In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets

Kosisochukwu Judith Madukwe, Xiaoying Gao, Bing Xue
School of Engineering and Computer Science
Victoria University of Wellington
PO Box 600 ,Wellington 6012, New Zealand
{kosisochukwu.madukwe,xiaoying.gao,bing.xue}@ecs.vuw.ac.nz

Abstract

Recently, a few studies have discussed the limitations of datasets collected for the task of detecting hate speech from different viewpoints. We intend to contribute to the conversation by providing a consolidated overview of these issues pertaining to the data that debilitate research in this area. Specifically, we discuss how the varying pre-processing steps and the format for making data publicly available result in highly varying datasets that make an objective comparison between studies difficult and unfair. There is currently no study (to the best of our knowledge) focused on comparing the attributes of existing datasets for hate speech detection, outlining their limitations and recommending approaches for future research. This work intends to fill that gap and become the one-stop shop for information regarding hate speech datasets.

1 Introduction

It is imperative to detect hateful speech on social media platforms and other online spaces because its real life implications are usually dire. The research community working towards achieving this goal spans from the Social Sciences to Computer Science. Under the field of Computer Science, Natural Language Processing (NLP) and Machine Learning (ML) techniques have been applied to this task of detecting hate speech by mostly framing it as a text classification task. Here, text is classified into different categories based on its innate content or features. Text classification is a supervised ML task; which means it requires a considerable amount of labelled data. Each data instance needs a label or a class/category that it belongs to. Although the majority of the studies in this research area use labelled data as they conduct a classification task, there are some that do not (Gao et al., 2017; Xiang et al., 2012).

In this study, we concentrate on datasets for hate speech detection in the English language while briefly highlighting other languages and similar concepts such as cyberbullying and abuse detection. The same issue discussed here are also true in other languages, thus all suggested solutions would persist.

The overall aim of this work is to provide insight into the existing datasets and a consolidated analysis into their strengths and weaknesses and most importantly suggest methods to forward research in this area. To achieve this, we ask several questions:

- What makes a dataset benchmark?
- How do we handle class imbalanced dataset? In its unbalanced form or not?
- What typology should we follow for hate speech research? What should or shouldn't it include?
- What is the best ethical format for collating and sharing such a sensitive dataset so as to avoid data degradation?

Although this work will be critiquing a few studies, it is not meant to be negative in any form.

2 Motivations

The importance of hate speech detection research cannot be overemphasised. Now, more than ever, with the current inflammatory political climate and discourse all around the world and minorities in various locations demanding for equality and equity, we cannot allow additional bias to be introduced into their lives through artificial intelligence. The problem of hate speech detection is one yet to be solved even to an acceptable level. It would be counter-productive if all the research efforts are not focused and channeled towards a better tomorrow

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 150–161

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

by building on top one another. So we were motivated to go back to a root of the problem: the data. One of the foundations of this research work (that we can easily make changes on) is the data set. We can only build solid structures on solid foundations. Furthermore, research efforts would be futile if the proposed state-of-the-art for this task fail to perform well on a realistic dataset.

3 Summary of Existing Datasets

In this section, we highlight the currently existing datasets used in literature for the task of detecting hate speech. In the broad area of abusive language detection, there exists several other datasets collected and annotated for cyberbullying, toxicity, aggression and so on (we would not discuss those in-depth as they are out of the scope of this work). As highlighted in (Fortuna and Nunes, 2018), the majority of the studies in this area of hate speech detection collected and annotated their own datasets, however, some were not made publicly available. The existing datasets are:

1. **BURNAP Dataset**: This dataset collected by (Burnap and Williams, 2016) comprises of cyber-hate targeted at four different protected characteristics (sexual orientation, race, disability and religion) in roughly equal amounts. Of the annotated sample, 10.15% of sexual orientation category, 3.73% of race category, 2.66% of disability category and 11.68% of religion category are considered offensive or antagonistic. The dataset was collected after different trigger events for each category.

2. **WASEEM Dataset**[1]**:** This dataset was published by (Waseem and Hovy, 2016). It contains 16k English tweets annotated into three classes (1972 are Racism, 3383 are Sexism and 11559 are Neither) and was made publicly available using TweetIDs. The authors annotated the data themselves, then used a third party to validate the annotations. They record an inter-annotator agreement of 0.84. This dataset is unbalanced and also biased toward specific users since all of the tweets labelled as racist where from 9 users only, while the other classes were from more than 600 users. This dataset was extended in (Waseem, 2016) by 4033 additional tweets, were they experimented with amateur and expert annotations to investigate their influence based on an existing knowledge of the research area.

