THEME ARTICLE: CIVILIZING AND HUMANIZING AI

Toward a Programmable Humanizing Artificial Intelligence Through Scalable Stance-Directed Architecture

Yusuf Mücahit Çetinkaya, *Middle East Technical University, Ankara, 06800, Türkiye, and Arizona State University, Tempe, AZ, 85281, USA*

Yeonjung Lee, *Arizona State University, Tempe, AZ, 85281, USA*

Emre Külah and İsmail Hakkı Toroslu, *Middle East Technical University, Ankara, 06800, Türkiye*

Michael A. Cowan, *Loyola University, New Orleans, LA, 70118, USA*

Hasan Davulcu, *Arizona State University, Tempe, AZ, 85281, USA*

The rise of harmful online content underscores the urgent need for artificial intelligence (AI) systems to effectively detect, filter, and foster safer and healthier communication. This article introduces a novel approach to mitigating toxic content generation propensities of large language models (LLMs) by fine-tuning them with a programmable stance-directed focus on core human values and the common good. We propose a streamlined keyword coding and processing pipeline that generates weakly labeled data to train AI models to avoid toxicity and champion civil discourse. We also developed a toxicity classifier and an aspect-based sentiment analysis model to assess and control the effectiveness of a humanizing AI model. We evaluate the proposed pipeline using a contentious real-world X (formerly Twitter) dataset on U.S. race relations. Our approach successfully curbs the toxic content generation propensity of an unrestricted LLM by a significant 85%.

The increasing use of generative artificial intelligence (AI) in digital communication highlights the need for AI systems to prioritize safety and ethics. This article introduces a novel pipeline to curb the generation of toxic narratives by training more humanized large language models (LLM). Community resilience building literature,[1] indicates that social cohesion, where communities foster a sense of belonging and shared values, acts as a bulwark against toxic speech. Unlike toxic, polarizing speech, cohesive language can empower individuals to understand the harm of hateful content and stand together against it. Therefore, we develop an approach to fine-tuning generative AI on positive, core human values and the common good, reinforcing arguments to promote healthier, more humanizing, and constructive discourse.

Existing reinforcement learning from human feedback-based approaches[2] have been criticized for being expensive, time-consuming, and prone to annotator bias, often leading to models that shy away from generating content on sensitive topics. This type of training leads to either overarching guardrails, which results in refusals to generate content, or overly contentious rhetoric whenever the prompt can escape the guardrails of the LLM, particularly when dealing with sensitive or controversial issues.

The pervasive challenge of toxic content generation by LLMs is an intrinsic defect that has been

1089-7801 © 2024 IEEE
Digital Object Identifier 10.1109/MIC.2024.3450090
Date of publication 29 August 2024; date of current version 25 November 2024.

increasingly studied. Xu et al.[3] pioneered a dual-model approach, training one Generative Pre-trained Transformer 2 to produce toxic content and another conventionally. By reranking the next-word predictions compared to the toxic model's suggestions, they effectively curtailed harmful outputs. Siegelmann et al.[4] introduced a dataset of toxic and nontoxic versions of text as pairs, fine-tuning LLMs on this corpus. Through human evaluation, they demonstrated a significant reduction in toxicity, highlighting the pivotal role of data curation in mitigating this issue. Li et al.[5] explored controlling text generation on toxicity prompts.[5] They utilized an autoregressive model that quantizes and adds noise to the data, subsequently learning to predict the original data. A reinforcement learning technique with token-level feedback has been used to reduce the toxicity. These studies illuminate a dual strategy for combating toxic content: dataset curation to inform the model about undesirable outputs and robust postfiltering mechanisms to intercept residual toxicity.

Our study introduces a scalable processing pipeline that balances labor payload with comprehensiveness. Annotators focus only on identifying camps and the top frequent phrases, creating a "concise" codebook with the most informative terms, coded for sentiment [positive (pos)/negative (neg)], stance (pro-/anti-), and societal considerations (common good versus othering). The coding is performed by a select panel of experts, including a psychologist with a Ph.D. who specializes in community organizing and relations, and two computer scientists with Ph.Ds. who specialize in social networks and computational linguistics. Our approach combines social network analysis and aspect-based sentiment analysis (ABSA) to generate extensive stance-aware training data for developing stance-directed, humanizing AI.

