begins and ends with the person, also known as the user in the digital world. It is very important to understand what data surrounds the user and how it is being collected and used. A normal user surfing the internet, whether on social media or other platforms, leaves a direct or indirect trace of his/her presence. Nowadays, this trace is being provided to corporations to study, manipulate, and target users in different ways and purposes.

2.3.1 Available user data

Today, Social Networking Sites (SNS) are becoming very versatile, servicing a wide range of functionalities [7] from posting messages, videos, and images, to shopping, finding a job, and playing online games. The use of social networking sites has become a daily routine and a meeting place to socialize with friends and family or even meet new people. Today more than two billion users use SNS, consuming those platforms as well as sharing and uploading hundreds of billions of data [7].

Following the variety of social networking sites, user data can differ greatly from one platform to another. This diversity of SNS platforms and services, along with the huge

appeal and high use of SNS, creates a wealth of information about users. User data can be separated into seven non-exclusive categories or groups [7], noting that some information may transcend multiple categories.

- 1- Personal details is the most shared category of all the user data being shared because for users to create a digital profile, they must fill the most basic questions about their personal details. This information can completely or partially identify a specific individual. Users willingly share their full names, pictures, dates of birth, birth places, home addresses, and most of their personal details to create their digital ID.
- 2- Interests and preferences: data are gathered indirectly; platforms gather and create hidden profile categorization for the users based on their likes and dislikes patterns of any topic, be it movies or books or politics and sexual preferences. These data are gathered by the click ratio on any suggested post, likes or dislikes for this post and even the time spent on it.
- 3- Social circles: communities or social circles are similar to those of the offline world. In the digital world, users create a pool of people that are labeled as their social circles. These people the users are friends with follow the latter's updates and interact directly or indirectly with them. This social circle can include family members, friends, professional contacts, partner(s), etc. The more people in the users' social circle, the more these typical users can influence others.
- 4- Shared content: composed of original posts being shared by users. Data are gathered from wall posts, blog posts, comments, opinions, likes, and shares based on timestamps, consistency, and originality of the content. This shared content differs if it is shared on public platforms or on personal pages.
- 5- Locations: data retrieved from users' geo-locations work on different layers. Users can create a post while tagging and mentioning the location they are in. Advanced artificial intelligence software can automatically detect the location from the image or video, and enabling location services on mobiles automatically updates users' locations with every online interaction they make. As such, active online users have their locations at specific times and events saved and shared with corporations.

- 6- Qualifications: personal details can have some points in common. Qualifications mainly are the data that cover the education, professional experience, training, certifications, and memberships in professional organizations ... user deliberately share such info with friends and surrounding.
- 7- Life events: Users share intentionally or unintentionally their life events such as marriage, pregnancy, birth, retirement, and other details with others online. These help the advanced categorization of their social lives.

2.3.2 Cognitive biases

Cognitive biases are related to judgment and behavior. Such biases do not necessarily entail incorrect behavior [6] but rather refer to “deviations” from the rational behavior prediction. The following are some of the cognitive biases that appear in the digital world: optimism bias and overconfidence, hyperbolic discounting, anchoring, and framing effect [13].

- 1- Optimism bias and over confidence: users’ tendency to accept being compared to others is not considered as risky as experiencing negative events. More specifically, the optimism bias for online data sharing breach negatively affects the adoption of high protective behaviors, effectively hindering individuals from self-protection. Moreover, having overconfidence in their knowledge is also shown when given more control over their privacy settings: users tend to expose and reveal more data, hence data flow increases.
- 2- Framing Effect: users’ decisions are influenced by how the available choices are framed via wording, situations, and settings. This affects privacy decision making in the same way as when alternatives are shown in a more positive light and in a more appealing setting.
- 3- Hyperbolic discounting: it is the users’ tendency to get a smaller reward sooner instead of a bigger reward later. This follows the saying, the sooner the better. Sharing data or information for a small reward on the spot causes people not to check the source or the background of that information.
- 4- Anchoring: involves points of references users rely on before making decisions. For example, they can share content from someone they know with the public without

double-checking what credible information the content includes simply because they know this person. This type of reliance on anchors, such as what others post and share and how many likes these posts get, affects users' judgment.

- 5- Bounded rationality: refers to the notion that, when taking decisions, users' rationality is constrained by the available data, their mindsets, cognitive limitations, and the available time to make that decision. Indeed, data and information acquisition requires users to evaluate, in a restricted amount of time, the consequences of making decisions based on highly uncertain information.

Such spots need great cognitive efforts and wider access to information. Consequently, in such cases, users react heuristically, with the rules of thumb, called shortcuts in decision making.

2.3.3 Visual Cues

Content clarity, speed to understand it, and attraction are the characteristics of a good content presentation. As the content is meant for the digital world, there are factors based on which the online post design is evaluated using a near-infrared spectroscopy (NIRS) where the factors [16] evaluated are the ratings of a specific post, its colors, and its placement.

A post rating depends on how much interaction it gets from the users it is appearing to. Whether a user likes, shares, or even reads it is considered an interaction. The study done by NIRS [16] shows that a user detects and interacts with high rated posts better than with low rated ones, hence the need to have good ratings on any piece of content.

Darker colors give better results compared to lighter ones, especially for verbal posts. In the NIRS, a dark background with light font color got the highest results [16]. Regarding pictorial posts, using colors in harmony with the picture itself and its background showed appealing results.

Post placements on digital platforms can differ; the post can be placed on the top, bottom, left or right of the pages. Most platforms use the header and footer (top and bottom of the pages) which leaves the right and left margins for the shared content. The NIRS showed that pictorial posts placed on the left of the pages get much better results than the posts

placed on the right [16]; there were no significant results for verbal or textual posts to make the same conclusion.

2.3.4 Social behavior

Propaganda aims to deceive people directly or indirectly in order to make them change their beliefs or opinions. Behavioral science plays an important role in this process. Scientists investigated the concept of deception and its process in modifying human behavior [6]. There are two main pillars of deception: functional deception which uses information disorder, and intentional deception which uses desires and/or beliefs [29]. From a psychological perspective, deception is defined as the act of providing misleading information to redirect people [31] or as the explicit misrepresentation of a fact aiming to mislead users.

- 1- Herd behavior: individuals' behaviors may be controlled or governed by their externalities. What everyone else is following is rational because they have information that other people do not have which leads them to their decision. Herd behavior is characterized by equilibrium, selecting problems based on the quality of information transferred by the whole group opinion [25]. Decisions taken by a user mainly depend on how choices are perceived. Such perceptions are affected by social elements; they are not independent of certain social environments where decisions are taken [25]. An individual user who belongs to a herd is considered to share that herd's social concerns and motivations.
- 2- Cultural effects: cultural variations are used to elaborate on the deception cues. Individuals from collectivist cultures are more prone to using deception when sharing any type of information than those from individualistic cultures [27]. This is a well-known classification of cultural values with two cultural dimensions [29]: individualism vs. collectivism. In the individualistic culture, the sense of 'I' and the individual's 'privacy' are valued, and individuals are loosely tied to one another. On the other hand, in collectivist cultures, 'we-ness' and 'belonging' to each other are highly shown where individuals are rightly connected one to another. The cultural effect is very different between the two dimensions, and each individual has his/her

own variances in their behavior in the digital world to which their cultural ties are transferred.

- 3- Social groups: entities formed based on mutual common issues, locations, cultures, and platforms. The same applies to the online presence of such groups, where they communicate and share information easily and acquire common grounds in terms of identities, likes, and dislikes. The importance of social categories or groups in shaping social perception lies in how and when they continue to evolve. Behavioral traits uncovering cognitive processes show group perceptions [2]. For example, new connections between emotions and social categories are being discovered [9]. Hence the evolution of social groups and variables that can define them is continuous, and their implementation in the digital world is increasing.

2.3.5 User behavior in Lebanon – Survey

The flow of information during crises or any sort of campaigns and events is enormous. Media agencies overflow offline and online platforms with content whether it is real or fake; as such, information disorder would be at its peak. We conducted a survey in Lebanon in 2020, a year during which the country was overwhelmed by a revolution, a very bad economic situation, a pandemic, and the explosion of Beirut port in February. The aim of the survey is to define how people interact with the information received on different platforms in times of crisis and to understand whether or not they can process the purpose of a targeted information disorder campaign and identify the real from the fake.

2.3.5.1 Instruments and data collection

An online survey was deployed in spring 2020 through Google forms platform to the general public. The survey was characterized by the snowball effect. Participants shared the survey with their entourage who would in turn share the survey again with their entourage and so on. No compensation was offered for the participants. After data cleaning, there were 382 usable unique responses.

The survey did not include any open-ended questions; all the questions were structured with nominal, ordinal, and interval measures. The questions were divided into 3 categories: the first category was to understand the respondents' personal information and

backgrounds; the second was to understand the media platforms they use and rely on; and the last was to see how much fake news they think they can identify, whether they share it and where, and whether they understand information disorder.

2.3.5.2 Data analysis and results

The answers were exported from Google form in a csv (excel) file format and exported into QlikView, a data analysis and visualization software, to simplify the analysis of the data.

Demographics and backgrounds of the respondents are shown in the figure below: the gender average is almost equally split between female (51%) and male (49%) respondents. 88% of the participants have a bachelor or master's degree, and 64% live in Mount Lebanon, a region characterized by an abundant presence of higher education institutions which can explain the high percentage of respondents with undergraduate and graduate degrees.

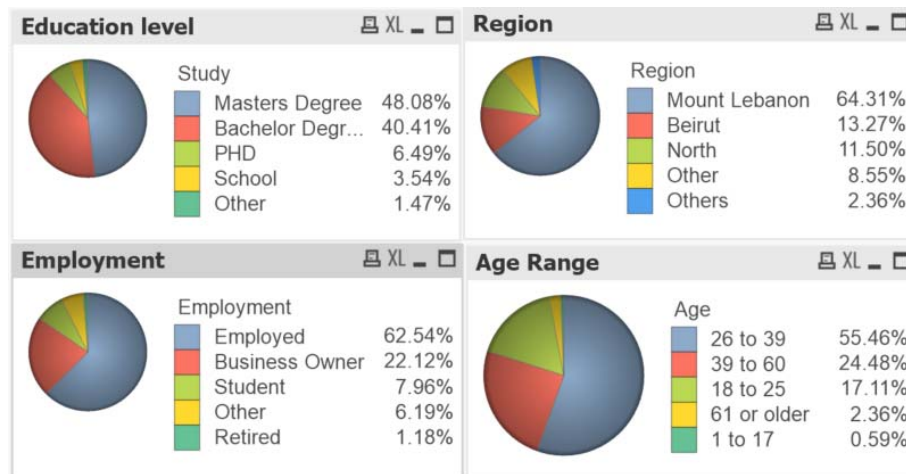


Figure 4: Survey respondent's demographics

When it comes to platform usage for online information access, of all the available digital platforms, 62% of the respondents receive their news on WhatsApp (highest) and 24% on Twitter (lowest). As for fake news identification, respondents affirmed being able to identify 56% of fake news on WhatsApp and 25% on Twitter.

For the traditional platforms, options were divided among the most known news platforms of which 3 are known to have indirect ties with specific Lebanese political parties and are watched almost equally by the 382 respondents while the remaining 3 are directly tied with political parties and are watched by 14% of the respondents.

