# Understanding Digital Ethnography: Socio-computational Analysis of Trending YouTube Videos

Muhammad Nihal Hussain[1], Kiran Kumar Bandeli[1], Serpil Tokdemir[1], Samer Al-khateeb[2], Nitin Agarwal[1]
[1]University of Arkansas at Little Rock (UALR)
Little Rock, Arkansas, USA
{mnhussain, kxbandeli, sxtokdemir, nxagarwal}@ualr.edu
[2]Creighton University
Omaha, Nebraska, USA
sameral-khateeb1@creighton.edu

*Abstract*— **The online video sharing website - YouTube, which was launched in February 2005 to help people share videos of well-known events, has rapidly grown to be a cultural phenomenon for its massive user-base. According to Alexa, the web traffic monitoring tool by Amazon, YouTube is the second most popular website globally. Still there is a lack of systematic research - both qualitative as well as quantitative - focusing on the video-based social networking site as compared to other social media sites. Video comments serve as a potentially interesting data source to mine implicit knowledge about users, videos, categories, and community's interests. In this research, we studied top 200 YouTube videos trending daily for a 40-day period separately in the United States of America (USA) and the Great Britain (GB) regions. We collected data for 7,998 videos trending in the USA throughout the 40-day time period and 7,995 videos trending in the GB regions. We studied content engagement behavior of users in the USA and GB regions by analyzing views, likes, dislikes, and comments on the set of trending videos. The study helped us glean some of the digital ethnographic behaviors of the users in these two regions. This paper presents highlights of the similarities and differences observed in such behaviors between the USA and GB regions.**

***Keywords-YouTube; digital ethnography; USA; GB; social network analysis***

## I. INTRODUCTION

YouTube, an online video sharing website was launched in early 2005 to help people share videos [1]. Since then sharing video content has become a cultural phenomenon. According to Alexa, the web traffic monitoring service owned by Amazon, YouTube accounts for 15.3% of traffic from search on Internet [2]. YouTube is also the second most popular website globally with over 300 hours of videos uploaded every minute and 5 billion videos watched every single day [3]. YouTube overall, and even YouTube on mobile alone, reaches more 18–34 and 18–49-year-olds than any cable network in the USA. YouTube has launched local versions of its platform in more than 88 countries. One can navigate YouTube in 76 different languages covering 95% of the Internet population [3].

Several investigations have reported digital communication tools especially social media plays an important role in how people interact, communicate, and share information. As evident in the mass protests and cyber campaigns, social media platforms help individuals to coordinate, spread messages, organize, and mobilize support for their efforts. While YouTube and other social media sites have helped frame and change the nature of online discourse, prolific linking of the content across multiple social media sites has helped spread narratives at lightning fast speeds. Hence, it is imperative to examine the emerging video-based social media platforms.

While significant body of work exists that analyze Twitter and other such social media platforms, there is a lack of systematic research focusing on video-based social networking sites. A few studies shed insights into the dynamics of online discussions on YouTube [4]. In addition to easy publishing and distribution of nearly any kind of video content, YouTube provides various features to engage with the video content, such as liking or disliking a video, commenting on a video, replying to a comment, liking or disliking a comment, or posting a video response. Comments on the videos may also be studied to extract insights into audience reactions to important issues or towards particular videos. Comments serve as a potentially interesting data source to mine implicit knowledge about the video's content, viewers, regions, and community interests. While some studies have focused on specific genre of videos (e.g., childbirth, coming out for lesbian, gay, bisexual, transgender and queer – LGBTQ – communities) [5] to understand the society's pulse, others have focused on the type of information in these videos [6] and the threats these videos pose to the societies [7].

In this research, we study the content engagement and consumption behaviors on YouTube that further helps develop a digital ethnographic mapping of user behaviors in terms of likes, comments, sentiments, and cross linking with other social media channels – compared across the United States of America (USA) and the Great Britain (GB). We analyzed top 200 YouTube videos trending daily for a 40-day time period separately in the USA and the GB regions. The dataset was obtained from Kaggle [8] and includes title, category, URL, ID, comments, views, likes, and dislikes for each video. Further, we enhanced the dataset by extracting associations with other social media platforms from the YouTube channels of these trending videos. This enhanced dataset was used to conduct a digital ethnographical

 ISBN: 978-1-61208-673-6 

mapping of users' video content generation, sharing, and consumption behaviors. We found interesting correlations between various ways users engage with videos for both USA and GB regions. We also found multiple similarities and stark differences in behavior of users of these regions. We further studied how prolific YouTubers leverage other social media platforms to disseminate their content more widely and gain audience. Detailed analysis and implications of the findings are presented in Section III. Next, we present literature review in Section II.

