Intelligent Parallel Mixed Method Approach for Characterising Viral YouTube Videos in Saudi Arabia

Abdullah Alshanqiti[1]
Faculty of Computer and Information Systems
Islamic University (IU)
Madinah, Saudi Arabia
ORCID:0000-0002-6080-5236

Ayman Bajnaid[2]
Faculty of Media and Communication
King Abdulaziz University (KAU)
Jeddah, Saudi Arabia

Abdul Rehman Gilal[3]
Department of Computer Science
Sukkur IBA University
Sukkur, Pakistan

Shuaa Aljasir[4]
Faculty of Media and Communication
Kin6g Abdulaziz University (KAU)
Jeddah, Saudi Arabia

Aeshah Alsughayyir[5]
College of Computer Science
and Engineering
Taibah University
Madinah, Saudi Arabia

Sami Albouq[6]
Faculty of Computer
and Information Systems
Islamic University (IU)
Madinah, Saudi Arabia

***Abstract*—In social networking platforms, comprehending virality, exemplified by YouTube, is of great importance, which helps in understanding what characteristics utilised to create content along with what dynamics involved in contributing to YouTube's strength as a platform for sharing content. The current literature surrounding virality problem appears sparse concerning development theories, investigations regarding empirical facts, and an understanding of what makes videos go viral. The overarching objective is to understand deeply the phenomena of viral YouTube videos in Saudi Arabia, hence we propose an intelligent convergent parallel mixed-methods approach that begins, as an internal step, by a qualitative thematic analyses method and an NLP-based quantitative method independently, followed by training an unsupervised clustering model for integrating the internal analysis outputs for deeper insights. We have empirically analysed some trended YouTube videos along with their contents, for studying such phenomena. One of our main findings revealed that boosting entertainments, traditions, politics, and/or religion issues when making a video, that is associated in somehow with sarcastic or rude remarks, is likely the preeminent impulse for letting a regular video go viral.**

***Keywords*—Virality; text mining; sentiment analysis; social media analysis; mixed method approach**

I. Introduction

Knowing how digital content can get rapidly spread worldwide, such as viral video, is of great importance in perfecting our e-services. In the scope of social networking platforms, virality can be loosely defined as the ability of content to spread rapidly in society from one person to another. Given the present time's propensity for communication via electronics, content spreads like wildfire thanks to the Internet. This virality is exemplified by YouTube, whose user-generated content allows users to freely create and share content both on its own platform and on other social media platforms [1].

Given YouTube's success, there exists an interest in understanding what characteristics utilised to create content along with what dynamics involved in contributing to YouTube's strength as a platform for sharing content. Although scholars agree with the characteristics that constitute/make up viral content, there exists less certainty with what makes a video extremely popular [2]. Despite a growing interest in this field, the current literature surrounding this topic also remains sparse with respect to development theories, investigations regarding empirical facts, and an understanding of what makes videos go viral.

Understanding why and how a video becomes extremely popular (i.e., how it goes viral) can maximise how consumers can benefit from a video's popularity along with how users can deal with the threats associated with virality such as spreading rumors or violating others' privacy. Analysing a large amount of data from YouTube's video collection would also allow for a deeper understanding of social behavior, dynamics, and processes at play when people consume and create content.

Broadly speaking, there exists two principal conceptual analysis when it comes to research on virality, formulated coherently in a valuable theoretical framework by [3]: a top-down mechanism which considers virality as the result of highly influential individuals who can use their power in promoting their videos by (e.g., existing mainstream media); a bottom-up mechanism, which argues that virality relies instead on the characteristics of the content that factually engage individuals to spread the content in a self-motivated way [4]. Interestingly, [5] (citied in [6]) mention that the latter mechanism is more often prompting virality.

In a general sense, this research attempts to contribute to the bottom-up mechanism by solely focusing on Arabic videos, particularly, videos that have gone viral. The overarching objective behind our attempt is to provide an intelligent based solution to help in understanding deeply the phenomena of viral YouTube videos in Saudi Arabia, which can be used in future research as a guideline or for comparison purposes. Thus, we propose a convergent parallel mixed-methods approach that begins, as an internal step, by a qualitative thematic analyses method and an NLP-based quantitative method independently, followed by training an unsupervised clustering model for integrating the internal analysis outputs for deeper insights. To be more precise, the proposed complex approach depends on (1) our optimised lexicon-based *Bag of Words* sentiment classifier for analysing viewer's shared

comments left on YouTube, and on (2) a manual inspiration method for qualitatively analysing video content. We report on experiments to understand the virality problem by examining several trended YouTube videos in Saudi Arabia. Summing up, the contributions of this research are:

- A qualitative study on a variety of video's categories and themes propagated in Saudi Arabia.
- A lexicon-based *Bag of Words* sentiment classifier, where the novelty here lies on our optimised algorithms, implemented in Java, that support any texts written in Arabic without translation.
- An innovative idea of utilising unsupervised machine learning technique, depending on distance matrix and hierarchical clustering, for integrating our internal findings. This thought could be a promising research paradigm that fundamentally contributes to social media intelligence approaches.

The next section seeks to investigate prior scholarship on phenomena that have gone viral, examine gaps in the literature regarding the virality process, and present noteworthy questions of the current research. The following section introduces our methodology utilised in the examination and presents the subsequent analysis and results. Lastly, the final section discusses the conclusions of this research and outline our intention for future research to take.

II. Review of Related Literature

This section provides an overview of the phenomenon of viral content by drawing on scholars who have sought to understand the processes and dynamics of virality. In 1997, the firm Draper Fisher Jurvetson coined the phrase "viral marketing" to describe Hotmail's use of advertisements to promote the fact that its emailing service was free [7]. [8] then noted that viral marketing was described as a type of marketing that infects its customers with an advertising message that passes from one customer to the next like a rampant flu virus (p. 93). More generally, "viral marketing" and "viral content" have since become catchphrases for online advertising success. A variety of other definitions have also been offered for virality, each coupled with a specific approach in examining its nature.

According to [9], examinations and definitions of virality can be categorised in three ways. The first seeks to examine how the content is accessed, disseminated, and propagated over a short time period. The second seeks to examine how virality is spread via electronic sharing (i.e., word of mouth) by focusing on the content shared. Lastly, the third concentrates on users' behaviors and engagement with the viral content in question and gauges their likes/dislikes, shares, and comments. [10] argued that the term "virality" includes a host of aspects and exchanges such as the number of people who have access to the content, the appreciation of the content, and how many people have liked or shared the content. The popularity of the content depends exclusively on those who share it and the reactions it garners (positive, negative, and, to a lesser extent, neutral). The current research defines viral content as that which spreads to the greatest degree possible over the shortest amount of time.

YouTube has been chosen as the topic of study for the present research due to the double-sided nature of its platform (i.e., the ability to share and participate through comments as well as to react to content via word of mouth). Sharing content on YouTube requires interacting with others online, which in turn affects the popularity of said content. Content spread online generates greater audience numbers than content spread through some other means. YouTube also affords the distinct opportunity to study both the activities of YouTube users' interactions and their social network ties. According to [10], a number of elements play a part in this sharing process, including the nature of the shared content, the nature of the user who shares it, the nature of the audience who receives it, and the structure of the network through which the content is spreading. The present research aimed to provide an AI-based solution to help in understanding the phenomena of viral YouTube videos.

