This is a repository copy of *Examining temporal bias in abusive language detection*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/208053/

Version: Submitted Version

Preprint:
Jin, M., Mu, Y., Maynard, D. orcid.org/0000-0002-1773-7020 et al. (1 more author) (Submitted: 2023) Examining temporal bias in abusive language detection. [Preprint - arXiv] (Submitted)

https://doi.org/10.48550/arxiv.2309.14146

© 2023 The Author(s). This preprint is made available under a Creative Commons Attribution 4.0 International License. (https://creativecommons.org/licenses/by/4.0/)

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

arXiv:2309.14146v1 [cs.CL] 25 Sep 2023

Examining Temporal Bias in Abusive Language Detection

Mali Jin, Yida Mu, Diana Maynard, Kalina Bontcheva

Department of Computer Science, The University of Sheffield, UK
{m.jin, y.mu, d.maynard, k.bontcheva}@sheffield.ac.uk

Abstract

The use of abusive language online has become an increasingly pervasive problem that damages both individuals and society, with effects ranging from psychological harm right through to escalation to real-life violence and even death. Machine learning models have been developed to automatically detect abusive language, but these models can suffer from temporal bias, the phenomenon in which topics, language use or social norms change over time. This study aims to investigate the nature and impact of temporal bias in abusive language detection across various languages and explore mitigation methods. We evaluate the performance of models on abusive data sets from different time periods. Our results demonstrate that temporal bias is a significant challenge for abusive language detection, with models trained on historical data showing a significant drop in performance over time. We also present an extensive linguistic analysis of these abusive data sets from a diachronic perspective, aiming to explore the reasons for language evolution and performance decline. This study sheds light on the pervasive issue of temporal bias in abusive language detection across languages, offering crucial insights into language evolution and temporal bias mitigation.

Introduction

The increasing use of social media platforms has given rise to a pervasive problem of online abusive language, which can cause harm to individuals and lead to societal polarization. In recent years, researchers have developed a huge variety of machine learning models that can automatically detect abusive language (Mishra et al. 2019; Aurpa, Sadik, and Ahmed 2022; Das and Mukherjee 2023; Alrashidi, Jamal, and Alkhathlan 2023). However, these models may be subject to temporal bias, which can lead to a decrease in the accuracy of abusive language detection models, potentially allowing abusive language to be undetected or falsely detected.

Temporal bias arises from differences in populations and behaviors over time (Olteanu et al. 2019). In natural language processing (NLP), it can result from various issues. Temporal concept drift refers to the problem of language evolving over time (Zhao et al. 2022). Languages change as new meanings develop for existing words and new words and topics come into use over time. Models trained on data from an earlier period can perform worse on chronologically newer data as they are unable to recognize new topics or linguistic features (Lukes and Søgaard 2018; Vidgen et al. 2019; Mu et al. 2023). Previous work has examined temporal bias in various tasks such as named entity recognition (Derczynski, Bontcheva, and Roberts 2016), sentiment analysis (Lukes and Søgaard 2018) and rumour detection (Mu, Bontcheva, and Aletras 2023).

In online abuse detection, words and expressions considered acceptable in the past may have an abusive or offensive connotation now due to the changing language or societal norms (Wich et al. 2022; McGillivray et al. 2022). Also, temporal bias occurs when the abusive content fluctuates based on the latest trends, popular topics or breaking news. As the online discussion evolves with new development, certain topics and forms of abuse might gain prominence while others become less prevalent. For example, in 2020 a fraudulently altered video was circulated on Twitter purporting to show Al Jazeera journalist Ghada Oueiss naked in a jacuzzi, as part of an orchestrated attack designed to discredit her (Posetti et al. 2021). The video and other photos were distributed with messages alleging she was an alcoholic, drug-addicted prostitute, which engendered in turn a large number of hateful messages connected with the alleged jacuzzi incident, a topic not typically associated with abuse.

Previous work identified temporal bias in an Italian hate speech data set associated with immigrants (Florio et al. 2020). However, they have yet to explore temporal factors affecting predictive performance from a multilingual perspective. In this paper, we explore temporal bias in 5 different abusive data sets that span varying time periods, in 4 languages (English, Spanish, Italian, and Chinese). Specifically, we investigate the following core research questions:

- *RQ1:* How does the magnitude of temporal bias vary across different data sets such as language, time span and collection methods?
- *RQ2:* What type of language evolution causes the temporal bias in our data sets and how?
- *RQ3:* Could domain adaptation models, large language models (LLMs) or a more robust data set help to mitigate the temporal bias in abusive language detection?

To answer these questions, we compare the predictive

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

performance between random and chronological data splits across data sets in different languages and with different temporal coverage. We also experiment with different transformer-based pre-trained language models (PLMs) using the original data set and a filtered data set. Finally, we present an in-depth analysis to investigate the factors for performance degradation.

Related Work

Bias in NLP

Bias refers to the presence of systematic and unfair favouritism or prejudice. In various contexts, bias can manifest as a skewed representation or inaccurate judgments that unfairly advantage or disadvantage certain individuals or groups (Garrido-Muñoz et al. 2021). Bias can arise from various sources such as data selection, annotation processes, models and research design. These biases can potentially lead to unfair or discriminatory outcomes through NLP applications (Hovy and Prabhumoye 2021). For instance, biased language models might generate discriminatory content or fail to accurately understand and respond to underrepresented languages. Consequently, addressing and mitigating bias in NLP has become a critical research endeavour. Researchers are exploring techniques to measure and mitigate bias across diverse domains and languages (Sun et al. 2019; Font and Costa-Jussa 2019; Zueva, Kabirova, and Kalaidin 2020; Czarnowska, Vyas, and Shah 2021). Common debiasing methods include data reweighing and resampling, debiasing word embeddings, counterfactual data augmentation and bias fine-tuning (Kamiran and Calders 2012; Zhao et al. 2018; Park, Shin, and Fung 2018).

Bias in Abusive Language Detection

Previous work has focused on identifying and mitigating different forms of social bias in abusive language detection, such as gender bias (Park, Shin, and Fung 2018), dialect bias (e.g. African-Americans English) (Davidson, Bhattacharya, and Weber 2019; Sap et al. 2019; Davidson and Bhattacharya 2020; Zhou 2021) and different forms of identity bias (e.g. transgender, black) (Dixon et al. 2018; Zueva, Kabirova, and Kalaidin 2020). Moreover, Elsafoury et al. (2022) measured systematic offensive stereotyping bias (i.e., associating slurs or profane terms with specific groups of people, especially marginalized people) in different word embeddings.

