WHY SO INFLAMMATORY? EXPLAINABILITY IN AUTOMATIC DETECTION OF INFLAMMATORY SOCIAL MEDIA USERS

Cuong Nguyen, Daniel Nkemelu, Ankit Mehta & Michael Best *
Georgia Institute of Technology
Atlanta, GA 30318, USA
{johnny.nguyen, dnkemelu, amehta318, mikeb}@gatech.edu

ABSTRACT

Hate speech and misinformation, spread over social networking services (SNS) such as Facebook and Twitter, have inflamed ethnic and political violence in countries across the globe. We argue that there is limited research on this problem within the context of the Global South and present an approach for tackling them. Prior works have shown how machine learning models built with user-level interaction features can effectively identify users who spread inflammatory content. While this technique is beneficial in low-resource language settings where linguistic resources such as ground truth data and processing capabilities are lacking, it is still unclear how these interaction features contribute to model performance. In this work, we investigate and show significant differences in interaction features between users who spread inflammatory content and others who do not, applying explainability tools to understand our trained model. We find that features with higher interaction significance (such as account age and activity count) show higher explanatory power than features with lower interaction significance (such as name length and if the user has a location on their bio). Our work extends research directions that aim to understand the nature of inflammatory content in low-resource, high-risk contexts as the growth of social media use in the Global South outstrips moderation efforts.

1 INTRODUCTION

Hate speech and disinformation spread on social networking services (SNS) continue to pose significant threats to the safety of people everywhere (UNESCO, 2021). Within the Global South, the spread of these types of problematic content has inflamed ethnic and political conflicts in several countries (the Rohingya genocide in Myanmar (Fink, 2018), anti-Muslim riots in Sri Lanka and India (Liebowitz et al., 2021), and anti-Hindu riots in Bangladesh (Hasan, 2021)). For this work, we refer to hate speech and misinformation used in these contexts as *inflammatory content* due to their ability to inflame existing tensions into real-life violence, as highlighted in the cases above.

Tackling these issues in low-resource language settings presents significant challenges due to the unavailability of linguistic resources to build effective text classifiers. Given that the signals extracted from posts and tweets are invariably linked to an entity's profile, we hypothesize that focusing on and leveraging user-level features (such as in (Ribeiro et al., 2018; ElSherief et al., 2018)) can help us understand and mitigate the spread of inflammatory content. These language-agnostic features bypass gaps in the availability and maturity of language-specific tools and resources.

In this paper, we explore an approach to detect users spreading inflammatory content using their interaction features and provide a detailed feature analysis of our classification model. We mainly focus on analyzing the explainability of classification models trained using these user-level features. Understanding model performance is required before deploying models that support tasks to moderate social media content. This engenders trust and accountability among platform owners,

*Presented at the PML4DC 2022 Workshop

moderators and users. Our case study looks at Ethiopian Twitter users and the spread of inflammatory content during its contemporary civil and ethnic conflict. As all of Ethiopia's official languages, including Amharic, Oromo, and Tigrinya, are considered low-resource languages (Hu et al., 2020), the context provides a prime example for situations where more conventional content-rich methods of analysis might fail thus motivating the methods we adopt in this work.

2 METHODS

2.1 DATA COLLECTION

This project was conducted in collaboration with the Center for Advancement of Rights and Democracy (CARD), an Ethiopian-based non-profit (ICNL, 2021). To capture social media content related to the current Ethiopian civil conflict, we first identified a set of search terms (keywords and hashtags) associated with the conflict. We selected search terms that reflect the major ethnic and political groups involved in the ongoing conflict. We then used Twitter API's Filtered Stream endpoint, which pulls approximately one percent of publicly available tweets and their metadata, to collect tweets that contained any of our search terms. We collected tweets from August 17th, 2020, until July 29th, 2021.