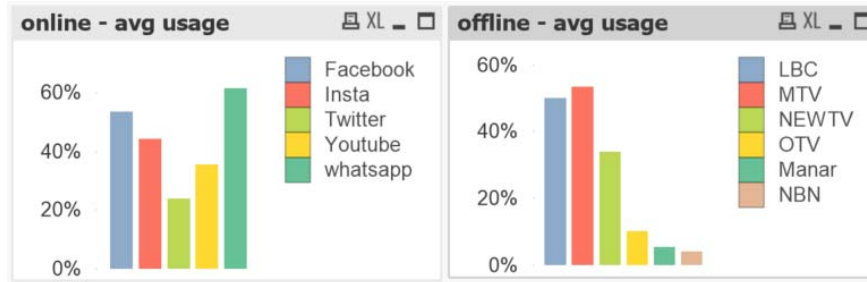


Figure 5: Offline and Online platforms usage

Participants were asked if they think that the content shared on social media is coherent with the content shared on the offline platforms: 13.4% think that it is never coherent while 37.2% think that half of the content is coherent. Only 4.7% of the participants think that there is high coherence between the offline and online platforms.

In the third category of the questions, participants were asked about whether they check the authenticity of the news when they receive it and if they share it or not. An average of 48% check the authenticity of the news, and an average of 32% share the news.

The last question of the form tackled participants' knowledge of the reason behind the creation and communication of fake news, in other words if they can identify whether certain news is part of a propaganda campaign that is clear to the public eye. 72% of respondents claim they understand the reason behind the creation and communication of fake news.

2.3.5.3 Patterns

Patterns that are related to respondents' demographics and behaviors were extracted and showed no correlation between the different questions and the respondents' answers.

17% of the respondents do not check whether the news is fake or not and do not share any news at all; these show an average distribution of awareness of the purpose behind fake news. As for the knowledge of the purpose of the fake news being shared, 18% of the respondents answered that they are knowledgeable in identifying fake news and showed average patterns of checking and sharing the news. Participants with higher results of knowledge showed higher patterns in checking the authenticity and news sharing.

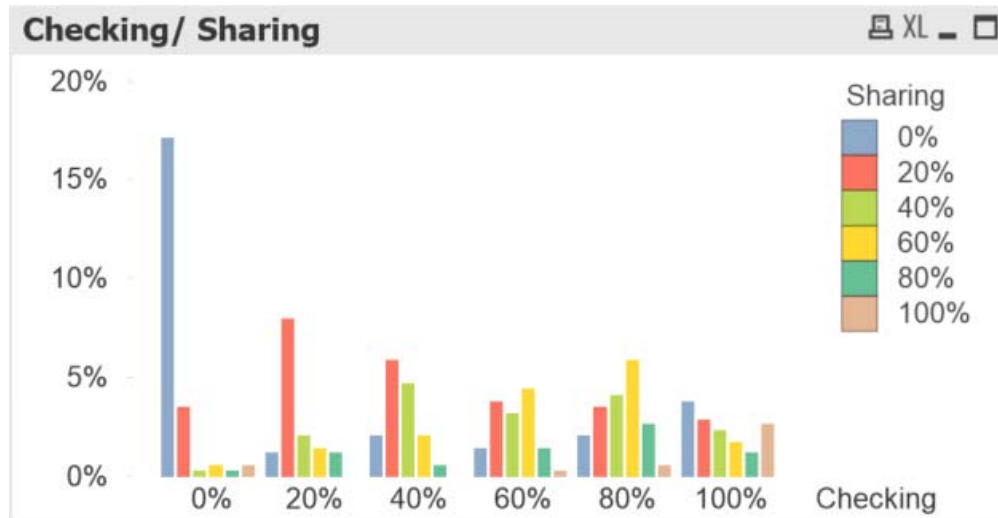


Figure 6: Participants answers analysis between checking and sharing of news

78% of elder participants, 61 years and above, claim awareness of the purpose of the fake news being communicated: that is 7% higher than the average. Elder participants were double the average in sharing news and lower than the average in checking the content of the news.

On the other hand, 85.71% of female participants show higher patterns in sharing news, of whom 80% check the news before sharing them and 50% believe they are aware of the purpose behind the fake news that they receive.

2.4 Social media used for information disorder

From the start, media outlets have been the leading disseminator and curator of news. Today, social media, especially Twitter [17], became the major source of breaking news and news trends. A glimpse at the size of users in 2018 shows that there were 3.2 billion users on social media, of which 2.4 billion were on Facebook [20]. Moreover, Twitter has been mostly used for online news generation and communication because of the ease of account creation and content communication. Hence, the emergence of social media platforms was greeted as a formidable challenger to the centralized publishing systems and their monopoly [5]. This rise of social media and today's advanced technology driven era boosted the use of information disorder to create propaganda.

2.4.1 Forms

There are many forms that information disorder can take on social platforms depending on the phase, false information being one of the main forms. Information disorder can be classified under misinformation or disinformation, such as fake news, rumors and even information manipulation [6]. Another well-known form is fake identities that could be fake profiles, profile cloning, or compromised accounts; these all serve the same target which is manipulation in order to get a certain intended result in return. Other forms like luring and even human targeted attacks are used as well[18]. Hybrid approaches are the mixture of holding fake identities, creating fake content, and tampering with communication platforms [20].

2.4.2 Applications

There is no one universal usability for the information disorder on social platforms [24] as it differs with every type, target, and form. Whether propaganda is launched on Facebook, Twitter, or even news websites, it has the same target. Each platform implements a variety of applications for content sharing. It may differ between a picture, video, text [13] and a combination of all, depending on the campaign implemented and used.

2.4.3 User Targeting

Social platforms give people the means and the way to share their message with others who are not part of the circle of people they follow or interact with. User targeting is shown to be the most reliable way to get a message to the right receiver [19].

User targeting is not done on a personal level where each person is targeted by their name or preferences; it is done based on classification and filtering mechanisms that target users in groups. A more advanced mechanism can target filter bubbles [4] which is a closed loop of news sharing based on cultural and common backgrounds. Paid advertisements use those structured targeting mechanisms that necessitate a payment to be made to the platform which then pushes the news and messages to a precise target audience.

Organic targeting requires such filter bubbles or groups to pass on the message organically between each other without pushing it to the users without paying for the platform to make it appear in their feeds.

In all cases, the user becomes the weakest link in the cycle, and the creators of the propaganda rely on social engineering techniques to trick such users into changing their behaviors, thus compromising all their decisions [21].

Chapter 3: State of the art of deception detection in social media

The results of the research undertaken and developed in the paper show that information disorder is affecting most if not all the online campaigns whether their intentions were to give a positive or negative aspect of the topic at hand. The focus of this chapter is on planning a predictive model for propaganda on social media. This module will tackle each phase of the communication disorder since understanding these phases separately and their players sheds light on the type of campaign being produced.

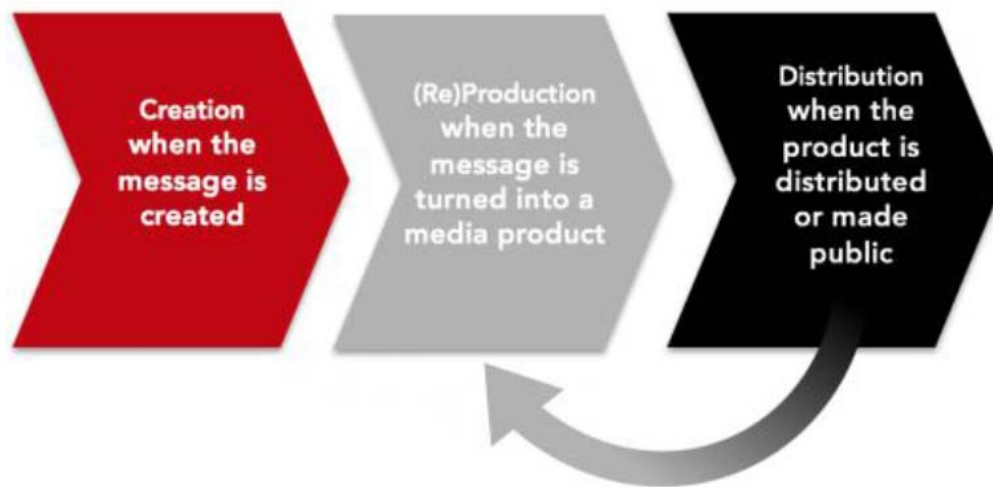


Figure 7: Information Disorder Main Phases [8]

The development of the research and the work on the state of the art of deception detection on social media focused more precisely on Twitter. Twitter is mostly used by media and professionals and gives access to its API's which can be worked on and analyzed.

The user experience is the last of the information disorder phases. The users' reception of the information is one of the most important layers to be elaborated on. Each section in this

chapter will tackle all the research done for each phase while going through the predictive modules already developed and related to each phase.

3.1 Creation (Agent)

The agent is mostly involved in the first phase of the information disorder. In the creation phase, the information disorder can be detected by analyzing the agent involved in the creation of the content. Many approaches have been developed for the detection of fake news based on its source. Three of these approaches will be elaborated on:

- 1- Account history approach: user profile information provides specific activities, features, and behaviors about each user; however, profile information is private. Thus, collecting private information is a violation of a user's privacy rights. So the use of such information is a violation no matter how the information is intended to be used. Besides, collecting profile and behavioral data occurs at high cost [6,8,20].
- 2- Behavioral indications: based on the account history and user profiling already discussed, agents' patterns in content creation can be detected based on their behaviors [2] that are analyzed for every agent [20], which is considered an added layer of detection. This is a sophisticated and complex model in terms of implementation since most of the variables extracted from the huge amount of data for every user are not clear and need too much processing power; in other words the cost is high.
- 3- Account credibility: it is built based on account history and is applied in different stages and levels based on the content being shared, user behavior, and sometimes followers' credibility [8,27]. It Could get very complex but is very efficient since all methods are combined in one module with huge amounts of data. Here, credibility is built based on the user's account itself and the people that follow this account or the people that this account follows [18]. This module has low accuracy in detecting sudden changes, for most of the data is studied over time, hence classified based on past actions.

3.2 Production (Message)

The production phase is when the content is developed into a visual, a video, or just a text before sending it to the communication phase. One content could be produced in different genres of messages and in different types, visual or video, depending on the target audience. Hence each produced message takes the signature of its producer; in other words, produced messages will always have special patterns and other significant points that each producer will imprint on them during the development process, and these differentiate producers from each other.

A common approach into predicting information disorder at this phase is data Base or dictionary approach. It is based on grouping and categorizing messages into forming a data base of messages and words and extracting patterns and significant differences in the production that later on will be used for comparison with new content [20]. On the other hand, the physical implementation of such a large data base or the formation of such a dictionary for the analysis is costly, especially that the data size could be enormous.

Message and Content cohesion focuses on the consistency to identify the information disorder within it [20]: this includes the whole content, different parts of the content, or the metadata attached with it. It focuses on identifying deceptive cues that are leaked by inappropriate or bad encoded functions in the content, so it reveals cues of misleading messages and not fake content directly.

Content analysis (qualitative, quantitative) Wide module of detection and prevention analyzes and categorizes vast amounts of data to identify users based on the content they share on a wide variety of topics [20]. The analysis could be based on quantitative or qualitative approach; for example, the number of words in a message is taken into consideration [7] as well as the use of symbols in that same message. Such a process can be based on the data base module to analyze the content or could be directly connected to the online platform analyzing the content directly from the platform itself. Its implementation is costly and needs huge amounts of processing power to keep it updated.

Sentiment analysis includes more emotional and background information, in addition to explicit content, which can increase the prediction accuracy [6]. However, the use of

sentiment analysis cannot fully leverage the linguistic information in the content where the lexicon is domain-specific [12]. There were no trusted sources for the sentiment analysis approach in Arabic lexicon to be implemented since most of the news shared in this region is in Arabic.

3.3 Communication (Platforms)

The communication phase is when the messages are being delivered after the production phase; depending on the chosen platform, information disorder is delivered for the users or the intended people in different forms and patterns. It could be done via personalized messages or they could be publicly shared.

