Evaluating Emotion and Morality Bias in YouTube's Recommendation Algorithm for the China-Uyghur Crisis

Obianuju Okeke, Mert Can Cakmak, Ugochukwu Onyepunuka, Billy Spann, and Nitin Agarwal

University of Arkansas at Little Rock, Little Rock, Arkansas, USA {oiokeke, mccakmak, uponyepunuka, bxspann, nxagarwal}@ualr.edu

Abstract. This study introduces a drift analysis methodology to explore patterns in YouTube's recommendation algorithm concerning the China- Uyghur crisis. Recognizing the influence of this conflict on a global discourse, we inspect the bias within YouTube's algorithm that can potentially affect information propagation about this crisis. Utilizing a dataset gathered from multiple layers of video recommendations, we apply emotion analysis and morality assessment to examine the algorithm's behavior. Our findings from the emotion analysis indicate a trend towards more positive emotions and a decline in negative emotions as users progress through recommended videos. Moreover, our morality assessment, based on the Moral Foundations Theory, indicates a decrease in moral vices and an emergent preference for certain moral virtues in the recommended content. This comprehensive analysis offers insights into how YouTube's recommendation algorithm might influence the perception of highly polarizing global issues, such as the China-Uyghur crisis.

Keywords: YouTube's recommendation algorithm · Drift analysis · Emotion analysis · Morality assessment · Bias in recommender systems.

1 Introduction

Recommendation algorithms are frequently associated with biases such as selection bias [15], position bias [13, 8], and popularity bias [16, 17]. Popular recommendation platforms have, in the past, been associated with patterns that lead users to highly homogeneous content, resulting in certain phenomena such as filter bubbles and echo chambers [12]. In such scenarios, users are isolated from diverse content and are instead exposed to a narrower band of information. This can pose the risk of reinforcing specific viewpoints [19, 18]. This study examines the emotion and morality bias present in YouTube's recommendation algorithm. By analyzing the evolution of emotion and morality across recommended YouTube videos related to the China-Uyghur crisis narrative, we aim to determine if YouTube's recommendation algorithm favors videos with certain emotions over others and to explore the distribution of moral content across YouTube's recommendation algorithm.

1.1 Background of Study

In this section, we discuss previous research related to our study, which includes previous works on morality assessment, emotion detection [1, 27], and bias analysis in recommender systems [26]. Recommendation bias has been researched extensively to understand its structure and effects, especially in the areas of radicalization and the spread of misinformation and disinformation [10]. These past works have studied the emergence of homophilic communities within content, such as recommended videos, and they have studied the factors leading to the emergence of these communities. Further investigations have revealed patterns of coordinated activity among YouTube commenters, potentially impacting engagement levels on certain videos [28] Insights from these studies have been crucial in identifying the emergence of homogeneity, the development of interconnected communities, and the potential bias in recommender systems. Drift is a technique that many researchers have used in studying how content evolves. By studying content evolution, we determine if the content remains the same or changes relative to a standard metric. O' Hare et al. [4] analyzed a sentiment-annotated corpus of textual data to determine topic drift among documents. Liu et al. [3] developed an LDA (Latent Dirichlet Allocation)-based method for topic drift detection in micro-blog posts. Akilaet al. developed a framework to identify the mood of the nation of India by analyzing real-time Twitter posts [21]. Results showed the trends of emotions as they evolved across the country within a selected date range (4 April 2020 to 4 May 2020). These trends were visualized using line graphs and radar maps. The authors' goal with this research was to understand the region's drift of emotions with respect to the number of Covid-19 cases reported in the region. In this research, we apply drift analysis techniques to assess emotion and morality, to determine the pattern of bias in YouTube's recommendation algorithm. By combining both emotion and morality assessment, we take a more holistic approach to understand the nature and impact on videos recommendations by YouTube's recommendation algorithm.

