Information Research - Vol. 31 No. iConf (2026)

From empathy to exclusion: how Immigrant groups are framed in online discourse

Kirin Mohile and Yiqi Li

DOI: https://doi.org/10.47989/ir31iConf64261

Abstract

Introduction. When examining moral foundations in immigration discourse on social media, most studies focus on ideology-based groups rather than across specific immigrant groups. This ignores moral framings that depict some communities with empathy and others as threats. This research explores how moral foundations vary in Twitter conversations about five immigrant groups.

Method. Tweets were sorted into five categories (based on immigrant groups being discussed: African, Asian, European, Latin American, or Middle Eastern) utilising keyword searches and AI LLM modeling. GPT-3.5 Turbo was employed and achieved a satisfactory performance (0.83) compared to manual human labeling.

Analysis. Scores for foundation variables (care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and purity/sanctity) were analysed using enhMFD1 dictionary. One-way ANOVA tested overall differences between groups and Tukey's HSD post-hoc test identified specific patterns.

Results. Latin American immigrant discourse emphasised care and authority. European-focused tweets featured stronger loyalty. African immigrant discourse highlighted loyalty with moderate authority, discourse about Middle Eastern immigrants showed elevated harm and betrayal, and Asian immigrant discourse portrayed higher fairness-vice.

Conclusion(s). Immigrant groups are framed differently through moral language in social media conversations, which may influence perceptions and can inform strategies for addressing harmful narratives on social media.

Introduction

We live in a digital age where social media has become a powerful space for people to share opinions, spread ideas, and influence political beliefs. Among the many issues debated online, immigration stands out as one of the most prominent and contentious (Conzo et al., 2021). Individuals from diverse backgrounds use these platforms not only to express their views, but also to persuade and inform others about immigration and its broader implications.

Immigration in social media discourse

Studies have quantitatively and qualitatively examined immigration and refugee-related discourse on social media. For example, Arunasalam et al. (2024) explored 1,400 tweets in seven languages and found that trolling and hate speech were the most prevalent forms of toxic content against refugees. Similarly, Khatua and Nejdl (2023) used machine learning to explore why migrants are targeted on social media and found that cultural concerns caused more toxicity than security or economic concerns. Researching the motives behind anti-immigrant sentiments online, Menshikova and Tubergen (2022) followed a panel of 28,000 Twitter users in the United Kingdom. They found that people tweet more negatively about immigrants in response to more significant news coverage of immigration.

While these studies highlight how immigration discourse is shaped by culture and media influence, they treat immigrants or refugees as homogenous categories. Little to no research has explored how discourse shifts when attention focuses on specific groups. Analysing the themes that emerge in discussions about different immigrant groups provides a valuable starting point for understanding how stereotypes, cultural narratives, and political anxieties affect each group differently. This motivates our first research question:

RQ1: How do dominant thematic categories differ across tweets referencing specific immigrant groups?

Moral foundations in online immigration context

To better understand immigration debates, moral foundations theory (MFT) provides a useful framework, as it allows us to explore how different moral appeals can reinforce or shift opinions and political attitudes. In MFT, moral foundations are the core dimensions people use to judge right and wrong: care/harm (well-being), fairness/cheating (justice), loyalty/betrayal (group belonging), authority/subversion (order), and purity/sanctity (sacredness) (Graham et al., 2013).

Many studies have explored MFT in relation to immigration attitudes. For example, Nath et al. (2022) researched whether pro-immigration messages framed around different moral foundations could influence U.S. immigration attitudes across the political spectrum. They found that responses varied depending on participants' political leanings. Hoewe et al. (2021) used moral foundations to assess Americans' perceptions of immigrants and refugees, finding that refugees were more commonly framed with the care/harm moral foundation and immigrants with loyalty/betrayal. Lastly, Grover et al. (2022) examined how liberals and conservatives in the U.S. use different moral foundations when discussing immigration. Using twitter data, they found that pro-immigration tweets emphasised harm, fairness, and loyalty, while anti-immigration tweets emphasised authority.

While studies have consistently explored moral foundations as it relates to immigration, few have assessed how moral language varies in discourse about different immigrant groups. We hypothesise that framings differ across groups, reflecting distinct strategies to appeal to audiences' values (Luttrell & Trentadue, 2024). This leads us to pose our second research question:

RQ2: How does moral framing vary across discourse about different immigrant groups?