ABSA plays a vital role in understanding the fine-grained sentiments expressed by users. Although stance detection is defined differently in different application settings, the most common definition is *automatic classification of the stance of the generator of a piece of text, towards a target, into one of these three classes: favor, oppose, and neither*.

Retweeting is a widely adopted feature on X (formerly Twitter) for spreading and making information go viral.[6] A recent study on political discourse on social media[7] confirms that X users are exposed mainly to opinions that agree with their own, and partisan users enjoy higher centrality and content endorsement. In other words, camps and echo chambers are very real.

Initially, we utilized community detection algorithms[8] on the retweet network to expose a pair of polarized camps alongside their community/subcommunity network structures (see Figure 1). Louvain community detection is being widely used in social networks.[9] To discriminate between partisan users and their more moderate, common-ground speaking counterparts, we identified two types of users in both camps: partisan *barrier bounds* versus more moderate *barrier crossers*. Barrier-crossing[10] users are defined as *those in each camp that interact (i.e., by retweeting or being retweeted by) with users from their own and with users from the other camp*. Barrier-bound[10] users are defined as *those who interact with users from their own camp only*.

To evaluate the efficacy of the previous concepts, we utilize a real-world X dataset that comprises 15,914,812 tweets from 296,540 users and 738,012 shares covering roughly one month between 18 April 2023 and 31 May 2023 on the topic of U.S. race relations.

In the following sections, we outline the architecture where our experiments demonstrate the success of our proposed architecture in curbing toxic content generation. The resultant model significantly reduces the propensity of an unrestricted LLM to produce toxic content by 85%. The experimental dataset and models used in our evaluations are online at https://github.com/tweetpie/stance-directed-humanizing-ai.

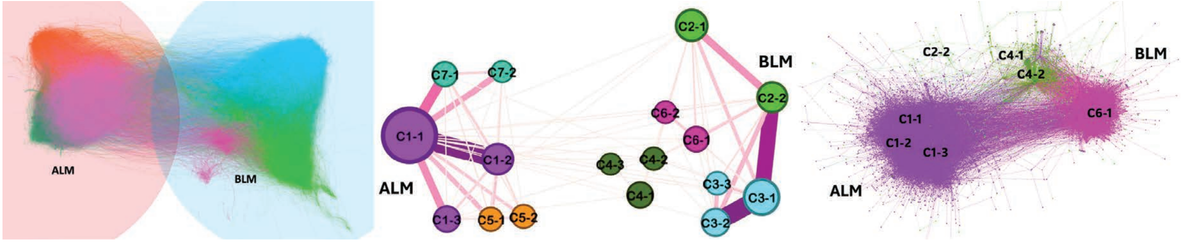

FIGURE 1. Retweet network graph of seven major communities, 17 subcommunities, and barrier-crossing users in the U.S. race relations network labeled with corresponding subcommunity IDs. BLM: Black Lives Matter; ALM: All Lives Matter.

DATASET AND CAMP LABELING

Camp Labeling of Users

Louvain community detection[9] yielded seven major communities with 17 subcommunities, which are represented as nodes with varying sizes that are depicted in different colors in Figure 1. Next, we experimentally identified an edge weight threshold value that partitions the network into a pair of polarized camps that our panel of experts labeled as either pro-Black Lives Matter (BLM) or pro-All Lives Matter (ALM) by inspecting their most viral 1000 tweets community by community.

Bias, Sentiment, and Common-Good Coding

Using named entity recognition and noun chunk processing, we extracted all noun phrases and named entities from the corpus. Our codebook includes unigrams, bigrams, and other n-grams. We noted that unigrams such as "health" might overshadow more informative bigrams like "mental health," even if "mental health" appears more frequently. To address this, we implemented a discounting mechanism, adjusting unigram frequencies by subtracting the frequencies of overlapping longer n-grams, provided that their cumulative frequency exceeds that of the unigram. Finally, we used the "elbow method" to determine the optimal cutoff frequency for identifying the most informative phrases for the codebook.