1.2 The China-Uyghur Crisis

The China-Uyghur crisis has garnered significant criticism from various organizations across the globe. According to the Council on Foreign Relations[9], more than a million Uyghurs, a Muslim, Uyghur-speaking Asian ethnic group have been detained in China's Xinjiang region since 2017. The United States and the UN Human Rights Office have described these acts as crimes against humanity. Despite reports from international journalists and researchers about the ongoing systems of mass detention throughout the region, which have been backed by satellite images and leaked Chinese government documents, individual testimonies from Chinese officials insist that the rights of Uyghur Muslims have not been violated. They assert that government crackdown measures, such as re-education camps, have been discontinued since 2019 [5]. According to Silverman, content-evoking polarization is propagated faster than non-polarizing

content [11]. In effect, since the emotions attached to our seed videos are negative, we also expect to see more negative emotions propagated through videos across recommendation depths. We will also be working towards exploring the moral content of these potentially negative emotions to study the moral nature of content distributed by the recommendation algorithm.

2 Methodology

For this research, we introduce a drift analysis methodology that allows us to monitor changes in video characteristics and explore the patterns of the recommendation algorithm. We apply the resulting approach to our dataset, which consists of a collection of videos recommended through various methodologies. To analyze the drift across recommendation depths, we calculated and compared the drift in emotion and morality within our target dataset.

2.1 Collection of Data

Video recommendations on YouTube are heavily influenced by the user's watch history, meaning that the algorithm personalizes the videos recommended to a user. To eliminate this personalization bias and control our experiment, we employed the following precautionary steps:

1. Video collection script prevented account login for each watch session.
2. A new browser instance was started for each level or depth of recommendation
3. Cookies from each previous recommendation depth were cleared to enable a fresh search for videos at the next depth of recommendation.

The data used in this research was composed of YouTube's **'watch-next'** videos which are found in the watch-next panel of the platform. These videos were collected using techniques employed in [10]. To begin our data collection process, we first conducted a series of workshops with subject matter experts to generate a list of relevant keywords related to the China-Uyghur conflict. These keywords were used as search queries on YouTube's search engine to generate the 40 seed videos. Video recommendations were gathered for each seed video using custom-made crawlers over levels or "depths" of recommended videos. Each of the 40 seed videos generated a first depth of recommendations, with subsequent depths serving as parent videos for generating further layers of recommended videos. This process continued until recommendations for four depths were generated, resulting in a total of 15,307 unique videos. In our data collection process, we extracted the video titles and video descriptions. For this research we chose to analyze the videos based on video text data (titles, description), as these levels of detail will provide information on the content of the videos. In future research we are further exploring the content of the video using transcipts from the videos as well as audio data to provide a more detailed level of analysis. The dataset was split by the depth and analyzed to study if and how video characteristics such as

emotion and morality changed (drifted) as the algorithm recommended videos to the user. Next, we describe the methodologies used to address the research questions posed in this study.

2.2 Emotion Analysis

We analyzed the emotions embedded in the video text data (title and description), focusing on seven emotions: anger, disgust, fear, joy, neutral, sadness and surprise. We used emotion drift to identify emotional bias across various depths of recommendations. The diversity of emotions resulting from the content was illustrated on a line graph, with each point on the depth axis representing a traversed depth of video recommendations. We utilized a fine-tuned version of transfer learning [6], Emotion-English-DistilRoberta-base [20], for Natural Language Processing (NLP) tasks to ensure the accuracy of results. Transfer learning aims to increase the accuracy and efficiency of the model training process by preserving information from prior models and applying it to related tasks.

2.3 Morality Assessment

Morality assessment is an effective social computing technique used to extract moral intuition from textual data. Using the Moral Foundations Theory [14], we worked on analyzing the moral content of our narrative and visualized the drift in morality within our dataset across five moral foundations; Care/harm (involving intuitions of sympathy, compassion, and nurturance), Fairness/cheating (including notions of rights and justice), Loyalty/betrayal (supporting moral obligations of patriotism and "us versus them" thinking), Authority/subversion (including concerns about traditions and maintaining social order), and Sanctity/degradation (including moral disgust and spiritual concerns related to the body) [7]. The resulting morality diversity in content was also illustrated on a line graph, with each depth representing a traversed depth of video recommendations.