CARD trackers labelled a sample of the collected Twitter posts using the Aggie platform[1] (Smyth et al., 2016). For consistency and contextual relevance, we adopted the definitions for hate speech and misinformation in the Ethiopian Government's 2020 Hate Speech Proclamation (Government, 2021). We additionally referenced definitions of hate speech and misinformation provided by Twitter (Ribeiro et al., 2018). At the end of the data collection process, we had collected 154 instances of inflammatory content that the trackers flagged. Out of those 154 instances, 58 posts (38 $\%$) had 'misinformation,' and 94 posts had 'hate speech' (61 $\%$) as specific type of inflammatory content. The remaining two posts (1 $\%$) either did not have a reason behind why it is inflammatory, or the reason did not fall into one of the two categories. We excluded these posts from further analysis.

2.2 DATASET CONSTRUCTION

From the inflammatory posts found by the CARD trackers, we identified a set of inflammatory users (IUs). We defined IUs as any user who tweets or retweets one or more inflammatory posts, per the definition adopted above. We further subdivided these IUs into three subtypes based on the type of inflammatory content they spread. Hate Users (HUs) are users who tweet or retweet one or more posts flagged by trackers as hate speech. Misinfo Users (MUs) are users who tweet or retweet one or more posts flagged by trackers as misinformation, and Hate+Misinfo Users (HMUs) are users who are both Hate and Misinfo users. This strategy resulted in 589 IUs (145 HUs, 419 MUs, and 25 HMUs).

To augment the set of IUs, we included users who used three or more instances of offensive and inflammatory terms from the PeaceTech Lab lexicon(Barrach-Yousef, 2021), which we found to be highly correlated with the action of spreading inflammatory content. The lexicon was collected via surveys distributed to PeaceTech Lab's civil partners in Ethiopia and evaluated through online focus groups and expert interviews. We used the main portion of the lexicon for this particular study containing 21 terms and their linguistics variants. In the end, this resulted in a final set of 865 IUs (415 HUs, 409 MUs, 41 HMUs)

Now that we have a set of inflammatory users (IUs), we constructed a set of non-inflammatory users (NIUs). This will enable us test differences between both groups based on our hypothesized user-level interaction features. This NIU sample captured users within our collected dataset that were active on Twitter during the time period but were not socially linked to the IUs previously identified. We assigned them to the non-inflammatory group based on homophily (Ribeiro et al., 2018; Mathew et al., 2019). We only selected users with ten or more unique activities (defined as either a tweet, a retweet, or a quoted tweet), which has been identified as a threshold for activeness in previous papers (Ribeiro et al., 2018; Mathew et al., 2019). This resulted in a smaller subset of 34,227 users. Then, we performed a diffusion process based on Degroot's Learning Model of the retweet graph similarly constructed from these users to (Ribeiro et al., 2018). In the end, we selected 18,978 users

[1] https://aggie.readthedocs.io/en/latest/index.html

with a belief score of 0 (i.e., did not retweet known IUs or people who retweet known IUs) to be our set of NIUs. The remaining 14,571 users with a non-zero belief score but do not belong to the IU group will be our set of borderline users (BUs).

2.3 Characterization And Analysis Of Inflammatory Users

We aim to identify and quantify the differences in profile metadata, usage patterns, and network centrality between NIUs, BUs, and IUs. To do this, we constructed a set of features that have been utilized in literature to identify differences between groups of Twitter users. Using the user feature classification framework presented in (Volkova & Bell, 2017), we divide these features into 3 categories: profile features (e.g *following* and *follower*), syntactic and stylistic features (such as *avg_mention*, *lex_diversity*), network features (e.g *eigencentrality*).

We used the Kruskal-Wallis test, alongside Dunn's posthoc test, to pinpoint where and how pairs of groups are significantly different for each feature. In addition to statistical significance, we are also interested in measuring the practical significance of any pairwise differences as they would inform how useful these features would be for building machine learning classifiers. To do so, we calculated Cliff's δ for each feature and group pairs. We report the results of these statistical tests in table 1.