Previous research on virality has primarily been drawn from five different fields: psychology (e.g., [9]; [11]; [12]; [13]; [4]), computer science (e.g., [14]; [15]; [16]; [1]; [10]; [17]; [18]; [19]; [20]; [21]; [22]), political science (e.g., [2]; [23]; [24]), marketing (e.g., [25]; [26]; [27]; [28]; [29]; [30]; [31]; [7]; [32]; [33]), and health (e.g., [34]). These studies have been mainly conducted in Western countries, such as the United States, Canada, Germany, Italy, and Australia. However, there have been a few studies conducted in less developed countries, such as [2] study in South Africa; [29] and [22] studies in China; [1] in South Korea; [18] in Brazil; [4] in Romania; and [21] in India. However, no studies have been conducted in Arab countries or even in the Middle East.

These studies used several methods to collect data. While some of them utilised questionnaires to obtain users responses, others used data-mining tools. A few studies manually conducted content analysis. Most of the previous researchers developed their own models to explain the phenomenon of virality. Only two studies have borrowed theories from other fields to explain virality; these theories included uses and gratifications theory, the persuasion model, and the memory-based model ([9] and [32]). Thus, the current study aims to fill the gap in the field by proposing a convergent parallel mixed-methods approach that begins, as an internal step, with a qualitative thematic analysis method and an NLP-based quantitative method (used independently), followed by training an unsupervised clustering model for integrating the internal analysis outputs for deeper insights in order to provide a deep understanding of the phenomena of viral YouTube videos in Saudi Arabia.

III. Research Methodology

The original work carried out in this paper was to better understand the rapid spread of viral YouTube videos in Saudi Arabia. We have considered a variety of video's categories and qualitative themes as input factors for our experiment that conducted on a dataset collected from the top 13 viral videos, trended between 2016 and 2017 as reported in *Think-with-Google*[1]

Through the stages of this study, we have investigated the importance of sentiment analysis and mining opinions from YouTube comments, which allows us to classify the viewer's

[1]Think-with-**Google**-https://www.thinkwithgoogle.com/intl/en-145/perspectives/local-articles/youtube-and-search-online-trends-mena-2016/

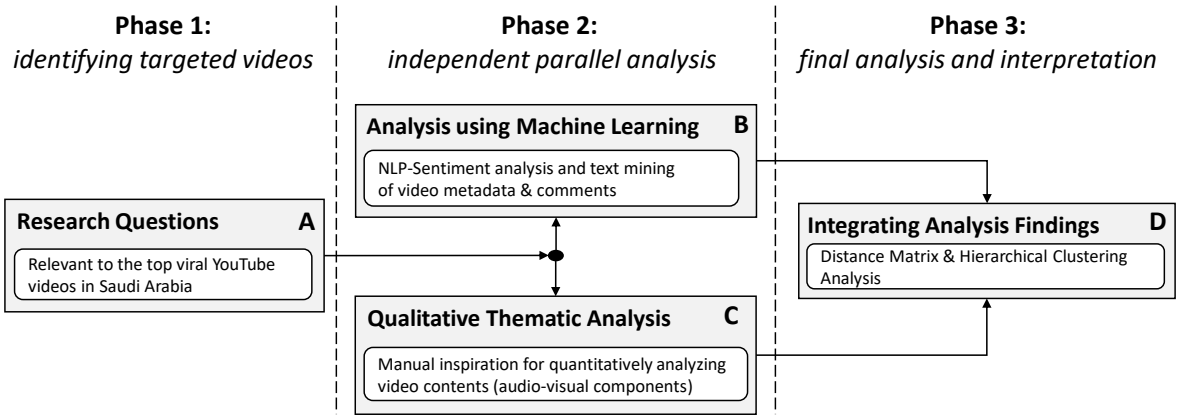

Fig. 1. Our proposed convergent parallel mixed-methods approach

concerns and behaviors in conjunction with video's themes. By considering timestamps as an additional dimension, we attempt to anticipate the future shift of community concerns in social circles.

The convergent parallel mixed-methods approach, presented in this paper, integrates a qualitative thematic analyses method for analysing video content view with a quantitative method of NLP-mining opinion for investigating viewer's textual comments. The outcomes from these independent methods (i.e., qualitative and quantitative methods) are then integrated and fed into our an unsupervised machine learning (i.e., Hierarchical Cluster Analysis) model for comprehensive understanding and more accurate predictions. The overall flow proposed approach is described in a three phases, illustrated in Fig. 1. In the rest of this section, we first discuss our data collection methodology, including the selection criteria of YouTube videos for experiments as shown under *Phase 1*. We next introduce our analysis methods for both viewer's comments and video's contents for answering our research questions, see *Phase 2* and *Phase 3* of Fig. 1.

A. Acquiring Data for Experiments

YouTube is the second-most popular video-sharing website in the world, according to Alexa website[2]. It provides an official API[3] Services to access and fetch specific data that are available under their authorisation credentials. The publicly available data (i.e., free to fetch with restrictions) include general video meta-data, comments thread, limited user profile details, etc. We have developed a Java application with a mySQL database as a back-end to fetch/store only available public data.

We crawled all obtainable data related to those top 13 trended YouTube's videos in Saudi Arabia[1], uploaded/posted within a one-year timeframe (i.e. between 2016 and 2017). The gathered datasets includes more than $51,697$ comments and all the available details about reviewer profiles, such as location and used devises for posting comments. Critical demographic variables such as user-age and gender are unfortunately not available for public use, and hence, we had to implement our own classifier to predict user-genders from user-nicknames. Statistical summary of the collected datasets for our experiments in this research is given in Table I.

TABLE I. STATISTICAL SUMMARY OF THE DATASETS GATHERED FROM THE TOP 13 TRENDED YOUTUBE'S VIDEOS IN SAUDI ARABIA BETWEEN 2016 AND 2017.

#Vid.	Video ID	Viewers	Comments	like	dislike
1	1rUn2j1hLOo	11134481	14672	107259	27645
2	U62F_sl-D	2064982	3315	14811	3966
3	tE22WlRdEek	3979358	1810	6847	3106
4	wHggs-hE16M	1722002	3774	19500	1882
5	3QS7j-jDATE	335536	216	411	105
6	lxp-HDSARXs	2515276	5584	39590	1701
7	1yVWXXWwgnM	3353240	7780	43852	6956
8	5U02EzUWDmc	3281900	2406	68162	8534
9	oIHuAwYLW-U	10770907	3231	223952	17345
10	gOOOhdXT6QU	1598790	3471	49100	4061
11	Bzveyqagqeo	565006	876	2104	1015
12	HLX6D1jDzCg	499442	852	3338	1173
13	NHkCN058yFE	1095252	3710	8239	2621
Total		**42916172**	**51697**	**587165**	**80110**

B. NLP-Sentiment Analysis for Classifying Textual Comments

As the standard YouTube's API does not provide sentiment information correlated with each posted comment, we implement our own multi-classes sentiment classification algorithm for text written in Arabic. Our sentiment classifier algorithm is modelled using an optimised version of *bag-of-words* approach and analyses deeply sentiment scores in five-pole scale (i.e. *Positive, Negative, Mixed, Criticism, Neutral*) taking into consideration their polarities. The bag-of-words approach is popular in natural language processing, which is a machine learning method of feature extraction with textual data [35]. Rather than measuring only the presence and/or frequency of known words for a given textual comment, we also consider the sentiment score of each matched word from our predefined lexicon dictionary. We build a rich Arabic lexicon dictionary that includes more than $72,000$ sentimentally classified units, some of them have been extracted from publicly available datasets such as SemEval [36, 37] and from review repositories of some domains[4], including Movies, Hotels, Restaurants and

[2]Alexa Internet, Inc June 2019. https://www.alexa.com/siteinfo/youtube.com

[3]YouTube Application Program Interface (API) Services - https://developers.google.com/youtube/

[4]Large Multi-Domain Resources for Arabic Sentiment Analysis - https://github.com/hadyelsahar/large-arabic-sentiment-analysis-resouces

Products [38]. In principle, these units have been generated by mining varieties of Arabic texts that are currently in use, and the average of their accuracy is approximately 72%.

