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]
{mohit.chandra, eesha.dutta, paryul.jain}@research.iiit.ac.in
ashwin.pathak@alumni.iiit.ac.in, manish.gupta@iiit.ac.in
m.shrivastava@iiit.ac.in, pk@iiitd.ac.in

Abstract

While extensive popularity of online social media platforms has made information dissemination faster, it has also resulted in widespread online abuse of different types like hate speech, offensive language, sexist and racist opinions, etc. Detection and curtailment of such abusive content is critical for avoiding its psychological impact on victim communities, and thereby preventing hate crimes. Previous works have focused on classifying user posts into various forms of abusive behavior. But there has hardly been any focus on estimating the severity of abuse and the target. In this paper, we present a first of the kind dataset with 7,601 posts from Gab[1] which looks at online abuse from the perspective of presence of abuse, severity and target of abusive behavior. We also propose a system to address these tasks, obtaining an accuracy of $\sim$80% for abuse presence, $\sim$82% for abuse target prediction, and $\sim$65% for abuse severity prediction.

1 Introduction

In recent times, Online Social Media (OSM) has become an indispensable part of our lives. Not only these websites connect billions of people around the world, but they also serve as a platform for expressing opinions and sharing information quickly. However, recently OSM platforms have been a subject for criticism over the propagation of fake (Shu et al., 2017) and hateful content (Fortuna and Nunes, 2018). Such cases of online abuse have also translated into real world hate crimes.[2]

Abuse in social media is spread across a wide spectrum from mild expressions of attitudes and beliefs to strong violent threats. Inspired by hate theories from Anti-Defamation League (ADL)[3], we broadly classify forms of abuse as 'Biased Attitude, 'Act of Bias and Discrimination' and 'Violence and Genocide'. Moreover, abusive content could be targeted at specific individuals (e.g., a politician, a celebrity, etc.) or particular groups (a country, LGBTQ+, a religion, gender, an organization, etc.). Detection of such abusive content is critical for avoiding its psychological impact on victim communities, and thereby preventing hate crimes. Prioritization of particular abuse cases can be done if severity of abuse can be automatically assessed. Further, identifying if the abuse target is a person or a large group is critical to predict potential impact set and thereby predict if it could lead to real world crimes along with its scale. Hence, in this paper, we propose three abuse prediction tasks: prediction of abuse presence, abuse severity prediction and abuse target prediction.

Since traditional OSM websites are reasonably moderated, finding broadly abusive content is possible. But finding abusive behaviour of differing severity is a 'needle in a haystack' kind of challenge. In contrast to the other OSM, Gab is relatively unexplored and presents a wider spectrum of online abusive behaviour due to its liberal moderation policy (Zannettou et al., 2018). Hence, we gathered a dataset from Gab and contribute the labeled posts to the community in the hope of promoting deeper research on abusive content analysis. Gab is an alt-right social media website launched in 2016, which has seen a

*The two authors contributed equally.

[1] https://Gab.com/

[2] https://www.justice.gov/hatecrimes/hate-crimes-case-examples

[3] https://www.adl.org/

significant rise in the number of registered users to 1,000,000 users along with a daily web traffic of 5.1 million visits per day by the end of July 2019.[4]

Our key contributions in this paper are as follows:

- We contribute an abuse analysis dataset comprising 7,601 Gab posts with finer classification labels associated with presence, severity and target of abuse. The code and dataset are publicly available here[5].
- We experiment with traditional machine learning (ML) classifiers with TF-IDF features, for the three abuse prediction tasks. We also experiment with two deep learning (DL) based methods. Our best method leads to high accuracy values of $\sim$80% for abuse presence, $\sim$82% for abuse target prediction, and $\sim$65% for abuse severity prediction.

Disclaimer: This paper contains examples of hate content used only for illustrative purposes, reader discretion is advised.