3. **DAVIDSON Dataset**[2]**:** This was published by (Davidson et al., 2017). The dataset contains 24,802 tweets in English (5.77% labelled as Hate speech, 77.43% as Offensive and 16.80% as Neither) and was published in raw text format. They report collecting this data from Twitter using a lexicon from HateBase[3] containing hateful words and phrases. They used a crowdsourcing platform (Figure-Eight[4] formerly CrowdFlower) for annotating the tweets into the 3 classes. The annotators were provided with the authors' definitions and specific instructions. They record an inter-rater agreement of 92% as provided by the crowdsourcing platform.

4. **FOUNTA Dataset**[5]**:** (Founta et al., 2018) published a dataset of 80k tweets, annotated for various abusive behaviors (abusive, hateful speech, spam, normal) and made publicly available using TweetIDs. They use a boosted random sampling technique through an iterative and incremental process to generate the final dataset in order to improve the number of derogatory samples. They use a larger number of annotators (20) through crowdsourcing. Their classes are None at 59%, Spam at 22.5%, Abusive at 11% and Hateful at 7.5%. Recently, as part of the ICWSM Data challenge, an updated version of this dataset, now containing 100k was made available in text format.

5. **WARNER Dataset**: The constituent data was collated by (Warner and Hirschberg, 2012) from Yahoo News Group and URLs from the American Jewish Society. It contains 9000 paragraphs, manually annotated into seven (7) categories (anti-semitic, anti-black, anti-Asian, anti-woman, anti-Muslim, anti-immigrant or other hate(anti-gay and anti-white)). It doesn't seem to be publicly available.

[1] https://github.com/ZeerakW/hatespeech

[2] https://github.com/t-davidson/hate-speech-and-offensive-language

[3] https://hatebase.org/

[4] https://www.figure-eight.com/

[5] https://dataverse.mpi-sws.org/dataset.xhtml?persistentId=doi:10.5072/FK2/ZDTEMN

6. **DJURIC Dataset**: (Djuric et al., 2015) collected comments from the Yahoo Finance website. 56,280 comments were labeled as hateful while 895,456 labeled as clean from 209,776 users.

7. **NOBATA Dataset**: The authors in (Nobata et al., 2016) collected data from Yahoo Finance and News comment section. Their definition of abusive language conflates hate speech, profanity and derogatory language. It was labelled as clean or abusive by Yahoo employees. In the primary dataset, 7.0% of Finance and 16.4% News comment were labelled as abusive. In the temporal dataset, 3.4% of Finance and 10.7% News comment were labelled as abusive. The dataset was reported to be at https://webscope.sandbox.yahoo.com/, however it currently cannot be found. They reported an annotation agreement rate of 0.867 and Fleiss Kappa of 0.401.

8. **ZHANG Dataset**[6]: The authors in (Zhang et al., 2018) created a dataset using refugee and muslim specific words and hashtags from Twitter. The dataset contains 2,435 tweets with 414 labelled as hate and 2,021 labelled as non-hate. The dataset was initially publicly available but not anymore due to the data sharing policy of the authors' institution.

9. **QIAN Dataset**[7]: (Qian et al., 2019) collected data from Reddit and Gab including intervention responses written by humans. Their data preserves the conversational thread as a way to provide context. From Reddit, they collect 5,020 conversations which includes a total of 22,324 comments labelled as hate or non-hate. 76.6% of the conversations contain hate speech while only 23.5% of the comments are labelled as hateful. They were mined from known toxic subbreddit using hate keywords. Similarly, from Gab, they collected 11,825 conversations containing 33,776 posts. 94.5% of the conversations contained hate speech while about 43.2% of the comments are labelled as hateful. Each entry in the dataset is a conversation of several indexed comments. The index (in another column) is used to identify which comment is considered hateful, then a response intervention is provided. The entries with no hate speech do not have an intervention response. The number of responses do not correspond to the number of hateful comments in the conversation. Therefore, a conversation with 5 hateful comments can have just 3 responses to intervene.

10. **HATEVAL Dataset**[8]: This is a very small dataset for detecting hate speech against women and immigrants. It contains English and Spanish tweets labelled into hateful or not hateful.