However, little attention has been paid to temporal bias in abusive language detection. One exception is the work of Florio et al. (2020), who identified temporal bias in an Italian hate speech data set associated with immigrants. They investigated the impact of data size and time spans on temporal robustness by using two strategies, namely a sliding window model and an incremental model. Their results showed that adding training data temporally closer to the testing set greatly improved the performance but simply increasing the size of training data did not lead to performance improvement. Also, they found that offensive language in online contexts experienced rapid changes in topics over different time periods. Moreover, McGillivray et al. (2022) made use of time-dependent lexical features to detect abusive language effectively by training on smaller and older data. To facilitate this, they obtained a list of words for semantic change (i.e. acquired or lost an offensive meaning between 2019 and 2020). Their results showed that semantic change impacts abusive language detection and it is feasible to improve the detection by considering this change instead of depending on large labeled data sets. However, both work restricted themselves only to a single data set or a single language and did not explore other languages.

Temporal Bias in Classification Tasks

Temporal bias occurs in classification tasks due to the variation and evolution of data patterns over time. This temporal variation can pose difficulties for machine learning models as patterns learned from one time period may not be applicable in another. Temporal bias was assessed in various classification tasks such as rumour detection (Mu, Bontcheva, and Aletras 2023), stance detection (Mu et al. 2023) and multi-label classification tasks related to legislation and biomedicine (Chalkidis and Søgaard 2022). Mu et al. (2023) found that domain-adapted pre-trained language models are less sensitive to time and thus are beneficial to temporal gap mitigation; while Chalkidis and Søgaard (2022) proposed group-robust algorithms to reduce the temporal bias in multi-label classification. Moreover, Alkhalifa, Kochkina, and Zubiaga (2023) investigated the impact of word representations and machine learning model choice on temporal performance of various classification tasks such as stance detection and sentiment analysis.

Data

We study two widely used English abusive data sets (*WASEEM* and *FOUNTA*). We also study a Chinese data set (*JIANG*), a Spanish data set (*PEREIRA*), and an Italian data set (*SANGUINETTI*), in order to explore the impact of temporality on different languages. We choose these data sets because the creation time of each post is provided or accessible (via tweet IDs). Details of the data sets are shown in Table 1.

WASEEM (Waseem and Hovy 2016) is an English abusive data set focusing on sexism and racism. They collect the tweets by manually searching common terms related to religious, sexual, gender, and ethnic minorities, and by using the public Twitter search API. They combine these two methods to ensure that non-offensive tweets that contain clearly or potentially offensive words are also obtained. The annotations are created by manual experts and then reviewed by an additional gender study expert. We merge the original *sexism* and *racism* labels into a single *abusive* label, and rename the *neither* label as *non-abusive*.

FOUNTA (Founta et al. 2018) is an English data set collected from Twitter containing two types of online abuse expressions: abusive and hateful. They randomly collect and sample the data, using text analysis and machine learning techniques to create the boosted set of tweets which are likely to belong to the two abusive classes. The data is then

Dataset	Language	Source	Time	Size	Labels
Waseem and Hovy (2016)	English	Twitter	07-04-2013 - 06-01-2016 (33 months)	16,914	neither, sexism, racism
Founta et al. (2018)	English	Twitter	30-03-2017 - 08-04-2017 (10 days)	80,000	normal, spam, abusive, hateful
Jiang et al. (2022)	Chinese	Weibo	06-04-2012 - 26-06-2020 (8 years)	8,969	sexism, not sexism
Pereira-Kohatsu et al. (2019)	Spanish	Twitter	04-02-2017 - 22-12-2017 (10 months)	6,000	hate speech, not hate speech
Sanguinetti et al. (2018)	Italian	Twitter	26-02-2015 - 25-04-2017 (26 months)	6,928	hate speech, not hate speech

Table 1: Data sets details.

annotated by crowdsourced workers. Similar to Leonardelli et al. (2021), we map the four labels in the data set into a binary offensive or non-offensive label. We exclude tweets labeled as *spam*, and merge *abusive* and *hateful* labels into *abusive*. The *normal* label is renamed *non-abusive*.

JIANG (Jiang et al. 2022) is a Chinese sexism data set collected from Sina Weibo (a Chinese microblogging platform). They first collect gender-related weibos by searching keywords such as 'feminism' and 'gender discrimination'. Then they extract the comments that link to these weibos and filter out the comments to produce the final data set, which is annotated by three PhD students.

PEREIRA (Pereira-Kohatsu et al. 2019) is a Spanish hate speech data set annotated by experts. They randomly collect the data using the Twitter Rest API and filter it using seven dictionaries, where six of them represent different types of hate speech (e.g., race, gender) and the last one contains generic insults.

SANGUINETTI (Sanguinetti et al. 2018) is an Italian hate speech data set targeting immigrants, Roma and Muslims. They obtain the tweets by selecting a set of neutral keywords related to each target. The data is annotated by a team of both expert and crowdsourced annotators.

Data Filtering

Since three is no time information or tweet content in the FOUNTA and SANGUINETTI datasets, we re-obtain the tweets with their created time using Twitter Academic API based on the provided tweet IDs. Given the provided tweet IDs and related texts in the PEREIRA corpus, we use them directly without re-collecting the data to avoid data loss as Twitter ids are time ordered[1], For all data sets, we remove the duplicates and any tweets with no created time information.

Data Splits

We divide the data into training and testing sets using two strategies, namely random splits and chronological splits. The statistics of each data set are shown in Table 2. We can see that two of the data sets cover only a short period (FOUNTA contains many tweets but only covers 10 days, while PEREIRA covers 10 months but is fairly small in size) while all the other datasets span several years.

Random Splits We randomly split the data sets into training and testing sets and keep class distribution the same as the original data sets.

Dataset	Training	Validation	Testing	All
WASEEM	12,214	2,156	2,536	16,906
FOUNTA	27,368	5,683	4,830	37,881
JIANG	6,335	1,118	1,316	8,769
PEREIRA	4,335	765	900	6,000
SANGUINETTI	2,861	595	506	3,962

Table 2: Data sets statistics.

