Predicting Tiktok Virality

Tanner Durant and Kasirha Goodman

A picture containing person, holding

Description automatically generated

(screenshot of a longitudinal deep learning analysis   
of the six “Ekman emotions” as expressed in the   
face of one popular TikTok influencer,  
Noah Beck, during a TikTok video)   
   
Source: the present researchers’   
Python TensorFow scan of  
(Put a Sock In It, 2022).

**Introduction**

TikTok has rapidly become one of the most popular social media platforms in recent years, with over a billion active users worldwide. The platform allows users to create and share short-form videos, often accompanied by popular music, and has created a new form of viral content that has quickly taken the internet by storm. In particular, TikTok has been a breeding ground for viral videos, where users create and share content that spreads rapidly across the platform and beyond. These viral TikTok videos have become a cultural phenomenon, capturing the attention of millions of people around the world, and influencing popular culture in new and exciting ways. Using R, elements of viral videos are analyzed in order to potentially make predictions about what may go viral in the future.

*Literature Review*

In an influential recent white paper, IBM introduced the concept of the “4 V’s of Big Data”—a framework for understanding the character of information technology pursuits. The Internet, IBM noted, generates a high *volume* of data that continues to expand in size at a high *velocity* and that appears in a complicated *variety* of both structured and unstructured data formats. These dyssynchronous forms of information must be organized in such a way as to verify the *veracity* of the data obtained and to be able to make true inferences about the world. Therefore, the message of the IBM paradigm fundamentally is this: we live in a brand new world—the 4 V’s of Big Data mean that 21st century research inquiries are categorically different from those of the 1990s and of any decade that came before (Perry, 2017).

Writing during the mid-2000s, in the early days of Big Data’s rise to ascendancy, a set of Canadian researchers discovered another possible meaning of the second V: *variety*. The researchers observed that certain kinds of inquiries would blend and frustrate traditional disciplinary divisions within the university, such that—more than ever before—STEM-oriented and humanities-oriented practitioners would find themselves working together in interdisciplinary teams that blended technology and the humanities in new ways. They noted that:

[t]hose who work with electronic systems need to be constantly alert to counter the deficit of creativity which will occur if they become passive in relation to their tools. Some of the more promising new electronic undertakings marry new media with relatively new approaches to disciplinary or cross-disciplinary knowledge, such as poststructuralism and feminism, taking risks with or departing from established methodologies in an attempt to represent scholarship in new ways (Brown et al., 2006).

In the decade and a half since these Canadians published their inquiry, the intellectual space in which they stood has grown from a nameless, tenuous, experimental theory into a common, popular buzzword—the *digital humanities* (Durant, 2020). Yet even amid rapid growth in both the technical tools available to STEM researchers and the lexicon available to humanities-oriented theorists of the digital age, some of the same fundamental questions remain almost as unanswered today as they did during the early 2000s.

What kinds of research questions are best suited for the extent to which new interdisciplinary and team-based research shifts the tenor and tone of knowledge production? Are we moving away from the 20th century, from the era of the atomistic, individual scholar striving to publish a book- or essay-form monograph for hierarchical accolades? Are we yielding toward 21st-century models of team-based achievement and collaborative identity, where information begins to be shared, accessed, and reified in democratic, accessible, global knowledge- networks? What styles of data cleaning, data management, and data transformation are less tied to 20th century understandings of the individual researcher’s identity, habitus, and place within the jockeying politics of academia? Is an academic research model that derives, ultimately, from the medieval university—or from Cartesian or other pre-modern conceptions of time and space—robust enough to keep up with a world that has adopted quantum and post-quantum understandings of time and space? Is a complete shift away from 20th century organizational behavior models appropriate, as the IBM white paper seems to imply, or to what extent do the academy’s classic traditions remain relevant for charting a course of continuity—IBM statements about a brand new world notwithstanding?