Classification for ethnicity-focused content

The majority of research that has explored ethnicity in the context of social media has been at the user-level, not at the text-level. For example, Hofstra and de Schipper (2018) provided a method for predicting the ethnicity of a user on social media utilising their first name. Moreover, Rubenzer (2016) examined how ethnic identity interest groups in the U.S. use social media, finding that more powerful groups were most active online, while those facing intergenerational challenges were not leveraging social media as much.

Only a limited number of studies have categorised social media text based on the ethnic group the text is referencing. For example, Whitfield et al. (2025) classified Reddit text mentioning Black and Asian communities using topic modeling and applied a custom named-entity recognition model to understand how the COVID-19 pandemic disproportionately affected these groups. Additionally, Koltsova et al. (2017) studied how ethnic groups in post-Soviet Russia are represented in user-generated content. They classified ethnicity-related discussions in Russian social media by developing a comprehensive lexicon of over 4,000 keywords covering 97 ethnic groups, using keyword matching to select relevant texts. While these studies provide a starting point, the lack of research on categorising social media text by ethnic group leads us to our third research question:

RQ3: What is the most effective method for classifying the ethnic group being discussed in tweets?

Significance

This study makes several key contributions to research on immigration discourse. First, rather than examining online conversations solely through the lens of issue stance or political ideology, it focuses on five specific immigrant groups in the U.S. (Asian, African, European, Latin American, and Middle Eastern) to uncover how ethnicity being discussed influences immigration debates. Second, it advances methodology by offering a cost-efficient approach for classifying tweets by immigrant groups that can be applied to future research. Third, it contributes to moral foundations theory by introducing ethnicity as an important variable, showing how moral values are applied unevenly across groups. Understanding these differences can provide insight into the role of morality in guiding public opinion about immigration and highlight the nuanced ways people support or criticise various immigrant communities.

Methodology

Data

The dataset chosen to answer our research questions consists of 305,878 total mention tweets and 151,803 tweets with unique content posted between January 2021 and March 2024, during the Biden Administration. All tweets are English-language, collected using a curated list of immigration-related keywords and hashtags. We removed non-U.S. tweets using manual coding combined with distilBERT modeling, reaching a satisfiable accuracy level (See Appendix for detail). We focus on mention posts because they reflect both individual views and interactions around ethnicity-focused framings in networked publics (Boyd, 2010; Papacharissi, 2015). Moral expressions in tweets often serve advocacy purposes, targeting both mentioned stakeholders and the broader public (Yang et al., 2021).

Morality calculation

We employed enhMFD1 dictionary (Rezapour & Shah, 2019) to detect moral dimensions in tweets. This dictionary is one of the most comprehensive annotated morality dictionaries based on extensive Twitter data containing 4636 entries. For each tweet, we calculated moral foundation scores for both virtue (praising) and vice (condemning) dimensions of care, fairness, loyalty, authority, and sanctity (Graham et al., 2013). These scores represent the proportion of words in

each tweet associated with a given moral foundation: higher scores indicate a greater presence of that moral dimension.

Classification

Five broad demographic categories of immigrants were chosen for this study: Asian, African, European, Middle Eastern, and Latin American. While these categories are not without limitations, they provide a useful framework because individuals within these group categories reflect shared regional and cultural backgrounds. Moreover, these groupings align with how immigrants are often represented and perceived in U.S. media and public discourse. To categorise tweets by immigrant groups, we created a comprehensive list of keywords associated with each group. These keywords include countries, ethnonyms, demonyms, and common terms associated with that immigrant group. Tweets containing at least one keyword were labelled for that group using a simple regex-based approach, consistent with prior Twitter content classification studies (Davidson et al., 2017; Koltsova et al., 2017).

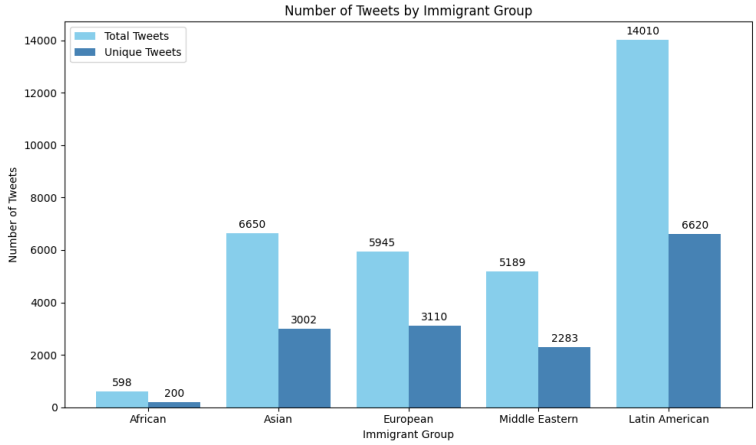

Figure 1. Number of total and unique tweets by Immigrant group after keyword filtering.