Applications implemented automatic prevention modules. For example, Facebook, Instagram, and Twitter have automatic identification systems for information disorder that can automatically identify fake news or that labels news as fakes after receiving too many complaints on a certain message, page, or user.

Another model of information disorder in the phase of communication is includes groups present on a certain platform employing personal approaches to check news shared online and certify whether it is legit or tag and report it if it is fake. Such a group exists in Greece and is called Ellinika Hoaxes Facebook group, a Greek community on Facebook whose members exchange information and insights and collaborate to tag and spot fake news and fight to counter its spread [26]. Such solution is organic and does not need big budgets, yet it needs a lot of personal effort and a good number of members as Ellinika Hoaxes groups members are around 2000.

3.4 User Experience (Receiver)

Most of the propaganda campaigns and the information disorder attacks are meant to deflect truth or change a user's point of view on any issue be it political or not. It makes the user or the receiver in this case the most important to study. Many approaches were developed to detect, predict, or identify information disorder even at higher stages of propaganda.

Phishing is considered one of the attacks that are technically difficult for users to detect [21]. Hence phishing prevention was developed with effective systems that are reliable to prevent online social deception by following traces of linked pages and other generated data [6]. However, delay issues may occur since the effectiveness of the developed systems needs high processing power to analyze the huge amounts of shared content.

Social honey pots work on the social media platforms exactly the same way as the honey pots attack prevention work on communication networks [6]. A well deployed honey pot could be very effective in dealing with attackers. The developed social honey pots function as a passive monitoring tool and mainly focus on detecting social media spammers, social bots, or malware. It uses the attackers' profiles to detect them based on variables collected from the social honeypots placed as fake social media accounts. However deploying an effective social honey pot is not going to deceive attackers easily since they would have already gained experience targeting users, not to mention that there is an ethical issue considering the fact that an act of deception is being committed against attackers and perhaps other users too.

Feature based deception detection includes raw features, such as word embedding, word vectors, URLs, and hashtags. Some advanced features include statistics, linguistic inquiry, word count, and other metadata, such as source, time, or location [6,8,14]. This method is characterized by high accuracy and low false positive rates, though the extraction of sophisticated features comes at a high cost [7].

The focus is more on Twitter as it is one of the major micro blogging service providers and the most trusted in Lebanon for the news. A number of researchers for tweet classification tasks were developed [10]. Machine learning methods were used based on training a classifier on a labeled text collection. Such supervised learning methods included Support Vector Machines (SVM), Naive Bayes (NB), Neural Networks (NN) and Random Forest approach to solve their classification problems [10]. Such approaches proved to be efficient, but the training data that was built needed a lot of time and showed flaws and was not always up to date.

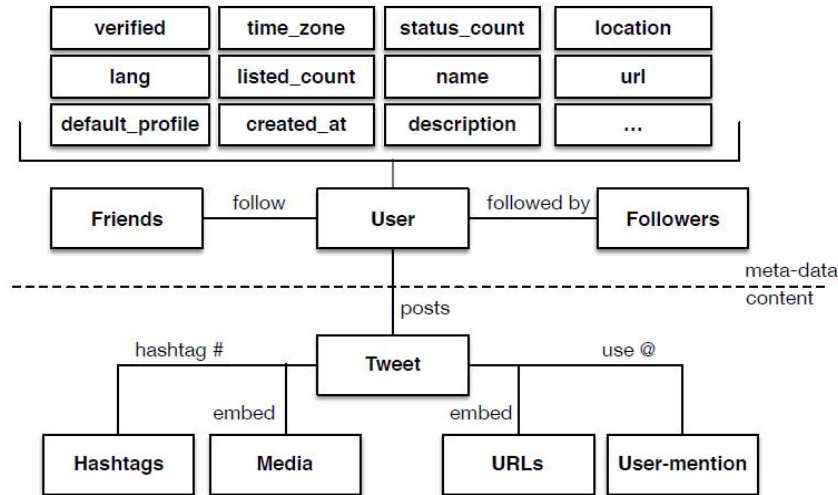


Figure 8: Twitter infographic

Many approaches were based on user categorization in propaganda detection. Understanding the types of users on Twitter is based on user classes while taking into consideration how the Twitter APIs, seen in figure 8 under Twitter infographic variables, lead to user classification [14,28]. A research based on a business perspective found six classes of users [28] identified as personal users, professional users, business users, spam users, feed/ news, and viral/ marketing services, based on their online behavior [10,17,28].

After the Trump and Brexit campaigns, BOT detection was on top of the prediction research, and the underlying assumption is that BOT accounts show a different social behavior than normal or than that of legitimate users. Specifically, machine learning techniques attempt to detect the signature of BOT behavior, generally based on features such as profile, geographical data, account creation date, as well as the content, sentiment of the posts, and their consistency [7].

Chapter 4: Twitter propaganda predictive module based on user behavior

Predicting propaganda on Twitter is based on a hybrid approach based on the phases of information disorder.

In the creation phase, the account history approach is considered [6,8,20] with the behavioral indications approach [20] to build the background of the agent [2] who is creating the content. Moving to the production phase, a database [20] is created for the content shared by the users while focusing on the content cohesion approach to detecting consistency in the content. In the final phase, most of the focus is on the prediction module, feature-based deception detection approach [6,8,14], while analyzing the content of the message and going deeper into tweets classification approach [10] and at the same time developing the module and the variables to focus on the users' online behavior. The user categorization approach [14,28] was adapted to fit a political aspect and not a business aspect [10,17,28]. Propaganda is established by multiple non-organic accounts pushing organic content to reach a certain target [5,8,13]; hence understanding those users based on their classification and based on their online behaviors shows whether propaganda is being pushed [15] starting with the trending keywords on Twitter. In the last stage of the module, propaganda prediction is based on those behavioral variables which are based on tweet categorization and user classification.

From a practical perspective, the keywords and comments extraction, user categorization, and propaganda detection are the three phases that have been developed in the module. The first phase starts with a manual extraction of top trending keywords or hashtags used in a recent time frame on Twitter for a certain region. The retrieval of the content is established by machine learning based on the keywords. The extraction of the top users in terms of interaction with the keywords is done by a simple pivot module based on the count of interactions.

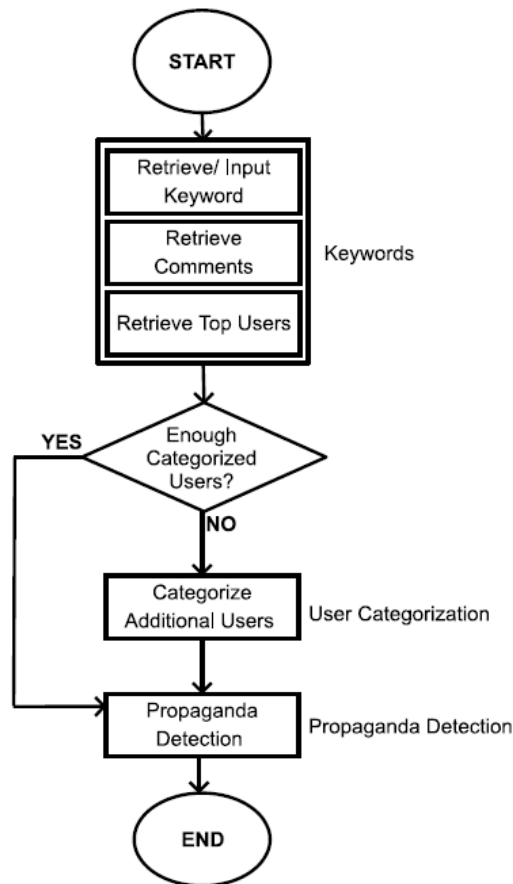


Figure 9: Module general flow chart

Before moving to the categorization layer, the extracted users from the first phase are reviewed to check whether they have been already categorized and classified in the database. The comparative process decides if it contains enough categorized users to move directly to the propaganda prediction phase. If the number of pre-categorized users is inappropriate, the process moves to the categorization layer.

The categorization layer classifies the retrieved users from the keywords layer based on 20 behavioral variables extracted with the machine learning approach based on Twitter API. Conditions are developed to help the tag each user and classify him/her in one of the 7 categories.

Following the categorization layer, the module moves to the propaganda prediction layer. Based on the users who are pushing a certain keyword, the module returns if the keyword

fits the propaganda conditions. It is based on the percentages and the types of the users sharing and posting the keyword.

4.1 Keywords

The first layer in the module (Figure 9) is divided into 3 phases: keyword extraction, content retrieval, and user extraction. Two databases are updated: the keywords database and the users database.

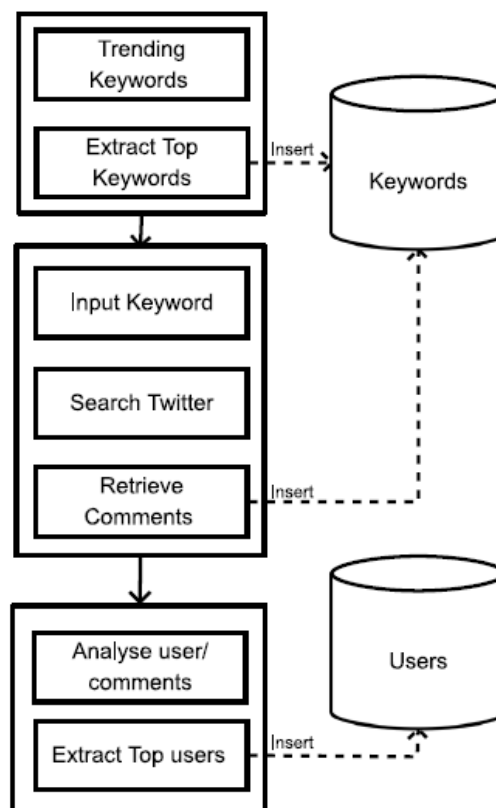


Figure 10: Keyword phase flow chart

4.1.1 Key words extraction

The first phase deals with extracting the top trending keywords from Twitter in a certain time frame. This process is done manually, through logging in to twitter.com and going to

the trending section to check and retrieve the trending keywords in the region. Twitter shows the number of mention for each keyword.

Following this phase and the keyword extraction, the keyword database is updated with the new keywords. A total of 29 keywords are analyzed in the module development phase to which are added 7 keywords in the testing and validation phase. The keyword extraction and analysis took 2 months and was done between 11 June 2020 and 11 August 2020.



Figure 11: Trending keywords from twitter sample

A sample of the keywords extracted from twitter.com on 15/3/2021 is shown in the figure above where the top keyword has 26.4K tweets appearing under the keyword on the left. Since the module is being tested in Lebanon, all the keywords are in Arabic and are related to trends/topics in the region.

4.1.2 Tweets retrieval

The second phase of the keywords layer is tweets retrieval based on the extracted keywords done in the first phase. The keywords are inserted in a machine learning application, RapidMiner, using search Twitter plugin where all the tweets containing specific keywords are retrieved and directly inserted in the keywords database.

The process will generate 12 variables for every tweet that is retrieved and extracted into a database which is in csv format. The extracted variables are:

- Created-At: Creation date of the tweet.
- From-User: User name of the user that created the tweet.

- From-User-Id: User ID of the user that created the tweet. Unique variable.
- To-User: In case the tweet was a reply to a certain tweet or a retweet, the username of the person that was replied to appears under it.
- To-User-Id: In case the tweet was a reply to a certain tweet or a retweet, the User ID of the person that was replied to appears under it. Unique variable.
- Language: In which language the tweet was created. Arabic in this case.
- Source: The direct link of the tweet, unique variable.
- Text: The content of the tweet.
- Geo-Location-Latitude: In case the location was enabled on the tweet, the latitude appears under it.
- Geo-Location-Longitude: In case the location was enabled on the tweet, the longitude appears under it.
- Retweet-Count: the number of retweets this tweet received.
- Id: Each tweet receives an ID, unique variable.

