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

Muhammad Nihal Hussain, Serpil Tokdemir, Samer Al-khateeb, Kiran Kumar Bandeli, and Nitin Agarwal

University of Arkansas at Little Rock,
Little Rock AR 72204, USA
{mnhussain, sxtokdemir, sxalkhateeb, kxbandeli, nxagarwal}@ualr.edu

Abstract. YouTube, the online video sharing website, was launched in February 2005 to help people host numerous amateur and professional videos. Since then, YouTube has rapidly grown to be a cultural phenomenon for its massive user base. Still there is a lack of systematic research focusing on the video-based social networking platform compared to other social media platforms. In this research, we study top 200 YouTube videos trending daily for a 40-day period separately in the United States of America (USA) and the Great Britain region (GB), resulting in close to a 8000-video dataset for each region. The dataset includes title, category, URL, video ID, comments, views, likes, and dislikes for the videos. The dataset is further enhanced by extracting associations with other social media platforms of the YouTube channels that are hosting these trending videos. Some of the findings from analyzing this dataset include, (1) the more comments a video has, the more neutral the sentiments are, and videos that have fewer comments have more polarized sentiments; (2) similar behavior is observed when the number of views is correlated with sentiments of comments; (3) videos that were associated with 10 online platforms (including a wide variety of social media platforms) received the maximum number of views - any more or less did not help in increasing the views. This study sheds light on the digital ethnographic behaviors in terms of video based content generation, sharing, and consumption, and further allows us to glean similarities and differences between USA and GB regions.

Keywords: YouTube comments, digital ethnography, sentiment analysis, cross-media analysis, user behavior, video content consumption behavior.

1 Introduction

YouTube was launched in early 2005 to help people share videos with a global audience [1]. Since then sharing video content has become a cultural phenomenon. Recent statistics show that the traffic to or from YouTube accounts for over 20% of the total web traffic and 10% of the whole Internet traffic [2]. According to Alexa, the web traffic monitoring service owned by Amazon, YouTube is the second most popular website globally with over 300 hours of videos uploaded every minute and 5 billion videos watched every single day [3].

Several investigations have reported social media plays an important role in how people interact, communicate, and share information. While significant bodies of work analyze Twitter and other such social media platforms, there is a lack of systematic research - both qualitative as well as quantitative - focusing on video-based social networking sites. A few

studies shed insights into the dynamics of online discussions on YouTube [4, 5]. 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. We analyzed top 200 YouTube videos trending daily for a 40-day time period separately in the United States of America (USA) and the Great Britain (GB), resulting in a 8000-video dataset for each region. The dataset was obtained from Kaggle 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* of the YouTube channels hosting these trending videos. Further details of our data collection and research methodology with findings are presented in Section 3. Next, we present literature review.

2 Literature Review

In 2009 a study was conducted to analyze the online video viewership of the US Internet users. The study found that 50% of adults in the US tend to watch funny videos, 38% watches educational videos, 32% watch TV shows or movies, and 20% watch political videos [6]. In addition to the ease of uploading nearly any kind of video content, YouTube also enables viewers to interact with the video content by liking or disliking a video, commenting on a video, commenting on a comment, liking or disliking a comment, or posting a video response. Comments on the videos can be studied to understand audience's reactions to important issues or toward particular videos. Comments can also be used to mine implicit knowledge about viewers, regions, videos' content, categories, and community interests.