Moreover, our text classifier algorithm allows performing a detailed analyses of viewer's comments by predicting user-genders as well as classifying comments into another three high level categories (i.e. *Information, Conversation, Non-response comments*) using a specific predefined keywords. In this paper, these high level categories, introduced and explained in [39], could give influential facts that help in understanding the currently dominated phenomena of Saudi society.

bag-of-words		NLP- Measurements		Lexicon dictionary (*sentiment probabilities*)				
Token	**occurrences**	**TF**	**IDF**	**Positive**	**Negative**	**Mixed**	**Criticism**	**Neutral**
1- تستاهلون	17	0.14	3041.00	0.13	0.26	0.31	0.17	0.13
2- شوف	281	0.14	183.98	0.16	0.13	0.37	0.16	0.18
3- التخلف	393	0.14	131.54	0.14	0.35	0.23	0.21	0.07
4- وصلكم	34	0.14	1520.50	0	0	0	0	0
5- اهل	509	0.14	101.57	0.20	0.13	0.26	0.19	0.22
6- البعران	19	0.14	2720.89	0.11	0.31	0.23	0.26	0.09
7- تفو	178	0.14	290.43	0.05	0.41	0.23	0.28	0.03
			Min probability	0.05	0.13	0.23	0.16	0.03
			Max probability	0.20	0.41	0.37	0.28	0.22
			Total probabilities	0.79	1.59	**1.63**	1.27	0.72
	The final predicted class is (***Negative***), determined by the highest score of $\sum_{i=1}^{7} cp_i$			359.83	**406.98**	-79.22	123.45	356.31

The original comment before the cleaning process: "تستاهلون .شوف التخلف وين وصلكم يا اهل البعرااان تفوووو"
The literal translation : *"Deserve it.. look at his backwardness to where it led you, O people of camels, petty insult on you"*

Fig. 2. A self-explanatory example for analysing a textual Arabic comment, represented by a two-dimensional array-like structure: *bag-of-word* across *lexicon dictionary*. The latter includes five sentiment probabilities for each word. Here, we should notice that the negative values (i.e. see -79.22) when calculating the total score $\sum_{i=1}^{bag_{size}} cp_i$, using Equation 3, can be a result of not finding tokens in the dictionary, such as the token number 4.

Algorithm 1 Creating a bag of Arabic words Algorithm.

Inputs: $Comments = \{c_1, \cdots, c_k\}$, $k \in \mathbb{N}$: a set of all extracted comments from the datasets.
Outputs: Bag: a set consisting of a cleaned bag of Arabic words, such that each word t has a numerical attribute t_{count} for holding the number of comments the t appear in.
Begin
1: $Bag := \emptyset$: initialising the empty bag set for creating distinct words.
2: **for each** c_i posted comment $\in$ *Comments* **do**
3: $t_{cleaned} \leftarrow$ clean (c_i) : remove all non-Arabic characters, conjunctions, punctuation, and repeated stressing characters from c_i except empty spaces.
4: $t_{tokenized} \leftarrow$ Tokenize (t_{clean}) : tokenizing the passed cleaned text by splitting it on single spaces.
5: **for each** t_i a cleaned token $\in t_{tokenized}$ **do**
6: **if** $t_i \notin Bag$ **then**
7: $Bag \leftarrow t_i$: append the token t_i to the list Bag.
8: **if** exist_and_first_count (c_i, t_i) **then**
8: count how many comments the t_i appear in *Comments*, i.e., at most once for each c_i.
9: set $t_{i_{count}} = t_{i_{count}} + 1$
10: return Bag.
End

The proposed algorithms are given explicitly in (Algorithms 1 and 2). Given a broad set of textual comments,

Algorithm 2 Lexicon-based Bag of Words Sentiment Classifier.

Inputs: Lex is a lexicon dictionary
Bag is the created bag of word from the Algorithm 1
tc and pc are the total number of comments and a specific posted-comment respectively.
Outputs: S_{class} is the classified class that has the maximum sentiment probabilistic scores from the five-pole scale (i.e., *Positive, Negative, Mixed, Criticism, Neutral*).
Begin
1: $dataFrame$ = makeMatrix (Bag, Lex)
2: $t_{cleaned} \leftarrow$ clean (pc)
3: $t_{tokenized} \leftarrow$ Tokenize ($t_{cleaned}$)
4: **for each** t_i a cleaned token $\in t_{tokenized}$ **do**
5: **if** $t_i \in dataFrame[Bag]$ **then**
6: $TF \leftarrow dataFrame$.TF (t_i, pc) : compute Term Frequency using Equation 1
7: $IDF \leftarrow dataFrame$.IDF (t_i, tc) : compute Inverse Document Frequency using Equation 2
8: $dataFrame[t_i][Lex]$.sentimentScores ($TF * IDF$): compute the sentiment score for each row in Lex according to their probabilities using formula Equation 3.
9: $S_{class} \leftarrow dataFrame$.maxSentimentScore() : summing the total sentiment scores for each class in Lex and then returns the class with the highest value.
10: return S_{class}
End

our approach begins by generating a bag of Arabic word using Algorithm 1 from all observed comments. We then generate a data-frame, representing a two-dimensional array-like structure, by mapping each token (word) from the bag with our predefined lexicon dictionary. Here, all columns are vectors of equal length, such that the first two vectors contain token-values and their occurrences respectively. The followed vectors correspond to measurements of the sentimental classes, obtained from our lexicon dictionary, see Figure 2 for illustration with a self-explanatory example. The generation of the data-frame is stated in Algorithm 2, see the first line.

We used Term-Frequency (TF) and Inverse-Document-Frequency (IDF) formulas for assessing how important a token is to a posted comment in corpus. These two statistical formulas, see Equation 1 and 2 are well-known measurements in text mining and information retrieval [40]. TF gives a scoring weight for each token in a document (i.e., how frequently a word appears in a comment), expressed as follows:

$$TF(t, c) = \frac{f_{t,c}}{c_{count}} \tag{1}$$

where $f_{t,c}$ is the number of times the token (or word) t appears in the posted comment c, and c_{count} is the total number of tokens. Whereas, IDF measures the score of how important a token is across documents (i.e., all observed comments), calculated by (2):

$$IDF(t, C) = \log \frac{C_{count}}{1 + |\{c \in C : t \in c\}|} \tag{2}$$

where C_{count} is the total number of extracted comments, and $|\{c \in C : t \in c\}|$ is the number of posted comments that the

TABLE II. COMMENTS CLASSIFICATION RESULTS AND THE PERCENTAGE OF IRRELEVANT COMMENTS (INCLUDING ADS) FOUND IN EACH VIRAL VIDEO

