Learning like human annotators: Cyberbullying detection in lengthy social media sessions

Peiling Yi
Queen Mary University of London
London, UK
p.yi@qmul.ac.uk

Arkaitz Zubiaga
Queen Mary University of London
London, UK
a.zubiaga@qmul.ac.uk

ABSTRACT

The inherent characteristic of cyberbullying of being a recurrent attitude calls for the investigation of the problem by looking at social media sessions as a whole, beyond just isolated social media posts. However, the lengthy nature of social media sessions challenges the applicability and performance of session-based cyberbullying detection models. This is especially true when one aims to use state-of-the-art Transformer-based pre-trained language models, which only take inputs of a limited length. In this paper, we address this limitation of transformer models by proposing a conceptually intuitive framework called LS-CB, which enables cyberbullying detection from lengthy social media sessions. LS-CB relies on the intuition that we can effectively aggregate the predictions made by transformer models on smaller sliding windows extracted from lengthy social media sessions, leading to an overall improved performance. Our extensive experiments with six transformer models on two session-based datasets show that LS-CB consistently outperforms three types of competitive baselines including state-of-the-art cyberbullying detection models. In addition, we conduct a set of qualitative analyses to validate the hypotheses that cyberbullying incidents can be detected through aggregated analysis of smaller chunks derived from lengthy social media sessions (H1), and that cyberbullying incidents can occur at different points of the session (H2), hence positing that frequently used text truncation strategies are suboptimal compared to relying on holistic views of sessions. Our research in turn opens an avenue for fine-grained cyberbullying detection within sessions in future work.

CCS CONCEPTS

• **Artificial intelligence** → *Natural language processing*; • **Human-centered computing** → Social media.

KEYWORDS

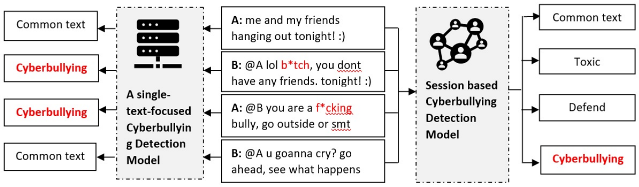
natural language processing, cyberbullying detection, long-text transformers, social media

ACM Reference Format:
Peiling Yi and Arkaitz Zubiaga. 2023. Learning like human annotators: Cyberbullying detection in lengthy social media sessions . In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3543507.3583873

1 INTRODUCTION

Cyberbullying is a form of bullying that is carried out through online devices [23]. Bullying is defined as the repeated, deliberate aggressive behaviour by a group or individual towards a more vulnerable person [24]. In the literature, there are two common traits of cyberbullying that are consistently referred to [45]: (1) repeatedly harming someone either physically or emotionally, and (2) a power imbalance between the parties. These are in turn used as the key criteria for identifying cyberbullying behaviour and to develop cyberbullying detection models [8, 9, 16, 21, 44].

Existing research in cyberbullying detection has predominantly focused on methods that analyse isolated social media posts. Research on models that analyse social media sessions –i.e. a sequence of posts and associated multimedia content– as a holistic view of how the abuse develops is however more limited. Figure 1 shows an example of a case of cyberbullying manifested as part of a social media session [35, 36], where modelling the session as a whole or as isolated posts can indeed make a difference. Indeed, the victim's and the bully's messages may contain "bad words" alike; where the victim's message (e.g. "You are a f*cking bully, go outside or smt") can be a defensive one, a single-post cyberbullying detection model may flag it as a case of cyberbullying due to its offensive words. In another example, such as "u gonna cry? go ahead, see what happens tomorrow", there are no offensive words, however it could constitute a case of cyberbullying if the surrounding social media session indicates so. Examples like these, as well as the inherent characteristic of cyberbullying being a recurring attitude of one's abusive behaviour over another, highlights the importance of modelling social media sessions as a whole for effective cyberbullying detection. This in turn makes cyberbullyihg detection different from other abusive language detection tasks, such as hate speech detection [17], where detection from isolated posts is more achievable.


Common text
Cyberbullying
Cyberbullying
Common text
A single-text-focused Cyberbullying Detection Model
Session based Cyberbullying Detection Model
Common text
Toxic
Defend
Cyberbullying

A: me and my friends hanging out tonight! :)

B: @A lol b*tch, you dont have any friends. tonight! :)

A: @B you are a f*cking bully, go outside or smt

B: @A u goanna cry? go ahead, see what happens

Figure 1: Single text based cyberbullying detection model vs Social media session based cyberbullying detection model.