A recent study investigated only particular topics or a specific type of video genre and focused on the purpose and/or reception of that genre (e.g., childbirth, coming out) [7]. Other studies have concentrated on the types of information [8] or potential threats to society from the information disseminated [9]. Another study found that videos which did not attract many viewers within the first few days of releasing were unlikely to grow an audience later on [10]. The rise in usage and popularity of social media sites, such as YouTube, have made them particularly vulnerable for abusive behaviors carried out by bot accounts and internet trolls that can post spam comments in large volumes [11]. According to O'Callaghan et. al [12], bots that post spam comments remain active for long periods of time, where the primary targets are popular videos' comments. What topics the video's commenters could bring in their comments, remains an open question. Moreover, it is unknown which videos are thought to deliver their aimed potential to users. Yet, a little is known about YouTube's discussions/comments in general including the role of sentiments [2]. YouTube comments are textual and many pieces of research have investigated the limitations and peculiarities of electronic text. An early study stated that the absence of the nonverbal channels, e.g., facial expressions, in textual communication would lead to widespread misunderstanding, particularly in short message formats, such as mobile phone texting [13]. In response, however, a number of conventions have emerged to express sentiment in short informal text, such as emoticons and deliberate non-standard spellings [14]. Work on sentiment classification and opinion mining such as [15, 16] 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 [17] to improve sentiment classification performance. 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 channels across the USA and the GB region.

3 Research Methodology

A three-phase data collection process was adopted. In the first phase, we obtained a list of top videos trending on YouTube. YouTube provides top 200 videos daily that are trending or popular. A dataset of these top 200 videos trending daily in United States of America (USA) and Great Britain (GB) region from September 13, 2017 to October 22, 2017 was obtained from Kaggle [18]. There were 7,995 videos trending in GB throughout the 40-day time period with 1,769 unique videos and 7,998 videos trending in the USA with 2,395 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*. In the second phase, 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, using YouTube API. It is a common practice among prominent YouTubers to associate their various social media accounts with their YouTube channel. In the third phase, we collected social media associations of the YouTubers using Web Content Extractor (WCE) [17] 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 conducted. The analyses presented next is conducted for both USA and GB region, however due to space limitations, findings for only the USA region are described.

We examined the role of integration of other social media platforms with YouTube for the trending videos in both regions. In the USA region, majority of the videos are affiliated with 5 social media sites and the most viral videos are affiliated with 10 social media sites. A weak correlation between the number of views on a video and the number of social media sites that video is affiliated with is observed in both regions. This debunks a common myth that the more social media presence you have, the more the views. Furthermore, we look at the network of various social media affiliations for the trending videos in both regions. The network map (Figure 1) helps us understand the integration of other social media sites with YouTube in the USA region. 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. It is clear that social media platforms related to the Music category (green colored nodes) 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 different family of social media sites.

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 assessed from comments*, and *sentiments assessed from video description*. We observed (Figure 2) that the more comments a video has the more neutral the sentiments are, and fewer comments a video has more polarized the sentiments are (column 1 and row 6). A similar behavior was observed between sentiments of comments and number of views on

videos. We observed a strong correlation (0.82) between number of comments and number of likes (column 4 and row 6) as well as between number of comments and number of views (column 3 and row 6 with correlation coefficient of 0.71). This implies that if a viewer likes or views a video then he/she is more likely to comment on that video. However, a weaker correlation was observed between 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 (0.49) was observed between sentiments from comments and sentiments from video description (column 1 and row 2). Further investigations are required to test the causal relationship of these correlations.

