HUSO 2020

The Sixth International Conference on Human and Social Analytics

ISBN: 978-1-61208-800-6

October 18 – 22, 2020

HUSO 2020 Editors

Nitin Agarwal, University of Arkansas – Little Rock, USA
Bourret Christian, University Gustave Eiffel /Paris East Marne la Vallée, France

HUSO 2020

Foreword

The Sixth International Conference on Human and Social Analytics (HUSO 2020), held between October 18–22, 2020 continued the inaugural event bridging the concepts and the communities dealing with emotion-driven systems, sentiment analysis, personalized analytics, social human analytics, and social computing.

The recent development of social networks, numerous ad hoc interest-based formed virtual communities, and citizen-driven institutional initiatives raise a series of new challenges in considering human behavior, both on personal and collective contexts.

There is a great possibility to capture particular and general public opinions, allowing individual or collective behavioral predictions. This also raises many challenges, on capturing, interpreting and representing such behavioral aspects. While scientific communities face now new paradigms, such as designing emotion-driven systems, dynamicity of social networks, and integrating personalized data with public knowledge bases, the business world looks for marketing and financial prediction.

We take here the opportunity to warmly thank all the members of the HUSO 2020 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to HUSO 2020. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the HUSO 2020 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that HUSO 2020 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the area of human and social analytics.

HUSO 2020 Chairs

HUSO 2020 Steering Committee

Els Lefever, Ghent University, Belgium
Dennis J. Folds, Lowell Scientific Enterprises (LSE), USA
Nitin Agarwal, University of Arkansas at Little Rock, USA
Christian Bourret, University of Paris East - Marne la Vallée (UPEM), France
Baden Hughes, Cognitiviti, Australia

HUSO 2020 Publicity Chair

Jose M. Jimenez, Universitat Politecnica de Valencia, Spain
Jose Luis García, Universitat Politecnica de Valencia, Spain

HUSO 2020 Industry/Research Advisory Committee

Yi Shan, SDE II, Electronic Arts, Seattle, USA
Fan Yang, eBay Inc., USA
Massimiliano Zanin, The Innaxis Foundation & Research Institute, Madrid, Spain
Zhiwen Fang, Microsoft, USA
Xiaolong Jin, 靳小龙, Chinese Academy of Sciences, China

HUSO 2020

Committee

HUSO 2020 Steering Committee

Christian Bourret, University of Paris East - Marne la Vallée (UPEM), France
Els Lefever, Ghent University, Belgium
Dennis J. Folds, Lowell Scientific Enterprises (LSE), USA
Nitin Agarwal, University of Arkansas at Little Rock, USA
Baden Hughes, Cognitiviti, Australia

HUSO 2020 Publicity Chair

Jose M. Jimenez, Universitat Politecnica de Valencia, Spain
Jose Luis García, Universitat Politecnica de Valencia, Spain

HUSO 2020 Industry/Research Advisory Committee

Yi Shan, SDE II, Electronic Arts, Seattle, USA
Fan Yang, eBay Inc., USA
Massimiliano Zanin, The Innaxis Foundation & Research Institute, Madrid, Spain
Zhiwen Fang, Microsoft, USA
Xiaolong Jin, 靳小龙, Chinese Academy of Sciences, China

HUSO 2020 Technical Program Committee

Paul Abbiati, Founding Fellow of the EUROPEAN LAW INSTITUTE, Austria
Rodrigo Agerri, University of the Basque Country UPV/EHU, Spain
Harry Agius, Brunel University London, UK
Hafizi Muhamad Ali, Yanbu University College, Saudi Arabia
Balbir Barn, Middlesex University, London, UK
Chidansh Amitkumar Bhatt, FX Palo Alto Laboratory Inc., USA
Christian Bourret, University of Paris East - Marne la Vallée (UPEM), France
Dickson Chiu, The University of Hong Kong, Hong Kong
Alexandra I. Cristea, University of Durham, UK
Chen Ding, Ryerson University, Canada
Birgitta Dresp-Langley, Centre National de la Recherche Scientifique, France
Thierry Edoh, RFW-Universität Bonn, Germany
Silvia Florea, Lucian Blaga University of Sibiu,Romania
Dennis J. Folds, Lowell Scientific Enterprises (LSE), USA
Denis Gracanin, Virginia Tech, USA
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Baden Hughes, Cognitiviti, Australia
Emilio Insfran, Universitat Politecnica de Valencia, Spain

Copyright Information

Table of Contents

Rebuilding Trust to Develop the Resilience of Weakened Territories in France

Elements for a New Approach to Territorial Intelligence concerning Information and Communication Stakes in a Context of Digital Transformation and Globalization

Christian Bourret
Dicen-IdF Research Team
Université Gustave Eiffel (Paris Est Marne- la-Vallée)
Serris, France
christian.bourret@u-pem.fr

***Abstract*—In an international context of globalization and digital transformation, social ties are in crisis, particularly in France. In an approach to Territorial Intelligence concerning information and communication stakes, we propose some ways to try to (re)build trust to promote resilience and sustainable development of weakened territories in France. This rebuilding of trust can be achieved through projects to develop a collective representation for shared understanding in communities of knowledge, associating all the actors and with a new role in public services. We will put forward an application of this approach in three sectors of activity: cultural tourism, local businesses, healthcare and social protection. We will consider the application of this approach to a specific territory (the 'Couserans'), in the heart of the French Pyrenees, whose specificities and opportunities of resilience we will present.**

Keywords - Information; Communication; Transformation; Society; Trust; Weakened Territories; France.

I. INTRODUCTION

Twenty years ago, A. Giddens outlined our « Runaway world » or « how globalization is reshaping our lives » [1]. This transformation is provoking increasingly strong reactions from those who feel they are victims of the changes.

In France, these consequences are very important in the weakened territories. The main topic of this paper is to show how rebuilding trust may help to develop the resilience of weakened territories in France, especially with cooperations through knowledge sharing to promote collective intelligence.

The issue of inequalities between territories is an old problem in France. It has taken on a new dimension with the crisis of social ties, in a global context of globalization and digital transformation [2][3]. In an approach to Territorial Intelligence centered on information and communication stakes, we propose some ways to promote the sustainable development of weakened territories insisting on a dynamic of trust based on shared projects, to (re) make society.

After an introduction, first of all, we will present the context of this work, specifying our approach to weakened territories. In a second step, we will present our scientific positioning and our research methodology in the French interdisciplinary field of information and communication sciences and our new approach to territorial intelligence in a constructivist approach with the aim of building knowledge communities to promote the attractiveness and sustainable development of these territories. Third, we will focus our approach on three sectors (local companies, cultural tourism, healthcare and social protection) and we will apply it to the 'Couserans' territory (French Pyrenees). Fourth, we will specify our territorial intelligence approach, based on interactions with and between all the inhabitants and actors. Furthermore, we will propose elements to develop a contributory intelligence around a synergy of projects. In addition, we will advocate for a new role of public services: State, local authorities and social protection organizations. We will end with a conclusion.

II. PURPOSE OF RESEARCH – SCIENTIFIC POSITIONING – METHODOLOGY

In this section, we first present our purpose of research about the specificities of weakened territories in France, then our scientific positioning and our methodology.

A. Weakened Territories in France

The issue of territorial inequalities is a long-standing one in France. In 1947, Gravier, in a book of great resonance, spoke of "Paris and the French desert" [2]. From a voluntarist perspective, the DATAR (*Délégation à l'Aménagement du Territoire et à l'Action Régionale*) was created in 1963, under De Gaulle's presidence.

The problem of weakened territories suddenly became apparent to the general public in France with the riots in the suburbs of large cities in the autumn of 2005. It reappeared two years ago with the Yellow Vests (*Gilets Jaunes*) revolt, in autumn 2018, this time principally concerning small towns or rural areas, described by Guilluy [3] as "peripheral France". It reflects a rupture between areas, very often affected by deindustrialization and unemployment, compared to areas where urban elites of power and wealth live, and between those who consider themselves to be the victims of globalization against those they consider to be the beneficiaries. According to Paugam [4], it is therefore a crisis of social bonds with a strong sense of abandonment or injustice accentuated by the isolation and withdrawal of public services.

ISBN: 978-1-61208-800-6

This feeling to be despised and forgotten by the power of the State (Paris) and by the European Union (Brussels), has progressively become a crisis of society [3]. This crisis of society is also a crisis of trust. According to Le Cardinal: "trust is at the foundation of society" [5]. It is built at different levels that interact with each other: trust in oneself, in others, in the future [6].

B. Scientific Positioning – Methodology

We position in a constructivist perspective (the construction of social reality by all actors) insisting on issues of meaning, interactions and social representations. We are part of a Research Team, DICEN IdF in a French University, associating the interdisciplinary Information and Communication Sciences that means information (data) perspectives with communication (links, interactions) perspectives. We present an approach named ICOE: Information and Communication Organizing Ecosystems. Ecosystems can be companies, organizations, social groups, and, of course, territories. We have an Action Research position (producing knowledge for action), with field observations and interviews with the main actors in the territories.

According to Bernard [7], we stress the importance of the researcher's commitment and communication to understand change as Carayol [8]. In the constructivist perspective explained, we attach great importance to creativity and innovation in territories like Godet [9].

We will insist on project dynamics to build trust between all the actors. The aim is to produce a collective intelligence and to learn how to better work together. To begin with, it is a question of exchanging ideas to build a collective representation of a shared future, then, to bring together all the relevant skills. The trust dimension is essential, both in the people and in the tools used.

We emphasize the fundamental notion of resilience. It originally concerned the Physical Sciences to define how a material could return to its original form after a shock. It was then used by psychology to explain how an individual could react to hard difficulties and get out of them. It has gradually taken on a collective dimension for social groups, organizations and also territories.

We particularly refer to the Situational and Interactionist Semiotics proposed by Mucchielli [10]. This method helps us to understand the meaning for actors in a specific situation. Mucchielli proposes to divide this situation serving as an interpretative background in different "frames": the intentions and the stakes of the actors, the culture and their norms of reference, their positions compared to other actors, the quality of relationships maintained, the historical and temporal frame, the sensory "frame", etc.

For us, communication may help to (re)create bonds, which are essential in these weakened territories. Data and its transformation into knowledge is also essential, with the issues of Big (by relying on the voluntary sector – *secteur associatif*) and Open data (institutional data) with GDPR (*General Data Protection Regulation*) challenges, with the new profession of Data scientist, applied to sustainable development and territorial marketing. But above all, it is a question of giving meaning to the data through intermediation tools in interface situations. According to Nesvijevskaia and Chartron [11], we insist on mediation between humans and data through interface tools.

Therefore, we consider the importance of socio-technical devices: Web sites and social networks as levers of resilience for these weakened territories. According to Ellul [12], we are aware of the ambivalence of technology and particularly of digitalization: it cannot do everything, but can make possible useful changes that create economic and cooperative values for sustainable development.

We focus on visibility issues (especially on social networks) to promote the attractiveness of these weakened territories, with e-reputation and territorial marketing issues. And additionally, with the importance of watching activities [13] in a Competitive Intelligence approach to anticipate and benchmark success stories developed in other territories.

After a mobilization of the existing literature, the proposed work is based on situations of participant observation. The author of this communication participates in the activities of local associations that serve as supports of this work. His membership in these associations has facilitated interviews, particularly with their leaders or essential actors and access to their documentation.

III. An Approach Focusing on Three Sectors (Local Companies, Tourism, Healthcare and Social Protection) Applied to 'Couserans' (French Pyrenees)

In this section, we propose a Territorial Intelligence approach focusing on three main sectors of analysis (local companies, tourism and cultural heritage, healthcare and social protection) and its application to the Couserans area in the heart of Central Pyrenees.

A. Three Sectors of Analysis: Local Companies, Cultural Tourism, Healthcare and Social Protection

Particularly interested in the territories of "*la France profonde*" or provincial France, often having a strong identity and worried about their future, we focus on three complementary converging sectors of analysis: one for economic development (local companies) and another for cultural activity and tourism, healthcare and social protection, with the main challenge of social and territorial inequalities, important for social cohesion and also for sustainable development.

First of all, we are interested in the sustainable economic development of these territories, through local companies, often focusing on authenticity and home-grown products (gastronomy), or on technological or specific product niches, employing local labor, with an innovative and creative dimension [9]. We must also mention the little-known role of the National Gendarmerie (*Gendarmerie Nationale*) in protecting both their physical (buildings) and their intangible (knowledge) assets.

We focus next on Tourism based on culture heritage and nature resources as a lever for the resilience of these weakened territories. We outline the prospects of "slow tourism", focusing on authenticity, valuing cultural heritage, in interaction with nature and gastronomy. The use of social networks can be a lever, as for "geocaching" for treasure hunting activities, with the example of Terra Aventura in the New Aquitaine Region. Also, with the enhancement of different “routes”: long-distance hiking trails (GR), the different Compostela paths, the paths of the Cathars (heretic people in the south of France during the Middle Ages) Stevenson's or Jacques Coeur's “routes” etc.

Healthcare and social protection issues (like education) are essential to maintain populations and attract newcomers (social link). We study them in a more global perspective of social and territorial inequalities in health: inequalities, both individual (poverty and isolation) and collective (remote territories): the question of medical deserts. We must also tackle the challenge of the “walls” dividing the hospital sector from the primary care sector [14]. New approaches with the development of interface organizations in healthcare (health networks or multi-professional healthcare houses, etc.), and the implementation of territorial support platforms for doctors in weakened territories with the construction of new "territorial health professional communities" can constitute interesting ways, particularly when they are based on local telemedicine projects and socio-technical devices (services platforms for doctors, EHR: Electronic Health Records, etc.). Safeguarding healthcare, as well as education structures (maintaining local hospitals, schools and high schools), implies preserving public services with a new approach (more in networks and with service platforms) to their activity. It implies also keeping public and private public transport to fight against the isolation of the poorest and most vulnerable people.

B. The Case of Couserans (Ariège / Pyrenees)

Our approach is applied to Couserans, corresponding to the district of Saint-Girons, in the department of Ariège in the heart of the French Pyrenees. This territory has a very strong identity that the daily newspaper *La Croix* has described as "an island in the Pyrenees" [15], with a strong tradition of dissent that is reflected in the importance of the Yellow Vests (*Gilets Jaunes*) movement in this area. This territory has been heavily affected by the rural exodus: 95,000 inhabitants in 1850 and 29,000 in 2015. Its “capital”, the small town (sub-prefecture) of Saint-Girons (6300 inhabitants) has lost most of its industries (paper mills) since 1975.

Couserans lost its passenger railway connection to Toulouse as soon as 1969 and its hospital has been threatening to close for several years. Classes and schools are closing in many villages. Live shows (*spectacles vivants*) such as *Autrefois le Couseran* at the beginning of August (more than 800 volunteers and 30.000 spectators) or that of the *Consorani* association keep the nostalgia of an idealized past and the pride of local traditions alive. This cultural dimension can be a lever for development with the repositioning of the tourist offices and the enhancement of cultural heritage, nature resources (mountain hikes, rivers), the Pyrenean Piedmont Compostela Way, etc. These live shows feed the dynamics of the voluntary sector (associations), constituting an essential asset. During the summer of 2020, most of the shows mentioned above, as well as a large number of local celebrations with their festive meals, important elements of identity and interaction between local inhabitants and tourists, were cancelled.

Local companies, particularly in the food sector (pork products, cheese dairies, cakes, jams, ice creams, etc.), focusing on authenticity and local products, with planned marketing on the Internet, constitute another important lever for sustainable development and local employment, also with competitive companies with niches and product strategies in industrial fields.

In the healthcare sector, interface organizations: Echo Healthcare Network, Home Hospitalization, MAIA for Alzheimer's patients, are now integrated into a CPTS (Territorial Professional Health Community) with the local hospital (CHAC or Hospital Community Ariege Couserans - *Centre Hospitalier Ariège Couserans*). After those existing in Prat-Bonrepaux, Castillon, Seix, Massat and La-Bastide-de-Sérou, a new multi-professional healthcare home (MSP) is under construction in Saint-Girons and will be operational in a few months. Training and education are also important assets. Saint-Girons has three high schools or *lycées* (one general and two vocational).

Since 2017, the 8 Communities of Communes of Couserans and two intercommunities unions have merged to create the Community of Communes "Couserans-Pyrénées" or "Com-Com". This now concerns 94 communes, with a population of around 30,000 inhabitants. It is concerned with a great number of sectors of local life: environment, waste, health, education, cultural life, healthcare, economic development, tourism, etc. It constitutes a major and federating actor for any territorial intelligence project and we are in contact with its main actors.

With its isolated and heavily depopulated high valleys, the Couserans is nevertheless only an hour and a half away from Toulouse and also has other assets that its inhabitants insist on: a quality of life and a rather preserved nature.

IV. For a Territorial Intelligence Approach in Interactions with and between Inhabitants

In this section, we propose some ways for a Territorial Intelligence approach focusing on interactions with and between all the inhabitants, all actors of these weakened territories, with of importance of rebuilding trust and the communication (relationship) and the use of data as possible levers.

A. The Importance of Territorial Intelligence

During a seminar of the French research network Org & Co – Organizations and Communication (2012) devoted to Territorial Intelligence, Le Moënne proposed a definition to which we subscribe: "Territorial intelligence is a form of collective intelligence developed on and around a territory in order to think and act there". Zara specified this notion of collective intelligence. For him, "it is the intelligence of the link, of the relation" [16].

We propose to insist on the dimension of cooperation, by involving all the actors even more. With this in mind, we meet Bernard and her approach to "engaging communication" [7], after having defined Information and Communication Sciences as interdisciplinary at the convergence of the four questions about creating links (interactions), meaning, knowledge and action [17]. Carayol [8] considered communication as a lever for change. We can also draw inspiration from Japan's "strategic knowledge communities", studied by Fayard and Moinet [18].

Bertacchini [19] considered "the territory as a collective intelligence enterprise to be organized towards the formation of local formal capital", promoting a culture of participation and project to federate all skills". We have considered [20] that the territory can be (re)constructed through a synergy of projects.

B. To Rebuild Trust

Girardot [21] within the international network of territorial intelligence INTI proposed the Catalyse method to make emerge and federate the initiatives of the actors of the territories, in particular of the inhabitants, with the use of the existing data and the production of new ones, with the possibility of creating territorial observatories.

The Catalyse method is based on many existing tools or tools to be built: information systems, especially geographical, with a strong territorial dimension, quantitative statistical methods or qualitative data analysis, project management and evaluation methods. Based on a diagnosis of needs and existing resources, its vocation is to build a territorial information system for decision support (proposal of indicators) of the main actors (local authorities, State services, associations, etc.) that can lead to the setting up of an observatory of the territory concerned.

The aim is to build solutions for and with the inhabitants. It proposes to articulate the needs of the territories and the available resources around a project dynamics federating the maximum number of actors, relying on the citizens' initiative and producing data to support their actions. This method has already been applied in different territories: Besançon (France), Liège (Belgium), Huelva (Spain), Salerno (Italy), in Quebec, Argentina, etc.

This method can be articulated with the FAcT - Mirror approach (Fears-Attractions-Temptations in Mirror method), proposed by Le Cardinal and his team [22], in particular to remove fears and develop trust around complex projects, but also in weakened territories such as the area of Belarus still affected by the Chernobyl disaster (1986) twenty years after the nuclear catastrophe [5]. Their challenge was to rebuild trust that had disappeared, as the populations had lost trust in the authorities, science, doctors and also in themselves.

Rather than trusting only the experts (those who said they know or think they know), Le Cardinal and his team went directly to the inhabitants, living with them for three years. They listened to their needs and hopes. Trust is time: it is built in the quality of relationships. Around the FAcT-Mirror method, this building of trust, constituting an accumulated capital, is based on a relationship ethic focusing on respect, loyalty and mutual commitment.

C. With Communication (relationship) and New Uses of Data as Levers?

Communication (relationships and co-operations) and new uses of data may constitute important levers of sustainable development of territories. “Living Labs" have been developed in this perspective. They insist on the project dimension, such as Brie Nov (Seine-et-Marne North), which proposes a PPPP approach (public-private partnerships and population), in particular to bring together and help work together native inhabitants and newcomers [23].

Data can be an important support for territorial development. With the creation and use of data, we enter the vast fashionable subject of smart cities, which, in the case of the Couserans, or other weakened territories, we prefer to approach through the notion of smart villages. Like Stiegler, we favor the dimension of human interactions. Faced with the risks of data use drift ("datacracy"), we prefer his "contributory learning territory" approach [24]. Insisting on the changes induced by massive data in human activities, Nesvijevskaia and Chartron underlined the stakes of the human/data interface. For us, this aspect is essential for the visibility and the development of the attractiveness of the territories [11].

But, for us, the uses of data and digital devices are not an end in themselves. They are not a miracle solution, but they can be an important lever for "resilience" and territorial development by promoting "reliance": development of interactions and social links between all stakeholders to promote innovation and creativity in the territories, as recommended by the Godet and al. report [9]. The aim is thus to produce data for action, in particular to promote collaborative innovations, which are promoted by Zacklad [25], by also making remote areas more visible thanks to digital technology.

These elements of reflection outline a cooperative and contributory intelligence approach that should be based on a synergy of local projects.

V. Trying to Develop a Contributory Intelligence Around a Synergy of Projects

In this section, we propose some ways to try to develop a contributory intelligence around a synergy of projects, with narrative shared experiences to build a form of Wise Territory. We analyze then the impact of the Covid pandemic in these territories.

A. The Importance of the Appropriation of the Territory through the Narration of Experiences and Projects

Following on from previous work [20], and in a socio-constructivist approach, we consider that the territory can be (re)constructed through a synergy of projects, including the sharing of knowledge to develop intangible territorial social capital [19]. Of course, we know how to take advantage of all the potentialities of new uses of data and digital socio-technical devices. The aim would be to foster a form of "resilience" of these territories, based on their assets of authenticity and identity around a new collective project of territorial dynamics, involving as many local and external actors as possible. In a way, "reliance" (network interactions) may help with "resilience", by making this territory more visible in the age of social networks and globalization and more attractive to newcomers: people and above all, companies.

We believe that, as in the Catalyse method, this project dynamics must begin with an appropriation of the territory by putting experiences and projects into narratives, combining memory and pride in the past (traditions) with future prospects, as D'Almeida [26] pointed out, as organizations always develop between projects and narratives. We think it may be the same possibility for these weakened territories in sharing experiences to build collective meaning and social representations.

B. Which Devices for Which Projects?

The Couserans Pyrenees Community of Communes or Com-Com and the Regional Natural Park of the Ariège Pyrenees including the Couserans and the various tourist offices that are in the process of merging may constitute interesting entry points.

The Web can also be an important element in making territories visible and contributing to their resilience. The University of Teramo, in Abruzzo (Italy), is thus trying to boost a new dynamic for the L'Aquila area, victim of an earthquake a few years ago, by relying on a Wikimedia project to give visibility to this territory by promoting the sharing of knowledge to foster its resilience, using open data in particular [27].

C. A Smart or rather Wise Territory ?

The Couserans could thus position itself as a "smart territory", with, by banking on the quality of life and the relative proximity of Toulouse, the possibility of "rural coworking" combining aspects of a break from the hectic life of today and remote work, as proposed for example by Mutinerie Village in the Perche (in the West France, in the south of Normandy), with, once again, the question of relays as levers for development and outreach, Mutinerie Village having also developed a network from Paris. This network operation, both internal and open to the outside world, can be an asset, drawing in particular on the dynamics of social networks.

This also raises the whole question of the role of the attractiveness agencies and the repositioning of the Couserans tourist offices, which, like the communities of communes a few years ago, are in the process of being grouped together. We are certainly moving towards a hybridization of their activities, with a platform dimension for an offer of services (with a strong interactive dimension) that are increasingly personalized.