## II. LITERATURE REVIEW

Steady rise in YouTube's popularity has attracted a surge of research. This section summarizes studies on YouTube users' watching preferences, the video recommendation system, users commenting behaviors and deviant behaviors.

For online video watching in general, a study of USA internet users in 2009 [9] focused on viewer's preferences of categories. The study found that 50% of adults had watched funny videos, 38% had watched educational videos, 32% had watched TV shows or movies, and 20% had viewed political videos. In terms of common content categories in YouTube, music videos are a significant presence in YouTube, probably accounting for about a quarter of videos with entertainment, comedy and sports categories [10]. 60% of videos are watched at least 10 times during the first day in which they are posted [11]. One of the very first studies also showed that videos that did not attract many viewers within the first few days of publication were unlikely to grow an audience later on [11]. Even so, once a viewer is on the YouTube website, by showing related videos (or recommended videos) to him/her, YouTube attempts to increase the time spent by individuals on the site itself. Davidson et. al [12] explain the video recommendation system of YouTube. YouTube uses association rule mining (videos watched in within same session) to generate sets of related videos. To compute personalized recommendations for a user, user's activities or interactions like videos watched, liked, rated, or added to playlist, are used to build a seed set of recommended videos. Videos related to these seed set are selected as relevant videos but to enhance user experience, a subset of these videos with/from diverse categories are ultimately recommended to the user. Zhou et. al [13] argue YouTube's recommendation system has significant impact on views and virality of videos.

Although there have been some large-scale quantitative investigations into YouTube [4], few have focused on discussions in comments. Most YouTube research seems to be small-scale and qualitative that give insights into how discussions can occur around videos without giving broad overall patterns that are of use. However, the study [14] was an exception, which identified patterns in user types that can be used to predict users' likely behaviors. Yet, a little is known about YouTube discussions/comments in general including the role of sentiments. YouTube comments are textual and much research has investigated the limitations and peculiarities of the electronic text. Early studies were particularly concerned that the absence of the nonverbal channel in textual communication would lead to widespread misunderstanding, particularly in short message formats, such as mobile phone texting [15]. In response, however, a number of conventions have emerged to express sentiment in short informal text, such as emoticons and deliberate non-standard spellings [16]. Work on sentiment classification and opinion mining such as [17][18] deals with the problem of automatically assigning opinion values (viz., positive, negative, or neutral) to documents or topics using various text-oriented and linguistic features. Recent works in this area also use SentiWordNet [19] to improve sentiment classification performance.

The rise in usage and popularity of social media sites, such as YouTube, has made them particularly vulnerable for abusive behaviors from bots and troll accounts that can post spam comments in large volumes [20]. According to O'Callaghan et. al. [21], bot-posted spam comments are often associated with orchestrated campaigns and can be detected by assessing the similarity in their comments. Authors noted that these cyber campaigns remain active for long periods of time and popular videos are usually targets of these spam comments. Although, it is unknown how far user comments in this context may promote hate speech or which videos are thought to deliver their aimed potential to users.

However, the problem setting in these papers differs from ours as we conduct a comparative behavioral analysis of users' engagement and consumption of content on YouTube in terms of likes, comments, sentiments, and cross linking with other social media platforms across the USA and GB regions.

## III. METHODOLOGY

Here, we describe the datasets used for the research, overall methodology and discuss results from our analysis.

### A. Data Collection

We obtained the dataset consisting of a list of top trending videos on YouTube from Kaggle [8]. YouTube provides top 200 videos daily that are trending or popular. A dataset of these top 200 videos trending daily in USA and GB regions from September 13, 2017 to October 22, 2017 was obtained from Kaggle [8]. There were 7,998 videos trending in the USA throughout the 40-day time period with 2,395 unique videos and 7,995 videos trending in the GB with 1,769 unique videos. Out of the 4,164 videos, 770 videos were trending in both the countries. The original dataset from Kaggle has the following attributes: *URL of the video*, *video ID*, *title of the video*, *title of the channel* that published the video, *category* in which the video belongs to, *number of views*, *number of likes*, *number of dislikes*, *number of comments* the video received at the time data was collected, and the *date the video was trending*.