2.4 Classification of Inflammatory Users

Building on our hypothesis of significant differences between IUs and NIUs, we examine the feasibility of classifying users belonging to these two groups with traditional ML models and learning over graphs models that exploit Twitter's retweet network. For traditional ML models, we use Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), CatBoost (CB), and XGBoost (XGB) models. For learning over graphs models, we use GraphSAGE (GS) due to its proven effectiveness in node classification tasks (Hamilton et al., 2017). For each user, we constructed a feature array of dimension 71 from these two categories of features: User features such as the number of followers, the number of activities, description length (21), and topic features, where each feature represents the prevalence of a given topic within the user's tweets (50). For GraphSage, we converted this user-feature matrix into a directed graph, with each node having a 71-dimensional vector representing the user.

For both traditional ML and learning-over-graph methods, we randomly selected 80 percent of the data to be the training data and reserved the remaining 20 percent to be used as testing data. The randomized training and testing data were stratified so that for both sets, the ratio between IUs and NIUs was roughly identical to one another. We performed the classification process over ten random train-test stratified splits with five-fold cross-validation to tune the parameters of the classifiers. We use accuracy, macro-weighted precision, recall, and F1-score as the evaluation measures. To determine how the features contributed to the NIU/IU classification, we calculate their average SHAP (SHapley Additive exPlanations) values (on a logit scale) across all test splits and visualize these values with a beeswarm plot. SHAP uses a game-theoretic approach to explain the output of a machine learning model (Lundberg & Lee, 2017).

3 Results

3.1 Statistical Results

We now present a synthesis of our statistical analysis results. See complete statistical results in table 1[2].

3.1.1 Inflammatory Users (IUs) Are Influential

The median IU has significantly higher *following* and *follower* compared to the median borderline user (BU) ($z = -36.83, p < 0.001$), and even more so comparing to the median non-inflammatory user (NIU) ($z = -19.33, p < 0.001$). At the median, IUs has three times the number of followings

[2]p-values are from Dunn's post-hoc tests

Feature Name	NIU-BU			NIU-IU			BU-IU		
	z-score	p	Cliff's δ	z-score	p	Cliff's δ	z-score	P-value	Cliff's δ
following	-36.83	< 0.001	-0.23	-19.33	< 0.001	-0.39	-7.61	< 0.001	-0.15
eigencentrality	7.45	< 0.001	0.048	-12.74	< 0.001	-0.27	-15.00	< 0.001	-0.28
follower	-52.05	< 0.001	-0.33	-27.75	< 0.001	-0.57	-11.19	< 0.001	-0.21
description_length	-39.38	< 0.001	-0.24	-15.21	< 0.001	-0.29	-2.72	0.01	-0.053
account_age	-27.79	< 0.001	-0.18	-16.64	< 0.001	-0.33	-7.78	< 0.001	-0.16
maxdate_ratio	30.01	< 0.001	0.19	21.15	< 0.001	0.43	11.56	< 0.001	0.23
avg_mention	3.93	< 0.001	0.024	11.39	< 0.001	0.239	10.08	< 0.001	0.19
avg_hashtag	0.45	0.98	0.0025	8.93	< 0.001	0.18	8.73	< 0.001	0.17
follower_following_ratio	-43.83	< 0.001	-0.28	-24.51	< 0.001	-0.50	-10.56	< 0.001	-0.20
name_length	10.84	< 0.001	0.069	3.87	< 0.001	0.077	0.44	0.99	0.01
avg_tweet_length	32.21	< 0.001	0.20	15.65	< 0.001	0.31	5.40	< 0.001	0.11
avg_url	10.28	< 0.001	0.065	-5.14	< 0.001	-0.10	-8.34	< 0.001	-0.17
Bot_score	20.71	< 0.001	0.13	2.72	0.001	0.061	-3.82	< 0.001	-0.086
activity_count	-18.67	< 0.001	-0.12	-20.39	< 0.001	-0.398	-14.38	< 0.001	-0.305
lex_diversity	3.43	< 0.001	0.022	-13.22	< 0.001	-0.27	-14.31	< 0.001	-0.28

Table 1: **z-score, p-value and effect size (using Cliff's δ) from the Dunn's post-hoc test (with Bonferroni correction) between pairs of groups. Here, yellow cells represents small effect size, orange represents medium effect size, and red cells represents large effect size in accordance to the reference scale mentioned above**

