Enriching Abusive Language Detection with Community Context

Jana Kurrek †
McGill University
School of Computer Science
jana.kurrek@mail.mcgill.ca

Haji Mohammad Saleem †
McGill University
School of Computer Science
haji.saleem@mail.mcgill.ca

Derek Ruths
McGill University
School of Computer Science
derek.ruths@mcgill.ca

Abstract

Uses of pejorative expressions can be benign or actively empowering. When models for abuse detection misclassify these expressions as derogatory, they inadvertently censor productive conversations held by marginalized groups. One way to engage with non-dominant perspectives is to add context around conversations. Previous research has leveraged user- and thread-level features, but it often neglects the spaces within which productive conversations take place. Our paper highlights how community context can improve classification outcomes in abusive language detection. We make two main contributions to this end. First, we demonstrate that online communities cluster by the nature of their support towards victims of abuse. Second, we establish how community context improves accuracy and reduces the false positive rates of state-of-the-art abusive language classifiers. These findings suggest a promising direction for context-aware models in abusive language research.

1 Introduction

Existing models for abuse detection struggle to grasp subtle knowledge about the social environments that they operate within. They do not perform natural language understanding and consequently cannot generalize when tested out-of-distribution (Bender et al., 2021). This problem is often the result of training data imbalance, which encourages language models to overestimate the significance of certain lexical cues. For instance, Wiegand et al. (2019) observe that "commentator", "football", and "announcer" end up strongly correlated with hateful tweets in the Waseem and Hovy (2016) corpus. This trend is caused by focused sampling, and it does not reflect an underlying property of abusive expressions.

When models rely on pejorative or demographic words, they can encode systemic bias through *false positives* (Kennedy et al., 2020). For example, research has established that detection algorithms are more likely to classify comments written in African-American Vernacular English (AAVE) as offensive (Davidson et al., 2019; Xia et al., 2020). Benign tweets like "*Wussup, n*gga*!" and "*I saw his ass yesterday*" both score above 90% for toxicity (Sap et al., 2019). Similarly, Zhang et al. (2020) analyze the Wikipedia Talk Pages Corpus (Dixon et al., 2018) and find that 58% of comments that contain the term "gay" are labelled as toxic, while only 10% of all comments are toxic. This enables the misclassification of positive phrases like "*she makes me happy to be gay*". Even Twitter accounts belonging to drag queens have been rated higher in terms of average toxicity than the accounts associated with white nationalists (Oliva et al., 2021). These findings underline how language models with faulty correlations can facilitate the censorship of productive conversations held by marginalized communities.

Productive conversations containing slurs are common, and they take many forms (Hom, 2008). Research inspired by the #MeToo movement has focused on the detection of sexual harassment disclosures by victims (Deal et al., 2020), but this research has not been sufficiently integrated into the literature on abusive language detection. The distinction between actual sexist messages and messages calling out sexism is rarely addressed in the field (Chiril et al., 2020). A similar trend is seen with sarcasm. Humor and self-irony can be employed as coping mechanisms by victims of abuse (Garrick, 2006), yet they constitute frequent sources of error for state-of-the-art classifiers (Vidgen et al., 2019). For example, the median toxicity score for language on `transgendercirclejerk`, a "parody [subreddit] for trans people", is as high as 90% (Kurrek et al., 2020). More broadly, transgender users are "excluded, harmed, and misrepresented in existing

† These authors made equal contributions.

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 131 - 142

platforms, algorithms, and research methods" related to network analysis (Stewart and Spiro, 2021).

Meaningful improvements in abusive language detection require a thoughtful engagement with the perspectives of marginalized communities and their allies. One way to ensure that machine learning frameworks are socially conscientious is to add context around conversations. Past research has explored the contextual information within conversation threads (Pavlopoulos et al., 2020; Ziems et al., 2020), user demographics (Unsvåg and Gambäck, 2018; Founta et al., 2019), user history (Saveski et al., 2021; Qian et al., 2018; Dadvar et al., 2013), user profiles (Unsvåg and Gambäck, 2018; Founta et al., 2019), and user networks (Ziems et al., 2020; Mishra et al., 2018) with varying degrees of success in improving performance. However, most modelling efforts for abusive language detection neglect one major aspect of online conversations: the community environment they take place within.

Online communities adhere to a variety of sociologistic norms that reinforce their identities. This phenomenon is easily observed on Reddit, where community structure is an explicit component of platform design. For example, the majority of comments on the pro-Trump subreddit `The_Donald` delegitimize liberal ideas (McLamore and Uluğ, 2020; Soliman et al., 2019). Similarly, a collection of "manosphere" subreddits espouse misogynistic ideologies (Stewart and Spiro, 2021; Ging, 2019). More broadly, communities can reinforce "toxic technocultures" (Massanari, 2017), and those technocultures are not limited to Reddit. Community structure is present across 4chan, Facebook, Voat, etc., and it exists in a less explicit manner on platforms like Twitter (Silva et al., 2017).