The Keyword is added as a variable and used as a metadata in order to keep track of the tweets. The process is repeated for all the keywords that have been extracted; a database of 40000 tweets is built, retrieved, and saved under keywords.

4.1.3 Users extraction

Extracting all comments related to certain keywords generated a humongous list of tweets with 13 variables for each, in addition to the keyword that was manually added. To distinguish the users that are engaging the most in every keyword, the top users of a certain keyword were analyzed.

Based on the retrieved database for all the tweets, a simple pivot using excel was developed to extract the sum of user IDs that have tweeted a specific keyword more than 1 time. Users were sorted in descending order based on their engagement rate with every specific keyword. The top 5 users for each keyword were extracted and were sent to the user database.

USER ID	Total Tweets
1271311127079604224	6
397199380	4
1484284405	3
216289357	3
701501905051197444	3

Figure 12: Sample of retrieved top users

The use of the user ID is essential as it is the main point of reference, a primary unique key, throughout all the module.

4.1.4 Keyword test

Following the keyword extraction phase and its top users, all the users attached to this keyword are tested with the database. After a data buildup of 2 weeks, repetitive users would show based on the keywords, hence the users would have already been categorized and analyzed, and their data would have already been gathered. Moreover, if more than 50% of the needed users were already present in the classified user database, the module would directly jump to the propaganda detection phase without going through the user categorization phase. The phase output generates the users' IDs of the accounts that are tweeting, retweeting or replying to a specific keyword that is being processed.

4.2 User categorization

The user categorization phase (Figure 9) is the longest phase which needs extensive processing and development. It uses 3 databases in which users, keywords, and users' activity is read and/or inserted. This phase is divided into 2 main layers: the first layer retrieves user background and user activity data, and the second phase, the analytical layer, processes the variable extraction and classification.

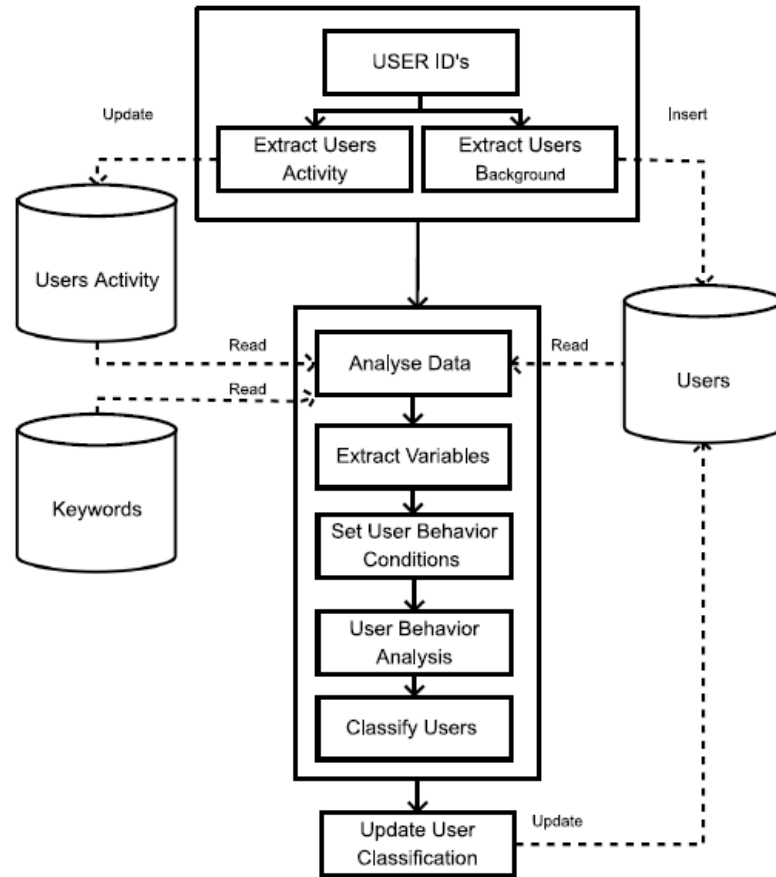


Figure 13: Categorization phase flow chart

4.2.1 User activity and background

Retrieved users' IDs that are transferred from the first phase are classified in order to determine their digital footprint behavior, and the background of the campaign is determined based on the engagement with trending keywords. This process employs two data extractions. The first retrieves the user profile background which is provided by Twitter API, and the second retrieves the users' activity footprints by retrieving all the user activity on Twitter.

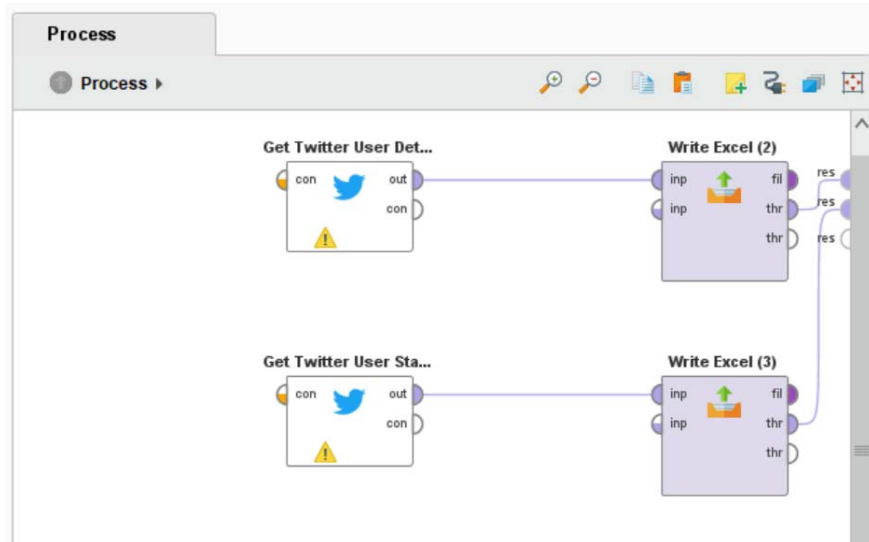


Figure 14: Twitter get user background and statuses processes

The user background extraction is done with machine learning using RapidMiner Studio via “Get Twitter User Details” component, where the user id is used to extract all the available user data on Twitter. The extracted user profile background generates 16 variables in text and in integer format. 150 users’ background were extracted in total and later saved into users database.

- Id: User account ID, unique variable.
- Name: User name as created by the user, unique text.
- ScreenName: User name as set to appear on the profile.
- Description: Profile description as set by the user.
- URL: Profile link, unique text.
- Created-At: Date of creation of the account.
- Location: Account location as set by the user.
- Verified: true/ false variables if the account was verified by Twitter.
- Protected: true/ false variables if the account owner has processed the protection layer with twitter.
- Followers: Number of followers of the profile.

- Friends: Number of profiles the user follows.
- Favorites: Number of topics or keywords the user is most interested in. The user enables each topic manually.
- Tweets: Total number of tweets of the profile since date of creation.
- Language: The language preference of the account.
- Profile-Image-URL: Profile picture direct link.
- Time-Zone: Time zone of the user if filled.

User activities extraction is done with the same machine learning platform but using the “Get User Statues” component in RapidMiner, where the user ID is used to extract all the user activities (tweets, retweets, and replies) based on a pre-set timeline and the number of activities that are set manually in the plugin. This extraction generates 12 variables identical to the extracted data in the keyword comments retrieval using the “Search Twitter” component. The user activity in this process is inserted into the user activity database in the form of csv document and kept separate from the keyword database, and the data is inserted without adding any variable. The user activity database was updated with an average of 3000 activities for every user summing up to 400000 activity in total based on the users’ IDs.

4.2.2 Analytical layer

The analytical layer of this phase is based on past research that laid the ground for the development of a module that tackles the political perspective. The Twitter infographic helps understand the types of users [14,28], and the user classification is based on user behavior on Twitter in the business perspective. Moreover, the propaganda analysis and detection on Twitter [10,17,28] are based on content analysis using machine learning.

This layer is divided to 5 steps, where all the databases created are added in the first step to QlikView application to be joined, analyzed, and visualized. The next steps are: analyze data, extract variables, set user behavior conditions, analyze user behavior, and classify users. This whole layer is treated as one block in QlikView and the last phase is updating the users data base.

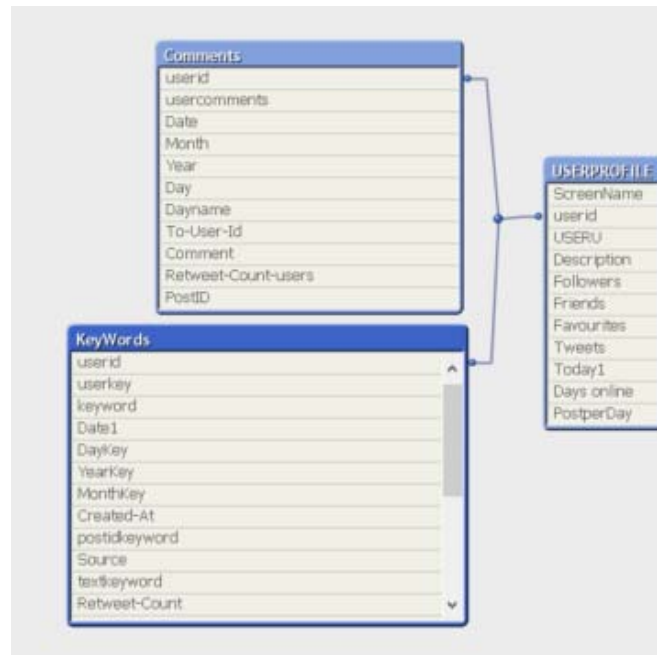


Figure 15: Data base join

4.2.2.1 Step 1 – Analyze data

In the data analysis, the three databases use the user ID as the main key. A simple graph to count the number of activities of all the users for a specific day in February 2020 showed some patterns and peaks for users as shown in the line graph below. Apparent is related engagement for different users on a specific day, such as day 6.

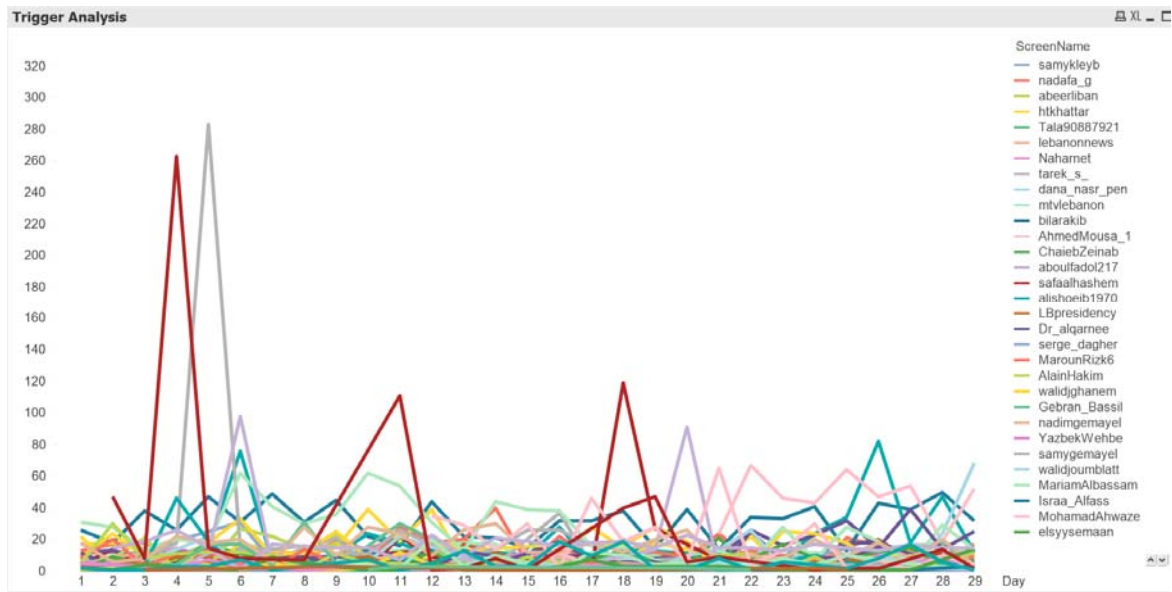


Figure 16: Activity count per day for users on February 2020

A clear connection between the users and their activity is present and is shown visually. To detect specific campaigns and relation between users over a time period, a sample was taken for the users that mentioned the keyword **الدرون_بالدرون** and extracted all the users' activities over a period extending from April 2020 to July 2020.