Both the overt successes and the overcome setbacks of the present research project situate it as a valuable artefact in the ongoing story of academia’s journey to upgrade in form and function to match the complexity of a datasphere defined by IBM’s 4 V’s: *volume*, *velocity*, *variety*, and *veracity*. Though not perfect its attempt to find answers to the questions that Brown et al. introduced at the start of the millennium, the present research artefact holds extra value, perhaps, because it occurs in the final weeks of a fierce American political debate about banning TikTok (C-SPAN, 2023).

Will people retain the freedom to explore questions of selfhood, appearance, and meaning in an unregulated free space of technological options (Bhandari, 2020; Manovich et al., 2017; Tifentale & Manovich, 2015; Tifentale & Manovich, 2018)? Or, in the event that TikTok does get banned in the United States, TikTok itself has created a research API uniquely for American researchers, designed to preserve access to the TikTok archive even after a ban (Durant & Surabhi, 2023). What kinds of modeling and guidance can future researchers glean from the present researchers—perhaps one of the last academic teams of any kind to perform research on TikTok during the actual, immediate era when new American TikTok content production was allowed (Durant & Dunlap, 2023; Jafarian & Park, 2021; Kohn, 2022; Surabhi et al., 2022)?

*Guiding Questions*

This experiment seeks to answer several questions about viral Tiktok videos. What qualifies as a viral video? What are the characteristics of these videos? Are there qualities that all viral videos share that a creator can use to make a new viral video? Using data scraped from Tiktok, R is used to analyze the metadata of a set of videos from 2021 and 2022. The analysis will either confirm or raise doubt regarding services like Modash.io that claim to be able to tell influencer the recipe for viral success.

**Analysis and Models**

*About the Data*

Apify was used to scrape metadata from Tiktok on several thousand videos for this experiment. 981 total attributes were gathered for each instance (video) and compiled into an excel document. A list of the collected attributes is provided below.

|  |  |
| --- | --- |
| authorMeta/avatar | authorMeta/bioLink |
| authorMeta/digg | authorMeta/fans |
| authorMeta/following | authorMeta/heart |
| authorMeta/id | authorMeta/name |
| authorMeta/nickName | authorMeta/privateAccount |
| authorMeta/roomId | authorMeta/signature |
| authorMeta/ttSeller | authorMeta/verified |
| authorMeta/video | commentCount |
| createTime | createTimeISO |
| diggCount | effectStickers/name/useCount/ID (0-5) |
| hashtags cover/id/name/title (0-216) | id |
| isAd | isMuted |
| locationCreated | mediaUrls |
| mentions (0-51) | musicMeta/coverMediumUrl |
| musicMeta/musicAlbum | musicMeta/musicAuthor |
| musicMeta/musicId | musicMeta/musicName |
| musicMeta/musicOriginal | musicMeta/playUrl |
| playCount | searchHashtag/name |
| searchHashtag/views | shareCount |
| text | videoMeta/coverUrl |
| videoMeta/definition | videoMeta/duration |
| videoMeta/format | videoMeta/height |
| videoMeta/width | WebVideoUrl |

*Experimental Design*

The experiment is divided into three parts: data acquisition, data preprocessing, and data analysis. Prior to data acquisition, hashtags and creators were chosen as targets based on online sources. Evergreen hashtags are those that are consistently popular on the platform.

The first set of data scrapes videos associated with 24 hashtags in 3 runs on the Apify tool. The targeted hashtags are listed below.

|  |  |  |
| --- | --- | --- |
| Evergreen | Top of 2021 | Top of 2022 |
| #TikTok | #wfh | #picasso |
| #fyp | #workfromhome | #renaissanceeyes |
| #foryoupage | #rentfree | #mymoneydontjiggle |
| #viral | #tellmewithouttellingme | #rotoscope |
| #funny | #sheeesh | #iwannagohome |
| #music | #cratechallenge | #horace |
| #fashion | #adultswim | #runningupthathill |
| #follow | #silhouette | #teenagedirtbag |

After the three Excel files were downloaded, each contained around 980 attributes and up to 8000 instances. For the experiment, only the data regarding the single outcome and the chosen predictors needed to be conserved. The chosen outcome for the experiment is playCount—the number of times that a video has been viewed. 23 predictor variables were chosen, separated into two categories:

|  |  |
| --- | --- |
| Author Metadata | Video Metadata |
| id | duration |
| name | date of creation |
| verified status | hashtags 0-10 |
| number of fans | diggcount (number of likes) |
| hearts (number of likes on the profile) | number of shares |
| number of videos made by the profile | number of comments |

All columns besides those containing the listed attributes were deleted from the files.