Keyword filtering alone can be inaccurate. For example, '*Jordan*' may refer to the politician Jim Jordan instead of the country, and tweets that contain '*China*' may be about politics and not immigration. Recent work has demonstrated that large language models (specifically ChatGPT) can be reliably used for social science text classification and annotation when combined with human validation (Gilardi et al., 2023; Törnberg, 2023). Therefore, we employed a second filtering method using OpenAI's GPT 3.5-turbo model to verify results. Each tweet was checked by prompting GPT to determine if it was about the specific immigrant group. Verified tweets were kept in the analysis, and the rest were excluded.

The prompt used was:

"You're reviewing tweets that may or may not be about {group_name} immigrants.

Tweet: '{*tweet*}'

Is this tweet about or related to {group_name} immigrants or immigration?

Reply '*Yes*' if it is, and '*No*' if it's unrelated.'"

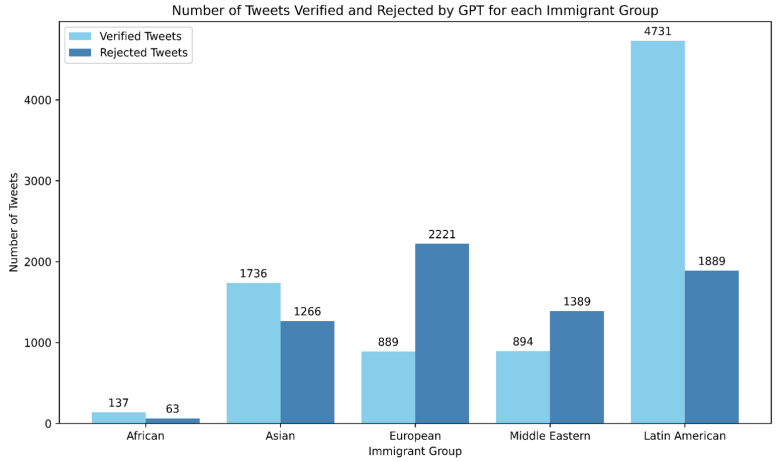

Figure 2. Number of tweets verified and rejected by GPT by Immigrant group.

The combination of keyword filtering and GPT verification offers a cost-efficient method for accurately categorising tweets by immigrant groups. Applying GPT to label the entire dataset would require significantly more resources and time. By first using keyword filtering to narrow down relevant tweets, we achieve a more efficient classification process.

Evaluation

To assess the accuracy of GPT, we randomly selected 25 rejected and 25 verified tweets from each of the five categories. We hand-labelled each tweet as '*yes*' (about the group) or '*no*' (not about the group) and compared these with GPT's labels. We generated a confusion matrix (shown below), which visually shows where GPT's predictions matched or differed from human labels. We calculated an accuracy score of 0.83, indicating satisfactory performance.

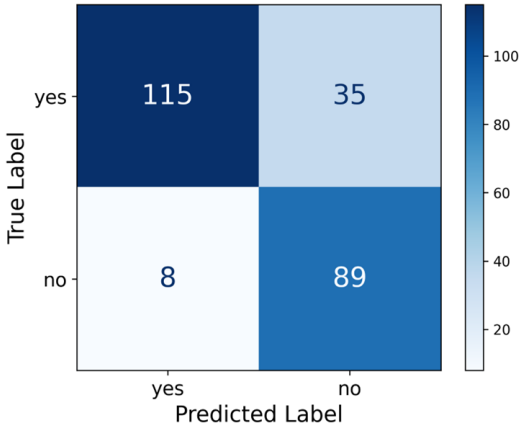

Figure 3. Confusion matrix.

Report	Precision	Recall	f1-score
yes	0.93	0.77	0.84
no	0.72	0.92	0.81
accuracy			0.83
Macro avg	0.83	0.84	0.82
Weighted avg	0.85	0.83	0.83

Table 1. Classification report.

Topic modeling

To better capture main themes in the dataset, we conducted topic modeling using BERTopic for each immigrant group (Grootendorst, 2022). This approach highlights the central issues driving immigration-related discourse during this specific time period and the contexts in which people are talking about the different ethnic groups.

Moral foundations analysis

Average levels of each moral dimension were analysed for each immigrant group. We used one-way ANOVA to test for overall differences between groups and Tukey's Honest Significant Difference (HSD) post-hoc test to identify which specific group pairs differed significantly.