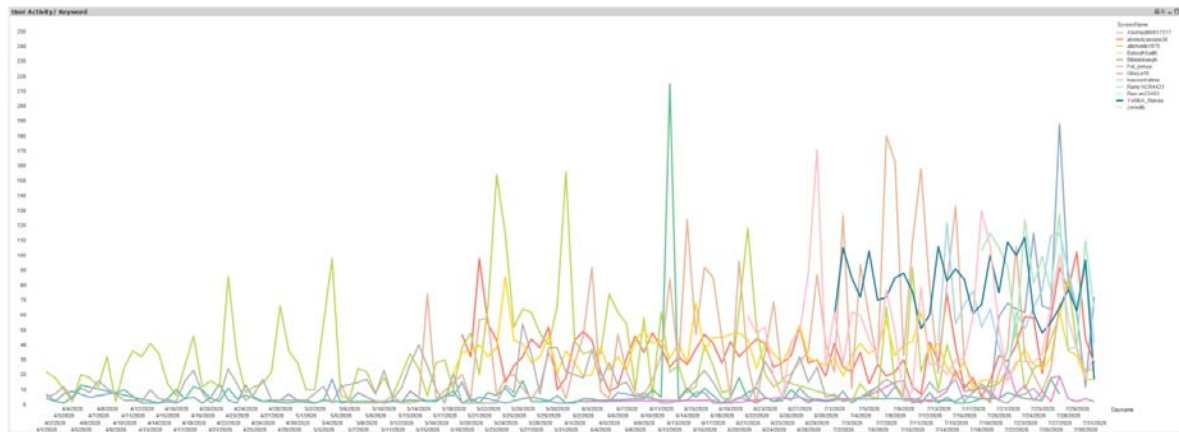


Figure 17: User Activity April to July 2020 **الدرون_بالدرون**

The data were extracted into a csv file, and the correlation analysis that was applied shows a clear pattern between the users over the time period and between the users and total number of activity of these users.

		1	2	3	4	5	6	7	8	9	10	11	12
ref	Total Activity	0.988	0.998	0.995	0.997	0.944	0.994	0.991	0.994	0.996	0.995	0.998	0.994
1	AboHadi90017277		0.984	0.979	0.980	0.986	0.981	0.979	0.979	0.987	0.988	0.981	0.979
2	ahmedyassine30			0.995	0.997	0.941	0.992	0.987	0.997	0.997	0.993	0.997	0.992
3	alishoeib1970				0.997	0.939	0.990	0.988	0.995	0.996	0.993	0.998	0.996
4	BatoulKhalil6					0.942	0.993	0.988	0.997	0.996	0.995	0.998	0.994
5	Bilalabbasgh						0.942	0.990	0.938	0.997	0.998	0.999	0.996
6	Fat_bekaa							0.988	0.990	0.995	0.996	0.995	0.990
7	Gheya10								0.988	0.985	0.994	0.992	0.983
8	hasounfatima									0.997	0.998	0.999	0.995
9	Ramy16294423										0.992	0.993	0.994
10	Rawan23493											0.995	0.995
11	YoMnA_Nanaa												0.994

Figure 18: Correlation Analysis #الدرون_بالدرون activity from April to July 2020

The correlation analysis shows clear highly correlated patterns between those users over the time period extending from April to July 2020. This correlation appears between the users with each other and between the users and the total activity of all the users of this pool on a specific day. The average correlation between the whole data is **0.9881** which clearly shows that this pool of users correlate with each other based on their activities which vary according to different triggers which, at one point of time in this case, was the keyword #الدرون_بالدرون.

4.2.2.2 Step 2 – Extract variables

Based on the findings and the clear relations between users over a certain keyword or in other words campaign, step 2 of the module focuses on understanding the behavioral patterns of the users to lay the ground for the classification process, by finding and extracting variables from databases in order to create the categorization module based on their parameters.

User ID is used as the main key for all of the data, and some variables were used without calculations, hence 16 variables were deduced and ready for analysis.

- AGE: user age in days since account creation

This variable is calculated with the present date minus the date of the creation date of the account provided from the users databases.

- Total Tweets: Total tweets since account creation

The number of total tweets is extracted directly from the users database.

- AVG Tweets: Average tweets per day

The calculation is made by dividing the total number of tweets by the Age.

- Median: Median of the tweets per day.

The median is generated directly on QlikView which calculates the value by separating the higher half from the lower half of a data sample; in other words, it gives the center of the data.

- Median/Average: Ratio to check if the user tweets constantly or on specific dates based on triggers; this value is created to identify if users are trigger activated. This value is calculated by dividing the median by AVG Tweets.

- Replies%: Out of all tweets, how much are replies to others.

This value is generated by QlikView, the replies are an activity that contains the user id. The % is the count of activity that contains the user id divided by the total count of posts by counting the posts ID.

- Share%: Out of all tweets, how much are retweets of other users' tweets ratio.

The tweets that start with RT are considered other users' retweets, sharing other users' content, this variable is extracted by QlikView with the use of the formula: $(\text{sum}(\text{if}(\text{SubStringCount}(\text{Comment}, 'RT '), 1, 0)))$ which returns the total number of comments that starts with RT. The ratio is then calculated by dividing the result of the formula over the total number of tweets.

- Organic Content: Out of all tweets, how many are the tweets that are created organically. This variable is calculated when the tweet is neither a reply nor a retweet.

- Tweets: How many tweets were gathered for a specific user.

It is calculated by the count of all the user IDs

- Following: How many users this profile follows. This variable is generated directly from the Users database.

- Followers: How many users follow this profile. This variable is generated directly from the Users database.
- Followers/Following: Ratio of Followers/ Following.
- Retweet total: How many retweets did the user get from other users in total.

It is calculated by the sum of all retweets the user's tweets have gained; it is generated by QlikView by using the sum formula of all retweets variable from the user activity database.

- Retweet AVG: Daily average of retweets from other users.

It is the division between the sum of total retweets over the C? of all his retrieved tweets.

- Retweets/ Followers: Ratio of retweets/ Total Followers.

It is the division between the sum of total retweets over the number of total followers.

- Topic count: How many topics did the user mention out of the trending topics already mined? It is calculated by the count of distinct keywords retrieved.

User	AGE	Total Tweets	AVG Tweets	Median	Median/Average	Replies%	Share%	Organic Content	Tweets	Following	Followers	Followers/Following	Retweets	Retweet AVG	Retweets/ Followers	Topic
sawtlebnan	3,096	291,621	94	105	11%	0%	12%	88%	3,213	1,229	61,251	4,984%	19,178	6	0.01%	13
LFofficalpage	3,610	677,227	188	119	37%	0%	0%	100%	3,206	11	164,073	1,491,573%	6,750	2	0.00%	12
Annahar	3,185	481,634	151	203	34%	0%	4%	96%	3,207	714	811,228	113,617%	4,941	2	0.00%	12
Mulhak	3,234	368,276	114	166	45%	1%	1%	99%	3,228	436	85,474	19,604%	2,408	1	0.00%	11
ALJADEEDNEWS	3,755	393,182	105	83	21%	0%	1%	99%	3,226	379	1,714,523	452,381%	16,337	5	0.00%	11
LarissaAounSky	3,142	29,606	9	28	197%	20%	20%	60%	3,245	2,114	32,906	1,557%	204,990	63	0.19%	9
AlHadath	3,493	219,048	63	188	200%	0%	0%	100%	3,219	43	8,674,292	20,172,772%	64,535	20	0.00%	8
EloMled	41	203	5	5	1%	48%	3%	49%	204	1,681	984	59%	709	3	0.35%	8
lebanondebate	3,688	758,439	206	228	11%	0%	2%	98%	3,201	5,968	478,387	8,016%	3,035	1	0.00%	8

Figure 19: Sample of the extracted variables via QlikView

Following the variables extraction, the extracted data is ready to move to the classification process. A thorough analysis of the data showed several groups that exhibit the same patterns of users.

4.2.2.3 Step 3 – Set user behavior conditions

Some of the variables were linked to understanding the types of users on Twitter which makes possible the analysis and detection of propaganda on Twitter [14,28]. For example, time and constant posting are two main patterns for bots [14]. Analyzing the current data

showed that there are no automated bots in Lebanon or targeting the Lebanon trending keywords. Taking a closer look at the date fields and the time of posts and different patterns shows that all the users take normal breaks in the day and have more than 5-hour breaks at night.

The analysis and identification of 8 types of users is based on the available behavioral and content variables. For some users, they are identical with one or two variables while for others, the whole constitution of variables is different.

Based on the research for understanding the types of users on twitter [28], the discussion of the types which were approached was based on a business perspective and were adapted to the political one.

- **Normal user:** It is considered the user that uses twitter to receive news and follow some interests in an average manner. Such a user shows low patterns in tweeting and most of their tweets are organic. There is a low average in retweets or replies. Followers and following ratio is very close to 1 as most followers are friends or part of their circle hence they will tend to follow them back.
- **Leader:** The leader is a normal person who already has leader patterns in real life which are reflected online via the presence of a huge number of followers, and people tend to share what he/she says a lot, and normally all his/her content is organic with no retweets or replies from his account.
- **Online Leader:** An online leader is only present in the digital world, where his/her trades are present and only established in the online world. He/she has an average follower background that tends to be high. His/her online presence is built over time so his/her account was created some time ago. The tendency to tweet, share and reply is high and contradicts the tendency of a normal leader who posts a little and only on specific events.
- **Media:** Similarly to leader, media is present in the offline world and shows the same patterns of a normal media outlet on TV. It is mostly a reflection of the offline world in the digital world. It shows a daily constant number of tweets with a low median/avgpost. Most of the content is organic with a very low, if not null, number in retweets and replies.

- **Online Media:** an online media outlet is only present in the digital world. Such category shows the same traits as an online leader's presence which shows a high average in daily content posting, and people tend to share their content more than the media category. They show a high ratio in Retweets/ followers where their followers are considered average comparing to the media outlets.
- **Backup:** A backup or spammer account is an account close to the normal user but with a very high reply rate compared to the very low organic content with a higher daily tweets average. A backup is event triggered and most of the replies are to back a leader, online leader, or a certain campaign which would be the trigger in this case.
- **Follower/ Feeder:** This category is close to a normal account and a backup but the main difference is the number of retweets or sharing patterns of content which are very high for this category. Such behavior is identical to that of people transferring information without even looking at it. They blindly trust the source or person to the extent that they share and trust it without looking and checking its content.
- **PRO:** Pro as in professional account is an advanced form of a normal account with all its variables boosted. It is the next level of a backup or follower where the users knows what they are doing without being detected as bots or spammers. They keep their posts organic, and their replying and sharing patterns are almost identical to a much higher follower/ following ratio.

For some users, categories are very close, and some users can be classified in more than one, but their behavior will always lead to only one at the end where one behavior overshadows the other.