Next, three domain experts were asked to code the noun phrases and named entities independently for *ambiguity*, *pos/neg sentiment*, *pro-/antibias* toward BLM or ALM camps, and the societal considerations associated with *common good versus othering*. Following that, three experts were impaneled to discuss and reach a common consensus by resolving all of their disagreements and finalizing the codebook.

The top 3000 phrases identified through the analysis were subject to coding, with 1789 deemed nonambiguous. The sentiment column has 496 phrases with negative connotations and 257 with positive ones. For the bias, 234 phrases were coded as "opposed," with 160 as "supported" by the BLM camp, whereas 177 phrases were coded as "opposed," with 90 as "supported" by the ALM camp. Also, 214 phrases were coded as "common good" versus 567 phrases as "othering." Furthermore, 85 phrases were identified as core human values across eight topics: family, life, truth, nation, God, democracy, and justice.

Weakly Labeled Dataset

Weakly labeled data were not manually tagged, but they were not entirely unlabeled either as it follows certain verifiable assumptions. They are preferred for their minimal labor requirements and scalability, especially when fully labeled data are too costly or impractical to obtain. Using weaker forms of supervision like camps, keywords, or noisy labels, these datasets help train AI models effectively, allowing the rapid collection of large training volumes and potentially enhancing generalization.

Barrier-bound users, who tend to produce more partisan content, exhibit a higher prevalence of "othering" language than do barrier-crossing users. This distinction is crucial for curating datasets for AI training. Our methodology utilizes tweets from barrier-bound users to train ABSA-based stance-aware reader and toxicity classifier models, while tweets from barrier-crossing users are used to test these models' efficacy. Previous research[11] has revealed that amplifying partisan barrier-bound messages within their camps polarizes the network via social media echo chambers. Conversely, amplification of barrier-crossing messages within their camps has a moderating effect.

To develop an effective toxic content classifier, we experimentally refined rules for identifying nontoxic tweets through an iterative process. The criteria required nontoxic tweets to contain at least two positive phrases, no more than one negative phrase, an absence of othering/divisive language, and inclusion of at least one common-good/public-interest phrase. Toxic content was defined by at most two positive phrases, at least three negative phrases, at least one othering/divisive phrase, and at most two common-good/public-interest phrases, which yielded a dataset of 33,787 training tweets: 12,757 from the ALM camp and 21,030 from the BLM camp. Among these, 29,005 were labeled toxic, and 4782 were labeled nontoxic. The BLM tweets had a slightly higher proportion of nontoxic content (17.4%) than did the ALM tweets (13.2%).

We prepared another weakly labeled dataset for ABSA training to capture aspects of sentiment and camp bias. The dataset prioritized camp bias when present and defaulted to sentiment labels if camp bias was neutral. We filtered tweets to include only those with at least three labeled phrases, resulting in 62,196 training samples. An example from the BLM camp for the sentence "reparations will help clear the bad karma of slavery" is represented as "[reparations/PRO] will [help/PRO] clear the bad karma of [slavery/ANTI]."

Following that, we adopted barrier-crossing users' content to develop a programmable stance-directed generative AI by employing two different strategies. The first strategy involved constructing a dataset devoid of explicit toxic content to guide the generation of a socially responsible and inclusive model. Specifically, tweets from barrier-crossing users were selected where there was no mention of any othering/divisive category or any negative sentiment category phrase, and there was the presence of at least one common-good/public-interest category phrase. The second strategy utilized an unfiltered barrier-crossing subset of the dataset, allowing for greater expressivity in content without imposing any explicit rules, thus exposing the model to a broader discourse.

Ultimately, 7966 tweets matching the first strategy criteria were identified, with 57.3% originating from the BLM camp, underscoring the diversity of perspectives captured within the dataset. For the second strategy, a combination of half of the tweets from the first strategy's dataset and a similar amount of random tweets from the barrier-crossing users was utilized to ensure a similar sample size for fine-tuning our experiments.

System Architecture

Our system architecture, depicted in Figure 2, demonstrates the process of generating nontoxic messages. It highlights stance-directed prompts on selected entities and issues to guide message creation. However, the intrinsic effects of LLMs necessitate a post-filtering process to ensure the generated contents' appropriateness. Following the content generation phase, *toxic content classifier* and *ABSA* models were employed to inspect the generated tweets' toxicity and stances as a postfilter.