Results

Topic modeling

The top three overall topics that emerged were discussions of Mexico and illegal border immigration (4,229 tweets), Indian immigrants seeking green cards (1,040 tweets), and Chinese/Asian immigration (586 tweets). The most prevalent topics about Middle Eastern immigrants include Muslim refugees (390 tweets), Afghanistan (217 tweets), and Israel/Palestine (102 tweets). The top topics in the European category consist of Irish immigration (187 tweets) and Ukrainian refugees (112 tweets).

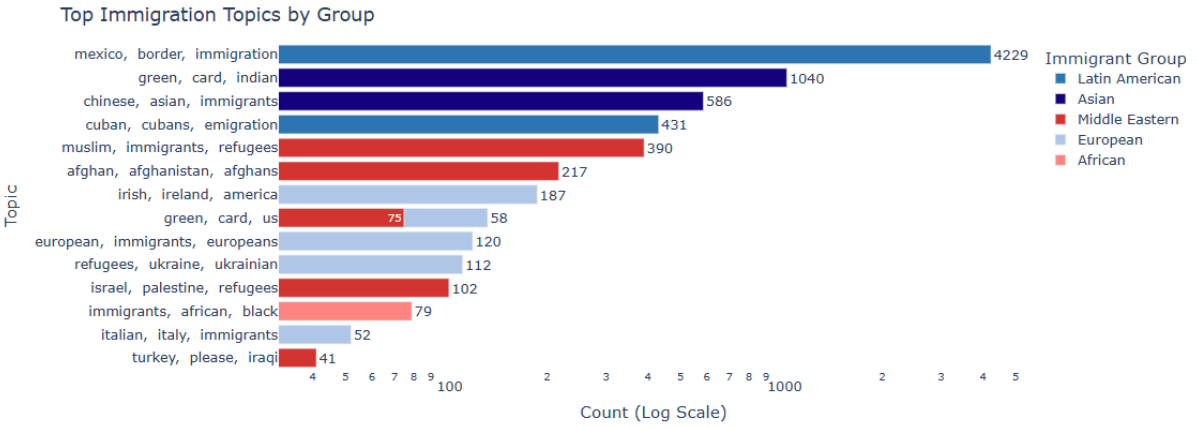

Figure 4. Top 15 Topics colour coded by immigrant group.

Moral foundations

Care-virtue ($F = 69.37$, $p < 0.001$) was significantly higher in tweets about Latin American immigrants, with an average score of 0.0136. Discourses about Middle Eastern, European Asian, and African immigrants had average care-virtue scores of 0.0097, 0.0052, 0.0054, and 0.0060 respectively. Loyalty-virtue ($F = 7.72$, $p < 0.001$) was highest in tweets about European immigrants (average score = 0.0142). There were no significant differences between loyalty-virtue scores in tweets about the other immigrant groups. Authority-virtue ($F = 4.60$, $p < 0.01$) was highest in tweets about Latin American immigrants with an average score of 0.0108. There were no significant differences in levels of fairness-virtue ($F = 1.23$, $p > 0.05$) and sanctity-virtue ($F = 2.28$, $p > 0.05$).

Care-vice ($F = 11.45$, $p < 0.001$) was significantly higher in tweets about Middle Eastern immigrants (average score = 0.0082). Fairness-vice ($F = 17.10$, $p < 0.001$) was highest in discourse about Asian immigrants (average score = 0.0018). Tweets about Middle Eastern immigrants showed elevated levels of loyalty-vice ($F = 4.63$, $p < 0.001$) with an average score of 0.0034. Discourse about Latin American immigrants showed the highest levels of authority-vice ($F = 40.23$, $p < 0.001$) with an average score of 0.0062. There were no significant differences in levels of sanctity-vice ($F = 0.41$, $p > 0.05$). Figure 5 depicts a heatmap showing the average moral foundations by group and Table 2 provides example tweets for each relevant moral foundation.