3 Results

3.1 Emotion Analysis

The goal of emotion drift analysis is to determine how emotions across the seven categories (anger, surprise, fear, joy, neutral, disgust, and sadness) evolve or drift across recommendation depths. To determine emotion drift, we computed and visualized the predominant emotions at each depth of recommendation, from seed to depth 4, on a line graph. In our analysis, we considered video text data at two different levels: video titles and video descriptions. This process allowed us to effectively apply emotion analysis and visualize emotion drift across different levels of video details. The neutral emotion effectively isolates text data with no identifiable emotion embedded. As a result, the neutral emotion in the

line graph was not considered in our result analysis. From the graph below Figure 1(a), we observe that on emotion drift analysis of the video titles, there was a significant presence of fear, anger, and disgust emotion at the seed level (depth 0), and a reduced presence of joy and surprise emotion. As we move from seed videos to depth 4 through the recommendations made by the algorithm, we observe an increase in the proportion of joy and surprise emotions. This trend is accompanied by a decrease in the previously heightened fear, anger, and disgust emotion as we approach depth 4. This trend of the emergence of positive emotions and decline of negative emotions across recommendations is also seen in the emotion drift analysis of video descriptions in Figure 1(b) but with a clear level of distinction. On analyzing video descriptions, we see a more distinct pattern of emergence and decline, this is most likely to the higher level of content and information in video descriptions as compared to video titles.

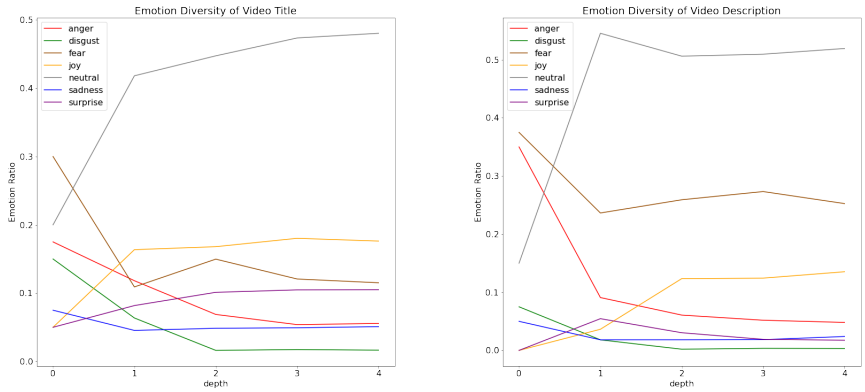

Fig. 1. Line graph showing the distribution of emotions across recommendations of videos using (a) video titles (b) video descriptions

3.2 Morality Assessment

While emotion analysis is a text analytics tool which has been used to determine the emotional expression in text data [22–24], morality assessment is another form of text analytic technique which has been used to discover ethical reasoning, moral concepts and decision-making in text data [14]. With the application of morality assessment, we can extract the underlying moral content in text data. This will enable us to understand the influence of moral judgments, especially in the areas of information dissemination. For our morality assessment analysis, the extended Moral Foundations Dictionary (eMFD), a dictionary-based tool for extracting moral content from textual corpora was used. This package was constructed from text annotations generated by a large sample of human coders

[25]. The morality drift analysis aimed to gain a more comprehensive understanding of the drift in embedded moral opinions across recommendations and to comprehend the moral implications of the bias in YouTube's recommendation algorithm. As was done in the emotion drift analysis phase, the predominant morality at each depth of recommendation from seed to depth four was computed and visualized on a line graph. Given the comprehensive nature of the morality assessment phase, the analysis was conducted in two parts across video titles and descriptions. These parts included: a Morality Virtue Assessment, which calculated the distribution of moral virtues such as care, fairness, loyalty, authority, and sanctity, and a Morality Vice Assessment, which computed the distribution of moral vices such as harm, cheating, betrayal, subversion, and degradation across recommendations.

Morality Virtue Assessment on Video Titles and Video Description The graph in Figure 2(a) illustrates the morality drift analysis of video titles. On analyzing our seed videos (depth 0), we observed a reduced expression of loyalty and sanctity and a heightened expression of fairness, authority, and calm. As we progressed through the recommendation depths, we observed a switch in the expression of virtues, where at depth 2, loyalty was the most expressed virtue, with a seemingly equal frequency occurrence of care and sanctity. Finally, at depth 4, all virtues are seen to plateau out at he same frequency level with no single virtue being significantly expressed. The morality drift analysis of video descriptions in Figure 2(b) shows that all virtues were expressed at the lowest frequency in the seed videos (depth 0), with sanctity being the least expressed. As we moved through the recommendations, we noticed a steady increase from depth 1 to depth 4. At depth 4, loyalty was the most frequently expressed morality, while sanctity was the least expressed.