The next steps of data preprocessing were done in R. Date of creation data was converted from characters into readable dates, and all videos made in years other than 2021 and 2022 were removed from the data set. Next, every attribute was discretized according to the table below.

|  |  |  |
| --- | --- | --- |
| *Attribute* | *Class* | *How to discretize* |
| authorMeta/fans | Num | Equal intervals |
| authorMeta/heart | Num | Equal intervals |
| authorMeta/id | Num | factor with levels |
| authorMeta/name | Character | factor with levels |
| authorMeta/verified | Logical |  |
| authorMeta/video | Num | Equal intervals |
| commentCount | Num | Equal intervals |
| createTimeISO | Date | factor with levels |
| diggCount | Num | Equal intervals |
| hashtags/0/name | Character | factor with levels |
| hashtags/1/name | Character | factor with levels |
| hashtags/2/name | Character | factor with levels |
| hashtags/3/name | Character | factor with levels |
| hashtags/4/name | Character | factor with levels |
| hashtags/5/name | Character | factor with levels |
| hashtags/6/name | Character | factor with levels |
| hashtags/7/name | Character | factor with levels |
| hashtags/8/name | Character | factor with levels |
| hashtags/9/name | Character | factor with levels |
| hashtags/10/name | Character | factor with levels |
| shareCount | Num | Equal intervals |
| videoMeta/duration | Num | Equal intervals |
| **playCount** | **Num** | **Binary factor** |

The outcome playCount was changed into a binary variable with all high performers (playCounts above the 3rd quartile of the data) classified as “viral” and all others classified as “not viral”. In the end, all the atrributes were logical or factors and could be inputted into the chosen algorithm.

After preprocessing, each set of data was analyzed using R. Association rules were mined from the data. Minimum support was set to 0.2 and minimum confidence to 0.6. The right hand side target was set to playCount. Due to technical issues with this version of Java and R, the J48 decision trees were not able to be run on the data.

*Complete R Code*

*#Apify for #tiktok*

*#Load data*

*getwd()*

*setwd(dir = "/Users/katgood/desktop")*

*library(readxl)*

*tt1x <- read\_excel("22hash.xlsx")*

*str(tt1x)*

*#Change time to readable date*

*tt1x$createTimeISO <- as.Date(tt1x$createTimeISO)*

*#What dates is this data from?*

*range(tt1x$createTimeISO)*

*#Let's delete anything pre 2021 or post 2022*

*install.packages("dplyr")*

*library("dplyr")*

*install.packages("data.table")*

*library("data.table")*

*arrange(tt1x,tt1x$createTimeISO)*

*tt1x<-tt1x[- grep("2016", tt1x$createTimeISO),]*

*tt1x<-tt1x[- grep("2017", tt1x$createTimeISO),]*

*tt1x<-tt1x[- grep("2018", tt1x$createTimeISO),]*

*tt1x<-tt1x[- grep("2019", tt1x$createTimeISO),]*

*tt1x<-tt1x[- grep("2020", tt1x$createTimeISO),]*

*tt1x<-tt1x[- grep("2023", tt1x$createTimeISO),]*

*#Let's delete any other dates that are unneccessary*

*tt1x<-tt1x[- grep("2021", tt1x$createTimeISO),] #for top 2022*

*tt1x<-tt1x[- grep("2021", tt1x$createTimeISO),] #for top 2021*

*#check that we have the correct data only*

*range(tt1x$createTimeISO)*

*#Visualize playCount*

*hist(tt1x$playCount, main="# Playcount 2021-2022", xlab="play count", breaks=10)*