In this paper, we study community context on Reddit, and we focus specifically on language that is collected using slurs. We demonstrate that subreddits cluster by the nature of their support towards marginalized groups, and we use subreddit embeddings to improve the accuracy and false positive rates of state-of-the-art abusive language classifiers. While our analysis is platform-specific, it suggests a promising new direction for context-aware models.

2 Related Work

2.1 Methods in Abusive Language Detection

Abusive language detection is a relatively new field of research, with "very limited" work from as recently as 2016 (Waseem and Hovy, 2016). Early methods featured Naive Bayes (Liu and Forss, 2014), SVMs (Tulkens et al., 2016), Random Forests (Warner and Hirschberg, 2012), Decision Trees (Del Vigna et al., 2017), and Logistic Regression (Burnap and Williams, 2014; Greevy, 2004).

However, recent developments in NLP have directed the field towards neural and Transformer-based approaches. CNNs, LSTMs (+ Attention), and GRUs have been widely used in the literature (Mathur et al., 2018; Meyer and Gambäck, 2019; Chakrabarty et al., 2019; Zhang et al., 2018; Modha et al., 2018). As of 2019, researchers have begun adopting pre-trained language models. Contemporary work leverages BERT, DistilBERT, ALBERT, RoBERTA, and mBERT (Alonso et al., 2020; Davidson et al., 2020). In fact, Bodapati et al. (2019) note that seven of the top ten performing models for offensive language identification at SEMEVAL-2019 were BERT-based. A similar trend was seen at SEMEVAL-2020, where "most teams used some kind of pre-trained Transformers" (Zampieri et al., 2020). Regardless of architecture, methods in abusive language detection can be divided into content- and context- based approaches.

Content-based approaches rely on comment text for feature engineering. Researchers have used TF-IDF weighted n-gram counts as well as distributional embeddings for text representation (Davidson et al., 2017; Nobata et al., 2016; Van Hee et al., 2018), POS tags or dependency relations for encoding syntactic information (Narang and Brew, 2020), and the frequencies of hashtags, URLs, user mentions, emojis, etc. for detecting platform-specific tokens. Lexicons are also popular for capturing sentiment, politeness, emotion, and hate words (Cao et al., 2020; Nobata et al., 2016; Markov and Daelemans, 2021; Koufakou et al., 2020). The central assumption behind content-based abusive language detection is that comments can be exclusively assessed using textual features. However, this assumption neither holds in theory nor in practice because linguistic structures are discourse-determined, and that discourse is shaped by social, historical, and political context (Bridges, 2017). Semantics cannot be completely interpreted using content cues alone. Even human annotators struggle to classify comments that involve satire or homonymy in the absence of broader context (Kurrek et al., 2020). In light of these concerns, researchers are increasingly identifying the impor-

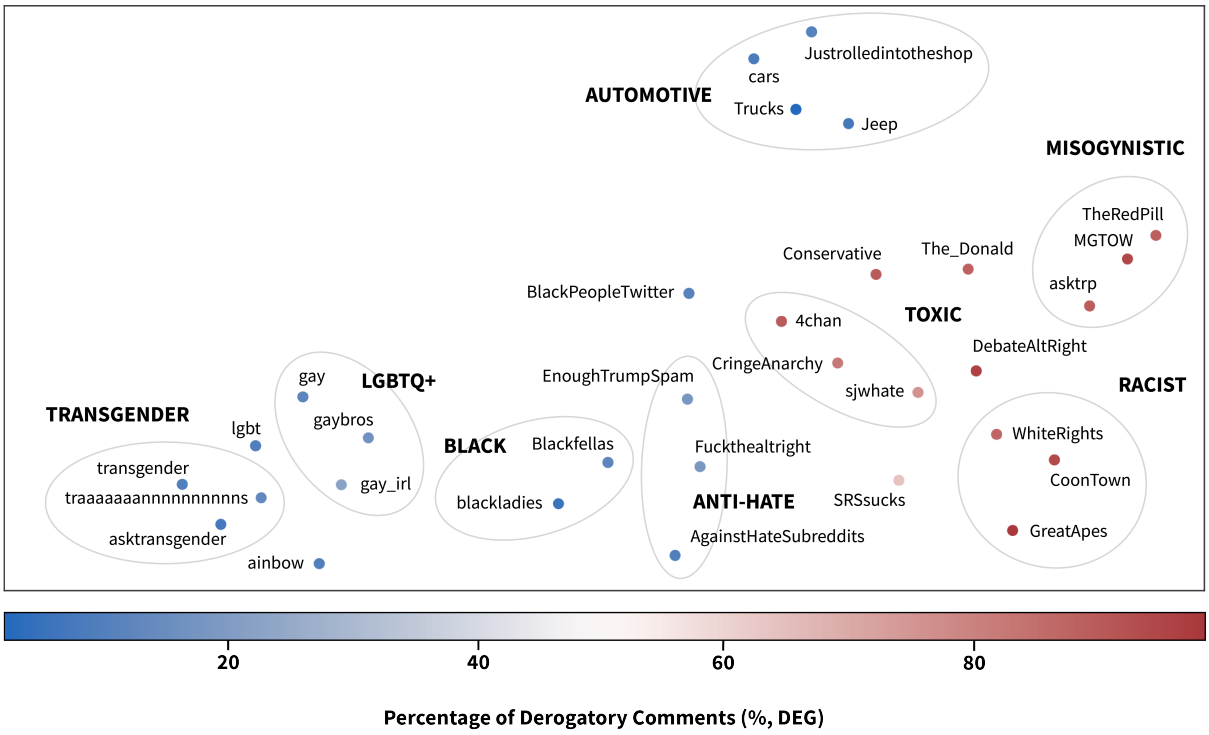

Figure 1: A subset of our subreddit embeddings plotted in two-dimensions using UMAP. Community clusters emerge based on the nature of users' support towards marginalized groups.

