From Innovation to Controversy: Identifying the Impact of Generative LLMs on Social Media Discourse - An Analysis of Toxicity and Public Concerns

Mohammad Ratul Mahjabin*, Sree Rushitha Santhoshi Mamidala*, Nishat Nayla Labiba*, Mahammed Kamruzzaman*, and Lokesh Reddy Dumpa*
Department of Computer Science and Engineering
University of South Florida
{mohammadratul,sreerushitha,labiba,kamruzzaman1,lokeshreddy1}@usf.edu
*All authors contributed equally to this work

I. Abstract

Large language models(LLMs), such as ChatGPT, demonstrate state-of-the-art capabilities but have generated discussions regarding their potential impact on different industries and job markets. Social media platforms have become prime spots where people share their concerns using hateful speech. In this paper, we have investigated public attitude towards LLMs by analyzing a sample of Twitter data. We employed several NLP techniques to identify various levels of toxicity and concern in tweets about LLMs. Using GPT-4 and few-shot technique, we annotated tweets for toxicity and concern, and compared the performance of generative LLMs with RoBERTa model. Our findings reveal significant toxicity in discussions about LLMs, particularly among individuals in Computer and Mathematics; Management and Arts, Design, Entertainment, Sports, and Media Occupations, predominantly from the United States. The results indicate superior performance of RoBERTa over generative models. Overall, our study provides insights into public attitude towards LLMs and a comparison of the performance of various LLMs.[1]

***Index Terms*—Toxicity classification, Twitter, Generative LLM, Occupation Extraction, Public discourse, Concern**

II. Introduction

In the evolving landscape of artificial intelligence, LLMs such as ChatGPT (1), Llama-2 (2), and Gemini (3) have become very important technologies that greatly affect many sectors and job markets. Their capabilities extend from simple text generation to complex problem-solving which includes different types of reasoning tasks (4), and their capabilities can amaze or worry people. As these models are widely used in various tasks, the risks associated with their use also increased, as the models can provide biased responses (5) and suffer from hallucinations (6). Also, their effect on social media conversations is especially significant, often leading to an increase in negative and toxic interactions. On platforms such as Twitter, discussions about LLMs frequently incorporate toxic language. This brings up important questions about how people communicate digitally in the time of advanced AI. The worries are not just about toxicity. Many people are also concerned about privacy and losing jobs to LLMs. These fears are shared by different groups of people and professionals across occupations, showing a worldwide feeling of concern. In this paper, we delve into these critical issues by addressing several research questions:

1) After the invention of LLMs especially ChatGPT, how do people use toxic words in their conversations on social media platforms (e.g., Twitter) while discussing LLMs? How does it change over time? Which part of the world has shown more toxicity?
2) What concerns do people have about ChatGPT? Which occupations are particularly worried about ChatGPT? People from which country have shown more concern?
3) How good are generative LLMs (e.g., Mistral-7B (7), Gemini) as a toxicity classifier vs baseline models (e.g., BERT, RoBERTa)?

To address these questions, our method involves using toxicity classification on tweets related to ChatGPT to measure the overall sentiment of toxicity. Also, we examine occupational data from user profiles associated with these tweets, aiming to find connections between occupations and their expressed concerns. Through this study, we help understand the complex social dynamics formed by LLMs, pointing out areas of public worry and the effects on future AI management.

[1]Datasets and code are available at https://github.com/Labiba1507045/Social-Media-Mining-Project