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04...$15.00
https://doi.org/10.1145/3543507.3583873

As a silver bullet for text classification, Transformer-based pre-trained language models (PLM) have recently received considerable attention, thanks to their ability to capture contextual information by getting rid of the reliance on word-for-word understanding. PLMs are also commonly used for state-of-the-art cyberbullying detection research [26, 34, 41]. However, PLMs are also limited in the length of the text inputs they can handle (usually 512 tokens), which poses a challenge for modelling lengthy social media sessions for cyberbullying detection. To the best of our knowledge, effective modelling of lengthy social media sessions for cyberbullying detection has yet to be studied, which addresses the following limitations of previous work: *(1) Information loss:* The loss of information after reducing from sessions to single posts may be negligible for tasks that only require a general understanding of the text, but can have a bigger impact on tasks requiring contextualised recognition, such as cyberbullying detection; *(2) Attention shift:* To enable feeding long texts as input, additional algorithms are generally incorporated into the pipeline [25, 46], potentially causing a shift from the original attention after processing embeddings through the pipeline; and *(3) Lack of task specific models:* To date, there is no research tackling the lengthy nature of social media sessions in cyberbullying detection. Existing solutions for handling long inputs are in turn not tailored for cyberbullying detection.

Proposed Approach: we propose a conceptually intuitive long text social media session cyberbullying detection framework, LS-CB, which can effectively and efficiently handle long social media sessions of unrestricted length. The framework is inspired by a human annotator workflow (see the upper part of Figure 2). When dealing with cognitive tasks, humans retain a small amount of key information referred to as “work memory” [1], which emphasises the importance of carefully crafting the description and key terms when explaining an annotation task. When judging whether a social media session constitutes a case of cyberbullying, a human annotator would skim through the session to locate the blocks where the abuse happens and then to further investigate if the context supports this judgement. Similarly, the lower part of Figure 2 illustrates the core component of the LS-CB framework where the ‘machine annotator’ scans blocks of text within a social media session in search of cyberbullying blocks to ultimately leave it to the ‘judger’ to determine if the whole session should be deemed a case of cyberbullying.

To evaluate the validity and robustness of the LS-CB framework, we test it with six different Transformer models: BERT, RoBERTa, MPNET, Electra, Xlent and Distilbert. We compare them with three types of competitive, Transformer-based baselines: (1) state-of-the-art session based cyberbullying detection models, (2) Transformer-based pre-trained language models, and (3) LongFormer, a long-sequence transformer model.

Contributions: By proposing LS-CB, we introduce the first-ever approach to tackle lengthy social media sessions for cyberbullying detection. LS-CB provides a new framework that offers the flexibility to be used with different Transformer models, of which we test six. Using two session-based cyberbullying datasets, we demonstrate substantial improvements in performance over three types of competitive baselines. In addition, we perform a set of qualitative analyses looking at validating two hypotheses that we set forth, looking at the potential of analysing social media sessions as the aggregate of smaller blocks for cyberbullying detection (**H1**) and at the positions where the cyberbullying occurs within a session, assessing the need for holistic analyses of sessions for cyberbullying detection (**H2**).

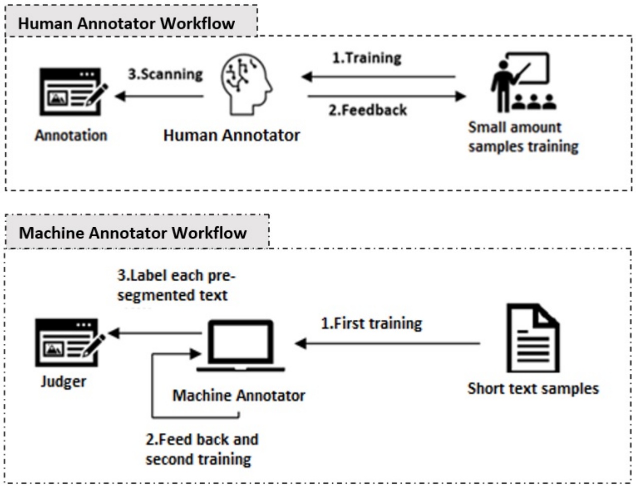


Figure 2: The general workflows of human annotation and the LS-CB framework.