tance of user and conversational features to their detection frameworks. We summarize five main trends in the literature below.

Conversational Context. Attempts have been made to situate abusive comments within conversation threads. Threads have been studied using preceding comments (Pavlopoulos et al., 2020; Karan and Šnajder, 2019), discussion titles (Gao and Huang, 2017), and counts for aggressive comments (Ziems et al., 2020). The position of a comment in a thread - start or end - has also been considered (Joksimovic et al., 2019). Finally, researchers have analyzed conversation graphs for topological indicators of abuse (Papegnies et al., 2017).

User Demographics. Researchers have attempted to incorporate user-level context through demographic signals for age, location, and gender. Age has been extracted from user disclosures, but these disclosures can be unreliable when users have an incentive to view adult-rated content (Dadvar et al., 2013). Previous work has inferred gender from user names (Waseem and Hovy, 2016; Unsvåg and Gambäck, 2018), expressions in user biographies (Waseem and Hovy, 2016; Unsvåg and Gambäck, 2018), and in-game avatar choices (Balci and Salah, 2015), but these methods can fail when names are gender-neutral. Location information obtained through geo-coding has also been used to analyze hateful tweets (Fan et al., 2020).

User History. Patterns in user behaviour, including daily logins (Balci and Salah, 2015), favourites (Unsvåg and Gambäck, 2018), and posting history (Saveski et al., 2021; Ziems et al., 2020), can be used as features in abusive language detection models. Some work focuses directly on the content of past comments. For example, Dadvar et al. (2013) look for the prevalence of profanity in text. Conversely, Qian et al. (2018) encode all historical posts by a user. Similarly, Ziems et al. (2020) create TF-IDF vectors derived from a user's timeline.

User Profiles. Several elements of profile metadata have been studied as a proxy for digital identity. These elements include usernames (Gao and Huang, 2017), user anonymity, the presence of updated profile pictures (Unsvåg and Gambäck, 2018), biographies (Miró-Llinares et al., 2018), verified account status (Ziems et al., 2020), counts for followers (Founta et al., 2019), and friends (Balci and Salah, 2015). Some other profile features include profile language (Galán-García et al., 2016) and account age (Founta et al., 2019).

User Networks. Homophily in social networks induces user clusters based on shared identities. These clusters have been shown to represent col-

lective ideologies and moralities (Dehghani et al., 2016), motivating researchers to examine local user networks for markers of abusive behaviour. Interaction and connection-based social graphs are analyzed using Jaccard's similarity and eigenvalue or closeness centrality (Ziems et al., 2020; Chatzakou et al., 2017; Founta et al., 2019; Unsvåg and Gambäck, 2018; Papegnies et al., 2017), which are also relevant for creating user embeddings.

2.2 Methods in Community Profiling

Network data may capture localized trends about individual users, but it often overlooks how groups of users behave as a whole. There are connection- and content-based solutions for explicit community profiling which, to the best of our knowledge, exist outside of contemporary abusive language research. Connection-based solutions evolved out of the idea that similar communities house similar users. In contrast, content-based solutions claim that similar communities contain similar content.

Connection-based Representations. Vector representations of online communities are known to encode semantics (Martin, 2017). Popular techniques for obtaining these representations require the construction of a community graph. Kumar et al. (2018) construct a bipartite multigraph between Reddit users and subreddits. An edge $u_i \rightarrow s_j$ is added for each post by a user u_i in a subreddit s_j. The graph is then used to learn subreddit embeddings by a "node2vec-style" approach.

Martin (2017) creates a symmetric matrix of subreddit-subreddit user co-occurrences, where $\mathbf{X}_{ij}$ is the number of unique users who have commented at least ten times in the subreddits i and j. Skip-grams with negative sampling or `GloVe` can then be used to obtain subreddit embeddings. Here, subreddits and user co-occurrences inherit the role of words and word co-occurrences respectively. Waller and Anderson (2019) also treat communities as "words" and users who comment in them as "contexts" and adapt word2vec for community representations. The subreddit graph proposed in Janchevski and Gievska (2019) contains edges weighted by the number of shared users between the two subreddits. They only consider users who participate in at least ten subreddits and use node2vec to generate node embeddings.

Content-based Representations. Content-based solutions for community profiling rely on methods for document similarity. Janchevski and Gievska (2019) average the word2vec representations for the top 30 words in each subreddit, ranked by TF-IDF score. This research is currently limited, relative to other techniques.