	Sentiment Classification					Keyword Classification			Ads and
#Vid.	Positive	Negative	Mixed	Criticism	Neutral	Info.	Conv.	Non-response	**irrelevant.**
1	1942	3056	1405	1125	7144	5023	2961	5817	40%
2	283	782	215	266	1769	1032	777	1390	42%
3	105	563	132	141	869	564	311	722	40%
4	861	953	745	342	873	1074	506	748	20%
5	25	69	19	23	80	74	71	67	31%
6	879	1576	939	555	1635	2179	975	1355	24%
7	603	2173	526	521	3957	2159	1104	3432	44%
8	341	709	336	413	607	881	106	482	20%
9	480	711	324	370	1346	1126	36	988	31%
10	508	703	330	298	1632	1265	561	1187	34%
11	80	194	74	155	373	238	228	309	35%
12	65	253	85	70	379	386	209	272	32%
13	304	1011	272	227	1896	878	764	1596	43%

TABLE III. RESULTS OF GENDER DETERMINATION IN EACH VIRAL VIDEO AND ESTIMATED RATE OF COMMUNICATION BETWEEN COMMENTERS

	Inferred Gender			
#Vid.	Male	Female	Unknown	**Interactions**
1	6956	3589	4127	20%
2	1607	813	895	23%
3	997	398	415	17%
4	2194	761	819	13%
5	121	43	52	33%
6	3294	1141	1149	17%
7	4182	1710	1888	14%
8	1236	544	626	4%
9	1645	745	841	1%
10	1920	787	764	16%
11	485	189	202	26%
12	464	193	195	25%
13	1850	862	998	21%

token t appears in it. We calculate *TF* and *IDF* for each token t in the posted comment c, see lines (4-7) of Algorithm 2. In line (8), we rescale data values in vectors that only correspond to the probabilities of the sentimental classes (i.e., columns with the header names: *Positive, Negative, Mixed, Criticism, Neutral* in Figure 2) using what so-called *feature scaling* multiplied by the calculated rates of token's importance $TF * IDF$ for each token t. To be more precise, rescaling these data values for the probabilities of each sentimental class is expressed as follows:

$$cp_t = TF_t * IDF_t \frac{cp_t - cp_{min}}{cp_{max} - cp_{min}} \quad (3)$$

where cp_{max} and cp_{min} are determined vertically in vectors that hold sentimental token's probabilities. Finally, the predicting sentimental class for c, stated in line (9) see Algorithm 2, is chosen according to the highest total scores of their $\sum_{i=1}^{b_{size}} cp_i$, where b_{size} is the size of the bag. Consider the example shown in Figure 2, the chosen sentimental class, the algorithm will classify the mentioned impolite comment to be *Mixed* in accordance with the total probabilities (i.e., 1.63). However, our optimised solution gives more precise classification as it takes into consideration the rates of token's importance, see the correct prediction by choosing *Negative* with total score of 406.98.

The same algorithms (i.e., Algorithms 1 and 2) are applied for classifying comments into three high level categories (i.e, *Information, Conversation, Non-response comments*), but with using different lexicon dictionary that is manually defined. Here, we carefully collect a large set of keywords that are often used in each category. For instance, comments that consist of WH-questions (as predefined keywords) at the beginning will likely be classified into *Information* category [39]. Furthermore, we have a rich database dictionary of male and female names, and we use it for predicting user-genders from user-nicknames.

Tables II and III show the generated predictions when applying our Algorithms 1 and 2 on the collected data, summarised in Table I.

C. Manual Inspiration for Quantitatively Analysing YouTube Content View

To our knowledge, there has been no idealistic method for performing video content analysis directly at the visual level. Accordingly, we have implemented a generic subjective method of interpretations by a panel of three reviewers, including the authors of this paper, moderately related to QualCA research method [41]. In essence, this subjective method involves three core independent phases:

1) The identification of the (global) most expressive themes and video categories that characterise the intentions deduced from the audio and/or visual components of video contents.
2) The coding frame, formulated in a two-dimensional thematic vector that maps the identified themes th_i with each observed video vid_i by a five-level Likert scale (i.e., from 1 to 5) [42].
3) Checking the validity of the constructed thematic vectors.

The selected 13 videos were distributed to the authors of this paper individually, and they were instructed to identify the global themes depending on what they observed in the video, regarded as a whole. Subsequently, the authors have held several remote meetings to unify all the agreed themes embedded in the video contents, wrapped into 10 distinct themes, described in Table IV. This phase was carried out during the month of January 2018. In the coding phase, the authors were requested individually to re-observe each video and scale all the 10-themes. To tackle the conflicting problems in scaling the same theme th_i vs. vid_i by the authors, the average scales has been calculated, and then rounded to the

nearest integer. After that the authors have sent the constructed thematic vectors (i.e., represented as a table shown in Table IV) to a panel of three reviewers for checking and validating the identified list of themes as well as the coding scales for each video, and no critical comments were noticed.

TABLE IV. RESULTS OF MEASURING THE 10-THEMES FOR EACH CONTENT, INCLUDING AUDIO-VISUAL COMPONENTS, OF THE SELECTED 13 YOUTUBE VIDEOS, BASED ON A SCALE FROM 1 TO 5

	#Video ID												
Themes	1	2	3	4	5	6	7	8	9	10	11	12	13
1	3	5	5	3	0	0	0	0	2	0	0	0	5
2	5	4	4	5	2	4	2	3	4	1	2	1	4
3	4	3	3	3	0	2	0	1	0	0	0	0	4
4	0	1	1	1	3	0	0	5	5	0	1	0	1
5	0	0	0	0	1	5	0	0	0	0	2	0	0
6	0	0	0	0	0	5	0	0	0	0	3	0	0
7	0	0	0	0	0	5	0	0	0	0	1	0	0
8	5	1	1	3	5	1	3	3	4	5	5	5	1
9	2	2	2	4	0	2	5	0	2	1	1	0	2
10	3	3	3	0	2	0	2	2	2	3	2	5	3

1 Opposite sex
2 Social and political issues
3 Religious issues
4 Celebrities and figures scandal
5 Defending the country
6 Supporting leaders
7 Feeling proud of the country
8 Sarcastic
9 Traditions
10 Sport and Entertainment

D. Unsupervised Learning Model for Integrating the Quantitative and Qualitative Findings

The third phase of our concept-level mixed-methods design, shown in section III, makes the quantitative and qualitative findings interdependent rationally. It involves the integration of the NLP-Sentiment analysis (cf. subsection III-B) and the thematic analysis (cf. subsection III-C) outputs as the centerpiece inputs to our unsupervised machine learning model. This unsupervised learning model gives a more in-depth insight into the relations between the quantitative and qualitative variables, allowing to better understand the nature of viral videos. The proposed model, in the third phase, is designed based on Distance Matrix[5] and Hierarchical Clustering [43]. In data mining, distance matrix is typically essential for building a hierarchy of clusters. Here, we consider *Cosine* similarity formula (i.e., usually used to measure the degree of angle between two variables) to generate our distance matrix. The formula is expressed by a dot product [44] as follows:

$$\text{Distance (A,B)} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (4)$$

where A_i and B_i are pairwise vectors containing values from two compared variables (e.g., a sentiment type as a variable vs. a specific user-gender).

```
# importing the required libraries ..
import pandas as pd
from scipy.spatial import distance_matrix
from sklearn.cluster import
    AgglomerativeClustering
...
# Loading the dataset as a CSV format and
# computing the distance matrix using the '
    scipy' library.
dataSet = pd.read_csv('cleaned-dataset.csv')
distanceMatrix = pd.DataFrame(distance_matrix(
    dataSet.values, dataSet.values), index=
    dataSet.index, columns=dataSet.index)
...
# Generating a hierarchical clustering model
    using the 'sklearn' library.
# We use the distance matrix as input to train
    the model.
hierarchicalClustering =
    AgglomerativeClustering (affinity='Cosine',
    linkage='ward')
hierarchicalClustering.fit_predict(
    distanceMatrix)
...
```