III. Literature Review

Vaidya et al. (8) observed that hate speech spreads in a complicated way through social networks, looking more like a complex spread than just a simple condition. They proposed a novel model to capture the evolution of toxicity spread on Twitter by classifying users into amplifiers, attenuators, and copycats based on how they propagate toxicity through the network. Moving forward, Fan et al. (9) used advanced machine learning models like BERT to identify and sort toxic content on social media platforms. Their study provides an important tool for moderating and stopping the spread of harmful content. Coming to the people's perceptions, Koonchanok et al. (10) analyzed public attitudes towards ChatGPT on Twitter using sentiment analysis and topic modeling to get a deeper understanding of public opinion. Their findings suggest that tweets related to ChatGPT are mostly neutral to positive. Their research showed the different topics people talked about, like Cybersecurity and Education, showing how ChatGPT has influenced various areas. In a similar research, Jangjarat et al. (11) looked into how ChatGPT, as a robotic helper, is seen by people. They paid special attention to its effects in the workplace and how it interacts with different professional areas. This study showed a careful hopefulness, where users recognized both the potential improvements and the challenges brought by such technologies.

IV. Methodology

We will discuss briefly about our data and analysis in this section.

A. *Dataset Description:*

We have used the dataset **"Large Language Models: the tweets"** from Kaggle which comprises of 3.5M tweets reflecting public opinions on large language models (LLMs). This dataset includes comprehensive user data such as names, locations, descriptions, tweeted texts, account creation times, tweet times, friend counts, follower counts, etc. For our analysis, we focused on the user location, user description, tweeted text, and timestamps of the tweets.

From 3.5M tweets, we curated a smaller set containing 10k tweets. To create this dataset, we have used two benchmark datasets (12; 13) and identified 400 words commonly associated with toxicity or concern. Using those 400 words, we randomly extracted 6,000 tweets featuring toxic remarks (at least one of the 400 words from the toxicity benchmark dataset present in these data) and 4,000 tweets reflecting concerned comments (at least one of the 400 words from the emotion benchmark dataset present in these data).

B. *Dataset Preprocessing:*

Data pre-processing is a crucial step for cleaning data. In our research, we performed basic pre-processing that includes removing URLs, punctuation, special characters, stop words, numbers, and applying lemmatization. Firstly, we removed URLs as they do not contribute to our analysis. We then eliminated punctuation marks, special characters, and numbers. Following that, we applied lemmatization, a process where words are converted to their base form; for example, 'eating' changes to 'eat'. This step helps reduce the complexity of our text data and makes it uniform. All preprocessing steps are vital as they help in making the data cleaner and more suitable for further analysis tasks.

C. *Toxic Tweets Categorization and Analysis*

We have categorized the toxic comments into the most common six classes which are 'Toxic', 'Severe Toxic', 'Obscene', 'Threat', 'Insult', and 'Identity Hate'. To label the tweets into these specific groups, we have performed human annotation on 1,000 tweets. The remaining tweets were labeled by employing few-shot learning techniques with GPT-4. We have used prompt engineering for this purpose.

After classifying the tweets into these categories, we counted the number of toxic tweets for each class which answered the type of toxicity people show towards LLMs in their tweets. To identify the change in toxicity over time, we have analyzed the time of the tweets over months and counted the number of toxic tweets for each month.

D. *Concern Tweets Extraction and Analysis*

To label concerning tweets from our dataset, we again used the few-shot capabilities of GPT-4. This time, we incorporated 500 examples of human-annotated data specifically related to concerning tweets. We labeled these tweets as 'fear' or 'not fear'. Tweets that display any signs of concern are labeled as 'fear'; all others are marked as 'not fear'. Later, we created a Word Cloud (14) to visually represent the most frequently used words in tweets associated with concern.

E. *Occupation Extraction and Analysis:*

To answer the second research question, we extracted the occupations of the users from the 'user_description' and combined it with occupation-related terms obtained from the Standard Occupational Classification (SOC) of the U.S. Bureau of Labor Statistics 2018 SOC. We preprocessed the 'user_description' to remove any URLs, emojis, punctuation, stop words, user handles; converted them to lower case, expanded abbreviations and contractions and performed lemmatization using python packages like 'NLTK', 'contractions' and 'spaCy'. We then extracted the unigrams and bigrams of the 'user_description'. From the SOC system we curated lists for title to major occupation mapping, where the titles are all unigrams. Unigrams from the 'user_description' and titles were matched to map to a major occupation. To disambiguate this mapping, we further curated modifier to occupation. E.g. To disambiguate

engineers like 'Software engineer' and 'Civil Engineer' the terms 'software' and 'civil' help as modifiers. We further curated a keyword to occupation list that has keywords of major areas of occupations to improve extraction (10).