2 Related Work

Several past works have explored different kinds of online abuse (like racism, sexism etc.) on traditionally studied platforms like Twitter (Kwok and Wang, 2013; Waseem and Hovy, 2016; Davidson et al., 2017; ElSherief et al., 2018) and on some newer web communities like 4chan and Whisper (Hine et al., 2017; Silva et al., 2016). But web communities differ from each other through subtleties in language and demographic differences. Gab poses an altogether different challenge as it differs from older web groups primarily in its use of online communities to congregate, organize, and disseminate information in weaponized form (Marwick and Lewis, 2017). Some previous papers (Zannettou et al., 2018; Lima et al., 2018; Mathew et al., 2019; Finkelstein et al., 2018) have presented basic statistical analysis of data extracted from Gab. Recently, Qian et al. (2019) presented a dataset of 33,776 posts on Gab annotated on binary labels hate/non-hate. While some papers have focused on racism versus sexism (Badjatiya et al., 2017), others have focused on sarcasm, cyber-bullying etc. (Founta et al., 2019). Initial works in this area focused on feature engineering based methods. With the emergence of deep learning, most of the recent works (Founta et al., 2019; Serrà et al., 2017; Park and Fung, 2017) have relied on deep learning techniques for abuse detection. To the best of our knowledge, there is no publicly available corpus or prediction system which focuses on fine-grained abusive language classification across all three tasks: prediction of abuse presence, abuse severity prediction and abuse target prediction.

3 Abuse Severity and Targets

3.1 Abuse Severity

Anti-Defamation League[6] presents a pyramid of hate from a sociology perspective (also supported by Sandu and Lyamouri-Bajja (2018)). Although the behaviors at each level of the hate pyramid negatively impacts individuals and groups, as one moves up the pyramid, the behaviors have more life-threatening consequences. Given the five levels in the pyramid, initially we considered five classes of abuse severity. However, we found that some of the classes were infrequent. Moreover, some classes were very similar. Hence, we finally recognized three clear classes of abuse as follows.

Biased Attitude: Contains posts related to trolling, accusations, sarcasm and insensitive remarks. These posts project minor hate and negation of intimacy towards individual/group through jokes or blaming the target for some mishappening. E.g., "At lease my choice of president waits till they are out of diapers.", "Good news. No wonder they tried to f*ck her nomination over hard."

Act of Bias and Discrimination: Consists of posts lying in the category of sexism, racism, xenophobia, homophobia and other forms of abuse which are based on some presumption or stereotype without well founded arguments. Additionally we consider aggression based on the ideologies (leftist, rightist,

[4] https://www.similarweb.com/website/gab.com

[5] https://github.com/mohit3011/AbuseAnalyzer

[6] https://www.adl.org/sites/default/files/documents/pyramid-of-hate.pdf

feminists etc.). This class also covers posts which induce a sense of inferiority complex on the basis of target's physical ('ugly', 'deformed' etc.), mental ('stupid', 'cretin' etc.) or moral deficiency ('sl*t', 'fraud' etc.). E.g., "Jews are tragic for the world.", "you're are a real f*cking braindead piece of trash b*tchboy."

Violence and Genocide: Contains abusive behaviours like violent threat, intimidation and extremism. These posts have statements of intent to inflict harm that would result in serious and lasting bodily impact, where an individual could die or be significantly injured. It reflects intense anger/fear towards the targeted individual/group and depicts them as a threat to society. In many cases, the posts refer to adverse activities like deportations etc. E.g., "antiwhite whites deserve public hangings", "I know what is going on! I f*ck his mammy and I hurt her. Sorry mam!"

3.2 Abuse Targets

A comment targeted towards a particular community impacts a larger audience as compared to one targeted towards a particular individual. Hence, it is important to predict target of the abusive post as one of these three classes.

Individual (Second-Person): Targets the person being mentioned in the post. Generally, there is usage of terms like '@username', 'you' and 'your' to refer the target. E.g., "No, but I do realize that you're full of sh*t and know it.", "@username is serving a purpose or just a load of hot air."

Individual (Third-Person): Target a third person. Usually, these posts use terms like 'he', 'she', etc. or many a times the posts mention the name/username of the target. E.g., "His predatory sexual behavior is still evident.", "Another pedophile circles the wagons."

Group: Target a group/organization based on ideologies, race, gender, religion, work industry or some other basis. Such posts contain terms like 'you all', 'they' or many a times refers to a group in an indirect manner. E.g., "We have some shit stirrers afoot today. Ignore them", "Why not set dead muslims on the curb in a trash bag?"

4 AbuseAnalyzer Dataset and Results

Our dataset contains 7601 Gab posts classified on three different aspects: abuse presence or not, abuse severity and abuse target. Of the 4120 abusive posts, distribution based on severity is – 'Biased Attitude': 1830, 'Act of Bias and Discrimination': 1807, and 'Violence and Genocide': 483. For the target classes – 389 are in 'Individual (Second-Person)', 1330 in 'Individual (Third-Person)', and 2401 in the 'Group' class. The code and dataset are publicly available here[7].

