On Cross-Dataset Generalization in Automatic Detection of Online Abuse

Isar Nejadgholi **Svetlana Kiritchenko**

National Research Council Canada
{isar.nejadgholi,svetlana.kiritchenko}@nrc-cnrc.gc.ca

Abstract

NLP research has attained high performances in abusive language detection as a supervised classification task. While in research settings, training and test datasets are usually obtained from similar data samples, in practice systems are often applied on data that are different from the training set in topic and class distributions. Also, the ambiguity in class definitions inherited in this task aggravates the discrepancies between source and target datasets. We explore the topic bias and the task formulation bias in cross-dataset generalization. We show that the benign examples in the Wikipedia Detox dataset are biased towards platform-specific topics. We identify these examples using unsupervised topic modeling and manual inspection of topics' keywords. Removing these topics increases cross-dataset generalization, without reducing in-domain classification performance. For a robust dataset design, we suggest applying inexpensive unsupervised methods to inspect the collected data and downsize the non-generalizable content before manually annotating for class labels.

1 Introduction

The NLP research community has devoted significant efforts to support the safety and inclusiveness of online discussion forums by developing automatic systems to detect hurtful, derogatory or obscene utterances. Most of these systems are based on supervised machine learning techniques, and require annotated data. Several publicly available datasets have been created for the task (Mishra et al., 2019; Vidgen and Derczynski, 2020). However, due to the ambiguities in the task definition and complexities of data collection, cross-dataset generalizability remains a challenging and understudied issue of online abuse detection.

Existing datasets differ in the considered types of offensive behaviour and annotation schemes, data sources and data collection methods. There is no agreed-upon definition of harmful online behaviour yet. Several terms have been used to refer to the general concept of harmful online behavior, including *toxicity* (Hosseini et al., 2017), *hate speech* (Schmidt and Wiegand, 2017), *offensive* (Zampieri et al., 2019) and *abusive* language (Waseem et al., 2017; Vidgen et al., 2019a). Still, in practice, every dataset only focuses on a narrow range of subtypes of such behaviours and a single online platform (Jurgens et al., 2019). For example, Davidson et al. (2017) annotated tweets for three categories, *Racist*, *Offensive but not Racist* and *Clean*, and Nobata et al. (2016) collected discussions from Yahoo! Finance news and applied a binary annotation scheme of *Abusive* versus *Clean*. Further, since pure random sampling usually results in small proportions of offensive examples (Founta et al., 2018), various sampling techniques are often employed. Zampieri et al. (2019) used words and phrases frequently found in offensive messages to search for potential abusive tweets. Founta et al. (2018) and Razavi et al. (2010) started from random sampling, then boosted the abusive part of the datasets using specific search procedures. Hosseinmardi et al. (2015) used snowballing to collect abusive posts on Instagram. Due to this variability in category definitions and data collection techniques, a system trained on a particular dataset is prone to overfitting to the specific characteristics of that dataset. As a result, although models tend to perform well in cross-validation evaluation on one dataset, the cross-dataset generalizability remains low (van Aken et al., 2018; Wiegand et al., 2019).

In this work, we investigate the impact of two types of biases originating from source data that can emerge in a cross-domain application of models: 1) task formulation bias (discrepancy in class definitions and annotation between the training and test sets) and 2) selection bias (discrepancy

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 173–183

https://doi.org/10.18653/v1/P17

in the topic and class distributions between the training and test sets). Further, we suggest topic-based dataset pruning as a method of mitigating selection bias to increase generalizability. This approach is different from domain adaptation techniques based on data selection (Ruder and Plank, 2017; Liu et al., 2019) in that we apply an unsupervised topic modeling method for topic discovery without using the class labels. We show that some topics are more generalizable than others. The topics that are specific to the training dataset lead to overfitting and, therefore, lower generalizability. Excluding or down-sampling instances associated with such topics before the expensive annotation step can substantially reduce the annotation costs.

We focus on the Wikipedia Detox or *Wiki*-dataset, (an extension of the dataset by Wulczyn et al. (2017)), collected from English Wikipedia talk pages and annotated for toxicity. To explore the generalizability of the models trained on this dataset, we create an out-of-domain test set comprising various types of abusive behaviours by combining two existing datasets, namely *Waseem*-dataset (Waseem and Hovy, 2016) and *Founta*-dataset (Founta et al., 2018), both collected from Twitter.