*summary(tt1x)*

*boxplot(tt1x$playCount)*

*#Discretize predictors*

*install.packages("arules")*

*install.packages("arulesViz")*

*library(arules)*

*library(arulesViz)*

*#Change verified to logical*

*tt1x$`authorMeta/verified`<-as.logical(tt1x$`authorMeta/verified`)*

*#Change all predictors to factors*

*tt1x$`authorMeta/fans`<-discretize(tt1x$`authorMeta/fans`, method = "interval", breaks=10)*

*tt1x$`authorMeta/heart`<-discretize(tt1x$`authorMeta/heart`, method = "interval", breaks=10)*

*tt1x$`authorMeta/id`<-as.factor(tt1x$`authorMeta/id`)*

*tt1x$`authorMeta/name`<-as.factor(tt1x$`authorMeta/name`)*

*tt1x$`authorMeta/video`<-discretize(tt1x$`authorMeta/video`, method = "interval", breaks=10)*

*tt1x$commentCount<-discretize(tt1x$commentCount, method = "interval", breaks=10)*

*tt1x$createTimeISO<-as.factor(tt1x$createTimeISO)*

*tt1x$diggCount<-discretize(tt1x$diggCount, method = "interval", breaks=10)*

*tt1x$`hashtags/0/name`<-as.factor(tt1x$`hashtags/0/name`)*

*tt1x$`hashtags/1/name`<-as.factor(tt1x$`hashtags/1/name`)*

*tt1x$`hashtags/2/name`<-as.factor(tt1x$`hashtags/2/name`)*

*tt1x$`hashtags/3/name`<-as.factor(tt1x$`hashtags/3/name`)*

*tt1x$`hashtags/4/name`<-as.factor(tt1x$`hashtags/4/name`)*

*tt1x$`hashtags/5/name`<-as.factor(tt1x$`hashtags/5/name`)*

*tt1x$`hashtags/6/name`<-as.factor(tt1x$`hashtags/6/name`)*

*tt1x$`hashtags/7/name`<-as.factor(tt1x$`hashtags/7/name`)*

*tt1x$`hashtags/8/name`<-as.factor(tt1x$`hashtags/8/name`)*

*tt1x$`hashtags/9/name`<-as.factor(tt1x$`hashtags/9/name`)*

*tt1x$`hashtags/10/name`<-as.factor(tt1x$`hashtags/10/name`)*

*tt1x$shareCount<-discretize(tt1x$shareCount, method = "interval", breaks=10)*

*tt1x$`videoMeta/duration`<-discretize(tt1x$`videoMeta/duration`, method = "interval", breaks=10)*

*#Find cutoff for viral/not viral playCount outcome*

*quantile(tt1x$playCount)*

*#discretize playCount outcome using breaks=c(min,3rd quartile, max)*

*tt1x$playCount <- cut(tt1x$playCount, breaks=c(0,547875,71500000), labels=c('not\_viral', 'viral'))*

*#check to make sure everything is logical or factor format*

*str(tt1x)*

*#Association Rule Mining*

*rules <- apriori(data=tt1x,parameter = list(minlen=2,supp=0.2,conf=0.6),appearance = list(rhs = c("playCount=not\_viral","playCount=viral"),default="lhs"), control = list(verbose=F))*