Data Extraction and Pre-processing: We obtained a collection of 8.4 million Gab posts from `http://files.pushshift.io/gab/` for a period of 4 months from Jul to Oct 2018. We used a high precision lexicon which consists of racial, sexist, xenophobic, extremist and other derogatory terminologies aggregated from multiple source.We used this to filter 7601 posts written in English for the annotation process. While we made efforts to strike a balance between abusive versus non-abusive posts, we made no efforts to maintain balance within abuse severity or abuse target classes.

Annotation Procedure: Four annotators with fluent English skills were provided clear guidelines (refined iteratively) for annotating the posts across all the three abuse prediction tasks. In case a post could belong to more than one severity classes, annotators were asked to mark the higher severity class (based on life-threatening consequences), to avoid multi-labels. Each example was annotated by exactly 3 annotators and all the disagreements were resolved after involving all the annotators. As a measure of inter-annotator agreement, we observed Cohen's Kappa Score (Cohen, 1960) as (1) 0.719 for presence/absence of abuse, (2) 0.720 for presence+target, and (3) 0.683 for presence+severity classification. In each case the Kappa score is near 0.7 which is a very good agreement among the annotators.

Dataset Statistics and Analysis: Table 1 shows the distribution of the 'Target' labels among each of the 'Severity' classes. We observe that majority of the abusive posts are against the 'Group' class, specifically for 'Act of Bias and Discrimination' class which is intuitive since this category covers the topics of racism, sexism etc.

[7] `https://github.com/mohit3011/AbuseAnalyzer`

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.

Table 2 shows popular unigrams and bigrams for various severity and target classes. We observe that: (1) Community related words and bigrams like 'jew', 'muslim', etc. are quite frequent for 'Act of Bias and Discrimination' class which is in line with the nature of posts on Gab. (2) violent ngrams like 'kill', 'the holocaust' are present in the 'Violence and Genocide' class. (3) Second person pronouns like "you", "yourself", etc. are frequent in the 'Individual (Second-Person)' class. (4) Third person pronouns and bigrams like "he", "she", "hes a", etc. are frequent in the 'Individual (Third-Person)' class. (5) Multiplicity indicating ngrams like "these people", "them", etc. are popular in the 'Group' class.

		Unigrams	Bigrams
Severity	Biased Attitude	lol, white, f*ck, against, killed, twitter, government, usermention, america	you are, they are, trying to, illegal alien, going to, to do, to get
	Act of Bias and Discrimination	jews, white, black, muslims, stupid, islam, b*tch, k*ke, evil, rape	you are, the jews, of sh*t, jews are, white people, muslims are, a race, white people, a n*gger, a k*ke
	Violence and Genocide	f*ck, kill, hell, die, b*tch, lol, fight, muslims, white, war	to hell, the f*ck, to kill, the b*tch, rid of, kill all, get rid, f*ck the, to die, the holocaust
Target	Second person	you, your, youre, f*ck, stupid, sh*t, jew, b*tch, yourself, @username	you are, if you, are you, you don't, do you, youre a, you just, your own
	Third person	he, her, she, his, you, this, b*tch, sh*t, trump, him, @username	she is, he is, hes a, he was, a jew, she was, he has, illegal alien
	Group	they, you, all, their, jews, them, people, f*ck, white, sh*t	they are, the jews, the left, jews are, these people, white people, they will, the US, all of, all the

Table 2: Frequent unigrams and bigrams for each of the abuse severity and abuse target classes.

Prediction Results: We experiment with multiple statistical ML methods (Support Vector Machines (SVM), XGBoost and Logistic Regression (LR)) using TF-IDF features. We also trained two Deep Learning based models: (1) Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) using transfer learning and (2) GloVe-based (Pennington et al., 2014) Long Short Term Memory (Hochreiter and Schmidhuber, 1997) networks (referred as GloVe+LSTM). With BERT, we use an additional 2-layer multi-layer Perceptron (MLP) for classification with a dropout value of 0.2.We trained both the DL networks using Adam optimizer (Kingma and Ba, 2014). Table 3 shows 5-fold cross validation accuracy (micro F1) and macro F1 for each of the methods. We observe that our BERT based model outperforms other methods with SVM being the best out of the ML models.

	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 3: AbuseAnalyzer Results for Presence, Target and Severity prediction across multiple classifiers.

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

Error Analysis: Table 6 presents the cases where AbuseAnalyzer mis-classifies the examples. We present some interesting cases for each of the three abuse prediction tasks. For the task of prediction of presence of abuse, we see that terms like 'black', 'muslims' which are prone to online abuse pose a challenge for the classifier. For example, the first post in Table 6 talks about the adoption of a girl belonging to the black community, this example is non-abusive but it is wrongly classified as abusive due

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

Table 4: Confusion matrix for Abuse Target prediction using BERT.