Geocaching, which is increasingly practiced in other territories, can be an interesting asset if it is integrated as a form of tourist entertainment to help people discover local heritage, for example as a treasure hunt. Geocaching could be coupled with "slow tourism". A form of tourism, where one knows how to take one's time, insisting on contacts with nature, culture and gastronomy, favoring non-polluting transport (bicycles, etc.), in a sustainable development perspective, which seems particularly well adapted to the Couserans. These projects could be based on other, more important ones, to relay and legitimize them, such as, for example, giving a cross-border dimension to the ski resort of Guzet-Neige with a pass road relationship with the Catalan valleys, thus recovering old traditional relationships.

Tourism is not the only thing, even if it can be a key lever for development. The local hospital or CHAC, has long been banking on telemedicine activities, including with Spain, and on a new range of networked services: Echo Santé Health Network, home hospitalization, etc. The coupling of their activities with the development of multi-professional healthcare houses in the various valleys and under construction for Saint-Girons area, is essential for improving healthcare in Couserans.

Rather than "Smart Territory", this term in relation to "Smart City", having a strong technical dimension, we prefer to speak of "Wise Territories", mobilizing the notion of "Wisdom", linked to that of experience with a strong collective intelligence approach associating all the actors to take their destiny into their own hands. It is a question of (re)building trust for the resilience of these vulnerable territories based on a new collective dynamics (trust and reliance for resilience).

D. The Impact of the Covid Pandemic Crisis

The Covid pandemic crisis and in particular the long period of containment (*confinement*) in France (two months from mid-March to mid-May), followed by numerous restrictive health measures (wearing masks) and the fear of new waves in the months or years to come, are leading the inhabitants of large cities to take a different look at these economically vulnerable territories, from which they often come and where they quite often have second homes.

As the pandemic is often carried by populations coming from other areas or the proximity of major airports for large cities, their relative isolation becomes an essential asset. Their quality of life can be coupled with distance working (teleworking) for new attractiveness in our services society, with the added advantage of renewed contact with nature and the rather low price of housing.

But there is one major prerequisite for developing these new projects: breaking the isolation of the Couserans. Isolation on two levels: transport and also digital. The development of broadband, including in remote valleys, is essential to promote activities with a strong intangible dimension. This is one of the major objectives of the new president of the Ariège Departmental Council.

VI. With a New Role for Public Services: the State, Local Authorities and Social Protection organizations

Algan’s point of view in relation to the Yellow Vests revolt [28] is that the feeling of unease amongst weakened territories (and for us particularly in Couserans) is very noticeable, "wounded relationships with others", both at the individual and collective level, on one hand with social and territorial inequalities in health, and on the other, the individual and more collective dimension with territorial specificities. Moreover, with reference to Algan, it is a question of "creating links and trust in the territories", this idea of rebuilding trust which is the guiding thread of our work.

The official discourse since the beginning of the Yellow Vests revolt is to promote new forms of public action in territories, especially vulnerable ones. A new National Agency for Territorial Cohesion (ANCT) was created in November 2019, bringing together former organizations, including the DATAR. Its coordination in the various departments should be ensured by the prefects (whose role is also strongly questioned by digital transformation), responsible for the State's deconcentrated services in the departments. Is this a new way of redefining the national State as a platform for services, particularly digital services (e-administration)? Is it simply a change of name? Making new out of old? Only time will tell us...

This new role of the State, local and regional authorities and public services, particularly in the areas of healthcare and social protection, is essential for restoring the trust of local players and "rebuilding society", in a dynamic of animation and partnerships, by making the most of existing data and producing new data together, in the service of collective projects.

The National Gendarmerie may also play an important role in protecting their buildings but also the data and immaterial capital of these companies, which are often vulnerable in terms of cybersecurity.

Data is essential in healthcare and social protection. More than 20 years after Rosanvallon [29], we believe that it is no longer a question of rethinking the Welfare State, but rather, in an even more difficult way, in the current context (mistrust, budgetary constraints, etc.) to rebuild it, especially in vulnerable territories, so long forgotten. The digital transformation can help, as it can also aggravate the social disruption, by continuing to maintain them as "digital deserts".

The question of infrastructures to stop isolation is fundamental at two levels: transport and digital. Here again, the role of public services (State, Occitania region, Ariège department, even Europe) must be essential.

Like progress or technology [12], data can be ambivalent: they can help rebuild social ties and improve or enable new services for inhabitants, just as they can derive into "datacracy" [24]. Like all our society, weakened territories are at a crossroads ...

VII. Conclusion

This paper corresponds to a work in progress in a whole context of globalization, of digital transformation and ecological transition that involves societal disruptions but can also help the resilience of weakened territories (ambivalence of technology and progress). In this context, we try to propose some ways to an approach of Territorial Intelligence in interaction with the inhabitants to develop a collective intelligence around a synergy of projects to build a shared future and give hope to these weakened territories.

We insist for a renewal of public services with new forms of presence and action in territories to be invented. It is thus a question of federating and creating new links to develop the resilience’s capacity of these weakened territories by associating all the actors. First of all, by making them better known in order to build a shared project, by focusing in particular on the wealth of the associative sector in these territories. To make them more attractive, it is also a question of making these territories more visible on social networks (websites of local authorities, such as local companies) and thus promoting their e-reputation in a territorial marketing approach.

We try to open new ways to develop a value chain process of attractiveness by having all the actors working together to build a shared future as proposed by Le Cardinal with trust as a key lever [5], [6]. At the end of this work devoted to the resilience of weakened territories in France, we must come back to its main thread with the essential dimension of trust: in oneself, in others, in the future, apprehended in a constructivist approach of information - communication.

We insisted on the importance of communication to create links and then trust in a collective intelligence dynamic [16] by producing new knowledge (importance of data) in a contributory perspective that can be part of strategic knowledge communities [18]. The State and public services of region and department have a major role to play, in particular with Social Security organizations, to create social links [4] around new forms of solidarity, by listening to and being at the service of all the inhabitants, by encouraging the emergence of situations in which trust can be (re)built. It is our manner, according to E. Morin [30] to try to build a new future.

This work is the first step. We must now try to act in interactive research with decision-makers actors and with the inhabitants in local situations, as for example in Couserans, or in other weakened territories.

Its originality and main goal is to propose to try to articulate new initiatives of public or private authorities and citizen initiatives to create a new dynamics of resilience rebuilding trust around a collective intelligence for shared projects. The affirmation of new territorial actors in France with a federative vocation, such as communities of communes, can be an important opportunity. This is particularly the case of that of Couserans - Pyrenees, with whom we are going to work over the next few months to improve the visibility and attractiveness of this territory with a strong identity and trying to mobilize as many actors as possible to build a shared project together: newcomers, or probably more easily, native people who may now, with the Covid crisis, want to come back to work remotely, and people already living and working in these territories.

REFERENCES

[1] Giddens, "Runaway World. How Globalisation is Reshaping our Lives". London: Profile Books, 2002.

[2] J.-F. Gravier, "Paris and the French Desert" / "Paris et le désert français", Paris, Flammarion, 1947.

[3] C. Guilluy, "Peripheral France. How the Working Classes were Sacrificed" / "La France périphérique. Comment on a sacrifié les classes populaires". Paris: Champs – Flammarion, 2015.

[4] S. Paugam, "The Social Link" / "Le lien social", Paris: PUF, 2010.

[5] G. Le Cardinal, " Trust in the Foundation of Society " / " La confiance au fondement de la société ", Ceras - Projet n°293, Juillet 2006. Available on: http://www.ceras-projet.com/index.php?id=983.

[6] G. Le Cardinal, " Built Trust is always Fruitful " / "Une confiance construite est toujours féconde". Paris : La Croix, 28 – 29 décembre, p. 9.

[7] F. Bernard, " For an Engaging Communication, towards sustainable development " / " Pour une communication engageante, vers un développement durable », In S. Tremblay, N. D'Almeida, T. Libaert, Sustainable development. Communication that stands out / Développement durable. Une communication qui se démarque. Québec, Canada: Presses de l'Université du Québec, 2018 : 215 - 232.

[8] V. Carayol, "Organizational Communication. An Allagmatic Perspective"/"Communication organisationnelle. Une perspective allagmatique", Paris : L'Harmattan, 2004.

[9] M. Godet, P. Durance, M. Mousli, "Unleashing Innovation in the Territories" / "Libérer l'innovation dans les territoire". Paris : Conseil d'Analyse Economique - La documentation Française, 2010.

[10] A. Mucchielli, "Situation and Communication" / "Situation et communication", Nice : Les éditions Ovadia, 2010.

[11] A. Nesvijevskaia, "Big Data Phenomenon in Companies: Project Process, Value Generation and Human Mediation – Data" / "Phénomène Big Data en entreprise : processus projet, génération de valeur et Médiation Homme– Données", Ph D in Information and Communication Sciences / Doctorat en Sciences de l'Information et de la Communication, G. Chartron dir., Paris: CNAM, 2019.

[12] J. Ellul, "The Technique or Challenge of the Century / "La technique ou l'enjeu du siècle", Paris: Economica, 1990.

[13] H. Dou and S. D. Manullang, Competitive Intelligence, Technology Watch and Regional Development. Indonesia: MUC Publishing,2004.

[14] C. Bourret, "Tackle the Challenge of Social and Territorial Inequalities in Health (ISTS) by Meeting Interface and Telehealth Organizations in a "Digital Humanism" Approach to Health? "/ "Relever le défi des Inégalités Sociales et Territoriales en Santé (ISTS) par la rencontre des organisations d'interface et de la télésanté dans une approche d'« humanisme numérique » en santé ? ", Contemporary Trends in Organizational Communication / Tendances contemporaines en communication organisationnelle, in S. Alemanno, C. Le Moënne, and G. Gramaccia, dir., Revue Française des Sciences de l'Information et de la Communication [En ligne], 9 | 2016, available on : http://rfsic.revues.org/2013 ; DOI : 10.4000/rfsic

[15] La Croix, " The Couserans, an island in the Pyrenees "/ "Le Couserans, une île dans les Pyrénées". 2017. Available on : http://www.la-croix.com/France/Le-Couserans-une-ile-dans-les-Pyrenees-2017-01-30-1200821036#

[16] O. Zara, The management of collective intelligence. Towards a new Governance / Le management de l'intelligence collective. Vers une nouvelle gouvernance, Paris: M21 Editions, 2008.

[17] F. Bernard, "Information and Communication Sciences (ICS) as a Discipline of Openness and Decompartmentalization"/ "Les SIC une discipline de l'ouverture et du décloisonnement" », in A. Bouzon dir., Organizational Communication in Debate. Fields, concepts, perspectives / La communication organisationnelle en débat. Champs, concepts, perspectives, Paris : L'Harmattan, 2006, pp. 33 – 46.

[18] N. Moinet, " Territorial Intelligence between Communication and Strategic Kowledge Community: the example of the Poitou-Charentes Regional System " / "L'intelligence territoriale entre communication et communauté stratégique de connaissance : l'exemple du dispositif régional Poitou-Charentes", Revue internationale d'intelligence économique, 2009 /1, pp. 30-38.

[19] Y. Bertacchini, " The Territory, a Collective Intelligence Enterprise to be Organized towards the Formation of Local Formal Capital / " Le territoire, une entreprise d'intelligence collective à organiser vers la formation du capital formel local", / Communication and Organization / Communication et Organisation. n° 25, 2004, Available on : http://journals.openedition.org/communicationorganisation/2948

[20] C. Bourret, " Elements for an Approach of Territorial Intelligence as a Synergy of Local Projects to Develop a Collective Identity " / " Eléments pour une approche de l'intelligence territoriale comme synergie de projets locaux pour développer une identité collective ", International Journal of Projectics, n° 1, Brussels: De Boeck, 2008, pp. 79-92.

[21] J.-J. Girardot, " Concepts, Principles and Tools of the Catalysis Method " / "Concepts, principes et outils de la méthode Catalyse ", Proceedings o/ European Territorial Intelligence Network (INTI), Liège, 2005, pp. 133-137. Available on: mti.univ-fcomte.fr/reit/REITDoc/docs/GirardotLiege2005.pdf –

[22] G. Le Cardinal, J. F. Guyonnet, B. Pouzoullic, and J. Rigby, "Intervention Methodology for complex problems: The FAcT-Mirror method", European Journal of Operational Research, Elsevier, n° 132, 2001, pp. 694-702.

[23] I. Fasshauer, " The Living Lab, a Device to Promote the Resilience of Territories "/ "Le Living Lab, un dispositif pour favoriser la résilience de territoires ? ", Research Workshop DICEN IdF, Paris, Val d'Europe, 2019, June 27th.

[24] B. Stiegler, "The philosopher Bernard Stiegler prefers "contributory learning territories" to "smart cities" / " Le philosophe Bernard Stiegler préfère les "Territoires apprenants contributifs" aux "smart cities", Paris : iNNovaPresse, 2018. Available on

https://innovapresse.com/acteurs/36375-le-philosophe-bernard-stiegler-prefere-les-territoires-apprenants-contributifs-aux-smart-cities.html

[25] M. Zacklad, "The Economics of User-Friendliness in Information and Service Societies" / "Les économies de la convivialité dans les sociétés de l'information et des services", Inaugural Lesson / Leçon inaugurale, Paris: CNAM, 2009, June 17th.

[26] N. D'Almeida, "Organisations between Projects and Stories" / "Les organisations entre projets et récits", in A. Bouzon, dir., Organizational Communication in Debate. Fields, concepts and perspectives / La communication organisationnelle en débat. Champs, concepts et perspectives, Paris: L'Harmattan, 2006, pp. 145 – 158.

[27] C. Colombati and P. Valocchi, "Multifaceted Interactions between Local Resources and Wikimedia Ecosystem to Boost Abruzzo. Territory and Tourism Promotion", Research Workshop DICEN IdF, Paris, Val d'Europe, 2019, June 27th.

[28] Y. Algan, "Creating Links and Trust in the Territories to Reduce Populism" / "Créer des liens et de la confiance dans les territoires pour faire reculer les populismes"/ Public Actors / Acteurs publics, 2019, November 26 th. Available on https : //www.acteurspublics.fr/webtv/emissions/semaine-de-linnovation-publique/yann-algan-creer-des-liens-et-de-la-confiance-dans-les-territoires-pour-faire-reculer-les-populismes

[29] P. Rosanvallon, "The new social question : rethinking the Welfare State" / "La nouvelle question sociale : repenser l'Etat providence", Paris: Le Seuil, 1998.

[30] E. Morin, Seven complex lessons in education for the future. Paris: UNESCO, 1999.

Smart Territories

Advocating for Smart Basic Entity (SBE) and the digital clone approach

Pierre Fournié, Pr Christian Bourret
Laboratoire Dicen-Idf
Université Gustave Eiffel Paris Est Marne la Vallée
Paris, France
pierre.fournie@u-pem.fr christian.bourret@u-pem.fr

Pr Jean-Pierre Caliste
Université de Technologie de Compiègne
Compiègne, France
jean-pierre.caliste@utc.fr

***Abstract*—Smart cities (SC) became, for a few years, a regular topic in the scientific literature, and both political and economic agendas. Indeed, the connection between urban development and Information and Communication Technologies (ICT) represents a large market. It is presented as a multi-dimension tool to face the challenges of the 21st century. We intend here to demonstrate that developing such a concept only on cities may reinforce the already existing fracture between rural and urban territories. The opportunity exists to bring smart technologies to a lower level, that we call the Smart Basic Entity (SBE). We advocate that, to do so, it could be wise, to experiment a Digital Clone Approach.**

Keywords; Territorial Intelligence, Smart City, Smart Basic Entity, Digital Clone, Rural Territories, Ariège, Angola

I. Introduction

In 2009, the number of people in urban areas surpassed the number of people living in rural areas. Although, we shall keep in mind that national definition of what is “urban” is not uniform across the World. For the World Bank, the rural population registered a sharp decline during the 1960-2018 period, from 66.4% to 44.7%. Such a figure hides huge disparities between regions and continents. In France, the rural population amounts only at 19.56% in 2018 (divided by 2 since 1960 (38.2%)) whereas it remains at 34.49% in Angola (from 89.6% 58 years earlier) and at 44.68% in Indonesia (85.41% in 1960) [1] .The interpretation of such a worldwide trend shall take into accounts local and regional specificities [2] .

Cities offer multiple advantages: access to electricity, sanitation, water, health and education. Incomes are also higher even they shall be related to living costs. Supported by infrastructures of transportation and communication, high density of individuals and businesses, the city is a territory for serendipity. Urban areas also tempt individuals fleeing war or environment disasters. The attractivity of towns, urban centres or urban clusters, do also impact the structure of employment as there is a shift from agriculture towards manufacturing or services.

Therefore, showing disinterest for the rural population may, on the medium-long term, result in dramatic social, economic and political consequences. At the same time, an opportunity exists, by customizing SC concepts, to create a smart rural development model. We will illustrate the issues and solutions, we currently work at, by taking the examples of France, China and Angola.

After an introduction, we present, in section II, the Smart cities as a dominant model. The long term debate about urban concentration and growth will follow (III). In a fourth step we will consider what makes the cities so unique before identifying the risks associated with abandoning the rural areas (V). We will study in section VI the case of Ariège before exploring government strategies to transform rural areas into attractive territories (VII). We will, by presenting our model, indicate how technology could support such a move (VIII) and develop further the digital clone approach (IX). Before concluding (XI), we will illustrate by our Angola and China experiments, the current status of our researchs (X).

II. Smart Cities: A Dominant Model

Smart Data, Smart People, Smart Technology and Smart Governance represent the four pillars of the Smart City. SC has become the dominant model of development for towns in the 21st century. Without any restriction related to traditions, culture and religion, almost all aspects of urban inhabitants' life do enter in the SC scope.

All over the world, political leaders need to answer to the combination of significant issues, namely the explosion of demography and the revolt of Earth. By 2025, the level of urbanization will reach 58.20% (4.7 Bn people) from 44.70% (2.57 Bn), twenty years earlier. The projections show that the two-third of humanity will live in an urban environment by 2050.

The explosion of demography and the high concentration of human beings into cities may have devastating effects. Pollution, security concerns, mental and health disorders, increased and concentrated needs for energy and resources are the most intensively documented. Such a massive trend always requires more substantial storage capacities and efficient distribution networks. It also increases the vulnerability of human centres to natural (volcanoes, earthquakes, floods, sea elevation, high tide) and health disasters as well as to terrorist and cyber-attacks.

The revolt of Earth takes multiple forms from climate changes to disappearance of fauna and flora species, from freshwater scarcity to pest invasions. It profoundly affects, together with human-generated conflicts, the living conditions worldwide and creates mutations and transformations at many levels as well as migrations and destruction of human settlements. In such a tense environment, Smart City appears as an “easy to sell” political tool. Expandable and flexible, it is a kind of “swiss knife urban concept” able to resolve all issues mentioned above.

The political discourse is supported mainly by the revolution of ICT (Information and Communication Technologies), the one of IoT (Internet of Things) and more recently the fast development of AI (Artificial Intelligence). They allow managing ever-increasing volumes of data that shall grow from 33 Zettabytes (ZB) in 2018 to 175 ZB by 2025 [3].

For a large share of the population, those signs of progress are synonym of job destructions and highly tricky adaptation. Political discourses attempt to calm fears. Thanks to technology, the city would become the place where, through monitoring, air and water purity would remain unchallenged, traffic and transports efficiency would be enhanced, garbage would be invisible and immediately recycled. Criminality and terrorist risks would be assessed, and new services would regularly appear. A smart city is also a place where available jobs would be managed and fulfilled. Still, most of all, its development would require a high number of qualified technicians to support every day appearing positions.

Whatever is the type of decision process (centralized or decentralized) used to drive the development of the Smart City, all insist on people empowerment. Many suggest the promotion of platforms for bottom-up participatory governance [4]. Then, through the access to data, the contribution by forwarding ideas, the development of software and applications by its inhabitants or local companies, the city of the future may appear as a new place for democracy. At least in the political discourse and notwithstanding the increased pressure and control that smart data and IoT may allow on the citizens. By becoming smart, the city will anticipate and satisfy all needs as defined by the Maslow Pyramid from physiological to self-actualization: “Cities can be the source of solutions to, rather than the cause of, the challenges that our world is facing today” [5].

III. Urban Concentration and Growth: A Long-term Debate

Does urban concentration support growth? The question remains a place for fierce debates. Jane Jacobs, the well-known author of The death and life of great American cities, appeared as a pathfinder when she claimed that “the understanding of cities, and also of economic development generally, has been distorted by the “dogma of agricultural primacy.” From initially being primary organs of cultural development, cities have become primary economic organs [6] and as such a centre of growth. As places of innovation and industrial production, cities offer to the rurality the services and products required to increase the output, to upgrade raw into transformed goods with added value. Polese [7] studies the pro and cons of such a theory extensively. He identifies the laudators of cities whether considering the relationships between per capita income and urbanization levels, the contribution of urban areas to national income and product, the definite link between productivity and the agglomeration of economic activities in cities. But he rejects a direct causal relationship as scientifically impossible to demonstrate. Tolley and Thomas conclude that “Urbanization as such is neither the source nor the enemy of development” [8]

IV. What Makes Cities So Unique ?

Urbanization mainly concerns the rural-urban shift. The denomination is also used when population growth is predominantly urban. As indicated earlier, agricultural productivity directly impacts urbanization pressure. Reversely effectiveness of agriculture may lower or slower urban growth. Urbanization, part of the farm workforce becoming free, constitutes an inferred-effect of the agricultural revolution [9]. Castells-Quintana and Royuela [10] edulcorate such a proposal by stressing,that rural population are often expelled from the rural areas.

The genuine attraction for cities shall be analyzed. Cities are the mothers of human progress [11] and, no development may take place without towns [12]. Often the change in status (from city to capital as it was the case for Jakarta in 1961-1964); the efforts from government or municipality; the development of the hinterland, accelerate the urbanization process. Thanks to the concentration of capital and means, the city offers a place for innovation. [13] “The city promotes the monetization of the economy, facilitates social mobility and the adequacy between offer and demand for qualified manpower, expands the markets for industrial and agricultural productions” [14]. Even a specific size shall be reached to boost technological progress [15], dense and marge populations constitute a fertile ground able to welcome the exchange of ideas [16]. City diversity supports employment growth [17]. By limiting the distance, offering efficient transportation and communication networks, they reduce the cost of transmitting information and increase efficiency and productivity. [18] [19] Unsurprisingly the patterns have changed with time also in the 90’s “Cities with high levels of human capital did well, and cities with large numbers of the poor did poorly” [20] Environmental concerns also require adaptative and innovative means of transportation. Intermodal platforms aim at improving efficiency. [21]

Cities are intricated into a complex system: the system of cities. For Pumain [22], cities also have an intrinsic quality to transform themselves: an evolutive capacity or (re)organization. Remembering the “General systems of cities” conceptualized by J.Reynaud in 1841; the works of Berry [23] and its famous “cities as systems within systems of cities” and the ones of Pred [24]; using analogies with physical systems and synergetics; Pumain compares cities systems to dynamic systems governed by an auto-regulation. She goes even further by defining an evolutional theory centred on the notion of “*system of cities*”; thus, ending the supremacy of a static vision of cities.

V. The Risk of Abandoning Rural Areas and “The Catastrophe Scenario”

Could any government decently abandon 60 to 20% of its population? The “yellow vest” movement in France, initiated in October 2018 and still active as of March 2020 in some rural areas, constitutes, for any government, a fierce reminder. The images of violence in Champs Elysees, the groups of rioters spreading in the capital and main cities, made people forget that many Yellow Vest were leaving in rural areas and were protesting against decisions taken by the central government that directly impacted their living conditions.

France is presented regularly as an heir of Jacobinism. Despite some tentative of decentralization, regionalism is refrained, and decisions centres are often far from the countryside.