Chronological Splits We adopt a stratified chronological split strategy following the method in Mu, Bontcheva, and Aletras (2023). We first sort the abusive and non-abusive texts separately in chronological order. Then, we extract the first 70% of posts from abusive and non-abusive sets separately and combine them as the training set. Similarly, we combine the last 15% of posts from abusive and non-abusive sets as the testing set. The middle part of the two sets is merged into the validation set. In this way, the distribution of labels in each set is consistent with the original data.

Predictive Models

LR We use Logistic Regression with bag-of-words using L2 regularization as our baseline (LR).

BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2018) is a transformer-based (Vaswani et al. 2017) language model, which is pre-trained on large corpora, such as the English Wikipedia and the Google Books corpus. During pre-training, it uses a technique called masked language modeling (MLM) where it randomly masks some of the words in the input text, aiming to predict the masked word based on the context (Devlin et al. 2018). We fine-tune the BERT model on abusive language detection by adding an output layer with a softmax activation function.

RoBERTa is an extension of BERT trained on more data with different hyperparameters and has achieved better performance in multiple classification tasks (Liu et al. 2019). We fine-tune RoBERTa in a similar way to BERT.

RoBERTa-hate-speech This domain adaptation model[2] is trained on 11 English data sets for hate and toxicity based on the RoBERTa-base model (Vidgen et al. 2020).

OA We use the OpenAssistant (OA) 30B model developed by LAIONAI, which fine-tunes the LLaMA (Large Language Model Meta AI; Touvron et al. 2023) 30B model using the OA dataset. Since the original LLaMA model is

[1] https://developer.twitter.com/en/docs/twitter-ids

[2] https://rb.gy/k5x9t

Model	Splits	WASEEM				FOUNTA			
		Acc	P	R	F1	Acc	P	R	F1
LR	*Random Splits*	81.94	79.27	79.08	79.18	92.54	83.66	84.69	84.16
	Chronological Splits	74.88	76.93	62.69	63.15	93.26	85.56	85.28	85.42
	Performance Drop	7.06↓	2.33↓	16.39↓	16.03↓	0.72↑	1.90↑	0.59↑	**1.26**↑
RoBERTa	*Random Splits*	85.73	84.10	82.65	83.26	94.95	90.98	86.43	88.49
	Chronological Splits	76.77	80.54	65.20	66.33	94.81	91.16	85.45	87.99
	Performance Drop	8.96↓	3.56↓	17.45↓	16.93↓	0.14↓	0.18↑	0.98↓	0.50↓
RoBERTa-hate-speech	*Random Splits*	89.20	87.50	87.82	87.64	96.42	93.16	91.11	92.09
	Chronological Splits	81.58	85.99	72.21	74.71	96.07	92.03	90.79	91.39
	Performance Drop	7.62↓	1.51↓	15.61↓	**12.93**↓	0.35↓	1.13↓	0.32↓	0.70↓
OA	*Random Splits*	64.47	68.96	70.88	64.26	80.43	68.11	81.93	70.54
	Chronological Splits	72.36	72.53	75.89	71.48	80.75	68.24	81.83	70.77
	Performance Drop	7.89↑	3.57↑	5.01↑	7.22↑	0.32↑	0.13↑	0.10↑	0.23↑

Table 3: Model predictive performance on English data sets using random and chronological splits. The smallest F1 performance drop (or rise) across models is in bold.

Model	Splits	JIANG				PEREIRA				SANGUINETTI			
		Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
LR	*Random*	76.14	73.24	72.81	73.01	77.00	70.35	70.95	70.64	86.22	73.93	77.19	75.36
	Chronological	71.50	68.28	68.76	68.49	80.67	76.08	69.86	71.83	85.21	71.71	71.71	71.71
	Performance Drop	4.64↓	4.96↓	4.05↓	4.52↓	3.67↑	5.73↑	1.09↓	**1.19**↑	1.01↓	2.22↓	5.48↓	**3.65**↓
BERT	*Random*	80.68	78.95	76.65	77.52	80.67	75.30	72.31	73.44	88.07	78.09	72.69	74.85
	Chronological	78.66	76.28	77.80	76.81	82.78	83.15	69.72	72.67	84.87	70.22	63.08	65.13
	Performance Drop	2.02↓	2.67↓	1.15↑	**0.71**↓	2.11↑	7.85↑	2.59↓	0.77↓	3.20↓	7.87↓	9.61↓	9.72↓

Table 4: Model predictive performance on a Chinese, Spanish and Italian data set using random and chronological splits. The smallest performance drops (or rise) across models are in bold.

not fully open-source, we obtain the xor weights from HuggingFace[3] and apply 8-bit quantisation techniques via BitsAndBytes (Dettmers et al. 2021) to decrease the inference memory requirements. We use OA for zero-shot classification where we provide the model with a sequence of texts and a prompt that describes what we want our model to do.

Experimental Setup

Tweet Pre-Processing For all data sets, we replace username mentions and hyperlinks with placeholder tokens, <USER> and <URL> respectively. For the Chinese data set, we use Jieba[4], a Chinese text segmentation, to tokenize the texts.

Hyperparameters For all the English data sets, we use RoBERTa-base[5]; for data sets in other languages, we use bert-base-chinese[6], bert-base-spanish-wwm-cased[7] and bert-base-italian-cased[8] respectively, which are trained on big corpora of the corresponding language based on the BERT-base model. We fine-tune all models with learning rate l = 3e-6, $l \in \{$1e-4, 1e-5, 5e-6, 3e-6, 1e-6, 1e-7$\}$. The batch size is set to 32 and the maximum sequence length is set to 128. All experiments are performed on a NVIDIA Titan RTX GPU with 24GB memory. We follow the official guidelines[9] to run the 30B OA model on a local server with two NVIDIA A100 GPUs.

Training and Evaluation We split the data sets into training, validation and testing sets with a ratio of 70:15:15. During training, we choose the model with the smallest validation loss value over 12 epochs. We run all models five times with different random seeds for both random and chronological split strategies. We report predictive performance using the average Accuracy, Precision, Recall and macro-F1 scores. For OA, we only input the prompt (i.e. *identify if the following text is abusive or non-abusive*) and the same testing sets using two data split strategies.