Listing 1. Python code fragments for computing distance matrix and generating a hierarchical clustering model

To clarify more, we have implemented a Python script to integrate all the inferred information acquired during the second phase (i.e., presented in Table II, Table III and Table IV). In Listing 1, we give a descriptive code fragment for creating a distance matrix between the qualitative thematic variables across the quantitative variables. Additionally, we have used an existing interactive data analysis tool called Orange[6] (i.e., a visual Python programing language for data analysis) to generate a distance matrix and hierarchical clustering figures, see them in Figure 3 and Figure 4. In principle, both figures illustrate the relations between the qualitative thematic variables across the quantitative variables. However, Figure 4 divides relations at different levels represented as a tree structure. We expand this and discuss our findings in the discussions section.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This paper is set out to investigate the types of comments posted on viral YouTube videos in Saudi Arabia, proposing a thematic classification schema to understand the rates of concern of social community in Saudi Arabia. Without paying to much attention to our technical contribution to this study, which includes implementations of our own optimised learning algorithms and metrics, the focus in this section is to give revealing insights into the figures reported in Figure 3 and Figure 4 from three perspectives that answer our research questions:

1) Exploring the categorisation and current concerns of commenters under our qualitative themes. Here, we deeply dig into what more or less concerned the commenters depending on their genders as well as the level of their interactions.
2) Understanding what thematic categorisations are more relevant in boosting the spread of videos.

[5]Distance Matrix is a mathematical square matrix that contains the numerical distances between the items in two-dimensional-array

[6]Orange tool is an interactive data analysis workflows https://orange.biolab.si/

Distance Matrix	Positive	Negative	Mixed	Criticism	Neutral	Information	Conversation	Non-response comments	Male	Female	Unknown Gender	User interactions	Ads and irrelevant comments
Positive									0.022	0.088	0.091	0.198	0.429
Negative									0.038	0.022	0.014	0.299	0.390
Mixed									0.041	0.096	0.104	0.190	0.404
Criticism									0.069	0.107	0.091	0.195	0.437
Neutral									0.069	0.014	0.022	0.363	0.247
Information									0.033	0.063	0.071	0.223	0.497
Conversation									0.110	0.063	0.099	0.481	0.242
Non-response comments									0.066	0.016	0.019	0.365	0.242
Male	0.022	0.038	0.041	0.069	0.069	0.033	0.110	0.066		0.036	0.044	0.255	0.437
Female	0.088	0.022	0.096	0.107	0.014	0.063	0.063	0.016	0.036		0.014	0.343	0.324
Unknown Gender	0.091	0.014	0.104	0.091	0.022	0.071	0.099	0.019	0.044	0.014		0.324	0.335
User interactions	0.198	0.299	0.190	0.195	0.363	0.223	0.481	0.365	0.255	0.343	0.324		0.338
Ads and irrelevant comments	0.429	0.390	0.404	0.437	0.247	0.497	0.242	0.242	0.437	0.324	0.335	0.338	
1 Opposite sex	0.467	0.343	0.479	0.471	0.297	0.480	0.377	0.297	0.348	0.344	0.340	0.474	0.313
2 Social and political issues	0.236	0.198	0.239	0.278	0.275	0.300	0.335	0.268	0.333	0.334	0.223	0.356	0.453
3 Religious issues	0.322	0.222	0.302	0.375	0.242	0.378	0.241	0.242	0.358	0.331	0.251	0.490	0.401
4 Celebrities and figures scandal	0.300	0.303	0.313	0.367	0.262	0.256	0.135	0.277	0.210	0.358	0.303	0.408	0.334
5 Defending the country	0.420	0.368	0.396	0.444	0.312	0.396	0.461	0.337	0.365	0.245	0.392	0.257	0.372
6 Supporting leaders	0.452	0.491	0.478	0.426	0.430	0.478	0.430	0.456	0.366	0.246	0.483	0.369	0.404
7 Feeling proud of the country	0.452	0.491	0.478	0.426	0.430	0.478	0.430	0.456	0.361	0.243	0.483	0.369	0.404
8 Sarcastic	0.415	0.291	0.395	0.416	0.314	0.425	0.341	0.314	0.327	0.343	0.301	0.373	0.402
9 Traditions	0.188	0.124	0.222	0.251	0.152	0.206	0.196	0.145	0.316	0.348	0.130	0.290	0.343
10 Sport and Entertainment	0.321	0.418	0.329	0.308	0.429	0.404	0.450	0.452	0.334	0.353	0.445	0.322	0.217

Fig. 3. Distance matrix, based on cosine similarity formula, representing the relations between qualitative thematic variables across the quantitative variables.

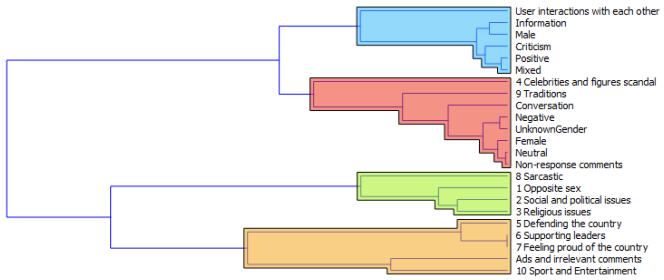

Fig. 4. Hierarchical clustering dendrogram, representing the relations between qualitative thematic variables across the quantitative variables.

3) Predicting the next shift wave of concerns of social commenters through observing all the event' times (i.e., time of posting or replying to a comment) on the timestamp.

After exploring the above-aforementioned perspectives in the next subsection, we discuss the threats that can impact the validity of our findings, and then we give a brief consideration regarding the ethical issues.

A. Understanding the Categorisations and Concerns of Saudi Society

Figure 5 shows four different distributions of our thematic categorisations, resulted by clustering our internal outputs (i.e., obtained after performing the qualitative and quantitative analysis parts independently). By taking a closer look at the mentioned figures as well as the disparity in percentages, one can observe, at a glance, the following points:

- The highest three categories in terms of social community concerns lie in (*Sport and Entertainment, Traditions* and *Sarcastic*), which constitute roughly half of the society's concerns in a total percentage of 49% (i.e. $18\% + 16\% + 15\%$), see (A) at the top left of Figure 5.
- The differences between males against females, as shown in (B), look slight by an average of about 11% except *Celebrities and figures scandal*, where they look more common among females than males by an approximate of 26%. This result is in line with the clustering dendrogram, presented in Figure 4. Here, the clustering figure gives different analytical readings, one of which is the overall behaviors of males against females. Roughly speaking, the produced clusters indicate that males appear more involved in making *positive, mixed* and *criticism* comments than females. These comments seem associated with all categorical themes apart from *Traditions* and *celebrities and figures scandal*. In contrast, however, females tend to post more *negative* and *neutral* comments associated with only *traditions* and *celebrities and figures scandal* categories.
- Interaction between commentators and their responses to each other is high in issues related to (*Political*, or *Sarcastic issues*), and gradually decrease in the other categories. Unsurprisingly, this is visibly analogous with the high presence of irrelevant comments, which can be a result of the exploitation of advertising owners in these categories, see (C) and (D).
- No much attempt is made by the commenters to delve into and engage in issues related to (*Political, Religious*, or *Opposite Sex issues*). However, there appears an advertising focus on these categories, which could be the reason behind boosting the level of communications between commenters.