Our main contributions are as follows:

- We identify topics included in the *Wiki*-dataset and manually examine keywords associated with the topics to heuristically determine topics' generalizability and their potential association with toxicity.
- We assess the generalizability of the task formulations by training a classifier to detect the *Toxic* class in the *Wiki*-dataset and testing it on an out-of-domain dataset comprising various types of offensive behaviours. We find that *Wiki-Toxic* is most generalizable to *Founta-Abusive* and least generalizable to *Waseem-Sexism*.
- We show that re-sampling techniques result in a trade-off between the True Positive and True Negative rates on the out-of-domain test set. This trade-off is mainly governed by the ratio of toxic to normal instances and not the size of the dataset.
- We investigate the impact of topic distribution on generalizability and show that general and identity-related topics are more generalizable than platform-specific topics.
- We show that excluding Wikipedia-specific data instances (54% of the dataset) does not affect the results of in-domain classification, and improves both True Positive and True Negative rates on the out-of-domain test set, unlike re-sampling methods. Through unsupervised topic modeling, such topics can be identified and excluded before annotation.

2 Biases Originating from Source Data

We focus on two types of biases originated from source data: task formulation and selection bias.

Task formulation bias: In commercial applications, the definitions of offensive language heavily rely on community norms and context and, therefore, are imprecise, application-dependent, and constantly evolving (Chandrasekharan et al., 2018). Similarly in NLP research, despite having clear overlaps, offensive class definitions vary significantly from one study to another. For example, the *Toxic* class in the *Wiki*-dataset refers to aggressive or disrespectful utterances that would likely make participants leave the discussion. This definition of toxic language includes some aspects of racism, sexism and hateful behaviour. Still, as highlighted by Vidgen et al. (2019a), identity-based abuse is fundamentally different from general toxic behavior. Therefore, the *Toxic* class definition used in the *Wiki*-dataset differs in its scope from the abuse-related categories as defined in the *Waseem*-dataset and *Founta*-dataset. Wiegand et al. (2019) converted various category sets to binary (offensive vs. normal) and demonstrated that a system trained on one dataset can identify other forms of abuse to some extent. We use the same methodology and examine different offensive categories in out-of-domain test sets to explore the deviation in a system's performance caused by the differences in the task definitions.

Regardless of the task formulation, abusive language can be divided into explicit and implicit (Waseem et al., 2017). Explicit abuse refers to utterances that include obscene and offensive expressions, such as *stupid* or *scum*, even though not all utterances that include obscene expressions are considered abusive in all contexts. Implicit abuse refers to more subtle harmful behaviours, such as stereotyping and micro-aggression. Explicit abuse is usually easier to detect by human annotators and automatic systems. Also, explicit abuse is more transferable between datasets as it is part of many

definitions of online abuse, including personal attacks, hate speech, and identity-based abuse. The exact definition of implicit abuse, on the other hand, can substantially vary between task formulations as it is much dependent on the context, the author and the receiver of an utterance (Wiegand et al., 2019).

Selection bias: Selection (or sampling) bias emerge when source data, on which the model is trained, is not representative of target data, on which the model is applied (Shah et al., 2020). We focus on two data characteristics affecting selection bias: topic distribution and class distribution.

In practice, every dataset covers a limited number of topics, and the **topic distributions** depend on many factors, including the source of data, the search mechanism and the timing of the data collection. For example, our source dataset, *Wiki*-dataset, consists of Wikipedia talk pages dating from 2004–2015. On the other hand, one of the sources of our target dataset, *Waseem*-dataset, consists of tweets collected using terms and references to specific entities that frequently occur in tweets expressing hate speech. As a result of its sampling strategy, *Waseem*-dataset includes many tweets on the topic of 'women in sports'. Wiegand et al. (2019) showed that different data sampling methods result in various distributions of topics, which affects the generalizability of trained classifiers, especially in the case of implicit abuse detection. Unlike explicit abuse, implicitly abusive behaviour comes in a variety of semantic and syntactic forms. To train a generalizable classifier, one requires a training dataset that covers a broad range of topics, each with a good representation of offensive examples. We continue this line of work and investigate the impact of topic bias on cross-dataset generalizability by identifying and changing the distribution of topics in controlled experiments.

The amount of online abuse on mainstream platforms varies greatly but is always very low. Founta et al. (2018) found that abusive tweets form 0.1% to 3% of randomly collected datasets. Vidgen et al. (2019b) showed that depending on the platform the prevalence of abusive language can range between 0.001% and 8%. Despite various data sampling strategies aimed at increasing the proportion of offensive instances, the **class imbalance** (the difference in class sizes) in available datasets is often severe. When trained on highly imbalanced data, most statistical machine learning methods exhibit a bias towards the majority class, and their performance on a minority class, usually the class of interest, suffers. A number of techniques have been proposed to address class imbalance in data, including data re-sampling, cost-sensitive learning, and neural network specific learning algorithms (Branco et al., 2016; Haixiang et al., 2017; Johnson and Khoshgoftaar, 2019). In practice, simple re-sampling techniques, such as down-sampling of over-represented classes, often improve the overall performance of the classifier (Johnson and Khoshgoftaar, 2019). However, re-sampling techniques might lead to overfitting to one of the classes causing a trade-off between True Positive and True Negative rates. When aggregated in an averaged metric such as F-score, this trade-off is usually overlooked.

