Humans Need Context, What about Machines? Investigating Conversational Context in Abusive Language Detection

Tom Bourgeade, Zongmin Li, Farah Benamara, Véronique Moriceau, Jian Su, Aixin Sun

Humans Need Context, What About Machines? Investigating Conversational Context in Abusive Language Detection

Tom Bourgeade[1,2], Zongmin Li[3,5,6], Farah Benamara[2,3,4], Véronique Moriceau[2], Jian Su[5], Aixin Sun[6]

[1] University of Turin, Italy, tom.bourgeade@unito.it
[2] IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France, {first.last}@irit.fr
[3] CNRS@CREATE LTD, Singapore
[4] IPAL, CNRS-NUS-A*STAR, Singapore
[5] Institute for Infocomm Research (I²R), A*STAR, Singapore, sujian@i2r.a-star.edu.sg
[6] School of Computer Science and Engineering, Nanyang Technological University, Singapore
zongmin001@e.ntu.edu.sg, axsun@ntu.edu.sg

Abstract

A crucial aspect in abusive language on social media platforms (toxicity, hate speech, harmful stereotypes, etc.) is its inherent contextual nature. In this paper, we focus on the role of conversational context in abusive language detection, one of the most "direct" forms of context in this domain, as given by the conversation threads (e.g., directly preceding message, original post). The incorporation of surrounding messages has proven vital for the accurate human annotation of harmful content. However, many prior works have either ignored this aspect, collecting and processing messages in isolation, or have obtained inconsistent results when attempting to embed such contextual information into traditional classification methods. The reasons behind these findings have not yet been properly addressed. To this end, we propose an analysis of the impact of conversational context in abusive language detection, through: (1) an analysis of prior works and the limitations of the most common concatenation-based approach, which we attempt to address with two alternative architectures; (2) an evaluation of these methods on existing datasets in English, and a new dataset of French tweets annotated for hate speech and stereotypes; and (3) a qualitative analysis showcasing the necessity for context-awareness in ALD, but also its difficulties.

Keywords: Abusive language detection, Conversational context, Context-aware classification

1. Introduction

Warning: *This paper contains examples of potentially offensive content.*

Abusive language have unfortunately become commonplace occurrences on various social media platforms. The sheer volume and often implicit nature of such unwanted content make manual moderation of these user spaces a formidable task. Consequently, numerous works have been proposed to create resources, datasets, and models aimed at automating the task of abusive language detection (henceforth ALD) (Talat and Hovy, 2016; Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Vidgen et al., 2019; Fortuna et al., 2020). As inspired by the framework proposed by Poletto et al. (2021), and for simplicity and conciseness, we use "Abusive Language" as an umbrella term to refer to the various forms of harmful language, such as toxic language, hate speech, and stereotypes (Vidgen et al., 2019; Madukwe et al., 2020).

One important aspect that warrants further exploration is the *contextual* nature of these various forms of abuse, including but not limited to: (a) *Conversational context*: What has been said before in a conversation thread (Karan and Šnajder, 2019; Menini et al., 2021; Vidgen et al., 2021; Pavlopoulos et al., 2020), (b) *Attitudinal or epistemic context*: The speaker/hearer's knowledge and the common ground, i.e., the situation in which the statement occurred (Mosca et al., 2021; Zhou et al., 2023), and (c) *Cultural context*: The cultural discrepancies and background of the people involved in communication as well as social bias (Arango Monnar et al., 2022; Lee et al., 2023; Davani et al., 2023). Incorporating such kinds of contextual information have shown to be essential in dealing with more subtle and implicit forms of online abuse (Wiegand et al., 2021; ElSherief et al., 2021; Ocampo et al., 2023; Vargas et al., 2021), detecting reported speech (Chiril et al., 2020) or discerning the intent of using offensive terms such as slurs, or group denominators (e.g., 'gay' or 'black') (Kennedy et al., 2020; Zhou et al., 2021). However, most works and resources in the field did not include any form of contextual information, during the annotation process, classification, or both which may put into question the high performance reported in them (Menini et al., 2021).

In this paper, we focus on the role of conversational context in ALD, which we may refer to as just "context" in the rest of this work. This is one of the most "direct" forms of context in this domain and we define it as *parts of the messages that precede*

LREC-COLING 2024, pages 8438–8452
20-25 May, 2024.

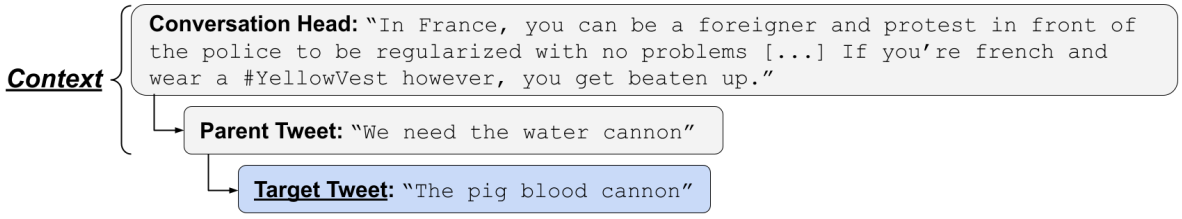


Figure 1: Conversational context in ALD, illustrated here on a tweet thread taken from our new French dataset and translated to English. The blue box is the target instance to be annotated/classified.

the one to be annotated/classified in a given conversation, such as the directly preceding message or the original post, also known as conversation head (see Figure 1). Conversational context, in the form of surrounding messages, has proven essential for the accurate manual annotation of harmful content (Jurgens et al., 2019; Vidgen et al., 2019; Prabhakaran et al., 2020; Bourgeade et al., 2023). For example, Vidgen et al. (2021) reported that $\sim$ 30% of their dataset was found by annotators to require this kind of contextual elements to be properly annotated. Similarly, Bourgeade et al. (2023), observed that annotators were not able to detect the presence of stereotypes without looking into the conversation thread in 74% of the annotated tweets.