 ISBN: 978-1-61208-673-6 

We enhanced the dataset obtained from Kaggle by adding the description of the video, date the channel was created, and the number of subscribers of the channel. It is a common practice among prominent YouTubers to associate their various social media accounts with their YouTube channel. Using Web Content Extractor (WCE) [22], we collected these social media associations and used them to study the cross-media integration. Due to the noisy nature of the data, several data processing steps such as data standardization, noise elimination, and data formatting were performed.

In Table I, we present high-level statistics of the attributes used from USA and GB regions.

TABLE I. YOUTUBE DATA STATISTICS FOR THE USA AND GB REGIONS.

| Attributes (total number of) | USA | GB |
|---|---|---|
| Trending videos | 7,998 | 7,995 |
| Unique trending videos | 2,364 | 1,736 |
| Views | 1,652,652,122 | 1,105,805,449 |
| Likes | 109,231,965 | 91,084,832 |
| Dislikes | 5,249,389 | 3,763,875 |
| Comments | 6,528,503 | 4,894,060 |
| Commenters | 3,824,862 | 3,083,770 |
| Likes on comments | 45,816,147 | 35,514,209 |
| Replies on comments | 2,112,200 | 1,605,248 |

## B. Analysis and Findings

We conducted comparative analysis on the trending videos. First, we examined differences and similarities between users' interests in USA and GB regions by analyzing the categories of the trending videos in each region. Videos in the *Entertainment*, *Music*, and *People & Blogs* categories tend to trend more for both USA and GB regions (see Figure 1). *Comedy* videos are watched more in USA, while *Sports* videos trend in GB. Further, *Political* videos trend more in USA than in GB.

Next, we study the lifespan of trending videos in the USA and GB regions. We analyzed the number of days a video would trend on YouTube in the USA and GB regions (see Figure 2). We found that in both the regions there is a sharp decline in the popularity of videos after a certain number of days, i.e., for USA, videos trend for the first 4 days, where their popularity increases gradually after which their popularity declines sharply. There was no video that was popular for more than 8 days, i.e., the longest lifespan. For GB region, a similar trend is observed; however, a sharp decline in popularity occurs after the 6th day. From Figure 2, we can observe that videos in the GB region have longer lifespan as compared to USA videos. The longest lifespan for videos trending in USA region was 8 days, while for GB region the longest lifespan was 13 days. This dissimilarity between the two regions raises multiple questions, do users in GB tend to watch trending videos more repeatedly? And/or does the information stays longer in the social networks in GB than USA? In other words, it takes longer for information to become obsolete in GB than USA, suggesting information propagation is either slower or more long-lasting in GB than USA. We investigate this phenomenon further and reflect on the possible reasons later in the paper.

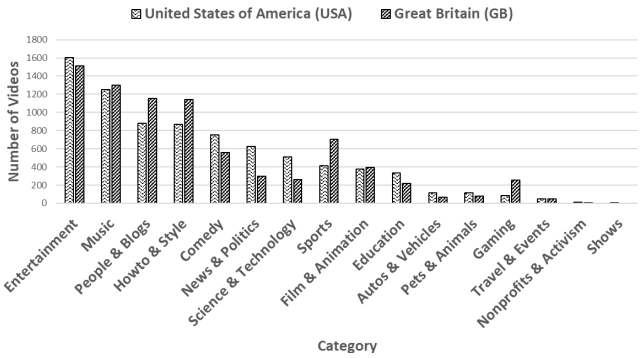


Figure 1. Number of trending videos per category for USA and GB regions.

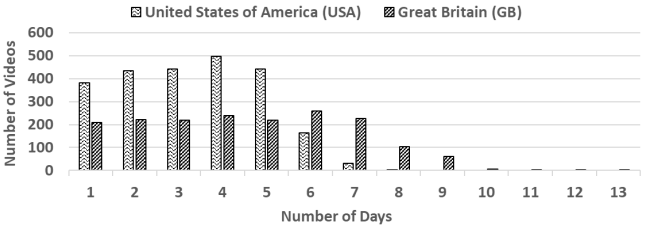


Figure 2. Lifespan of trending videos in USA and GB regions.