*summary(rules)*

*rules<-sort(rules, decreasing=TRUE,by="confidence")*

*inspect(rules)*

*#Decision Tree*

*#Not functional on my computer due to issues locating Java*

*install.packages("rJava")*

*library("rJava")*

*install.packages("RWeka")*

*library("RWeka")*

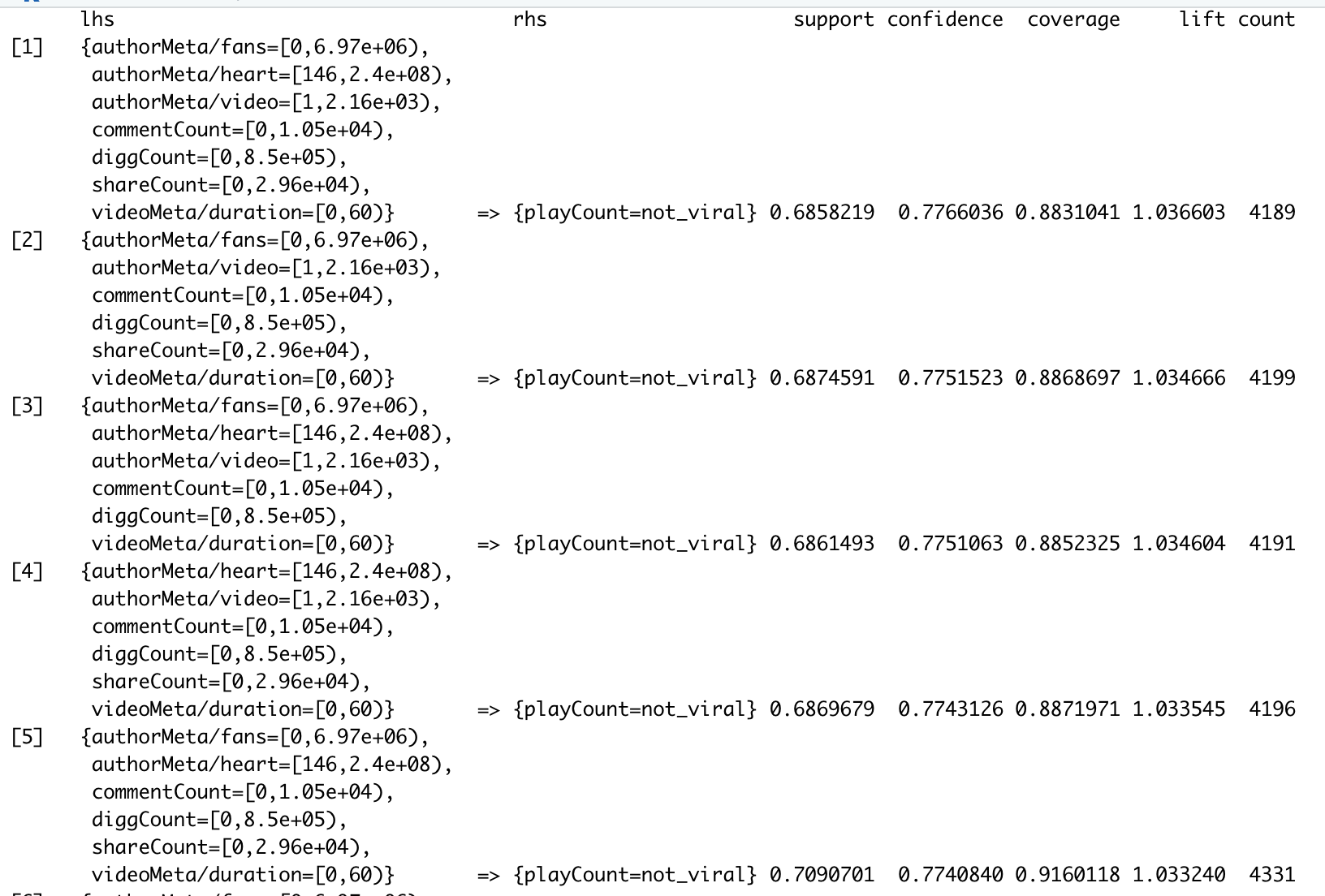
*install.packages("party")*

*library("party")*

*tree<-J48(playCount~., data=tt1x)*

**Results**

The first dataset is an analysis of 8 evergreen hashtags over the period January 2021 to December 2022. 175 rules were produced. The resulting association rules did not produce any meaningful predictions regarding virality. The second dataset is an analysis of 8 top hashtags from 2021 over the period of January 2021 to December 2021. 127 rules were produced. None produced meaningful predictions of virality. The third dataset is an analysis of 8 top hashtags from 2022 over the period of January 2022 to December 2022. 127 rules were produced. None produced meaningful predictions of virality. Below is a selection of some of the rules with the highest confidence:



Every rule produced by the algorithm could only show how to produce non-viral videos. While this is an unfortunate result, it confirms the conventional wisdom regarding viral videos—there is no recipe for success, virality comes at random.