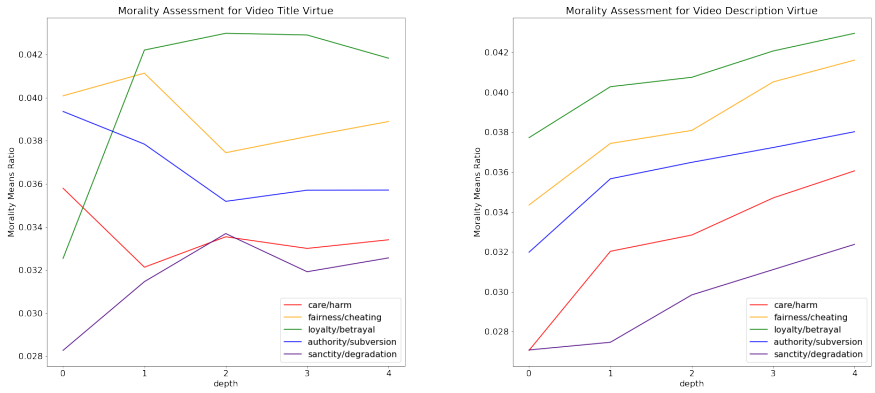

Fig. 2. Line graph showing the distribution of virtue morality across recommendations of videos using (a) video titles (b) video descriptions

Morality Vice Assessment on Video Titles and Description From Figure 2(a), which illustrates the morality drift analysis of video titles, we observe a significant presence of all vices (harm, cheating, betrayal, subversion, and degradation) at depth 0, representing our seed videos. Harm appears as the most prevalent moral vice. As we move from the seed videos to depth 4, there is a significant decline in all moral vices, with a plateau observed from depth 3 to 4. This declining trend in the presence of moral vices across recommendations is also evident in the video descriptions analysis depicted in Figure 2(b). However, the decline appears to occur at a slower pace compared to that in the video titles.

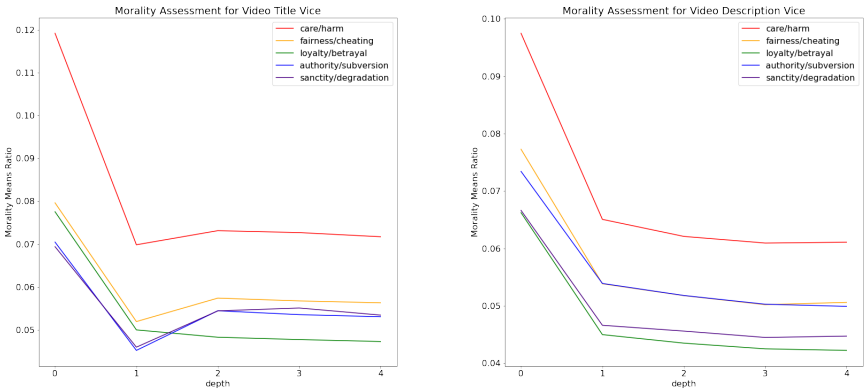

Fig. 3. Line graph showing the distribution of vice morality across recommendations of videos using (a) video titles (b) video descriptions

4 Discussion and Conclusion

In this research, we collected videos from four recommendation stages, using relevant seed videos related to the China-Uyghur crisis. On collection of our data, emotion analysis and morality assessment were conducted to determine the nature and impact of YouTube's recommendation algorithm as it relates to the China-Uyghur narrative. Upon conducting emotion analysis on both video titles and descriptions, we identified and isolated videos expressing neutral emotions to gain more effective insights. Results from our emotion analysis showed that there is an emergence of positive emotions and a decline of negative emotions as we progress through recommended videos. This emotion pattern drift suggests that as the algorithm encounters videos related to the China-Uyghur narrative, the algorithm reduces the recommendation of videos expressing a negative emotion while increasing recommendations of videos with positive emotions. On assessing the distribution of morality across recommendations, we see that within our seed videos, the least virtues expressed were loyalty and sanctity but as more videos were recommended, these moral virtues steadily increased over other virtues,