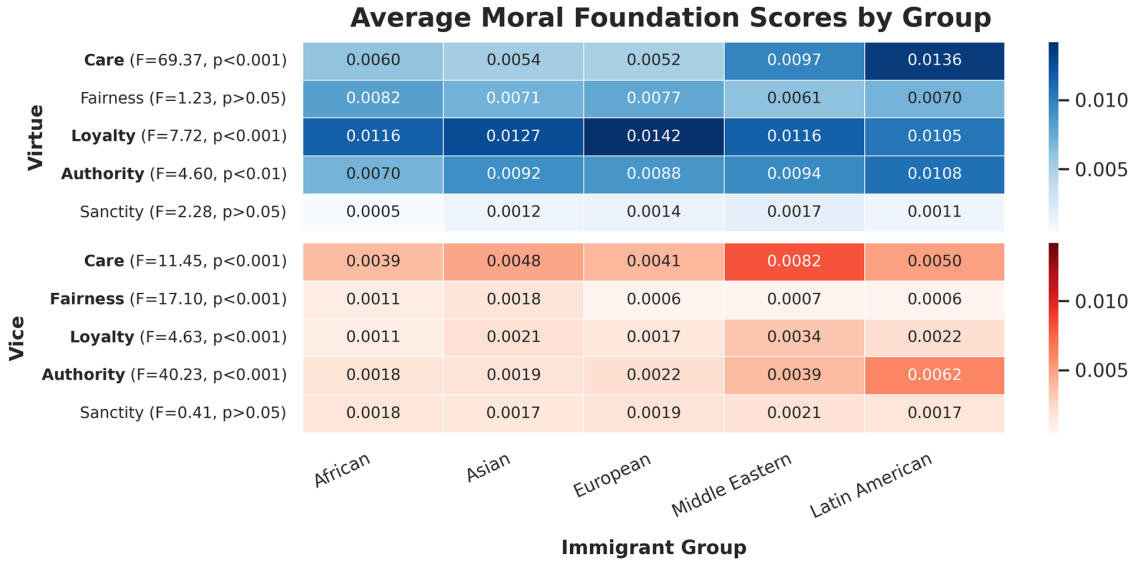

Figure 5. Average moral foundation scores by group.

Moral Foundation	Immigrant Group	Tweet	Score	Author Username	Date Posted
Care (Virtue)	Latin American	'What is he talking about, it's pro-Latino love! Sanctuary City Summer of Love Latino Style!'	0.1875	griptiger	Aug 31, 2022
Care (Vice)	Middle Eastern	'America fights for democracy and freedom. What does the Middle East and Islam fight for? The right to oppress women, throw gay people off buildings, terrorism, tyranny, etc. You should emigrate to the Middle East.'	0.1142	hucklehux23	Apr 26, 2022
Fairness (Vice)	Asian	'Deportation of documented dreamers. The root cause is country-based caps discriminating against Indians.'	0.0700	trotterrrrr	Jan 15, 2023
Loyalty (Virtue)	European	'How #Irish #emigration is celebrated in the home #country #Ireland #immigration #economy #StPatricksDay.'	0.1670	justinpowellweb	Mar 17, 2024
Loyalty (Vice)	Middle Eastern	'Disgusting terrorists raising terrorists in our country. Immigration reform now! Deport and block any more Muslim radicals from coming.'	0.1050	michaelwestgat7	Nov 20, 2023
Authority (Virtue)	Asian	'Meanwhile, legal aliens (educated, law abiding, paying $$$ taxes... mostly Indian professionals) waiting for decades for green cards.'	0.1170	plugpulr1	Feb 22, 2024
Authority (Vice)	Latin American	'You can't discuss illegal immigration without 'talkin' bout' Mexicans, who represent the VAST MAJORITY of illegal immigrants!'	0.1170	rutledgecharle1	Sep 08, 2022

Table 2. Notable examples of tweets from the immigrant groups, showcasing different moral foundations.

Discussion

This dataset is conducive for studying how online discourse about immigrants works not only as conversation between individuals but also as a form of public advocacy. Even when tweets are directed at specific users through mentions, they remain publicly visible and take place in a broader network environment (O'Connor & Shumate, 2020). As Yang and Saffer (2020) describe, the way issues are framed in such spaces shapes how much attention they receive from the public. Thus, online discourse is not only a reflection of private opinions but also a tool to influence others, amplify grievances, or legitimise exclusion (Chouliaraki & Zaborowski, 2017). Focusing on specific immigrant groups shows that some communities are singled out with distinct moral framings, revealing how discourses are distributed unevenly across groups.

Differences in moral foundations between immigrant groups

The findings of this research imply that online discourse about different immigrant groups contain varying levels of each moral foundation. Tweets about Latin American immigrants show a significantly higher care-virtue score and authority-vice score, as well as the highest authority-virtue score. This is most likely because Latin American immigrants are often framed in humanitarian contexts (Santa Ana, 2002; Khatua & Nedjl, 2023). Conversations about Latin Americans contain words like '*families*', '*children*', '*safety*', and '*opportunity*', which may be related to an elevated moral foundation of care. Additionally, language used to talk about Latin American immigrants is often framed in the context of following the law and illegality. Words like '*borders*', '*illegal*', and '*deportation*' increase the levels of authority framing. The presence of both authority-virtue and authority-vice framing indicates that both supporters and critics of this immigrant group utilise this foundation to justify their positions.