We used the dataset that has concern annotated and extracted upto 73% of the occupations of the users. We then visualized the concern frequency by occupation to identify which occupations have highest concerned tweets discussed in Section V.

F. *Location Extraction and Analysis:*

Location data from the twitter was not always available and accurate. We first cleaned 'user_location' by removing null values. We then extracted unigrams which are potential City names and checked their validity by attempting to geocode it using the Nominatim geocoder. We then extracted the country from valid cities and marked the unresolved locations as 'Unknown'.

G. *Generative LLMs vs Baseline Models as Toxicity Classifier:*

We evaluated the performance of two generative LLMs, specifically Mistral and Gemini, against two pre-trained baseline models, RoBERTa and BERT to classify the tweets into the six toxic groups. Our labeled dataset was divided into two parts where we kept 8,000 tweets(80%) for training the model and 2,000 tweets(20%) for testing the model.

For the baseline models, we employed a multi-label classification approach. We began by cleaning and pre-processing our dataset to let data-loader to fit in the model. Next, we fine-tuned the baseline models. During fine-tuning, we used the *Adam* optimizer and *Binary Cross Entropy(BCE)* loss function. After fine-tuning, we evaluated the performance of fine-tuned model on test set. To assess the effectiveness, we calculated metrics like *Accuracy, Precision, Recall, and F1-score* for each label.

We used generative LLMs, Mistral-7B (`TheBloke/Mistral-7B-Instruct-v0.1-GGUF`) checkpoint on Huggingface, and Gemini (`gemini-1.0-pro-latest`) using Google API and utilized prompt engineering to classify the 2,000 test tweets. We did not use any dataset to train the model and solely used the inferencing capabilities of these models. This approach allowed us to test how effectively the models could perform classification tasks based on pre-existing knowledge. After that, we compared their classifications with the ground truth to analyze their accuracy.

V. Results and Discussion

1) ***Research Question 1:***

a) Toxicity Categorization: From the labeled dataset, we found the result shown in Figure 1. From the result we can say that people mostly use toxic, obscene and insulting words while discussing about LLMs in their tweets.