    In other languages, hate speech detection research have also progressed.

11. **ROSS Dataset**[9]: (Ross et al., 2017) collected and annotated 541 German tweets with key hashtags on the refugee crisis that could be offensive. The tweets were rated on their level of offensiveness on a 6 point Likert scale. They reported a Krippendorff's alpha from 0.18 to 0.29.

12. **BENIKOVA Dataset**[10]: (Benikova et al., 2018) contains 36 German tweets with 33% labelled as hatespeech and 67% as non-hatespeech.

13. **VIGNA Dataset**: (Vigna et al., 2017) labeled 17,567 Facebook comments from 99 posts as No hate, Weak hate and Strong hate. They recorded a Fleiss' kappa inter-annotator agreement metric of 0.19 with 5 annotators. It doesn't seem to be publicly available.

14. **EVALITA Dataset**[11]: EVALITA[12] published two datasets in Italian in 2018 and 2020 for a shared task in hate speech detection.

15. **TULKENS Dataset**: (Tulkens et al., 2016) crawled and collected data from comments on Dutch Facebook pages most likely to contain derogatory statements such as a Belgian anti-islamic organization and a right-wing organization. They recorded an inter-annotator

[6] https://github.com/ziqizhang/data#hate

[7] https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech

[8] https://competitions.codalab.org/competitions/19935#phases

[9] UCSM-DUE/IWG_hatespeech_public

[10] github.com/MeDarina/HateSpeechImplicit

[11] https://github.com/msang/haspeede

[12] "http://www.evalita.it/2020"

agreement using the Cohen Kappa score of 0.60. The train set contains 5,424 comments while the test set contains 607 comments labelled as non-racist and racist. The dataset is not publicly available, however the dictionary used can be accessed at https://github.com/clips/hades

Finally, since hate speech can occur in different modes such as text, images, audio and video, there are some multimodal datasets to address this issue:

16. **MMHS150K Dataset**[13]: (Gomez et al., 2019) made publicly available a multimodal (image and text) dataset collected from Twitter using Hatebase terms. It contains 150,000 tweets manually annotated into six classes of No attacks to any community, Racist, Sexist, Homophobic, Religion based attacks or Attacks to other communities.

17. **HATEFUL MEMES Dataset**: Facebook AI (Kiela et al., 2020) collected a multimodal dataset for detecting and classification of hate speech containing images and text. It was annotated using their specific definition of hate speech. It contains 10k memes with a 5% dev and 10% test set. The memes belong to the following classes: multimodal hate (benign confounders were found for both modalities), unimodal hate (one or both modalities were already hateful on their own), benign text confounder, benign image confounder, random non-hateful. A benign confounder is defined as "*a minimum replacement image or replacement text that flips the label for a given multimodal meme from hateful to non-hateful.*" They record a Cohen's kappa score (inter annotators reliability) of 67.2%. The dataset is available upon joining a currently ongoing competition[14].

Datasets	Availability	Classes/Labels	Size	Format
1	No	Sexual Orientation Race Disability Religion	-	-
2	Yes	Racism Sexism Neither	11.69% 20.00% 68.33% 16,914 tweets	TweetID
3	Yes	Hate Speech Offensive Neither	5.77% 77.43% 16.80% 24k tweets	Raw text
4	Yes	Abusive Hateful Spam Normal/None	11% 7.5% 22.5% 59% 80,000 tweets	TweetID
5	No	Anti-Semitic Anti-Black Anti-Asian Anti-Woman Anti-Muslim Anti-Immigrant Other hate	-	-
6	No	Hate Speech Clean	5.91% 94.08% 951,736 comments	- -
7	No	Abusive Clean	7 %of F + 16.4% of N 3.4 %of F + 10.7% of N	- -
9	Yes	Hate Speech Non-Hate Speech	23.5% 76.5% 22,324 Reddit comments	Raw text
9	Yes	Hate Speech Non-Hate Speech	43.2% 51.8% 33,776 Gab comments	Raw text
11	Yes	6 Point Likert Scale	- 541 tweets	-
12	Yes	Hate Speech Non-Hate Speech	33% 67% 33 tweets	-
13	No	No Hate Weak Hate Strong Hate	- 6,031 Facebook comments	-
15	No	Racist Non-Racist	- 17,567 Facebook comments	-

