AbuseAnalyzer: Abuse Detection, Severity and Target Prediction for Gab Posts

Mohit Chandra[1], Ashwin Pathak[2], Eesha Dutta[3], Paryul Jain[4], Manish Gupta[5], Manish Shrivastava[6], Ponnurangam Kumaraguru[7]
International Institute of Information Technology, Hyderabad [1,2,3,4,5,6] Indraprastha Institute of Information Technology, Delhi[7] , Microsoft[5]
Contact: mohit.chandra@research.iiit.ac.in

Motivation

•Social Media platforms have become an indispensable part of our lives but at the same time there are growing concerns over the propagation of hateful content.

•Prioritization of certain cases of online abuse can be done if severity of abuse can be automatically assessed. Moreover, identifying the target of abuse can lead us to estimate the potential impact set.

Our Contribution

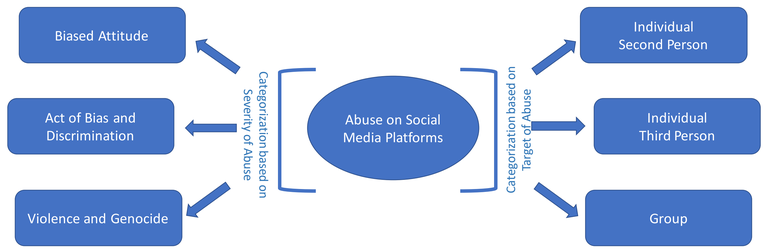
•We contribute an abuse analysis dataset comprising 7,601 Gab posts with finer classification labels associated with presence, severity and target of abuse.

•We experiment with traditional machine learning (ML) classifiers as well as Deep Learning based methods on our dataset and propose a BERT based classifier.

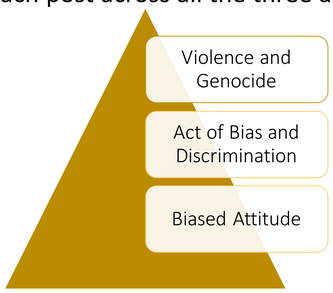
• The proposed classifier gives as accuracy values of ~80% for abuse presence, ~82% for abuse target prediction, and ~65% for abuse severity prediction

Abuse Severity and Targets

- In this work we look at the posts collected from the social media platform GAB. We categorize the abusive content based on severity and the target of abuse.



- The severity-based categorization follows the concept of pyramid of hate (based on life-threatening consequences).
- Target-based categorization differentiates whether the abuse is directed to the person being mentioned (Individual Second Person) or directed to a third person or to a particular group.
- We introduce 3 abuse prediction tasks: 1) Presence of Abuse, 2) Abuse Severity Prediction, 3) Abuse Target Prediction.
- We annotate each post across all the three abuse prediction tasks (Presence, Severity and Target)



Severity ↓ Target →	Individual Second P.	Individual Third P.	Group	Total
Biased Attitude	226	650	954	1830
Act of Bias and Discrimination	129	543	1135	1807
Violence and Genocide	34	137	312	483
Total	389	1330	2401	4120

Table 1: Distribution of posts across various abuse severity and abuse target classes.

Experiments and Results

•We experiment with multiple statistical ML methods (Support Vector Machines (SVM), XGBoost and Logistic Regression (LR)) using TF-IDF features along with two Deep Learning based models: (1) BERT based classifier using transfer learning and (2) GloVe-based LSTMs (referred as GloVe+LSTM).

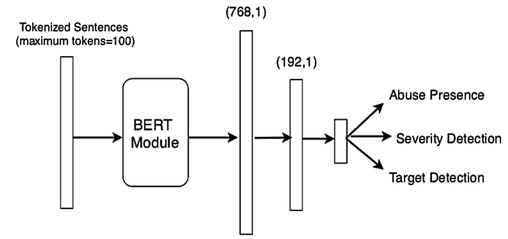


Figure 1: Architecture for AbuseAnalyzer text classifier (BERT)

	Presence		Target prediction		Severity prediction	
Classifier	Macro F1	Micro F1/Acc	Macro F1	Micro F1/Acc	Macro F1	Micro F1/Acc
SVM	0.7277 ± 0.0112	0.7279 ± 0.0113	0.7085 ± 0.0207	0.7619 ± 0.0120	0.5787 ± 0.0211	0.6238 ± 0.0236
XGBoost	0.7157 ± 0.0097	0.7165 ± 0.0096	0.6750 ± 0.0236	0.7405 ± 0.0126	0.5296 ± 0.0141	0.6238 ± 0.0084
LR	0.7235 ± 0.0135	0.7239 ± 0.0135	0.6961 ± 0.0185	0.7558 ± 0.0094	0.5674 ± 0.0132	0.6201 ± 0.0168
BERT	0.7985 ± 0.0110	0.8015 ± 0.0105	0.7893 ± 0.0104	0.8201 ± 0.0086	0.6244 ± 0.0465	0.6500 ± 0.0443
GloVe+LSTM	0.5261 ± 0.2365	0.6396 ± 0.1332	0.4009 ± 0.0324	0.6097 ± 0.0097	0.4253 ± 0.0480	0.4726 ± 0.0150

Table 2: Results for Presence, Target and Severity prediction across multiple classifiers.

•We present the confusion matrices for abuse target and severity prediction tasks in Tables 3 and 4 respectively. The entries denote the sum of examples in the 5-fold cross validation.

		Predicted		
		Second-Person	Third-Person	Group
Actual	Second-Person	319	34	36
	Third-Person	61	1078	191
	Group	111	308	1982

Table 3: Confusion matrix for Abuse Target prediction

		Predicted		
		Biased Attitude	Act of Bias and Discrimination	Violence and Genocide
Actual	Biased Attitude	1252	386	192
	Act of Bias and Discrimination	503	1104	200
	Violence and Genocide	98	63	322

Table 4: Confusion matrix for Abuse Severity prediction

Future Work

- An interesting direction of work can be related to context-based abuse detection especially with post and replies threads.
- Another direction can be to annotate the multimodal data using the presented annotation scheme and use it for the task of abuse detection.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

precog @ IIIT-D