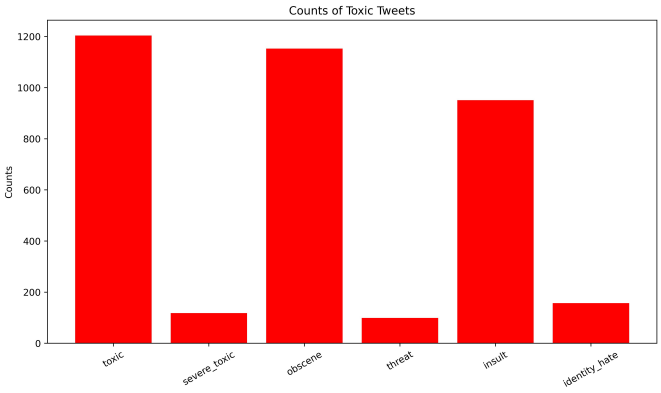

Figure 1: Categorized Toxic tweets Counts

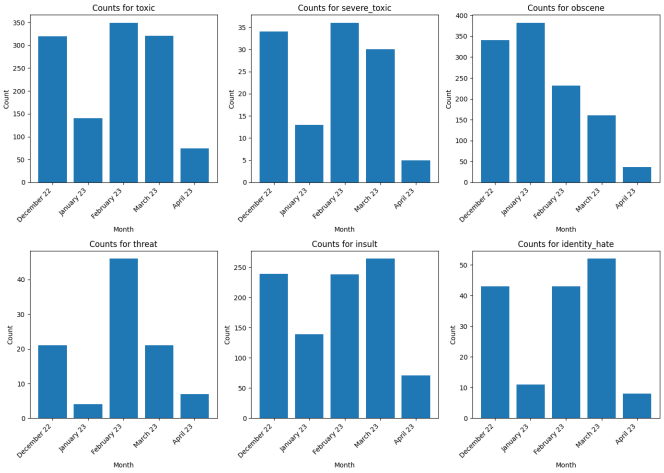

Figure 2: Monthly Analysis of Toxic Tweets

b) Change over time: The count of toxic tweets over five months period is shown in Figure 2. The result is very noisy and we couldn't draw any further conclusions.

c) Geography: After extracting the country based on users location, the count for the sub category of toxic tweets results are shown in Figure 3. The results show the countries with the highest levels of toxic count across world. The United States had the highest number of concern related tweets in our dataset, likely due to it having the largest number of Twitter users.

2) ***Research Question 2:***

a) What concerns?: We presented the most used words when people are concerned about LLMs in Figure 4. From the Figure 4, we can see that people are mainly worried about their employment, and the spread of false information through large models. They also worried about the issues related to the security, privacy, and bias of these models.

b) Occupations with most concern: After extracting the occupations of the user we visualized count of concerning tweets grouped by Occupations and the results are in Figure 5.

The Computer and Mathematical Occupations; Arts, Design, Entertainment, Sports, and Media Occupations

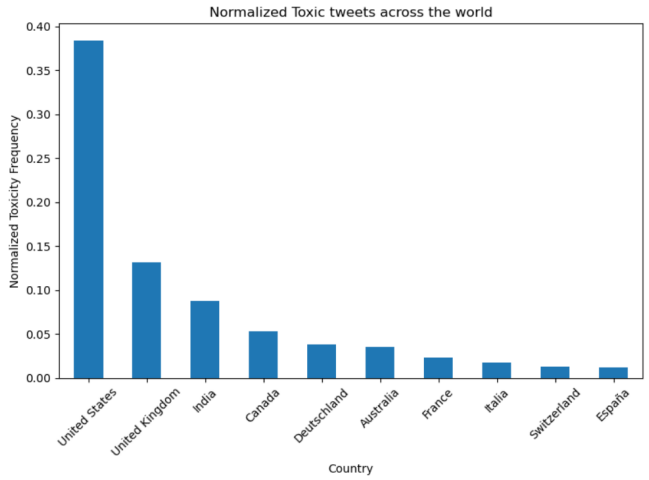

Figure 3: Graph showing the normalized count of toxic tweets by Country

Figure 4: Word Cloud representation of the most used words when people are concerned