Next, we examine the role of integration of other social media platforms with YouTube for the trending videos in both regions. From the enhanced data, we know the various social media platforms that are associated with the trending videos. Top social media platforms that were found to be linked to the trending videos in USA and GB regions are Twitter, Facebook, Instagram, Google Plus, and Tumblr, in that order. In the USA region, most videos are affiliated with 5 social media sites, while most views were obtained by videos that were affiliated with 10 social media sites. This implies having 5 social media platforms associated with the video increases its possibility of going viral but associating more social media platforms increases the probability of video getting more views. Having other social media platforms associated with the video certainly brings more visibility through shares or other cross-media

 ISBN: 978-1-61208-673-6 

information spillovers. This might be one of the reasons for weak correlation, in Figure 3, between number of views and number of social media sites affiliated. A similar finding was observed for the GB region (Figure 4), where most videos are affiliated with 5 social media sites, while most views were obtained by videos that were affiliated with 20 social media sites.

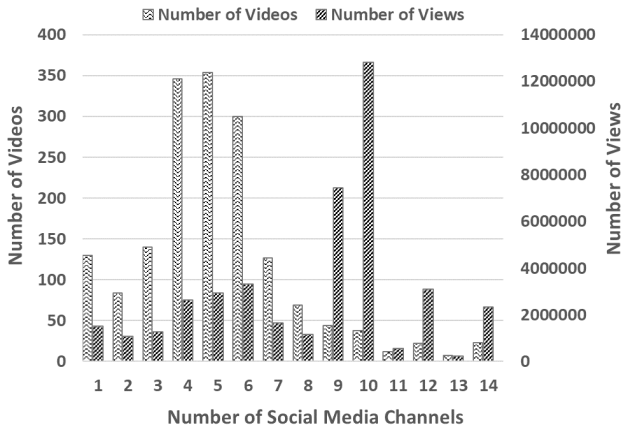


Figure 3. Social media affiliation of the trending videos in USA.

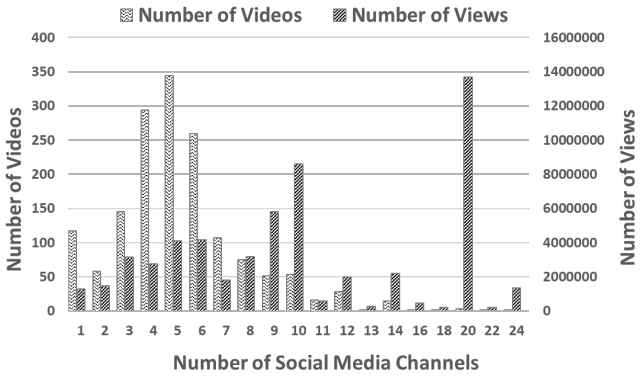


Figure 4. Social media affiliation of the trending videos in GB.

Next, we look at the network of various social media affiliations for the trending videos in both regions. The network map presented in Figure 5 helps us understand the integration of other social media sites with YouTube in the USA region. Different colors denote different clusters. Clustering analysis is done by Gephi network visualization using modularity algorithm. The network shows us the clusters of the most commonly used social media websites and how those social media platforms connect/relate and interact with each other. We found that Facebook, Instagram, Twitter, Google Plus, Pinterest, and Tumblr are the most popular social media sites. Same set of social media sites are also more frequently associated with the trending videos in the GB region (Figure 6). It is clear from the USA region's social media map that social media platforms related to the Music category (green colored nodes in top left corner of the network) clustered independently compared to other social media sites. This implies that the videos posted in the Music category were, quite understandably, shared or affiliated to completely.