Results

The predictive results are shown in Table 3 (English data sets)[10] and Table 4 (data sets in Chinese, Spanish and Italian). Values in the *Performance Drop* column are calculated by subtracting the results of chronological splits from that of random splits, where ↓ indicates a positive value and ↑ indicates a negative value. In other words, performance drop refers to the performance decreases using chronological splits compared to random splits with the same model.

[3] https://rb.gy/qfpc9

[4] https://github.com/fxsjy/jieba

[5] https://huggingface.co/roberta-base

[6] https://huggingface.co/bert-base-chinese

[7] https://rb.gy/br2ys

[8] https://huggingface.co/dbmdz/bert-base-italian-cased

[9] https://huggingface.co/OpenAssistant

[10] We only report results of English data sets using OA as those of other languages are not good.

Random vs. chronological splits In general, we observe performance degradation using chronological splits compared to random splits across all pretrained language models (PLMs). This is in line with previous work on other classification tasks such as document classification (Chalkidis and Søgaard 2022), stance detection (Mu et al. 2023) and rumour detection (Mu, Bontcheva, and Aletras 2023). Furthermore, the longer the time span, the greater the performance degradation. For the data sets with long time spans, we observe 16.93↓ F1 on WASEEM using RoBERTa and 9.72↓ F1 on SANGUINETTI using BERT); while for the data sets with short time spans we observe only 0.5↓ F1 on FOUNTA using RoBERTa and 0.77↓ F1 on PEREIRA using BERT.

However, although the performance of LR is not as good as that of PLMs, it has a smaller performance drop (or even performance rise) on data sets with small time spans (e.g., 1.26↑ F1 on FOUNTA compared with 0.50↓ F1 using RoBERTa).

Interestingly, we observe only a slight performance drop on the data set of JIANG (0.71↓ F1 using BERT) despite the eight-year time span. This may be due to the differences in the expression of abusive language online in Chinese and English (JIANG vs. WASEEM) or different collection methods between these two data sets. Another speculation is that JIANG only focuses on sexist abuse (sexism or not) which is one of the domains of abusive language. In this case, it covers fewer topics than other abusive data sets, which makes the performance less affected by temporalities (we will further investigate it in the following section).

Vanilla vs. domain adaptation models We compare the vanilla RoBERTa model with the domain adaptation model (RoBERTa-hate-speech) on two English data sets. We found that RoBERTa-hate-speech not only outperforms RoBERTa across two data sets using both random and chronological splits as expected but also has a smaller performance drop on WASEEM (12.93↓), where tweets span three years. This suggests that domain adaptation models can help mitigate temporal bias in abusive language detection, especially over long time spans. However, there are no domain-specific models for other languages, suggesting that further efforts are needed to develop such models.

Zero-Shot Classification Since OA is trained after the year of WASEEM (2016) and FOUNTA (2018), we hypothesize that the difference of predictive results between two data split strategies using OA will be negligible (e.g. smaller than 1). The performance drop of FOUNTA is as expected (0.23↑ F1); while the F1 performance on WASEEM using chronological splits is 7.22 higher than using random splits. We speculate that the large performance difference between these two splitting ways on WASEEM is due to the more explicit abusive content in the testing set using chronological splits as temporalities are less likely to be an influencing factor for OA. To investigate this, we calculate the swearing rates (the percentage of tweets containing at least one swear word among all tweets) of these two testing sets using an English swearword list from Wiktionary (words considered taboo and vulgar or offensive)[11]. The swearing rate of WASEEM using random and chronological splits is 5.60% and 8.40%; while that of FOUNTA is 4.64% and 5.51% respectively. The performance of OA is more likely to be influenced by the explicitness of abusive expressions instead of temporal factors based on the results of two English data sets. However, more abusive data sets are needed to make a more robust conclusion.

	Random Split			Chronological Split		
	Precision	Recall	F1	Precision	Recall	F1
	WASEEM					
Non-abusive	92.6	91.6	92.0	77.6 (15.0↓)	97.4 (5.8↑)	86.6 (5.4↑)
Abusive	82.6	84.0	83.0	88.0 (5.4↑)	40.2 (43.8↓)	55.6 (27.4↓)
Overall	87.5	87.8	87.6	86.0 (1.5↓)	72.2 (5.6↓)	74.7 (12.9↓)
	FOUNTA					
Non-abusive	97.6	98.2	98.0	96.8 (0.8↓)	98.0 (0.2↓)	97.0 (1.0↓)
Abusive	88.6	83.8	86.4	86.2 (2.4↓)	77.4 (6.4↓)	81.6 (4.8↓)
Overall	93.2	91.1	92.1	92.0 (1.2↓)	90.8 (0.3↓)	91.4 (0.7↓)
	JIANG					
Non-abusive	83.2	88.8	85.8	86.2 (3.0↑)	80.6 (8.2↓)	83.2 (2.6↓)
Abusive	74.6	64.2	69.0	66.2 (8.4↓)	74.8 (10.6↑)	70.2 (1.2↑)
Overall	79.0	76.7	77.5	76.3 (2.7↓)	77.8 (1.1↑)	76.8 (0.7↓)
	PEREIRA					
Non-abusive	85.0	89.6	87.4	82.8 (2.2↓)	97.0 (7.4↑)	89.2 (1.8↑)
Abusive	65.6	55.0	59.6	83.6 (18.0↑)	42.4 (12.6↓)	55.8 (3.8↓)
Overall	75.3	72.3	73.4	83.2 (7.9↑)	69.7 (2.6↓)	72.7 (0.7↓)
	SANGUINETTI					
Non-abusive	91.4	95.0	93.2	88.2 (3.2↓)	94.6 (0.4↓)	91.6 (1.6↓)
Abusive	64.8	50.4	56.8	52.2 (12.6↓)	31.4 (19.0↓)	39.0 (17.8↓)
Overall	78.1	72.7	74.9	70.2 (7.9↓)	63.1 (9.6↓)	65.1 (9.8↓)

Table 5: Model predictive performance of each class as well as the overall performance using random and chronological splits.