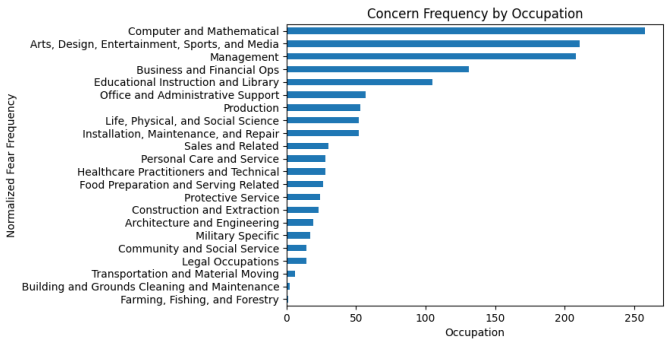

Figure 5: Graph showing the count of concerned tweets by Occupation

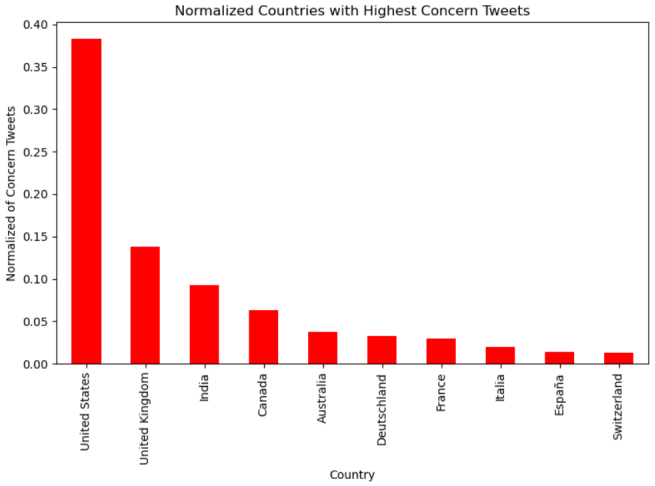

Figure 6: Graph showing the normalized count of concerned tweets by Country

and Management Occupations have high frequency of concerned tweets. This combined with the word cloud we could deduce that there are concerns about privacy and the possibility of loosing jobs to LLMs and AI. In 2023, American actors' union, Screen Actors Guild - American Federation of Television and Radio Artists (SAG-AFTRA) went on a strike, one of the reasons being studio usage of artificial intelligence (AI) to scan actors' faces to generate performances digitally which the artists were against.

c) Geography: After extracting the country based on users location the count for the concerns of toxic tweets results are shown in Figure 6. The United States had the highest number of concern related tweets in our dataset, likely due to it having the largest number of Twitter users.

3) ***Research Question 3:*** In this section, we present the result of RQ3. We include average accuracy, precision, recall, F1-score for the models [I]. Label-wise result can be found in appendix [II]. From the table, we can conclude that RoBERTa relatively performed better than all other models as expected.

VI. Conclusions

Our study provides a detailed analysis of how people use toxic and concerning language in their discussions about large language models (LLMs). By categorizing these comments by location and occupation, we gained valuable insights into the demographics and the specific occupations most actively discussing LLMs. The United States emerged as the predominant country for toxic commentary, likely reflecting the higher proportion of Twitter users from this region. To remove this geographic bias, future studies could benefit from a more diverse dataset drawn from various social media platforms.

Additionally, the smaller size of our dataset may have contributed to the variability in the observed toxicity trends over time. Regarding model performance, while RoBERTa outperformed the generative LLMs overall, Gemini showed commendable inferencing capabilities among its peers, suggesting potential areas for further refinement.

Overall, our research successfully addressed the three

Table I: Average Score

Model	Accuracy	Precision	Recall	F1-Score
Mistral 7B	0.79	0.08	0.24	0.10
Gemini	0.91	0.28	0.40	0.31
BERT	0.97	0.85	0.80	0.82
RoBERTa	0.97	0.83	0.81	0.82

posed questions, shedding light on the patterns of toxic language use in LLM-related discussions and highlighting the effectiveness of different models in classifying such content. Future research could expand upon these findings by incorporating a larger, more varied dataset to enhance the robustness and generalizability of the results.

VII. Contribution

- Data pre-processing - Mahammed Kamruzzaman
- Human Annotation - Everyone
- Few-shot learning - Mahammed Kamruzzaman, Sree Rushitha Santhoshi Mamidala, Nishat Nayla Labiba
- RQ1 - Nishat Nayla Labiba, Mahammed Kamruzzaman, Lokesh Reddy Dumpa
- RQ2 - Mohammad Ratul Mahjabin, Mahammed Kamruzzaman, Nishat Nayla Labiba
- RQ3 - Sree Rushitha Santhoshi Mamidala, Mohammad Ratul Mahjabin, Lokesh Reddy Dumpa

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[3] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[4] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[5] M. Kamruzzaman, M. M. I. Shovon, and G. L. Kim, "Investigating subtler biases in llms: Ageism, beauty, institutional, and nationality bias in generative models," *arXiv preprint arXiv:2309.08902*, 2023.

[6] V. Rawte, A. Sheth, and A. Das, "A survey of hallucination in large foundation models," *arXiv preprint arXiv:2309.05922*, 2023.