Table 1: Analysis of some of the existing hate speech datasets

4 The Need For A Benchmark Dataset

In the field of ML, benchmark datasets are datasets used to evaluate or compare the performance of ML methods on a particular task. It is used by researchers to test how their new ideas perform against existing ones (Caruana and Niculescu-Mizil, 2006) and to objectively measure progress on a particular problem. The dataset is usually the only necessary consistent/constant aspect of a study. Benchmark datasets have been shown in areas like image processing to be of paramount importance in enabling research progress and a fair/objective comparison between studies and proposed methods. Datasets like CIFAR10, CIFAR100 (Krizhevsky, 2009) and MNIST (LeCun and Cortes, 2010) for image processing and computer vision were published and are maintained by a large research institution. The CIFAR10 and CIFAR100 have designated train and test sets, which makes comparison between studies and proposed methods fair.

[13] https://gombru.github.io/2019/10/09/MMHS/

[14] https://www.drivendata.org/competitions/64/hateful-memes/page/205/

4.1 Dataset Accessibility and Availability

In Table 1, we show the state of availability and accessibility of some of the discussed datasets. Making datasets available on personal repositories is problematic because the user can take it down at anytime. For example, a hate speech dataset listed in (Fortuna and Nunes, 2018) on Annie Thorbun's personal github page[15] does not exist anymore. This problem can also occur when a website address changes. For example, in (Watanabe et al., 2018), one of the dataset used was listed to be at `www.crowdflower.com/data-for-everyone/` which now redirects to `https://appen.com/resources/datasets/`. However, the dataset cannot be found as at 19th June, 2020

Data degradation occurs when a dataset, published in an encrypted format, needs to be regenerated by the researcher on-demand, does not produce the same number/amount of data as on the publication date. This phenomenon occurs with hate speech data harvested from Twitter and published in form of tweetIDs which are identification number that linked to each individual tweet. In some cases, the author of the tweet deletes it, or the account owner deactivates the account, or it might be reported to Twitter as breaking one of their guidelines and Twitter takes it down. This has been reported in (Zhang and Luo, 2018; Arango et al., 2019). Also (Watanabe et al., 2018), noted that the WASEEM dataset had only 6,655 tweets left, out of the 6,909 initially published. (Osho et al., 2020) reported that for FOUNTA dataset they only found 69k out of 80k tweets. As compared to the distribution highlighted in Table 1, the new distribution over the classes were now 62% normal, 20% as abusive, 14% as spam and 4% as hateful. The hateful class was even more reduced. Both the FOUNTA and WASEEM data suffer from data degradation. As at June 2020, we found that the first batch of WASEEM data was completely degraded while the second batch has only 2,412 out of 6,090 tweets left. We also found that the FOUNTA data has 18,943 tweets out of the 80,000 left. The already minute class of interest bears the brunt of this phenomena.

For a persistent benchmark dataset to succeed, we need to make data available in a better format. The nature of the data and the fact that it provides a consolidated source of harmful information makes it very tricky. Therefore, we suggest a submission portal for the data, where each researcher can request for a copy of the data using a verifiable email address and then a copy of the benchmark dataset is sent to them. This might restrict access for those that might want to use this data for malicious purposes. This service can be provided by large institutional data repositories like Dataverse[16] or ICPSR[17].

4.2 Class Imbalance Issue

Unlike most text classification task such as sentiment analysis; hate speech detection suffers from a severe class imbalance issue, with the hate class being in most cases less than 12% for the multi-class datasets and less than half of the total dataset for the binary datasets (Table 1) .

Usually when the classes in a dataset are unbalanced, it is because one of the following reasons: Its either

- the data is rarely occurring (more specifically the class of interest is rare compared to the other class(es))
- or the data collection and labelling is difficult, time consuming and expensive;
- or the overlap between the classes is high.

