On Analyzing Annotation Consistency in Online Abusive Behavior Datasets

Md Rabiul Awal,[1] Rui Cao,[2] Roy Ka-Wei Lee,[1] Sandra Mitrović[3]
[1]University of Saskatchewan, Canada
[2]University of Electronic Science and Technology of China, China
[3]Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Switzerland
mda219@mail.usask.ca, caorui0503@gmail.com, roylee@cs.usask.ca, sandra.mitrovic@idsia.ch

Abstract

Online abusive behavior is an important issue that breaks the cohesiveness of online social communities and even raises public safety concerns in our societies. Motivated by this rising issue, researchers have proposed, collected, and annotated online abusive content datasets. These datasets play a critical role in facilitating the research on online hate speech and abusive behaviors. However, the annotation of such datasets is a difficult task; it is often contentious on what should be the true label of a given text as the semantic difference of the labels may be blurred (e.g., abusive and hate) and often subjective. In this study, we proposed an analytical framework to study the annotation consistency in online hate and abusive content datasets. We applied our proposed framework to evaluate the consistency of the annotation in three popular datasets that are widely used in online hate speech and abusive behavior studies. We found that there is still a substantial amount of annotation inconsistency in the existing datasets, particularly when the labels are semantically similar.

Introduction

Misbehavior in online social media such as cyberbullying, propagation of hate speeches, and abusive content have become an increasing problem. Such online misbehavior has not only sowed discord among individuals or communities online but also resulted in violent hate crimes Williams (2019); Relia et al. (2019); Mathew et al. (2019). Therefore, it is a pressing issue to detect and curb such misbehavior in online social media.

Traditional machine learning and deep learning approaches have been proposed to detect online misbehavior automatically. The recent surveys (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018) have comprehensively summarized these methods. Most of the automatic online misbehavior detection methods are supervised text classification methods trained and tested on annotated datasets. As such, the quality of the annotation has direct implications on detection algorithms' performance and the insights gained from the online misbehavior research studies.

Three popular datasets are widely used in online misbehavior studies: **WZ** (Waseem and Hovy, 2016; Waseem, 2016), **DT** (Davidson et al., 2017), and the recently published **FOUNTA** Founta et al. (2018) dataset. Waseem and Hovy (2016) first collected and annotated the **WZ** Twitter dataset into four classes: racism, sexism, both, and neither. Waseem and Hovy (2016) subsequently enhanced the dataset by controlling the bias introduced by annotators. Davidson et al. (2017) argued that hate speech should be differentiated from offensive tweets; some tweets may contain hateful words but should be labeled as offensive as they did not meet the threshold of classifying them as hate speech. The researchers collected the **DT** dataset and manually annotated the dataset into three categories: offensive, hate, and neither. In a recent study, Founta et al. (2018) proposed the **FOUNTA** dataset. This dataset went through two rounds of annotations. In the first round, annotators are required to classify tweets into three categories: normal, spam, and inappropriate. Subsequently, the annotators were asked to refine further the labels of the tweets in the "inappropriate" category. Specifically, the final version of the dataset consists of four classes: normal, spam, hate, and abusive.

While these datasets have facilitated many online misbehavior studies, few analyses have been done to evaluate and benchmark the quality of these datasets. The annotation of online misbehavior datasets is a challenging tasks. Firstly, the difference between certain labels may be subtle (Davidson et al., 2017; Founta et al., 2018). Secondly, the manual annotation process is often subjected to the annotator's biasness (Waseem, 2016). Therefore, we proposed an analytical framework to examine the annotation consistency in online misbehavior datasets. Included in our proposed framework is a two-step pipeline, which enables us to identify potential mislabeling and contentious annotation in the datasets.

We summarize our main contributions as follows:

- We proposed a novel analytical framework to examine the annotation consistency in the online misbehavior dataset.
- We applied our proposed framework to analyze three popular real-world and publicly available datasets. To the best of our knowledge, we are the first study that quantitatively and qualitatively compares existing online misbehavior datasets.
- Our analysis showed that there is a substantial amount of annotation inconsistency in the existing datasets. We also