Stance-Directed Humanized LLM

In our application, we utilized Flan-T5[12] to fine-tune humanized and unrestricted LLMs. For the humanized LLM, we fine-tuned Flan-T5 using a dataset of barrier-crossing user tweets that adhered to specific guidelines outlined earlier as strategies 1 and 2, ensuring fairness and inclusivity in the generated content. The model was trained to generate tweets framed around specific camp biases while avoiding divisive language and maintaining a balanced representation of diverse viewpoints. Conversely, the unrestricted LLM was fine-tuned using a combination of tweets from the humanized dataset and a random subset of tweets from the barrier-crossing users, allowing for a broader exploration of language patterns and expressions.

The model is prompted through a programmable template that requests *"Write me a tweet for following camp: '<ALM or BLM>'; framing pro entities: []; anti-entities: []; neutral entities: [] based on following issues pro issues: []; anti-issues: []; neutral issues: []."* The distinction between entities and issues lies in the

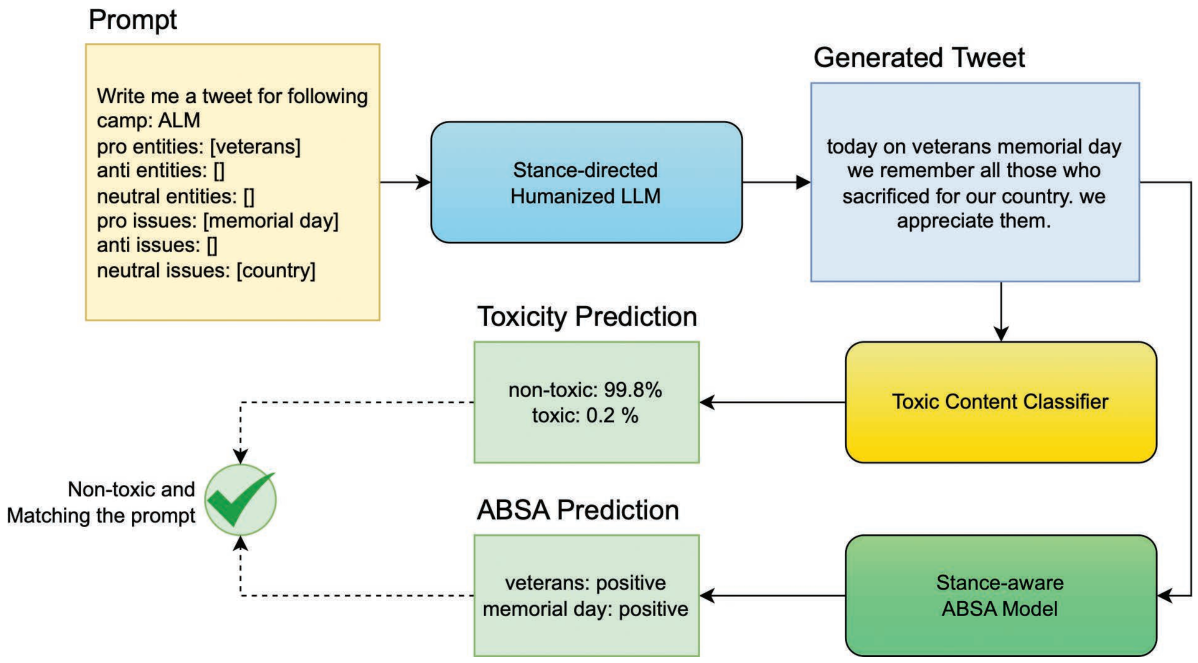

FIGURE 2. The process of generating a nontoxic message thru a stance-directed prompt on certain entities and issues.

type of noun phrases mentioned in the codebook, with entities being standardized named entities and issues being topic-coded noun phrases. Including neutral aspects aims to help the model mention specific topics without necessarily taking a clear stance as pro- or anti-.