3 Datasets

We exploit three large-scale, publicly available English datasets frequently used for the task of online abuse detection. Our main dataset, *Wiki*-dataset (Wulczyn et al., 2017), is used as a training set. The out-of-domain test set is obtained by combining the other two datasets, *Founta*-dataset (Founta et al., 2018) and *Waseem*-dataset (Waseem and Hovy, 2016).

Training set: The *Wiki*-dataset includes 160K comments collected from English Wikipedia discussions and annotated for *Toxic* and *Normal*, through crowd-sourcing[1]. Every comment is annotated by 10 workers, and the final label is obtained through majority voting. The class *Toxic* comprises rude, hateful, aggressive, disrespectful or unreasonable comments that are likely to make a person leave a conversation[2]. The dataset consists of randomly collected comments and comments made by users blocked for violating Wikipedia's policies to augment the proportion of toxic texts. This dataset contains 15,362 instances of *Toxic* and 144,324 *Normal* texts.

Out-of-Domain test set: The toxic portion of our test set is composed of four types of offensive language: *Abusive* and *Hateful* from the *Founta*-dataset, and *Sexist* and *Racist* from the *Waseem*-dataset. For the benign examples of our test set, we use the *Normal* class of the *Founta*-dataset.

[1] https://meta.wikimedia.org/wiki/Research:Detox/Data_Release

[2] https://github.com/ewulczyn/wiki-detox/blob/master/src/modeling/toxicity_question.png

The *Founta*-dataset is a collection of 80K tweets crowd-annotated for four classes: *Abusive*, *Hateful*, *Spam* and *Normal*. The data is randomly sampled and then boosted with tweets that are likely to belong to one or more of the minority classes by deploying an iterative data exploration technique. The *Abusive* class is defined as content with any strongly impolite, rude or hurtful language that shows a debasement of someone or something, or shows intense emotions. The *Hateful* class refers to tweets that express hatred towards a targeted individual or group, or are intended to be derogatory, to humiliate, or to insult members of a group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender. *Spam* refers to posts consisted of advertising/marketing, posts selling products of adult nature, links to malicious websites, phishing attempts and other unwanted information, usually sent repeatedly. Tweets that do not fall in any of the prior classes are labelled as *Normal* (Founta et al., 2018). We do not include the *Spam* class in our test set as this category does not constitute offensive language, in general. The *Founta*-dataset contains 27,150 of *Abusive*, 4,965 of *Hateful* and 53,851 of *Normal* instances.

The *Waseem*-dataset includes 16K manually annotated tweets, labeled as *Sexist*, *Racist* or *Neither*. The corpus is collected by searching for common slurs and terms pertaining to minority groups as well as identifying tweeters that use these terms frequently. A tweet is annotated as *Racist* or *Sexist* if it uses a racial or sexist slur, attacks, seeks to silence, unjustifiably criticizes or misrepresents a minority or defends xenophobia or sexism. Tweets that do not fall in these two classes are labeled as *Neither* (Waseem and Hovy, 2016). The *Neither* class represents a mixture of benign and abusive (but not sexist or racist) instances, and, therefore, is excluded from our test set. *Waseem*-dataset contains 3,430 of *Sexist* and 1,976 of *Racist* tweets.

4 Topic Analysis of the *Wiki*-dataset

We start by exploring the content of the *Wiki*-dataset through topic modeling. We train a topic model using the Online Latent Dirichlet Allocation (OLDA) algorithm (Hoffman et al., 2010) as implemented in the Gensim library (Řehůřek and Sojka, 2010) with the default parameters. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a Baysian probabilistic model of a collection of texts. Each text is assumed to be generated from a multinomial distribution over a given number of topics, and each topic is represented as a multinomial distribution over the vocabulary. We pre-process the texts by lemmatizing the words and removing the stop words. To determine the optimal number of topics, we use a coherence measure that calculates the degree of semantic similarity among the top words (Röder et al., 2015). Top words are defined as the most probable words to be seen conditioned on a topic. We experimented with a range of topic numbers between 10 and 30 and obtained the maximal average coherence with 20 topics. Each topic is represented by 10 top words. For simplicity, each text is assigned a single topic that has the highest probability. The full list of topics and their top words are available in the Appendix.