Conscious that a balance shall exist across the entire national territory, President Charles de Gaulle created in 1963 the Délégation à l’Aménagement du Territoire et à l’Action Régionale (DATAR). Under the Prime Minister, such a structure had to impulse, coordinate, reequilibrate the actions of the state whereas in developing rural areas, reinforcing transportation networks, meshing the country. By using DATAR, the state aimed at organizing and modernizing France; at preserving cohesion, at making territories more attractive.

On 2014, DATAR merged with the Comité Interministériel des Villes and l’Agence Nationale pour la Cohesion Sociale et l’Egalité des Chances into CGET, le Commissariat à l’Egalité des Territoires. Such a combination of agencies was lately replaced by the Agence Nationale pour la Cohésion des Territoires (ANCT) under the ministry of Cohesion of Territories and Relationships with Territorial Collectivities.

Terminology matters here: cohesion, equality constitute the fertile soil of a territory that any disorder may spoil forever.

In the 1970s, DATAR prepared a prospective study forecasting what will be France by the end of the 20th century. One of the output, referred to as the “Catastrophe scenario”, concentrating all developments into large metropoles and consequently creating desertification of rural areas had to be avoided. Dou and Fournié [29] have shown that, notwithstanding expert’s advice, France has been developed under such a configuration. The sizeable interstitial space created suffers from insufficient means, infrastructures and is dramatically abandoned by its population.

The state agency France Strategy proposed that national investments be channelled to the 15 largest French cities through metropolitan pacts of innovation. Such a policy, justified by a lack of resources, may lead in territories located at the fringe, to the reinforcement of inequalities as regards as public services, access to medical services, connectivity between others.

France suffers from a triple fracture: a Territorial Fracture, a Technological Fracture and “Data-consciousness” fracture.

France is not the sole country to face problems with its rurality. The phenomena concerns almost all countries of the continent. The European Union, now aware of the situation, initiated some action process with the Cork declaration, Ireland, 2016; the UE Action Plan for Smart Villages (11/04/2017) that aims at « investing in the viability and vitality of rural areas »; the Bled declaration, Slovenia, 2018.

VI. The Case of Ariege

Separated from Spain by the Pyrenees mountains, Ariège is one of the 13 departments composing the Occitanie region, second largest province in France (72 724 sq/km). With 5.8 million inhabitants, organized around 18 urban poles, two large metropoles (Toulouse and Montpellier), the region shelters a strong industry recognized at international level (aeronautics, spatial, in-vehicle systems, agro-industries, biotech) supported by 15 poles of competitivity and several large universities.

Ariège is a department limited in size, home of 152 724 inhabitants of which 46.04% are over 50 years old compared to 33.08 % in the neighbour department of Haute Garonne (central city: Toulouse). Largest cities in 2017 were Pamiers (15675 habs), Foix (9 532 habs), Saint Girons (6 383 habs) and Lavelanet (6 137 habs). Only 31.1% of Ariege’s inhabitants have a graduate-level, far from the 50.8% of Haute Garonne and 97% of the companies have less than 50 employees.

The territory, called in some media the anti-startup nation [25], has suffered from the closure of few large factories (in particular of paper, an industry-supported by hydroelectric capacities and forests) and mines in the 90’s, of “green tourism” being impacted by fierce competition with other regions and international destinations since the year 2000, and more recently of “white tourism” being affected by climatic change.

Still, Ariège benefits of several assets. A vast cultural and historical heritage (between others Cathar castles), an immaculate nature with 55 000 ha of regional Natura 2000 park that welcomes bears and wolves, high peaks over 3000 m, thermalism are only parts of them.

In this French department the disappearing of state presence, mass transportations and private services alter the living conditions and destroy the efforts to promote tourism On the long term, they might be the synonym of exode and increased poverty. A feeling of exclusion may prevail that could be transformed into social unrest and affect social and national cohesion. State presence disappearance takes multiple forms: the closing of taxation and perception offices, of classes in schools and colleges, of tribunals and legal offices, the reorganization of beds in hospitals or health services. In parallel, traditional shops are impacted by slow local and touristic activities. Low traffic and profitability condemn branches of banks and post offices. Bank Automated Distributors are suppressed. As in many rural regions, people have to drive 20 to 40 km to access essential services. Not the least, doctors and specialists abandon those areas, and there is a lack of professionals in both private sector (-26% between 2014 and 2018) and hospitals (-16% for the same period).

Moreover, new regulations made possible the absence of controllers in the trains creating tense situations as regards as security and law enforcement in transports. On many lines, tickets cannot be sold into trains whereas at the same time commercial offices are being suppressed or opened during a short period of the day. Automatic machines are destroyed if located outside of the station and remain not accessible out of office hours or during weekends. The law referred to as LOM (Loi d'Organisation des Mobilités) signed December 24th, 2019, creates a right to mobility. It may also, despite allowing through open data the access to information on transports and reinforcing the role of region, broadly impact rural territories.

As regards as communication networks, 48.3% of housing have access to a high-speed network whereas over 10% still have connection problems. At the same time, only 22.6% are eligible to optic fibre. [26]

Thus, at the same time authorities are pushing for more eco-friendly means of transportation and numeric transformation, an ageing population, suffering from a lack of knowledge and adaptative capacities, has difficulties in adapting to tools not fully available and to which it has not been trained. Besides the rupture of equality between territories, such a situation may endanger- this a guess that shall be assessed on the field- on a large scale, mental health, create deep feelings of exclusion and increase the vulnerability of rural populations.

VII. Could Rural Areas Become Attractive Again?

Could rural areas become attractive again? Or in other words could we, create in those regions, at a time environmental concerns become a priority, the conditions to make the soil fertile again for living, to invert the rural-urban shift, to attract capital and means, to transform rural areas into places of innovation and serendipity.

The report of Cour des Comptes dated March 2019 and entitled Accessing public services in rural territories concludes that, despite multiple initiatives, rural areas have required a long-term effort and remain a permanent failure of the central state. Multiplication of policies, overlapping of competences, and lousy coordination represent only a few aspects of the problem. Several laws have been voted with little positive impact: the legislation « Montagne » (Mountain) of 1985; the law for the Orientation and Development of Territories dated 04/02/1995; the bill for the Development of Rural Territories (23/02/2005). Circular letters addressed to the prefects by the Prime Minister and interministerial committees for rurality took place between 2015 and 2016 defining a set of 104 measures to promote rural areas and ensure local development. Were considered as priorities: the access to health and the fight against « medical desert »; the access to services through Maisons de Services Publics (MSP) (Public Services Office); « Nomade » Public; and « mobile » Postmen Services; the fight against school weaknesses; the numeric coverage and implementation of networks in rural areas; the execution of « Contracts of Rurality ». Those committees were replaced by CGET soon to merge into ANCT as mentioned earlier.

The strategy of President Macron government for rurality should be read in the continuation of previous initiatives. Still 18% of the total French population live in medical deserts [27], a figure to compare to the 19.56% representing the total rural population. The same remedies have been proposed over the years without success. Rural territories continue to see their young population disappearing, the closure of public and private services and the constant degradation of their images in a never-ending vicious circle.

VIII. A New Strategy Supported by Technology

Boosting the attractivity of rural territories is related to both a change in image, the development of light infrastructures of utilities and communications, the development of inner innovation capacities and value, the acceptation of the challenge by local people. Technology could be a fantastic chance to invert such an ineluctable destiny. Whatever we call it "Smart Village" or "Smart Basic Entity" (SBE), we advocate that a new frame of organization and development shall be studied. It may become a potential area of growth and allow the inversion of the concentration process into cities. The SBE model shall, of course, and provided customization, benefit of the technical innovations and successful realizations that would be implemented in Smart Cities. SBE and SC models shall not be competing with each other, but live side by side, completing each other through exchanges on data, technologies and experiences, through existing or to be developed networks.

What we foresee today is nothing else than a downsizing process; to go from mainframe (the state or the region) to a connected smart unit, the SBE. Such a reorganization of the territory underpins, on the medium term, a global rethinking of administrative and political organization.The SBE could be defined as an evolutive and complex system, without exact physical limit but characterized by a logic of flows (persons, assets and information) always looking for efficiency improvement. By analogy to the system of cities; SBE will be connected through them as a network: the system of SBEs. Flows exchanges will exist within the SBE and with the exterior.

A way to define Smart Cities is by using a tangible (Hard)/ intangible (Soft) domains approach [28]. By customizing such an approach, we may consider two axes:

- Tangible/Hard: Water, Energy, Land and Environment Resources (Forests, Mines, Parks…), Transportation, Buildings (including health care and education) infrastructures, Security systems
- Intangible/Soft: Governance, Education, Health, Economy, Culture systems and data

And seven related applications: Utility management, Land and Environment Ressources management, Mobility, Buildings, Economy, Security, People (Cultural, social, education and health)

New technology facilitates communication, remote financial operations, distance learning, online medicine. As a consequence, cities are on the way to lose most of their competitive advantages.

IX. The Smart Digital Clone Approach

By collecting data related to the 2 axes, we shall be able to create a digital clone of the SBE. It shall represent its tangible and intangible assets and allow to analyse flows within the SBE as well as exchanges with outside, mainly, within the system of SBE. Such an approach that we call SMART (for Systemic Modeling and Advance Reengineering of Territory) shall facilitate improvements, reinforce governance efficiency

The digital clone would allow, by simulating any structural or logical change, to evaluate it. It shall promote, as well, the definition and implementation of contingency plans; and make possible dynamic stress tests on all tangible and intangible dimensions. It appears mandatory to respect when implementing such an approach, several rules that we summarized as the HAWKS principles. H: Holistic: the digital clone covers all aspects and enlightens even shadow areas A: Accepted: by the population W: Wise K: Creative S: Secured

The digital clone terminology is used by analogy with recent development in medical technology. If a digital clone can be created to save or cure a human patient, why not applying the same to study, monitor, optimize the living conditions in a rural SBE?

X. Angola and China Cases

In Southern Africa and more specifically in Angola, the question of the economic development based on a systemic (system and system of systems) approach is crucial. The problem is multidimensional from two different points of view. Multidimensional, according to the aspects which constitute the domains where solutions must be designed, developed, implemented and evaluated: production and consumption of energy, wastes treatments, water control (drinking and wastewater), forest preservation and agriculture development, artisanal (handmade) and industrial activities, mobility … . Multidimensional, because at the same time, cultural, societal, technological, collective and individual.

Under the umbrella of the DNDTI (National Direction for the Development of Technology and Innovation) of the MESCTI (Ministry for Higher Education, Sciences, Technology and Innovation) of Angola a “SMART Villages” project is close to being launched. The question is how new technologies can be used for assuring a real, sustainable development of a small city (village)? How could a global approach be defined and implemented? As we noticed at the beginning, the project aims to consider each field of challenge (energy, wastes, water …) as a system and to consider the global interactions between all these systems. The objective is not only to solve problems; it is more the optimization of different solutions under a set of balanced criteria (economic development, alphabetisation rate, safety and security …).

This pilot project will consist of two main parallel parts. The first will be focused on scientific and technological studies to propose and implement selected solutions, for example, about production and consumption of energy. But the most original part could be the second one. This part will consist of developing a “digital ghost” (or digital clone) of the small city (village) based on the capture of data concerning all the aspects to take into account even the feelings and the opinions of the citizens, visitors tourists…. Also if at the origin the ”digital ghost” will be more a database than a real digital representation of the city, each development will be the occasion to reinforce the amount of data (information) creating step by step a new digital structure able to facilitate original representations of the city based on the selection of data according to specific criteria (as mobility by example).

On the other hand, to be able to assess the project itself and its results, a model of excellence, as the EFQM model used for evaluating the efficiency of enterprises or public organizations will be used. This approach will allow to assure or to analyze the results for different stakeholders. According to this last point, it will be interesting to compare the evolutions (transformations) of the “digital ghost” of the city with variations of the stakeholder’s opinions reinforcing by this way the capacity to understand better how a SMART integrated development impacts positively or negatively the citizens’ lives.

Another exploratory work was carried out in the city of Shanghai (China). In partnership with the company Mobike (free bicycle), students worked on the development of an onboard pollution capture system (integrated into the bike frame). The aim was to transmit this information to a platform to advise cyclists on "greener" alternative routes.

XI. Conclusion

There is no inevitable future for rural territories. We firmly believe that the rural-urban shift is not a fatality and could be inverted provided rural areas modify their image and become land of innovation and investment. Thanks to technology, distance is no more a concern as regards as accessing information, finance and education. Environmental and security issues could reinforce the position of territories towards cities that have lost part of their competitive advantages. Such a revolution could take place at the SBE level and that the development of SMART Digital Clones constitutes a significant leap into the future for rural territories. Its application has no frontier from China to Angola, from rural territories of France to the ones of Indonesia. And the current ongoing pandemy of Corona virus, that constitutes a real stress test for the economy and states structures at the international level may question the future of megapoles. Isn’t it time to invest in rural territories, therefore keeping in mind that their specificities and environment shall be preserved ?

References

[1] https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS

[2] M.Pesaresi et al., “Atlas for the Human Planet 2016: Mapping human presence on Earth with the Global Human Settlement Layer”, European Commission, 2016

[3] D.Reinsel, J.Gantz and J.Rydning, “The Digitization of the World: From Edge to Core”, An IDC White Paper #US44413318, 2018.

[4] UN Economic and Social Council, Smart cities and infrastructures. Geneva, 2016.E/CN.16/2016/2.

[5] New Urban Agenda - Habitat III. United Nations. Quito, Ecuador, 2016. ISBN: 978-92-1-132731-1.

[6] J. Jacobs, “The economy of the cities”, New York, Vintage Books, 1969.

[7] M.Polèse, “Cities and National Economic Growth: A Reappraisal”, Urban Studies, Vol. 42, 8, pp. 1429-1451, July 2005.

[8] G.S.Tolley and V. Thomas, “The economics of urbanization and urban policies in developing countries”, The World Bank,1987, p. 185.

[9] J.Véron, “Enjeux économiques, sociaux et environnementaux de l'urbanisation du monde”, Mondes en développement, Vol. 2008/2, 142, 2008, pp. 39-52.

[10] D.Castells-Quintana and V. Royuela, “Malthus living in a slum:Urban concentration, infrastructures and economic growth”, St. Petersburg, Russia : European Regional Science Association (ERSA), Louvain-la-Neuve, 26-29 August 2014.

[11] P.Bairoch, “De Jéricho à Mexico, Villes et économie dans l’histoire”, Paris, Gallimard, 1985.

[12] N.Keyfitz, “International Migration and Urbanization”, in P. Demeny and M.F. Perutz, Resources and Population, Clarendon Press, 1996, pp. 269-285.

[13] K.F.Sokoloff, “Inventive Activity in Early Industrial America: Evidence From Patent Records, 1790–1846”, The Journal of Economic History, Vol. 48, 04, 1998, pp. 813-850.

[14] P.Bairoch, P. “Cinq millénaires de croissance urbaine” in I. Sachs. Quelles villes, pour quel développement ?, Paris, PUF, 1996, pp. 17-60.

[15] J.Henrich, “Demography and cultural evolution: How adaptive cultural processes can produce maladaptive losses-The Tasmanian case”, American Antiquity, Vol. 69, 2, 2004, pp. 197-214.

[16] P.M. Romer, “Endogenous Technological Change”, The Journal of Political Economy, Vol. 98, 5 Part 2, October 199., pp. S71-S102.

[17] E.L.Glaeser et al., "Growth in Cities”, Journal of Political Economy, Vol. 100, 6, 1992, pp. 1126-1152.

[18] A.Marshall, “Principles of Economics”, Macmillan and Co, New York, 1890.

[19] G.Duranton and M.A. Turner, “Urban growth and transportation”, The review of economic studies, Vol. 79, 4, October 2012, pp. 1407-1440.

[20] E.L. Glaeser and J.Shapiro, “Is there a new urbanism ? The growth of the US cities in the 1990's”, Cambridge, MA, National Bureau of Economic Search, 2001. Working Paper 8357.

[21] V.Vuchic, “Transportation for livable cities”, New York, Routledge, 1999. 378 pages.

[22] D.Pumain, “Pour une théorie évolutive des villes”, Espace géographique, Vol. 26, 2, 1997, pp. 119-134.

[23] B.J.L Berry, “Cities as systems within systems of cities”, Papers of the Regional Science Association, 1964, Cited by Pumain, 1997.

[24] A.Pred, “City systems in advanced societies”, London, Hutchinson, 1977.

[25] A.Rousseau, “L'Ariège, l'anti-Start-Up Nation”, March 22nd 2019, Les Echos.

[26] ZoneAdsl, Couverture internet de l'Ariège, retrieved 1st March 2020. https://www.zoneadsl.com/couverture/ariege/.

[27] C.Maisonneuve, “Nouvelle carte des déserts médicaux : votre commune est-elle concernée ? ”, La gazette.fr, 2017.

[28] P.Neirotti et al., “Current trends in Smart City initiatives: Some stylised facts”., Cities, June 2014, pp. 25-36.

Articles in conference proceedings:

[29] H.Dou and P.Fournié, “Les Smart Villages dans le contexte européen” ,Val d'Europe, Codata 2019: Data value chain in Science and Territories, 2019, pp. 19-26.

YouTube Video Categorization Using Moviebarcode

Recep Erol, Rick Rejeleene, Richard Young, Thomas Marcoux, Muhammad Nihal Hussain, and Nitin Agarwal
Collaboratorium for Social Media and Online Behavioral Studies (COSMOS),
University of Arkansas at Little Rock,
Little Rock, Arkansas, USA
{rxerol, rrejeleene, rbyoung, txmarcoux, mnhussain, nxagarwal}@ualr.edu

***Abstract*—Every minute more than five-hundred hours of video content is uploaded to YouTube, and we can only expect this number to increase. Although YouTube is the most popular video sharing website, studies conducted on this platform are sparse. The lack of effective video analysis techniques presents a tedious challenge for researchers and has hindered overall research on this platform. Due to this, research conducted on YouTube primarily focuses on analyzing text-based content or video metadata. With recent advancements in the development of *moviebarcode*, a technique that shrinks a movie or video into a barcode, we have developed a tool designed to extend the capabilities of moviebarcode as a forensic technique for systematically categorizing YouTube videos. We use moviebarcode to summarize an entire YouTube video into a single image to help users understand a video without even watching it and later use cluster them based on similarity. We analyzed six video collections and using moviebarcode only and without looking at the video content, we were able to achieve an accuracy of 75%. Using our method, an analyst can quickly group videos into bin computationally reducing the overhead of manually doing it.**

***Index Terms*—Moviebarcode, Video Categorization, YouTube, Social Computing Tool**

I. Introduction

In recent years, social media has become ubiquitous among the lives of people who seek to consume content from social media. With respect to content creation and data analysis, it is fair to compare the promises of social media to a modern-day gold rush. Although there are numerous platforms classified as social media; videos have been proven to be the most popular medium for sharing content among users. The most popular platform for video-based content is YouTube.

For every minute, more than five-hundred hours of video is being uploaded to YouTube. We can only expect that number to grow as YouTube focuses on expanding its global reach and making the platform more profitable for content creators [1]. As digital content and consumption is increasing at an incredible rate all over the globe, YouTube video processing becomes computationally intensive. Prior to 2010, YouTube videos could not exceed a video length of 10 minutes. When this restriction was removed, a user published a single video with over 600 hours, which would take 24 days to watch the video [2].

There are many available deep learning based video categorization studies [3, 4]. These studies show great contribution to the research community. However, the length of a video is the major limitation for available video processing tools such as computer vision and deep learning based algorithms as they require extensive computational power, time and human effort. In addition to cost and power requirements, currently available video processing tools have a steep learning curve for social computing researchers. Moreover, these tools do not directly provide information to use in identification of cyber activities on videos. Due to these limitations, we extend moviebarcode, a state of the art video summarization tool that provides linear or close-to-linear processing time regardless of video length.

Moviebarcode is a technique that uses color theory to summarize videos by compressing an entire video into a single image [5]. The result of this technique is a single barcode consisting of generated colors for every frame of the movie. Moviebarcode shows the color transitions within videos, gives an overall idea about the video content, and enables comparison with other videos without watching the video, thereby saving time.

In this paper, we extend previously described moviebarcode into an implementation and prototype as a tool to identify similarities among videos, capturing the visual patterns in a video and extract insightful knowledge efficiently. In addition to implementation and prototyping, our novel idea is categorizing videos with moviebarcode. For this purpose, we created six different video collections, namely APAC, BalticOps, FifaUnder17Games, ManuGinobiliGames, SpongeBobSquarePants, and HBOSiliconValleyTrailer. Using categorization algorithm to group the moviebarcodes and got promising results that are explained under section 4. With moviebarcode, researchers can interact with YouTube video without watching an entire video through summarization. A user is able to optimize important resources such as time to condense each video. The details of the dataset and analysis can be found in Section 4.

The rest of the paper is organized as follows: In Section 2, we describe related works of moviebarcode. In Section 3, we explain Moviebarcode, its generation process and representation. We describe our dataset, categorization of videos using moviebarcode, discuss our findings and their significance in Section 4. We conclude with major contributions and future direction for this research in section 5.

II. Related Work

Moviebarcode was made popular by Clark [5], a Tumblr blogger, who generated moviebarcode for numerous movies, and each movie could be filtered by title, director, genre, year. Blogger would capture color patterns in a movie to summarize it, irrespective of its length, to a single barcode.

There are several researchers that used moviebarcodes for visual video analysis [6] such as ColorBrowser [7]. Burghardt

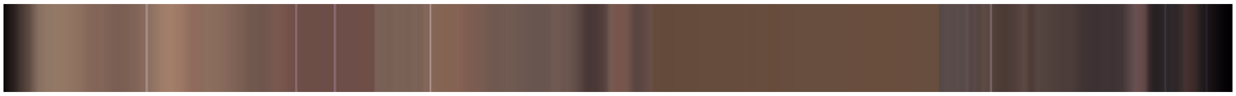


Fig. 1. A moviebarcode illustration from a basketball game video.

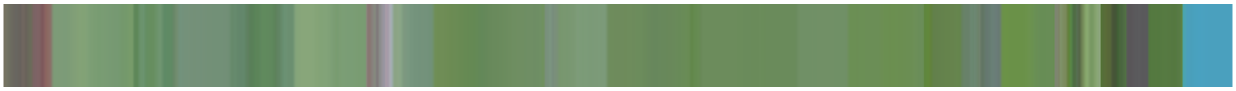


Fig. 2. A moviebarcode illustration from a soccer game video.

et al. [6] present an approach that can automatically extract and analyze the language and color parameters from movies by visualizing the most frequent colors in movies. In their approach to visualize, they used clustering algorithms and moviebarcodes. However, we find that there is no summarization tool provided by this research, and their idea falls short of searching and comparing multiple videos. Another study [8] introduced a pictorial summary that summarizes a segment of a video for visual representations. Otto et al. [9] presented moviebarcodes and long exposure images to visualize the colours present in a movie by calculation color population in a frame and stack them together in a moviebarcode format. However, they also normalize the color values to 100. But, their computational cost is more expensive and their research does not include categorization.

Our work is different from aforementioned works as we use moviebarcode for categorizing videos. Also, the method to group videos based on similarity has been done empirically and requires an analyst to manually watch the videos to group them. Our method reduces the effort significantly by using computational methods.

III. Moviebarcode

In this section, we describe moviebarcode, its generation process and its representation as vector, matrix, tensor. We also explain step by step process used to generate moviebarcodes for videos on YouTube.

A. Moviebarcode

Moviebarcode is a technique to represent a video or a movie as an image by stacking mean values of each frame. Video is a sequence of frames, and there are approximately 30 to 60 frames in each second of a video. When the video is longer than 10 minutes, the number of frames in a video will be greater than 18,000 frames which makes the video analysis even harder because of the high computation requirements. However, Moviebarcode can easily handle any video for analysis.