ABSA

InstructABSA[13] builds upon the instruct-based model, a framework designed for instruction tuning LLMs to perform better on domain-specific tasks. InstructABSA extends this concept specifically to the ABSA domain by leveraging the inherent capabilities of Tk-Instruct to grasp sentiment expressions related to different aspects of a subject. The core mechanism of InstructABSA involves instruction tuning, where the model is fine-tuned with specific instruction prompts that are related to ABSA tasks. These prompts encompass task definitions and examples that represent positive, negative, and neutral sentiments. This method aims to refine the model's ability to identify, extract, and classify the sentiments that are associated with particular aspects of a text.

Toxic Content Detection

Our pipeline leverages the BERT[14] classifier in toxicity/nontoxicity detection, which has the advantage on deep understanding of language context. To enhance BERT's performance in toxicity detection, we fine-tuned a pretrained hate speech detection model.[15] This model comprises numerous examples of hate speech on race, religion, gender, sports, and politics, enabling the model to learn the specific characteristics and linguistic patterns that are associated with offensive content. The fine-tuning approach significantly improved the model's ability to detect a wide range of toxic language, from overt hate speech to more subtle forms of toxicity.

Experimental Evaluation

Our experimental evaluations are predicated on a novel approach to dataset segmentation, wherein tweets from barrier-bound users are allocated for training, and tweets from barrier-crossing users are reserved for testing. This division is designed to simulate a realistic scenario where the model is trained on a homogeneous group of users but tested on a diverse set of users that spans different communities, thereby assessing the model's capabilities across a variety of social orientations within the domain.

To mitigate overfitting and the memorization of linear relationships in the toxic content classifier, we implemented a codebook-based *phrase removal* strategy in tweets. By probabilistically removing phrases, we aimed to reduce the undue influence of high-frequency words in the codebook on classification outcomes. This approach prevents high-frequency keywords like "family" from skewing toxicity detection in inherently toxic content. Although the complete removal of high-frequency phrases would eliminate valuable insights, retaining them entirely leads to overfitting. After testing removal probabilities from 0% to 100%, we found that an 80% removal probability produced the most accurate results. To enhance dataset diversity and counteract potential biases, we integrated it with the hate speech dataset from the hate speech detection study,[18] which includes neutral, offensive, and hate categories. We merged the offensive and hate categories into a single "toxic" label and designated the neutral category as "nontoxic."

We developed a gold test dataset of 1000 expert-labeled tweets to evaluate the toxic content classifier. Toxic content was defined as *stereotyping, fostering division, employing offensive language, or demeaning/dehumanizing individuals or groups*. Using Krippendorff's alpha coefficient, we measured interannotator agreement, achieving a strong nominal score of 0.73, indicating high consistency in the assigned labels.

In the final phase of our experiment, we applied the toxic content classifier to LLM-generated tweets. An expert panel labeled these tweets, achieving an interrater reliability score of 0.75 using Krippendorff's alpha. We compared the classifier's predictions against the expert-coded labels to assess the pipeline's effectiveness in a real-world scenario.

The toxic content classifier model achieves an accuracy of between 76% and 81% in accurately categorizing tweets as toxic or nontoxic. Our assessment revealed that the ABSA model demonstrates proficiency in extracting 63% of the issues and entities from the generated text. Moreover, the model accurately matches 83% of the identified stances with the directives given in the prompt, indicating a high level of alignment.

The unrestricted LLM generates toxic content 73% of the time. Significantly, our humanized LLM achieved a 64% reduction in toxicity when responding to prompts on sensitive issues and an impressive 83% on general domain-related discourse. Even more compelling, integrating the toxic content classifier with the humanized LLM further boosted these reductions to 85% for the sensitive issue-related generation and 86% for general domain-related generation.

Comparison With Pretrained LLMs

The findings presented in this section illustrate the responses of popular LLMs to generating tweets on sensitive topics with left or right ideological leanings based on 30 balanced prompts. This analysis highlights the capabilities and limitations of contemporary LLMs in content moderation and bias mitigation. The humanized LLM model achieved a 97% response rate, with only one instance of toxicity, showcasing its robustness in handling sensitive content with minimal adverse effects. By comparison, ChatGPT-4o produced relevant responses 60% of the time without any toxic content, indicating its tendency to avoid sensitive entities like racial groups when prompts request antimentions. Gemini and Claude showed significant reluctance to engage with right-wing ideologies, with Gemini refusing 63% of the time and Claude 40%. However, both models were more responsive to left-leaning prompts, indicating a variance in engagement based on ideological context. Additionally, replacing right-wing entities and ideologies with their left-wing counterparts resulted in increased responsiveness, underscoring the impact of annotator bias.