Topics	Top words
Category 1	
topic 0	know, like, thank, think, want
topic 1	time, like, peopl, think, life
Category 2	
topic 2	suck, year, school, c*ck, p*ssi
topic 7	english, countri, american, nation, german
topic 8	kill, die, jewish, islam, israel
topic 9	god, christian, cast, presid, japanes
topic 12	person, editor, attack, accuse, user
topic 14	f*ck, sh*t, *ss, stupid, bastard
topic 16	team, footbal, gay, match, station
Category 3	
topic 3	redirect, talk, categori, film, episod
topic 4	page, wikipedia, edit, talk, articl
topic 5	sourc, claim, cite, wikipedia, publish
topic 6	link, list, page, inform, articl
topic 10	delet, articl, imag, tag, copyright
topic 11	univers, law, scienc, theori, definit
topic 13	page, discuss, review, talk, templat
topic 15	articl, section, discuss, refer, editor
topic 17	http, com, www, org, wiki
topic 18	edit, block, vandal, user, account
topic 19	style, align, color, background, border

Table 1: Topics identified in the *Wiki*-dataset. For each topic, five of ten top words that are most representative of the assigned category are shown.

We group the 20 extracted topics into three categories based on the coherency of the top words and their potential association with offensive language. This is done through manual examination of the 10 top words in each topic. Table 1 shows five out of ten top words for each topic that are most representative of the assigned category.

Category 1: incoherent or mixture of general topics

The top words of two topics (topic 0 and topic 1) are general terms such as *think*, *want*, *time*, and *life*. This category forms 26% of the dataset. Since

these topics appear incoherent, their association with offensiveness cannot be judged heuristically. Looking at the toxicity annotations we observe that 47% of the *Toxic* comments belong to these topics. These comments mostly convey personal insults, usually not tied to any identity group. The frequently used abusive terms in these *Toxic* comments include *f*ck*, *stupid*, *idiot*, **ss*, etc.

Category 2: coherent, high association with offensive language

Seven of the topics can be associated with offensive language; their top words represent profanity or are related to identity groups frequently subjected to abuse. Topic 14 is the most explicitly offensive topic; nine out of ten top words are associated with insult and hatred. 97% of the instances belonging to this topic are annotated as *Toxic*, with 96% of them containing explicitly toxic words.[3] These are generic profanities with the word *f*ck* being the most frequently used word.

The top words of the other six topics (topics 2, 7, 8, 9, 12, and 16) include either offensive words or terms related to identity groups based on gender, ethnicity, or religion. On average, 16% of the comments assigned to these topics are labeled as *Toxic*. We manually analyzed these comments, and found that each topic (except topic 12) tends to concentrate around a specific identity group. Offensive comments in topic 2 mostly contain sexual slur and target female and homosexual users. In topic 7, comments often contain racial and ethnicity based abuse. Topic 8 contains physical threats, often targeting Muslims and Jewish folks (the words *die* and *kill* are the most frequently used content words in the offensive messages of this topic). Comments in topic 9 involve many terms associated with Christianity (e.g., *god*, *christian*, *Jesus*). Topic 16 has the least amount of comments (0.3% of the dataset), with the offensive messages mostly targeting gay people (the word *gay* appears in 67% of the offensive messages in this topic). Topic 12 is comprised of personal attacks in the context of Wikipedia admin–contributor relations. The most common offensive words in this topic include *f*ck*, *stupid*, *troll*, *ignorant*, *hypocrite*, etc. 20% of the whole dataset and 35% of the comments labeled as *Toxic* belong to this category.

[3]Following Wiegand et al. (2019), we estimate the proportion of explicitly offensive instances in a dataset as the proportion of abusive instances that contain at least one word from the lexicon of abusive words by Wiegand et al. (2018).

Dataset/Class	Cat.#1	Cat.#2	Cat.#3
Training Set			
Wiki-Toxic	48%	34%	18%
Wiki-Normal	24%	18%	58%
Test Set			
Founta-Abusive	58%	33%	8%
Founta-Hateful	54%	37%	9%
Waseem-Sexist	50%	35%	15%
Waseem-Racist	23%	67%	10%
Founta-Normal	51%	28%	21%

Table 2: Distribution of topic categories per class

Category 3: coherent, low association with offensive language

The remaining eleven topics include top words specific to Wikipedia and not directly associated with offensive language. For example, keywords of topic 4 are terms such as *page*, *Wikipedia*, *edit* and *article*, and only 0.4% of the 10,471 instances in this topic are labeled as *Toxic*. These eleven topics comprise 54% of the comments in the dataset and 18% of the *Toxic* comments.

5 Topic Distribution of the Test Set

We apply the LDA topic model trained on the *Wiki*-dataset as described in Section 4 to the Out-of-Domain test set. As before, each textual instance is assigned a single topic that has the highest probability. Table 2 summarizes the distribution of topics for all classes in the three datasets.