Moviebarcode is unique to each video. For instance, when the same scene is recorded with the same camera two different times, the moviebarcode will be different from each other. Furthermore, if a scene is recorded from two different angles, Moviebarcode will again be different. So, it can be said that Moviebarcode is a good technique to catch replicated videos or short clips within a video.

Moviebarcode gives dominant colors in each frame. From these dominant colors, significant information about a video can be learned without watching it. For instance, there are two videos; one from a basketball court, and the other from a soccer video. Fig. 1 and 2 show the moviebarcode of a basketball game and the soccer game videos, respectively, and both moviebarcode are easily distinguishable. This is important because getting an idea about a video requires significant time to watch and categorize. Moviebarcode technique eliminates this process and shortens the time required for categorizing and filtering videos without watching them.

B. Moviebarcode generation and structure

A moviebarcode can be generated for any video or movie, not just limited to YouTube. Since a video has a sequence of frames, each frame is extracted from a video. Then, the mean value of Red (R), Green (G), Blue (B) channels for each frame is calculated. So, after getting a mean value of a frame, a vector of three color values (RGB) is generated (Fig. 3.1). By using RGB channels, the gray scale image can also be generated if needed so that the moviebarcode can be represented with a gray scale and used for quantitative analysis. After all these vectors are stacked, we get a matrix of RGB values. So, we can represent a video as a matrix of RGB values (Fig. 3.2).

Besides vector and matrix representations, moviebarcodes can also be shown as a tensor. As seen in Fig. 3.3, when the RGB matrix is converted to tensor, it can be displayed as an image which means representing a video as an image. The width of the image is equal to the number of frames, and the length is to the number of pixels that a user can assign. This number of pixels is 224 in our experiments.

The most important question to ask here is what kind of information can be extracted from a moviebarcode. Moviebarcodes use color theory to represent a video. Dominant colors of each frame are stacked on a moviebarcode which means that a moviebarcode keeps dominant colors of the video that are easily identifiable. This sequence of dominant colors and their transitions can give information about the video such as changes in the scene, the subject, the narratives of the video within time without watching the video.

C. Generating moviebarcodes from YouTube

For data collection and streaming, we use public data API of YouTube [10] to download data from YouTube, and

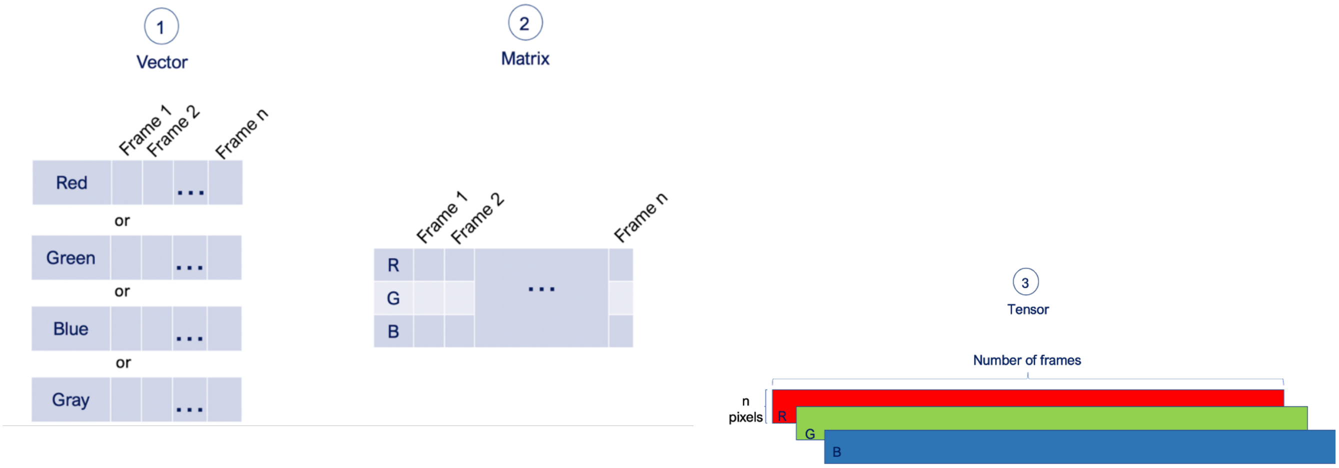


Fig. 3. Moviebarcode representations as a vector (1), matrix (2) and tensor (3).

OpenCV computer vision framework [11] to stream videos from YouTube. However, due to YouTube's policy, we do not save the original videos. Moviebarcode of a video is generated through a process shown in Fig. 4. If a user enters a YouTube video URL, the procedure first checks the availability of this video. If the video is online and still public to download, we stream the video and generate a moviebarcode on the fly which means that the video is not saved locally. If the URL is for a playlist, the same steps are applied recursively for each video. The algorithm only saves mean values of each frame of a video as .json file.

IV. Datasets and Categorization

For this study, categorization of videos using moviebarcode, we carefully curated a dataset of six different collections of videos. Subject Matter Experts (SME) helped us identify and group the videos that were later collected using data collection method described in our previous studies [12,13]. These collections of videos are "APAC", "BalticOps", "FifaUnder17Games", "ManuGinobiliGames", "HBOSiliconValleyTrailers", and "SpongeBobSquarePants". APAC collection consists of conspiracy theories and misinformation videos being disseminated related to various events and issues in the Asia Pacific region. BalticOps collection consists of videos with misinformation about NATO's 2019 BALTOPS exercise. FifaUnder17Games collection consists of videos about soccer. ManuGinobiliGames collection consists of videos about highlights from the NBA games that Manu Ginobili plays. HBOSiliconValleyTrailers collection consists of trailers of a hit television series called Silicon Valley. SpongeBobSquarePants collection comprises of videos of the cartoon show called Sponge Bob Square Pants. The number of videos in each collection is shown in Table 1. The lengths of videos in video collections ranges from 3 minutes to 20 minutes.

To construct moviebarcode images, we used matrix representation. Moviebarcodes have three channels, RGB, and different widths. Each moviebarcode's width is equal to the number of frames.

TABLE I
VIDEO COLLECTION DATASET INFORMATION

Collection Name	Number of Videos
APAC	14
BalticOps	14
FifaUnder17Games	15
ManuGinobiliGames	15
HBOSiliconValleyTrailers	15
SpongeBobSquarePants	15

TABLE II
MOVIEBARCODE VIDEO CATEGORIZATION RESULTS

	Precision	Recall	F1-Score	Accuracy
Red channel only	0.79	0.64	0.59	0.64
Green channel only	0.82	0.71	0.69	0.71
Blue channel only	**0.83**	**0.75**	**0.73**	**0.75**
Gray channel only	0.82	0.71	0.69	0.71
All channels together	0.8	0.68	0.64	0.68

The video categorization pipeline consists of these steps: (1) image pre-processing to align all moviebarcodes to the same shape in terms of width and length, (2) applying dimensionality reduction algorithm to all input datasets, (3) applying a clustering algorithm to group similar moviebarcodes into the same clusters, and (4) comparing cluster results with the video collection labels of videos for evaluation. The performance of the process is measured with confusion matrix [14]. We tried using many different pre-trained convolutional neural network models to extract features with only convolutional layers. However, the result matrix was sparse and did not give us good results on video categorization. Since our moviebarcode images are not natural images like ImageNet dataset [15], moviebarcodes require custom feature extraction algorithm. Instead, we decided to use one of the most important features of an image which is pixel value directly on the clustering part of the pipeline. The fine tuning of convolutional neural networks for better feature extraction and alternative video

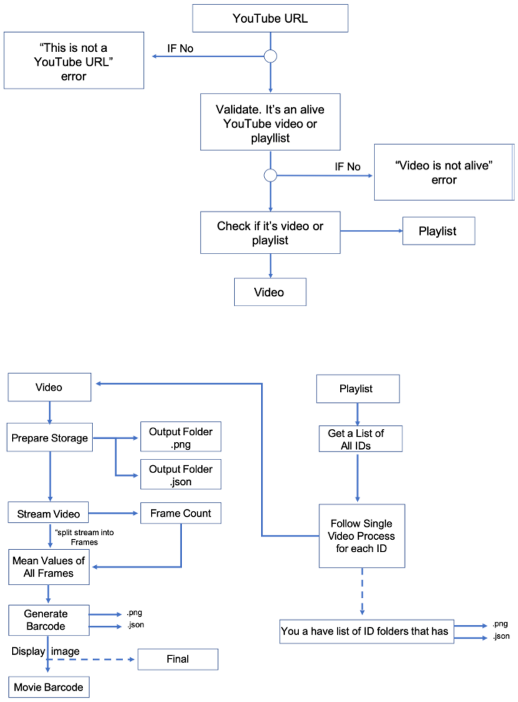

YouTube URL
IF No
"This is not a YouTube URL" error
Validate. It's an alive YouTube video or playllist
IF No
"Video is not alive" error
Check if it's video or playlist
Playlist
Video
Video
Prepare Storage
Output Folder .png
Output Folder .json
Stream Video
Frame Count
"split stream into Frames
Mean Values of All Frames
Generate Barcode
.png
.json
Display image
Final
Movie Barcode
Playlist
Get a List of All IDs
Follow Single Video Process for each ID
You a have list of ID folders that has
.png
.json

Fig. 4. Algorithm used to generate moviebarcode.

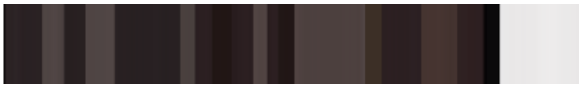


Fig. 5. A sample movidebarcode image of a video from HBOSiliconValley collection.

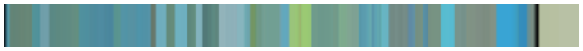


Fig. 6. A sample moviebarcode image of a video from SpongeBob-SquarePants collection.

categorization with a moviebarcode dataset is left for future studies.

Next, due to high dimension of images, we applied Principal Component Analysis (PCA) dimensionality reduction algorithm [16]. The results of this step were used during the clustering step. Due to its simple nature to implement and run, we utilized K-means clustering algorithm [17] with the cluster value as the number video collections for the clustering step. Next, we applied model evaluation with confusion matrix. We repeated the process of k-means and model evaluation 10 times and calculated the average of these experiments for the final result. The clustering results are analyzed and compared using the collection labels of videos.

This pipeline was applied on tensor of moviebarcodes which are all color channels together. After that, we repeated the process for individual channels and gray scale. The all results of moviebarcode video categorization are shown in Table 2.

Table 2 shows that red channel in moviebarcode is not a good feature to distinguish the clusters. On the contrary, blue channel has the highest scores on all metrics including precision, recall, f1-score, and accuracy. The scores for all other channels and their combinations are between red and blue channels.

Fig. 5 shows the moviebarcode of a video from the HBOSiliconValley collection. And in contrast, Fig. 6 shows the moviebarcode of a video from the SpongeBobSquarePants collection. These moviebarcodes show that it is simple to distinguish one collection from another. Also, changes in the scenes and patterns of similar frames can be clearly observed from moviebarcodes.

Moviebarcodes are useful images that can be used for information retrieval applications such as filtering or grouping images based on their color population. Additionally, the number of different colors in a moviebarcode image can be a good indicator of the pace of the video.

Copyright (c) IARIA, 2020. ISBN: 978-1-61208-800-6

V. Conclusions and Future Work

In this paper, we introduced the use of moviebarcode for video categorization and summarization. We also demonstrated that the video processing is easier with moviebarcode for social computing researchers, especially if they deal with YouTube which is the most popular video sharing platform. Our experiments focus on reducing the video to colors. Traditional techniques for video categorization are resource intensive and time consuming. Moviebarcode is a great methodology to extract insightful features by capturing visual patterns in a video without watching, and grouping or categorizing same or similar videos together in fast and efficient manner.

Results show that using individual channels of moviebarcode image helps video categorization by differentiating one video from another or grouping them. Each channel carries different features about an image. Splitting the channels of an image increased the performance of video categorization. Our findings suggests that analyzing only the colors within the video without looking the video content in detail gives the accuracy of 75%.

Video length is one of the most important features about the video. But, it is also one of the limitations of moviebarcode technique because it is difficult to align long videos with short videos. In this paper, we experimented with six video collections. In order to make our model more generalized, future research could examine the experiment pipeline of our model on other video collections. Since we use moviebarcode pixels directly on the categorization pipeline, it would be better to have a custom feature extraction method to extract more features from the moviebarcodes.

Moviebarcode technique can be used for further analysis of videos. With the acceleration of new deep learning techniques, it is easy to generate new videos artificially. To identify these artificially generated videos, moviebarcode might be a great tool to identify similar or same videos, as well as pieces of these videos as a short clip. Even though we currently use moviebarcod only video categorization, we could use them to detect scene changes and narratives by detecting changes in colors.

RGB channels are used in this study, but YCbCr or HSV color channels could also be used to categorize videos. Each color channel has different features about a video. Color theory techniques show that different color channels can be used for different purposes. With this motivation, video categorization could be examined by using other color channels different from RGB. Other data models such as transcription of a video or metadata could also be combined for video categorization. These multiple data models might boost performance of video categorization.

Acknowledgment

This research is funded in part by the U.S. National Science Foundation (IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1- 2016, N00014-16-1-2412, N00014-17-1-2605, N00014-17- 1-2675), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059) and the Jerry L. Maulden/Entergy Fund at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations.

Nomenclature

RGB: Red, Green, Blue
YCbCr: Luma, Blue-difference chroma, Red-difference chroma components
HSV: Hue, Saturation, Value
PCA: Principal Component Analysis

References

[1] P. Suciu. ”Is It Possible To Become The Next Big YouTube Star In 2020?,” *Forbes*, 03-Jan-2020.

[2] Y. Press. ”Up, Up and Away - Long videos for more users.” *YouTube*. [Online]. Available: https://youtube.googleblog.com/2010/12/up-up-and-away-long-videos-for-more.html. [Accessed: 16-Mar-2020].

[3] Y. Jiang, Z. Wu, J. Wang, X. Xue and S. Chang, ”Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks,” *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 352-364, 1 Feb. 2018, doi: 10.1109/TPAMI.2017.2670560.

[4] M. Liu, L. Nie, X. Wang, Q. Tian and B. Chen, ”Online Data Organizer: Micro-Video Categorization by Structure-Guided Multimodal Dictionary Learning.” *in IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1235-1247, March 2019, doi: 10.1109/TIP.2018.2875363.

[5] ”Moviebarcode,” *Tumblr*. [Online]. Available: https://moviebarcode.tumblr.com/. [Accessed: 16-Mar-2020].

[6] M. Burghardt, M. Kao, and C. Wolff. ”Beyond Shot Lengths – Using Language Data and Color Information as Additional Parameters for Quantitative Movie Analysis.” *In Digital Humanities 2016: Conference Abstracts*. Jagiellonian University and Pedagogical University, Kraków, pp. 753-755, 2016.

[7] M. Barbieri, G. Mekenkamp, M. Ceccarelli, and J. Nesvadba. ”The color browser: a content driven linear video browsing tool.” *IEEE International Conference on Multimedia and Expo, (ICME 2001)*, Tokyo, Japan, pp. 627-630, 2001.

[8] M. M. Yeung and Y. Boon-Lock. “Video visualization for compact presentation and fast browsing of pictorial content.” *IEEE Trans. Circuits Syst. Video Techn*, vol 7, pp. 771-785, 1997.

[9] I. Otto, A. Plutino, M. Lanaro, and A. Rizzi. ”All the colours of a film: A study on the chromatic variation of movies.” *AIC Interim Meeting*, Lisbon, Portugal, 2018.

[10] YouTube Data API. ”Add YouTube functionality to your app” *YouTube*. [Online]. Available: https://developers.google.com/youtube/v3/

[11] *OpenCV*. [Online]. Available: https://opencv.org/

[12] J. Kready, S. A. Shimray, M. N. Hussain, and N. Agarwal. ”YouTube data collection using parallel processing.” *In 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. (pp. 1119-1122), IEEE, May 2020.

[13] M. N. Hussain, S. Tokdemir, N. Agarwal, and S. Al-Khateeb. ”Analyzing disinformation and crowd manipulation tactics on YouTube.” *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018.

[14] S. Visa, B. Ramsay, A. Ralescu, and E. van der Knaap. ”Confusion matrix-based feature selection.” *MAICS*. 710, pp.120-127, 2011.

[15] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li and L. Fei-Fei. ”ImageNet: A Large-Scale Hierarchical Image Database.” *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[16] Scikit Learn. ”Decomposing signals in components (matrix factorization problems).” *Scikit Learn*. [Online]. Available: https://scikit-learn.org/stable/modules/decomposition.html#pca. [Accessed: 22-Mar-2020].

[17] Scikit Learn. ”K-means.” *Scikit Learn*. [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#k-means. [Accessed: 22-Mar-2020].

An Investigation of Twitter Users

Who Disclosed Their Personal Profile Items in Their Tweets Honestly

Yasuhiko Watanabe, Hiromu Nishimura, Yuuya Chikuki, Kunihiro Nakajima, and Yoshihiro Okada
Ryukoku University
Seta, Otsu, Shiga, Japan
Email: watanabe@rins.ryukoku.ac.jp, t160405@mail.ryukoku.ac.jp, t160389@mail.ryukoku.ac.jp,
nakajima.k216@gmail.com, okada@rins.ryukoku.ac.jp

***Abstract*—These days, many people use a Social Networking Service (SNS). Most SNS users are careful in protecting the privacy of personal information: name, age, gender, address, telephone number, birthday, etc. However, some SNS users disclose their personal information that can threaten their privacy and security even if they use non-real name accounts. In this study, we investigated tweets disclosing submitters' personal profile items which many of us think are not true. We collected 565 tweets where submitters used non-real name accounts and made promises to disclose their personal profile items, surveyed the details of their personal profile items disclosed by themselves, especially their ages, genders, and heights, and analyzed them statistically, to be specific, applied the Shapiro-Wilk test of normality and the Welch's test to them. The results of these tests showed that most of the submitters disclosed their ages, genders, and heights honestly.**

***Keywords*–personal information; Twitter; SNS; privacy risk; Shapiro-Wilk test of normality; Welch's test.**

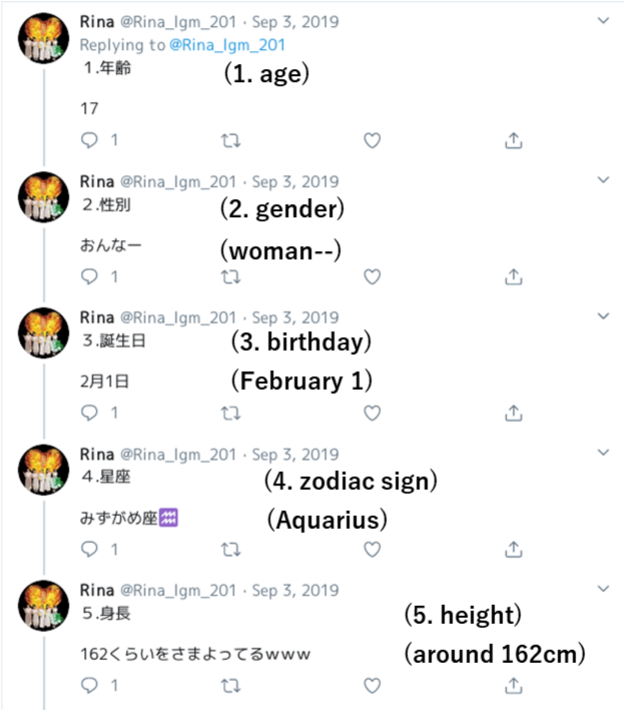
I. INTRODUCTION

These days, many people use a Social Networking Service (SNS) to communicate with each other and try to enlarge their circle of friends. SNS users are generally concerned about potential privacy risks [1]. To be specific, they are afraid that unwanted audiences will obtain information about them or their families, such as where they live, work, and play. As a result, SNS users are generally careful in disclosing their personal information. They disclose their personal information only when they think the benefits of doing it is greater than the potential privacy risks. However, some SNS users, especially young users, disclose personal information on their profiles, for example, real full name, gender, hometown and full date of birth, which can potentially be used to identify details of their real life, such as their social security numbers. In order to discuss this phenomenon, many researchers investigated how much and which type of information are disclosed in SNSs, especially, in Facebook [2] [3]. Researchers might think that personal information disclosed in Facebook is reliable, or it is possible to check whether personal information disclosed in Facebook is true. This is because

- Facebook users are required to register and disclose their real names when they first start using Facebook.
- Facebook users would be criticized by their friends if they disclose their information dishonestly.

On the other hand, a small number of researchers investigated how much and which type of information disclosed by non-real name account users, such as Twitter users. Researchers might think that personal information disclosed by non-real name account users is unreliable. This is because

- nobody criticizes non-real name account users when they disclose their personal information dishonestly.
- true personal information can threaten their privacy and security even if they use non-real name accounts.



Rina @Rina_lgm_201 · Sep 3, 2019
Replying to @Rina_lgm_201
1.年齢 (1. age)

17

1

Rina @Rina_lgm_201 · Sep 3, 2019
2.性別 (2. gender)

おんなー (woman--)

1

Rina @Rina_lgm_201 · Sep 3, 2019
3.誕生日 (3. birthday)

2月1日 (February 1)

1

Rina @Rina_lgm_201 · Sep 3, 2019
4.星座 (4. zodiac sign)

みずがめ座♒ (Aquarius)

1

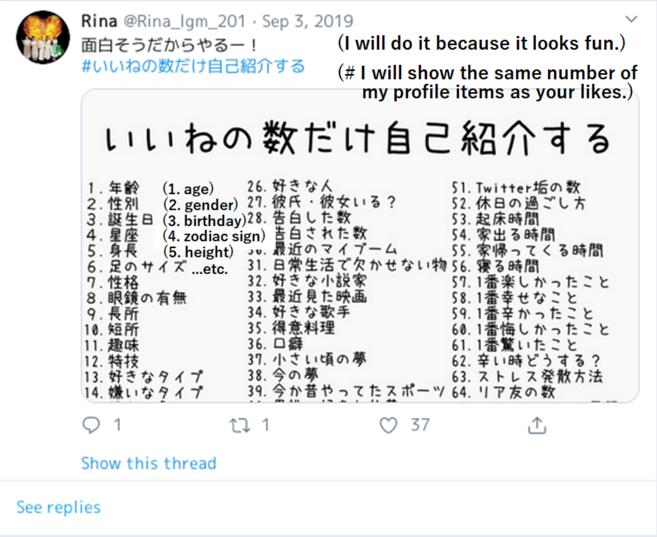
Rina @Rina_lgm_201 · Sep 3, 2019
5.身長 (5. height)

162くらいをさまよってるwww (around 162cm)

1

Figure 1. A non-real name account user, *Rina*, disclosed her personal profile items in her tweets.

As a result, many of us think that it is natural for non-real name account users not to disclose their personal information honestly. Figure 1 shows tweets submitted by non-real name account user, *Rina*. In these tweets, *Rina* disclosed her personal profile items: her age, gender, birthday, zodiac sign, and height. Many of us think that these personal profile items were not true. However, we do not check whether *Rina* disclosed her personal profile items honestly because it is difficult to do it. In this paper, we collect tweets where non-real name account users disclosed their personal profile items, analyze them statistically, and show that it is likely that most of the non-real name account users, especially young users, disclosed their



Rina @Rina_lgm_201 · Sep 3, 2019

面白そうだからやるー！ (I will do it because it looks fun.)

#いいねの数だけ自己紹介する (# I will show the same number of my profile items as your likes.)