Mistral showed a higher related response rate of 73%, comparable to the humanized LLM model, but produced abstract and fewer domain-specific tweets. Like the humanized LLM model, this model had a single instance of toxic content generation, underscoring the challenge of maintaining nontoxic outputs across diverse prompts. Despite these differences, all the LLMs managed to generate predominantly nontoxic tweets. The comparison underscores the humanized LLM model's efficacy in handling sensitive topics with high relevance and low toxicity, outperforming other LLMs in consistency and engagement. Detailed prompts and outputs are accessible in the public GitHub repository.

DISCUSSION

Our open-domain discourse exploration revealed that the models, especially the humanized AI, demonstrate a preference for nontoxic language. This highlights their ability to engage in complex conversations without resorting to harmful expression, likely due to robust pretraining. However, a challenge arises when prompts contain intentionally mentioned toxic elements as this can unintentionally trigger the generation of subtly toxic content, illustrating the need for a toxic content classifier in any generative pipeline.

Alongside randomly removing frequent phrases, approaches such as masked training and joint training could also be explored as promising research avenues for enhancing ABSA's accuracy in discerning the entity (or issue) alongside its stance. Additionally, it is crucial to study newly emerging toxic keywords as words that were once benign can evolve to be used offensively. A valuable follow up to this study could involve comparing discussions on similar topics across different years to observe these changes.

Our experimental results reveal a substantial decrease in toxic language generation when using the proposed stance-directed humanized LLM compared to the unrestricted model. This underscores the significant practical impact of our approach in curbing harmful language generation across contentious rhetoric. These findings affirm the potential of our pipeline as a promising step toward fostering safer and more inclusive digital environments.

The proposed approach is adaptable to any social media discourse where communities can be distinctly identified. By leveraging the communities and their most frequently used phrases to train the model, it is essential that the favored and opposed phrases be unambiguous. This methodology can be effectively applied to various discussions, such as climate change, vaccine safety, gun control, and similar high-profile topics where stances are clearly delineated.

CONCLUSION

In conclusion, our codebook-driven approach for weakly labeled data generation, LLM training, and refined classification techniques significantly reduced the toxic content generation of an unrestricted LLM while preserving stance accuracy. This demonstrates the potential of our pipeline to improve digital communication standards, fostering safety, respect, and shared understanding across diverse and contentious issues.

Looking ahead, we see several promising paths for refining our models' efficacy and generalizability. Data augmentation strategies that use synonyms and contextual cues could enhance classification accuracy. Additionally, incorporating context-aware metadata into weakly labeled data for ABSA could improve the model's sensitivity to stances and the common good.

These advancements pave the way for models that navigate complex discourse, generating tailored, contextually appropriate content that champions social responsibility and inclusivity. This approach aligns with the democratic principle of consensus building by fostering understanding through seeking common ground among diverse viewpoints.

ETHICAL CONSIDERATIONS

Our diverse expert panel ensures fairness and nondiscrimination, aiming to develop AI that promotes positive societal values. We maintain data confidentiality and strive for transparency and accountability in our methods, contributing to society through responsible research. Acknowledging the potential misuse of our pipeline to generate harmful content, we emphasize the importance of responsible application. We remain vigilant to these risks, continually working to mitigate them through rigorous evaluation and ethical oversight.

ACKNOWLEDGMENTS

The work of Yusuf Mücahit Çetinkaya was supported by TUBITAK-2214-A.