resulting in loyalty becoming the most expressed virtue at depth 4. Furthermore, morality vice assessment of video titles and description showed that all vices were significantly expressed in our seed videos, but as more videos were recommended by the algorithm, these vices continually reduced till they reached minimum levels at depth 4. Our finding suggest that our seed videos had high amounts of negative emotions and vices and reduced amounts of virtues, especially loyalty and sanctity morals. As more videos are recommended by the algorithm from the China-Uyghur seed videos, we see an increase in positive emotions and a decrease in moral vices. These recommendations also lead to an increase in the presence of loyalty and sanctity virtues. For this research, we employed the use of drift analysis to identify emotion and morality bias across recommended videos. Our results showed that YouTube's recommendation system tends to lessen negative emotions such as anger and amplify positive emotions such as joy across recommended videos on the platform. We also see the presence of moral vices being reduced and relevant virtues are promoted across recommendations as it relates to the China-Uyghur narrative. In our future works, we plan to expand this research into exploring an alternate narrative which could express an opposing mix of vice and virtues in the seed videos (reduced vice and increased virtue) to determine the role of a narrative in the propagation of bias across recommendations.

Acknowledgements This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS 1636933, ACI-1429160, and IIS-1110868), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), 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, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189, W911NF-23-1-0011), U.S. Defense Advanced Research Projects Agency (W31P4Q 17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). 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. S. N. Shivhare and S. Khethawat, "Emotion Detection from Text." arXiv, May 22, 2012. doi: 10.48550/arXiv.1205.4944.
2. J. M. Garcia-Garcia, V. M. R. Penichet, and M. D. Lozano, "Emotion detection: a technology review," in Proceedings of the XVIII International Conference on Human Computer Interaction, New York, NY, USA, Sep. 2017, pp. 1–8. doi: 10.1145/3123818.3123852.

3. Q. Liu, H. Huang, and C. Feng, "Micro-blog Post Topic Drift Detection Based on LDA Model," in Behavior and Social Computing, Cham, 2013, pp. 106–118
4. N. O'Hare et al., "Topic-dependent sentiment analysis of financial blogs," in Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, New York, NY, USA, Nov. 2009, pp. 9–16. doi: 10.1145/1651461.1651464.
5. M. Suhasini and S. Badugu, "Two Step Approach for Emotion Detection on Twitter Data," International Journal of Computer Applications, vol. 179, pp. 12–19, Jun. 2018, doi: 10.5120/ijca2018917350.
6. C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.
7. F. R. Hopp, J. T. Fisher, D. Cornell, R. Huskey, and R. Weber, "The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text," Behav Res, vol. 53, no. 1, pp. 232–246, Feb. 2021, doi: 10.3758/s13428-020-01433-0.
8. A. Agarwal, I. Zaitsev, X. Wang, C. Li, M. Najork, and T. Joachims, "Estimating Position Bias without Intrusive Interventions," in Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Jan. 2019, pp. 474–482. doi: 10.1145/3289600.3291017.
9. "China's Repression of Uyghurs in Xinjiang | Council on Foreign Relations." https://www.cfr.org/backgrounder/china-xinjiang-uyghurs-muslims-repression-genocide-human-rights [Accessed Jan. 09, 2023].
10. M. Faddoul, G. Chaslot, and H. Farid, "A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos." arXiv, Mar. 06, 2020. doi: 10.48550/arXiv.2003.03318.
11. C. Silverman, "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook," BuzzFeed News. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook [Accessed Jan. 09, 2023].
12. B. Kitchens, S. L. Johnson, and P. Gray, "Understanding Echo Chambers and Filter Bubbles: The Impact of Social Media on Diversification and Partisan Shifts in News Consumption," Aug. 26, 2020. https://misq.umn.edu/understanding-echo-chambers-and-filter-bubbles-the-impact-of-social-media-on-diversification-and-partisan-shifts-in-news-consumption.html [Accessed Jan. 09, 2023].
13. X. Wang, N. Golbandi, M. Bendersky, D. Metzler, and M. Najork, "Position Bias Estimation for Unbiased Learning to Rank in Personal Search," in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, in WSDM '18. New York, NY, USA: Association for Computing Machinery, Feb. 2018, pp. 610–618. doi: 10.1145/3159652.3159732.
14. J. Haidt, "The New Synthesis in Moral Psychology," Science, vol. 316, no. 5827, pp. 998–1002, May 2007, doi: 10.1126/science.1137651.
15. Z. Ovaisi, R. Ahsan, Y. Zhang, K. Vasilaky, and E. Zheleva, "Correcting for Selection Bias in Learning-to-rank Systems," in Proceedings of The Web Conference 2020, Apr. 2020, pp. 1863–1873. doi: 10.1145/3366423.3380255.
16. H. Abdollahpouri, R. Burke, and B. Mobasher, "Controlling Popularity Bias in Learning-to-Rank Recommendation," in Proceedings of the Eleventh ACM Conference on Recommender Systems, in RecSys '17. New York, NY, USA: Association for Computing Machinery, Aug. 2017, pp. 42–46. doi: 10.1145/3109859.3109912.
17. R. Cañamares and P. Castells, "Should I Follow the Crowd? A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems," in The 41st International ACM SIGIR Conference on Research and Development in Information

