INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

R&D SHOWCASE 2020

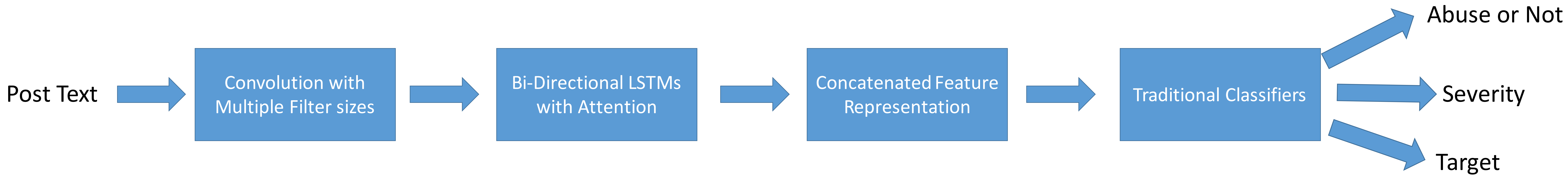
AbuseAnalyzer: Abuse Detection, Severity and Target Prediction for Gab Posts

ABSTRACT

- Online social media platforms extensively popular in the recent times. However, this popularity 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.
- In this paper, we propose mechanisms for predicting presence, severity and target of abusive behavior.
- To address these challenges, we propose a two-stage hybrid attention-based deep learning system called AbuseAnalyzer.
- To demonstrate the efficacy of our proposed methods, we contribute a dataset with 7820 Gab posts, each of which is manually labeled comprehensively across all such aspects.

METHOD

- Besides the Gab post text, we also extracted meta-data related to posts over a period of 4 months to create a corpus of 7820 annotated posts.
- After the creation of the dataset, we used multiple dataset analysis techniques to understand the nature of the data, this essentially helped us to get insight on the nature of the platform as well as on the variety of abuse present on it.
- The second part of work involves development of a deep neural network pipeline system called AbuseAnalyzer which takes in the post text as the input and makes prediction across each of the aforementioned task. We then performed multiple experiments to check the effectiveness of the proposed model.



Post Text

Convolution with Multiple Filter sizes

Bi-Directional LSTMs with Attention

Concatenated Feature Representation

Traditional Classifiers

Abuse or Not

Severity

Target

EXPERIMENTS & ANALYSIS

Models	Accuracy	Recall	Precision	F-1
Stat. Method	0.7345 ± 0.0129	0.7318 ± 0.0131	0.7353 ± 0.0130	0.7323 ± 0.0132
W−CNN	0.7109 ± 0.0290	0.7083 ± 0.02906	0.7384 ± 0.0080	0.6993 ± 0.0377
W−CNN + ELMo	0.7130 ± 0.0405	0.7059 ± 0.0466	0.7505 ± 0.0136	0.6942 ± 0.0644
C−CNN	0.6053 ± 0.0081	0.6013 ± 0.0117	0.6057 ± 0.0076	0.5985 ± 0.0145
W−LSTM	0.7227 ± 0.0197	0.7187 ± 0.0229	0.7308 ± 0.0110	0.7169 ± 0.0259
W−LSTM + ELMo	0.7166 ± 0.0212	0.7169 ± 0.0180	0.7269 ± 0.0123	0.7128 ± 0.0229
C−LSTM	0.6057 ± 0.0117	0.6052 ± 0.0098	0.6107 ± 0.0088	0.5998 ± 0.0123
AbuseAnalyzer	**0.8758 ± 0.0060**	**0.8767 ± 0.0059**	**0.8759 ± 0.0058**	**0.8757 ± 0.0060**

[illegible]

Models	Accuracy	Recall	Precision	F-1
Stat. Method	0.7208 ± 0.0151	0.5414 ± 0.0228	0.6575 ± 0.0316	0.5732 ± 0.0260
W−CNN	0.6732 ± 0.0315	0.4229 ± 0.0543	0.4437 ± 0.1861	0.3910 ± 0.0745
W−CNN + ELMo	0.6826 ± 0.0373	0.4147 ± 0.0939	0.4422 ± 0.2210	0.3868 ± 0.1377
C−CNN	0.6471 ± 0.0003	0.3333 ± 0.0000	0.2157 ± 0.0001	0.2619 ± 0.0001
W−LSTM	0.6535 ± 0.0146	0.3476 ± 0.0319	0.2541 ± 0.0859	0.2863 ± 0.0546
W−LSTM + ELMo	0.6471 ± 0.0003	0.3333 ± 0.0001	0.2157 ± 0.0000	0.2619 ± 0.0001
C−LSTM	0.6471 ± 0.0003	0.3333 ± 0.0000	0.2157 ± 0.0001	0.2619 ± 0.0001
AbuseAnalyzer	**0.8807 ± 0.0071**	**0.8621 ± 0.0174**	**0.8364 ± 0.0123**	**0.8480 ± 0.0116**