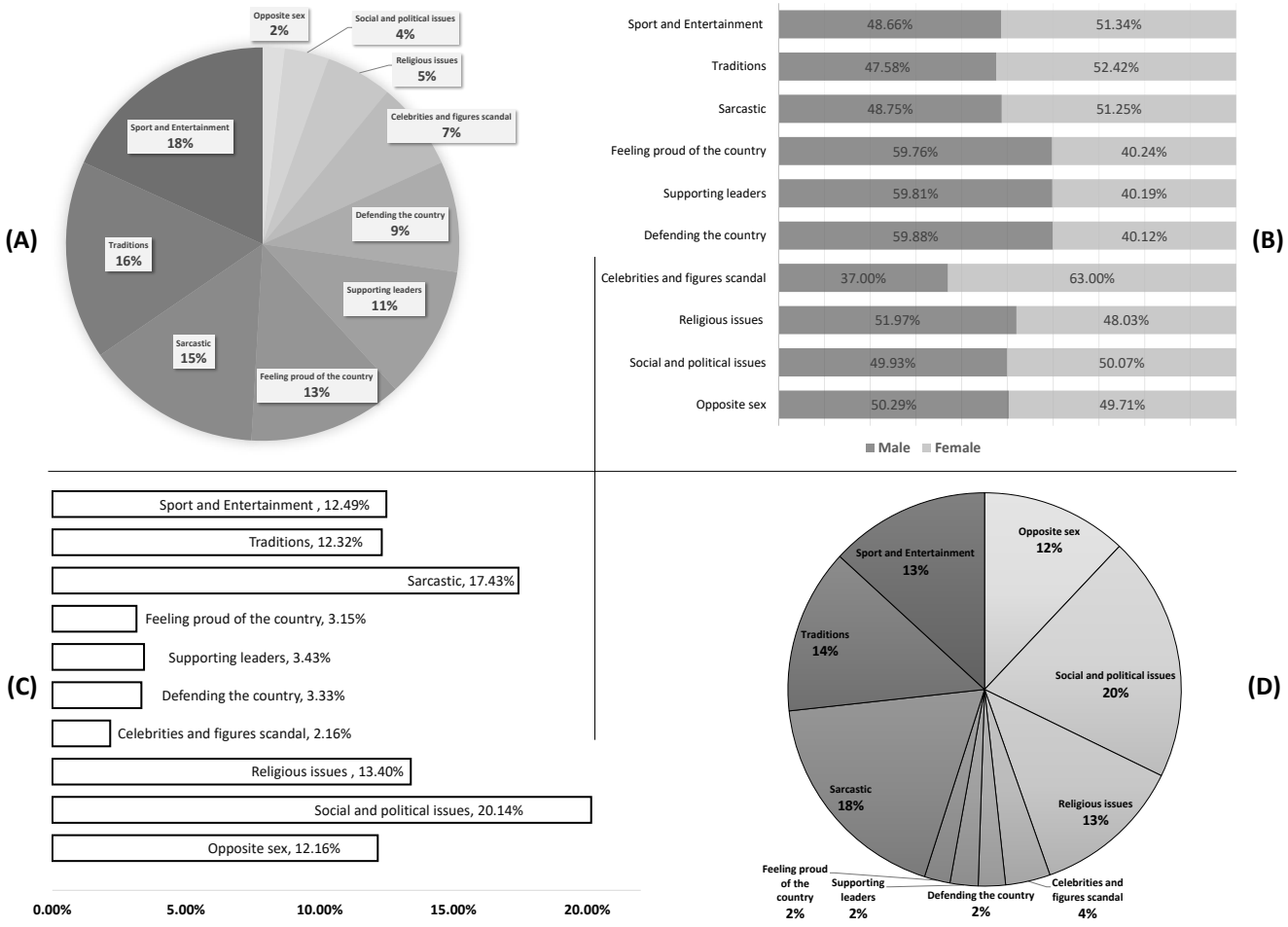

Fig. 5. Different distributions of our thematic qualitative categorisations. In (A), we show the distribution with percentages based on shared and/or posted comments. (B) shows the distribution in accordance with the percentage of male vs. female commentators. (C) focuses on the percentage of commentator-interactions with each other. The distribution, in (D), is determined based on the number of shared irrelevant comments, including textual Ads.

In order to entirely understand the phenomena of viral YouTube videos, one should collect data from several reliable resources that provide, e.g., tracing data of sharing video-links through external social networking platforms (or through existing mainstream media) or providing data that describe how much robot software tools being used for spreading video-links globally. Since such data are outside the scope of YouTube platform, let us assume *a hypothesis with a typical scenario where the genuine reason that led a particular video to go viral is just the content*. This trivial hypothesis simplifies our understanding of this phenomena by preciously examining one aspect (i.e., video's content in addition to its comments) while neglecting all other aspects that are difficult to obtain. In this context, we see the leading cause, confined to having an attractive positive or negative content, is the implication of what is in line with (1) the main interests of regular viewers or (2) with things that advertising organisations care about. Referring to such rational grounds, we figure out, from the results reported in Figure 5, that the prevalent categorical themes are *Traditions* and *Sarcastic*, thus supporting these categorises may contribute significantly to make an extremely viral video. Furthermore, what seems attracts the social community, in particular, is the promotion of entertainments and/or traditions issues associated with sarcastic or rude remarks. Advertisers, however, are keen to exploit contents correlated to politics, religion, opposite-sex issues and, in the meanwhile, surrounded by also rude remarks. Therefore, boosting these circumstances are likely the main reasons behind letting regular videos go viral.

Concerning our prediction for the changes in the distributed thematic categorisations, we have conducted a specific experiment to measure the changes. The concept of this experiment lies in adding event's times as an additional dimension to our dataset. To avoid the ambiguity, we firstly have broken the time-line down into several equal intervals, such that all our selected videos were accessible on-line during the first interval. Then, for each interval i, we generate a distance matrix by extracting comments, posted within i, and analysing them using our sentiment classifier. In (A) of Figure 6, we describe how the classified comments are fluctuated over time. By computing the generated set of distance metrics, using Forecast and Trendline equation[7], we were able to estimate the percentages of the change in each category, see (B) of Figure 6. This figure here reports that the changes, whether up or down,

[7]Forecast and Trendline are popular equation in MS-Excel tool.