[7] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[8] A. Vaidya, S. Nagar, and A. A. Nanavati, "Analysing the spread of toxicity on twitter," in *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, pp. 118–126, 2024.

[9] H. Fan, W. Du, A. Dahou, A. A. Ewees, D. Yousri, M. A. Elaziz, A. H. Elsheikh, L. Abualigah, and M. A. Al-qaness, "Social media toxicity classification using deep learning: real-world application uk brexit," *Electronics*, vol. 10, no. 11, p. 1332, 2021.

[10] R. Koonchanok, Y. Pan, and H. Jang, "Public attitudes toward chatgpt on twitter: Sentiments, topics, and occupations," 2024.

[11] K. Jangjarat, T. Kraiwanit, P. Limna, and R. Sonsuphap, "Public perceptions towards chatgpt a s the robo-assistant," *Jangjarat, K., Kraiwanit, T., Limna, P., & Sonsuphap*, 2023.

[12] J. E. L. D. M. M. n. W. C. cjadams, Jeffrey Sorensen, "Toxic comment classification challenge," 2017.

[13] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547*, 2020.

[14] L. Oesper, D. Merico, R. Isserlin, and G. D. Bader, "Wordcloud: a cytoscape plugin to create a visual semantic summary of networks," *Source code for biology and medicine*, vol. 6, pp. 1–4, 2011.

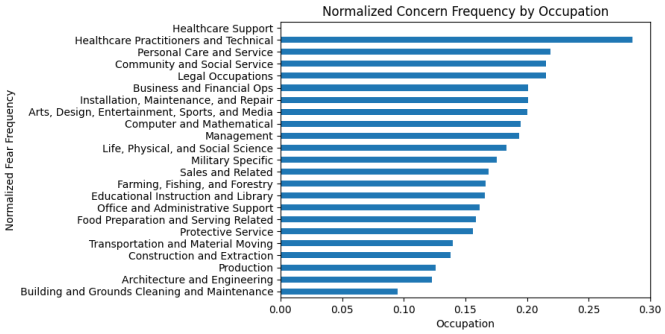

Figure 7: Graph showing the normalized count of concerned tweets by Occupation

Appendix

a) Normalized visualization for concern by occupation: Visualized the normalized frequencies by occupation, Figure 7. However, given the distribution of the subset of data, the normalized frequencies were not very insightful to draw conclusions.

b) Additional Results for RQ3: Label-wise result for different models can be found in Table II.

c) Sub category of tweets across world: The results shows the toxicity count of different sub category across countries, were the united states have the highest count across all category's, this might be due to the more users of twitter are from united states Figure 8.

Table II: Label-wise Evaluation Metrics for Different Models

Metric	Model	Label					
		Toxic	Severe Toxic	Obscene	Threat	Insult	Identity Hate
Accuracy	Gemini	0.75	0.98	0.94	0.98	0.85	0.98
	Mistral 7B	0.65	0.78	0.82	0.82	0.78	0.89
	RoBERTa	**0.95**	**0.98**	**0.95**	**0.98**	**0.98**	**0.99**
Precision	Gemini	0.21	0.07	**0.78**	0.00	0.25	0.36
	Mistral 7B	0.10	0.01	0.21	0.02	0.10	0.04
	RoBERTa	**0.88**	**0.72**	0.75	**0.38**	**0.90**	**0.86**
Recall	Gemini	0.50	0.07	0.66	0.00	0.49	0.68
	Mistral 7B	0.30	0.25	0.18	**0.26**	0.18	0.26
	RoBERTa	**0.72**	**0.59**	**0.88**	0.24	**0.90**	**0.86**
F1	Gemini	0.30	0.07	0.72	0.00	0.49	0.68
	Mistral 7B	0.15	0.02	0.19	0.03	0.13	0.08
	RoBERTa	**0.79**	**0.65**	**0.81**	**0.29**	**0.90**	**0.86**