For the hate speech detection task, it is all of the above. It becomes increasingly difficult to train ML algorithms on such small samples, which leads to subpar performance. The class imbalance problem is probably inevitable when collecting data, as there is an estimated maximum of 3% derogatory tweets on Twitter (Founta et al., 2018). Thus, the open question of whether to work with the dataset in its unbalanced form or to look into methods to make it balanced remains unanswered. It is desirable to develop a model that does a good job in identifying hateful instances even with the small sample size. Certainly, such a model will perform well in real life scenarios during deployment. Therefore a naturally occurring question is; *Are the methods for learning with a small data size more easily accessible and less computationally expensive than methods for reducing the class imbalance?* It is worthwhile to look into both and compare. Several studies (Davidson et al., 2017; Founta et al., 2019;

[15] `https://github.com/anniethorburn/Hate-Speech-M`

[16] `https://dataverse.org/`

[17] `https://www.icpsr.umich.edu/web/pages/index.html`

Madukwe and Gao, 2019; Mozafari et al., 2020; Zhang et al., 2018) have used the datasets in its unbalanced form with the claim that since this is the naturally occurring state, it shouldn't be altered. However, we argue that this is not advantageous to existing supervised ML algorithms that depend on a large supply of data with balanced classes for optimum performance. Similarly, (Swamy et al., 2019) showed that models generalize better when trained on data containing a high amount of samples in the positive class which also unfortunately the minority class in most datasets.

Since the collection and annotation of data for this task is time-consuming, expensive, error-prone with low yield, we recommend more studies into the best way to augment existing data. This would assist in increasing the data size and inadvertently solving the class imbalance problem. A few studies have discussed and proposed solution for augmenting related datasets (Chung et al., 2019; Karatsalos and Panagiotakis, 2020; Sharifirad et al., 2018). However, employing data augmentation as a preprocessing step to cater to the class imbalance problem will lead to an unfair comparison amongst other proposed solutions as there are wide of augmentation techniques. Also, data augmentation methods such as oversampling the minority class not done right (Agrawal and Awekar, 2018), will introduce bias into the model (Arango et al., 2019). Another suggestion is to look into ML methods that are unaffected by the class size such as one-class and active learning. Rigorous investigations are required to answer the question of how to handle class imbalance in hate speech datasets.

4.3 Varying Definitions and How it Affects Annotation

It is known that there are varying definitions of hate speech, however there are some consistencies amongst them. (Fortuna and Nunes, 2018) have analysed some available definitions of hate speech and highlighted the major similarities amongst them. Specifically, hate speech:

- has a specific target.
- incites violence or hate.
- attacks or diminishes.
- can contain humor or sarcasm.

Varying definitions imply that, of course, it might be impossible to rid social media platforms completely of hateful instances. Despite this fact, the agreed upon similarities is a good place to start. Currently, existing datasets are affected by these variations because the annotations are powered by the definitions. Thus, similar instances can fall under different annotation categories. (Ross et al., 2017) investigated the effects of the presence and absence of a definition during annotation on the annotation reliability of a hate speech dataset. They conclude that hate speech requires a stronger definition. Similarly, (Fortuna et al., 2020) empirically find that most of the publicly available datasets are incompatible due to different definitions assigned to similar concepts.

In order to measure the annotation reliability of the labels in a dataset, a numerical index known as the Inter-Rater/Inter-Coder/Inter-Annotator Agreement (Artstein and Poesio, 2008) is usually adopted. The studies that collected data, use it to measure the level of agreement among their annotators on the labels they chose for each text or sentence. Examples of this score are Fleiss (Fleiss, 1971) or Cohens (Cohen, 1960) Kappa. This score is affected by annotator bias and imbalance in the classes making it unreliable. In addition, different studies suggest different thresholds for acceptable annotation (Di Eugenio and Glass, 2004; Artstein and Poesio, 2008). As can be seen from the datasets highlighted in Section 3, the annotation reliability is relatively low. In (Awal et al., 2020), the authors propose a framework to analyse the annotation inconsistency in the WASEEM, DAVIDSON and FOUNTA dataset. They found major inconsistencies in the labels of all the three dataset most especially in FOUNTA dataset where duplicate tweets exists in great number and the exact same tweet can have opposing labels. ML models built on this data will find it difficult to learn anything useful. Additionally, using different names for the same concept can be misleading. (Waseem et al., 2017) examined the relationship between abusive language, hate speech, cyberbullying and trolling. A lax use of typology affects annotation. For example in (Wiegand et al., 2019) they conflated the racism and sexism class in WASEEM data into one class and changed the labels to Abuse and No Abuse.