3 Methodology

3.1 Corpus

We select the `Slur-Corpus` by Kurrek et al. (2020). It consists of 40k human-annotated Reddit comments. Every comment contains a slur and is labelled as either derogatory (`DEG`), appropriative (`APR`), non-derogatory non-appropriative (`NDNA`), or homonym (`HOM`). The corpus is nearly evenly split between derogatory and non-derogatory (`APR, NDNA, HOM`) slur usages, with 51% of comments labelled `DEG` (see Table 1).

The `Slur-Corpus` is one of few community-aware resources for abusive language detection. The data is sampled over the course of a decade (October 2007 to September 2019), reflecting a variety of users and language conventions. Every comment is published with the subreddit from which it was sourced, and the authors curate content across a number of antagonistic, supportive, and general discussion communities. As opposed to random sampling, this method guarantees the representation of targeted and minority voices. We see this as crucial for investigating the role of social context within abusive language conventions.

3.2 Definitions

Subreddits are niche communities dedicated to the discussion of a particular topic, with users participating in subreddits that engage their personal interests. As a result, subreddits often exhibit language specificity that can be leveraged for making inferences about slur usages.

Consider the slur *tr*nny*. The comment, "*I am genuinly surpised at a suicidal tr*nny*" from `CringeAnarchy` is derogatory. In contrast, "*So do I. Just that the tr*nny is dying on me lol.*" from `Honda` is non-derogatory because *tr*nny* is being used as a homonym. Both of these subreddits adhere to different linguistic norms and appeal to different user bases. Quantifying these differences is important. Niche or small automotive subreddits are likely to be related to `Honda`, and their users may also use *tr*nny* to mean *transmission*.

Label	Count	%	Stats	Count
DEG	20531	51%	Users	36962
NDNA	16729		Posts	34610
HOM	1998	49%	Subreddits	2691
APR	553			
Total	39811			

Table 1: Characteristics of the Slur-Corpus. The split between DEG and NDG comments is nearly equal.

3.3 Constructing Subreddit Embeddings

We construct subreddit embeddings based on user comment co-occurrence. This method aligns with prior work on the subject (Martin, 2017; Kumar et al., 2018; Waller and Anderson, 2019), but extends it by considering data collected at a much larger scale. We use all publicly available Reddit comments prior to September 2019 in order to generate lists of users that comment in each found subreddit (Baumgartner et al., 2020). We then store frequency counts for each list and, in total, identify 998K unique subreddits and 42.7M unique authors over the course of 12 years. There is a long tail because many subreddits have low participation.

Next, we identify active users, defined as being any users with at least ten comments in a subreddit. We exclude bot accounts and focus on top subreddits by activity. This leaves 10.4K subreddits and 12.2M unique users. With this data, we build a subreddit adjacency matrix $\mathbf{A}$, where $\mathbf{A}_{ij}$ is the number of co-occurring users in subreddits i and j. We use GloVe to generate dense embeddings from $\mathbf{A}$, and we run it over 100 epochs with a learning rate of 0.05 and a representation size of 150.

3.4 Evaluating Subreddit Embeddings

Our tests for subreddit similarity seek to capture two conditions: (1) compositionality: similar subreddits have similar constituent subreddits; and (2) analogy: subreddit similarity is preserved under analogical argument. We rely on vector algebra to model each of these two conditions.

3.4.1 Similarity

The similarity between subreddits S_i and S_j is simply the cosine similarity of their representations:

$$sim(S_i, S_j) = \frac{\vec{S_i} \cdot \vec{S_j}}{|\vec{S_i}||\vec{S_j}|}$$

3.4.2 Composition Tests

We find a subreddit S_k that represents the sum of S_i and S_j. We create $\vec{V} = \vec{S_i} + \vec{S_j}$, and then compute $S_k := max_x(\{sim(\vec{V}, \vec{S_x})\})$. We run the composition test to identify local sports team subreddits from combinations of sport and city subreddits ($\overrightarrow{sport} + \overrightarrow{city} = \overrightarrow{team}$). We base these tests on the evaluations of Martin (2017).

3.4.3 Analogy Tests

We find a subreddit S_n such that $\vec{S_i} : \vec{S_j} :: \vec{S_m} : \vec{S_n}$ for a triad of subreddits S_i, S_j and S_m. We create $\vec{V} = \vec{S_i} - \vec{S_j} + \vec{S_m}$ and then compute $S_n = max_x(\{sim(\vec{V}, \vec{S_x})\})$. The analogy tests, based on Waller and Anderson (2019), identify:

1. A local team given a city and sport:
   $\overrightarrow{city} : \overrightarrow{team} :: \overrightarrow{city'} : \overrightarrow{team'}$
2. A sport given a team and its city:
   $\overrightarrow{team} : \overrightarrow{sport} :: \overrightarrow{team'} : \overrightarrow{sport'}$
3. A city given a university
   $\overrightarrow{university} : \overrightarrow{city} :: \overrightarrow{university'} : \overrightarrow{city'}$

In total, we ran 157 composition tests and 6349 analogy tests. In 81% of cases, the correct answer to a composition test was in the top five most similar subreddits. Similarly, in 84% of cases, the correct answer to an analogy test was in the top five most similar subreddits. Examples are highlighted in Table 2, and we note that they are in line with the results reported in the original paper.