Method	Precision	Recall	F1	Accuracy
Support Vector Machine (SVM)	0.478	0.500	0.489	0.956
Random Forest (RF)	0.930	0.753	0.816	0.975
Logistic Regression (LR)	0.804	0.579	0.619	0.959
XGBoost (XGB)	0.929	**0.767**	**0.827**	**0.976**
CatBoost (CB)	**0.957**	0.705	0.779	0.973
GraphSage (GS)	0.837	0.515	0.519	0.959

Table 2: **Evaluation metrics for the task of predicting if a user is a IU or a NIU. The values reported are the means over the 10 random train-test splits for each of the evaluation metric (macro-weighted)**

and fourteen times the number of followers compared to NIUs. The size of this disparity is further backed up by the Cliff's δ for the NIU-IU pair from table 1, which indicates medium effect size for *following* (δ = -0.392) and large effect size for *follower* (δ = -0.57). In addition, we see that IUs are more influential within the retweet network compared to NIUs. They have significantly higher *eigencentrality* than the other groups, scoring twice as much as NIUs (z = -12.74, p < 0.001 δ = -0.275) on median.

3.1.2 Inflammatory Users Are Consistently Active

We investigated differences in activity patterns between the groups. Not only did we calculate the raw number of activities per user, but we also calculated the ratio of their peak daily activity over their total number of unique activities or *maxdate_ratio*. We propose this feature as a simple proxy to users' continual engagement with topics relating to the current Ethiopian civil conflict. Users with a high *maxdate_ratio* concentrate most of their activities regarding the conflict on a single date. They are thus indicative of either fleeting engagement with the topic or bot-like activity (i.e., spamming tweets/retweets over a short period). As shown in table 1, we found that IUs have significantly higher *activity_count* and lower maxdate_ratio (z = 21.15, p < 0.001, δ = 0.43), with five times as many unique activities and half the maxdate_ratio compared to a NIU on median.

3.2 Classification Results

We used the trained model to predict whether users in the test set belonged to the IU or the NIU group. The results were aggregated from the 10 train-test splits, and the means of aforementioned evaluation measures are reported in Table 2.

We got the best performance from the XGBoost model using topic and user features, achieving

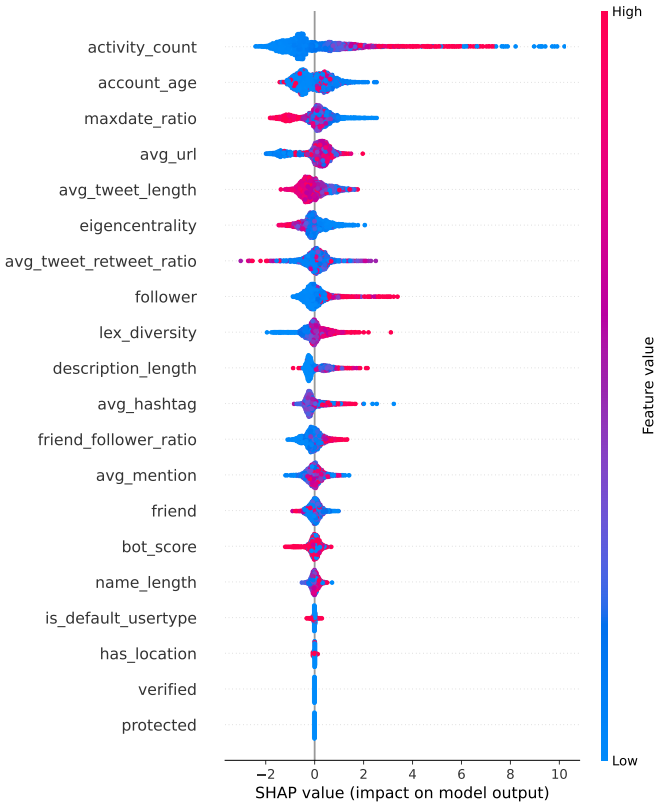

Figure 1: **Beehive plot representing SHAP values for user-level features generated on the test-set by the XGBoost model. Each dot represents an user with the test set. The x-axis represents the SHAP value the model assigned to that feature given the user classification. Positive values indicate increased likelihood of IU prediction and negative values indicate increased likelihood of NIU prediction (on a logit scale)**