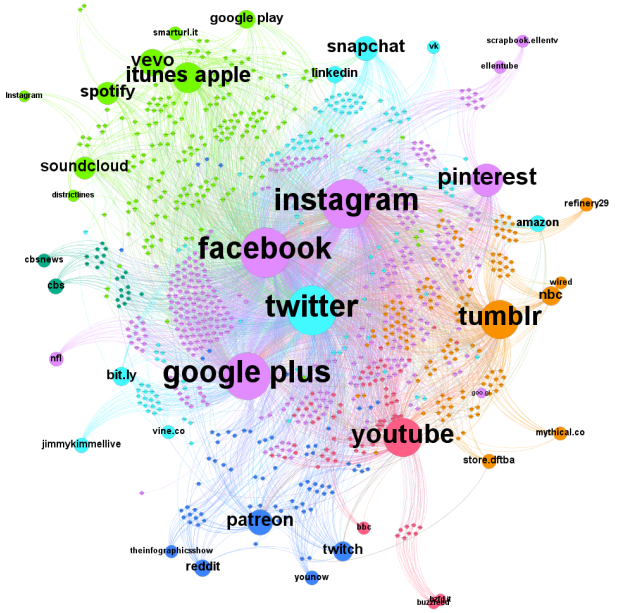


Figure 5. Social media map of the trending videos in USA

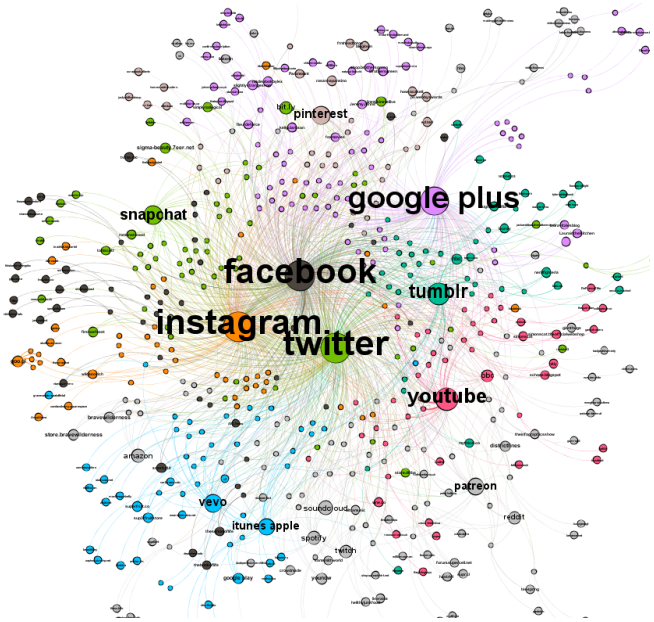


Figure 6. Social media map of the trending videos in GB

different family of social media sites. However, the media map of the GB region seems more well-connected than the USA region implying that the trending videos in GB region tend to have a richer cross-media integration on their YouTube channels that could further help explain the longer lifespan of trending videos in GB as compared to USA

 ISBN: 978-1-61208-673-6 

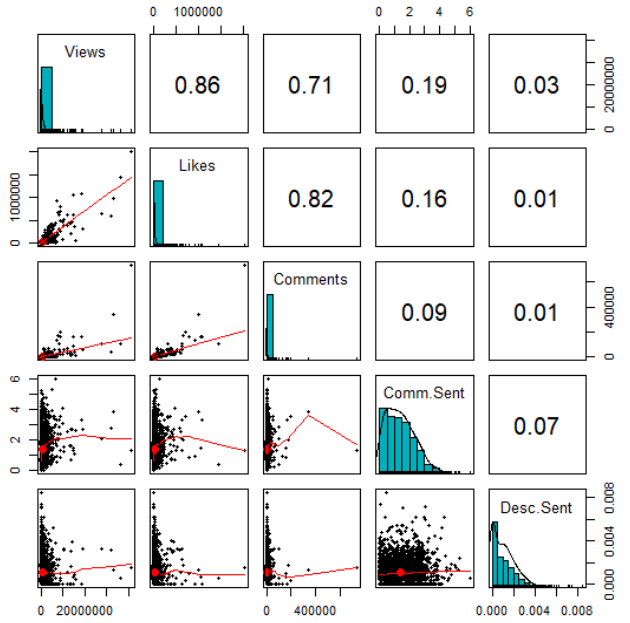


Figure 7. Correlation analysis of the trending videos in USA.

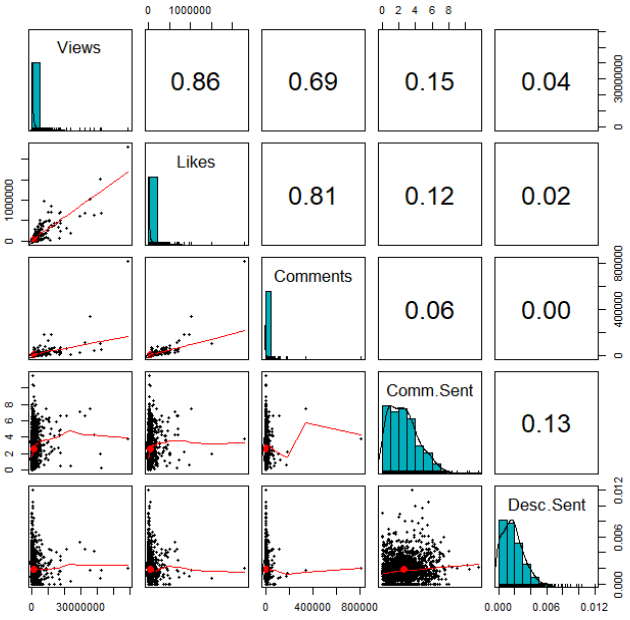


Figure 8. Correlation analysis of the trending videos in GB.