If humans need context in inferring abusive content in online conversations, what about machines? When looking into recent context-aware approaches, most existing works have largely relied on *simple injection* methods where additional information is feed into Transformer-based architectures, by concatenating the target message with preceding messages in the conversation. However, this type of approach has produced inconsistent findings, across different studies and datasets (Zhang et al., 2022b; Pavlopoulos et al., 2020). For instance, while Menini et al. (2021) found that parent (directly preceding) tweets as context was not helpful, Markov and Daelemans (2022) reported enhanced performance when incorporating previous Facebook comments. The absence of a universal solution to the problem we are investigating, in our view, underscores the relevance of this research. These disparities raise questions about the underlying factors influencing the efficacy of different methods: *Is it a matter of the architecture of the models, the datasets they are fine-tuned on, or both?*

To answer these questions, we propose to investigate ALD using different strategies for incorporating conversational context information, while varying social media platforms, languages, and targets of abuse. As far as we know, this is the first study that provides such a meta-review of the impact of conversational context in ALD. Our contributions are as follows:

1. An analysis and overview of previous works on contextualized datasets and context-aware classification, followed by a proposal of two alternative architectures to the more widely used concatenation approach;

2. An evaluation of these architectures on existing abusive language datasets in English, but also a new dataset of French tweets annotated for both hate speech and harmful stereotypes[1];

3. A qualitative analysis of our results, highlighting some of the difficulties with the task at hand, and showcasing the necessity for context-awareness in ALD, as well as the limitations of the concatenation approach.

This paper is organized as follows. We first present a state of the art on ALD in conversations, focusing in particular on recent context-aware approaches. Section 3 describes the datasets used in our experiments. Section 4 gives the architectures we designed to account for contextual phenomena. Quantitative and qualitative results are presented in Section 5. We end the paper highlighting the main findings of this study and drawing some perspectives for future work.

2. Related Work

ALD has emerged as a significant and well-established research area in NLP, with a substantial body of literature. For comprehensive overviews of this field, we recommend surveys such as Fortuna and Nunes (2018); Vidgen and Derczynski (2020) and Yin and Zubiaga (2021). We focus in this section on the prevailing context-based methodologies. Subsequently, we delve into the principal conclusions drawn from these studies, underscoring the timely need for a more systematic examination of contextual phenomena.

Table 1 summarizes most existing context-aware datasets showing the source of the messages, the type of targeted abusive language phenomenon,

[1]The dataset can be made available upon request to the authors.

Authors	Source	Phenomena	A	B	C	D
▷ **(Vidgen et al., 2021)**	Reddit	Abuse	Yes	Yes	Yes	–
(Cercas Curry et al., 2021)	AI agents	Abuse	No	Yes	Unknown	No
(Zhang et al., 2020)	Twitter	Malevolence	No	Yes	Unknown	Yes
▷ **(Menini et al., 2021)**	Twitter	Abuse	Yes	Yes	Yes	No
▷ **(Yu et al., 2022)**	Reddit	HS, Counter Speech	Yes	Yes	Yes	Yes
(Zhang et al., 2022a)	PERSON-CHAT	Offensive language	Yes	Yes	Unknown	Yes
(Qian et al., 2019)	Reddit, Gab	HS	Yes	Yes	Unknown	–
(Pavlopoulos et al., 2020)	Wikipedia	Toxicity	Yes	Yes	Yes	No

Table 1: Comparison of context-aware abusive language datasets. **A**: Is it about conversation between human beings? **B**: Is context available during annotation? **C**: Is context helpful for human annotation? **D**: Does context improve the performance of classifiers, if any? ▷: Datasets used here (see Subsection 3.1).

and more importantly highlighting whether context helps during human vs. automatic annotations.

2.1. Do Humans Need Context?

Regarding annotation campaigns, some studies explicitly evaluate the importance of context during human annotations by comparing human labels with and without contextual information (marked "Yes" in the table). In other studies, annotators had either access to the message alone, or the message and its context together, which does not allow for measuring the specific impact of the latter (marked "Unknown" in column **C**). Overall, context has been shown to be crucial, regardless of the data source and the type of abusive phenomenon.

2.2. Context-aware ALD

A number of works from the Data Mining literature have investigated graph-based approaches to parse conversational threads for the early-detection of abusive content (Dahiya et al., 2021; Sahnan et al., 2021; Lin et al., 2021; Meng et al., 2023). Though these tasks differ in nature and end-goals with the one we are focusing on this work, the use of graph mining methods to model and extract information from the flow of conversation threads on social media may be of interest in future hybrid methods, for example. In the remainder of this section, we focus on context-based approaches as reported in the recent NLP literature.

Karan and Šnajder (2019) investigated whether including all comments from a conversation thread could enhance toxic content detection, as opposed to analyzing individual comments. Utilizing a pre-existing dataset of Wikipedia contributors' discussions, they discovered that context-aware models did not significantly surpass context-agnostic models. Nonetheless, they noted that the performance of both models was subpar for practical application.

Pavlopoulos et al. (2020) examined how the presence or absence of context—defined as the comment preceding the target comment—impacts both human and machine judgment of toxicity in Wikipedia conversations. They discovered that context could sway human perception of a comment's toxicity in either direction. Surprisingly, using various classification architectures (BiLSTM, CNN, BERT), no performance improvement was observed with context-aware models.