Discourse about Middle Eastern immigrants contains the highest levels of care-vice and elevated loyalty-vice, signaling concerns about harm and betrayal. Words such as '*risk*,' '*danger*,' '*security*,' and '*outsider threat*' are often used to describe this group, which may account for these elevated foundations. They are frequently cast through the lens of potential danger, disloyalty, and betrayal, and are framed as outsiders (Alsutany, 2012; Bail, 2012). While positive care framing does appear, the elevated vice scores suggest that negative, threat-based interpretations are especially salient. Tweets about Asian immigrants have the highest levels of fairness-vice. This suggests a moral framing around cheating and unfair advantage. Asian immigrants may be seen as economic competition, with discourse centering around complaints about violated rules or unequal treatment (Wu, 2014).

Discourse about European immigrants contain the highest levels of loyalty-virtue, with low vice overall. This may reflect Europeans' perceived cultural proximity and ingroup status (Bonilla-Silva, 2004). Tweets therefore use more language of solidarity, shared heritage, and belonging. Because European immigrants are framed closer to the ingroup, loyalty is used positively to express inclusion and unity rather than exclusion or betrayal. African-focused discourse features loyalty-virtue most prominently, with moderate authority and care. This may stem from solidarity (diasporic ties, shared struggles) combined with authority rhetoric about order and respectability (Pierre, 2004). Vice signals are relatively low, suggesting less moralised condemnation than for Middle Eastern or Latin American targets, but more ambivalence than for Europeans.

Implications

For scholars, these findings highlight the value of analysing group-specific moral framings to move beyond generalised accounts of online hostility. By examining how immigrant groups are linked to distinct moral narratives, scholars can gain a better understanding of what drives public discourse and conflict. This will allow research to capture nuances in the ways that morality shapes group-specific perceptions. For policymakers and advocacy groups, these findings suggest the need for targeted strategies to address harmful narratives. Because moral framings can legitimise hostility against specific immigrant groups, interventions must be adapted to each group's specific dynamics. Designing communication and advocacy efforts around the moral dimensions and vocabularies relevant to each group may offer a path toward reducing harm and fostering inclusive dialogue towards immigrants online.

Limitations and future work

While these findings provide valuable insight, there are several limitations that should be noted. First, the analysis relies on dictionary-based measures of moral foundations, which capture explicit language but may miss sarcasm or more subtle moral cues. Second, the grouping of immigrants into broad regional categories may obscure important within-group differences and reinforce overly broad labels. As Brubaker (2004) notes, treating immigrant or ethnic groups as uniform ignores important differences in history, legal status, religion, and social position, which may shape how different subgroups are discussed and moralised online. Third, because the data are drawn from a specific online platform (twitter) and time period (the Biden administration), the results may reflect contextual events and platform-specific dynamics rather than general patterns of discourse.

Future research could examine whether these trends persist across other social media platforms and within a broader range of Twitter posts. It would also be valuable to analyse how discourse has shifted over time by incorporating more current data. In addition, focusing on immigrant subgroups (e.g., individual countries) could provide a more nuanced understanding of how discourse varies. Finally, extending the analysis to other linguistic dimensions, such as emotional

tone or metaphorical framing, could offer deeper insights into how immigrants are discussed online.

Conclusion

This study highlights the importance of examining immigration discourse through the lens of ethnicity. By analysing tweets referencing five distinct immigrant groups in the U.S., we uncover how moral values are applied differently across communities and identify the nuanced ways online conversations reflect support and criticism for different immigrant groups. Methodologically, our combination of keyword filtering and GPT verification offers a scalable, cost-efficient approach for classifying social media text by immigrant group. Overall, these findings contribute to a deeper understanding of how morality and ethnicity intersect in online debates about immigration, emphasising that scholars and policymakers should consider group-specific dynamics when analysing public opinion and designing interventions online.

Acknowledgements

This work is funded by the National Science Foundation through the Research Experience for Undergraduates (REU) program at the Syracuse University School of Information Studies (iSchool).