4.4 Conflating Classes/Labels

Hate speech datasets sometimes have very similar labels and some studies merge some of them together into one class, often as a way to combat the

level of class imbalance. However, this conflation could negatively affect research progress as distinction between them is very necessary. One example is the DAVIDSON data with the Offensive and Hate class or the WASEEM data with the Racist and Sexist class. Classes in the DAVIDSON data were conflated in (Zhang and Luo, 2018; Zhang et al., 2018) where they merged the Hate and Offensive class into one class while (Miok et al., 2019) conflated the Offensive and Neither class into a Non-hate class. (Watanabe et al., 2018; Wiegand et al., 2019) conflated classes in the WASEEM data and for the FOUNTA data, (Davidson and Bhattacharya, 2020) deleted the Spam class and conflated the Hate and Abusive Class into Abusive.

These last two sections (4.3 and 4.4) affect the typology used in this research. There aren't any enforced or strict demarcations, therefore the use of varying terms to mean one thing negatively affects research progress. An author searching for hate speech data or studies, might miss out on ones that used abusive language or toxic comment as an umbrella term encompassing several paradigms. We suggest that the terms be used strictly following the available definitions. Similar to suggestion in (Davidson et al., 2017), offensive language is not the same as hate speech and should not be merged. Also, abusive language and cyberbullying should not be merged with hate speech.

4.5 Varying Preprocessing Steps and Train-Test Splits

Social media data is often very noisy since it is a user-generated data. Different researchers have employed varying steps to clean the data in preparation for an ML algorithm. We show that these choice of steps can affect the data size, therefore obstructing an objective comparison between studies even more. Table 2 shows a few papers using three commonly used hate speech datasets and the preprocessing applied which leads to variations that negatively affect a fair comparison. Some of the existing studies select different train-test splits such as 70:30 or 80:20, some do a train-test-validation split of 70:15:15 or 60:20:20 or 80:10:10 while some do a 10-fold or 5-fold cross validation. This varying setting means that fair comparison amongst studies is not possible except if every researcher reruns all existing studies they wish to compare with. This is both impractical and costly.

Datasets	Paper	Stem or Lemmatize	Username	URLs	Lowercase	Hashtags	Remove Punctuation	Remove Stopwords	Train-Test Split	Final DataSize
WASEEM	**(Badjatiya et al., 2017)**	-	replaced	replaced	added <allcaps> after an all capitalized word	replaced # sign with <hashtag>	No. Repetition replaced with <repeat>	-	-	-
	(Founta et al., 2019)	both	counted	counted	No. counted all capital words	counted	-	Yes	-	16,059
	(Mozafari et al., 2020)	-	replaced with placeholder <user>	replaced with placeholder <url>	Yes	removed # sign only	Yes	No	-	-
DAVIDSON	**(Davidson et al., 2017)**	stem	counted	counted	Yes	counted	-	-	5-fold CV	24,802
	(Malmasi and Zampieri, 2017, 2018)	-	removed	removed	Yes	-	-	-	10-fold CV	14,509
	(Founta et al., 2019)	both	counted	counted	No. counted all capital words	counted	-	Yes	-	24,783
	(Madukwe and Gao, 2019)	lemmatize	removed	removed	Yes	removed	Yes	Yes. Custom stop words	75/25	-
	(Mozafari et al., 2020)	-	replaced with placeholder <user>	replaced with placeholder <url>	Yes	removed # sign only	Yes	No	-	-
	(Miok et al., 2019)	lemmatize	remove	remove	-	expanded into words	Yes	Yes	-	3000
FOUNTA	**(Verma et al., 2020)**	-	replaced	replaced	Yes	Dropped # sign only	No	Yes	80/10/10	-
	(Liu et al., 2020)	-	-	removed	-	removed	Yes	-	80% 20%	99603 from 100000
	(Davidson et al., 2017)	-	replaced	replaced	-	-	Yes	Yes	-	75,023 from 100000
	(Kim et al., 2020)	Stem	-	-	Yes	-	-	-	80 /20	-

Table 2: Varying Pre-processing Steps

4.6 What Makes a Dataset Benchmark

Here, we highlight factors that qualifies a dataset to be considered as benchmark.

- A publicly available dataset: The dataset should be considerably easy to access by potential researchers. This will increase the chances of these researchers to use the dataset to measure the performance of their proposed methods.

- Consistent Train-Test-Validation Split: Likewise, this will contribute to fairer comparison between studies.

- Accessible data format: The data should preferably be in a format that does not degrade or change in time. Therefore the exact same dataset is available to Researcher A now and Researcher Z later.