Menini et al. (2021) conducted a similar analysis for abusive language detection by re-annotating the English tweets dataset initially proposed by Founta et al. (2018), both with and without preceding messages in a conversation. This setup ensured that human annotators and machine learning algorithms had access to identical information for judgment. Their findings were in line with prior studies; they encountered difficulties in persuading abusive language detection models to consider this contextual information and saw no enhancements with context-aware models.

Markov and Daelemans (2022) delved into various forms of contextual information that could supplement target social media comments, such as directly preceding comments, original posts, and potentially more distant comments annotated by humans as contextually relevant. They precisely discovered that while directly preceding comments did not improve model performance, the latter two forms of context (original post and human-annotated context) contained useful information for hate speech detection.

Lastly, Anuchitanukul et al. (2022) presented an in-depth analysis of the influence of conversational context in toxicity detection. They focused on English datasets annotated for hate speech phenomena involving humans and humans vs. bots, including counter-speech generation. They proposed three context-aware architectures, based on different ways to construct a representation for the conversational context of a target message, by using BERT `[CLS]` token embeddings as representations of individual utterances (or the entire concatenated context) in the conversation thread, which are then aggregated to form the context rep-

resentation (through summation, or the use of a GRU model). Their results show that context-aware architectures outperformed the context-agnostic ones. However, and matching our own findings, across most datasets, the best performing model not only changes for each dataset, but context-aware models usually underperformed compared to even the agnostic one, showing that, so far, no universally reliable context-aware architecture has been found for ALD. Their results also confirm the relative weakness of simple concatenative approaches.

2.3. Do Machines Need Context?

When we compare datasets in terms of machine classification (see column **D** in Table 1), several hypotheses can be drawn to explain why context might not always be beneficial, and in fact, could be detrimental to some models. In particular, these include:

- Not all forms of context are necessarily relevant to understand a particular instance. For instance, the original post or the "head" of a conversation may provide more pertinent background context than the directly preceding (*parent*) message. The relevance of the parent message to its responses can vary depending on the specific usage patterns of a social media platform's users.
- The prevalent approaches of implementing context-aware NLP architectures through the elementary concatenation of the context with the target message's content often results in an inflated number of tokens for single models to attend to. Many of these tokens may act as "distractors", which can be detrimental to a model's overall ability to learn the subtleties of abusive language.
- The effectiveness of various context-aware approaches has predominantly been assessed on individual datasets, thus raising pertinent questions about their dependency on specific elements such as source of data, language, and specific abusive phenomena (i.e., target).

In this paper, we propose for the first time a systematic evaluation of easy-to-implement approaches to inject context for improving ALD. In particular, we propose two methods which attempt to address some of the issues described above. Compared to Anuchitanukul et al. (2022), we propose a BART-based approach together with a more *hierarchical* usage of Transformer architectures, by first separately encoding the context of a message as a whole, then using the resulting vector representation as though it was the embedding vector of a single token, which we concatenate to the embeddings of the message, and feed as input to a second Transformer. In addition, we newly investigate varying phenomena (i.e., hate speech but also stereotypes), as well as varying languages (i.e., English and French), aiming to expand upon prior contributions in the field.

3. Datasets

For this study, we selected three existing contextualized datasets from the relevant literature, and also introduce our own dataset. To allow us to more effectively compare different resources annotated with various phenomena related to abusive language, for all the datasets considered here, we focused solely on the binary abusive/non-abusive classification task, thereby collapsing additional classes (e.g., Counter-Hate Speech → negative class), and/or additional annotation layers (e.g., finer-grained categories), into the relevant binary classes. The characteristics of these datasets are summarized in Table 2 and Table 3.

3.1. Existing Datasets

From Table 1, we discarded all datasets that were not sourced from purely user-produced social media content, as well as counter-speech generation, machine-generated content and detection of target of hate speech (Schäfer and Burtenshaw, 2019; Cercas Curry et al., 2021; Zhang et al., 2022a). We also took into account annotation quality (i.e., good Inter-Annotator Agreement, see **IAA** row in Table 2), focusing on datasets where context has been proved to be useful for human annotations. Finally, datasets have to come with context information and substantial data samples.

(1) The `Contextual Abuse Dataset (CAD)` (Vidgen et al., 2021): It consists of Reddit conversations from different controversial (or related) Subreddits, annotated for six categories of abusive content grouped into three abusive and three non-abusive classes, as well secondary, finer-grained categories. Context includes all the previous comments and original posts, and the annotators specified whether this context was necessary or not, for each instance, and also provided rationales (highlighted passages) for their decisions. They, however, did not experiment with context-aware baselines and left it to future work.

(2) FOUNTA (Menini et al., 2021): This is a re-annotations of a subset (8K instances) of the large dataset introduced by Founta et al. (2018) where the presence of abusive or hateful content has been collapsed into a single abusive class. Annotators had access to the full Twitter thread as context for each instance. The results of the annotation campaign indicate that annotators frequently require context to comprehend the intent of tweet

	CAD	Reddit	Founta	FR_{hate}	FR_{stereo}
IAA	0.583	>=0.6	0.713	0.59	0.66
Size	23,947	5,223*	4,949	5,357	5,357
Splits	13,931; 4,624; 5,392	3,795; 883; 545	3,357; 373; 1,219	3,827; 473; 1,057	3,827; 473; 1,057
Pos %	20.25%	36.74%	8.45%	7.63%	11.83%

Table 2: Overview of the datasets used in this study. Splits sizes are respectively *train; validation; test*. $*$ Only Gold labelled instances, 6,845 including silver labelled.

authors, leading to instances initially labeled as abusive being re-annotated as the negative class, even in the presence of profanities. Models trained on the contextualized subset struggled compared to those trained without context, suggesting that the high performance of prior works may have been overly optimistic, as they did not account for realistic conditions. After crawling the data with Tweet IDs, we finally obtained 4,949 samples out of 8K which unfortunately makes proper comparisons between different works impossible (see Madukwe et al. (2020) for a discussion). Nevertheless, as it represents a very popular resource in abusive language detection, and one of the few contextualized datasets matching our criteria available, we keep it for our experiments.