REFERENCES

1. L. Moustakas, "Social cohesion: Definitions, causes and consequences," *Encyclopedia*, vol. 3, no. 3, pp. 1028–1037, 2023, doi: 10.3390/encyclopedia3030075.
2. L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 27730–27744.
3. C. Xu, Z. He, Z. He, and J. McAuley, "Leashing the inner demons: Self-detoxification for language models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 10, 2022, pp. 11530–11537, doi: 10.1609/aaai.v36i10.21406.
4. R. Siegelmann et al., "MICo: Preventative detoxification of large language models through inhibition control," in *Proc. Findings Assoc. Comput. Linguistics: NAACL,* 2024, pp. 1696–1703, 2024.
5. W. Li, W. Wei, K. Xu, W. Xie, D. Chen, and Y. Cheng, "Reinforcement learning with token-level feedback for controllable text generation," 2024, *arXiv:2403.11558*.
6. D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *Proc. 43rd Hawaii Int. Conf. Syst. Sci.,* 2010, pp. 1–10.
7. K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship," in *Proc. World Wide Web Conf.,* 2018, pp. 913–922.
8. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Statist. Mech.: Theory Exp.,* vol. 2008, no. 10, 2008, Art. no. P10008, doi: 10.1088/1742-5468/2008/10/P10008.
9. N. Dakiche, F. B. S. Tayeb, Y. Slimani, and K. Benatchba, "Tracking community evolution in social networks: A survey," *Inf. Process. Manag*, vol. 56, no. 3, pp. 1084–1102, 2019, doi: 10.1016/j.ipm.2018.03.005.
10. M. D. Buhrmester, M. A. Cowan, and H. Whitehouse, "What motivates barrier-crossing leadership?," *New England J. Public Policy*, vol. 34, no. 2, 2022, Art. no. 7.
11. Y. Lee, M. Ozer, S. R. Corman, and H. Davulcu, "Identifying behavioral factors leading to differential polarization effects of adversarial botnets," *ACM SIGAPP Appl. Comput. Rev.*, vol. 23, no. 2, pp. 44–56, 2023, New York, NY, USA: ACM, doi: 10.1145/3610409.3610412.
12. C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.
13. K. Scaria, H. Gupta, S. Goyal, S. A. Sawant, S. Mishra, and C. Baral, "InstructABSA: Instruction learning for aspect based sentiment analysis," 2023, *arXiv:2302.08624*.
14. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
15. C. Toraman, F. Şahinuç, and E. H. Yilmaz, "Large-scale hate speech detection with cross-domain transfer," in *Proc. Lang. Resour. Eval. Conf.*, Marseille, France, Jun. 2022, pp. 2215–2225, European Language Resources Association. Accessed: Sep. 9, 2024. [Online]. Available: https://aclanthology.org/2022.lrec-1.238

YUSUF MÜCAHIT ÇETINKAYA is a Ph.D. candidate in computer engineering at Middle East Technical University (METU), Ankara, 06800, Türkiye, and a researcher at Arizona State University. His research interests include social media analysis, natural language processing, and information retrieval. Çetinkaya received his M.S. degree in computer engineering from METU. Contact him at yusufc@ceng.metu.edu.tr or ycetinka@asu.edu.

YEONJUNG LEE is a research assistant with Arizona State University (ASU), Tempe, AZ, 85281, USA. Her research interests include machine learning, graph neural networks, and natural language processing. Lee received her Ph.D. degree in computer science from ASU. Contact her at ylee197@asu.edu.

EMRE KÜLAH is a Ph.D. candidate in computer engineering at Middle East Technical University (METU), Ankara, 06800, Türkiye. His research interests include social media analysis, natural language processing, and sports analytics. Külah received his M.S. degree in computer engineering from METU. Contact him at kulah@ceng.metu.edu.tr.

İSMAIL HAKKI TOROSLU is a professor in the Department of Computer Engineering, Middle East Technical University, Ankara, 06800, Türkiye. His research interests include data mining, information retrieval, and intelligent data analysis. Toroslu received his Ph.D. degree in computer science from Northwestern University. Contact him at toroslu@ceng.metu.edu.tr.

MICHAEL A. COWAN is an Emerity faculty member at Loyola University, New Orleans, LA, 70118, USA. His research interests include community organizing, racial and ethnic issues, and urban issues. Cowan received his doctoral degree in psychology from The Ohio State University. Contact him at mcowan@loyno.edu.

HASAN DAVULCU is a professor in the School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, 85281, USA. His research interests include data mining, sociocultural modeling, and persuasive artificial intelligence. Davulcu received his Ph.D. degree in computer science from Stony Brook University. Contact him at hdavulcu@asu.edu.