3.5 Context Insensitive Classifiers

To assess the importance of community context, we run a series of context sensitive and context insensitive experiments. We run all experiments using a 5-fold cross validation in order to label the entire corpus. Moreover, we use stratified sampling to ensure a uniform distribution of slurs, subreddits, and labels across all folds. Below, we describe the models used for our context insensitive experiments.

(LOG-REG) Our first classifier is a Logistic Regression with L2 regularization. We preprocess the corpus by lowercasing and stemming the text, removing stop words, and masking user mentions and URLs prior to tokenization. Each token is then weighed using TF-IDF to create unigram, bigram, and trigram features. We use scikit-learn to create our classification pipeline.

city	+	sport	=	team
toronto	+	baseball	=	Torontobluejays
chicago	+	baseball	=	CHICubs
chicago	+	hockey	=	hawks
chicago	+	nba	=	chicagobulls
boston	+	baseball	=	redsox
boston	+	hockey	=	BostonBruins
boston	+	nba	=	bostonceltics
boston	+	nfl	=	Patriots

city	:	team	::	city	:	team
boston	:	BostonBruins	::	toronto	:	leafs
boston	:	Patriots	::	chicago	:	CHIBears
team	:	**sport**	::	**team**	:	**sport**
redsox	:	baseball	::	BostonBruins	:	hockey
redsox	:	baseball	::	Patriots	:	nfl
university	:	**city**	::	**university**	:	**city**
mcgill	:	montreal	::	UBC	:	vancouver
mcgill	:	montreal	::	UofT	:	toronto

Table 2: Examples of subreddit embedding evaluation, based on our composition and analogy tests.

gaybros	**Blackfellas**	**trans**	**AgainstHateSubreddits**
askgaybros	blackladies	transpositive	Fuckthealtright
gay	BlackHair	ask_transgender	TopMindsOfReddit
gaymers	racism	MtF	beholdthemasterrace
4chan	**CoonTown**	**GenderCritical**	**MGTOW**
ImGoingToHellForThis	GreatApes	itsafetish	WhereAreAllTheGoodMen
classic4chan	WhiteRights	GCdebatesQT	TheRedPill
CringeAnarchy	AntiPOZi	Gender_Critical	asktrp
changemyview	**hiphop**	**cars**	**relationships**
PoliticalDiscussion	90sHipHop	Autos	AskWomen
bestof	rap	BMW	relationship_advice
TrueAskReddit	hiphop101	carporn	offmychest

Table 3: Top three subreddits by cosine similarity to each subreddit in bold (experiments run on top five).

(BERT) Our second classifier is BERT. We use BERT-BASE pre-trained on uncased data with the AdamW optimizer, which has a final linear layer. It takes the top-level embedding of the [CLS] token as input. We do fine-tuning over four epochs with a batch size of 32, and we choose a learning rate of 2e-05 and epsilon 1e-8[1].

$$[\text{CLS}]\ c\ [\text{SEP}]$$

(PERSPECTIVE) We use a publicly available commercial tool for toxicity detection[2]. It is a CNN-based model that is trained on a high volume of user-generated comments across social media platforms. While the tool is updated by PERSPECTIVE, the API cannot be retrained, fine-tuned, or modified. We use 0.8 as our threshold for DEG.

3.6 Context Sensitive Classifiers

Below, we describe the models used for our context sensitive experiments.

(LOG-REG-COMM) We use the same setup as in LOG-REG, but we include an additional feature for the name of each subreddit that comments are sourced from. This is done with the purpose of incorporating a social prior with which the algorithm can contextualize the comment text.

(BERT-COMM) We concatenate the name of each source subreddit to the beginning of each text before passing the comment to BERT.

$$[\text{CLS}]\ s + c\ [\text{SEP}]$$

(BERT-COMM-SEP) In our second variant for context sensitivity, we use the sentence entailment format for BERT. This model concatenates the comment with the source subreddit, separated by BERT's [SEP] token. The model is fine-tuned in the same way as our other BERT models.

$$[\text{CLS}]\ c\ [\text{SEP}]\ s\ [\text{SEP}]$$

(BERT-COMM-NGH) We use our trained GloVe embeddings (see Section 3.3) to obtain the five most similar subreddits to each source subreddits. This allows us to build a direct community neighborhood that we concatenate to the source subreddit. We train this variant of BERT using the same sentence entailment format as was described above.

$$[\text{CLS}]\ c\ [\text{SEP}]\ s_1\ s_2\ ...\ s_6\ [\text{SEP}]$$

4 Results

4.1 Subreddits Cluster around Social Polarity

Prior work has established that communities cluster around topics like music, movies, and sports (Martin, 2017). We want to examine how subreddit neighbourhoods behave based on the nature of their support towards marginalized groups. We identify

[1]All BERT experiments were performed on Google Colab with Tesla V100-SXM2-16GB GPU, and we use BERTForSequenceClassification from huggingface for our implementation.