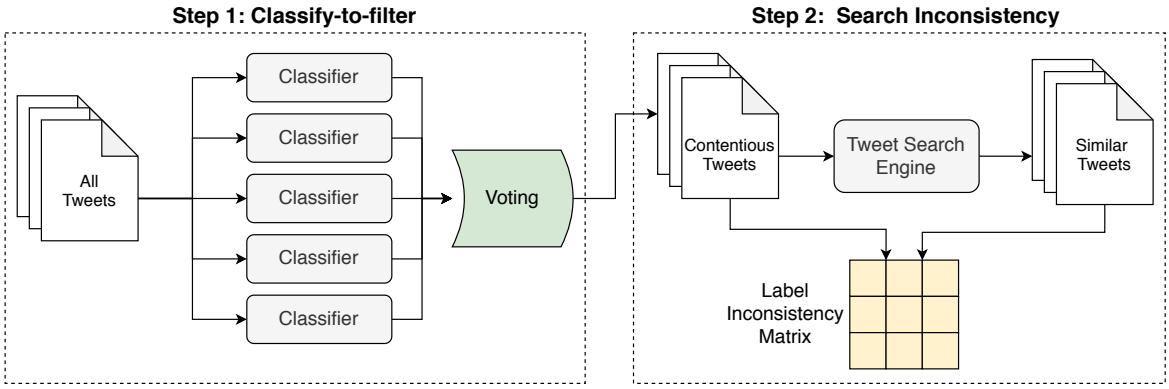

Figure 1: Overall annotation consistency analysis framework

empirically demonstrate case studies where the annotation inconsistency is likely to occur in the datasets.

Annotation Consistency Analysis Framework

Figure 1 shows our proposed annotation consistency analysis framework. Included in the analytical framework is a two-step process. In the first step, we train a set of classifiers to predict the labels of a given dataset of tweets. Voting will then be performed to vote for contentious tweets, i.e., tweets that are wrongly classified by more than half of the classifiers. The intuition is that it is more challenging to classify tweets that are annotated with contentious labels. For example, in Table 1, the tweet *t1* is identified as contentious when more than half of the classifiers mispredicted its label. A potential reason for the wrong classification may be due to *t1*, which is labeled as *Hate*, sharing very similar attributes with other tweets that are labeled as *Offensive*. Such contentious labeling is likely to confuse the classifier, resulting in the wrong prediction. In the second step, the set of retrieved contentious tweets are used as input queries into a search engine to find similar tweets in the dataset. Finally, we construct an annotation inconsistency matrix by comparing the labels of the contentious tweets and the retrieved similar tweets. The underlying assumption is that potential inconsistencies arise when the labels of the contentious tweet and its similar tweet are different. For example, in Table 1, the search engines return *t2* as the most similar tweet to the contentious tweet *t1*. When we compare the label of *t1* and *t2*, we notice that the two tweets have different labels, flagging a potential annotation inconsistency for the tweet *t1*.

Step 1: Classify-to-filter

The *classify-to-filter* step can be further broken down into two stages: *classification* and *voting*.

In the *classification* stage, we adopt an ensemble approach to train five different text classifiers on a given online misbehavior dataset. The commonly-used traditional machine learning and deep learning models are selected for our text classification task. Specifically, we use Logistic Regression (LR), Naive Bayes (NB), Single-layer Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Convolutional Long-Short Term Memory network (C-LSTM) as the classifiers in this step. For LR and NB, we trained these classifiers using the tweets' word-level term frequency-inverse document frequency (tf-idf) features. For the deep learning models, we use pre-trained GloVe word embeddings (Pennington, Socher, and Manning, 2014) to represent the words in the tweets, which are subsequently used as input for the classifiers. Each classifier is trained using 5-fold cross-validation, and the predictions on the tweets in the validation set are recorded for voting.

Table 1: Tweets example

Id	Tweet	Label	Contentious
t1	You are such a b*tch	Hate	Yes
t2	Don't be such a b*tch	Offensive	No
t3	B*tch please, try hard!	Offensive	No

In the *voting* stage, we consolidate the predictions made by five classifiers and identify the contentious tweets. Specifically, given a tweet, if three or more classifiers predicted its label wrongly, we would place this tweet into the *contentious tweets* set. While the incorrect prediction may be attributed to inconsistency in annotation, there could also be other reasons. For example, a tweet may contain rare words, and there are insufficient data to train the models well to classify this tweet. Therefore, we perform another step to further verify whether it is annotation inconsistency that led to incorrect predictions.