- Absence of bias: A benchmark data lacks (for the most part) bias. A benchmark dataset for hate speech detection needs to be devoid of racial (Davidson et al., 2019; Sap et al., 2019), gender (Park et al., 2018) or intersectional (Kim et al., 2020) biases. Bias introduced by the data collection process was discussed in (Wiegand et al., 2019). Likewise, (Waseem et al., 2018) noted that more than 2k tweets in the DAVIDSON dataset, written in African American Vernacular English were labeled as hateful or offensive simply because they used the n-word. A diverse group of annotators would have significantly reduced this bias. In (Arango et al., 2019), they showed that a bias in user distribution adversely affected the generalization ability of the proposed models. Therefore, it is important that benchmark dataset are not biased towards particular users and that information on the distribution of the users whose tweets make up the dataset are provided in an anonymized format. (Davidson and Bhattacharya, 2020) reported that in the FOUNTA dataset, there are several duplicated tweets which can introduce a strong bias in the model as some instances are contained in both the training and testing sets.

- A common evaluation method/metric: Different studies use different metrics which affect comparison without re-implementation which might not be feasible if the said method is expensive to re-implement. Also, some metric choices do not reflect the true performance of the proposed methods.

  (Olteanu et al., 2017) argues for evaluation metrics that are directly proportional to user perception of correctness, thus more human-centered.

- It should be preferably pre-processed to an extent. If this is not feasible, then the authors should endeavor to make their pre-processing code public so that other researchers can apply it to keep the resulting dataset consistent and uniform.

Datasets	Publicly Available	Consistent Split	Accessible data format	Common Evaluation Metric	Unbiased	Pre-processed
WASEEM	✓	✗	✗	✗	✗	✗
DAVIDSON	✓	✗	✓	✗	✗	✗
FOUNTA	✓	✗	✗	✗	✗	✗
QIAN	✓	✗	✓	✗	✗	✗
HATEVAL	✓	✗	✓	✗	✗	✗

Table 3: Benchmark criteria met by datasets

Table 3 highlights the existing publicly available datasets and the benchmark criteria they fulfil. From this summary, it is clear that there currently exits no benchmark hate speech detection dataset.

5 Discussions and Implications for future research

First, we want to encourage researchers to put in more efforts towards a less biased, benchmark dataset taking the prior discussed factors into consideration.

Second, we also implore social media platforms to make the access easier for researchers.

Collaboration with these platforms is also another way to ensure better data sharing. Twitter has been known to release datasets for research purposes[18] (Vidgen et al., 2019).

We suggest that all datasets are anonymized before release because some of the username left in the dataset have ended up in research publications; which is a glaring ethical breach. Although some studies have extracted user information as a feature, we argue that it constitutes some ethical concerns and should be avoided. For a more in-depth survey on the issues surrounding social data bias see (Olteanu et al., 2019).

[18] https://www.wired.com/story/twitters-disinformation-data-dumps-helpful/

Also, we purport that the specific terms be used to avoid confusions and conflations of ideas. Even better, a clear definition should be provided on what the researcher defines a term as, e.g. what is offensive, abusive, or hate speech for the researcher. Unnecessary conflations dampens the research efforts. Moreover, a clear demarcation should be made for for proposed methods to solve hate speech, abusive language and cyberbullying detection. Their characteristics differ and proposed solutions might not generalize.

Finally, making codes public is always in best interest of the research community and when that is not possible, the hyperparameter choices and other necessary settings should be reported to support replicability of research work.

6 Conclusion

This work assisted in understanding the limitations of existing hate speech data for future research and the way forward. The contributions of this work include:

- Recommendation on a better approach to make datasets publicly available in the future.
- Requirements for any future researcher/organization interested in collecting and labelling data
  - Persistently publicly available
  - Consistent train-test split
  - Less bias
  - Lack of data degradation
  - Common evaluation metric
  - Basic pre-processing

These suggestions can be easily applied to other NLP applications apart from hate speech detection that require real-world datasets. We acknowledge the fact that an unbiased dataset does not exist, however, there are steps to be taken to make them less biased. Finally, even though we might have highlighted limitations in datasets and approaches, it is not meant as a negative criticism of the authors or their work. We acknowledge that their individual and collective efforts have brought us so far in this research area.

Acknowledgements: The authors are grateful for the insightful comments from the reviewers that helped improve this work.