[2]www.perspectiveapi.com

	Performance				% Classified DEG			
	Accuracy	**Precision**	**Recall**	**F1**	DEG	NDNA	APR	HOM
PERSPECTIVE	0.6132	0.6147	0.6102	0.6079	70.75%	53.10%	53.16%	10.71%
LOG-REG	0.8003	0.8009	0.7994	0.7997	82.85%	22.46%	61.30%	16.67%
LOG-REG-COMM	0.8002	0.8001	0.7999	0.8000	81.10%	20.53%	58.95%	15.67%
BERT	0.8856	0.8854	0.8857	0.8855	88.06%	10.26%	47.20%	6.31%
BERT-COMM	0.8905	0.8904	0.8908	0.8905	88.08%	9.38%	42.31%	5.36%
BERT-COMM-SEP	**0.8930**	**0.8930**	**0.8934**	**0.8930**	**88.12%**	8.95%	**39.60%**	5.11%
BERT-COMM-NGH	0.8923	0.8924	0.8928	0.8923	87.82%	**8.80%**	39.78%	**4.75%**

Table 4: Results from our classification task. We report the percentage of each gold label that is classified as DEG. This indicates the percentage of true positives for DEG and the percentage of false positives for the other three labels.

eight supportive and antagonistic subreddits and use our GloVe embeddings to extract the three most similar communities to each of them (see: Table 3). We make two main observations.

First, we observe that supportive subreddits are most similar to other supportive subreddits that cater towards the same marginalized community. For instance, the neighbourhood of gaybros, a subreddit built for the LGBTQ+ community, contains other subreddits based on pride and support: askgaybros, gay, and gaymers. A similar trend is observed with the neighbours of Blackfellas and trans.

Second, we see that antagonistic subreddits are most similar to other antagonistic subreddits. GenderCritical is contained in a cluster of anti-trans subreddits, MGTOW is near misogynistic subreddits, and CoonTown is surrounded by racist subreddits. This highlights how polarizing communities tend to cluster around other communities with the same, or similar, polarities.

Figure 1 shows the embeddings of a sample of subreddits from Slur-Corpus plotted in two-dimensions using UMAP. There are independent groups for misogynistic, racist, toxic, anti-hate, black, gay, and trans subreddits.

4.2 Subreddit Context Reduces False Positives

We present the results from our classification experiments in Table 4[3]. The results will be discussed through two lenses: (1) overall performance; and (2) performance by label.

BERT-based models outperformed classifiers based on Logistic Regression. This is unsurprising, given that Transformers are the current state-of-the-art in NLP. However, LOG-REG achieves nearly 20% higher accuracy than PERSPECTIVE. While this performance gap is likely the result of the data used to train both models, it is concerning given that the Perspective API is widely used as a tool for toxicity detection with both commercial[4] and academic applications (Cuthbertson et al., 2019).

		BERT	∩	BERT-COMM-SEP	
FP		765	1339	480	
TP	1067	587	17492	599	1364
TN		480	16696	765	
FN		599	1853	587	
	2.68%	6.11%	93.89%	6.11%	3.43%

Table 5: The effect of community context on BERT classification outcomes. The column ∩ counts the number of comments with identical labels from BERT and BERT-COMM-SEP, while the columns relating to each classifier only describe comments with different labels. The percentages 2.68% and 3.43% represent the share of true positives and negatives for BERT and BERT-COMM-SEP, respectively.

For both BERT and LOG-REG, the addition of subreddit context reduced the number of false positives across all three non-derogatory labels. Performance on DEG comments remained relatively unchanged. The highest increase in performance was seen with BERT-COMM-SEP, which had each source subreddit concatenated to the comment with a middle [SEP] token. Adding subreddit context led to a significant improvement for appropriative text, across which the false positive rate decreased by almost 8%. For example, "*Tr*nny* here, some of us are actually really cool." was originally misclassified without community context.

Surprisingly, BERT-COMM-NGH, our model with expanded neighbourhood context, showed little improvement over BERT-COMM-SEP. While the identification of NDNA and HOM improved marginally, the false positive rate for appropriative language increased. One possible explanation is that smaller communities did not have a significant presence in the Slur-Corpus (8% of all subreddits accounted for 80% of all comments), and consequently the performance gains associated

[3]We report Macro F1.

[4]Trusted partners include Reddit, The New York Times, The Financial Times, and the Wall Street Journal.

with comments belonging to these subreddits was marginal. We still believe that neighbourhood context is important for determining the nature of niche communities based on their proximity to larger, established supportive or antagonistic communities. Further analysis of this model is required to understand its full potential.