an F1 of 0.827. Classifiers such as Random Forest and CatBoost also performed comparably well. The standard deviation for the evaluation metrics ranges from 0.001 to 0.02, from which we conclude that our results are relatively consistent across train-test splits. Learning-over-graph methods such as GraphSage did not perform as well as traditional ML methods (0.519 vs. 0.827 for XGBoost). We attribute this to the class imbalance (1:20) between IUs and NIUs, which seemed to affect learning-over-graph methods compared to traditional ML methods. We also notice that the macro-averaged precision is higher across the board than macro-averaged recall, signaling that our models were accurate yet conservative when assigning users into the IU group. Furthermore, an analysis of the beeswarm plot in Figure 1 reveals that features with high practical significance (such as activity_count, follower and maxdate_ratio) also have high SHAP values in the appropriate direction. In contrast, features with low practical significance (such as has_location, name_length) have SHAP values closely bunched around 0 (i.e little impact on model output).

3.3 Conclusion

This paper presents an approach to characterize and detect inflammatory content on Twitter using user-level interaction metrics. We argue that this method is particularly effective for contexts in the Global South where natural language ground truth data and processing resources are scarce. We find that inflammatory users - users that have been identified to share hate speech and disinformation posts - show significant distinctions from non-inflammatory users on key interaction metrics. They are more influential, active, and non-botlike in their interaction. They also tend to post more diverse content within the discussion surrounding the conflict. Extending our statistical analyses of group differences with SHAP analysis of our XGBoost model, we found that the best model trained

to distinguish IUs from NIUs utilized features with high practical significance from the statistical analysis. This work extends research directions that aim to understand the nature of inflammatory content in low-resource contexts as the growth of social media use in the Global South overwhelms platform moderation.

REFERENCES

N. Barrach-Yousef. *Ethiopia hate speech lexicon: PeaceTech lab*. PeaceTech Lab — Putting the Right Tools in the Right Hands to Build Peace, 2021.

M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *In Proceedings of the International AAAI Conference on Web and Social Media*. AAAI, June 2018.

C. Fink. Dangerous speech, anti-muslim violence, and facebook in myanmar. *Journal of International Affairs*, 71(1):5, 2018.

Ethiopian Government. (2020, march 23). hate speech and disinformation prevention and suppression proclamation. *FEDERAL NEGARIT GAZETTE OF THE FEDERAL DEMOCRATIC REPUBLIC OF ETHIOPIA*, 12, January 2021.

W. L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1025–1035, December 2017.

M. Hasan. Minorities under attack in bangladesh. *The Interpreter*, November 2021.

J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *In Proceeedings International Conference on Machine Learning . PMLR*, pp. 4411–4421, 2020.

ICNL. 2021. URL `https://www.icnl.org/about-us/bio/befekadu-hailu`.

J. Liebowitz, G. Macdonald, S. Vivek, and V. Sanjendra. The digitalization of hate speech in south and southeast asia: Conflict-mitigation approaches. *Georgetown Journal of International Affairs.*, April 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pp. 173–182, 6 2019.

M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. Almeida, and W. Meira Jr. Characterizing and detecting hateful users on twitter. In *Twelfth international AAAI conference on web and social media*, Hune 2018.

T. N. Smyth, A. Meng, A. Moreno, M. L. Best, and E. W. Zegura. Lessons in social election monitoring. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*. ACM, 2016.

UNESCO. Addressing hate speech on social media: contemporary challenges. 2021.

S. Volkova and E. Bell. Identifying effective signals to predict deleted and suspended accounts on twitter across languages. In *Proceedings of the International AAAI Conference on Web and Social Media*, May 2017.