いいねの数だけ自己紹介する

1. 年齢 (1. age)
2. 性別 (2. gender)
3. 誕生日 (3. birthday)
4. 星座 (4. zodiac sign)
5. 身長 (5. height)
6. 足のサイズ ...etc.
7. 性格
8. 眼鏡の有無
9. 長所
10. 短所
11. 趣味
12. 特技
13. 好きなタイプ
14. 嫌いなタイプ

26. 好きな人
27. 彼氏・彼女いる？
28. 告白した数
告白された数
30. 最近のマイブーム
31. 日常生活で欠かせない物
32. 好きな小説家
33. 最近見た映画
34. 好きな歌手
35. 得意料理
36. 口癖
37. 小さい頃の夢
38. 今の夢
39. 今か昔やってたスポーツ

51. Twitter垢の数
52. 休日の過ごし方
53. 起床時間
54. 家出る時間
55. 家帰ってくる時間
56. 寝る時間
57. 1番楽しかったこと
58. 1番幸せなこと
59. 1番辛かったこと
60. 1番悔しかったこと
61. 1番驚いたこと
62. 辛い時どうする？
63. ストレス発散方法
64. リア友の数

Figure 2. A tweet promising to disclose the same number of submitters' personal profile items as likes to it.

personal information honestly.

The rest of this paper is organized as follows: In Section II, we survey the related works. In Section III, we show how to collect tweets disclosing submitters' personal profile items. In Section IV, we survey the details of submitters' personal profile items, analyze them statistically, and show that it is likely that most of the submitters disclosed their personal profile items honestly. Finally, in Section V, we present our conclusions.

II. Related work

Personally identifiable information is defined as information which can be used to distinguish or trace an individual's identity such as social security number, biometric records, etc. alone, or when combined with other information that is linkable to a specific individual, such as date and place of birth, mother's maiden name, etc. [4] [5]. Internet users are generally concerned about unwanted audiences obtaining personal information. Fox et al. reported that 86% of Internet users are concerned that unwanted audiences will obtain information about them or their families [1]. Also, Acquisti and Gross reported that students expressed high levels of concern for general privacy issues on Facebook, such as a stranger finding out where they live and the location and schedule of their classes, and a stranger learning their sexual orientation, name of their current partner, and their political affiliations [2]. However, Internet users, especially young users, tend to disclose personal information on their profiles, for example, real full name, gender, hometown and full date of birth, which can potentially be used to identify details of their real life, such as their social security numbers. As a result, many researchers discussed the reasons why young users willingly disclose personal information on their SNS profiles. Dwyer concluded in her research that privacy is often not expected or undefined in SNSs [6]. Barnes argues that Internet users, especially teenagers, are not aware of the nature of the Internet and SNSs [3]. Hirai reported that many users had troubles in SNSs because they did not mind that strangers observed their communication with their friends [7]. Viseu et al. reported that many online users believe the benefits of disclosing personal information in order to use an Internet site is greater than the potential privacy risks [8]. On the other hand, Acquisti and Gross explain this phenomenon as a disconnection between the users' desire to protect their privacy and their actual behavior [2]. Also, Livingstone points out that teenagers' conception of privacy does not match the privacy settings of most SNSs [9]. Joinson et al. reported that trust and perceived privacy had a strong affect on individuals' willingness to disclose personal information to a website [10]. Also, Tufekci found that concern about unwanted audiences had an impact on whether or not students revealed their real names and religious affiliation on MySpace and Facebook [11]. The authors also think that most students are seriously concerned about their privacy and security. However, they often underestimate the risk of their online messages and submit them. For example, Watanabe et al. reported that many students submit tweets concerning school events and these tweets may give a chance to other people, including unwanted audiences, to distinguish which schools students go to [12].

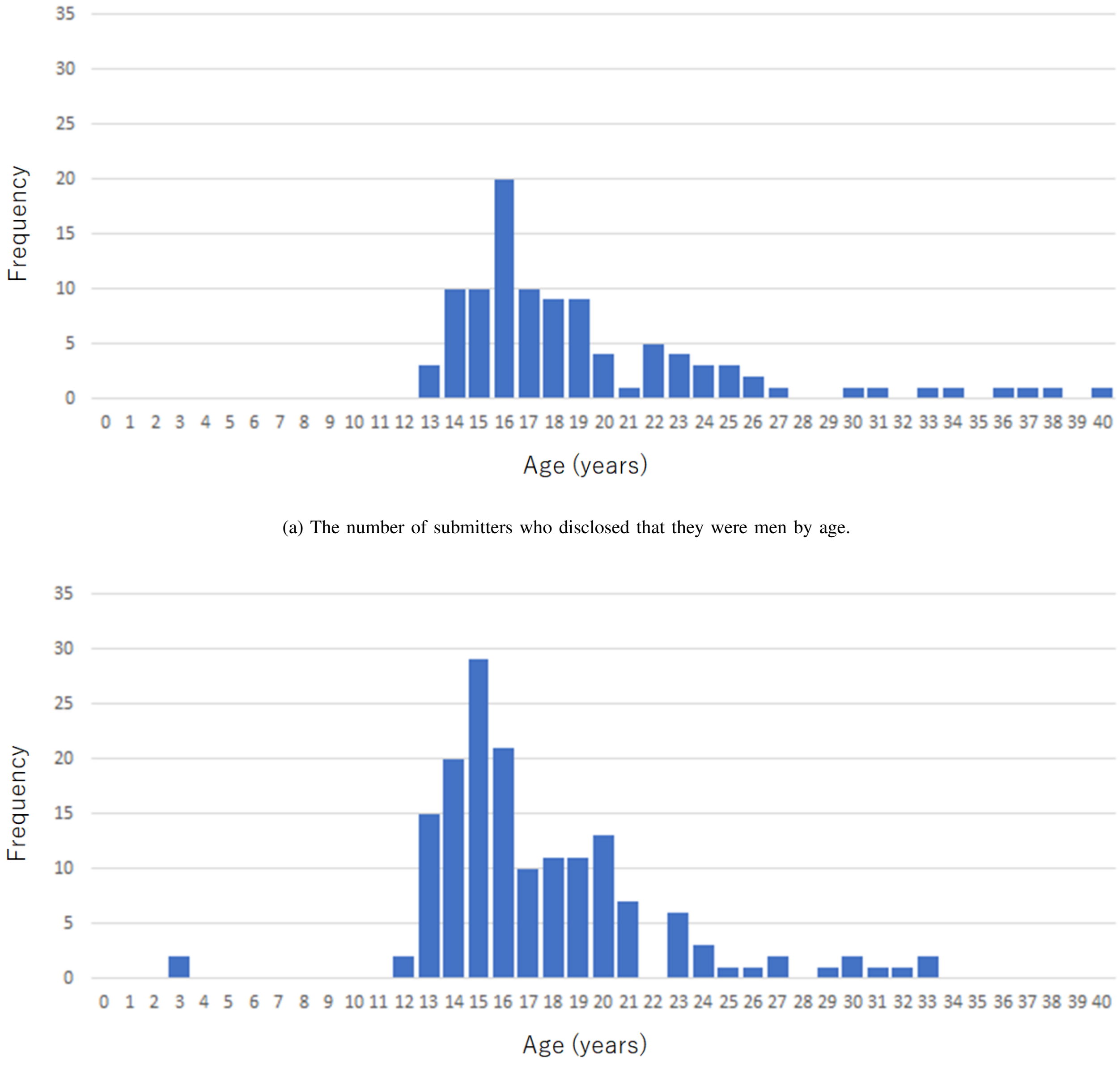
III. A collection of tweets disclosing submitters' personal profile items

It is difficult to collect tweets disclosing submitters' personal profile items, such as tweets in Figure 1, directly. To solve this problem, we focused on tweets where submitters promised their followers to disclose the same number of their own personal profile items as likes to their tweets. Figure 2 shows a tweet submitted by *Rina* on September 3, 2019. In this tweet, *Rina* promised her followers to disclose the same number of her personal profile items as likes to her tweet. Actually, *Rina* submitted 35 replies disclosing her personal profile items to her tweet shown in Figure 2 from September 3 to 9, 2019. The five tweets shown in Figure 1 were the first five replies submitted by *Rina* to her tweets shown in Figure 2. As of November 20, 2019, we confirmed that 37 likes were given to her tweet shown in Figure 2. Furthermore, we found many tweets promising to disclose the same number of their own personal profile items as likes to their tweets. As a result, it is easy to collect tweets disclosing submitters' personal profile items when we collect tweets promising to disclose submitters' personal profile items. The reasons why many Twitter users submitted tweets promising to disclose submitters' personal profile items might be

- they thought they looked fun,
- they wanted to draw attention, and
- they wanted to know how much attention was paid to their tweets.

In order to collect tweets promising to disclose submitters' personal profile items, we focused on images attached to these tweets. This is because many submitters attached the same image to their tweets and many personal profile items were listed in the image. As shown in Figure 2, *Rina* attached an image to her tweet and showed the list of personal profile items that she promised her followers to disclose in the image. Many twitter users attached the same image to their tweets promising to disclose their personal profile items. As a result, we used these shared images as key to collect tweets promising to disclose submitters' personal profile items. To be specific, we

Copyright (c) IARIA, 2020. ISBN: 978-1-61208-800-6



(a) The number of submitters who disclosed that they were men by age.

(b) The number of submitters who disclosed that they were women by age.

Figure 3. The number of submitters who disclosed their genders clearly by age.

collected these tweets by using Twigaten [13]. Twigaten helps us to collect tweets to which the same image is attached. By using Twigaten, we collected 565 Japanese tweets promising to disclose submitters' personal profile items on November 20, 2019. The obtained tweets were submitted from October 3, 2018 to November 20, 2019.

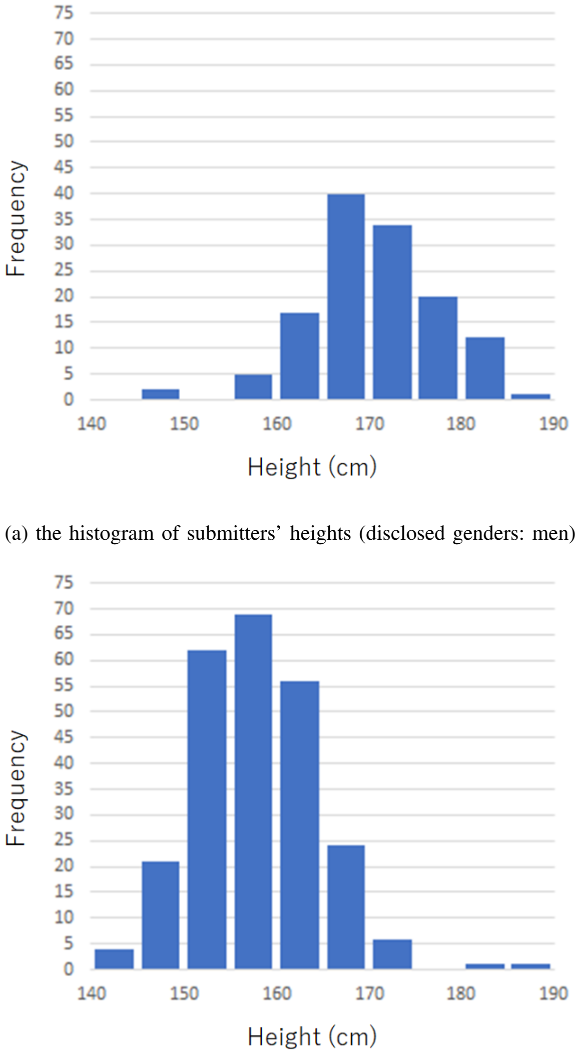
IV. An analysis of tweets disclosing submitters' personal profile items

It is difficult to determine whether an individual submitter disclosed his/her personal profile items honestly. For example, it is difficult to determine whether *Rina*, who submitted tweets in Figure 1 and Figure 2, was a woman. In this study, we discuss whether submitters disclosed their personal profile items honestly when they made promises to disclose them. In order to discuss this problem, we analyze submitters' genders, ages, and heights statistically.

A. Submitters' genders

As mentioned in Section III, we obtained the 565 tweets promising to disclose their personal profile items. We surveyed these 565 tweets and their replies and, according to submitters' genders disclosed in the replies, classified them into

- 282 tweets (women)
- 156 tweets (men)
- 27 tweets (unclear)
- 100 tweets (no replies)



(a) the histogram of submitters' heights (disclosed genders: men).

(b) the histogram of submitters' heights (disclosed genders: women).

Figure 4. The histogram of heights of submitters who disclosed their genders, men or women, clearly. (bin width = 5cm)

B. *Submitters' ages*

We also surveyed the 565 tweets and their replies and, according to whether submitters' ages were disclosed in their replies clearly, classified them into

- 276 tweets (clearly)
- 60 tweets (unclearly)
- 229 tweets (no replies)

When submitter's age was disclosed such as "early 20s" and "over thirty", we determined that submitter's age was disclosed unclearly. Among the 276 tweets where submitters' ages were disclosed clearly, we found 102 and 161 tweets where submitters' genders were also disclosed clearly, men and women, respectively. Figure 3 shows the number of submitters, who disclosed their genders clearly, men and women, by age. As shown in Figure 3, the most popular age of men and women were 16 and 15 years old, respectively.

TABLE I. The results of the Shapiro-Wilk test of normality

gender	age	sample size	W value	p-value
men	15	10	0.885	0.147
men	16	18	0.929	0.190
men	17	9	0.977	0.946
women	14	17	0.933	0.244
women	15	24	0.971	0.697
women	16	19	0.961	0.587

C. *Submitters' heights*

We also surveyed the 565 tweets and their replies and, according to whether submitters' heights were disclosed in their replies clearly, classified them into

- 401 tweets (clearly),
- 8 tweets (unclearly), and
- 156 tweets (no replies).

Among the 401 tweets where submitters' heights were disclosed clearly, we found 131 and 244 tweets where submitters' genders were disclosed clearly, men and women, respectively. Figure 4 shows the histogram of heights of submitters who disclosed their genders, men or women, clearly.

It is difficult to determine whether an individual submitter disclosed his/her personal profile items honestly. In this study, we statistically examine whether submitters disclosed their personal profile items honestly when they made promises to disclose their personal profile items and disclosed them in the same way as *Rina* did.

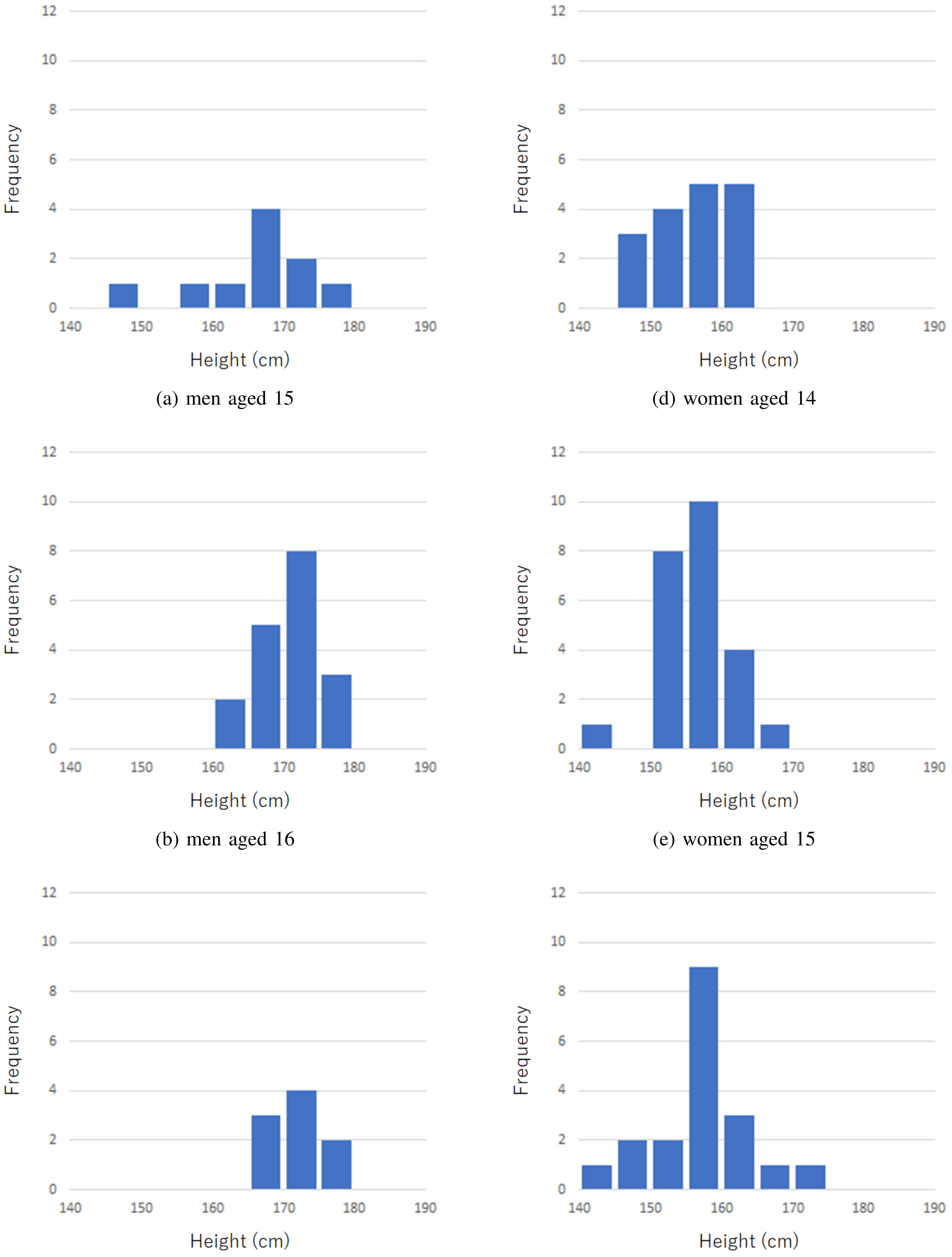
It is well known that our heights follow a normal (Gaussian) distribution [14]. As a result, if most of submitters disclose their ages, genders, and heights honestly, their heights would follow a normal distribution. Also, the average of their heights would be equal to the national average height in Japan. To solve this problem, in this paper, we conduct the statistical analysis on

- 37 submitters who disclosed their genders (men), ages (15-17 years old), and heights clearly, and
- 60 submitters who disclosed their genders (women), ages (14-16 years old), and heights clearly.

As shown in Figure 3, men aged 15-17 and women aged 14-17 were the most popular segments in the submitters' ages.

First, we discuss whether submitters' heights followed a normal distribution. Figure 5 shows the histograms of their heights. In order to discuss whether submitters' heights followed a normal distribution, we conducted the Shapiro-Wilk test of normality. The null hypothesis in this study was that submitters' heights followed a normal distribution. Table I shows the results of the Shapiro-Wilk test of normality. As shown in Table I, the p-value in each case was greater than 0.05. As a result, the null hypothesis in each case was not rejected. In other words, submitters' heights, in each case of men aged 15-17 and women aged 14-16, followed a normal distribution.

Next, we discuss whether the average of submitters' heights was equal to the national average height in Japan. Table II shows the average of submitters' heights. Table III shows the national average height in Japan [15]. In order to discuss



(a) men aged 15

(b) men aged 16

(c) men aged 17

(d) women aged 14

(e) women aged 15

(f) women aged 16

Figure 5. The histograms of heights of submitters who disclosed that they were men aged 15-17 and women aged 14-16 (bin width = 5cm).

TABLE II. THE AVERAGE AND STANDARD DEVIATION OF SUBMITTERS' HEIGHTS

gender	age	sample size	average	standard deviation
men	15	10	165.5	7.58
men	16	18	169.2	4.08
men	17	9	171.3	3.84
women	14	17	155.0	6.11
women	15	24	155.7	4.99
women	16	19	156.9	6.54

TABLE III. THE NATIONAL AVERAGE AND STANDARD DEVIATION OF HEIGHTS IN JAPAN

gender	age	sample size	average	standard deviation
men	15	1411	168.37	5.75
men	16	1428	169.59	5.70
men	17	1427	170.46	5.82
women	14	1386	156.36	5.24
women	15	1413	156.76	5.36
women	16	1419	157.16	5.17

TABLE IV. THE RESULTS OF WELCH'S TEST

gender	age	Degrees of freedom	test statistic T	p-value
men	15	9.07	1.195	0.262
men	16	17.84	0.380	0.708
men	17	8.23	-0.675	0.518
women	14	16.29	0.914	0.374
women	15	23.91	1.060	0.300
women	16	18.30	0.179	0.860

whether the average of their heights was equal to the national average height in Japan, we conducted the Welch's test. The null hypothesis in this study was that the average of submitters' heights was equal to the national average height in Japan. Table IV shows the results of the Welch's test. As shown in Table IV, the p-value in each case was greater than 0.05. As a result, the null hypothesis in each case was not rejected. In other words, in each case of men aged 15-17 and women aged 14-16, the average of submitters' heights was equal to the national average height in Japan.

The results of the Shapiro-Wilk test of normality and the Welch's test rarely happened when many submitters disclosed their ages, genders, and heights dishonestly. As a result, it is assumed that most of the submitters disclosed their ages, genders, and heights honestly. Furthermore, age, gender, and height were important personal information. It is likely that they disclosed not only their ages, genders, and heights but also other personal profile items honestly.

V. CONCLUSION

In this paper, we investigated tweets disclosing submitters' personal profile items and analyzed submitters' ages, genders, and heights statistically. The results of the statistical analysis showed that it is likely that most of the submitters disclosed their personal profile items honestly. These personal profile items can threaten their privacy and security even if they use non-real name accounts. We are investigating whether submitters were concerned about their privacy and security risks caused by submitting tweets disclosing their personal profile items honestly. Furthermore, we intend to conduct the same statistical analysis on tweets in languages other than Japanese.

REFERENCES

[1] S. Fox et al., Trust and Privacy Online: Why Americans Want to Rewrite the Rules, The Pew Internet & American Life Project, 2000. [Online]. Available: http://www.pewinternet.org/2000/08/20/trust-and-privacy-online/ [accessed: 2020-09-01]

[2] A. Acquisti and R. Gross, Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 36–58.

[3] S. B. Barnes, "A privacy paradox: Social networking in the United States." First Monday, vol. 11, no. 9, 2006. [Online]. Available: http://firstmonday.org/article/view/1394/1312 [accessed: 2020-09-01]

[4] C. Johnson III, Safeguarding against and responding to the breach of personally identifiable information, Office of Management and Budget Memorandum, 2007. [Online]. Available: http://www.whitehouse.gov/omb/memoranda/fy2007/m07-16.pdf [accessed: 2016-10-04]

[5] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," Computer Communication Review, vol. 40, no. 1, 2010, pp. 112–117. [Online]. Available: https://doi.org/10.1145/1672308.1672328 [accessed: 2020-09-01]

[6] C. Dwyer, "Digital relationships in the "myspace" generation: Results from a qualitative study," in Proceedings of the 40th Annual Hawaii International Conference on System Sciences, ser. HICSS '07. Washington, DC, USA: IEEE Computer Society, 2007, p. 19.

[7] T. Hirai, "Why does "Enjyo" happen on the Web? : An Examination based on Japanese Web Culture," Journal of Information and Communication Research, vol. 29, no. 4, mar 2012, pp. 61–71. [Online]. Available: http://doi.org/10.11430/jsicr.29.4_61 [accessed: 2020-09-01]

[8] A. Viseu, A. Clement, and J. Aspinall, "Situating privacy online: Complex perception and everyday practices," Information, Communication & Society, 2004, pp. 92–114.

[9] S. Livingstone, "Taking risky opportunities in youthful content creation: teenagers' use of social networking sites for intimacy, privacy and self-expression." New Media & Society, vol. 10, no. 3, 2008, pp. 393–411.

[10] A. N. Joinson, U.-D. Reips, T. Buchanan, and C. B. P. Schofield, "Privacy, trust, and self-disclosure online." Human-Computer Interaction, vol. 25, no. 1, 2010, pp. 1–24. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/07370020903586662 [accessed: 2020-09-01]

[11] Z. Tufekci, "Can You See Me Now? Audience and Disclosure Regulation in Online Social Network Sites," Bulletin of Science, Technology & Society, vol. 28, no. 1, 2008, pp. 20–36.