4.3 Understanding Context Sensitivity

We call a comment "context sensitive" if the addition of context changed its classification label. `BERT` and `BERT-COMM-SEP` have comparable performance on the majority of the corpus: 94% of comments are context insensitive (see Table 5). However, 1364 of the total classification errors made by `BERT` were rectified with the inclusion of social context. These classifications represented $> 3\%$ of the actual corpus, but 56% of context-sensitive comments within it. In Table 6, we present examples of top subreddits for both *true positive* and *true negative* context sensitive comments, along with comments for each. The *true positive* comments largely belonged to antagonistic subreddits, while the *true negative* ones belonged to supportive subreddits. Community context helped `BERT-COMM-SEP` identify community polarity.

True Positives
`CringeAnarchy` I am genuinly surpised at a suicidal *tr*nny*
`4chan` This is basically everyday in Atlanta. It's *n*gger/sp*c* central. Give a useful warning next time.
True Negatives
`BlackPeopleTwitter` Shit Britney rides for us too, idk if you seen when she was about to let the hands fly on some dude for calling her security a *n*gger*
`askgaybros` Masc bear here. Twinks are my favorite and *f*ggot* is a pretty funny word :b

Table 6: Top subreddits across comments whose labels were correctly classified with the addition of context.

5 Discussion

Our analysis points to two key resources that would benefit future abusive language research.

Subreddit embeddings for community sampling. Systems for abuse detection should reliably identify different variations of abuse (e.g. sexism, racism, etc.), while still exhibiting sensitivity towards non-derogatory comments (e.g. appropriation, reclamation, etc.). One way to achieve this is to ensure content diversity in training data. Kurrek et al. (2020) specifically use community sampling to achieve this kind of diversity. The authors collect comments from various Reddit communities, but their work is limited by the absence of resources that identify and consolidate supportive or antagonistic subreddits. Instead, they rely on manual data exploration. There are several issues with this approach. First, knowing which communities to look for (and how to find them) requires a high degree of domain knowledge. Second, manual comment analysis is an expensive task, which makes it difficult to scale or reuse as new communities form. Third, this method is prone to overlooking smaller, niche subreddits that would otherwise have been found using a neighborhood exploration of community clusters. We propose the use of subreddit embeddings in future research to further extend efforts on diverse and representative content collection.

Community context for protecting productive conversations. One of our primary research objectives was to ensure that detection frameworks do not mistakenly classify productive conversations as abusive. Community contextualized models, based on Logistic Regression and `BERT`, better identified non-derogatory comments than their context-insensitive counterparts. Context was found to be particularly helpful for identifying appropriative language, resulting in an 8% increase in accuracy with the addition of a subreddit name. Appropriation is a tool used by marginalized populations to counteract oppression. When abuse detection frameworks misclassify reclamation, they censor the empowerment tools of the very communities that they are installed to protect. Our analysis of the `Slur-Corpus` suggests that productive conversations tend to happen in safe and supportive social spaces. It is therefore crucial that these spaces be considered for nuanced classification of abuse.

6 Conclusion and Future Work

The subjectivity of abuse makes it challenging to annotate and detect reliably. One method for making the problem tractable is to position online conversations within a larger context. This paper was an exploration of one type of contextual in-

formation: community identity. We found that the context derived from community identity can help in the collection and classification of abusive language. We therefore believe that community context is integral to all stages of abusive language research. We leave as future work the inclusion of community information in existing, platform-agnostic, ensemble detection frameworks.

References

Pedro Alonso, Rajkumar Saini, and György Kovács. 2020. Hate speech detection using transformer ensembles on the hasoc dataset. In *International Conference on Speech and Computer*, pages 13–21. Springer.

Koray Balci and Albert Ali Salah. 2015. Automatic analysis and identification of verbal aggression and abusive behaviors for online social games. *Computers in Human Behavior*, 53:517–526.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Sravan Bodapati, Spandana Gella, Kasturi Bhattacharjee, and Yaser Al-Onaizan. 2019. Neural word decomposition models for abusive language detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 135–145.

Judith Bridges. 2017. Gendering metapragmatics in online discourse:"mansplaining man gonna mansplain…". *Discourse, Context & Media*, 20:94–102.

Peter Burnap and Matthew Leighton Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. *Internet, Policy & Politics*.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. Deephate: Hate speech detection via multi-faceted text representations. In *12th ACM Conference on Web Science*, pages 11–20.

Tuhin Chakrabarty, Kilol Gupta, and Smaranda Muresan. 2019. Pay "attention" to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 70–79.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. He said "who's gonna take care of your children when you are at acl?": Reported sexist acts are not sexist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4055–4066.

Lana Cuthbertson, Alex Kearney, Riley Dawson, Ashia Zawaduk, Eve Cuthbertson, Ann Gordon-Tighe, and Kory Wallace Mathewson. 2019. Women, politics and twitter: Using machine learning to change the discourse. In *Proceedings AI for Social Good workshop at NeurIPS*.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *Proceedings of the 35th European conference on Advances in Information Retrieval*, pages 693–696.

Sam Davidson, Qiusi Sun, and Magdalena Wojcieszak. 2020. Developing a new classifier for automated identification of incivility in social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 95–101.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Bonnie-Elene Deal, Lourdes S Martinez, Brian H Spitzberg, and Ming-Hsiang Tsou. 2020. "i definitely did not report it when i was raped...# webelievechristine# metoo": A content analysis of disclosures of sexual assault on twitter. *Social Media+ Society*, 6.