We further explore whether temporal bias has a greater influence on abusive texts or non-abusive texts. Table 5 shows the performance of each class as well as the overall performance on five data sets using their best-performing models (*RoBERTa-hate-speech* for English data sets and *BERT* for other language data sets). In general, the performance drop in abusive classes is larger than that in non-abusive classes. Also, the larger the time span of the data sets, the greater the difference in performance degradation between abusive and non-abusive classes (e.g. F1 1.8↑ vs. 27.4↓ for PEREIRA with ten-month time span and F1 1.6↓ vs. 17.8↓ for SANGUINETTI with two-year time span). However, Jiang et al. (2022) is an exception where F1 scores of abusive classes increase by 1.2. We also notice that the degradation of precision for non-abusive content is larger than that of recall using chronological splits (e.g. 3.2↓ precision and 0.4↓ recall in SANGUINETTI); while for abusive content, the performance drop in precision and recall is reversed (e.g. 5.4↑ precision and 43.8↓ in recall in WASEEM). This indicates that by using chronological splits, non-abusive texts are more likely to be detected; fewer abusive texts can be detected but the detected ones are more likely to be correct.

Analysis

Text Similarities

We hypothesize that the drop in performance is due to a larger difference between training and testing sets using chronological splits. To verify this, we use three methods to

[11] https://en.wiktionary.org/wiki/Category:English_swear_words

Data Set	Splits	Jarccard	DICE	OC
WASEEM	*Random*	.278	.435	.777
	Chronological	.216	.355	.733
	Similarity Drop	.062↓	.080↓	.044↓
FOUNTA	Random	.203	.337	.672
	Chronological	.199	.332	.668
	Similarity Drop	.004↓	.005↓	.004↓
JIANG	*Random*	.243	.391	.748
	Chronological	.211	.349	.717
	Similarity Drop	.032↓	.042↓	.031↓
PEREIRA	*Random*	.185	.312	.653
	Chronological	.167	.286	.602
	Similarity Drop	.018↓	.026↓	.051↓
SANGUINETTI	*Random*	.190	.320	.657
	Chronological	.173	.295	.636
	Similarity Drop	.017↓	.025↓	.021↓

Table 6: Text similarities between training and testing sets using Jarccard, DICE and OC.

Random Splits				**Chronological Splits**			
Training		**Testing**		**Training**		**Testing**	
Unigram	r	Unigram	r	Unigram	r	Unigram	r
mohammed	.030	countless	.056	sexist	.163	kat	.550
liar	.029	chipotle	.056	women	.116	#mkr	.459
#mkr2015	.028	rapes	.056	islam	.089	andre	.230
job	.028	fault	.054	#notsexist	.078	face	.165
day	.027	lower	.050	call	.071	annie	.161
truth	.026	distraction	.049	female	.070	#cuntandandre	.147
kat	.026	forget	.047	men	.065	#katandandre	.121
everything	.026	terrorist	.047	girls	.060	celine	.112
death	.025	consider	.047	religion	.051	karma	.109
fight	.023	appears	.047	prophet	.049	cunt	.099

Table 7: Unigram feature correlations with abusive tweets between training and testing sets from WASEEM using random splits (left) and chronological splits (right), sorted by Pearson correlation (r). All correlations are significant at $p < .001$, two-tailed t-test.

Training		**Testing**	
		Random Splits	
Unigram	r	Unigram	r
性别 (gender)	.053	压迫者 (oppressor)	.084
小 (small)	.044	多点 (more)	.083
叫 (shout)	.040	取决于 (depending on)	.074
搞 (do)	.039	并非 (not)	.073
样子 (looks)	.034	发出 (sending)	.072
先 (first)	.033	废话 (nonsense)	.072
之前 (before)	.033	承担责任 (take responsibility)	.071
		Chronological Splits	
人 (people)	.063	厌女 (misogyny)	.132
多 (more)	.051	厌 (hate)	.122
那么 (then)	.042	女上司 (female manager)	.098
人家 (people)	.042	下属 (subordinate)	.092
需要 (need)	.040	1	.088
或者 (or)	.040	介意 (mind)	.079
只是 (just)	.0.039	我 (I)	.079

Table 8: Unigram feature correlations with abusive tweets between training and testing sets from JIANG using random splits (left) and chronological splits (right), sorted by Pearson correlation (r). All correlations are significant at $p < .001$, two-tailed t-test.

calculate text similarities: (a) Jaccard similarity coefficient; (b) DICE coefficient (Dice 1945) and (c) overlap coefficient (OC).

Jaccard similarity coefficient is defined as the size of the intersection divided by the size of the union of two sets, A and B,

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

DICE cofficient is defined as twice the size of the intersection divided by the sum size of two sets, A and B,

$$DICE(A,B) = \frac{2 * |A \cap B|}{|A| + |B|} \tag{2}$$

Overlap coefficient is defined as the size of the intersection divided by the smaller size of the two sets, A and B,

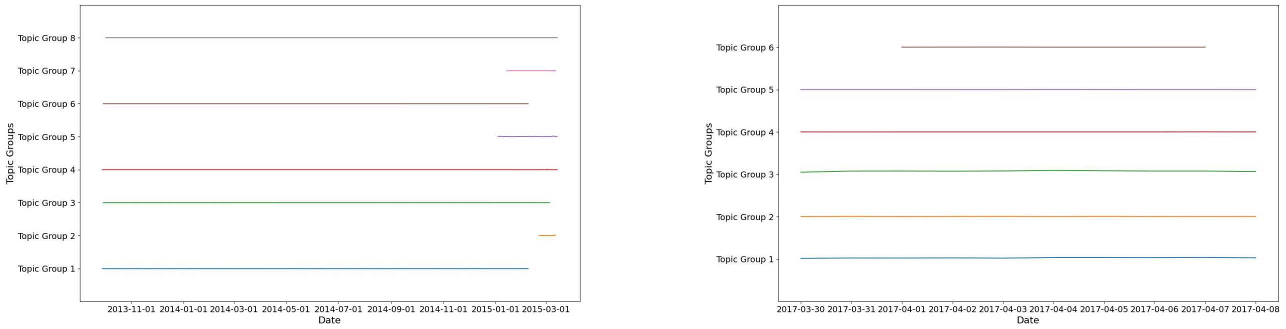
$$OC(A,B) = \frac{|A \cap B|}{\min(|A|, |B|)} \tag{3}$$

where A and B denote the set of distinctive words from training and test sets, respectively. $|A \cap B|$ and $|A \cup B|$ indicate the sum of distinctive words that appear in the intersection and union of the two subsets respectively. When the two subsets have no shared vocabulary, these three coefficient values will be zero, while if they are identical, the two values will be 1.