(3) REDDIT (Yu et al., 2022): This is a dataset of Reddit comments annotated for hate speech, counter speech, or neutral content. Each instance comes with the directly preceding comments in conversation threads. The study demonstrated that neural networks achieve significantly better results when context is considered.

3.2. A new French Dataset

To explore the influence of context across languages and culture, we introduce a new contextualized French dataset composed of tweets, annotated for the presence of abusive stereotypes and hate speech, specifically aimed at immigrants. To ensure that our dataset carries contextual information, we have collected conversational threads of tweets using the Twitter API. First, conversation heads (the initial tweets in threads) have been collected from two different sources:

(1) A large list of hashtags used by groups propagating anti-migrant ideology: *#grandremplacement* [lit. *great replacement*], *#remigration*, *#retouraubled* [lit. *back to the bled*], *#STOPimmigration*, *#Cespaslislam* [lit. *it's not islam*], *#encoreunsuedois* [lit. *yet again a Swede*], *#unechancepourlafrance* [lit. *an opportunity for France*], *#encoreundesequilibre* [lit. *yet again a mentally unstable person*]; and

(2) A smaller list of highly public and heavily followed Twitter accounts (~ 20 accounts) tied to journalism or politics, and prone to using these hashtags, as well as sharing fake news.

A total of 42 conversation heads were obtained. All their replies, whether direct or indirect, have been collected, leading to a corpus of 5,357 tweets. Each tweet is associated with its conversation head text, as well as its direct parent (the tweet to which it replies) when available.

Following the annotation scheme proposed by Bourgeade et al. (2023), our dataset has been annotated for the presence of stereotypes at the tweet level ($\mathbf{FR_{stereo}}$) and if yes, annotators have to determine whether, in order to understand the meaning of the racial stereotype expressed, they need to look through the context. In addition to these two annotation layers, annotators are asked to determine the presence of hate speech ($\mathbf{FR_{hate}}$). The structure of an instance of our dataset is shown in Figure 1 where the conversation head compares illegal immigrants vs. yellow vests rights to protest, assuming that the latter are more legitimate. In this example, the target message has been annotated as follows: `Does the target message convey an anti-immigrant stereotype?: yes; Is the context necessary to understand the stereotype?: yes; Does the target tweet contain hate speech?: yes.`

The dataset was labelled separately by two annotators, and to verify cross-annotator and temporal consistency, inter-annotator agreement (IAA Cohen's kappa) was measured on two common validation sets (approx. ~ 200 instances each), once at the beginning of the annotation process, and once at the end (see IAA in Table 2).

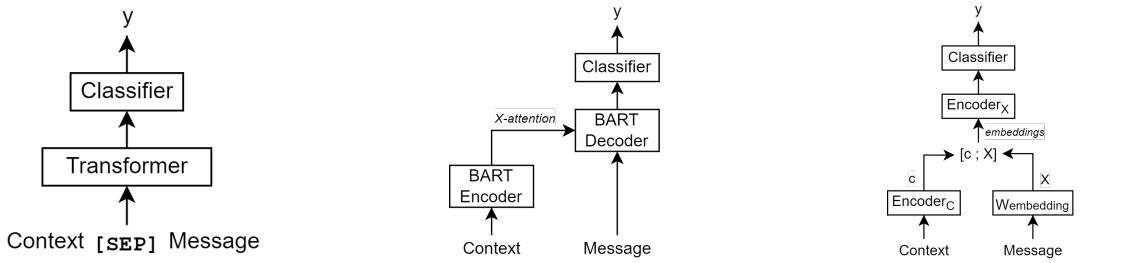
Table 3 presents the distribution of labels in our dataset. It shows that for 67.98% of the tweets, the annotators needed the context to decide whether a stereotype is present and understand its meaning. Besides, although 11.83% of the tweets contain a stereotype, only 7.63% contain hate speech, indicating that the presence of a stereotype does not necessarily imply hate speech.

4. Context-Aware Approaches

Social media platforms such as Reddit, Twitter, and Facebook employ diverse methods for structuring user interactions. However, we can generally view

FR_{stereo} Presence of a stereotype			Is Context needed to infer stereotypes?			FR_{hate} Presence of hate speech			Total
No	Yes	% Yes	No	Yes	% Yes	No	Yes	% Yes	
4,723	634	11.83%	203	431	67.98%	4,948	409	7.63%	5,357

Table 3: Distribution of **FR**, our new French dataset.



(a) `Concat` architecture (b) `BART` architecture (c) `Context-Embed` architecture

Figure 2: Overview of architectures for context-aware abusive language detection.

their content as conversation "trees", where each message is either a direct response to a *parent message* or an original post, which we refer to as a "*conversation head*". Similarly, contextualized datasets vary in the types of context they provide, including full conversations, direct parents, and conversation heads. In this study, we examine four different configurations to classify tweets: (1) without any context, (2) using the direct parent as context, (3) using the conversation head as context, and (4) concatenating both the conversation head and parent as context. We also investigate various strategies for incorporating contextual information in abusive language detection, as follows (see Figure 2 for an overview).