		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 5: Confusion matrix for Abuse Severity prediction using BERT.

to the presence of potentially racial terms. Similar is the case with the second post which reports a news of arrest of *muslim jihadists*. In example 4 in Table 6 the presence of the pronoun 'you' along with the overall tone of the post of being sarcastic confused the system to predict the target class as 'Individual (Second-Person)' where the ground truth label was 'Group' as the post conveys a racist ideology against Jews. Example 5 presents an interesting case which trolls the concerned person while making a general statement about the world, due to the presence of terms like 'evil' along with 'world', the system got confused. In example 6 the reference to the third person has been made using '@username' but later the pronoun 'you' has been used to refer to this person, this change in the way of referencing confused the system. Example 7 in Table 6 is a sexist comment on the target which blames her for making false accusation of rape. But the presence of an extremist term like *rape* made the classifier to commit error. Example 8 presents a case of an extremist post which propagates the hate in a subtle way. The post talks about killing immigrants from across the border. This phenomenon was common with other posts where the hate was expressed in a very subtle way without using any explicit terms. In example 9 we have a case of trolling, where the person posting has trolled national socialists.

Task	Post	Our Prediction	Ground Truth
Presence	Under a video about a black girl finding out shes being adopted by her white parents	Abusive	Non-Abusive
	On Friday the FBI arrested all five Muslim jihadists connected to the New Mexico compound The suspects were charged with violating federal firearms and conspiracy laws.	Abusive	Non-Abusive
	Thats how you know she is down for anything	Non-Abusive	Abusive
Target	Ill give you a little secret Go after Jews You wouldnt last minutes I didnt	Individual Second Person	Group
	You just cant make this crap up Evil sure has a strong presence in this world.	Group	Individual Second Person
	My tweet to this creature usermention You scrubbed your Social Media history but its too late The FBI is investigating you now You better lawyer up You wont do well in Prison.	Individual Second Person	Individual Third Person
Severity	Rape Im sure she was begging for it Doesnt look like a rape scene to me	Violence and Genocide	Act of Bias and Discrimination
	As immigrants flow across US border American guns go south	Act of Bias and Discrimination	Violence and Genocide
	How do yall national socialists feel now that the democrats are adopting national socialist policies instead of marxist policies	Act of Bias and Discrimination	Biased Attitude

Table 6: Sample cases where AbuseAnalyzer predicts incorrectly in comparison to the ground truth.

5 Conclusion

In this paper, we presented a novel dataset with 7,601 Gab posts labeled for abuse presence, target and severity. We experimented with both statistical and deep learning based models for each of these tasks and showed that the BERT based model performs the best. There are several open avenues for the presented work like exploring context based abuse detection especially in social media post and reply threads. Another interesting direction can be to use data from multiple modalities like images, videos and speech along with the text for the task of abuse detection.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *WWW*, pages 759–760.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. In *Twelfth International AAAI Conference on Web and Social Media*.

Joel Finkelstein, Savvas Zannettou, Barry Bradlyn, and Jeremy Blackburn. 2018. A quantitative approach to understanding online antisemitism. *CoRR*, abs/1809.01644.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM CSUR*, 51(4):85.

Antigoni-Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *WebSci*, pages 105–114.

Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *ICWSM*, pages 92–101.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *AAAI*.

Lucas Lima, Julio C. S. Reis, Philipe F. Melo, Fabricio Murai, Leandro Araújo Silva, Pantelis Vikatos, and Fabrício Benevenuto. 2018. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. *ASONAM*, pages 515–522.

Alice Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. *New York: Data & Society Research Institute*.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *WebSci*, pages 173–182.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. In *Workshop on Abusive Language Online*, pages 41–45.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China, November. Association for Computational Linguistics.

Oana Nestian Sandu and Nadine Lyamouri-Bajja. 2018. *T-Kit 4-Intercultural learning*, volume 4. Council of Europe.

Joan Serrà, Ilias Leontiadis, Dimitris Spathis, Gianluca Stringhini, Jeremy Blackburn, and Athena Vakali. 2017. Class-based prediction errors to detect hate speech with out-of-vocabulary words. In *Workshop on Abusive Language Online*, pages 36–40.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Expl.*, 19(1):22–36.

Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *ICWSM*, pages 687–690.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Student Research Workshop@HLT-NAACL*, pages 88–93.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2018. What is gab: A bastion of free speech or an alt-right echo chamber. In *WWW*, pages 1007–1014.