Step 2: Search Inconsistency

In this step, we utilize the retrieved set of contentious tweets as input into our search engine to retrieve similar tweets. Specifically, given a query contentious tweet, t_q, the search engine aims to retrieve its most similar tweet, t_s, from the dataset. To measure the similarity between tweets, we compute the cosine similarity between the tweets' tf-idf representation. The cosine similarity between two tweets are computed as follows:

$$cos\_sim(t_q, t_s) = \frac{\sum_{w \in t_q \cap t_s} t_q^w t_s^w}{\sqrt{\sum_{w \in t_q} t_q^{w2}} * \sqrt{\sum_{w \in t_s} t_s^{w2}}} \quad (1)$$

where t_q^w is the tf-idf weight of term w in the query tweet t_q,

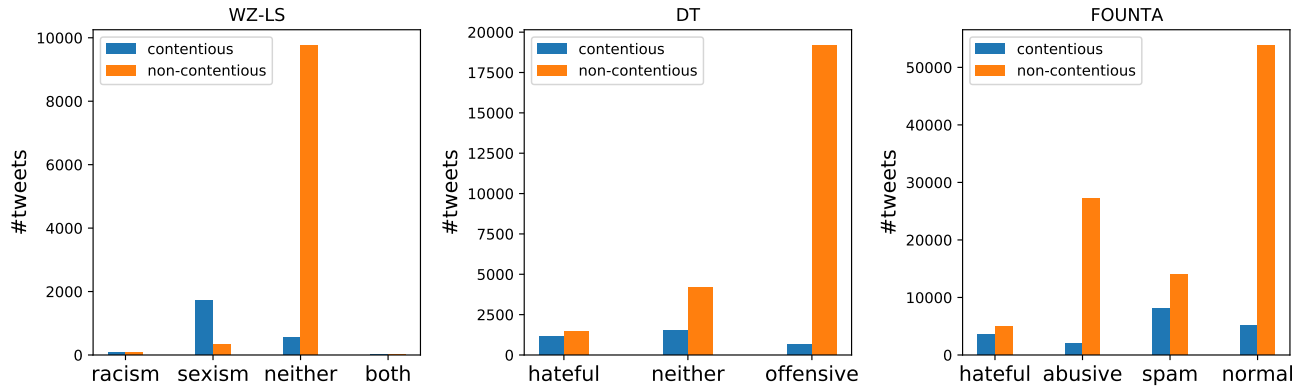

Figure 2: Breakdown distributions of contentious and non-contentious tweets from **WZ** (left), **DT** (middle), and **FOUNTA** (right) retrieved in step 1 of the annotation consistency analysis framework.

and t_s^w is the tf-idf weight of term w in the similar tweet t_s. We compute the cosine similarity between each query tweet t_q and all tweets in the dataset, i.e., $t_s \in T$, and select the tweet with the highest cosine similarity score as the similar tweet to the query tweet.

Finally, we compare the annotated labels of the t_q and t_s: if the two annotated labels disagree, we flag out that t_q might have an annotation inconsistency as the (a) classifiers find it hard to classify this tweet, and (b) its annotated label is different from its most similar tweet. The annotation inconsistencies in the contentious tweets are subsequently reported in annotation inconsistency matrices in the next step.

Evaluation and Discussion

We applied our proposed annotation consistency analysis framework on the three popular datasets, which are widely used in online misbehavior studies: **WZ** (Waseem and Hovy, 2016; Waseem, 2016), **DT** (Davidson et al., 2017), and **FOUNTA** Founta et al. (2018). The summary statistics of the datasets are presented in Table 2. Note that we combined the number the tweets in (Waseem and Hovy, 2016; Waseem, 2016) to form the current **WZ** dataset.

Table 2: Summary statistics of datasets

Dataset	#tweets	Classes (#tweets)
WZ	13,202	racism (82), sexism (3,332), both (21), neither (9,767)
DT	24,783	hate (1,430), offensive (19,190), neither (4163)
FOUNTA	99,999	normal (53,851), abusive (27,150), spam (14,029), hate (4,965)

Figure 2 shows the breakdown distributions of contentious and non-contentious tweets retrieved in step 1 of our proposed analytical framework. We observe that contentious tweets are found from all labels in the three datasets, i.e., the five classifiers made mistakes in predicting the true label of all kinds of tweets. Specifically, in **WZ**, the classifiers have incorrectly predicted most of the sexism tweets. In **DT**, almost half of the hateful tweets are wrongly predicted. Similar observations are made in **FOUNTA**, with hateful and spam tweets seeing a higher percentage of misclassification.