Observe that Category 3 is the least represented category of topics across all classes, except for the *Normal* class in the *Wiki*-dataset. Specifically, there is a significant deviation in the topic distribution between the *Wiki-Normal* and the *Founta-Normal* classes. This deviation can be explained by the difference in data sources. Normal conversations on Twitter are more likely to be about general concepts covered in Category 1 or identity-related topics covered in Category 2 than the specific topics such as *writing* and *editing* in Category 3. Other than *Waseem-Racist*, which has 67% overlap with Category 2, all types of offensive behaviour in the three datasets have more overlap with the general topics (Category 1) than identity-related topics (Category 2). For example, for the *Waseem-Sexist*, 50% of instances fall under Category 1, 35% under Category 2 and 15% under Category 3. Topic 1, which is a mixture of general topics, is the dominant topic among the *Waseem-Sexist* tweets. Out of the topics in Category 2, most of the sexist tweets are matched to topic 2 (focused on sexism and homophobia) and topic 12 (general personal insults).

Dataset/Class	Test Subset All	Cat.#1	Cat.#2	Cat.#3
Out-of-Domain - Toxic				
Founta-Abusive	0.94	0.94	**0.96**	0.91
Founta-Hateful	0.62	**0.65**	0.62	0.43
Waseem-Sexist	0.26	**0.29**	0.26	0.17
Waseem-Racist	0.35	**0.37**	0.36	0.20
Out-of-domain - Normal				
Founta-Normal	0.96	0.95	0.97	**0.99**

Table 3: Accuracy per test class and topic category for a classifier trained on *Wiki*-dataset. Best results in each row are in bold.

6 Generalizability of the Model Trained on the *Wiki*-dataset

To explore how well the *Toxic* class from the *Wiki*-dataset generalizes to other types of offensive behaviour, we train a binary classifier (*Toxic* vs. *Normal*) on the *Wiki*-dataset (combining the train, development and test sets) and test it on the Out-of-Domain Test set. This classifier is expected to predict a positive (*Toxic*) label for the instances of classes *Founta-Abusive*, *Founta-Hateful*, *Waseem-Sexism* and *Waseem-Racism*, and a negative (*Normal*) label for the tweets in the *Founta-Normal* class. We fine-tune a BERT-based classifier (Devlin et al., 2019) with a linear prediction layer, the batch size of 16 and the learning rate of 2×10^{-5} for 2 epochs.

Evaluation metrics: In order to investigate the trade-off between the True Positive and True Negative rates, in the following experiments we report accuracy per test class. Accuracy per class is calculated as the rate of correctly identified instances within a class. Accuracy over the toxic classes (*Founta-Abusive*, *Founta-Hateful*, *Waseem-Sexism* and *Waseem-Racism*) indicates the True Positive rate, while accuracy of the normal class (*Founta-Normal*) measures the True Negative rate. Note that given the sizes of the positive and negative test classes, all other common metrics, such as various kinds of averaged F1-scores, can be calculated from the accuracies per class. In addition, we report macro-averaged F-score, weighted by the sizes of the negative and positive classes, to show the overall impact of the proposed method.

Results: The overall performance of the classifier on the Out-of-Domain test set is quite high: weighted macro-averaged $F_1 = 0.90$. However, when the test set is broken down into the 20 topics of the *Wiki*-dataset and the accuracy is measured within the topics, the results vary greatly. For example, for the instances that fall under topic 14, the explicitly offensive topic, the F1-score is 0.99. For topic 15, a Wikipedia-specific topic, the F1-score is 0.80. Table 3 shows the overall accuracies for each test class as well as the accuracies for each topic category (described in Section 4) within each class.

For the class *Founta-Abusive*, the classifier achieves 94% accuracy. 12% of the *Founta-Abusive* tweets fall under the explicitly offensive topic (topic 14), and those tweets are classified with a 100% accuracy. The accuracy score is highest on Category 2 and lowest on Category 3. For the *Founta-Hateful* class, the classifier recognizes 62% of the tweets correctly. The accuracy score is highest on Category 1 and lowest on Category 3. 8% of the *Founta-Hateful* tweets fall under the explicitly offensive topic (topic 14), and are classified with a 99% accuracy. For the *Founta-Normal* class, the classifier recognizes 96% of the tweets correctly. Unlike the *Founta-Abusive* and *Founta-Hateful* class, for the *Founta-Normal* class, the highest accuracy is achieved on Category 3. 0.1% of the *Founta-Normal* tweets fall under the explicitly offensive topic, and only 26% of them are classified correctly.

The accuracy of the classifier on the *Waseem-Sexist* and *Waseem-Racist* classes is 0.26 and 0.35, respectively. This indicates that the *Wiki*-dataset, annotated for toxicity, is not well suited for detecting sexist or racist tweets. This observation could be explained by the fact that none of the coherent topics extracted from the *Wiki*-dataset is associated strongly with sexism or racism. Nevertheless, the tweets that fall under the explicit abuse topic (topic 14) are recognized with a 100% accuracy. Topic 8, which contains abuse mostly directed towards Jewish and Muslim people, is the most dominant topic in the *Racist* class (32% of the class) and the accuracy score on this topic is the highest, after the explicitly offensive topic. The *Racist* class overlaps the least with Category 3 (see Table 2), and the lowest accuracy score is obtained on this category. The definitions of the *Toxic* and *Racist* classes overlap mostly in general and identity-related abuse, therefore higher accuracy scores are obtained in Categories 1 and 2. Similar to *Racist* tweets, *Sexist* tweets have the least overlap and the lowest accuracy score on Category 3. The accuracy score is the highest on the explicitly offensive topic (100%) and varies substantially across other topics.