4.1. `Concat`: the Elementary Approach

Firstly, given that not all datasets were originally evaluated with context in mind, we evaluate the most straightforward and prevalent method of integrating context information using Transformer architectures (Vaswani et al., 2017). In this implementation, referred to as **`Concat`** (see Figure 2a), a Transformer-based classifier like BERT (Devlin et al., 2019) is employed to encode and classify both the target message and its context. This is achieved by providing them as a pair input via the separation token (*[SEP]*) mechanism, which is built into and learned by most of these architectures. For this context-aware architecture, we explore five different models from the literature, two domain-generic, and three pre-fine-tuned to domains related to abusive language detection in English: BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2020), fBERT (Sarkar et al., 2021), HateBERT (Caselli et al., 2021), and ToxDect-RoBERTa (Zhou et al., 2021). For our own French dataset, we investigate the CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020) models.

A potential limitation of this architecture concerning context-handling could be the input sequence length. Indeed, BERT-like Transformers have been demonstrated to underperform with longer contexts in some instances (Joshi et al., 2019). As a single Transformer is responsible for managing all the context tokens along with the main message's tokens (with the former often being more numerous), the attention mechanism might be adversely affected. Furthermore, certain context elements, particularly original posts or conversation heads on Reddit, are frequently too lengthy for these architectures to handle, especially when concatenated with a message, and thus often require truncation to fit into the models' input sequence length.[2]

4.2. Context without Distractions

We explore two additional architectures in this study to address these potential limitations.

`Bart`: The second approach involves using an Encoder-Decoder Transformer architecture, as originally defined by Vaswani et al. (2017), such as BART (Lewis et al., 2020). In this approach, the Encoder processes the context, and the Decoder processes the main message separately (see Figure 2b). The Encoder generates a representation of the context, which is then used in the Decoder's cross-attention layers to guide the prediction. In the original publication, pair classification tasks (such as NLI) are executed by feeding

[2]Note that we only truncate the context part for these instances, ensuring that the main messages to be classified remain fully within the input window of the models.

a similar concatenated input as the Concat architecture to both the Encoder and Decoder. However, to investigate the effects of processing the context and message separately, we instead draw inspiration from the use of BART for Machine Translation, where the two inputs are distinct. For the English datasets, we use the original BART model (Lewis et al., 2020), whereas for our own dataset, we employ the BARThez (Eddine et al., 2021) variant for French.

Context-Embed: In this method, we also process the context and message separately, but in a more "hierarchical" manner (see Figure 2c). The context is first passed through a "bare" Transformer model (Encoder_C), such as RoBERTa (Liu et al., 2019), to obtain a single vector representation (c) of the same dimensions as a word embedding vector. Subsequently, the message is first embedded (X) using a second Transformer's (Encoder_X) word embedding matrix ($\text{W}_\text{embedding}$). The context vector and message vectors are then concatenated ($[c; X]$) and directly fed as input word embeddings to the remaining layers of the second Transformer. The outputs of this Transformer are then fed into a standard classification head module. For this approach, for the English datasets, we use the original pre-trained RoBERTa model (Liu et al., 2019), and for French, the equivalent with CamemBERT (Martin et al., 2020).

4.3. Experimental Settings

For our experiments, we train the three aforementioned architectures in three contextualized setups: with the conversation head (+ head), with the parent message (+ parent), and with both of them (+ both). Additionally, we evaluate against the Context-Agnostic configuration, which employs the same Transformer classifier as in Concat, but without the context as input. For the Reddit dataset, only the parent message is available, making it the sole configuration we could evaluate. Due to the variable class imbalance inherent to ALD (see Table 2), we use a cross-entropy loss with class weights.

Due to hardware limitations and to ensure comparability of results, we present the outcomes only for the -base variants of all the models used. Although we experimented with some of the -large variants of these architectures, significant improvements were not consistently observed in most cases. Additional implementation and technical details for these experiments can be found in Appendix A.

5. Results and Discussion

5.1. Quantitative Results

Table 4 presents the results of our experiments, in terms of Macro F1 (**M**) and positive (abusive) class F1 (**P**) scores. For space reasons, we only present one best-performing (across all datasets at once[3]) base Transformer model per configuration: for the Context-Agnostic and Concat setups, ToxDect-RoBERTa performed the best overall on the English datasets while CamemBERT did for the French dataset.

Firstly, we can observe that for all three existing English datasets, the Concat approach delivers similar or lower performance compared to the Context-Agnostic models. This could be attributed to the input length issue discussed in Subsection 4.1. However, we were unable to find any statistical correlations between the total lengths, or the differences in lengths between the messages and contexts, and the predictions (binary or logits) of the models analyzed. For the models trained and evaluated on the French dataset, for both the *hate* and *stereotypes* tasks, we observe a slight improvement with the Concat + parent setup. This could be attributed to the high context-dependency of our dataset as identified by its annotators: for the entire dataset, 67.98% of instances were annotated as requiring context for annotation (see Table 3).

For comparison, in the CAD dataset (the only other dataset with this type of annotation), only 29.79% of instances were labelled as requiring context. For the re-annotated Founta dataset, we observed no improvements using any of the context-aware architectures we examined in this study. Interestingly, the Concat + head configuration yielded results almost identical to the best-performing Context-Agnostic equivalent model. This is overall in-line with the previous work by Menini et al. (2021) (though our results are not comparable, as we could not retrieve as many tweets as them; see Subsection 3.1). This could suggest a convergent local optimum for this specific dataset. Generally, and in relation to this, training on the three smaller datasets (Reddit, Founta, FR) presents challenges, due in parts to the diminished variety of data, in particular for the positive (abusive) class which is almost always significantly smaller than the negative class (see Table 2).

For the remaining datasets, we observe varying degrees of improvement using either or both of the more advanced BART and Context-Embed context-

[3]For example, for the Reddit dataset, for BERT, DistilBERT, HateBERT, and fBERT (in the Concat architecture), the macro F1 scores are as follows: 68.90, 68.00, 71.10, 68.90.