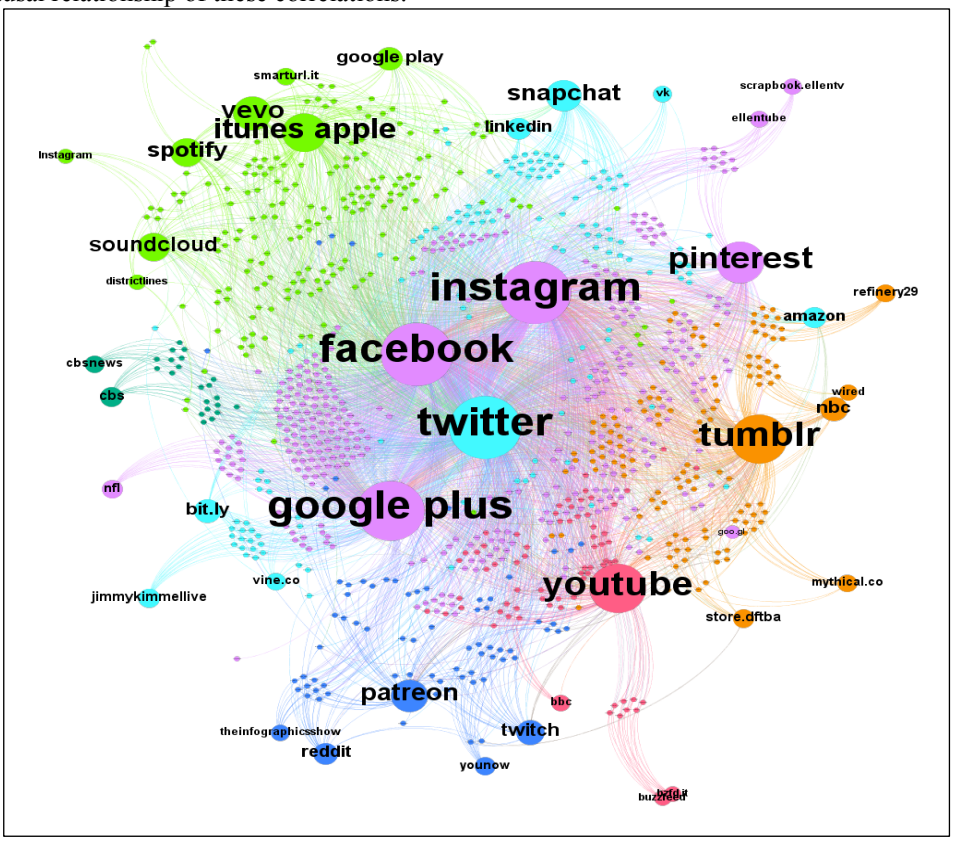

Fig. 1. Social media map of the trending videos in USA.

4 Conclusion and Future Work

Although YouTube is the second most popular website globally, there is a lack of systematic research focusing on the video-based social networking sites 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 trending YouTube videos in the USA and GB regions. Our research attempts to study

content engagement and consumption on YouTube and further helps develop a digital ethnographic mapping of user behaviors compared across the USA and the GB region. In both the regions, number of likes and number of views show a stronger positive correlation as compared to number of dislikes and number of views. Further, the fewer comments a video has the more polarized the sentiments are. This implies that if a viewer watches the whole video he/she is more likely to like the video than dislike it. 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 examination for a likely presence of spam comments from bots or troll accounts.