References

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *Advances in Information Retrieval*, pages 141–153, Cham. Springer International Publishing.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 45–54, NY, USA. ACM.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596.

Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2020. On analyzing annotation consistency in online abusive behavior datasets. In *Proceedings of the 14th International AAAI Conference on Web and Social Media*.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*.

Darina Benikova, Michael Wojatzki, and Torsten Zesch. 2018. What does this imply? examining the impact of implicitness on the perception of hate speech. *Lecture Notes in Computer Science*, 10713 LNAI:171–179.

Pete Burnap and Matthew L. Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, 5(1).

Rich Caruana and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 161–168, NY, USA. Association for Computing Machinery.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. pages 2819–2829.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Thomas Davidson and Debasmita Bhattacharya. 2020. Examining racial bias in an online abuse corpus with structural topic modeling. In *Proceedings of the 14th International AAAI Conference on Web and Social Media*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Third Abusive Language Workshop, Annual Meeting for the Association for Computational Linguistics 2019*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Comput. Linguist.*, 30(1):95–101.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings. In *Proc. 24th Int. Conf. World Wide Web*, pages 29–30.

JL Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378—382.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):1–30.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

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*, WebSci '19, page 105–114, New York, NY, USA. Association for Computing Machinery.

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 *AAAI International Conference on Web and Social Media (ICWSM)*.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 774–782, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2019. Exploring hate speech detection in multimodal publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467.

Christos Karatsalos and Yannis Panagiotakis. 2020. Attention-based method for categorizing different types of online harassment language. *Communications in Computer and Information Science*, page 321–330.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *ArXiv Preprint*.

Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. In *ArXivPreprint*.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. Technical report.

Yann LeCun and Corinna Cortes. 2010. MNIST handwritten digit database.

Ruibo Liu, Guangxuan Xu, and Soroush Vosoughi. 2020. Enhanced offensive language detection through data augmentation. In *ICWSM'20 Safety Data Challenge*.

Kosisochukwu Judith Madukwe and Xiaoying Gao. 2019. The Thin Line Between Hate and Profanity. In *AI 2019: Advances in Artificial Intelligence*, pages 344–356, Cham. Springer International Publishing.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 467–472, Varna, Bulgaria.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental and Theoretical Artificial Intelligence*, 30(2):187–202.

Kristian Miok, Dong Nguyen-Doan, Blaž Škrlj, Daniela Zaharie, and Marko Robnik-Šikonja. 2019. Prediction uncertainty estimation for hate speech classification. *Lecture Notes in Computer Science*, page 286–298.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII*, pages 928–940, Cham. Springer International Publishing.

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, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Alexandra Olteanu, Kartik Talamadupula, and Kush R. Varshney. 2017. The limits of abstract evaluation metrics: The case of hate speech detection. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 405–406, NY, USA. ACM.

Abiola Osho, Ethan Tucker, and George Amariucai. 2020. Implicit crowdsourcing for identifying abusive behavior in online social networks. In *ArXiv PrePrint*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Bxl, Belgium. ACL.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4757–4766.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis.

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, FLR, Italy. ACL.

Sima Sharifirad, Borna Jafarpour, and Stan Matwin. 2018. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 107–114, Brussels, Belgium. Association for Computational Linguistics.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

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 LREC 2016 Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)*. European Language Resources Association (ELRA).

Gaurav Verma, Niyati Chhaya, and Vishwa Vinay. 2020. "to target or not to target": Identification and analysis of abusive text using ensemble of classifiers. In *ICWSM'20 Safety Data Challenge*.

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, FLR, Italy. ACL.

Fabio Del Vigna, Andrea Cimino, and Felice Dell Orletta. 2017. Hate me, hate me not: Hate speech detection on facebook. In *ITA-SEC 17*, Venice.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the 2012 Workshop on Language in Social Media*, pages 19–26.

Zeerak Waseem. 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*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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, San Diego, California. Association for Computational Linguistics.

Zeerak Waseem, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Golbeck J. (eds) Online Harassment. Human–Computer Interaction Series*, pages 29–55, Cham. Springer International Publishing.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. *IEEE Access*, 6:13825–13835.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. *Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pages 602–608.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 1980–1984, NY, USA. ACM.

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, page 925 – 945.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web: European Semantic Web Conference*, pages 745–760, Cham. Springer International Publishing.