Model	Setup	CAD		Reddit		Founta		FR_{hate}		FR_{stereo}	
		M	**P**	**M**	**P**	**M**	**P**	**M**	**P**	**M**	**P**
No Context	-	71.10	50.30	69.10	52.00	**77.00**	**58.00**	54.78	13.17	63.37	33.49
Concat	+ head	68.40	45.60	-	-	**77.00**	**58.00**	49.74	2.63	60.78	28.40
	+ parent	70.00	48.20	67.40	50.60	76.00	56.20	55.69	15.03	65.59	37.84
	+ both	69.40	47.40	-	-	75.20	54.90	49.06	1.19	60.44	27.57
BART	+ head	66.09	46.15	-	-	70.80	46.73	77.46	60.17	69.93	**49.71**
	+ parent	67.44	48.83	**69.46**	**55.92**	68.90	43.44	*46.80*	*0.00*	47.14	4.08
	+ both	68.20	49.14	-	-	69.58	45.42	75.81	57.37	**69.95**	47.86
Context-Embed	+ head	72.54	55.31	-	-	65.57	37.21	75.16	55.79	64.50	41.60
	+ parent	73.02	**56.37**	59.57	48.85	64.17	33.13	**77.92**	**64.18**	64.64	43.10
	+ both	**73.04**	56.23	-	-	68.13	43.48	73.31	53.85	60.37	37.39

Table 4: Results of our experiments, in terms of Macro F1 (**M**) and positive (abusive) class F1 (**P**) scores. The scores in italics gray indicate models that failed to learn the task.

aware architectures. For instance, on the Reddit dataset and the stereotype detection task of our French dataset, the BART architecture delivers relatively important improvements compared to the simpler Concat and the Context-Agnostic models. Given that the role of context is to provide broader conversational and background knowledge to aid in understanding a given message, the ability to encode it separately with different weights may prove beneficial for numerous datasets.

While no single method emerged as universally superior in handling context, aside from the problematic Founta dataset (refer to Subsection 3.1), the methods we investigated here (BART and Context-Embed) consistently outperformed the more widely used Concat approach. Given these findings, we would advise going beyond simple approaches based on concatenation in future works, and aiming towards methods which can more adequately process contextual information in a more dedicated way, particularly when using similarly sized Transformers. Further research on alternative context-aware models will however be necessary, in order to potentially find a more universal solution, across genres, languages, and abusive phenomena.

Finally, the lower scores for the Concat approach on our French dataset appears to be due to overfitting on the negative (non-hateful/no stereotypes) class, which is highly unbalanced, a problem unfortunately common to many datasets in ALD. In addition, our dataset was found by its annotators to be highly contextual (for the stereotypes task, 68% of instances were judged as requiring the preceding context to be correctly interpreted), especially for positive class instances (see Subsection 5.2). As such, without a proper treatment of context within the classification models, the already small subset of positive class instances cannot be correctly classified as such. CamemBERT and FlauBERT also appear to be less prone to learning such pairs classification tasks, unlike BARThez, which could be explained by differences in these models pretraining, and pre-finetuning on some task(s) (which the first two did not go through).

5.2. Qualitative Analysis

To investigate how context-aware models handle the presence and nature of the contextual information provided to them, we perform a qualitative error analysis.

We first consider 4 configurations of the Context-Embed architecture for the CAD dataset (+ *head*, + *parent*, + *both*, and a "dummy" variant where the context input is replaced with "[No Context]"). First, we look at instances of the test set which were mispredicted by all 4 configurations: we find 549 out of 5392 such instances, and within those, a large number contain slurs (e.g., $\sim$ 50 occurrences of the f-slur and variations), and many other keywords often associated with controversial and/or inflammatory topics on social media platforms, especially on the Subreddits which compose the CAD corpus (e.g., "gender", "left"/"right", "liberal"/"conservative", and variations). Out of these 549, 207 were constantly misclassified as the positive class, against 342 misclassified as the negative class, which aligns with the training data distribution. These types of errors may indicate some easy-to-learn keywords-based shortcuts in this dataset, and this also extends in varying capacities to the other datasets.

When we examine instances that were correctly classified by a context-aware variant, such as + *parent*, but incorrectly classified by the no-context variant, we find that while the context indeed allows for the "proper" interpretation of the main message in some instances, a significant number of them remain challenging to understand. This difficulty

primarily arises due to the lack of broader background knowledge and, in the case of Subreddits for instance, the heavy use of "in-group" knowledge (e.g., inside jokes/references, meta-humor/irony, etc.). Furthermore, many abusive instances are highly implicit, as social media platforms have their own moderation mechanisms, which likely filtered out most of the more explicit instances. These challenges could potentially be mitigated by adapting models to the specific domain of the social media platform(s) in question, using methods such as Continued Pretraining (Gururangan et al., 2020), or by injecting a learned representation of the "background domain" context, to assist models in learning the specificities of user interactions on a given platform.

In our French dataset, we can remark practical differences between occurrences of hate speech and the propagation of harmful stereotypes: namely, tweets displaying hate speech tend to be "in direct response" to their parent tweet, whereas many occurrences of stereotypes tend to reference the conversation head, where a subject or situation was initially introduced for the conversation, which may explain the higher effectiveness of the + *parent* configurations for the former and + *head* for the latter (see Table 4).

6. Context in ALD: Main Findings

Overall, our study confirm the necessity of context for humans to properly understand the potentially harmful content, in a majority of online conversations. Furthermore, in many cases, both extensive and specific world knowledge is essential for capturing intricate elements of internet communication such as irony, allusions to current events, cultural references, and platform-specific phenomena like memes or community-specific humor. With this in mind, how could we ever hope for classification models to perform ALD tasks, without proper access and ability to handle the preceding messages in a conversation, let alone these more complex forms of contextual information? In line with the observations from Menini et al. (2021), we posit that proper ALD cannot feasibly take place without context-aware models, and that more research is therefore necessary to investigate the underlying reasons behind different approaches performing better or worse on different datasets.