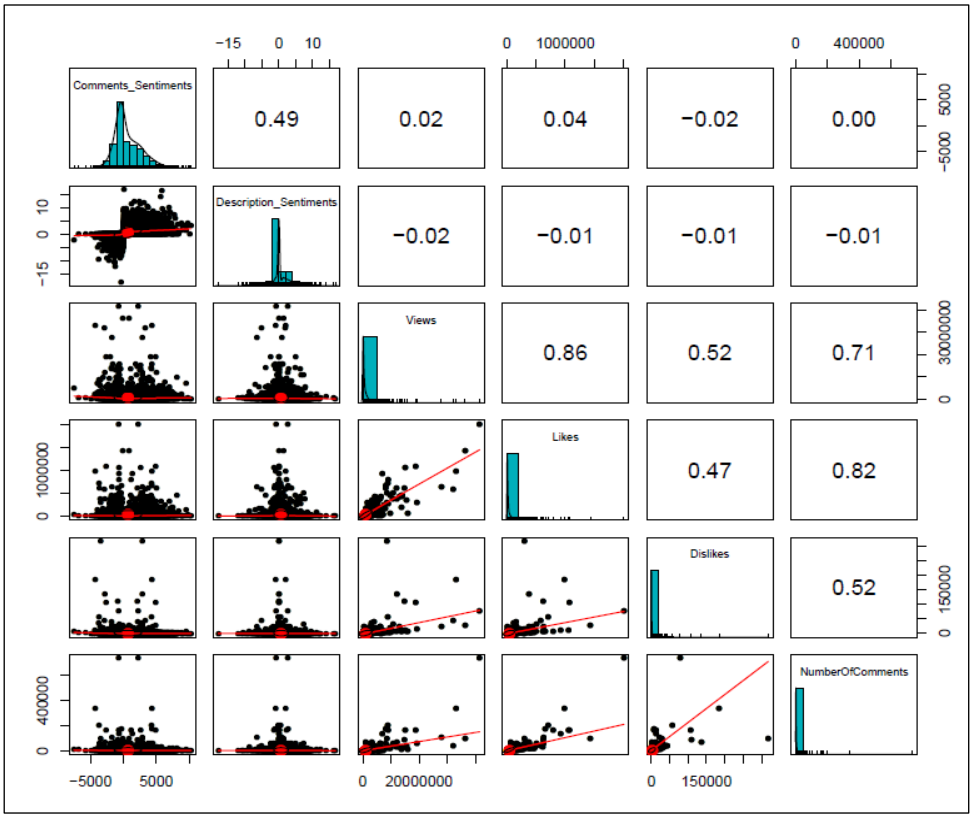

Fig. 2. Correlation analysis of the trending videos in USA.

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 Endowment 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. Hopkins, J, Surprise! There's a third YouTube co-founder, USA Today, July 15, 2011
2. Thelwall, Mike & Sud, Pardeep & Vis, Farida. (2012). Commenting on YouTube Videos: From Guatemalan Rock to El Big Bang. Journal of the American Society for Information Science and Technology. 63. 616-629. 10.1002/asi.21679
3. Bain Staistic, 2017 Statistic Brain Research Institute, September 2016
4. Ding, W. R., Gan, Y. M., Guo, X. S., & Yang, F. Y. (2009). A review on studies and applications of near infrared spectroscopy technique (NIRS) in detecting quality of hay. *Guang pu xue yu guang pu fen xi= Guang pu*, *29*(2), 358-361.
5. Gill, P., Arlitt, M., Li, Z., & Mahanti, A. (2007, October). Youtube traffic characterization: a view from the edge. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (pp. 15-28). ACM.
6. Purcell, K., The State of Online Videos, Pew Internet (2011)
7. Thorson, K., Ekdale, B., Borah, P., Namkoong, K., & Shah, C. (2010). YouTube and Proposition 8: A case study in video activism. *Information, Communication & Society*, *13*(3), 325-349.
8. Steinberg, P. L., Wason, S., Stern, J. M., Deters, L., Kowal, B., & Seigne, J. (2010). YouTube as source of prostate cancer information. *Urology*, *75*(3), 619-622.
9. Lewis, S. P., Heath, N. L., St Denis, J. M., & Noble, R. (2011). The scope of nonsuicidal self-injury on YouTube. *Pediatrics*, *127*(3), e552-e557.
10. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y. Y., & Moon, S. (2009). Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking (TON)*, *17*(5), 1357-1370.
11. Sureka, A. (2011). Mining user comment activity for detecting forum spammers in YouTube. *arXiv preprint arXiv:1103.5044*.
12. O'Challaghan, D., Harrigan, M., Carthy, J., Cunningham, P., (2012), Network Analysis of Recurring YouTube Spam Campaigns. *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media,* 531-534
13. Walther, J. B., & Parks, M. R. (2002). Cues filtered out, cues filtered in. *Handbook of interpersonal communication*, *3*, 529-563.
14. Derks, D., Bos, A. E., & Von Grumbkow, J. (2008). Emoticons and online message interpretation. *Social Science Computer Review*, *26*(3), 379-388.
15. Denecke, K. (2008, April). Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on* (pp. 507-512). IEEE.
16. Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79-86). Association for Computational Linguistics.
17. Thomas, M., Pang, B., & Lee, L. (2006, July). 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* (pp. 327-335). Association for Computational Linguistics.
18. Mitchell, J., Trending YouTube Video Statistics and Comments, https://www.kaggle.com/datasnaek/youtube/data, 2017.