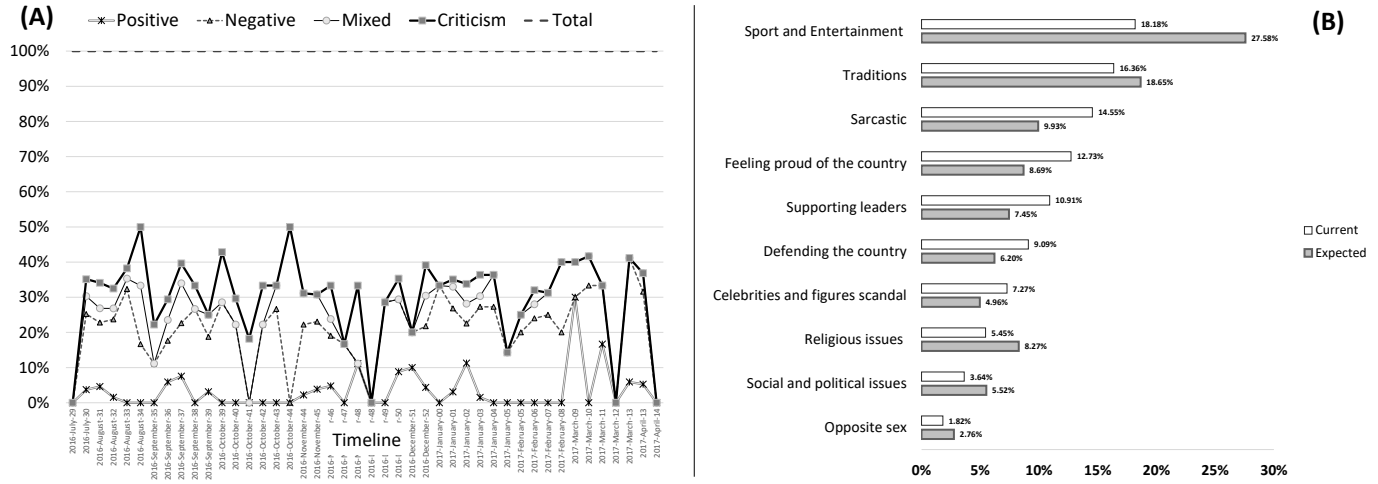

Fig. 6. Predicting the change in the current categorisations and concerns of social community. In (A), we describe the classified comments sentimentally for the first video (shown in Table I) over time. (B) illustrates the predicted change in each thematic category

would be inconsiderable of around 6%, except *Sport and Entertainment* that is expected to get progressed by almost 10%.

B. Discussion and Threats to Validity

The feasibility of using our complex AI-based approach in analysing the behaviors of YouTube communities depends primarily on the quality of the collected data, and this fact probably applies to most of in-use machine learning solutions. Consequently, our thought here is that viral YouTube videos can be considered as a fertile place for extracting high-quality dataset, resulting in producing accurate readings after correctly conducting required analysis. Concerning the soundness of our experiment, we discuss the threats that can impact the internal as well as the external validity of our results.

In internal validity (i.e., related to aspects that could have affected our finding), the threats may include (1) inaccurate predictions by our sentiment classifier, and (2) the improper use of cosine metric for generating distance matrices and hierarchical clustering (i.e., different metric may fit better in our approach such as *Manhattan* or *Euclidean*). For inaccurate predictions issue, we are not claiming that our sentiment classifier would give 100% correct predictions (no text-classifiers could reach this percentage), but accepting a specific prediction would often be based on a predefined threshold for a particular domain. The threshold considered in this paper is relatively close to the lowest accuracy of our lexicon units (i.e., about 63%) as we did exclude all lexicon units that have poor accuracy. This mean, the accuracy of our predictions should be above 63%, and such percentage is acceptable from the author's point of view. Regarding the use of cosine formula, intuitively using different formula will generate different results. However, cosine metric has been widely used for measuring preciously lexical similarity, and it is a typical metric for examining short text [45].

Threats to external validity investigate the scope of generalising the research findings. Here, a potential threat is represented by having incomplete data, collected from a limited number of (13) videos. While our approach deals with a single social networking platform (YouTube) in collecting data, there still relevant data left unconsidered, e.g., data from other social networking platforms as well as from chatbots software tools. However, as explained in the previous subsection, obtaining such data from external resources (i.e., outside the scope of YouTube platform) is not possible. Hence, we attempted to apply a robust and sophisticated research methodology using unsupervised machine learning for in-depth analysis and understanding. Besides, we are aware that our findings are based on a small number of carefully selected viral videos, but for ensuring a proper generalisation of our findings, we have extracted all shared comments (i.e., more than $51,697$ comments, see the details in Table I) within a one-year timeframe.

C. Ethical Considerations

Emotional feeling is individual privacy, and mining such individual privacy of a particular person evokes a significant concern regarding ethic legitimately. As intelligent machine learning systems are becoming more powerful and superior at understanding a human conversation, and their relationships, they could go beyond human ability in revealing their privacies, and hence raising critical questions to be addressed around security/privacy [46]. Technically speaking, mining what people express emotionally in the virtual social media worlds, as conducted in our experiments, is prone to random errors in disclosing the reality of the physical world. This means the predicted information, by mining algorithms, is not highly reliable and, therefore, could result in making ill-informed decisions.

Text mining and sentiment analysis approach on public resources of social media, as a knowledge-driven technique, are meant to give high societal level analytics. Despite this fact, our proposed approach is not designed to support oppressive regimes for identifying dissents and/or applying censorship. In this research, the collected datasets from YouTube's API are public and do not contain any details related to the identity of commenters. However, we have no attempt to use the

inferred information, such as user-genders, to evaluate the private intellectual orientations of commenters.

To the best of our knowledge, no standard ethical guidelines exist to be implemented during the development of an artificial intelligence tool. However, there appears a promising attempt, which is not finalised yet, by a research group called Partnership on AI (PAI) to study the regulations and create such important guidelines [47].

V. Conclusion

This paper contributes a convergent analysing approach that can be applied, with negligible customisation, to any social video platform for in-depth analyses and comprehension. The principle underlying this approach depends on an unsupervised machine learning technique that integrates the internal outputs, obtained by applying qualitative and quantitative methods independently. For the latter method, we have introduced an optimised version of a well-known Bag of Words algorithm to sentimentally classify any given Arabic text into a five-pole scale using a rich lexicon dictionary. Our work also rationalised the importance of artificial intelligence (including NLP and machine learning) when dealing with a complex dataset that requires text mining analysis or analysing user behaviors.

We have empirically analysed $51,697$ comments, left on 13 trended YouTube videos along with their contents, for studying the phenomena of virality in Saudi Arabia. One of our main findings revealed that boosting entertainments, traditions, politics, and/or religion issues when making a video, that is associated in somehow with sarcastic or rude remarks, is likely the preeminent impulse for letting a regular video goes viral.

In the future, we intend to further optimise our parallel mixed-methods by semi-automating all the internal parts in a web-based application. We will be investigating on also optimising our sentiment classifier by taking into consideration the linguistic structure and grammar of texts.

References

[1] G. Feroz Khan and S. Vong, "Virality over youtube: an empirical analysis," *Internet research*, vol. 24, no. 5, pp. 629–647, 2014.

[2] E. Botha, "A means to an end: Using political satire to go viral," *Public Relations Review*, vol. 40, no. 2, pp. 363–374, 2014.

[3] K. Nahon and J. Hemsley, *Going viral*. Polity, 2013.

[4] R. A. STAN and C. Ana, "Emotions–drivers of online virality content characteristics of viral blog articles in romania," *Local versus Global*, p. 694, 2015.

[5] R. Wang, W. Liu, and S. Gao, "Hashtags and information virality in networked social movement: Examining hashtag co-occurrence patterns," *Online Information Review*, vol. 40, no. 7, pp. 850–866, 2016.

[6] M. Castells, "Networks of outrage and hope: Social movements in the internet age polity press," 2012.

[7] A. J. Mills, "Virality in social media: the spin framework," *Journal of public affairs*, vol. 12, no. 2, pp. 162–169, 2012.

[8] A. L. Montgomery, "Applying quantitative marketing techniques to the internet," *Interfaces*, vol. 31, no. 2, pp. 90–108, 2001.