[12] Y. Watanabe, H. Onishi, R. Nishimura, and Y. Okada, "Detection of school foundation day tweets that can be used to distinguish senders' schools," in Proceedings of the Eleventh International Conference on Evolving Internet (INTERNET 2019), Nov 2016, pp. 30–35. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=internet_2019_2_30_40026 [accessed: 2020-09-01]

[13] twigaten.204504byse.info. TwiGaTen. [Online]. Available: https://twigaten.204504byse.info/ [accessed: 2020-09-01]

[14] National Centre for Research Methods (NCRM). Using Statistical Regression Methods in Education Research. [Online]. Available: http://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/index.html [accessed: 2020-09-01]

[15] the Ministry of Education, Culture, Sports, Science and Technology (MEXT). the survey on physical strength and sporting ability (2018). [Online]. Available: https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00402102&tstat=000001088875&cycle=0&tclass1=000001133904&stat_infid=000031872003 [accessed: 2020-09-01]

Emoji as Sentiment Indicators: An Investigative Case Study in Arabic Text

Shatha Ali A. Hakami
School of & Dept. of Computer Science
University of Birmingham &
Jazan University
Birmingham, UK / Jazan, Saudi Arabia
e-mails: sah624@cs.bham.ac.uk
sahakami@jazanu.edu.sa

Robert Hendley
School of Computer Science
University of Birmingham
Birmingham, UK
e-mail: r.j.hendley@cs.bham.ac.uk

Phillip Smith
School of Computer Science
University of Birmingham
Birmingham, UK
e-mail: p.smith.7@cs.bham.ac.uk

***Abstract*—With the explosion of social media usage, researchers have become interested in understanding and analysing the sentiment of the language used in textual digital communications. One particular feature is the use of emoji. These are pictographs that are used to augment the text. They might represent facial expressions, body language, emotional intentions or other things. Despite the frequency with which they are used, research on the interpretation of emoji in languages other than English, such as Arabic, is still in its infancy. This paper analyses the use of emoji in Arabic social media datasets to build a better understanding of sentiment indicators in textual contents. Seven benchmark Arabic datasets containing emoji were manually and automatically annotated for sentiment value. A quantitative analysis of the results shows that emoji are sometimes used as true/direct sentiment indicators. However, the analysis also reveals that, for some emoji and in some contexts, the role of emoji is more complex. They may not act as sentiment indicators, they may act as modifiers of the sentiment expressed in the text or, in some cases, their role may be context dependent. It is important to understand the role of emoji in order to build sentiment analysis systems that are more accurate and robust.**

***Keywords—Emoji*; *Social Media*; *Arabic*; *NLP*; *Sentiment Analysis*.**

I. Introduction

Natural human communication involves both verbal (natural language) and nonverbal channels. In face-to-face communication, nonverbal cues are often the meta-messages that instruct receivers on how to interpret verbal messages. These cues can be either visual/mimogestual (the use of the body), like head nodding, facial expressions, posture, mime, gaze, and eye contact [1]; or oral/prosodic (the use of the voice), like pitch contour, tone, stress, pause, rhythm, tempo and vocal intonation [2]. Ambady et al. [3] also consider these nonverbal cues as reliable indicators for attributes of the speaker, such as gender, personality, abilities, and sexual orientation. The main feature of nonverbal cues, however, is their "ability to convey emotions and attitude" as well as to "emphasize, contradict, substitute or regulate verbal communication" [4]. From a Psycholinguistic perspective, Mehrabian [5] argues that 93% of human communication takes place non-verbally.

In text-based communication, it has been argued that many of these nonverbal cues are missed, which potentially makes the communication ambiguous and inefficient and can lead to misunderstandings [6]. To address this issue, people often use many kinds of text-based surrogates, such as nonstandard/multiple punctuation (e.g., '...', or '!!!'), lexical surrogates (e.g., 'hmmm', or 'yummm'); asterisks (e.g., '*hug*' or '*grin*'), emoticons (e.g., ':)' or ':('), and emoji (e.g., '😊' and '😡'). Carey [7] categorized these nonverbal cues into five types: vocal spelling, lexical surrogates, spatial arrays (e.g., using the textual layout to aid understanding or provide emphasis), manipulation of grammatical markers, and minus features. Emoticons, and later emoji, are sometimes considered as examples of spatial arrays that are used to convey emotion or sentiment. Sentiment analysis can be defined as a process that analyses text and builds an interpretation of the sentiment that it is intended to convey. Usually, this is a one dimensional measure from negative to positive and often it is quantized to just three values: negative, neutral or positive. Sentiment analysis has become an important tool in classifying and interpreting text. It has important applications in social media analysis, consultation systems, text classification and many other areas.

Generally, there are two broad approaches to analyzing sentiment in text: a machine learning approach and a lexicon-based approach. The conventional automated sentiment analysis, that takes account of emoji, especially in the Arabic language, works as follows: the text is analysed to calculate a value representing the sentiment of the text, any emoji are analysed to derive their sentiment values, and then the two values are combined to build an overall interpretation of the sentiment of the whole text.

This conventional assumption might not always be correct. Emoji do not always just indicate additional emotional content. It has been noticed in [8]–[11] that emoji often play sentiment roles other than as a direct indication. For instance, a negative emoji (e.g., broken-heart 💔) can disambiguate an ambiguous sentiment in a text (i.e., add negativity to neutral sentiment texts), it can also complement it in a relatively positive text. Kunneman et al. [11] discussed a similar duality of sentiment role in the use of emotional hashtags such as #nice and #lame. Since this information is not explicit, we assume that the role of emoji as a sentiment signal needs to be examined using various approaches and in different contexts, in order to build a better understanding.

In this work, we seek to investigate the interpretation of the sentiment expressed in informal Arabic texts, which contain emoji and are drawn from a Twitter dataset. This is done by trying to answer, from a broad perspective, the following questions:

> **Q1:** When is it appropriate, in sentiment analysis, to use the conventional techniques for interpreting emoji (i.e., when are they a true sentiment indicator within the text)?
>
> **Q2:** What are the other, unconventional, cases of emoji in sentiment analysis, and when do they apply?

To answer these questions, we borrow from [8] the argument that each emoji has three different norms of sentiment within itself. These are positivity, neutrality, and negativity. Thus, we cannot merely consider a single emoji to be a representative or an indicator of one absolute sentiment (positive, negative, or neutral) unless we examine its sentiment state within that related context. Indeed, arguably, within a textual context, some emoji can mislead the sentiment analysis process.

Here, we propose an investigation that uses a comparison between the sentiment of text with and without emojis as well as of the sentiment of the emoji on its own. We apply this approach with 496 different emojis that are used in a corpus of 5204 Arabic texts, annotated with sentiment labels. As a result, we identify four cases for the roles of emoji as sentiment indicators. These cases are as: true sentiment indicators, multi-sentiment indicators, ambiguous sentiment indicators, and not sentiment indicators.

The rest of this paper is organized as follows. Section II reviews related work upon which we build; Section III presents the study's design; Section IV presents the results, analysis and discussion. Finally, in Section V we draw conclusions from this work along with its weaknesses and limitations as well as some recommendations for future work.

II. Related Work

Previous studies on emoji within texts have attempted to explore their roles as nonverbal cues and as sentiment indicators.

A. *Emoji as Textual Nonverbal Cues*

Emoticons are a sequence of keyboard characters (ASCII characters) that represent nonverbal behaviors, such as facial expressions. Emojis are, in many ways, a successor to emoticons with more sophisticated rendering and a wider repertoire but they often play a similar role. In practice, emoji are actual icons that appear on physical or virtual keyboards and can be used across various platforms, such as WhatsApp, Twitter, Facebook, Instagram, and others. These icons can represent facial expressions, body language, food, animals, places, and natural objects like flowers and trees. As discussed by Denis [12] and Zwaan and Singe [13], the human brain instantly analyzes image elements whilst it processes language linearly. That is to say: the human brain processes visual elements faster than written text. Many major technology companies, like Apple and Microsoft, have realized this importance of emoji and have taken considerable strides towards developing them in their systems.

Dresner and Herring [14] and Skovholt et al. [15] have observed that including emoticons, as well as emojis, in text not only helps the receivers to infer some contextual information, but it also eases understanding of the expressed sentiment. Therefore, it has become necessary to integrate the analysis of textual content and emoji in order to properly undertake sentiment analysis. Accordingly, Evans [16] defined emoji as a form of developed punctuation (the way of encoding nonverbal prosody cues in writing systems) that supplements written language to facilitate the writers articulating their emotions in text-based communication.

Also, Miller et al. [17] considered the use of emoji to be understood as ”visible acts of meaning”. As defined by Bavelas and Chovil [18], visible acts of meanings are analogically encoded symbols that are sensitive to a sender-receiver relationship, and they are fully integrated with the accompanying words. Indeed, the sender-receiver cultural background is one of the essential contextualization aspects that might affect emoji-text sentiment analysis. For that, Gao and VanderLaan [19] presented a study suggesting that Eastern and Western cultures are different in their use of mouth versus eye cues when interpreting emotions. According to the study, the norm in Western cultures is to display the overt emotion while in Eastern cultures, the norm is to present more subtle emotion to other people. Westerners interpret facial emotional expressions through the mouth region. Conversely, Eastern cultures focus more on the eyes. The researchers of the study also found that such differences extend to written paralinguistic signals such as emojis and, consequently, this has implications for digital communication.

B. *Emoji as Textual Sentiment Indicators*

Studies on emoji within textual context mainly focus on three directions: the usages of emoji, their meaning and the sentiment they convey. Researchers have found that emoji can be used to disambiguate the intended sense [20], manipulate the original meaning [21][22], or add sentiment to a message [23].

Regarding sentiment analysis, some studies' findings suggest that the level of sentiment perceived from a text increases with the inclusion of facial-emojis [8][23][24] and [25]. Moreover, Rathan et al. [26] considered facial-emoji as a direct sentiment indicator. In their approach, they used emoji as a sentiment source to evaluate social media messages containing particular brands' names. Furthermore, Riordan [20] found that even non-facial emoji can increase the sentiment and improve the clarity of texts.

Going a step further, many studies have assumed emoji to be a reliable ground truth for the sentiment. For example, researchers in the work [27]–[29] followed the same approach by constructing datasets for sentiment prediction and using a set of emoji to label their datasets automatically. Despite its intuitiveness, this assumption seems insufficient since it ignores that the emoji-text sentiment correlation is context-sensitive. Therefore, approaches relying on such an assumption might yield arbitrary and inaccurate sentiment annotation. Besides, it

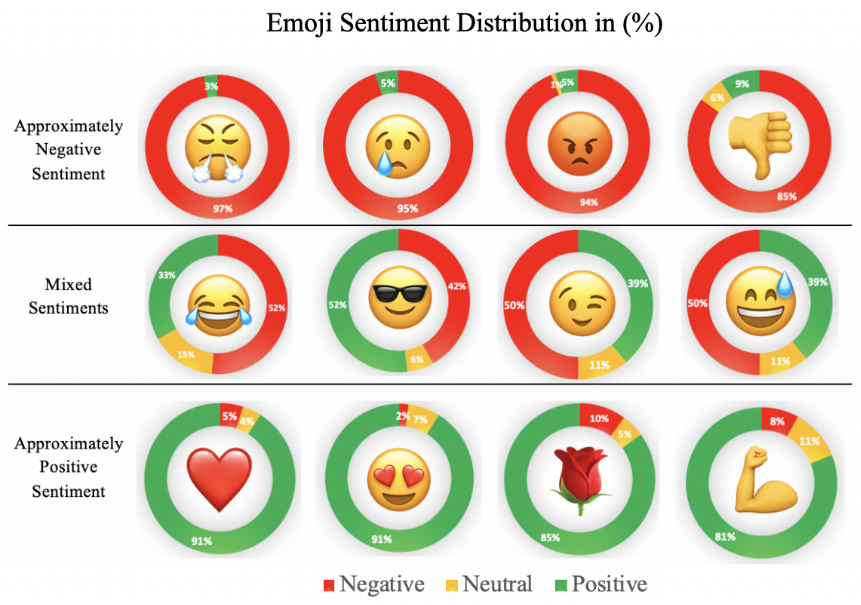


Figure 1. Examples of the Most Representative Emoji for Each Sentiment in Emoji-Text Dataset. The Percentage (%) Shows the Relative Frequency of the Sentiment Class of the Text within Which Each Emoji Occurs.

has been shown that the sentiments of surrogates for nonverbal cues (like emoji) and verbal messages (the accompanying text) are not isolated, and they should be integrated as a whole forming a context with a particular sentiment [30][31].

In line with this hypothesis, Novak et al. [8] conducted a study, which considers context-sensitivity when analyzing the sentiment of emoji and texts. In the study, the researchers annotated a collection of tweets containing at least one emoji, with sentiment labels (negative, neutral, positive). From that textual content, the researchers computed and presented sentiment ranking scores for 751 emoji. Their work illustrated that while some emoji have very high sentiment scores with little variance, others were often used to denote both positive and negative sentiment. These observations suggest that treating emoji as a direct sentiment signal is misleading because they are often full of nuanced details that are highly context dependent.

Overall, it is clear that the conventional approach of performing separate sentiment analysis of text and emoji and then combining the two to generate an overall value, is inadequate. Sometimes this approach will work. However, often and in particular with some frequently used emoji and in some critical cases, this approach fails. Furthermore, in some language such as Arabic, there is little research and also there is evidence that emoji play an especially strong sentiment indication role. The aim of this research is to close that gap.

III. Study Design

We argue that each emoji can have a different sentiment effect on a text, depending upon the context in which it appears. This is a micro-level linguistic phenomenon so, along with the standard natural language processing approach (sentiment analysis), we also used a technique from computer-mediated discourse analysis: “Coding and Counting” [32]–[34]. This is defined by Herring et al. [35] as consisting of three phases: observe, code, and count. It starts with purely qualitative observation and ends with a set of relative frequencies.

A. Data for Observation

To observe how emoji behave as a sentiment indicator for a text, content with specific criteria is needed. The content should be from a social media platform, written in the Arabic language, multi-dialect, multi-aspect, and, more importantly, should contain emoji. Therefore, the main focus of our observation was on 5402 texts (tweets from the Twitter platform), each with at least one emoji. These were extracted from seven different public datasets of Arabic social media [36]–[43]. We refer to this as the Emoji-Text dataset.

Then, we extracted all of the emoji from the Emoji-Text dataset to form a collection of 496 unique emoji. We refer to this as the Emoji-only dataset.

Lastly, a third dataset was constructed, which consists of all the texts in the Emoji-Text dataset, with the emoji removed. We refer to this dataset as the Plain-Text dataset.

B. Coding with Sentiment

In order to understand the way in which the emoji affects the interpretation of the sentiment of each text, we need to have a sentiment annotation for each item in each of the datasets.

All of the texts in the Emoji-Text dataset were human annotated with either sentiment labels (negative, positive, or neutral), or emotional labels (angry, sadness, or joy). For simplicity, we unified all the labels to be in the sentiment label form. The negative emotional labels ‘angry’ and ‘sadness’ were labelled as *negative*, and the positive emotional label ‘joy’ as *positive*.

ISBN: 978-1-61208-800-6

TABLE I. THE TOP 5 EMOJI IN EMOJI-ONLY DATASET WITH SENTIMENT FREQUENCY (Fr.) AND RELATIVE FREQUENCY (RelFr.).

Emojis	Name	Class	Sentiment	Total	W/ Negative Texts Fr.(RelFr.)	W/ Neutral Texts Fr.(RelFr.)	W/ Positive Texts Fr.(RelFr.)
😂	Face with Tears of Joy	Facial Expression	Positive	2,270	1,229 (54.14%)	92 (4.05%)	949 (41.80%)
❤️	Red Heart	Heart	Positive	765	45 (5.88%)	20 (2.61%)	700 (91.50%)
🇸🇦	Saudi Arabia	Flag	Positive	733	89 (12.14%)	29 (3.95%)	615 (83.90%)
😍	Smiling Face with Heart-Eyes	Facial Expression	Positive	426	21 (4.93%)	15 (3.52%)	390 (91.55%)
💔	Broken Heart	Heart	Negative	410	286 (69.75%)	16 (3.90%)	108 (26.34%)

TABLE II. THE FREQUENCY (Fr.) AND RELATIVE FREQUENCY (RelFr.) OF SENTIMENTS IN THE PLAIN-TEXT, EMOJI-TEXT AND EMOJI-ONLY DATASETS.

Sentiment Label	Plain-text Fr.(RelFr.)	Emoji-text Fr.(RelFr.)	Emoji-only Fr.(RelFr.)
Negative	2045 (39%)	1885 (36%)	4016 (31%)
Neutral	1119 (22%)	965 (19%)	2547 (20%)
Positive	2040 (39%)	2354 (45%)	6244 (49%)
Total	5,204	5,204	12,807

TABLE III. THE FREQUENCY (Fr.) AND RELATIVE FREQUENCY (RelFr.) OF SENTIMENTS IN THE EMOJI-TEXT DATASET WITH DIFFERENT EMOJI LOAD.

Emoji Load	Total Text Fr(RelFr.)	Neg. Text Fr.(RelFr.)	Neut. Text Fr.(RelFr.)	Pos. Text Fr.(RelFr.)
1	2283 (44%)	908 (40%)	436 (19%)	939 (41%)
2	1358 (26%)	467 (34%)	233 (17%)	658 (48%)
3	652 (12%)	261 (40%)	77 (12%)	314 (48%)
4	393 (8%)	112 (28%)	94 (24%)	187 (48%)
5 or more	518 (10%)	137 (26%)	125 (24%)	256 (49%)

For the emoji, each emoji in the Emoji-only dataset was manually annotated. This was done independently by three native Arabic speaking annotators, two females and one male. To test the reliability of this coding process, we used the inter-rater Fleiss' Kappa agreement test [44]. The test resulted in k = 0.85, which is interpreted as a general high agreement among the three annotators. In cases where two annotators disagreed on a specific sentiment, the annotation from the third annotator was considered to determine the decision.

Lastly, for the text only, we labelled each text in the Plain-text dataset with sentiment. An automatic sentiment annotation process was applied using the Python based Arabic sentiment analysis model, Mazajak [45].

C. Frequency and Relative Frequency Counting

To understand how each emoji is associated with each sentiment class, we undertook a frequency analysis of the Emoji-Text dataset. This identifies the frequency with which each emoji is associated with (human annotated) text labelled as negative, neutral and positive. We calculate two measures, the frequency (Fr), which is the absolute number of times that that emoji occurred within text of that sentiment class and also the relative frequency (RelFr), which is the proportion of the occurrences of that emoji that fall into that class. Table I shows the results for the 5 most common emojis in our data.

A similar process was repeated for each of the datasets Emoji-Text, Plain-Text and Emoji-only, in order to understand how the distribution of the sentiment annotation varied between the three sentiment classes. The results are shown in Table II.

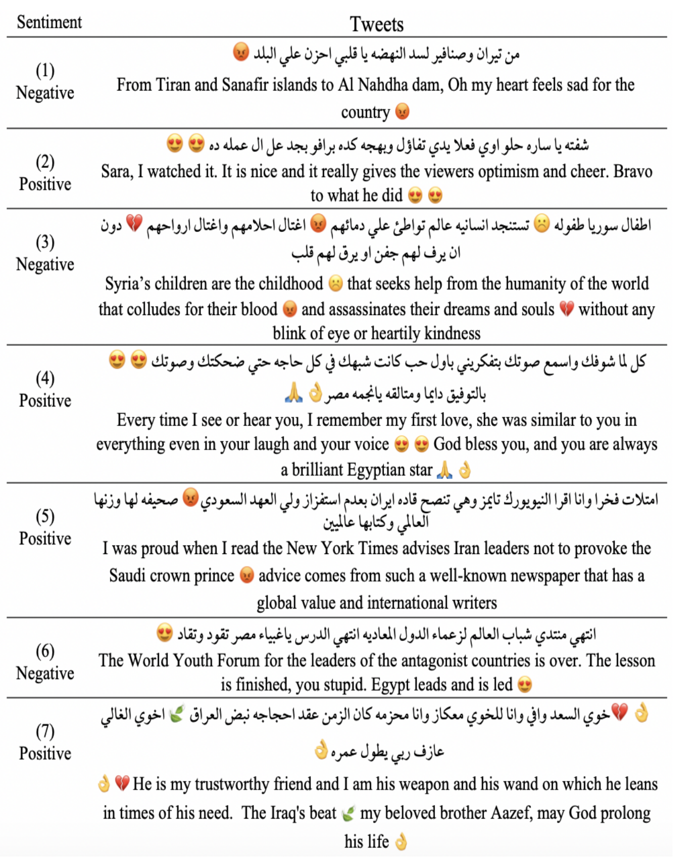
Finally, the number of emoji occurring in each text is counted. This is referred to as the "emoji load" of that text. The Fr and RelFr distributions of each emoji load for each of the three sentiment norms is then calculated to explore how sentiment varies with emoji load. This is shown in Table III.

IV. Results Analysis and Discussion

Table II shows the results of counting the frequency of texts in each sentiment class, both with and without emoji, besides the counting of the emoji only. The results show that for the negative and neutral classes there was a decrease in frequency of 3% when the emoji were included in the text. However, the number of texts classified as positive was increased by 6% when the emoji were included. In Table III, we show the emoji load across all texts and broken down by sentiment class. It is clear that the most usual usage is to include just one or sometimes two emoji in a text. The number of texts in the dataset with three or more emoji is much lower. It is also clear that, as the number of emoji in a text increases, the balance between the sentiment classes changes significantly. The proportion of negative texts is much lower when there are 3 or more emoji than when there are just 1 or 2. Similarly, the proportion of neutral or positive texts increases. This may reflect that, for negative texts, it is sufficient to use one emoji to signal the negative sentiment in Arabic. Whereas, for a positive sentiment, additional emoji are used to provide more emphasis.

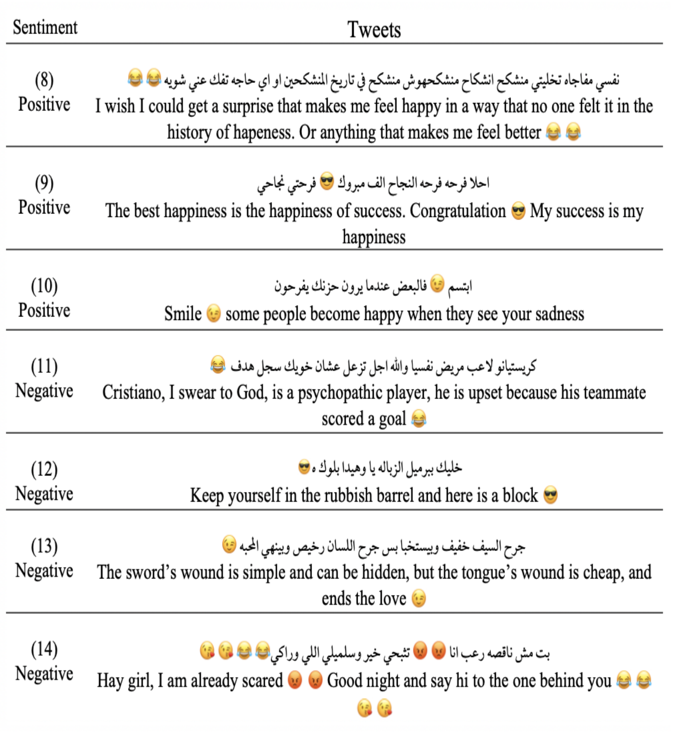
Based on this quantitative observation, we analyzed the emoji textual behavior as sentiment indicators and noticed the following significant cases.

TABLE IV. EXAMPLES FROM EMOJI-TEXT DATASET (1).