About the authors

Kirin Mohile is an undergraduate student at Duke University, studying Computer Science and Linguistics. His research interests centre on how computational methods can enrich research in the social sciences and humanities, with a particular focus on ensuring that AI and machine learning are applied thoughtfully and inclusively. He can be contacted at kirin.mohile@duke.edu

Yiqi Li (https://orcid.org/0000-0002-3730-5743) is an assistant professor at the School of Information Studies (iSchool) at Syracuse University (SU). Her research examines the intersection of social networks, computational communication, and online community organising. Integrating advanced computational methods, including large language models, network analysis, and machine learning, Dr. Li investigates phenomena such as incivility diffusion, political polarisation, and social influence in social media communities. She can be reached at yli360@syr.edu

References

Arunasalam, A., Farrukh, H., Tekcan, E., & Celik, Z. B. (2024). An exploration of online toxic content against refugees. In Symposium on Usable Security and Privacy (USEC).

Bail, C. A. (2012). The fringe effect: Civil society organisations and the evolution of media discourse about Islam since the September 11th attacks. American Sociological Review, 77(6), 855-879.

Bonilla-Silva, E. (2004). From bi-racial to tri-racial: Towards a new system of racial stratification in the USA. Ethnic and racial studies, 27(6), 931-950.

Boyd, D. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In A networked self (pp. 47-66). Routledge.

Brubaker, R. (2004). Ethnicity without groups. Harvard university press.

Chouliaraki, L., & Zaborowski, R. (2017). Voice and community in the 2015 refugee crisis: A content analysis of news coverage in eight European countries. International Communication Gazette, 79(6-7), 613-635.

Conzo, P., Fuochi, G., Anfossi, L., Spaccatini, F., & Mosso, C. O. (2021). Negative media portrayals of immigrants increase ingroup favoritism and hostile physiological and emotional reactions. Scientific Reports, 11(1), 16407.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In Proceedings of the international AAAI conference on web and social media (Vol. 11, No. 1, pp. 512-515).

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30), e2305016120.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In Advances in Experimental Social Psychology (Vol. 47, pp. 55–130). Academic Press.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Grover, T., Bayraktaroglu, E., Mark, G., & Rho, E. H. R. (2019). Moral and affective differences in US immigration policy debate on Twitter. Computer Supported Cooperative Work (CSCW), 28(3), 317–355. https://doi.org/10.1007/s10606-019-09365-0

Hoewe, J., Panek, E., Peacock, C., Sherrill, L., & Wheeler, S. (2021). Using moral foundations to assess stereotypes: Americans' perceptions of immigrants and refugees. Journal of Immigrant & Refugee Studies, 20(4), 501–518. https://doi.org/10.1080/15562948.2021.1949657

Hofstra, B., & de Schipper, N. C. (2018). Predicting ethnicity with first names in online social media networks. Big Data & Society, 5(1), 2053951718761141. https://doi.org/10.1177/2053951718761141

Khatua, A., & Nejdl, W. (2023, September). Why do we hate migrants? A double machine learning-based approach. In Proceedings of the 34th ACM Conference on Hypertext and Social Media (pp. 1–10).

Koltsova, O., Nikolenko, S., Alexeeva, S., Nagornyy, O., & Koltcov, S. (2017, June). Detecting interethnic relations with the data from social media. In International Conference on Digital Transformation and Global Society (pp. 16–30). Cham: Springer International Publishing.

Luttrell, A., & Trentadue, J. T. (2024). Advocating for mask-wearing across the aisle: Applying moral reframing in health communication. Health Communication, 39(2), 270-282.

Menshikova, A., & van Tubergen, F. (2022). What drives anti-immigrant sentiments online? A novel approach using Twitter. European Sociological Review, 38(5), 694–706. https://doi.org/10.1093/esr/jcac021

O'Connor, A., & Shumate, M. (2020). A multidimensional network approach to strategic communication. In Future directions of strategic communication (pp. 71-88). Routledge.

Pierre, J. (2004). Black immigrants in the United States and the' cultural narratives' of ethnicity. Identities: Global studies in culture and power, 11(2), 141-170.

Papacharissi, Z. (2015). Affective publics: Sentiment, technology, and politics. Oxford University Press.

Rezapour, R., Shah, S. H., & Diesner, J. (2019, June). Enhancing the measurement of social effects by capturing morality. In Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 35–45).

Rubenzer, T. (2016). Social media foreign policy: Examining the political use of social media by ethnic identity groups in the United States. Politics, 36(2), 153–168.

Santa Ana, O. (2002). Brown tide rising: Metaphors of Latinos in contemporary American public discourse. University of Texas Press.https://doi.org/10.1111/1467-9256.12109

Törnberg, P. (2023). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. arXiv preprint arXiv:2304.06588.