**Conclusions**

In this experiment, a viral video is defined to be a video with a playcount above the normal range of other videos from the same time frame and using similar hashtags. While some viral videos have superficial similarities—popular hashtags, trending sounds, relatable humor, famous creators—there is no evidence to suggest that any one of these characteristics is a predictor for guaranteed success. Additional factors such as timing, audience, and algorithmic favoritism also play a significant role in determining the success of a TikTok video. Despite the lack of a surefire recipe, creativity and authenticity are essential for standing out on the platform and increasing the likelihood of going viral. Moreover, the impact of viral TikTok videos extends beyond the platform, often influencing popular culture and shaping social discourse. Therefore, understanding the characteristics and factors involved can be advantageous for creators striving to produce engaging and impactful content on the platform.

**References**

Bruney, G. (2021, December 31). 15 of the Best TikTok Moments That Defined 2021. Elite Daily. <https://www.elitedaily.com/lifestyle/best-tiktok-moments-2022-viral>

Bhandari, A., & Bimo, S. (2020). Tiktok and the "Algorithmized Self": A New Model of Online Interaction. *AoIR Selected Papers of Internet Research*. <https://doi.org/10.5210/spir.v2020i0.11172>.

Brown, S., Clements, P., & Grundy, I. (2006, May/June). Sorting things in: Feminist knowledge representation and changing modes of scholarly production. Retrieved March 24, 2023. <https://doi.org/10.1016/j.wsif.2006.04.010>.

C-SPAN. (2023). *TikTok Ceo Shou Chew Testifies Before Congress*. *YouTube*. CSPAN. Retrieved March 24, 2023, from <https://www.youtube.com/watch?v=_E-4jtTFsO4>.

D’Amelio, C. (2023. *Wanted to see if I remembered this one dc @chuckygonewild @nickanthonyy*. *TikTok*. Retrieved March 24, 2023, from <https://www.tiktok.com/t/ZTRvAa8BP/>.

Donkor, N. (2021, December 1). The Most Followed TikTok Accounts of 2021. We Got This Covered. <https://wegotthiscovered.com/videos/the-most-followed-tiktok-accounts-of-2021/>

Dunlap, K. (2023, March 14). *TikTokFER*.

Durant, T. (né Ekpenyong, A.). (2021). Digital Humanities Scholarship: A Model for Reimagining Knowledge Work in the 21st Century. In: *Diversity, Divergence, Dialogue: 16th International Conference, iConference 2021, Beijing, China, March 17–31, 2021, Proceedings, Part I* 16 (pp. 435-445). Springer International Publishing. Retrieved February 1, 2023, from <https://link.springer.com/chapter/10.1007/978-3-030-71292-1_33>.

Durant, T., & Dunlap, K. (2023, March 14). Stanford TikTokFER Code Implementation, personal.

Durant, T., & Pekowski, L. (2023). Stanford TikTokFER Code Implementation and Cluster-Based Cloud Computing Solutions. personal.

Durant, T., & Surabhi, K. (2023, March 14). Stanford TikTokFER Code Implementation, personal.

Favikon. (2022, February 9). The 20 Most Famous TikTok Influencers in the World. <https://www.favikon.com/blog/the-20-most-famous-tiktok-influencers-in-the-world#TOP8km>

Geyser, W. (2023, February 14). TikTok Stats that Matter for Your Business in 2022. Influencer Marketing Hub. <https://influencermarketinghub.com/tiktok-stats/>

Global Grind. (2021, December 28). Viral Videos: A List of 2021's Most Popular TikTok Trends. <https://globalgrind.com/playlist/viral-videos-a-list-of-2021s-most-popular-tiktok-trends/item/3>

Hughes, D. (2021, August 3). How to Go Viral on TikTok: 7 Essential Tips for 2021. Digital Marketing Institute. <https://digitalmarketinginstitute.com/blog/how-to-go-viral-on-tik-tok>

Jafarian, Y., & Park, H. S. (2021). *Learning High Fidelity Depths of Dressed Humans by Watching Social Media Dance Videos*. Computer Vision Foundation. Retrieved March 24, 2023, from <https://openaccess.thecvf.com/content/CVPR2021/papers/Jafarian_Learning_High_Fidelity_Depths_of_Dressed_Humans_by_Watching_Social_CVPR_2021_paper.pdf>.