4.2.2.4 Step 4 – User behavior analysis

The conditions were inserted into QlikView and followed a basic equation of 0 and 1 approach as a result. If a user matches a condition of a category for one of the variables it will be counted as 1. The system will then count the 1s, and the user will be graded for each category separately.

The rules and conditions of each category were set into the code to meet the required conditions for each, which were set and followed based on the following brief.

	AGE	Total Tweets	AVG Tweets	Median	AVG-Median	Replies%	Share%	
MAX	4,138	219,048	503	883	88%	97%	98%	
Min	32	60	0	1	-2465%	0%	0%	
AVG	1,892	26,446	26	33	-133%	36%	29%	
Median	2,003	12,638	8	15	-37%	34%	20%	

	Organic	Tweets	Following	Followers	Followers/Following	Retweets	Retweets AVG	Retweets/ Followers
MAX	100%	6,462	13,846	19,628,203	29946000%	7,333,183	2,873	4466.98113%
Min	0%	0	0	6	10%	0	0	0.00023%
AVG	34%	2,833	1,764	353,603	648298%	292,979	123	52.65601%
Median	27%	3,193	978	7,622	520%	91,836	31	0.28442%

Figure 20: Variable results analysis

The code was created for each category in QlikView backed with one if-statement for each variable, and at the end of each category code, all the numbers of conditions were added to get a total result for each user.

	AGE	AVG Tweets	Median	AVG-Median	Replies%	Share%	Organic	Following	Followers	FLR/FLO	Retweets AVG	Retweets / Followers
MEDIA	>2500	>10	>10	<30%	<10%	<10%	>80%	<1500	>500K	>1000%	> 300	<0.00001
Online Media	>1500	>15	>10	>10%	<10%	<10%	>70%	>500	20K< and <500K	>1000%	<50	>0.00001
Leader	>1825	<5	<7	<20%	<15%	<20%	>90%	<500	>100000 =2	>1000%	> 50	<0.001
Online Leader	>2500	>10	>10	>20%	>10%	>30%	<50%	<10K	>20000	>1000%	> 200	>0.0001
Super Follower	>90	>30	>12	>40%	<20%	>50%	<20%	<1500	<1500	>10%	>50	<0.0001
Backup	>90	>90	>12	>60%	>80%	<20%	<10%	<1500	<1500	>10%	>50	<0.0001
Normal	> 90	<5	<7	<20%	<30%	<30%	>40%	<1500	<1500	>10%	<5	<0.1
PRO	>90	<150	<150	<5%	>10%	>10%	>40%	500> and <20000	500> and <20000	>10%	>7	>0.0045

Figure 21: Categories conditions for the variables

Each category followed the results in the table for each user. Some variables were assigned 2 instead of 1 to give more importance to the variable in a certain category. For example, the leader’s followers’ results could get 0, 1, or 2 depending on the number of followers. For better results in the conditions, some variables were set as mandatory for other variables to be true. These important variables for each category are placed in squares in the tables above. For example, a leader must have more than 100000 followers and the retweet/followers average should be less than 0.001.

	AGE	AVG Tweets	Median	Median/Average	Replies %	Share %	Organic Content	Followin g	VFollowers	Followers/ Following	Retweet AVG	Retweets/ Followers	LEADER Result
Gebran_Bassil	1	1	1	0	1	1	0	1	2	1	1	1	11
LBpresidency	1	1	1	0	1	1	0	1	2	1	1	1	11
samygemayel	1	1	1	0	1	0	0	1	2	1	1	1	10
nadimgemayel	1	1	1	1	1	1	0	1	2	1	0	0	10
majjdaelroumi	1	1	1	1	1	0	0	1	2	1	0	0	9
mayadiab	1	1	1	1	1	0	0	1	2	1	0	0	9
walidjoublatt	1	1	1	0	1	1	0	0	2	1	1	0	9
YazbekWehbe	1	1	1	0	0	1	0	0	2	1	1	1	9
PaulaYacoubian	1	0	1	0	1	0	0		2	1	1	1	8
MohamadAhwaze	1	0	0	1	0	1	0	0	2	1	1	1	8
Neshan	1	0	1	0	0	1	0	1	2	1	1	0	8

Figure 22: Leader variables grading results after the application of the conditions

The condition development was based on the testing and sampling of the data; the tweet categorization research was based on a business perspective and on the political analysis patterns over twitter [14,28].

4.2.2.5 Step 5 – Classify users

Each category was graded based on the conditions that were received from the categorization analysis; the users were classified in the categories based on the highest grade they got.

USER	ScreenName	LEADER	O Leader	MEDIA	BACKUP	SUPER	PRO	O MEDIA	CAT
	3bsonn	1	2	0	3	3	11	0	PRO
	4zahri	4	12	0	3	2	9	6	O LEADER
	70sul	6	10	0	5	4	6	8	O LEADER
	AAlhblany	4	6	0	3	5	8	6	PRO
	ABDALLAH_B2	3	3	0	5	3	10	0	PRO
	abeerliban	1	3	0	1	1	0	0	NORMAL
	aboadnan1020	1	2	0	3	7	7	0	SUPER...
	AboHadi90017277	0	2	0	2	2	11	0	PRO
	aboulfadol217	3	3	0	1	0	10	0	PRO
	AdhamMG	2	3	0	2	1	10	0	PRO
	AhmedMousa_1	5	8	0	0	2	8	5	O LEADER
	ahmedyassine30	1	1	0	6	4	10	0	PRO
	Ake29mailaubed1	1	2	0	2	5	1	0	NORMAL
	akel_I_lohoun	1	3	0	6	5	0	0	BACKUP...
	akhbar	5	6	3	2	1	0	1	NORMAL
	alaa_naserdine	1	2	0	2	3	7	0	NORMAL
	alaaeid177	1	2	0	3	6	9	0	PRO
	AlainHakim	2	3	0	1	2	8	0	PRO
	AlArabiya	5	6	3	2	1	0	1	NORMAL
	AlGhadTV	5	8	0	3	3	3	7	O LEADER

Figure 23: Grading of categories and the classification process

As shown in the table above, some users received the same grade for different categories; for example, the user *aboadnan1020* received 7 on two categories. In this case the module chooses to classify him based on which of the two categories met the conditions first. Following the categorization process, the Users' database is updated, and a column is added for the categorization variable.

4.3 Propaganda detection

Propaganda starts with an agent who sets the purpose of the propaganda, which is then transferred to a certain circle that produces and reproduces the content in many forms, and all is communicated to the receivers or the users, in this case, to push a certain idea/content [15,23], mislead them, or even change their beliefs and mindsets. The same steps are transferred and generated through Twitter, where the original agent who sets the target and main content is not apparent unlike all the sub-agents who created, shared, and tweeted for a certain campaign. This campaign is boosted till it reaches every possible receiver, and this automatically shows on Twitter in the trending keywords. These represent the behaviors of an agent or a group of agent pushing for a certain propaganda in a form of a keyword.

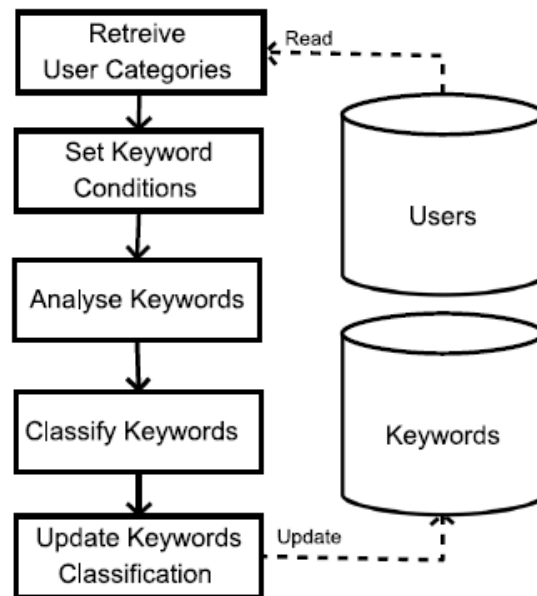


Figure 24: Final phase, propaganda detection

The PRO, backup, and follower categories in the module have almost the same traits and behaviors, and they evolve showing team work for generating and boosting a certain keyword. The three show very high activity based on a certain trigger in a certain timeframe. In the case of the discussed #drone, they showed a very high correlation in their activity triggered by the keyword which translated into the boost of reach of the content, in other words propaganda.

ScreenName	LEADER	O Leader	MEDIA	BACKUP	SUPER	PRO	O MEDIA	CAT
Ake29mailaubed1	1	2	0	2	5	1	0	NORMAL
AmmarKouchak	1	2	0	1	1	0	0	NORMAL
BahaaHajir2	1	2	0	1	1	1	0	NORMAL
dana_nasr_pen	1	1	0	0	1	5	0	NORMAL
DineJihane	1	2	0	2	5	1	0	NORMAL
kordab_karen	2	2	0	1	4	1	0	NORMAL
Irmouawad	1	2	0	1	1	1	0	NORMAL
MarwaMakarem3	1	2	0	0	1	0	0	NORMAL
Sarahm555	1	2	0	1	1	0	0	NORMAL
Tala90887921	2	2	0	1	0	1	0	NORMAL

Figure 25: Normal campaign categorized users keyword #fnflebanon

These categories play the main role in the propaganda creation, hence they will appear as the top users who are tweeting the propaganda. Hence the module will detect propaganda based on the categories of its top users. The module will be based on 5 simple layers leading to the detection of propaganda.

This phase will start by reading from the users' database: all the users that showed activity on a certain keyword will be categorized and filtered based on the keyword.

	AGE	Total Tweets	AVG Tweets	Median	Median/Average	Replies %	Share %	Organic Content	Tweets	Following	Followers	Followers/Following	Retweets	Retweet AVG	Retweets/Followers	Category
x7PskDFwybB9gu7	70	6,699	95	31	67%	2%	93%	6%	3,227	834	562	67%	93,059	29	5.13%	PRO
AminInaya	217	915	4	14	232%	5%	79%	16%	1,792	1,395	833	60%	58,342	33	3.91%	PRO
AboHadi90017277	306	4,241	14	5	64%	20%	70%	10%	3,193	617	1,009	164%	120,687	38	3.75%	PRO
alaaeid177	216	14,854	69	42	39%	5%	90%	5%	3,090	1,037	1,915	185%	198,805	64	3.36%	PRO
Bilalabbasgh	219	873	4	4	12%	15%	39%	46%	814	1,909	2,362	124%	27,793	34	1.45%	PRO
Hussein95310181	94	16,682	178	173	3%	97%	0%	3%	3,244	2,098	2,539	121%	7,168	2	0.09%	BACKUP Follower
Rawan23493	138	13,062	94	99	5%	77%	13%	10%	3,191	3,133	3,128	100%	89,029	28	0.89%	PRO
Hassank198	296	145,838	492	883	79%	1%	98%	0%	3,189	1,124	3,572	318%	208,928	66	1.83%	PRO
mourad_alii	1,017	32,935	32	141	335%	86%	4%	9%	3,236	3,534	3,889	110%	11,750	4	0.09%	PRO
housaini82	3,064	6,892	225%	750%	233%	41%	0	0	2,966	0%	4,100	-	57,293	1,932%	0%	PRO
alishoeib1970	2,123	28,463	1,341%	2,000%	49%	50%	0	0	6,389	142,100%	153,290	108	326,992	5,118%	0%	O LEADER

Figure 26: Users retrieval based on the keyword #مقبره_الميركافا

As seen in the sample for the keyword #مقبره_الميركافا 22 users appeared, and out of these, 3 are not categorized as PRO users: 1 is a backup, 1 online media, and one online leader.