Sentiment	Tweets
(1) Negative	من تيران وصنافير لسد النهضه يا قلبي احزن علي البلد 😡 From Tiran and Sanafir islands to Al Nahdha dam, Oh my heart feels sad for the country 😡
(2) Positive	شفته يا ساره حلو اوي فعلا يدي تفاؤل وبهجه كده برافو بجد عل ال عمله ده 😍 😍 Sara, I watched it. It is nice and it really gives the viewers optimism and cheer. Bravo to what he did 😍 😍
(3) Negative	اطفال سوريا طفوله ☹️ تستنجد انسانيه عالم تواطئ علي دمائهم 😡 اغتال احلامهم واغتال ارواحهم 💔 دون ان يرف لهم جفن او يرق لهم قلب Syria’s children are the childhood ☹️ that seeks help from the humanity of the world that colludes for their blood 😡 and assassinates their dreams and souls 💔 without any blink of eye or heartily kindness
(4) Positive	كل لما شوفك واسمع صوتك بتفكريني باول حب كانت شبهك في كل حاجه حتي ضحكتك وصوتك 😍 😍 بالتوفيق دايما ومتالقه يانجمه مصر 👌 🙏 Every time I see or hear you, I remember my first love, she was similar to you in everything even in your laugh and your voice 😍 😍 God bless you, and you are always a brilliant Egyptian star 🙏 👌
(5) Positive	امتلات فخرا وانا اقرا النيويورك تايمز وهي تنصح قاده ايران بعدم استفزاز ولي العهد السعودي 😡 صحيفه لها وزنها العالمي وكتابها عالميين I was proud when I read the New York Times advises Iran leaders not to provoke the Saudi crown prince 😡 advice comes from such a well-known newspaper that has a global value and international writers
(6) Negative	انتهي منتدي شباب العالم لزعماء الدول المعاديه انتهي الدرس ياغبياء مصر تقود وتقاد 😍 The World Youth Forum for the leaders of the antagonist countries is over. The lesson is finished, you stupid. Egypt leads and is led 😍
(7) Positive	👌 💔خوي السعد وافي وانا للخوي معكاز وانا محزمه كان الزمن عقد احجاجه نبض العراق 🍃 اخوي الغالي عازف ربي يطول عمره👌 👌 💔 He is my trustworthy friend and I am his weapon and his wand on which he leans in times of his need. The Iraq's beat 🍃 my beloved brother Aazef, may God prolong his life 👌

TABLE V. EXAMPLES FROM EMOJI-TEXT DATASET (2).



Sentiment	Tweets
(8) Positive	نفسي مفاجاه تخليتي منشكح انشكاح منشكحهوش منشكح في تاريخ المنشكحين او اي حاجه تفك عني شويه 😂😂 I wish I could get a surprise that makes me feel happy in a way that no one felt it in the history of hapeness. Or anything that makes me feel better 😂 😂
(9) Positive	احلا فرحه فرحه النجاح الف مبروك 😎 فرحتي نجاحي The best happiness is the happiness of success. Congratulation 😎 My success is my happiness
(10) Positive	ابتسم 😉 فالبعض عندما يرون حزنك يفرحون Smile 😉 some people become happy when they see your sadness
(11) Negative	كريستيانو لاعب مريض نفسيا والله اجل تزعل عشان خويك سجل هدف 😂 Cristiano, I swear to God, is a psychopathic player, he is upset because his teammate scored a goal 😂
(12) Negative	خليك ببرميل الزباله يا وهيدا بلوك ه😎 Keep yourself in the rubbish barrel and here is a block 😎
(13) Negative	جرح السيف خفيف ويستخبا بس جرح اللسان رخيص وينهي المحبه 😉 The sword’s wound is simple and can be hidden, but the tongue’s wound is cheap, and ends the love 😉
(14) Negative	بت مش ناقصه رعب انا 😡 😡 تبحي خير وسلميلي اللي وراكي😂😂😘😘 Hay girl, I am already scared 😡 😡 Good night and say hi to the one behind you 😂 😂 😘 😘

A. True Sentiment Indication

In Figure 1, the analysis shows the relationship between particular emoji and the sentiment of the text. The table uses the most representative examples of each sentiment class for illustration. It is clear that some emoji are overwhelmingly negative indicators, for instance: 😤, 😢, 😡 and 👎. Others are mostly positive indicators, like ❤️, 😍, 🌹 and 💪. With this kind of emoji, the indicated sentiment is usually explicit and clear, for two reasons:

First, the messages delivered within the text are, themselves, clear and unambiguous. So, these messages do not express irony, sarcasm or other more complex phenomena. Moreover, most of the cases in our dataset where these emoji occur, include sentiment words or phrases, like the words: “love”, “hate”, or the phrases: “I agree with” or “I am against”. We find that Arabic speakers (perhaps, like others) usually use these emoji to directly articulate their feelings of sadness or anger (example 1) or love, cheerfulness, and satisfaction (example 2) in Table IV.

The second reason is that these emoji often co-occur with other emoji from the same sentiment class (i.e., positive with positive and negative with negative). Thus, the combination of these emoji works together to strengthen the sentiment indication (examples 3, and 4) in Table IV.

Note that, in examples 1 and 3, the sentiments of the text-only, the emoji-only, and the text with emoji (i.e., the tweet) are identical, and they all are negative. The same occurs in examples 2 and 4, but with positive sentiment. This means that when all the components of a tweet (i.e., text and each emoji) share the same sentiment class, they will end up reinforcing the effect and so the result will, clearly, belong to that same sentiment class. Therefore, in this condition, emoji can be considered as direct (true) sentiment indicators for a tweet.

B. No-Sentiment Indication

For some of the emoji in our dataset, they do not appear to convey any sentiment indication. This is the case for examples 5 and 6 in Table IV. This may be because, in our examples, the sentiment of the text (i.e., the sentiment of the words) or of the other emoji in the same text dominates.

However, often these emoji are used randomly with some other emoji in a way that is not intended to convey any sentiment. For instance, they may be used as ’decoration’ rather than to serve any real purpose. Example 7 in Table IV, which uses the emoji 💔, is an example.

C. Multi-Sentiment Indication

In Figure 1, there are examples of emoji that we classify as ”Mixed Sentiment”. We considered emoji, like 😂, 😎, 😉, and 😅 as multi-sentiment indicators.

These emoji can be considered as being true sentiment indicators, but with cases with two opposite sentiments, exemplified in Table V. As positive indicators, these emoji have been found playing a significant role in cases similar to example 8 where the 😂 emoji indicates being funny. In example 9, the 😎 emoji indicates being proud, and example 10 where the 😉 emoji indicates being a positive adviser.

In other cases, the same emoji as in examples 8, 9 and 10 are found playing the opposite sentiment role (i.e., a negative sentiment). This can be seen, in Table V, in example 11 where the 😂 emoji indicates being a mocker, example 12 where the 😎 emoji indicates being arrogant, and example 13 where the 😉 emoji indicates embedded threatening advice.

D. Ambiguous Sentiment Indication

Beyond the cases mentioned above, there can also be an ambiguous sentiment indication for a text arising where an emoji exists, not only as a single, stand-alone emoji, but also in combination with emoji with different sentiments. For instance, in example 14 in Table V, human annotators agreed on annotating this tweet with negative sentiment. However, when re-reading the tweet, it could also be interpreted as a positive tweet, depending on context.

This confusion in judging the tweet sentiment is because of the complexity of the sentiment of the text itself. In this example, the sentence “Hey girl, I am already scared” is negative, while the following sentence, “Good night and say hi to the one behind you”, is positive. Besides, the combination of the negative emoji (i.e., 😡), the positive emoji (i.e., 😘), and the multi/mixed-sentiments emoji (i.e., 😂) increase the complexity of deciding the sentiment of the tweet as a whole. Hence, none of the involved emoji can be considered the true/direct sentiment indicator for this tweet.

V. Conclusion, Limitations, and Future Work

In this work, we have undertaken an empirical investigation of the phenomenon of emoji as a sentiment indicator within text. We have applied this in a study of an Arabic, social media corpus using the “Coding and Counting” approach.

Emoji can be a true sentiment indicator, which is the conventional assumption of existing sentiment analysis approaches with emoji. This is the approach used by most of the existing work and implementations of software to perform sentiment analysis of text with embedded emoji. There are many cases in our data where this interpretation is the correct one.

However, some of the most frequently used emoji also occur in many other, unconventional, cases. They may either act as multi-sentiment indicators or as ambiguous sentiment indicators. This is because, according to the context, emoji sometimes are very negative, and sometimes are very positive. Besides, in some cases, our investigation identified examples where the sentiment of an emoji can be neglected within a text. They may be dominated by the sentiment of the text or be dominated by the sentiment of the other emoji in that text. In this case, we considered such emoji as No-sentiment indicators.

It is worth mentioning that the emoji sentiment indications stated above have been found within the dataset that we collected and sampled for this investigation. We are aware that the sentiment behavior of emoji is context-sensitive. This means that in a different context, (for instance, in a different country or in a different social group), the emoji sentiment might reflect the sentiment or usage of that context. Therefore, one of the weaknesses of this work is that, if the same investigative approach was applied on a different dataset, from a different context, then these emoji may be found to behave differently as sentiment indicators.

What is clear, is that the sentiment role of emoji in Arabic social media is complex. Our analysis shows that the conventional approach is sometimes appropriate. However, it also shows that (especially for some of the most frequently used emoji) the conventional approaches are inadequate and that a more sophisticated technique is needed.

Another constraint of this work is the source of the text that was analysed. Whilst Twitter provides a useful source for data, there may be differences between different social media platforms. Furthermore, different classes of conversation (e.g., purely social, political, business and so on), may have an influence upon how emoji are used. Again, further research is required to investigate this.

In conclusion, using emoji solely, as a feature of sentiment indication for text is not a reliable approach, and it might yield arbitrary, noisy, and incorrect sentiment annotation. For that, we need to understand, in detail, the different sentiment states in which emoji can occur, and also the associated sentiment roles that emoji can play within different textual and social contexts.

In the future, our work will expand upon the analysis presented here, develop a model based upon this understanding and then evaluate it, empirically, against human annotated text, and compare the performance of this approach against existing methods. Also, the focus of the work presented here has been on the interpretation of the sentiment effect of emoji in Arabic text. We would expect that similar phenomena would be found in other languages. However, there are likely to be some differences with language and culture. Further work is necessary to confirm whether this is true.

References

[1] A. Kendon, “Gesticulation and speech: Two aspects of the”, *The relationship of verbal and nonverbal communication*, no. 25, p. 207, 1980.

[2] B. Altenberg, *Prosodic patterns in spoken English: Studies in the correlation between prosody and grammar for text-to-speech conversion*. Lund University Press Lund, 1987, vol. 76.

[3] N. Ambady, F. J. Bernieri, and J. A. Richeson, “Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream”, in *Advances in experimental social psychology*, vol. 32, Elsevier, 2000, pp. 201–271.

[4] A. Chen Yuet Wei, “Emoticons and the non-verbal communication: With reference to facebook”, PhD thesis, Christ University, 2012.

[5] A. Mehrabian, *Silent messages*, 152. Wadsworth Belmont, CA, 1971, vol. 8.

[6] S. Kiesler, J. Siegel, and T. W. McGuire, “Social psychological aspects of computer-mediated communication.”, *American psychologist*, vol. 39, no. 10, p. 1123, 1984.

[7] J. Carey, “Paralanguage in computer mediated communication”, in *18th Annual Meeting of the Association for Computational Linguistics*, 1980, pp. 67–69.

[8] P. K. Novak, J. Smailović, B. Sluban, and I. Mozetič, “Sentiment of emojis”, *PloS one*, vol. 10, no. 12, 2015.

[9] L. Rezabek and J. Cochenour, “Visual cues in computer-mediated communication: Supplementing text with emoticons”, *Journal of Visual Literacy*, vol. 18, no. 2, pp. 201–215, 1998.

[10] E. Braumann, O. Preveden, S. Saleem, Y. Xu, and S. T. Koeszegi, “The effect of emoticons in synchronous and asynchronous e-negotiations”, in *Proceedings of the 11th Group Decision & Negotiation Conference (GDN 2010)*, 2010, pp. 113–115.

[11] F. Kunneman, C. Liebrecht, and A. van den Bosch, “The (un) predictability of emotional hashtags in twitter”, 2014.

[12] M. Denis, “Imaging while reading text: A study of individual differences”, *Memory & Cognition*, vol. 10, no. 6, pp. 540–545, 1982.

[13] R. A. Zwaan and M. Singer, “Text comprehension”, in *Handbook of discourse processes*, Routledge, 2003, pp. 89–127.

[14] E. Dresner and S. C. Herring, “Functions of the nonverbal in cmc: Emoticons and illocutionary force”, *Communication theory*, vol. 20, no. 3, pp. 249–268, 2010.

[15] K. Skovholt, A. Grønning, and A. Kankaanranta, “The communicative functions of emoticons in workplace e-mails::-”, *Journal of Computer-Mediated Communication*, vol. 19, no. 4, pp. 780–797, 2014.

[16] V. Evans, *The emoji code: The linguistics behind smiley faces and scaredy cats*. Picador USA, 2017.

[17] H. Miller, D. Kluver, J. Thebault-Spieker, L. Terveen, and B. Hecht, “Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication”, in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[18] J. B. Bavelas and N. Chovil, “Visible acts of meaning: An integrated message model of language in face-to-face dialogue”, *Journal of Language and social Psychology*, vol. 19, no. 2, pp. 163–194, 2000.

[19] B. Gao and D. P. VanderLaan, “Cultural influences on perceptions of emotions depicted in emojis”, *Cyberpsychology, Behavior, and Social Networking*, 2020.

[20] M. A. Riordan, “The communicative role of non-face emojis: Affect and disambiguation”, *Computers in Human Behavior*, vol. 76, pp. 75–86, 2017.

[21] G. Donato and P. Paggio, “Investigating redundancy in emoji use: Study on a twitter based corpus”, in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2017, pp. 118–126.

[22] K. Njenga, “Social media information security threats: Anthropomorphic emoji analysis on social engineering”, in *IT Convergence and Security 2017*, Springer, 2018, pp. 185–192.

[23] M. Shiha and S. Ayvaz, “The effects of emoji in sentiment analysis”, *Int. J. Comput. Electr. Eng.(IJCEE.)*, vol. 9, no. 1, pp. 360–369, 2017.

[24] N. Na’aman, H. Provenza, and O. Montoya, “Varying linguistic purposes of emoji in (twitter) context”, in *Proceedings of ACL 2017, Student Research Workshop*, 2017, pp. 136–141.

[25] D. Rodrigues, D. Lopes, M. Prada, D. Thompson, and M. V. Garrido, “A frown emoji can be worth a thousand words: Perceptions of emoji use in text messages exchanged between romantic partners”, *Telematics and Informatics*, vol. 34, no. 8, pp. 1532–1543, 2017.

[26] M. Rathan, V. R. Hulipalled, K. Venugopal, and L. Patnaik, “Consumer insight mining: Aspect based twitter opinion mining of mobile phone reviews”, *Applied Soft Computing*, vol. 68, pp. 765–773, 2018.

[27] B. Guthier, K. Ho, and A. El Saddik, “Language-independent data set annotation for machine learning-based sentiment analysis”, in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2017, pp. 2105–2110.

[28] H. Abdellaoui and M. Zrigui, “Using tweets and emojis to build tead: An arabic dataset for sentiment analysis”, *Computación y Sistemas*, vol. 22, no. 3, pp. 777–786, 2018.

[29] W. A. Hussien, Y. M. Tashtoush, M. Al-Ayyoub, and M. N. Al-Kabi, “Are emoticons good enough to train emotion classifiers of arabic tweets?”, in *2016 7th International Conference on Computer Science and Information Technology (CSIT)*, IEEE, 2016, pp. 1–6.

[30] J. B. Walther and K. P. D’addario, “The impacts of emoticons on message interpretation in computer-mediated communication”, *Social science computer review*, vol. 19, no. 3, pp. 324–347, 2001.

[31] D. Derks, A. E. Bos, and J. Von Grumbkow, “Emoticons and social interaction on the internet: The importance of social context”, *Computers in human behavior*, vol. 23, no. 1, pp. 842–849, 2007.

[32] M. T. Chi, “Quantifying qualitative analyses of verbal data: A practical guide”, *The journal of the learning sciences*, vol. 6, no. 3, pp. 271–315, 1997.

[33] D. W. Shaffer, *Quantitative ethnography*. Lulu. com, 2017.

[34] J.-W. Strijbos, R. L. Martens, F. J. Prins, and W. M. Jochems, “Content analysis: What are they talking about?”, *Computers & education*, vol. 46, no. 1, pp. 29–48, 2006.

[35] S. C. Herring, S. Barab, R. Kling, and J. Gray, “An approach to researching online behavior”, *Designing for virtual communities in the service of learning*, vol. 338, pp. 338–376, 2004.

[36] M. Salameh, S. Mohammad, and S. Kiritchenko, “Sentiment after translation: A case-study on arabic social media posts”, in *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2015, pp. 767–777.

[37] S. Rosenthal, N. Farra, and P. Nakov, “Semeval-2017 task 4: Sentiment analysis in twitter”, in *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 2017, pp. 502–518.

[38] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, “Semeval-2018 task 1: Affect in tweets”, in *Proceedings of the 12th international workshop on semantic evaluation*, 2018, pp. 1–17.

[39] F. Barbieri *et al.*, “Semeval 2018 task 2: Multilingual emoji prediction”, in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 24–33.

[40] R. Baly, A. Khaddaj, H. Hajj, W. El-Hajj, and K. B. Shaban, “Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets”, *arXiv preprint arXiv:1906.01830*, 2019.

[41] H. Mulki, H. Haddad, C. B. Ali, and H. Alshabani, “L-hsab: A levantine twitter dataset for hate speech and abusive language”, in *Proceedings of the Third Workshop on Abusive Language Online*, 2019, pp. 111–118.

[42] S. N. Alyami and S. O. Olatunji, “Application of support vector machine for arabic sentiment classification using twitter-based dataset”, *Journal of Information & Knowledge Management*, vol. 19, no. 01, p. 2 040 018, 2020.

[43] A. Elmadany, H. Mubarak, and W. Magdy, “Arsas: An arabic speech-act and sentiment corpus of tweets”, *OSACT*, vol. 3, p. 20, 2018.

[44] J. L. Fleiss *et al.*, “The measurement of interrater agreement”, *Statistical methods for rates and proportions*, vol. 2, no. 212-236, pp. 22–23, 1981.

[45] I. A. Farha and W. Magdy, “Mazajak: An online arabic sentiment analyser”, in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 192–198.

Modelling the Role of Social Media in Hotel Selection Using Bayesian Networks

Alexandros Bousdekis
Business Informatics Lab, Department of Business Administration, School of Business
Athens University of Economics and Business
Athens, Greece
e-mail: albous@mail.ntua.gr

Dimitris Kardaras
Business Informatics Lab, Department of Business Administration, School of Business
Athens University of Economics and Business
Athens, Greece
e-mail: dkkardaras@yahoo.co.uk

Stavroula G. Barbounaki
Merchant Marine Academy of Aspropyrgos,
Aspropyrgos, Greece
e-mail: sbarbounaki@yahoo.gr

Abstract— **Consumers increasingly use social media to search for information, compare alternative products and services, and make decisions for activities, such as travel planning and hotel selection. In this context, social media have gathered the research interest as a major form of electronic Word-Of-Mouth (eWOM) to prospective travelers. Existing literature is rich on research works about the influence of travel-oriented online media, such as TripAdvisor, to consumers' decisions with several approaches for sentiment analysis. However, travelers are also widely affected by online comments posted on social media, such as Facebook, Twitter, etc. This paper proposes a methodology for modelling the role of social media in hotel selection using Bayesian Networks (BN). Specifically, it enables identifying the relationships between the way travelers use social media and the criteria for selecting hotels. The proposed approach is demonstrated on a dataset of 360 social media users.**

Keywords-belief network; data mining; e-tourism; tourism management.

I. INTRODUCTION

Consumers increasingly use online media to search for information, compare alternative products and services, and make decisions for activities, such as travel planning and hotel selection [1][2]. Not surprisingly, high ratings in social media have a direct impact on sales [3][4]. Due to the experiential nature of travel-related products and their instantaneous nature, online reviews have become an increasingly popular information source in travel planning and have a profound effect on consumers' buying decisions, particularly in hotel booking [5]. According to Travel Industry Association of America, the evidence shows that 64% of travelers use search engines for their travel planning [6][7].

In this context, social media have gathered the research interest as a major form of electronic Word-Of-Mouth (eWOM) to prospective travelers facilitating the sharing and seeking of experiences [5,8,9,10]. Hotel-related decision-making has fundamentally changed, as social media are used in every stage of the consumers' decision-making process. They play a key role before, during and after the trip [11]. Prospective tourists are influenced by social media, as content from other travellers can shape, guide and redirect their initial decisions [12][13].

In the business perspective, social media are perceived as effective tools and fruitful platforms for deepening customer engagement and enhancing customer–business interactions [14]. In fact, they have provided a new distribution channel for businesses to communicate with their customers [7]. In the consumer perspective, consumers use social media for a wide spectrum of scenarios, e.g., sharing their travel-related experiences, engaging with others, connecting with people from different destinations and buying travel-related products and services [15][16].

Existing literature is rich on research works about the influence of travel-oriented online media, such as TripAdvisor, to consumers' decisions. However, travelers are also widely affected by online comments posted on social media, such as Facebook, Twitter, etc. as well as by hotels' marketing campaigns [17]. Therefore, the identification of the relationships between the way travelers use social media and the criteria for selecting hotels is of outmost importance. This paper proposes a methodology for modelling the role of social media in hotel selection using Bayesian Networks (BN). To the best of our knowledge, despite their applicability in a wide range of problems and scenarios, BNs have not been used for identifying the influence of social media to the decisions of travelers about the hotel selection.

The rest of the paper is organized as follows: Section II presents the related work on methods and approaches for evaluating the effect of online reviews on social media on hotel booking. Section III describes the research methodology and the proposed approach for modelling the role of social media in hotel selection using BNs. Section IV presents the results from the adoption of the proposed methodology on a dataset of 360 users. Section V concludes the paper and outlines our plans for future work.

II. RELATED WORK

Online comment has become a popular and efficient way for sellers to acquire feedback from customers and improve their service quality [18]. These online reviews generate an eWOM effect, which influences future customer demand and hotels' financial performance [19]. However, apart from the hotels' websites and official social media pages, prospective travelers are increasingly interacting through social media in order to gather and share information about hotels and to select the one that matches their criteria. To this end, a vast amount of research has focused on travel-oriented platforms and social media, such as TripAdvisor, aiming at investigating their influence to hotel booking decisions [7][11][14][20][21]. Moreover, such works are conducted from a tourism management perspective resulting in the use of descriptive statistical methods instead of exploiting the advancements of data analytics and machine learning. On the other hand, the role of social media such as Facebook and Twitter on hotel selection is rarely investigated [8].

In [22], the authors examined the effects of traditional customer satisfaction relative magnitude and social media review ratings on hotel performance and explored which online travel intermediaries' review ratings serve as the most reliable and valid predictor for hotel performance. The results of this study indicate that social media review rating is a more significant predictor than traditional customer satisfaction for explaining hotel performance metrics. The research work in [23] assessed social media content produced by customers and related review-management strategies of domestic and international hotel chains with the use of descriptive statistics and multilevel regression.

In [11], the authors proposed the use of multi-criteria ratings provided by the travelers in social media networking sites for developing a new recommender system for hotel recommendations in e-tourism platforms. Reference [3] applied multilevel regression analysis in order to quantify the extent to which differences in client satisfaction with hotels can be attributed to the destination in which the hotels are located. They measured this through ratings provided through social media outlets. In [24], the authors also investigated the influence of social media on destination choice. In [5], the presented work is based upon homophily and similarity-attraction theory in order to prove that review valence significantly affects hotel booking intention, and that reader-reviewer demographic similarity moderates this effect. This three-way interaction reveals a substituting moderation effect between demographic similarity and preference similarity.