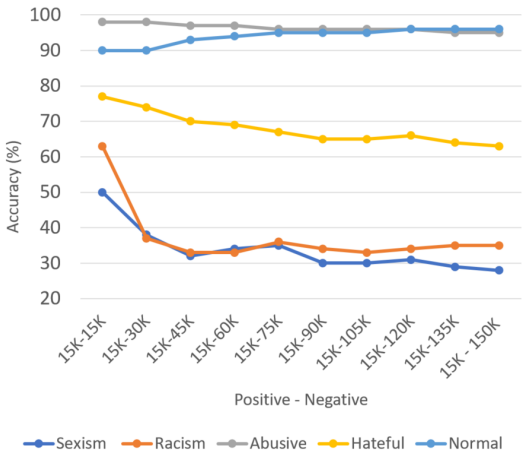


Figure 1: The classifier's performance on various classes when trained on subsets of the *Wiki*-dataset with specific class distributions.

6.1 Discussion

The generalizability of the classifier trained on the *Wiki*-dataset is affected by at least two factors: task formulation and topic distributions.

The impact of task formulation: From task formulations described in Section 3, observe that the *Wiki*-dataset defines the class *Toxic* in a general way. The class *Founta-Abusive* is also a general formulation of offensive behaviour. The similarity of these two definitions is reflected clearly in our results. The classifier trained on the *Wiki*-dataset reaches 96% accuracy on the *Founta-Abusive* class. Unlike the *Founta-Abusive* class, the other three labels included in our analysis formulate a specific type of harassment against certain targets. Our topic analysis of the *Wiki*-dataset reveals that this dataset includes profanity and hateful content directed towards minority groups but the dataset is extremely unbalanced in covering these topics. Therefore, not only is the number of useful examples for learning these classes small, but the classification models do not learn these classes effectively because of the skewness of the training dataset. This observation is in line with the fact that the trained classifier detects some of the *Waseem-Racist*, *Waseem-Sexist* and *Founta-Hateful* tweets correctly, but overall performs poorly on these classes.

The impact of topic distribution: Our analysis shows that independent of the class labels, for all the abuse-related test classes, the trained classifier performs worst when test examples fall under Category 3. Intuitively, this means that the platform-specific topics with low association with offensive language are least generalizable in terms of learning offensive behaviour. Categories 1 and 2, which include a mixture of general and identity-related topics with high potential for offensiveness, have more commonalities across datasets.

7 Impact of Data Size, Class and Topic Distribution on Generalizability

Our goal is to measure the impact of various topics on generalization. However, modifying the topic distribution will impact the class distribution and data size. To control for this, we first analyze the impact of class distribution and data size on the classifier's performance. Then, we study the effect of topic distribution by limiting the training data to different topic categories.

Impact of class distribution: The class distribution in the *Wiki*-dataset is fairly imbalanced; the ratio of the size of *Wiki-Toxic* to *Wiki-Normal* is 1:10. Class imbalance can lead to poor predictive performance on minority classes, as most of the learning algorithms are developed with the assumption of the balanced class distribution. To investigate the impact of the class distribution on generalization, we keep all the *Wiki-Toxic* instances and randomly sample the *Wiki-Normal* class to build the training sets with various ratios of toxic to normal instances.

Figure 1 shows the classifier's accuracy on the test classes when trained on subsets with different class distributions. Observe that with the increase of the *Wiki-Normal* class size in the training dataset, the accuracy on all offensive test classes decreases while the accuracy on the *Founta-Normal* class increases. The classifier assigns more instances to the the *Normal* class resulting in a lower True Positive (accuracy on the offensive classes) and a higher True Negative (accuracy on the *Normal* class) rates. The drop in accuracy is significant for the *Waseem-Sexist*, *Waseem-Racist* and *Waseem-Hateful* classes and relatively minor for the *Founta-Abusive* class. Note that the impact of the class distribution is not reflected in the overall F1-score. The classifier trained on a balanced data subset (with class size ratio of 1:1) reaches 0.896 weighted-averaged F1-score, which is very close to the F1-score of 0.899 resulted from training on the full dataset with the 1:10 class size ratio. However, in practice, the designers of such systems need to decide on the preferred class distribution depending on the distribution of classes in the test environment and the significance of the consequences of the False Positive and False Negative outcomes.