To examine users' content engagement behavior on YouTube, next, we conduct correlation analysis among number of views, number of likes, number of dislikes, number of comments, sentiments of comments, and sentiment of video description for the trending videos in both regions. The correlation charts for the USA region are shown in Figure 7. We observed that the more comments a video has the more neutral the sentiments are. On the other hand, fewer comments a video has more polarized the sentiments are (column 3 and row 4). A similar behavior was observed between sentiments of comments and number of views on videos, i.e., the more views a video has the more neutral the sentiments are, and the fewer views a video has more polarized the sentiments are (column 1 and row 4).We observed a strong correlation (0.82) between the number of comments and number of likes (column 2 and row 3) as well as between number of comments and number of views (column 1 and row 3 with correlation coefficient of 0.71). A weak correlation (0.07) was observed between sentiments from comments and sentiments from video description (column 4 and row 5). Further research is needed to investigate the reasons for a relatively weak correlation between sentiments from the comments and video description, which might be attributed to unrelated comments generated by bots or troll accounts. A similar analysis is conducted for the trending videos in GB region. The correlation charts for the GB region are shown in Figure 8. Comparing the correlation analysis of the trending videos in both regions resulted in similar findings. In both the regions, number of likes and number of views show a stronger correlation as compared to number of dislikes and number of views. This implies that if a viewer watches the whole video he/she is more likely to like the video than dislike it. A weak correlation (0.07) was observed between sentiments from comments and sentiments from video description (column 4 and row 5). Further investigations are required to test the causal relationship of these correlations.

## IV. Conclusion and Future Work

Although YouTube is the second most popular website globally with over 300 hours of videos uploaded every minute and 5 billion videos being watched every single day, there is a lack of systematic research focusing on the video-based social networking site as compared to other social media sites. There are a few studies that shed insights into the dynamics of online discussions on YouTube. Video comments serve as a potentially interesting data source to mine implicit knowledge about users, videos, categories, and community interests.

In this research, we studied top 200 YouTube videos trending in the USA and GB regions. We did a comparative study of users' video consumption behaviors and engagement to find stark difference in preferences of users of these regions. Trending videos in each region provided a glimpse of the interests of the viewers. For instance, viewers in USA are more interested in Comedy and Politics, while viewers in GB are more interested in Sports. Additionally, videos in the GB region tend to have longer lifespan compared to the USA region. Furthermore, videos in GB region are shared on more social media platforms than in USA.

We found correlation between various aspects of user engagement, such as number of comments a video received, and the polarity of these comments was inversely correlated, number of views of a video and its comments were strongly correlated. We did not test the causality of this relationship and intend to collect more data to do the same. We also

 ISBN: 978-1-61208-673-6 

found correlation coefficient for above mentioned relationships are different for USA and GB regions. Since the data for this analysis was collected only on a 40-day period, more data samples are needed for validation. Although we observed a strong correlation between number of comments and number of likes and number of views, a weak correlation was observed between sentiment of comments and sentiments of video description. The fact that commenting behavior is somewhat unrelated to the videos' content, warrants further analysis of a likely presence of spam comments from bots or troll accounts.

This research has attempted to shed light on content consumption and engagement behaviors on one of the most prominent video-based social media platforms, i.e., YouTube, in USA and Great Britain regions. As exciting as these findings are, the analysis is limited to the 40-day study period and result might vary for different time-periods. We plan to investigate the reasons for the observations in more depth and identify the ethno-digital cultural factors that help manifest the differences in the content consumption and sharing behaviors. We envision this study will help open doors for innovative and foundational work in analyzing digital behaviors of our societies (good, bad, and the ugly) as the Information and Communication Technology (ICT) landscape evolves. For instance, we can study complex questions such as, what is the role of the platform and its users in disseminating misinformation? Which societies or strata within a society are more vulnerable to misinformation? More broadly, our study motivates the need for development of methodologies to diagnose novel pathologies of online social media.