Table 6 shows the similarity coefficient between training and testing sets using *random* and *chronological* splits. Firstly, we notice that values from three similarity calculation methods drop across all data sets, indicating that using chronological splits leads to a larger difference between training and testing sets. Secondly, the longer the time span of data sets, the larger the similarity drop. For example, OC of WASEEM (three years) drops 0.044 while that of FOUNTA (one week) drops 0.004. Also, there tends to be a positive correlation between the magnitude of similarity reduction and the performance drop. However, considering the minor decline (drop 0.71 F1) in the predictive performance of JIANG (eight years), the text similarity drop is not consistent (e.g. OC drops 0.31). This can be explained by the fact that text similarity calculation is granular down to words; while topics might be limited (number, variety) in a sexist data set (i.e. JIANG).

Linguistic Analysis

We hypothesize that when using chronological splits, there are more new events or topics in the testing set that are not present in the training set, which leads to a decrease in model performance. In contrast, when using random splits, topics are evenly distributed between the training and testing sets. We present the linguistic patterns (unigram) of abusive tweets in training and testing data sets using two splitting strategies involving univariate Pearson correlations. Table 7 shows unigram feature correlations of WASEEM and Table 8 shows that of JIANG. We compare these two data sets be-



(a) Topic distribution of WASEEM.

(b) Topic distribution of FOUNTA.

Figure 1: Topic distribution over time.

cause their time spans are both long (three years vs. eight years) while their predictive performance drops vary widely (16.93 vs. 0.71).

For WASEEM, most abusive tweets in the testing set using chronological splits involve an Australian TV show, My Kitchen Rules (MKR) (e.g. *#mkr*, *#cuntandandre*, *#kateandandre*, *kat*, *andre*, *annie*), which is one of the queried terms for data collection. Our speculation is that the discussion about this show began to emerge during the later timeframe of the data set (within the time covered by the testing set when using chronological splits). However, there are hardly any new topics in the testing set when using random splits (e.g. *countless*, *lower*, *forget*).

For JIANG, testing sets using both split strategies mainly contain basic or gender-related terms (e.g. *more*, *not*, *misogyny*, *female manager*) and do not involve terms related to specific events. This is also correlated to how they collect the data: searching gender-related keywords such as 'feminism' and 'gender discrimination' for sexist content instead of using specific events as keywords. This suggests that collecting data using generic terms as keywords instead of terms associated with current hot events is likely to introduce less temporal bias.

Topic Distribution

We also explore topic distribution over time across two English data sets. We first use a topic modelling technique, BERTopic[12], to extract the 10 most important topic groups in a data set. Then we manually remove repeated or commonly used words (e.g. 'this', 'said') in these topic groups and combine similar groups into one group (e.g. combining 'women', 'men', 'she', and 'girls' into *gender-related* group). The generated topic groups of each data set are shown as follows[13]:

WASEEM: Group 1: {*sexist*, *women*, *men*, *bitch*, *her*, *she*, *girls*, *female*, *woman*, *notsexist*}; Group 2: {*kat*, *mkr*, *face*, *mkr2015*, *karma*}; Group 3: {*drive*, *drivers*, *driving*, *driver*}; Group 4: {*blondes*, *blonde*, *pretty*, *hot*, *dumb*}; Group 5: {*israel*, *hamas*, *palestinians*, *israelis*, *palestinian*, *palestine*, *gays*, *destroy*, *muslims*}; Group 6: {*sports*, *announcers*, *commentators*, *announcer*, *football*, *stand*, *commentator*}; Group 7: {*feminism*, *feminists*, *feminist*, *equality*, *movement*, *hypocrisy*, *rights*, *emma*, *modern*}; Group 8: {*funny*, *comedians*, *comedian*, *jokes*}.

FOUNTA: Group 1: {*trump*, *president*, *obama*, *voted*, *republicans*, *idiot*}; Group 2: {*nigga*, *niggas*}; Group 3: {*hate*, *bitch*, *bad*, *fucking*, *bitches*, *she*}; Group 4: {*syria*, *assad*, *syrian*, *chemical*, *trump*, *missiles*, *attack*, *obama*, *war*, *refugees*}; Group 5: {*pizza*, *eat*, *pineapple*, *disgusting*, *food*, *home*, *taco*}; Group 6: {*wrestlemania*, *wwe*, *match*, *rawaftermania*, *wrestlemania33*}.

Figure 1 shows the topic distributions over time of these two data sets. For WASEEM, Group 2 (MKR TV show related), 5 (race and religion related) and 7 (feminism related) appear only after 2015, which is also the starting time of the testing data set using chronological splits (March 2015). This results in the models barely seeing these words in the training set and a lack of knowledge in these three topics during training, especially for Group 2. Thus, it would be easier for models to fail when predicting text involving these topics using chronological splits. All topic groups are evenly distributed in FOUNTA except for Group 6 (wrestling match related). However, Topic Group 6 rarely appears in the testing set using chronological splits (starting from 7th April 2017), which is less likely to influence the performance.

Filtered Data Set

We explore whether removing words related to specific topics or events will enhance the robustness of the models when predicting abusive content. We hypothesize that the model performance will drop slightly while the difference between random and chronological splits will be more minor by removing these words. We experiment with WASEEM as its performance drop has room to reduce. We filter the data set by excluding three types of words: (1) words in all eight groups extracted by BERTopic (**D1**); (2) words se-

[12] https://github.com/MaartenGr/BERTopic

[13] We also try to extract topics of data sets with other language using BERTopic but the results are not good.