Morteza Dehghani, Kate Johnson, Joe Hoover, Eyal Sagi, Justin Garten, Niki Jitendra Parmar, Stephen Vaisey, Rumen Iliev, and Jesse Graham. 2016. Purity homophily in social networks. *Journal of Experimental Psychology: General*, 145.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC)*, pages 86–95.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Lizhou Fan, Huizi Yu, and Zhanyuan Yin. 2020. Stigmatization in social media: Documenting and analyzing hate speech for covid-19 on twitter. *Proceedings of the Association for Information Science and Technology*, 57(1):e313.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*, pages 105–114.

Patxi Galán-García, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. 2016. Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying. *Logic Journal of the IGPL*, 24(1):42–53.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266. INCOMA Ltd.

Jacqueline Garrick. 2006. The humor of trauma survivors: Its application in a therapeutic milieu. *Journal of aggression, maltreatment & trauma*, 12(1-2):169–182.

Debbie Ging. 2019. Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and Masculinities*, 22(4):638–657.

Edel Greevy. 2004. *Automatic text categorisation of racist webpages*. Ph.D. thesis, Dublin City University.

Christopher Hom. 2008. The semantics of racial epithets. *The Journal of Philosophy*, 105(8):416–440.

Andrej Janchevski and Sonja Gievska. 2019. A study of different models for subreddit recommendation based on user-community interaction. In *International Conference on ICT Innovations*, pages 96–108. Springer.

Srecko Joksimovic, Ryan S Baker, Jaclyn Ocumpaugh, Juan Miguel L Andres, Ivan Tot, Elle Yuan Wang, and Shane Dawson. 2019. Automated identification of verbally abusive behaviors in online discussions. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 36–45.

Mladen Karan and Jan Šnajder. 2019. Preemptive toxic language detection in wikipedia comments using thread-level context. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 129–134.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43. Association for Computational Linguistics.

Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference*, pages 933–943.

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149.

Shuhua Liu and Thomas Forss. 2014. Combining n-gram based similarity analysis with sentiment analysis in web content classification. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management-Volume 1*, pages 530–537.

Ilia Markov and Walter Daelemans. 2021. Improving cross-domain hate speech detection by reducing the false positive rate. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*.

Trevor Martin. 2017. community2vec: Vector representations of online communities encode semantic relationships. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 27–31.

Adrienne Massanari. 2017. # gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New media & society*, 19(3):329–346.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26.

Quinnehtukqut McLamore and Özden Melis Uluğ. 2020. Social representations of sociopolitical groups on r/the_donald and emergent conflict narratives: A qualitative content analysis. *Analyses of Social Issues and Public Policy*.

Johannes Skjeggestad Meyer and Björn Gambäck. 2019. A platform agnostic dual-strand hate speech detector. In *ACL 2019 The Third Workshop on Abusive Language Online Proceedings of the Workshop*. Association for Computational Linguistics.

Fernando Miró-Llinares, Asier Moneva, and Miriam Esteve. 2018. Hate is in the air! but where? introducing an algorithm to detect hate speech in digital microenvironments. *Crime Science*, 7(1):1–12.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th international conference on computational linguistics*, pages 1088–1098.

Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2018. Filtering aggression from the multilingual social media feed. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 199–207.

Kanika Narang and Chris Brew. 2020. Abusive language detection using syntactic dependency graphs. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 44–53.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture*, 25(2):700–732.

Etienne Papegnies, Vincent Labatut, Richard Dufour, and Georges Linares. 2017. Graph-based features for automatic online abuse detection. In *International conference on statistical language and speech processing*, pages 70–81. Springer.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2, pages 118–123.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Martin Saveski, Brandon Roy, and Deb Roy. 2021. The structure of toxic conversations on twitter. In *Proceedings of the Web Conference 2021*, pages 1086–1097.

Wendel Silva, Ádamo Santana, Fábio Lobato, and Márcia Pinheiro. 2017. A methodology for community detection in twitter. In *Proceedings of the International Conference on Web Intelligence*, pages 1006–1009.

Ahmed Soliman, Jan Hafer, and Florian Lemmerich. 2019. A characterization of political communities on reddit. In *Proceedings of the 30th ACM conference on hypertext and Social Media*, pages 259–263.

Leo G. Stewart and Emma S. Spiro. 2021. Nobody puts redditor in a binary: Digital demography, collective identities, and gender in a subreddit network. In *Proceedings of the 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. Association for Computing Machinery.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media. In *Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety*, page 11. LREC.

Elise Fehn Unsvåg and Björn Gambäck. 2018. The effects of user features on twitter hate speech detection. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 75–85.

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PloS one*, 13.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93.

Isaac Waller and Ashton Anderson. 2019. Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. In *The World Wide Web Conference*, pages 1954–1964.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 602–608.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14. ACL.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447. International Committee for Computational Linguistics.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer.

Caleb Ziems, Ymir Vigfusson, and Fred Morstatter. 2020. Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 808–819.