## 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. The researchers gratefully acknowledge the support.

## REFERENCES

[1] J. Hopkins, "Surprise! There's a third YouTube co-founder - USATODAY.com." [Online]. Available: https://usatoday30.usatoday.com/tech/news/2006-10-11-youtube-karim_x.htm. [Retrieved: Aug 19, 2018].

[2] "Youtube.com Traffic, Demographics and Competitors - Alexa." [Online]. Available: https://www.alexa.com/siteinfo/youtube.com. [Retrieved: August 19, 2018].

[3] S. Brain, "YouTube Company Statistics," *2017 Statistic Brain Research Institute*, 01-Sep-2016. [Online]. Available: https://www.statisticbrain.com/youtube-statistics/. [Retrieved: August 16, 2018].

[4] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: a view from the edge," in Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, 2007, pp. 15–28.

[5] K. Thorson, B. Ekdale, P. Borah, K. Namkoong, and C. Shah, "YouTube and Proposition 8: A case study in video activism," Information, Communication & Society, vol. 13, no. 3, pp. 325–349, 2010.

[6] P. L. Steinberg, S. Wason, J. M. Stern, L. Deters, B. Kowal, and J. Seigne, "YouTube as source of prostate cancer information," Urology, vol. 75, no. 3, pp. 619–622, 2010.

[7] S. P. Lewis, N. L. Heath, J. M. St Denis, and R. Noble, "The scope of nonsuicidal self-injury on YouTube," Pediatrics, p. peds. 2010-2317, 2011.

[8] M. J, "Trending YouTube Video Statistics and Comments," *Kaggle*, Oct-2017. [Online]. Available: https://www.kaggle.com/datasnaek/youtube/data. [Retrieved: August 18, 2018].

[9] K. Purcell, "The State of Online Video," *Pew Research Center: Internet, Science & Tech*, 03-Jun-2010. [Online]. Available: http://www.pewinternet.org/2010/06/03/the-state-of-online-video/. [Retrieved: August 18, 2018].

[10] X. Cheng, J. Liu, and C. Dale, "Understanding the characteristics of internet short video sharing: A youtube-based measurement study," IEEE Transactions on Multimedia, vol. 15, no. 5, pp. 1184–1194, 2013.

[11] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," Ieee/Acm Transactions On Networking (Ton), vol. 17, no. 5, pp. 1357–1370, 2009.

[12] J. Davidson et al., "The YouTube video recommendation system," in Proceedings of the fourth ACM conference on Recommender systems, 2010, pp. 293–296.

[13] R. Zhou, S. Khemmarat, and L. Gao, "The impact of YouTube recommendation system on video views," in Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, 2010, pp. 404–410.

[14] M. Maia, J. Almeida, and V. Almeida, "Identifying user behavior in online social networks," in Proceedings of the 1st workshop on Social network systems, 2008, pp. 1–6.

[15] J. B. Walther and M. R. Parks, "Cues filtered out, cues filtered in," Handbook of interpersonal communication, vol. 3, pp. 529–563, 2002.

[16] D. Derks, A. E. Bos, and J. Von Grumbkow, "Emoticons and online message interpretation," Social Science Computer Review, vol. 26, no. 3, pp. 379–388, 2008.

[17] K. Denecke, "Using sentiwordnet for multilingual sentiment analysis," in Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on, 2008, pp. 507–512.

[18] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002, pp. 79–86.

[19] M. Thomas, B. Pang, and L. Lee, "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts," in Proceedings of the 2006 conference on empirical methods in natural language processing, 2006, pp. 327–335.

[20] A. Sureka, "Mining user comment activity for detecting forum spammers in YouTube," arXiv preprint arXiv:1103.5044, 2011..

[21] D. O'Callaghan, M. Harrigan, J. Carthy, and P. Cunningham, "Network Analysis of Recurring YouTube Spam Campaigns.," in ICWSM, 2012.

[22] Newprosoft, "Web Content Extractor," *Newprosoft Web Data Extraction Software*, 2004. [Online]. Available: http://www.newprosoft.com/web-content-extractor.htm. [Retrieved: August 18, 2018].

 ISBN: 978-1-61208-673-6 