Table 3: Annotation Inconsistency matrix for WZ

		Contentious Tweet Label			
		Racism	Sexism	Both	Neither
Similar Tweet Label	Racism	16	0	0	1
	Sexism	9	662	10	222
	Both	0	4	0	1
	Neither	26	754	5	218

As discussed earlier in the section, there could be multiple reasons for the misclassification. For instance, the hate speech detection problem may be hard as the tweets within the same label have high variance, or there might be insufficient training data. In this paper, we are interested to understand how much of the misclassification can be attributed to annotation inconsistency. Table 3, 4, and 5 shows the annotation inconsistency matrix generated in step 2 of our analytical framework for **WZ**, **DT**, and **FOUNTA** respectively.

Table 4: Annotation Inconsistency matrix for DT

		Contentious Tweet Label		
		Offensive	Hate	Neither
Similar Tweet Label	Offensive	282	760	282
	Hate	84	133	16
	Neither	105	41	74

From Table 3, we observe that 662 sexism contentious tweets have their most similar tweets sharing the same label, while 745 of the sexism contentious tweets have their most similar tweets labeled as normal tweets (i.e., neither). This suggests that there could be inconsistencies in the annotation of sexism tweets as two similar tweets may have different labels, one labeled as sexism while another as normal. Similar observations are made in other class labels, although the inconsistency in sexism tweet annotation is ob-

Table 5: Annotation Inconsistency matrix for FOUNTA

		Contentious Tweet Label			
		Abusive	Hate	Spam	Normal
Similar Tweet Label	Abusive	491	1547	736	1062
	Hate	347	370	93	192
	Spam	109	62	790	1024
	Normal	758	1133	3170	915

served to be the highest. Similar observations are also made for the **DT** dataset in Table 4. A majority of the contentious hate tweets have their most similar tweets labeled as offensive. This is unsurprising as even for human annotators it is often difficult to differentiate hateful tweets from offensive ones (Davidson et al., 2017). Nevertheless, such challenges in annotation also highlight the difficulty in the hate speech detection task.

Comparing the annotation inconsistency matrix of **FOUNTA** against the other two datasets, we noted that there could be significantly more annotation inconsistencies in the **FOUNTA** dataset. As shown in Table 5, there is a high amount of annotation inconsistencies observed in all labels. For instance, we observed that 758 contentious abusive tweets have their most similar tweets labeled as normal, and a significant number of contentious hate tweets have their most similar tweets labeled as abusive or normal. We further verify the annotation inconsistencies in **FOUNTA** dataset by retrieving some samples of the **FOUNTA** tweets. Table 1 shows three examples of the **FOUNTA** contentious tweets and their most similar tweets. Surprisingly, we notice that the most similar tweets retrieved for contentious tweets C1 and C2 are retweets, and the retweets are annotated with different class labels. This exposes an issue in FOUNTA's annotation strategy. We postulated that the identical tweets (i.e., the retweets) are annotated by different human annotators, resulting in the inconsistencies. We further investigated and found that more than 10% of the tweets are retweets, and most of which have inconsistencies in their annotation.

Table 6: Examples of tweets from **FOUNTA** dataset. C*x* denotes the contentious tweet and S*x* denotes the corresponding most similar tweet.

Id	Tweet	Label
C1	RT:[USER_1] How about we f**king hire trans boys to play trans boys	hate
S1	RT:[USER_1] How about we f**king hire trans boys to play trans boys	normal
C2	RT:[USER_2] I wish I wasn't so annoying like I even p*ss myself off	normal
S2	RT:[USER_2] I wish I wasn't so annoying like I even p*ss myself off	abusive
C3	RT:[USER_3] f**king faggot	hate
S3	[USER_4] f**king faggot	abusive

Conclusion

In this paper, we proposed an analytical framework to examine annotation consistency in online misbehavior datasets. We applied our proposed framework to analyze three popular online misbehavior datasets. Our analysis showed that annotation inconsistencies in all three datasets, illustrating the challenges in online misbehavior data collection. Specifically, in the **FOUNTA** dataset, we found a significant amount of annotation inconsistency where identical tweets are annotated with different class labels. We also provided the updated datasets[1] with annotation inconsistency information so that researchers may perform the necessary data preprocessing in future online misbehavior studies.

References

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Fortuna, P., and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51(4):1–30.

Founta, A. M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Mathew, B.; Dutt, R.; Goyal, P.; and Mukherjee, A. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, 173–182.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Relia, K.; Li, Z.; Cook, S. H.; and Chunara, R. 2019. Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 us cities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 417–427.

Schmidt, A., and Wiegand, M. 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*, 1–10.

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

Waseem, Z. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, 138–142.

Williams, M. 2019. The connection between online hate speech and real-world hate crime.

[1] https://gitlab.com/bottle_shop/abusive/annotation_framework.