Kaplanis, K. (Host.) (2022). TikTok Creator Rant and Understanding Your Worth as a Content Creator with Nimay Ndolo. [Audio podcast episode.] In *BizTok for*  *TikTok*. Retrieved March 24, 2023, from <https://open.spotify.com/episode/3CBf2rDLWrgufteTcPb6lj?si=ea3ab28921c04600>.

Kohn, M. (2022). Clearview AI, TikTok, and the Collection of Facial Images in International Law. *Chicago Journal of International Law*, *23*(1). Retrieved March 24, 2023, from <https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1832&context=cjil>.

Malik. (2023). *2021 Or 2023 prefer or Charli?* *TikTok*. Retrieved March 24, 2023, from <https://www.tiktok.com/t/ZTRvAS6td/>.

Manovich, L., Ferrari, V., & Bruno, N. (2017). Selfie-takers prefer left cheeks: Converging evidence from the (extended) selfiecity database. *Frontiers in Psychology*, *8*. <https://doi.org/10.3389/fpsyg.2017.01460>.

Napierala, B. (2021, August 10). TikTok: The Metadata Machine. <https://www.linkedin.com/pulse/tiktok-metadata-machine-brian-napierala/?trk=articles_directory>

Paul, B. (2022, October 3). Best TikTok Hashtags to Get More Views. We Got This Covered. <https://wegotthiscovered.com/social-media/best-tiktok-hashtags-to-get-more-views/>

Perry, J. S. (2017, May 22). *What Is Big Data? More than volume, velocity, and variety.* IBM Developer Blog. Retrieved March 24, 2023, from <https://developer.ibm.com/blogs/what-is-big-data-more-than-volume-velocity-and-variety/>.

Put a Sock in It Podcast. (2022). *Larri is the master of games*. *TikTok.* Put A Sock In It Podcast. Retrieved March 24, 2023, from <https://www.tiktok.com/t/ZTRneaQN8/>.

Surabhi, S. (2022, April 22). TikTok\_FER. GitHub. Retrieved March 24, 2023, from <https://github.com/walllab/tiktok_FER>.

Surabhi, S. (2023, March 14). *TikTokFER*

Surabhi, S., Shah, B., Washington, P., Mutlu, O. C., Leblanc, E., Mohite, P., Husic, A., Kline, A., Dunlap, K., McNealis, M., Liu, B., Deveaux, N., Sleiman, E., & Wall, D. P. (2022). *Tiktok for Good: Creating a Diverse Emotion Expression Database*. CVF Open Access. Retrieved March 24, 2023, from <https://openaccess.thecvf.com/content/CVPR2022W/ABAW/html/Surabhi_TikTok_for_Good_Creating_a_Diverse_Emotion_Expression_Database_CVPRW_2022_paper.html>.

Tifentale, A., Manovich, L. (2015). Selfiecity: Exploring Photography and Self-Fashioning in Social Media. In: Berry, D.M., Dieter, M. (eds) *Postdigital Aesthetics.* Palgrave Macmillan, London. <https://doi.org/10.1057/9781137437204_9>.

Tifentale, A., & Manovich, L. (2018). *Competitive Photography and the Presentation of the Self (unedited working draft).* Manovich.net. Retrieved March 24, 2023, from <http://manovich.net/index.php/projects/competitive-photography-and-the-presentation-of-the-self>.

TikTok Ads. (n.d.). Popular Hashtags. <https://ads.tiktok.com/business/creativecenter/inspiration/popular/hashtag/pc/en>

TikTok Hashtags. (n.d.). The Best TikTok Hashtags. <https://tiktokhashtags.com/>

**Special Acknowledgement**

The researchers would like to thank the Syracuse University HTC Grid and NSF award ACI-1341006 for making possible some of the computing resources and guidance that were used in the development of this research.