[9] S. Alhabash and A. R. McAlister, "Redefining virality in less broad strokes: Predicting viral behavioral intentions from motivations and uses of facebook and twitter," *new media & society*, vol. 17, no. 8, pp. 1317–1339, 2015.

[10] M. Guerini, C. Strapparava, and G. Ozbal, "Exploring text virality in social networks," in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[11] S. Alhabash, J.-h. Baek, C. Cunningham, and A. Hagerstrom, "To comment or not to comment?: How virality, arousal level, and commenting behavior on youtube videos affect civic behavioral intentions," *Computers in human behavior*, vol. 51, pp. 520–531, 2015.

[12] R. E. Guadagno, D. M. Rempala, S. Murphy, and B. M. Okdie, "What makes a video go viral? an analysis of emotional contagion and internet memes," *Computers in Human Behavior*, vol. 29, no. 6, pp. 2312–2319, 2013.

[13] K. Nelson-Field, E. Riebe, and K. Newstead, "The emotions that drive viral video," *Australasian Marketing Journal (AMJ)*, vol. 21, no. 4, pp. 205–211, 2013.

[14] Q. Bai, Q. V. Hu, L. Ge, and L. He, "Stories that big danmaku data can tell as a new media," *IEEE Access*, vol. 7, pp. 53 509–53 519, 2019.

[15] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "Characterizing and modelling popularity of user-generated videos," *Performance Evaluation*, vol. 68, no. 11, pp. 1037–1055, 2011.

[16] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer, "Catching a viral video," *Journal of Intelligent Information Systems*, vol. 40, no. 2, pp. 241–259, 2013.

[17] L. Jiang, Y. Miao, Y. Yang, Z. Lan, and A. G. Hauptmann, "Viral video style: A closer look at viral videos on youtube," in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, p. 193.

[18] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using early view patterns to predict the popularity of youtube videos," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 365–374.

[19] D. A. Shamma, J. Yew, L. Kennedy, and E. F. Churchill, "Viral actions: Predicting video view counts using synchronous sharing behaviors," in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[20] A. Susarla, J.-H. Oh, and Y. Tan, "Influentials, imitables, or susceptibles? virality and word-of-mouth conversations in online social networks," *Journal of Management Information Systems*, vol. 33, no. 1, pp. 139–170, 2016.

[21] A. Vaish, R. Krishna, A. Saxena, M. Dharmaprakash, and U. Goel, "Quantifying virality of information in online social networks," *International Journal of Virtual Communities and Social Networking (IJVCSN)*, vol. 4, no. 1, pp. 32–45, 2012.

[22] R. Zhou, S. Khemmarat, L. Gao, J. Wan, J. Zhang, Y. Yin, and J. Yu, "Boosting video popularity through keyword suggestion and recommendation systems," *Neurocomputing*, vol. 205, pp. 529–541, 2016.

[23] K. English, K. D. Sweetser, and M. Ancu, "Youtube-ification of political talk: An examination of persuasion appeals in viral video," *American Behavioral Scientist*, vol. 55, no. 6, pp. 733–748, 2011.

[24] K. Nahon and J. Hemsley, *Going viral*. Polity, 2013.

[25] R. Miller and N. Lammas, "Social media and its implications for viral marketing," *Asia Pacific Public Relations Journal*, vol. 11, no. 1, pp. 1–9, 2010.

[26] I. Mohr, "Going viral: An analysis of youtube videos," *Journal of Marketing Development and Competitiveness*, vol. 8, no. 3, p. 43, 2014.

[27] D. Southgate, N. Westoby, and G. Page, "Creative determinants of viral video viewing," *International Journal of Advertising*, vol. 29, no. 3, pp. 349–368, 2010.

[28] J. Hautz, J. Füller, K. Hutter, and C. Thürridl, "Let users generate your video ads? the impact of video source and quality on consumers' perceptions and intended behaviors," *Journal of Interactive Marketing*, vol. 28, no. 1, pp. 1–15, 2014.

[29] J. Huang, S. Su, L. Zhou, and X. Liu, "Attitude toward the viral ad: Expanding traditional advertising models to interactive advertising," *Journal of Interactive Marketing*, vol. 27, no. 1, pp. 36–46, 2013.

[30] O. F. Koch and A. Benlian, "Promotional tactics for online viral marketing campaigns: how scarcity and personalization affect seed stage referrals," *Journal of Interactive Marketing*, vol. 32, pp. 37–52, 2015.

[31] J. M. Leonhardt, "Tweets, hashtags and virality: Marketing the affordable care act in social media," *Journal of Direct, Data and Digital Marketing Practice*, vol. 16, no. 3, pp. 172–180, 2015.

[32] E. Shehu, T. H. Bijmolt, and M. Clement, "Effects of likeability dynamics on consumers' intention to share online video advertisements," *Journal of Interactive Marketing*, vol. 35, pp. 27–43, 2016.

[33] C. Tucker, "Virality, network effects and advertising," *The Networks, Electronic Commerce, and Telecommunications*, 2011.

[34] H. S. Kim, "Attracting views and going viral: How message features and news-sharing channels affect health news diffusion," *Journal of Communication*, vol. 65, no. 3, pp. 512–534, 2015.

[35] A. Kao and S. R. Poteet, *Natural language processing and text mining*. Springer Science & Business Media, 2007.

[36] S. Kiritchenko, S. M. Mohammad, and M. Salameh, "Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases," in *Proceedings of the International Workshop on Semantic Evaluation*, ser. SemEval 16, San Diego, California, June 2016.

[37] S. K. Mohammad Salameh, Saif M. Mohammad, "Arabic sentiment analysis and cross-lingual sentiment resources," http://saifmohammad.com/WebPages/ArabicSA.html, 2019, [Online; accessed 1-September-2019].

[38] H. ElSahar and S. R. El-Beltagy, "Building large arabic multi-domain resources for sentiment analysis," in *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II*, 2015, pp. 23–34. [Online]. Available: https://doi.org/10.1007/978-3-319-18117-2_2

[39] A. Madden, I. Ruthven, and D. McMenemy, "A classification scheme for content analyses of youtube video comments," *Journal of Documentation*, vol. 69, no. 5, pp. 693–714, 2013. [Online]. Available: https://doi.org/10.1108/JD-06-2012-0078

[40] A. Rajaraman and J. D. Ullman, *Data Mining*. Cambridge University Press, 2011, p. 1–17.

[41] B. M. Wildemuth, *Applications of social research methods to questions in information and library science*. ABC-CLIO, 2016.

[42] W. L. Neuman, "Social science methods: Quantitative and qualitative approaches," 2011.

[43] T. Hastie, T. Robert, and J. Friedman, "The elements of statistical learning: Hierarchical clustering," 2009.

[44] P. Dangeti, *Statistics for machine learning*. Packt Publishing Ltd, 2017.

[45] G. Sidorov, A. F. Gelbukh, H. Gómez-Adorno, and D. Pinto, "Soft similarity and soft cosine measure: Similarity of features in vector space model," *Computación y Sistemas*, vol. 18, no. 3, 2014.

[46] N. Hutchins, Z. Kirkendoll, and L. Hook, "Social impacts of ethical artifical intelligence and autonomous system design," in *2017 IEEE International Systems Engineering Symposium (ISSE)*, Oct 2017, pp. 1–5.

[47] "Tenets - partnership on ai," https://www.partnershiponai.org/tenets/, 2019, [Online; accessed 1-September-2019].