Based on conditions that have been set to the module, the sum of PRO, followers, and backup users must be more than 70% out of the detected users for a keyword to be classified as propaganda. This number appeared after analyzing all the keywords, and most of the propaganda keywords that were analyzed contained an online leader as well. In other words, the tendency of these 3 categories to have very high activity on a certain keyword will automatically make it a boosted keyword or campaign that aims to have a certain target.

The analysis of the keywords from Lebanon as location, where twitter is mostly used for political perspectives, was as follows: more than 90% of the keywords were positive for

propaganda, 8% returned with high media and online users, and only 2% returned of normal users.

Following the classification of the keyword based on the conditions set, the module will update the keywords database and will use as metadata the time of the classification. The time of the classification is essential as some keywords could be repeated with different intentions or background.

4.4 Testing and validation

The test data included 5 keywords and was extracted on the 31st of August 2020. 25 users were extracted as top users and out of those 25, few were already classified as they appeared in the main data keywords.

To further validate the data, the extracted users' background was manually checked. Manual validation was feasible for all the leaders and media; as for the remaining types of categories, they were validated by their variables.

4.4.1 Test sample

The test sample included 5 keywords, 25 new users, and 60000 tweets. It was able to predict 3 propaganda campaigns with more than 70% related to PRO, Backup, and follower users. 2 normal campaigns, with only around 25% propaganda users.

As a test sample, the keyword Feyrouz was used, and the users were analyzed. Feyrouz is a known Lebanese singer which explains the presence of a lot of MEDIA and OMEDIA accounts tweeting about it. Moreover, the background of the OLeaders most of whom work in media outlets and who created their online presence in parallel with their offline presence on TV or other platforms was checked.

	AGE	Total Tweets	AVG Tweets	Median	Median/Average	Replies %	Share %	Organic Content	Tweets	Following	Followers	Followers/Following	Retweets	Retweet AVG	Retweets/ Followers	Category
marysaadeh19	-	-	-	-	-	49%	27%	24%	6,444	-	-	-	462,410	72	-	NORMAL
Nour50006572	277	2,743	10	15	52%	9%	5%	86%	2,732	200	43	22%	21,501	8	18.30%	NORMAL
elie_freyha	17	98	6	4	40%	12%	43%	44%	97	173	45	26%	271	3	6.21%	NORMAL
nadafa_g	572	4,713	8	9	9%	42%	13%	45%	3,230	745	262	35%	49,874	15	5.89%	PRO
NabilaHammami	1,965	5,089	3	47	1,715%	27%	22%	50%	3,170	185	379	205%	76,123	24	6.34%	PRO
Eliea112	485	2,610	5	5	7%	32%	17%	51%	2,594	530	502	95%	692,381	267	53.17%	NORMAL
YounanWaddah	246	8,628	35	63	78%	80%	10%	10%	3,225	811	568	70%	28,745	9	1.57%	PRO
ElioMiled	41	203	5	5	1%	48%	3%	49%	204	1,681	984	59%	709	3	0.35%	PRO
ABDALLAH_B2	2,555	20,185	8	30	280%	75%	12%	14%	3,240	347	2,561	738%	352,992	109	4.25%	PRO
SAOUD1st	1,744	23,756	14	42	205%	90%	1%	9%	3,214	978	5,358	548%	8,757	3	0.05%	PRO
AmalNadhereen	4,131	324,713	79	45	43%	28%	30%	42%	3,201	84	22,323	26,575%	84,561	26	0.12%	O LEADER
LarissaAounSky	3,142	29,606	9	28	197%	20%	20%	60%	3,245	2,114	32,906	1,557%	204,990	63	0.19%	O LEADER
SufianSamarrai	2,639	26,879	10	20	96%	23%	36%	41%	2,994	4,688	80,036	1,707%	1,812,766	605	0.76%	O LEADER

Figure 27: User categorization on the keyword Feyrouz

It is worth noting that the normal campaigns included a lot of media categorized users which means that both keywords were organically trending or based on an event happening in Lebanon on that date.

Keywords	Propaganda users	Prediction
Feyrouz	23%	Normal
Macron	24%	Normal
بيت بيك سيد عون	73%	Propaganda
امام العيش المشترك	75%	Propaganda
قاطعوا قنوات الفتن	100%	Propaganda

Figure 28: Sample keywords with predictions

The validation of the users that are related to media and leaders was successful, hence the validation of the module and code was also successful. A high number of users recurrence was seen in the data number categorized in the main data set, hence the more keywords and users extracted and categorized, the more the recurrence of data.

4.4.2 Module Adaptation

The module was built based on behavioral variables of the users in Lebanon. the conditions that were applied are established based on the numbers and analytical extractions of the database that was collected in Lebanon. Users' digital footprints and behaviors differ by a percentage from one county to another. The conditions were built based on the Lebanese

market whose population was estimated at 6.8 Million (worldmeters.info) of whom more than 50% penetrated Facebook in 2020 compared to not more than 5% of new Twitter users based on “medialandscapes.org/country/Lebanon”.

The behaviors and patterns for each category do not differ from one country to another: a leader will always have a low average in tweeting whereas an online leader will have higher patterns in tweeting. Whether these leaders are in Lebanon, Europe, or another country in the Middle East, these patterns will not differ. Compared to other countries, Lebanon is a small country with a population of 6.7million people only. France’s population, for example, in the year 2020, was around 65 million with a much higher number of 35% of new Twitter users.

The two main variables that affect the conditions are the penetration of social platforms for the adapted country, knowing that penetration is based on the number of the population. Another main factor is the use of Twitter for each country. In Lebanon Twitter is mainly used for politics while in France its main use is for business and the secondary one for politics and in the UAE it is strictly used for business.

A comparative approach was applied to study the top leaders in Lebanon, France and the UAE. Since the research is based on a political aspect, the followers for the presidential office and their personal accounts were retrieved for every country. An important point to take into consideration, is that the UAE is a monarchy hence the President’s personal account is used as the office account also.

		Twitter Penetration (A)	Population that uses Twitter (B)	Followers (C)	Ratio Followers (C)/ Twitter users (B)	Factor of adaptation
Lebanon	Lebanese Presidency	3%	201328	300000	149%	1
	Michel Aoun			357000	177%	1
France	Élysée	21%	13497242	2500000	19%	8
	Emmanuel Macron			5700000	42%	16
UAE	Mohamed Bin Zayed	53%	5241913	10300000	196%	29

Figure 29: Adaptation table for France and UAE

The adaptation factor is based on 4 variables:

- Percentage of Twitter penetration in the country

- Number of Twitter users from a specific population, which is twitter penetration x total population of the country.
- Account followers
- Ratio of Followers/ Twitter users for this county, to have an idea of the account popularity and its importance in the culture of the country they belong to
- The factor of adaptation is the division of the number of followers of the leader account of the new country over the base case for the same category which in this case is Lebanon. The personal accounts of the leaders, the presidents of the countries, in this case are divided by the personal account of the Lebanese presidents and same for the office accounts.

As seen for France, the average gfactor of adaptation is 11, so the conditions of the categorization module must be multiplied by this factor for them to work in France. As for the UAE, the factor of adaptation is 29, so the module must be multiplied by 29 in order for it to work.

4.5 Findings and discussion

The survey results showed that females have higher patterns in sharing the news with an average ratio for checking its authenticity; moreover, the respondents showed low interest in using Twitter for information and news, and there was a low percentage of respondents who refer to some offline media outlets which are related to a certain political party. Finally, the deduction from the survey that more than 70% of the participants understand the reason behind the communicated fake news is not reflected on Twitter as seen in the graph below: few participants have identified fake news on Twitter, and even those who have identified high percentages of fake news were not aware of their purpose, hence their low knowledge patterns.

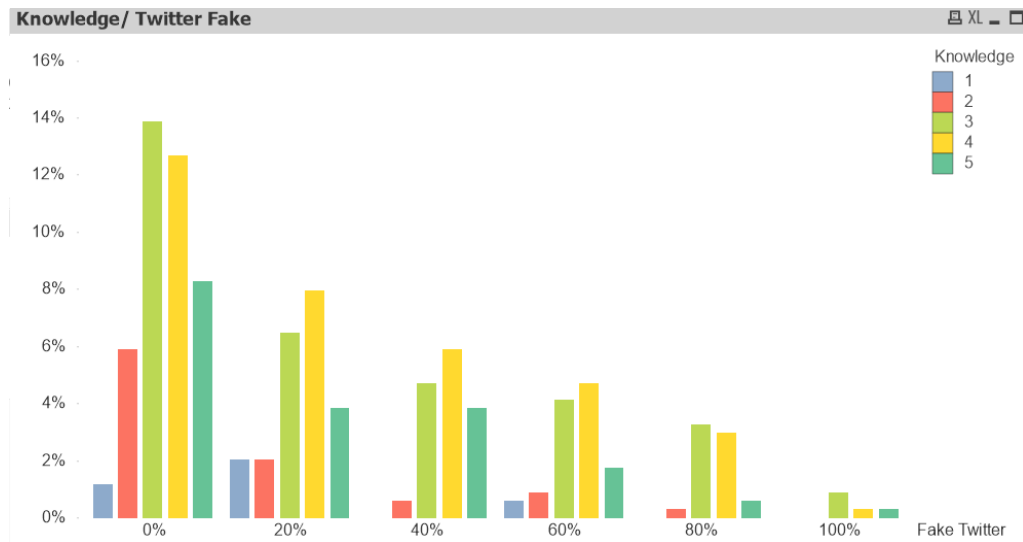


Figure 30: Questionnaire answers of Knowledge/ Twitter detected fake

In the analytical part of the module, the entities that have very high correlation patterns based on certain keywords were detected with a correlation average that was as high as 0.9881. This finding proves the presence of entities pushing and boosting for certain political agendas based on the keywords. This kind of entities proved the presence of such campaigns that could be used to start or boost propaganda.

The categorization module was developed in the political perspective to understand the behavior of the 3 categories that form the buildup pushing propaganda inorganically. Those categories are triggered by certain events and work in very organized and controlled groups. Out of the 184 users that were analyzed, 50% were PRO, Backup, or super followers while normal users were only 17%. The presence of high numbers of propaganda agents and the low presence of normal users based on the trending keywords can only be explained by the number of Lebanese respondents who use Twitter for information. Twitter in Lebanon is mainly used for political purposes.

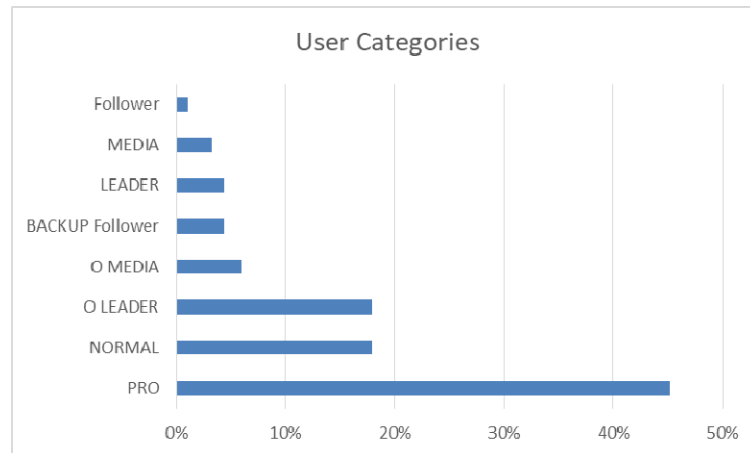


Figure 31: User Categorization

As for the propaganda prediction module, the analysis was done for 36 keywords. The results were discussed under 3 categories: organic, being the first category, shows the normal trending keywords and whose top users are mostly normal users; the second, categorized as propaganda, is pushed by the 3 agents that form propaganda; and the last category is the media which is a mix between media, leaders, and propaganda users in all their forms. This category can be a normal campaign or news that underwent a lot of discussion on the media outlets or could be propaganda that garnered a lot of noise even among online and offline outlets and leaders. The last two categories are pushing information inorganically and are formed by agencies and other entities for political gain.