We could not find a universally helpful context-aware approach, nor identify which specific contextual element(s) can automatically improve the detection of abusive language, which seems to match the results of prior related work. It may be that a universal solution to context injection might not exist: instead, it appears likely that each architecture and dataset/task may require specific elements of conversational context.

However, one core insight which we believe to be worth future exploration is the overall relative weakness of `Concat`-related approaches. Indeed, the "distractors" hypothesis may be a good candidate explanation, as we have shown that circumventing it, through the separate processing of the context elements, appears to improve the performance of context-aware models. Ultimately, however, it would be desirable for models to have access to the fullest and richest contextual information possible, as even at the instance level, each may require referring to different elements of the context to be properly understood and classified. Ideally, models should be made able to filter or select the precise parts which are necessary for each instance, and as such, different architectures and configurations should be investigated.

7. Conclusion and Future Work

In this study, we explored various approaches to develop context-aware classification models for detecting abusive language in conversations. Both previous studies and our qualitative analysis indicate that humans heavily rely on context to correctly decipher the intent and abusive characteristics of a message. Consequently, from our and previous findings, it is evident that algorithms can and should be designed to harness this information as well. These models were evaluated on several existing English datasets, as well as our novel French dataset, and our results challenge the efficacy of the commonly adopted `Concat` method, suggesting that it struggles to effectively leverage context tokens in a supporting role for the primary message. This shortcoming may contribute to the inconclusive results observed in prior research relying on similar techniques. Overall, we believe it would be important to explore prior works anew, with differing implementations of context-aware models, as the choice of the proper approach appears to be key to enabling context-aware classification on a given dataset.

Indeed, we successfully demonstrated here two alternative easy-to-implement approaches, and which could be further expanded upon. For instance, an architecture similar to Siamese BERT-Networks (Reimers and Gurevych, 2019, 2020; Thakur et al., 2021) could be used to create richer context representations. Dealing with the often substantial size discrepancy between a message and its context presents a fundamental challenge: a potential solution could be the use of a summarization model, such as one of the originally fine-tuned BART models from Lewis et al. (2020), to condense the number of tokens to attend to in the context.

Acknowledgements

This work has been supported by the STERHEOTYPES project funded by the Compagnia San Paolo 'Challenges for Europe'. It has also been supported by DesCartes: The National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program.

Limitations

This work involves experimenting with automated models for detecting abusive language. However, the performance and qualitatively analyzed outputs of these models, as well as those found in most other works, are currently too unreliable for practical applications. They should thus only be considered as a starting point for further research. The concept of "Abusive Language" is often poorly defined in theory and difficult to annotate in practice due to its subjectivity and sensitivity to spatio-temporal and cultural-linguistic context differences (van Aken et al., 2018; Malmasi and Zampieri, 2018; Vidgen et al., 2019; Madukwe et al., 2020; Poletto et al., 2021).

More specifically to this study, we only considered a very limited subset, (partially) common to all considered datasets, of context elements which could be attended to with the explored architectures (due in particular to input size constraints). Ideally, the entire conversation and its background, particularly on social media platforms, should be considered before attempting to classify the intent behind individual messages.

As is typical in abusive language detection, the positive (abusive) class is the minority class, resulting for these datasets in a very limited number of instances for fine-tuning (400-2000 instances). This results in models which are very sensitive to hyperparameters fine-tuning, as well as initial weights for the non-pretrained layers (e.g., the classification heads). In future work, we may explore different methods to attempt to alleviate these issues, such as cross-domain multitask learning, to augment the amount of data available at fine-tuning time.

Ethics Statement

In this work, we use social media users' posts to train and evaluate machine learning models. We also propose here a dataset of French tweets annotated for the presence of hate speech and harmful stereotypes towards immigrants. Following these platforms' data usage policies, we only use the textual content of messages and posts, discarding all metadata, including but not limited to the authors' user accounts, posting dates and times, etc. In addition, we perform a number of preprocessing steps on the textual content of messages before annotating, training, or evaluating (automatically and for manual analyses): we anonymize user mentions (`@USER`), replace URLs (`[URL]`), and, when possible, transform emojis into their textual code form (e.g., `:smile:`).

For all existing datasets used here, we only use the data publicly provided and linked to by the authors in their respective original publications, except for Menini et al. (2021) re-annotation of the original dataset proposed by Founta et al. (2018), for which only the tweets IDs and labels are provided, which we thus used to retrieve and reconstitute a dataset of text messages (with the IDs then dropped) given the currently available subset of tweets retrievable via the Twitter API (some of which having been deleted).

For our dataset's annotation process, four French native-speaking annotators were employed, of various academic levels (1 PhD, 2 Master's students, 1 undergraduate student), who were all compensated monetarily and/or through university credits applicable to their studies paths.

References

Atijit Anuchitanukul, Julia Ive, and Lucia Specia. 2022. Revisiting Contextual Toxicity Detection in Conversations. *Journal of Data and Information Quality*, 15(1):6:1–6:22.

Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. Resources for multilingual hate speech detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 686–696, Dubrovnik, Croatia. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. In *Proceedings of the 5th Workshop on*

Online Abuse and Harms (WOAH 2021), pages 17–25, Online. Association for Computational Linguistics.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Abuse Detection in Conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Snehil Dahiya, Shalini Sharma, Dhruv Sahnan, Vasu Goel, Emilie Chouzenoux, Víctor Elvira, Angshul Majumdar, Anil Bandhakavi, and Tanmoy Chakraborty. 2021. Would Your Tweet Invoke Hate on the Fly? Forecasting Hate Intensity of Reply Threads on Twitter. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 2732–2742, New York, NY, USA. Association for Computing Machinery.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Moussa Kamal Eddine, Antoine J.-P. Tixier, and Michalis Vazirgiannis. 2021. BARThez: A Skilled Pretrained French Sequence-to-Sequence Model. *arXiv:2010.12321 [cs]*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):85:1–85:30.

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 Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

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 *Twelfth International AAAI Conference on Web and Social Media*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for Coreference Resolution: Baselines and Analysis. 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 5803–5808, Hong Kong, China. Association for Computational Linguistics.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Mladen Karan and Jan Šnajder. 2019. Preemptive Toxic Language Detection in Wikipedia Comments Using Thread-Level Context. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 129–134, Florence, Italy. Association for Computational Linguistics.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing Hate Speech Classifiers with Post-hoc Explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, 18(185):1–52.

Ken-Yu Lin, Roy Ka-Wei Lee, Wei Gao, and Wen-Chih Peng. 2021. Early Prediction of Hate Speech Propagation. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 967–974.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.

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

Ilia Markov and Walter Daelemans. 2022. The Role of Context in Detecting the Target of Hate Speech. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 37–42, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Qing Meng, Tharun Suresh, Roy Ka-Wei Lee, and Tanmoy Chakraborty. 2023. Predicting hate intensity of twitter conversation threads. *Knowledge-Based Systems*, 275(C).

Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2021. Abuse is Contextual, What about NLP? The Role of Context in Abusive Language Annotation and Detection.

Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102, Online. Association for Computational Linguistics.

Nicolas Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An In-depth Analysis of Implicit and Subtle Hate Speech Messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity Detection: Does Context Really Matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Vinodkumar Prabhakaran, Zeerak Waseem, Seyi Akiwowo, and Bertie Vidgen. 2020. Online Abuse and Human Rights: WOAH Satellite Session at RightsCon 2020. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 1–6, Online. Association for Computational Linguistics.

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 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Dhruv Sahnan, Snehil Dahiya, Vasu Goel, Anil Bandhakavi, and Tanmoy Chakraborty. 2021. Better Prevent than React: Deep Stratified Learning to Predict Hate Intensity of Twitter Reply Chains. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 549–558.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fBERT: A Neural Transformer for Identifying Offensive Content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Johannes Schäfer and Ben Burtenshaw. 2019. Offence in Dialogues: A Corpus-Based Study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1085–1093, Varna, Bulgaria. INCOMA Ltd.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Zeerak Talat 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.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Francielle Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Pardo. 2021. Contextual-lexicon approach for abusive language detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1438–1447, Held Online. INCOMA Ltd.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

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, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: The Contextual Abuse Dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly Abusive Language – What does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2022a. Rethinking Offensive Text Detection as a Multi-Hop Reasoning Problem. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3888–3905, Dublin, Ireland. Association for Computational Linguistics.

Taolin Zhang, Chengyu Wang, Nan Hu, Minghui Qiu, Chengguang Tang, Xiaofeng He, and Jun Huang. 2022b. DKPLM: Decomposable Knowledge-Enhanced Pre-trained Language Model for Natural Language Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11703–11711.

Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. 2020. Detecting and Classifying Malevolent Dialogue Responses: Taxonomy, Data and Methodology.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in Automated Debiasing for Toxic Language Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. COBRA frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

A. Implementation and Technical Details

For all our experiments, we used the HuggingFace `transformers` (Wolf et al., 2020) and `datasets` libraries, and their provided model checkpoints. For all transformer models, we used the `-base` variants (e.g., `roberta-base`, etc.) due to hardware, time, and cost constraints. For the same reasons, and due to the number of different configurations to be trained and evaluated, we only proceeded with automatic hyperparameters fine-tuning, using the Weights & Biases (Biewald, 2020) AI platform's Bayesian hyperparameters optimization system, with the Hyperband early-stopping algorithm (Li et al., 2018), for a limited number of configurations: namely, for the `BART` and `Context-Embed` approaches, we automatically tuned only the + *head* configuration (as + *both* required more processing time due to the augmented size of the instances), and **FOUNTA** and **REDDIT** were tuned simultaneously, and thus use the same hyperparameters (due to their smaller combined size, compared to **CAD**).

The hyperparameters tuned were the learning rate (lr), the hardware training batch size (bs), and the number of gradient accumulation steps (*ga*), the latter two multiplied corresponding to the effective mini batch size used during training. Their final values after tuning can be found in Table A1. These models were trained for 6 epochs (determined during initial experimentation to be sufficient to reach stable performance scores), with the best performing epoch checkpoint kept at the end (measured by macro F1-score), with a warm-up ratio of 0.2 (linear warm-up from 0 to the initial learning rate over 20% of the training set, determined during initial experiments).

The other configurations (`Concat` and *No Context*) were hyperparameters-tuned manually, on a per-model basis. A mixture of hardware configurations were used, ranging from NVidia T4, V100, and A100, as required by each experiment.

Approach	Dataset	*lr*	*bs*	*ga*
`BART`	**CAD**	3.880E-05	8	12
	F&R	3.880E-05	8	12
	French	7.846E-05	4	24
`Cxt-Emb`	**CAD**	4.035E-05	4	24
	F&R	1.000E-05	4	16
	French	3.577E-05	4	24
`ToxDect`	***All***	0.00001	2	0

Table A1: Automatically fine-tuned hyperparameters (***lr***: learning rate; ***bs***: batch size; ***ga***: gradient accumulation steps; **F&R**: **FOUNTA** and **REDDIT**)