Retrieval, in SIGIR '18. New York, NY, USA: Association for Computing Machinery, Jun. 2018, pp. 415–424. doi: 10.1145/3209978.3210014.
18. A. J. B. Chaney, B. M. Stewart, and B. E. Engelhardt, "How algorithmic confounding in recommendation systems increases homogeneity and decreases utility," in Proceedings of the 12th ACM Conference on Recommender Systems, in RecSys '18. New York, NY, USA: Association for Computing Machinery, Sep. 2018, pp. 224–232. doi: 10.1145/3240323.3240370.
19. K. Hosanagar, D. Fleder, D. Lee, and A. Buja, "Will the Global Village Fracture Into Tribes? Recommender Systems and Their Effects on Consumer Fragmentation," Management Science, vol. 60, no. 4, pp. 805–823, Apr. 2014, doi: 10.1287/mnsc.2013.1808.
20. Hartmann, J. "Emotion English DistilRoBERTa-base". https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, (2022).
21. "Mood of India During Covid-19 - An Interactive Web Portal Based on Emotion Analysis of Twitter Data | Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing." https://dl.acm.org/doi/10.1145/3406865.3418567 (accessed Jun. 02, 2023).
22. Vo, Bao-Khanh Ho, and N. I. G. E. L. Collier. "Twitter emotion analysis in earthquake situations." Int. J. Comput. Linguistics Appl. 4.1 (2013): 159-173.
23. D. Xu, Z. Tian, R. Lai, X. Kong, Z. Tan, and W. Shi, "Deep learning based emotion analysis of microblog texts," Information Fusion, vol. 64, pp. 1–11, Dec. 2020, doi: 10.1016/j.inffus.2020.06.002.
24. A. Jamdar, J. Abraham, K. Khanna, and R. Dubey, "Emotion Analysis of Songs Based on Lyrical and Audio Features," IJAIA, vol. 6, no. 3, pp. 35–50, May 2015, doi: 10.5121/ijaia.2015.6304.
25. "GitHub - medianeuroscience/emfd: The Extended Moral Foundations Dictionary (E-MFD)." https://github.com/medianeuroscience/emfd (accessed Jun. 04, 2023).
26. Okeke, O.I., Cakmak, M.C., Spann, B., and Agarwal, N.: Examining Content and Emotion Bias in YouTube's Recommendation Algorithm. In the Ninth International Conference on Human and Social Analytics, Barcelona, Spain, (2023).
27. Banjo, D. S., Trimmingham, C., Yousefi, N., & Agarwal, N. Multimodal Characterization of Emotion within Multimedia Space. (2022)
28. Shajari, S., Agarwal, N., & Alassad, M. Commenter Behavior Characterization on YouTube Channels. arXiv preprint arXiv:2304.07681. (2023)