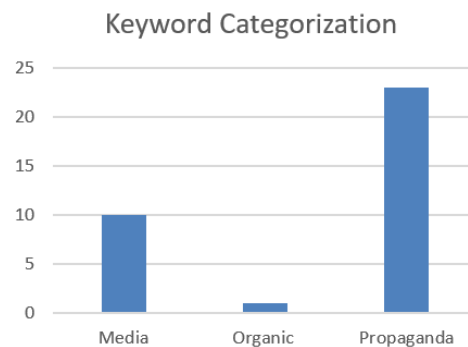


Figure 32: Keywords Summary

As a conclusion, more than 95% of the trending keywords online on Twitter in Lebanon is not organic. Around 70% of the gathered data are all propaganda in its normal form and pushed by entities for political purposes and gain. Hence more than 95% of the campaigns are not organic which shows that Twitter is used with purely controlled content only.

Chapter 5: Conclusion and future works

The world has faced propaganda in all its forms since the start of offline communications. Communications can take different forms; it is mainly the exchange of information between 2 entities. Communications started offline with the TV; it then evolved into the digital form we know today. The exchange of information is evolving exponentially with the advancement in digital platforms, hence people are receiving and sending information in constant and very rapid ways on the digital. Information is not being checked, verified or authenticated because of the amounts and speed of information being shared. As a result, Information disorder has reached forms and results that have affected the world as we know it. For example, information disorder over the digital platforms was able to affect the United States presidential elections in the propaganda campaign managed by Cambridge Analytica.

The research undertook information disorder in its different phases, forms, and elements. It explained and elaborated them in order to try and find a solution to predict and understand propaganda before it happens. Moreover, the human aspect was added to understand how people act and react in such circumstances. A study was done in Lebanon via a questionnaire form targeting information disorder and propaganda more deeply. A research on propaganda and information disorder predictive models was also completed. The modules found were classified according to the phases of information disorder that the predictive model was based on. Given all the modules that were previously developed, a common factor among all of them was identified. The user or human factor was found to be the main contributor and agent.

5.1 Main Contributions and results of the Thesis

The focus of the module started with the study of the human digital footprint in propaganda. The data extraction was done via shared APIs on Twitter, the main news sharing platform among the digital platforms. The database was built with over 500,000 tweets from 150 users. The users' engagement and behaviors were studied and analyzed.

The analysis led to the classification and categorizations of these users over Twitter. Based on 17 calculated variables, the module was able to categorize all the top users that are discussing the trending keywords which detect propaganda. In fact, understanding the inorganic push of certain keywords leads to propaganda, and identifying those keywords leads to predicting propaganda in its very early stages.

The extraction of 3 sub-categories from the 8 categorized users was concluded. The first sub-category, which was the main focus of the research, is the propaganda related users. The second sub-category was all media related outlets, and finally, the third sub-category was the normal users.

The propaganda sub-category includes users that worked in closed groups, in a professional and event-triggered manner. As for the media sub-category, it is formed by media agencies and agents and shared news in the same way it was shared on the offline platforms. Finally, the normal users sub-category includes the users that had normal patterns and did not show any immoral behaviors. The normal users sub-category included organic users or leaders that are reflected in the digital.

As for the adaptation, the application of the module was developed in Lebanon, but a simple adaptation of ratio for the conditions makes it fully adaptable and applied to any country. Alternatively, the use of Twitter in Lebanon, mainly for political campaigning, was proven with the high presence of PRO accounts. In other countries, those PRO accounts will show the same digital footprint, hence digital behaviors, but with purposes other than political.

5.2 Possible Extensions and Future Work

This work can be extended by further exploring sentiment analysis through text analysis of the user content shared in Arabic.

Additionally, the classification module of propaganda agents can be evolved in terms of identification to which groups these agents belong, through a group analysis based on interactions between the users.

No factors concluded the presence of digital bots in the research. A deeper research will be conducted based on content sharing linking the time factor and events. This method will allow linking the grouping mechanism more efficiently, and this will lead into network analysis.

To have a clearer study of the online shared content, relating Facebook and Twitter users over related content between the two platforms will be looked into in order to extend the module and the research.

Bibliography

- [1] Esma Aïmeur, Nicolás Díaz Ferreyra, and Hicham Hage. 2019. Manipulation and Malicious Personalization: Exploring the Self-Disclosure Biases Exploited by Deceptive Attackers on Social Media. *Front. Artif. Intell.* 2, (2019). DOI:<https://doi.org/10.3389/frai.2019.00026>
- [2] Galen Bodenhausen, S.K. Kang, and D. Peery. 2012. Social categorization and the perception of social groups. . 311–329. DOI:<https://doi.org/10.4135/9781446247631.n16>
- [3] Silvia Covacio. 2003. Misinformation: Understanding the Evolution of Deception. DOI:<https://doi.org/10.28945/2656>
- [4] Dominic Difranzo and Kristine Gloria-Garcia. 2017. Filter bubbles and fake news. *XRDS: Crossroads, The ACM Magazine for Students* 23, (April 2017), 32–35. DOI:<https://doi.org/10.1145/3055153>
- [5] Johan Farkas and Marco Bastos. 2018. IRA Propaganda on Twitter: Stoking Antagonism and Tweeting Local News. In *Proceedings of the 9th International Conference on Social Media and Society (SMSociety '18)*, Association for Computing Machinery, New York, NY, USA, 281–285. DOI:<https://doi.org/10.1145/3217804.3217929>
- [6] Zhen Guo, Jin-Hee Cho, Ing-Ray Chen, Srijan Sengupta, Michin Hong, and Tanushree Mitra. 2020. Online Social Deception and Its Countermeasures for Trustworthy Cyberspace: A Survey. *arXiv:2004.07678 [cs]* (April 2020). Retrieved June 3, 2020 from <http://arxiv.org/abs/2004.07678>
- [7] Hicham Hage, Esma Aïmeur, and Amel Guedidi. 2020. Understanding the Landscape of Online Deception. *Navigating Fake News, Alternative Facts, and Misinformation in a Post-Truth World*, 290–317. DOI:<https://doi.org/10.4018/978-1-7998-2543-2.ch014>
- [8] Claire Wardle Hossein Derakhshan. 2017. INFORMATION DISORDER : Toward an interdisciplinary framework for research and policy making. Retrieved April 28, 2020 from <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c>
- [9] Jan De Houwer. Similarities and Differences. 48.
- [10] Ben Ltaifa Ibtihel, Hlaoua Lobna, and Ben Jemaa Maher. 2018. A Semantic Approach for Tweet Categorization. *Procedia Computer Science* 126, (January 2018), 335–344. DOI:<https://doi.org/10.1016/j.procs.2018.07.267>
- [11] J. Isaak and M. J. Hanna. 2018. User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer* 51, 8 (August 2018), 56–59. DOI:<https://doi.org/10.1109/MC.2018.3191268>
- [12] Marc Jones. 2019. Propaganda, Fake News, and Fake Trends: The Weaponization of Twitter Bots in the Gulf Crisis. *International Journal of Communication* 13, (March 2019), 1389–1415.

- [13] Alireza Karduni. 2019. Human-Misinformation interaction: Understanding the interdisciplinary approach needed to computationally combat false information. *arXiv:1903.07136 [cs]* (March 2019). Retrieved April 9, 2020 from <http://arxiv.org/abs/1903.07136>
- [14] Ansgar Kellner, Lisa Rangosch, Christian Wressnegger, and Konrad Rieck. 2019. Political Elections Under (Social) Fire? Analysis and Detection of Propaganda on Twitter. *arXiv:1912.04143 [cs]* (December 2019). Retrieved November 22, 2020 from <http://arxiv.org/abs/1912.04143>
- [15] Ansgar Kellner, Christian Wressnegger, and Konrad Rieck. 2020. What’s all that noise: analysis and detection of propaganda on Twitter. In *Proceedings of the 13th European workshop on Systems Security*, ACM, Heraklion Greece, 25–30. DOI:<https://doi.org/10.1145/3380786.3391399>
- [16] Chikaho Kurahashi, Tadanobu Misawa, and Kazuya Yamashita. 2018. Evaluation of Online Advertisement Design Using Near-infrared Spectroscopy. *Sensors and Materials* 30, 7 (July 2018), 1487. DOI:<https://doi.org/10.18494/SAM.2018.1879>
- [17] Quanzhi Li, Xiaomo Liu, Rui Fang, Armineh Nourbakhsh, and Sameena Shah. 2016. User Behaviors in Newsworthy Rumors: A Case Study of Twitter.
- [18] Marcelo Maia and Virgilio Almeida. 2008. Identifying User Behavior in Online Social Networks. In *Proceedings of the 1st Workshop on Social Network Systems*. DOI:<https://doi.org/10.1145/1435497.1435498>
- [19] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proc Natl Acad Sci USA* 114, 48 (November 2017), 12714. DOI:<https://doi.org/10.1073/pnas.1710966114>
- [20] Tsikerdekis Michail and Zeadally Sherali. 2020. Detecting Online Content Deception. (April 2020). Retrieved from <https://ieeexplore.ieee.org/abstract/document/9049293>
- [21] Kaan Onarlioglu, Utku Ozan Yilmaz, and Engin Kirda. Insights into User Behavior in Dealing with Internet Attacks. 14.
- [22] Jukka Pietiläinen. Diffusion of the News Paradigm 1850-2000. 15.
- [23] Philip N. Howard Samuel C. Woolley. 2017. Computational Propaganda Research Project. Retrieved April 9, 2020 from https://ora.ox.ac.uk/objects/uuid:d6157461-aefd-48ffa9a9-2d93222a9bfd/download_file?file_format=pdf&safe_filename=Casestudies-ExecutiveSummary.pdf&type_of_work=Record
- [24] Ben Shneiderman. 2003. Promoting Universal Usability with Multi-Layer Interface Design. Retrieved April 9, 2020 from <http://citeseerx.ist.psu.edu/viewdoc/citations?doi=10.1.1.202.7069>
- [25] Shinji Teraji. 2003. Herd behavior and the quality of opinions. *The Journal of Socio-Economics* 32, 6 (December 2003), 661–673. DOI:<https://doi.org/10.1016/j.socec.2003.10.004>
- [26] Master’s Thesis, Margareta Melin, and Ilkin Mehrabov. Fake news and Social Media. 55.
- [27] Michail Tsikerdekis and Sherali Zeadally. 2014. Online deception in social media. *Commun. ACM* 57, 9 (September 2014), 72–80. DOI:<https://doi.org/10.1145/2629612>
- [28] Muhammad Moeen Uddin, Muhammad Imran, and Hassan Sajjad. 2014. Understanding Types of Users on Twitter. *arXiv:1406.1335 [cs]* (June 2014). Retrieved November 22, 2020 from <http://arxiv.org/abs/1406.1335>

[29] Andrew Whiten and Richard W. Byrne (Eds.). 1997. *Machiavellian Intelligence II: Extensions and Evaluations*. Cambridge University Press, Cambridge.

DOI:<https://doi.org/10.1017/CBO9780511525636>

[30] Propaganda: The Formation Of Men's Attitudes By Jacques Ellul. Retrieved November 22, 2020 from <http://www.whenthebeststops.org/2017/04/propaganda-formation-of-mens-attitudes.html>

[31] Lying and Deception: Theory and Practice, by Thomas L. Carson. | Mind | Oxford Academic. Retrieved February 17, 2021 from <https://academic.oup.com/mind/article-abstract/120/480/1232/955846>

Appendix A: Questionnaire