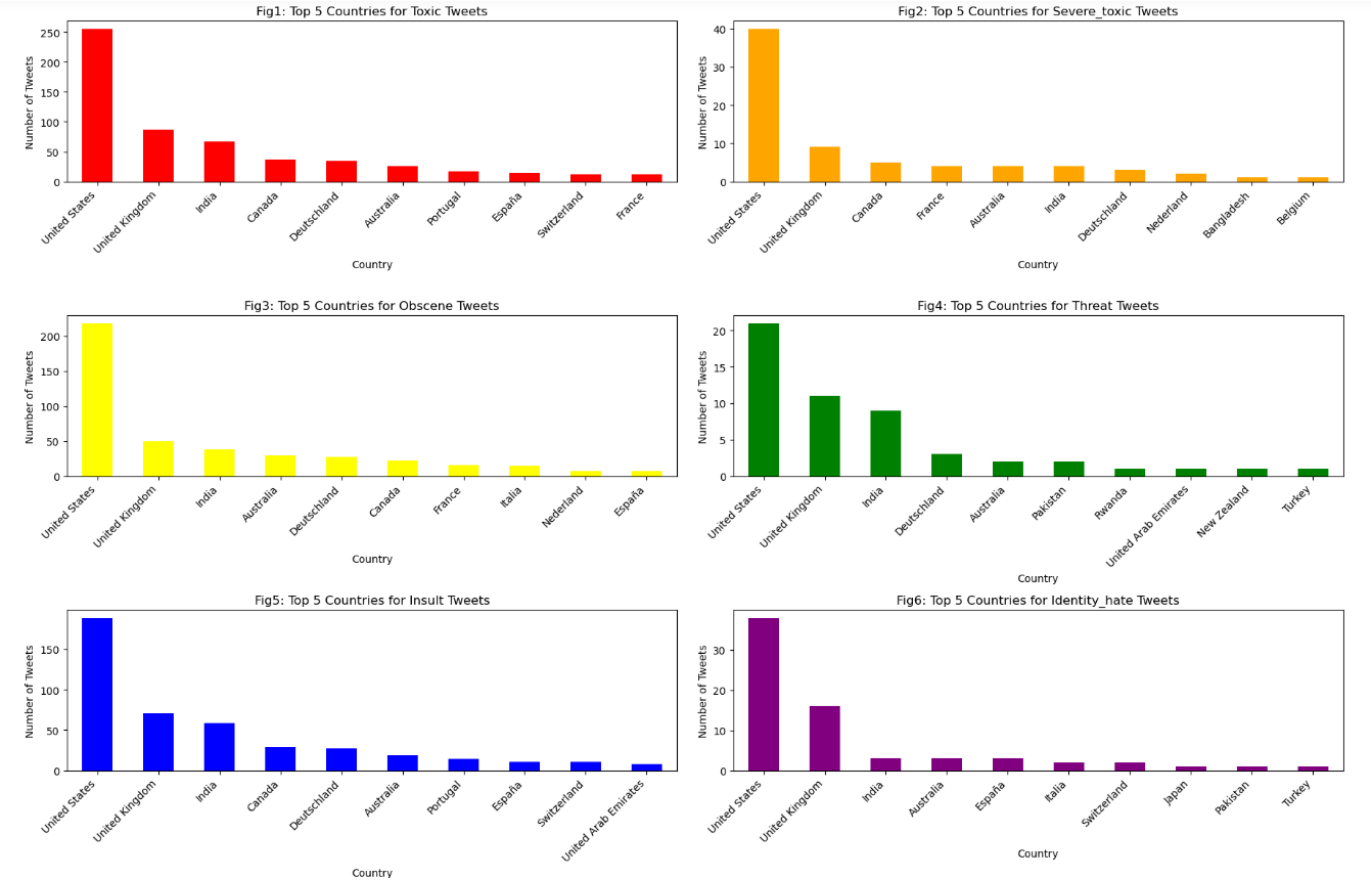

Figure 8: Graph showing the sub category of the toxic count across countries