In [12], the authors explored how social media influence the way consumers search, evaluate and select a hotel within the 'evaluation stage' of the wider hotel decision-making process, i.e., in the pre-travel stage during which social media unfold their most critical role. In [6], the authors examined tourists' knowledge sharing behavior in social media for two different types of social media: Facebook and TripAdvisor. They proposed a structural model that connects homophily and knowledge sharing through posting. Finally, the research work in [13] investigated the influencing role of social media in the consumer's hotel decision-making process and identified the advantages and disadvantages. They concluded that the advantages of utilizing social media in hotel selection outperform the disadvantages.

III. RESEARCH METHODOLOGY

A. *Data Collection and Structuring*

The data was collected in the form of a questionnaire completed by 360 social media users. The questions lay on three categories: generic questions, questions related to the reasons of searching information on social media, and questions related to the criteria according to which the users select a hotel for vacation. The first category of questions was in the form of multiple choice, while the last two were in the form of Likert scale.

B. *Modelling the Relationships between Social Media and Hotel Selection Criteria Using Bayesian Networks*

In order to model the relationships between the reasons of searching information on social media and the criteria according to which the users select a hotel for vacation, we applied BNs. A Bayesian Network (BN) [25], also known as belief network, is defined as a pair $B = (G, \Theta)$. $G = (V, E)$ is a Directed Acyclic Graph (DAG) where $V = \{v_1, \ldots, v_n\}$ is a collection of n nodes, $E \subset V \times V$ a collection of edges and a set of parameters Θ containing all the Conditional Probabilities (CP) of the network.

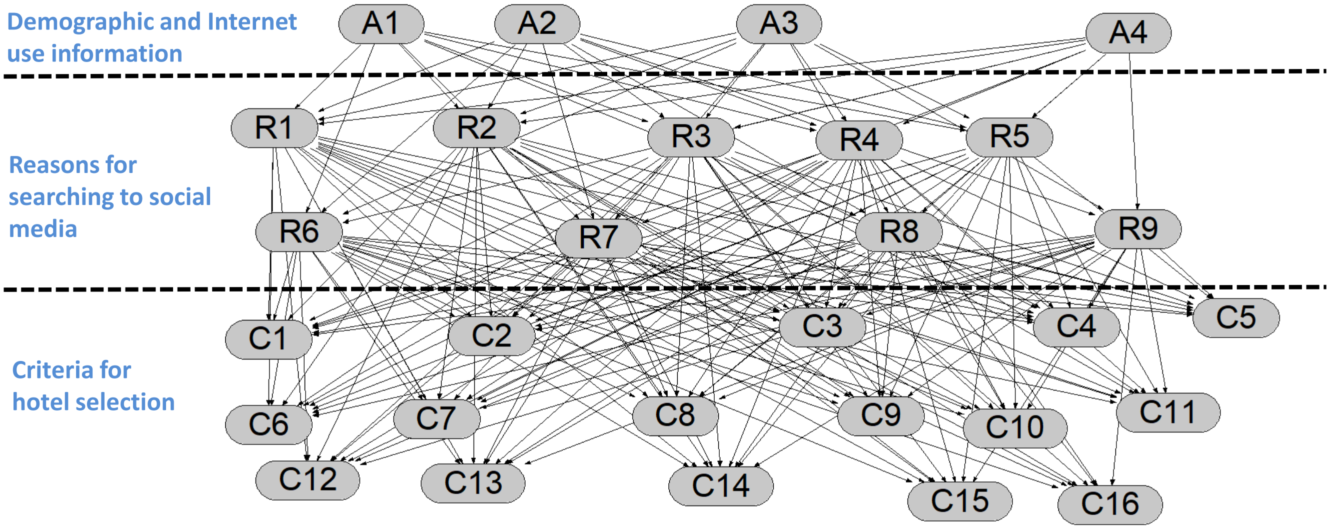

Demographic and Internet use information
A1
A2
A3
A4
Reasons for searching to social media
R1
R2
R3
R4
R5
R6
R7
R8
R9
Criteria for hotel selection
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
C13
C14
C15
C16

Figure 1. The Bayesian Network structure for modelling the role of social media in hotel selection.

Each node $v \in V$ of the graph represents a random variable X_V with a state space $\boldsymbol{X_V}$ which can be either discrete or continuous. An edge $(v_i, v_j) \in E$ represents the conditional dependence between two nodes v_i, $v_j \in V$ where v_i is the parent of child v_j. If two nodes are not connected by an edge, they are conditional independent. Because a node can have more than one parent, let π_v the set of parents for a node $v \in V$.

Therefore each random variable is independent of all nodes $V \setminus \pi_v$. For each node, a Conditional Probability Table (CPT) contains the CP distribution with parameters $\theta_{xi|\pi i} := P(x_i|\pi_i) \in \Theta$ for each realization x_i of X_i conditioned on π_i. The joint probability distribution over V is visualized by the BN and can be defined as

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i|\pi_i) \qquad (1)$$

With BN, inference for what-if analysis can be supported, either top-down (predictive support) or bottom-up (diagnostic support). If a random variable which is represented by a node is observed, the node is called an evidence node; otherwise, it is a hidden node [26]. Based on the categories of the questions included in the questionnaire, a BN with three layers was developed, as shown in Figure 1. The nodes per each layer of the BN are presented in Table I.

The top layer of the BN includes 4 nodes related to generic information (A1-A4). These nodes along with their alternative values are: the respondent's age group = {*15-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, >50*), the frequency of vacations = {*once per 2 years, once per year, twice per year, three times per year, more than three times per year*}, the frequency of staying at hotel in vacations = {*always, very often, often, rarely, never*}, and the frequency of using social media for hotel information = {*always, very often, often, rarely, never*}.

The intermediate layer includes nodes related the reasons of searching information on social media in general and consists of 9 nodes (R1-R9). In other words, it indicates the behaviour and the attitude of the users with respect to the use of social media.

The bottom layer includes nodes related to the criteria according to which the users select a hotel for vacation and consists of 14 nodes (C1-C16). Their candidate values are {*Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree*}.

Based upon this structure, the BN is subject to reasoning in order to compute all the CPTs. The BN was constructed in a way that all the nodes of the intermediate and the bottom layer are potentially affected by all the nodes of the top and the intermediate layer respectively. Therefore, the CPTs are calculated accordingly.

The outcome of the proposed methodology indicates the probability assigned to each selection criterion (bottom layer) given the reasons a user searches for information in social media (intermediate layer) and some generic information (top layer).

TABLE I. BAYESIAN NETWORK NODES PER LAYER

Layers	Nodes		Node Values
Top Layer ***(General Information)***	**A1**	Age group	{15-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, >50}
	A2	Frequency of vacations	{once per 2 years, once per year, twice per year, three times per year, more than three times per year}
	A3	Frequency of staying at hotel in vacations	{always, very often, often, rarely, never}
	A4	Frequency of using social media for hotel information	{always, very often, often, rarely, never}
Intermediate Layer ***(Reasons of searching to social media)***	**R1**	Trust the social media users	{Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree}
	R2	Possibility of asking opinions	
	R3	Search engines are not helpful	
	R4	Socializing	
	R5	Quick responses	
	R6	Easy procedure	
	R7	Better quality of responses	
	R8	Costless	
	R9	Funny	
Bottom Layer ***(Criteria for hotel selection)***	**C1**	Personnel	{Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree}
	C2	Reliable booking procedure	
	C3	Fast check-in / check-out	
	C4	Immediate service and problem solving	
	C5	Hotel security and privacy assurance	
	C6	Cleanliness	
	C7	Reasonable price	
	C8	Convenient parking	
	C9	Comfortable bed	
	C10	Comfortable public spaces	
	C11	Interior design	
	C12	Location	
	C13	External environment	
	C14	Quality of hotel restaurant	
	C15	Availability of mini bar in the rooms	
	C16	Belonging to a reputable hotel chain	

Therefore, the model can answer questions such as: "What is the probability that a user will select a hotel according to the criteria of the reliable booking procedure

(C2) and the cleanliness (C6) given that he/she uses the social media for socializing (R4) (referring to node values *"Strongly Agree"* and *"Agree"*) and for receiving better quality of responses (R7), while he/she belongs to the age group *31-35* (A1), he/she goes for vacations *once per year* (A2), he/she stays at a hotel *often* (A3) and he/she *often* uses social media for hotel information (A4)?". In order to answer such questions, the model computes all the CPTs for all its nodes and for all their alternative values.

The model is able to identify, represent and store in the database complex relationships aiming at supporting marketing and hotel operations in response to different customers' profiles. Upon request, the model can compute the CPTs of every possible relationship based upon the resulting CPT in order to provide insights on the hotel selection criteria. In this way, the hotels can focus on specific target groups according to their strengths as well as to improve their operations that result in lower rating of certain criteria. Moreover, it is able to serve as a model for predicting the criteria according to which a social media user will select a hotel among various alternatives. The model is extensible to additional nodes per each layer in case more information needs to be incorporated.

IV. Results

The proposed approach was applied on a dataset of 360 social media users. The implementation and execution of the experiments were performed using the BN functionalities of the pgmpy (Probabilistic Graphical Models using Python) package in Python [27]. We developed the associated BN and we calculated the CPTs for all the nodes.

Table II presents the criteria (C_i) and their associated values with the highest CPs, given the values of the reasons of searching information in social media (R_i) and the generic information (A_i). Table III presents the criteria (C_i) and their associated values with the lowest CPs, given the values of the reasons of searching information in social media (R_i) and the generic information (A_i).

For this specific analysis, we have grouped the values *Strongly Agree* and *Agree* in order to identify the most probable criteria in the first columns of the aforementioned Tables. The results show that the criterion C6 given the values of the R_i and A_i nodes that are shown in the first row of Table II is the one with the highest CP, which is equal to 39.5%. The criterion C15 given the values of the R_i and A_i nodes that are shown in the first row of Table III is the one with the lowest CP, which is equal to 1.2%.

Based upon these results, the hotels are able to identify the most important criteria according to which a social media user selects a hotel given some generic information, such as the age group, the frequency of vacations, etc., and their attitude towards the use of social media for searching information. In this way, the hotels may design more specialized marketing strategies, e.g., focusing on specific target groups, and to improve their operations in order to achieve higher service quality and increased customer satisfaction with respect to certain criteria.

TABLE II. Criteria C_i with the Highest CPs Given R_i and A_i

Criteria (Child Nodes)	**Parent Nodes**		***CP***
C_i	***R_i***	***A_i***	

C6	R1={Neutral},
	R2={Agree},
	R3={Disagree},
	R4={Agree},
	R5={Strongly Agree},
	R6={Neutral},
	R7={Disagree},
	R8={Neutral},
	R9={Strongly Disagree}
		A1={36-40},
		A2={once per year},
		A3={very often},
		A4={often}
			0.395

C1	R1={Disagree},
	R2={Agree},
	R3={Strongly Disagree},
	R4={Strongly Disagree},
	R5={Agree},
	R6={Neutral},
	R7={Neutral},
	R8={Agree},
	R9={Strongly Disagree}
		A1={46-50},
		A2={twice per year},
		A3={very often},
		A4={rarely}
			0.362

C7	R1={Agree},
	R2={Strongly Agree},
	R3={Disagree},
	R4={Agree},
	R5={Agree},
	R6={Strongly Agree},
	R7={Neutral},
	R8={Neutral},
	R9={Agree}
		A1={31-35},
		A2={once per 2 years},
		A3={often},
		A4={very often}
			0.294

C12	R1={Strongly Agree},
	R2={Strongly Agree},
	R3={Neutral},
	R4={Strongly Agree},
	R5={Neutral},
	R6={Neutral},
	R7={Neutral},
	R8={Neutral},
	R9={Agree}
		A1={26-30},
		A2={once per year},
		A3={rarely},
		A4={always}
			0.285

C11	R1={Neutral},
	R2={Agree},
	R3={Disagree},
	R4={Neutral},
	R5={Agree},
	R6={Agree},
	R7={Disagree},
	R8={Agree},
	R9={Neutral}
		A1={41-45},
		A2={twice per year},
		A3={very often},
		A4={often}
			0.239

C2	R1={Neutral},
	R2={Strongly Agree},
	R3={Neutral},
	R4={Agree},
	R5={Strongly Agree},
	R6={Neutral},
	R7={Disagree},
	R8={Neutral},
	R9={Neutral}
		A1={36-40},
		A2={once per year},
		A3={very often},
		A4={often}
			0.217

C4	R1={Disagree},
	R2={Agree},
	R3={Neutral},
	R4={Agree},
	R5={Strongly Agree},
	R6={Agree},
	R7={Neutral},
	R8={Agree},
	R9={Neutral}
		A1={36-40},
		A2={twice per year},
		A3={very often},
		A4={very often}
			0.208

TABLE III. Criteria C_i with the Lowest CPs Given R_i and A_i

Criteria (Child Nodes)	Parent Nodes		*CP*
C_i	R_i	A_i	

C15	R1={Strongly Agree},
	R2={Neutral},
	R3={neutral},
	R4={Strongly Disagree},
	R5={Agree},
	R6={Neutral},
	R7={Neutral},
	R8={Disagree},
	R9={Agree}
		A1={21-25},
		A2={three times per year},
		A3={never},
		A4={very often}
			0.012

C14	R1={Strongly Agree},
	R2={Strongly Disagree},
	R3={Neutral},
	R4={Agree},
	R5={Neutral},
	R6={Strongly Disagree},
	R7={Disagree},
	R8={Neutral},
	R9={Strongly Disagree}
		A1={26-30},
		A2={once per year},
		A3={rarely},
		A4={always}
			0.023

C3	R1={Strongly Agree},
	R2={Strongly Agree},
	R3={Neutral},
	R4={Agree},
	R5={Agree},
	R6={Agree},
	R7={Disagree},
	R8={Strongly Agree},
	R9={Neutral}
		A1={21-25},
		A2={once per 2 years},
		A3={rarely},
		A4={rarely}
			0.025

C10	R1={Agree},
	R2={Agree},
	R3={Strongly Disagree},
	R4={Neutral},
	R5={Neutral},
	R6={Strongly Agree},
	R7={Disagree},
	R8={Agree},
	R9={Agree}
		A1={31-35},
		A2={once per 2 years},
		A3={rarely},
		A4={often}
			0.031

C16	R1={Strongly Agree},
	R2={Agree},
	R3={Neutral},
	R4={Strongly Agree},
	R5={Agree},
	R6={Agree},
	R7={Neutral},
	R8={Agree},
	R9={Strongly Agree}
		A1={21-25},
		A2={once per 2 years},
		A3={rarely},
		A4={very often}
			0.044

C8	R1={Disagree},
	R2={Neutral},
	R3={Disagree},
	R4={Neutral},
	R5={Agree},
	R6={Strongly Agree},
	R7={Strongly Agree},
	R8={Agree},
	R9={Neutral}
		A1={21-25},
		A2={once per year},
		A3={very often},
		A4={often}
			0.046

C5	R1={Agree},
	R2={Neutral},
	R3={Strongly Disagree},
	R4={Neutral},
	R5={Strongly Agree},
	R6={Agree},
	R7={Strongly Agree},
	R8={Agree},
	R9={Neutral}
		A1={36-40},
		A2={three times per year},
		A3={rarely},
		A4={always}
			0.052

TABLE IV. Confusion Matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP) = 41	False Negative (FN) = 9
Actual Negative	False Positive (FP) = 3	True Negative (TN) = 32

As already mentioned, the model can also serve as a classifier for predicting the class attribute of criteria (C_i) as soon as new records of R_i and A_i are inserted into the database. In order to evaluate its classification effectiveness, we inserted additional records, derived from more questionnaires addressed to social media users, and we created the confusion matrix according to Table IV in order to estimate the precision and the recall of the classifier using the (2) and (3) [28].

$$Precision = \frac{TP}{TP+FP} = \frac{41}{41+3} = 93.1\% \qquad (2)$$

$$Recall = \frac{TP}{TP+FN} = \frac{41}{41+9} = 82\% \qquad (3)$$

The Precision results are quite satisfactory, while the Recall results can be further improved. The BN model sticks to the initially identified relationships, i.e., the ones that have been mined during the model training. Therefore, when new relationships, not previously identified, are added, they are not classified correctly. These records include values that are not frequent (e.g., A1={>50} and A4={always}), so they are not critical for decision making.

V. Conclusions and Future Work

Consumers increasingly use social media to search for information, compare alternative products and services, and make decisions for activities such as travel planning and hotel selection. In this context, social media have gathered the research interest as a major form of eWOM to prospective travelers. In this paper, we proposed a BN model for modelling the role of social media in hotel selection. More specifically, we developed a 3-layered BN corresponding to generic information, reasons for searching information to social media, and criteria for hotel selection respectively. In this way, the model is able to mine relationships and to compute the CPTs in order to reveal meaningful insights and predictions about the criteria of hotel selection given the use of social media and other information.

The BN model was applied to a dataset of 360 social media users, derived from an associated questionnaire. According to the defined BN structure, all the CPTs were computed. We presented indicative examples of the outcome, i.e., the criteria with the highest and the lowest CPs. We also validated the model in terms of its precision and recall in predicting the most important hotel selection criteria when new records are inserted into the database.

Regarding our future work, we plan to use more data analytics and machine learning methods and algorithms in order to mine hidden relationships among various attributes.

Moreover, we aim to use fuzzy pattern matching methods for mining also online review comments, as well as clustering and fuzzy sets qualitative analytics algorithms for extracting user profiling insights of hotel customers. These directions have the potential to further enhance decision making process in hotel management from both a marketing (e.g., revealing key groups of customers and target groups) and an operations management (e.g., for improving service quality if it receives negative review rating) perspective.

REFERENCES

[1] B. A. Sparks, K. K. F. So, and G. L. Bradley, “Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern,” Tour. Man., vol. 53, pp. 74-85, 2016.

[2] T. Radojevic, N. Stanisic, and N. Stanic, “Inside the rating scores: a multilevel analysis of the factors influencing customer satisfaction in the hotel industry,” Cornell Hosp. Quart., vol. 58, no. 2, pp. 134-164, 2017.

[3] J. Bulchand-Gidumal, S. Melián-González, and B. G. Lopez-Valcarcel, “A social media analysis of the contribution of destinations to client satisfaction with hotels,” Int. J. of Hosp. Man., vol. 35, pp. 44-47, 2013.

[4] W. G. Kim and S. A. Park, “Social media review rating versus traditional customer satisfaction,” Int. J. of Cont. Hosp. Man., vol. 29, no. 2, pp. 784-802, 2017.

[5] I. C. C. Chan, L. W. Lam, C. W. Chow, L. H. N. Fong, and R. Law, “The effect of online reviews on hotel booking intention: The role of reader-reviewer similarity,” Int. J. of Hosp. Man., vol. 66, pp. 54-65, 2017.

[6] S. Okazaki, L. Andreu, and S. Campo, “Knowledge sharing among tourists via social media: A comparison between Facebook and TripAdvisor,” Int. J. of Tour. Res., vol. 19, no. 1, pp. 107-119, 2017.

[7] M. Nilashi, O. Ibrahim, E. Yadegaridehkordi, S. Samad, E. Akbari, and A. Alizadeh, “Travelers decision making using online review in social network sites: A case on TripAdvisor,” J. of Comp. Sc., vol. 28, pp. 168-179, 2018.

[8] W. Duan, Y. Yu, Q. Cao, and S. Levy, “Exploring the impact of social media on hotel service performance: A sentimental analysis approach,” Cornell Hosp. Quart., vol. 57, no. 3, pp. 282-296, 2016.

[9] P. De Pelsmacker, S. Van Tilburg, and C. Holthof, “Digital marketing strategies, online reviews and hotel performance,” Int. J. of Hosp. Man., vol. 72, pp. 47-55, 2018.

[10] Q. Ye, R. Law, B. Gu, and W. Chen, “The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings,” Comp. in Hum. Behav., vol. 27, no. 2, pp. 634-639, 2011.

[11] M. Nilashi, E. Yadegaridehkordi, O. Ibrahim, S. Samad, A. Ahani, and L. Sanzogni, “Analysis of Travellers’ Online Reviews in Social Networking Sites Using Fuzzy Logic Approach,” Int. J. of Fuz. Sys., vol. 21, no. 5, pp. 1367-1378, 2019.

[12] E. Varkaris and B. Neuhofer, “The influence of social media on the consumers’ hotel decision journey,” J. of Hosp. and Tour. Tech., vol. 8, no. 1, pp. 101-118, 2017.

[13] V. Gupta, “The influencing role of social media in the consumer’s hotel decision-making process,” Worl. Hosp. and Tour. Them., vol. 11, no. 4, pp. 378-391, 2019.

[14] X. Y. Leung, B. Bai, and M. Erdem, “Hotel social media marketing: a study on message strategy and its effectiveness,” J. of Hosp. and Tour. Tech., vol. 8, no. 2, pp. 239-255, 2017.

[15] A. M. Munar and J. K. S. Jacobsen, “Motivations for sharing tourism experiences through social media,” Tour. Man., vol. 43, pp. 46-54, 2014.

[16] B. Zeng and R. Gerritsen, “What do we know about social media in tourism? A review,” Tour. Man. Persp., vol. 10, pp. 27-36, 2014.

[17] Z. Xiang, Q. Du, Y. Ma, and W. Fan, “A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism,” Tour. Man., vol. 58, pp. 51-65, 2017.

[18] X. Wei, X. Luo, X., Q. Li, J. Zhang, and Z. Xu, “Online comment-based hotel quality automatic assessment using improved fuzzy comprehensive evaluation and fuzzy cognitive map,” IEEE Trans. on Fuzzy Sys., vol. 23, no. 1, pp. 72-84, 2015.

[19] K. L. Xie, C. Chen, and S. Wu, “Online consumer review factors affecting offline hotel popularity: evidence from tripadvisor,” J. of Trav. & Tour. Mark., vol. 33, no. 2, pp. 211-223, 2016.

[20] S. Molinillo, J. L. Ximénez-de-Sandoval, A. Fernández-Morales, and A. Coca-Stefaniak, “Hotel assessment through social media: The case of TripAdvisor,” Tour. & Man. Stud., vol. 12, no. 1, pp. 15-24, 2016.

[21] M. L. Yadav and B. Roychoudhury, “Effect of trip mode on opinion about hotel aspects: A social media analysis approach,” Int. J. of Hosp. Man., vol. 80, pp. 155-165, 2019.

[22] W. G. Kim and S. A. Park, “Social media review rating versus traditional customer satisfaction,” Int. J. of Cont. Hosp. Man., vol. 29, no. 2, pp. 784-802, 2017.

[23] M. Schuckert, S. Liang, R. Law, and W. Sun, “How do domestic and international high-end hotel brands receive and manage customer feedback?,” Int. J. of Hosp. Man., vol. 77, pp. 528-537, 2019.

[24] A. Tham, G. Croy, and J. Mair, “Social media in destination choice: Distinctive electronic word-of-mouth dimensions,” J. of Trav. & Tour. Mark., vol. 30, pp. 144-155, 2013.

[25] J. Pearl, “Probabilistic reasoning in intelligent systems: networks of plausible inference,” Elsevier, 2014.

[26] T. D. Nielsen, and F. V. Jensen, “Bayesian networks and decision graphs,” Springer Science & Business Media, 2009.

[27] A. Ankan, and A. Panda, “pgmpy: Probabilistic graphical models using python,” in Proceedings of the 14th Python in Science Conference (SCIPY 2015), Citeseer, vol. 10, 2015.

[28] C. Goutte and E. Gaussier, “A probabilistic interpretation of precision, recall and F-score, with implication for evaluation,” in European conference on information retrieval, pp. 345-359), Springer, Berlin, Heidelberg.