	Acc	P	R	F1
	Without removal			
Random	89.20	87.50	87.82	87.64
Chronological	81.58	85.99	72.21	74.71
Performance drop	7.62↓	1.51↓	15.61↓	12.93↓
	D1: Remove words by BERTopic			
Random	86.96	85.42	84.20	84.75
Chronological	80.45	83.86	70.93	73.21
Performance drop	6.51↓	1.56↓	13.27↓	11.54↓
	D2: Remove words by attention			
Random	87.16	85.75	84.30	84.95
Chronological	81.50	84.61	72.61	75.03
Performance drop	5.66↓	**1.14**↓	11.69↓	9.92↓
	D3: Remove words by both			
Random	84.73	83.37	80.70	81.76
Chronological	79.35	80.89	70.10	72.11
Performance drop	**5.38**↓	2.48↓	**10.60**↓	**9.65**↓

Table 9: Model predictive performance using RoBERTa-hate-speech on WASEEM with and without filtering. The smallest performance drops across filtering strategies are in bold.

lected by attention mechanisms (**D2**) and (3) the union of the words extracted by (1) and (2) (**D3**). For (2), we first use the RoBERTa-hate-speech model to produce attention scores that represent a probability distribution over each text. We then manually remove topic-related tokens among the top five tokens with the highest probability in each abusive tweet. Most of the removed tokens are names or hashtags related to the cooking TV show.

The results of filtered data sets are shown in Table 9. Similar to the previous experiment, we run five times for each method. First, all three strategies for removing topic-related words hurt performance in most cases, especially for chronological splits (e.g. 87.64 vs. 84.75 F1 using random splits, 74.71 vs. 72.11 F1 using chronological splits). However, the performance on D2 using chronological splits improves by 0.32 F1. Second, using more robust data sets leads to more minor performance drops. We achieve the smallest performance drop (9.65↓ F1) using D3. Also, using D2 achieves a comparable performance drop but only slightly hurts the performance. This suggests that filtering out specific topic-related words in a data set (i.e. a more robust data set) helps reduce temporal bias.

Error Analysis

Additionally, we perform an error analysis on two data sets containing sexist abuse, WASEEM and JIANG, using chronological splits. For WASEEM, we found that most errors happen when content involves the TV show (MKR). Also, when names from the show are mentioned, it is easy for models to misclassify the texts as non-abusive. We guess this is because the model cannot associate names in the testing set with male, female (gender-related) or abusive if it has not seen those names in the training set. However, the annotators of this data set have prior knowledge of this TV show and its characters. Thus, they are able to classify dissatisfaction or hatred toward specific characters as *sexist*. In the following two examples, tweets belonging to *abusive* are misclassified as *non-abusive* (names are highlighted in bold)[14]:

> T1: ***Kat*** *on #mkr is such a horrible person.. I wish* ***Kat*** *and* ***Andre*** *would just get eliminated already.*

> T2: *#MKR-I am seriously considering not watching just because I have to see* ***Kats*** *face. God. I want to slap it with a spatula!*

However, when gender-related words also appear in the content, models are more likely to classify them correctly. The following tweets are correctly classified as *abusive*:

> T3: *#katandandre gaaaaah I just want to slap* ***her*** *back to WA #MKR*

> T4: *#MKR* ***Girls****, thank you for filling the slapper quotient on this years series... we no longer have a need for* ***bitchy blondes****! Au Revoir!*

For JIANG, it is easy for models to fail to understand the actual meaning of a text without knowing traditional Chinese cultural viewpoints related to gender and marriage (e.g. some people value sons more than daughters). The following text belong to *abusive* (sexism originally) is wrongly classified as *non-abusive*:

> T5: 什么垃圾父亲，大女儿*16*岁就嫁人生子，没达到法定结婚年龄真的没问题？拿法律规定是用来干嘛的？又接着逼*15*岁二女儿去相亲赚彩礼？养猪啊？尽早出栏降低自己的成本是吗？ (*What a terrible father! Marrying off his eldest daughter and letting her have a child at the age of 16, without meeting the legal marriage age requirement? What's the point of having laws if they're not followed? And now he's pressuring his 15-year-old second daughter to go on blind dates to earn a dowry? Is he treating them like livestock? Trying to reduce his own costs by selling them off early?*)

Furthermore, objective discussions that contain words closely related to abuse are more likely to be misclassified as *abusive*. The following text is an example (potential abusive words are in bold):

> T6: 家暴不分男女！精神和身体上的暴力是同等的！ (***Domestic violence*** *knows no gender!* ***Mental and physical violence*** *are equally harmful!*)

To conclude, for WASEEM, models tend to misclassify tweets containing terms that implicitly link to gender or sexist where models have no prerequisite knowledge; while for JIANG, most errors happen when involving Chinese culture or terms that are more likely to appear in abusive content.

[14] Note that WASEEM is originally a sexist and racist data set, so other abusive content will be labeled as neither (*non-abusive* in our paper).

Limitations

This work aims to investigate the impact and causes of temporalities across different abusive data sets. In our work, we can only evaluate limited data sets that provide time information (e.g. 2 English ones, 2 data sets spanning more than 3 years) which limits control experiments for more sound comparisons. Also, all debiasing methods can only applied to English abusive data sets due to the imperfect implementation of techniques in other languages (i.e. domain adaptation models, BERTopic, OA). Moreover, our studies on temporal bias only explore topic changes and lack a comprehensive understanding of language evolution over time.

Conclusion

In this work, we investigate the impact of temporal bias on abusive language detection. We compare the predictive results using two data split methods (i.e. random and chronological splits) across different data sets (*RQ1*). The results indicate that temporal bias has a larger influence on data sets with a larger time span and collected using keywords, especially specific event-related keywords. Languages (or culture) may also be a factor but due to insufficient data sets, we can not draw concrete conclusions. We also conduct extensive analysis including text similarities, feature analysis and topic distribution to explore the causes of temporalities (*RQ2*). We found that performance degradation is mostly because of topic changes in our data sets. To provide a complete answer to *RQ3*, we filter a data set by removing topic-related words that appear in abusive texts. The predictive results suggest that using domain adaptation models and LLMs and training on a more robust data set can effectively reduce temporal bias in abusive language detection.

In the future, we plan to study temporal bias patterns in abusive data sets across different languages or platforms, aiming to understand the importance of considering the specific nature of the target variable when collecting the data sets and developing models. It can also be expanded to other text classification tasks.

Ethics Statement

This work has received ethical approval from our Research Ethics Committee. All datasets are acquired either through the URLs provided in the original papers or by requesting them from the respective authors. Note that we did not gather any fresh data from Twitter for this study. Additionally, we can verify that the data has been completely anonymized prior to its utilization in the Language Model Inference process.