Table 1: Experimental Results for Abuse Detection and Severity Prediction Tasks

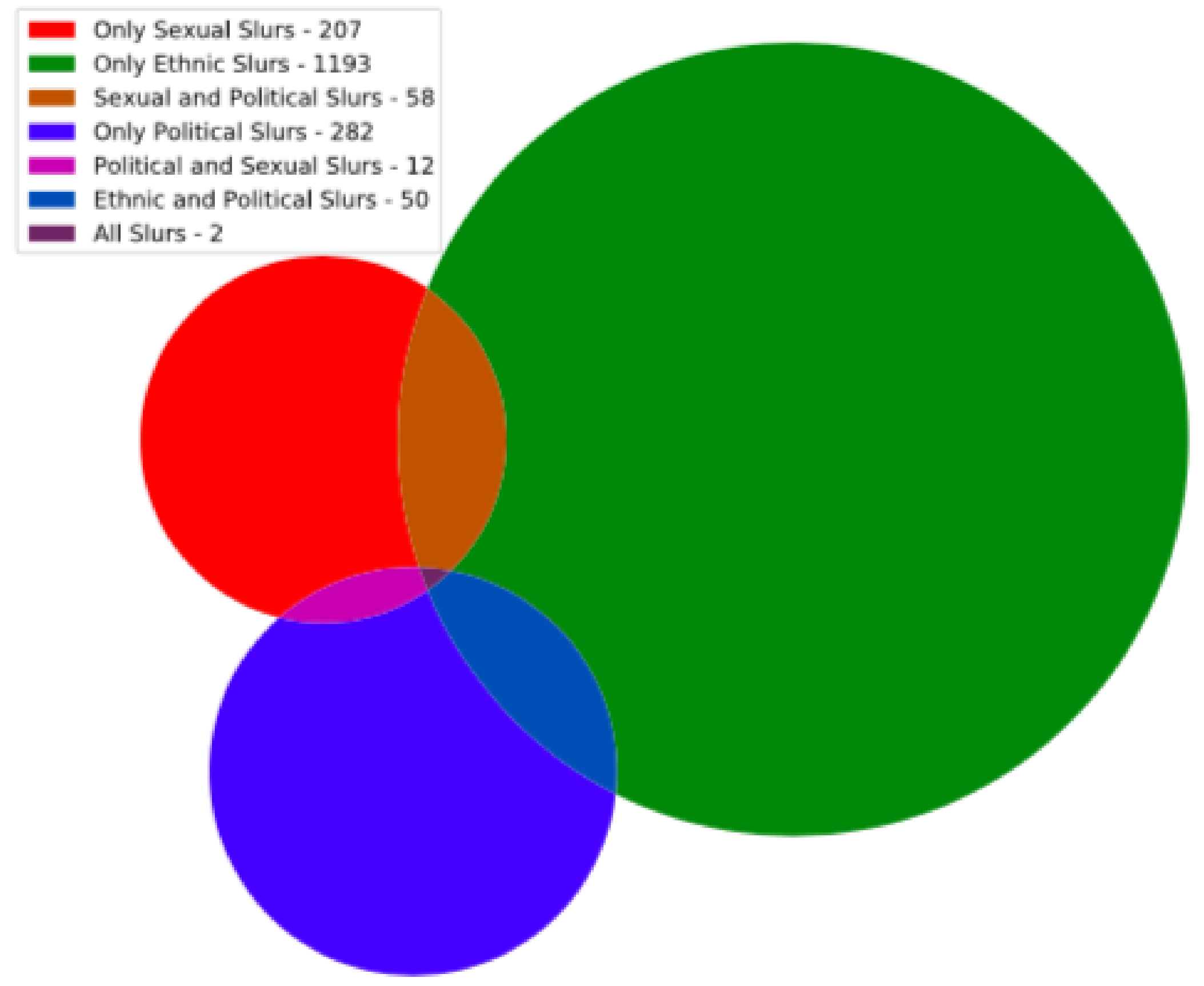


Figure 1: Distribution of different types of slurs among the posts

Task	Post	Our Prediction	Baseline Prediction
Detection	tony blair a great british traitor	Abusive	Non-Abusive
	yes retards always speculate	Abusive	Non-Abusive
	finally the blacks in nfl chiefs won the match	Non-Abusive	Abusive

Table 2: Sample cases where AbuseAnalyzer predicts correctly but the best baseline system fails.

Authors: Mohit Chandra, Ashwin Pathak, Eesha Dutta, Paryul Jain, Manish Gupta, Manish Shrivastava, Ponnurangam Kumaraguru Research Center Name: LTRC