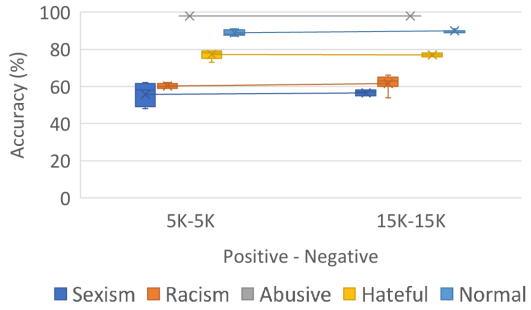


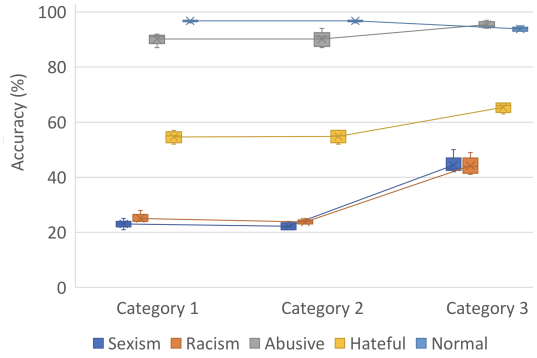
Figure 2: The classifier's average performance on various classes when trained on balanced subsets of the *Wiki*-dataset of different sizes.

Impact of dataset size: To investigate the impact of the size of the training set, we fix the class ratio at 1:1 and compare the classifier's performance when trained on data subsets of different sizes. We randomly select subsets from the *Wiki*-dataset with sizes of 10K (5K *Toxic* and 5K *Normal* instances) and 30K (15K *Toxic* and 15K *Normal* instances). Each experiment is repeated 5 times, and the averaged results are presented in Figure 2. The height of the box shows the standard deviation of accuracies. Observe that the average accuracies remain unchanged when the dataset's size triples at the same class balance ratio. This finding contrasts with the general assumption that more training data results in a higher classification performance.

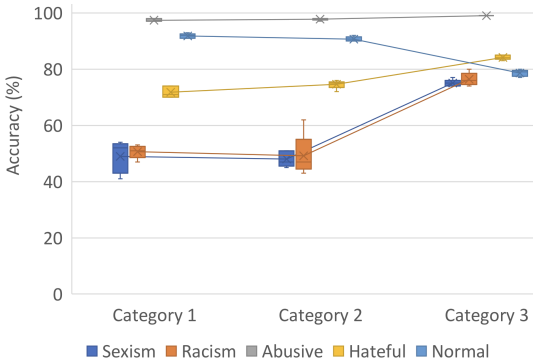
Impact of topics: In order to measure the impact of topics covered in the training dataset, we compare the classifier's performance when trained on only one of the three categories of topics described in Section 4. To control for the effect of class balance and dataset size, we run the experiments for two cases of toxic-to-normal ratios, 3K-3K and 3K-27K. Each experiment is repeated 5 times, and the average accuracy per class is reported in Figure 3.

For both cases of class size ratios, shown in Figures 3a and 3b, we notice that the classifier trained on instances belonging to Category 3 reaches higher accuracies on the offensive classes, but a significantly lower accuracy on the *Founta-Normal* class. The benign part of Category 3 is overwhelmed by Wikipedia-specific examples. Therefore, utterances dissimilar to these topics are labelled as *Toxic*, leading to a high accuracy on the toxic classes and a low accuracy on the *Normal* class. This is an example of the negative impact of topic bias on the detection of offensive utterances.

In contrast, the classifiers trained on Categories 1 and 2 perform comparably across test classes. The



(a) toxic - normal : 3K - 27K



(b) toxic - normal : 3K - 3K

Figure 3: The classifier's performance on various classes when trained on specific topic categories.

classifier trained on Category 2 is slightly more effective in recognizing *Founta-Hateful* utterances, especially when the training set is balanced. This observation can be explained by a better representation of identity-related hatred in Category 2.

8 Removing Platform-Specific Instances from the Training Set

We showed that a classifier trained on instances from Category 3 suffers a big loss in accuracy on the *Normal* class. Here, we investigate how the performance of a classifier trained on the full *Wiki*-dataset changes when the Category 3 instances (all or the benign part only) are removed from the training set. Table 4 shows the results. Observe that removing the domain-specific benign examples, referred to as 'excl. C3 *Normal*' in Table 4, improves the accuracies for all classes. As demonstrated in the previous experiments, this improvement cannot be attributed to the changes in the class balance ratio or the size of the training set, as both these factors cause a trade-off between True Positive and True Negative rates. Removing the Wikipedia-specific topics from the *Wiki*-dataset mitigates the topic bias and leads to this improvement.