Acknowledgements

This research is supported by a UKRI grant ES/T012714/1 ("Responsible AI for Inclusive, Democratic Societies: A cross-disciplinary approach to detecting and countering abusive language online").

References

Alkhalifa, R.; Kochkina, E.; and Zubiaga, A. 2023. Building for tomorrow: Assessing the temporal persistence of text classifiers. *Information Processing & Management*, 60(2): 103200.

Alrashidi, B.; Jamal, A.; and Alkhathlan, A. 2023. Abusive Content Detection in Arabic Tweets Using Multi-Task Learning and Transformer-Based Models. *Applied Sciences*, 13(10): 5825.

Aurpa, T. T.; Sadik, R.; and Ahmed, M. S. 2022. Abusive Bangla comments detection on Facebook using transformer-based deep learning models. *Social Network Analysis and Mining*, 12(1): 24.

Chalkidis, I.; and Søgaard, A. 2022. Improved Multi-label Classification under Temporal Concept Drift: Rethinking Group-Robust Algorithms in a Label-Wise Setting. *arXiv preprint arXiv:2203.07856*.

Czarnowska, P.; Vyas, Y.; and Shah, K. 2021. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9: 1249–1267.

Das, M.; and Mukherjee, A. 2023. Transfer Learning for Multilingual Abusive Meme Detection. In *Proceedings of the 15th ACM Web Science Conference 2023*, 245–250.

Davidson, T.; and Bhattacharya, D. 2020. Examining Racial Bias in an Online Abuse Corpus with Structural Topic Modeling. *arXiv preprint arXiv:2005.13041*.

Davidson, T.; Bhattacharya, D.; and Weber, I. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. *arXiv preprint arXiv:1905.12516*.

Derczynski, L.; Bontcheva, K.; and Roberts, I. 2016. Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1169–1179.

Dettmers, T.; Lewis, M.; Shleifer, S.; and Zettlemoyer, L. 2021. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Dice, L. R. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3): 297–302.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Elsafoury, F.; Wilson, S. R.; Katsigiannis, S.; and Ramzan, N. 2022. SOS: Systematic Offensive Stereotyping Bias in Word Embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1263–1274.

Florio, K.; Basile, V.; Polignano, M.; Basile, P.; and Patti, V. 2020. Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media. *Applied Sciences*, 10(12): 4180.

Font, J. E.; and Costa-Jussa, M. R. 2019. Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques. *arXiv preprint arXiv:1901.03116*.

Founta, A.; 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 *Proceedings of the international AAAI conference on web and social media*, volume 12.

Garrido-Muñoz, I.; Montejo-Ráez, A.; Martínez-Santiago, F.; and Ureña-López, L. A. 2021. A Survey on Bias in Deep NLP. *Applied Sciences*, 11(7): 3184.

Hovy, D.; and Prabhumoye, S. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8): e12432.

Jiang, A.; Yang, X.; Liu, Y.; and Zubiaga, A. 2022. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27: 100182.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1): 1–33.

Leonardelli, E.; Menini, S.; Aprosio, A. P.; Guerini, M.; and Tonelli, S. 2021. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. *arXiv preprint arXiv:2109.13563*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Lukes, J.; and Søgaard, A. 2018. Sentiment analysis under temporal shift. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 65–71.

McGillivray, B.; Alahapperuma, M.; Cook, J.; Di Bonaventura, C.; Meroño-Peñuela, A.; Tyson, G.; and Wilson, S. 2022. Leveraging time-dependent lexical features for offensive language detection. In *Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP)*, 39–54.

Mishra, P.; Del Tredici, M.; Yannakoudakis, H.; and Shutova, E. 2019. Abusive language Detection with Graph Convolutional Networks. *arXiv preprint arXiv:1904.04073*.

Mu, Y.; Bontcheva, K.; and Aletras, N. 2023. It's about Time: Rethinking Evaluation on Rumor Detection Benchmarks using Chronological Splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, 724–731.

Mu, Y.; Jin, M.; Bontcheva, K.; and Song, X. 2023. Examining Temporalities on Stance Detection Towards COVID-19 Vaccination. *arXiv preprint arXiv:2304.04806*.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kıcıman, E. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2: 13.

Park, J. H.; Shin, J.; and Fung, P. 2018. Reducing Gender Bias in Abusive Language Detection. *arXiv preprint arXiv:1808.07231*.

Pereira-Kohatsu, J. C.; Quijano-Sánchez, L.; Liberatore, F.; and Camacho-Collados, M. 2019. Detecting and Monitoring Hate Speech in Twitter. *Sensors*, 19(21): 4654.

Posetti, J.; Shabbir, N.; Maynard, D.; Bontcheva, K.; and Aboulez, N. 2021. The chilling: Global trends in online violence against women journalists. *New York: United Nations International Children's Emergency Fund (UNICEF)*.

Sanguinetti, M.; Poletto, F.; Bosco, C.; Patti, V.; and Stranisci, M. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. Florence, Italy: Association for Computational Linguistics.

Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.-W.; and Wang, W. Y. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. *arXiv preprint arXiv:1906.08976*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All you Need. *Advances in neural information processing systems*, 30.

Vidgen, B.; Harris, A.; Nguyen, D.; Tromble, R.; Hale, S.; and Margetts, H. 2019. Challenges and frontiers in abusive content detection. Association for Computational Linguistics.

Vidgen, B.; Thrush, T.; Waseem, Z.; and Kiela, D. 2020. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. *arXiv preprint arXiv:2012.15761*.

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.

Wich, M.; Eder, T.; Al Kuwatly, H.; and Groh, G. 2022. Bias and comparison framework for abusive language datasets. *AI and Ethics*, 1–23.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *arXiv preprint arXiv:1804.06876*.

Zhao, Z.; Chrysostomou, G.; Bontcheva, K.; and Aletras, N. 2022. On the Impact of Temporal Concept Drift on Model Explanations. *arXiv preprint arXiv:2210.09197*.

Zhou, X. 2021. *Challenges in Automated Debiasing for Toxic Language Detection*. University of Washington.

Zueva, N.; Kabirova, M.; and Kalaidin, P. 2020. Reducing Unintended Identity Bias in Russian Hate Speech Detection. *arXiv preprint arXiv:2010.11666*.