Whitfield, C., Liu, Y., & Anwar, M. (2025). Impact of COVID-19 pandemic on social determinants of health issues of marginalised Black and Asian communities: A social media analysis empowered by natural language processing. Journal of Racial and Ethnic Health Disparities, 12(3), 1641–1656. https://doi.org/10.1007/s40615-024-01927-7

Wu, E. D. (2014). The colour of success: Asian Americans and the origins of the model minority. Princeton University Press.

Yang, A., Choi, I. M., Abeliuk, A., & Saffer, A. (2021). The influence of interdependence in networked publics spheres: How community-level interactions affect the evolution of topics in online discourse. Journal of Computer-Mediated Communication, 26(3), 148-166.

Yang, A., & Saffer, A. J. (2020). Standing out in a networked communication context: Toward a network contingency model of public attention. New Media & Society, 23(10), 2902–2925. https://doi.org/10.1177/1461444820939445

Appendix

DistilBERT

DistilBERT was trained and evaluated on a dataset of 3,654 human-coded and verified tweets. Of these, 76.5% were used for training, 15% for validation, and 8.5% for testing.

On the test set, the model achieved:

- Class 0 (negative cases): precision = 0.86, recall = 0.91, F1 = 0.88 (141 examples)
- Class 1 (positive cases): precision = 0.92, recall = 0.88, F1 = 0.90 (170 examples)

Overall performance:

- Accuracy: 0.89 on 311 test examples
- Macro average (both classes equally): precision = 0.89, recall = 0.89, F1 = 0.89
- Weighted average (accounting for class sizes): precision = 0.89, recall = 0.89, F1 = 0.89

Harmful Language Analysis

In addition to Moral Foundations, we also analyzed average levels of toxicity, obscenity, identity attacks, threat, and insult across immigrant groups. Toxicity ($F = 3.69$, $p < 0.01$) was greatest in discourse about Middle Eastern immigrants with an average score of 0.1257 and in tweets about Asian immigrants with an average score of 0.1155. It was lowest in tweets about European immigrants with an average score of 0.0899. Identity Attacks ($F = 23.20$, $p < 0.001$) were most prevalent in tweets about African immigrants (0.0838), Asian immigrants (0.0708), and Middle Eastern immigrants (0.0696). They were less common in discourse about Latin American immigrants (0.0440) and European immigrants (0.0371). Threat ($F = 2.41$, $p < 0.05$), while much less common than the other variables, was significantly higher in tweets about Middle Eastern immigrants (0.0026). Tweets about Asian and Latin American immigrants had an average threat score of 0.0012. Tweets about European and African immigrants had threat scores of 0.0006 and 0.0005 respectively. There were no significant differences in levels of obscenity ($F = 2.17$, $p > 0.05$) and insult ($F = 0.65$, $p > 0.05$) between groups.

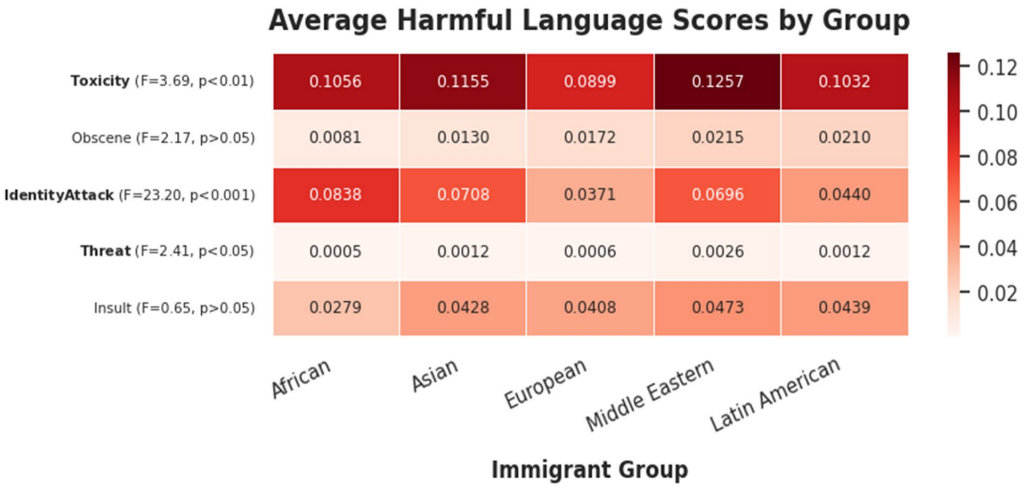

Figure 6. Average Harmful Language Scores by Group.