Dataset/Class	Training Set		
	Wiki	***Wiki* excl. C3 *Normal***	***Wiki* excl. C3 all**
Toxic			
Founta-Abusive	0.94	0.96	0.95
Founta-Hateful	0.62	0.67	0.65
Waseem-Sexist	0.26	0.30	0.28
Waseem-Racist	0.35	0.40	0.32
Normal			
Founta-Normal	0.96	0.97	0.97

Table 4: Accuracy per Out-of-Domain test class for a classifier trained on the *Wiki*-dataset, and the *Wiki*-dataset with Category 3 instances (*Normal* only or all) excluded.

Test Set	Training Set		
	Wiki	***Wiki* excl. C3 *Normal***	***Wiki* excl. C3 all**
Out-of-Domain	0.90	0.91	0.91
In-Domain	0.97	0.97	0.97

Table 5: Weighted macro-averaged F1-score for a classifier trained on portions of the *Wiki*-dataset and evaluated on the in-domain and out-of-domain test sets.

Similarly, when all the instances of Category 3 are removed from the training set ('excl. C3 all' in Table 4), the accuracy does not suffer and actually slightly improves on all classes, except *Waseem-Racist*. This is despite the fact that the training set has 58% less instances in the *Normal* class and 18% less instances in the *Toxic* class. The overall weighted-averaged F1-score on the full Out-of-Domain test set also slightly improves when the instances of Category 3 are excluded from the training data (Table 5). Removing all the instances of Category 3 is particularly interesting since it can be done only with inspection of topics and without using the class labels.

To assess the impact of removing Wikipedia-specific examples on in-domain classification, we train a model on the training set of the *Wiki*-dataset, with and without excluding Category 3 instances, and evaluate it on the full test set of the *Wiki*-dataset. We observe that the in-domain performance does not suffer from removing Category 3 from the training data (Table 5).

9 Discussion

In the task of online abuse detection, both False Positive and False Negative errors can lead to significant harm as one threatens the freedom of speech and ruins people's reputations, and the other ignores hurtful behaviour. Although balancing the class sizes has been traditionally exploited when dealing with imbalanced datasets, we showed that balanced class sizes may lead to high misclassification of normal utterances while improving the True Positive rates. This trade-off is not necessarily reflected in aggregated evaluation metrics such as F1-score but has important implications in real-life applications. We suggest evaluating each class (both positive and negative) separately taking into account the potential costs of different types of errors. Furthermore, our analysis reveals that for generalizability, the size of the dataset is not as important as the class and topic distributions.

We analyzed the impact of the topics included in the *Wiki*-dataset and showed that mitigating the topic bias improves accuracy rates across all the out-of-domain positive and negative classes. Our results suggest that the sheer amount of normal comments included in the training datasets might not be necessary and can even be harmful for generalization if the topic distribution of normal topics is skewed. When the classifier is trained on Category 3 instances only (Figure 3), the *Normal* class is attributed to the over-represented topics, leading to high misclassification of normal texts or high False Positive rates.

In general, when collecting new datasets, texts can be inspected through topic modeling using simple heuristics (e.g., keep topics related to demographic groups often subjected to abuse) in an attempt to balance the distribution of various topics and possibly sub-sample over-represented and less generalizable topics (e.g., high volumes of messages related to an incident with a celebrity figure happened during the data collection time) before the expensive annotation step.

10 Conclusion

Our work highlights the importance of heuristic scrutinizing of topics in collected datasets before performing a laborious and expensive annotation. We suggest that unsupervised topic modeling and manual assessment of extracted topics can be used to mitigate the topic bias. In the case of the *Wiki*-dataset, we showed that more than half of the dataset can be safely removed without affecting either the in-domain or the out-of-domain performance. For future work, we recommend that topic analysis, augmentation of topics associated with offensive vocabulary and targeted demographics, and filtering of non-generalizable topics should be applied iteratively during data collection.

References

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Paula Branco, Luís Torgo, and Rita P. Ribeiro. 2016. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):1–50.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. 2017. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239.

Matthew Hoffman, Francis R. Bach, and David M. Blei. 2010. Online learning for latent Dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the Instagram social network. In *Proceedings of the International Conference on Social Informatics*, pages 49–66.

Justin M. Johnson and Taghi M. Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy.

Miaofeng Liu, Yan Song, Hongbin Zou, and Tong Zhang. 2019. Reinforced training data selection for domain adaptation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1957–1968, Florence, Italy.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the International Conference on World Wide Web*, pages 145–153.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 16–27.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 399–408.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 372–382.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *arXiv preprint arXiv:2004.01670*.

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

Bertie Vidgen, Helen Margetts, and Alex Harris. 2019b. How much online abuse is there? *Alan Turing Institute. November*, 27.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada.

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, Minneapolis, Minnesota.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1046–1056, New Orleans, Louisiana.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.