2 RELATED WORK

Despite the substantial body of research in cyberbullying detection [29, 30], research into session-based detection has been more scarce [45], which we discuss next. We then follow with a discussion of work tackling long texts with transformers.

2.1 Session-based cyberbullying detection

Existing models for session-based cyberbullying detection generally construct a representation of the interaction between the bully and the victim to detect cyberbullying incidents. There are three main approaches in the literature to model the hierarchy and temporality of social media sessions.

Hierarchical networks with attention: This approach leverages the hierarchical network to reflect the structure of a social media session and utilises the hierarchical attention mechanism to automatically capture word-level and sentence-level hidden embeddings. The bidirectional GRU is employed to capture the sequence of contents [3, 4]. In addition, the temporal ordering of comments is considered by Cheng et al. [5] in a hierarchical network.

Multimodal learning: This approach considers representing the joint representations of different multimodal data. A straightforward method for encoding multi-modal context is to simply concatenate the raw feature vectors of each modality (e.g., locations, comments, images, timestamps). However, this method overlooks both structural dependencies among different social media sessions and cross-modal correlations among different modalities. MMCD [40] integrates them into a hierarchical attention network to capture hierarchical relationships. XBully [6] is another presentive model, which reformulates multimodal social media data as a heterogeneous network and then aims to learn node embeddings.

User interaction modelling: The third approach relies on the assumption that cyberbullying events often take place in the form of a series of interactions. Therefore, approaches incorporating sequences of user interactions have also been studied. For example, Ventirozos et al. [38] use a rule-based classifier to tag a conversation session as a sequence of sentiment words to reflect user interactions. Yao et al. [43] proposed an unsupervised cyberbullying detection method, which incorporates comment inter-arrival times for social media sessions, allowing the use of the full comment history to classify instances of cyberbullying. A graph neural network is used by Ge et al. [13] for modelling topic coherence and temporal user interactions to capture the repetitive characteristics of bullying behaviour.

Current state-of-the-art models can automatically acquire features and learn natural variations from complex data structures. However, the problem of enabling the feeding of entire, lengthy social media sessions into models has not been studied, particularly so in combination with state-of-the-art Transformer models which struggle with long inputs [25]. In recent research, it is a common practice to truncate the inputs (see next subsection), thereby making a trade-off between computational complexity and loss of information, with the potential risk of removing parts where the cyberbullying events occur. Our work in fact goes further into exploring the positions within a social media session where the cyberbullying incidents occur, showing that they can happen at any point, hence stressing the need to handle full sessions effectively.

2.2 Long text classification through Transformers

Since Transformer models use self-attention in a stack of encoders and decoders, time and memory complexity increase quadratically with sequence length [37]. Consequently, most Transformers set the input length limit of 512. This has attracted research into solutions to circumvent this limitation.

Text Truncation: Truncating texts to the required length is a simple approach that sacrifices parts of the content to make it manageable by the model; while this approach can be reasonable where we know the location of the core part of the text (e.g. first paragraph in news articles), it is arguably not the case in social media sessions. Among the studies focused on this direction, Sun et al. [33] is the most influential study, who fine-tune BERT for sentiment analysis, question classification and topic classification using three different methods for truncating text: (1) taking the first 510 tokens from the text, (2) taking the last 510 tokens, and (3) taking the first 128 tokens and the last 382 tokens. The second method performed the best in the original authors' experiments. However, from our empirical experiments, the first method outperforms the other two methods in the session-based cyberbullying detection task, so we adopt into our baselines. In some cases, such as Ganhotra and Joshi [12], results suggest that truncating has marginal impact on model performance, however we may be removing crucial information in the case of social media sessions.

Selecting relevant sentences: Rather than arbitrarily truncating texts, these approaches perform more careful sentence selection to shorten the input text. Different methods have been proposed to select sentences, such as Min et al. [22] using an encoder-decoder architecture, Wang et al. [39] using classical feature selection methods, and Ding et al. [11] using unsupervised learning. The change of loss during learning is used to judge if the sentence is a task-relevant sentence. Performance of this approach highly relies on the accuracy of the task-specific feature selection algorithm, which our proposed method avoids.

Hierarchical transformers: This approach preserves all the input text. The long input is segmented into small chunks and fed into transformers to generate the representation for each part, ultimately combining the representations. The combined representation is input to a single recurrent layer, or another transformer [25]. For this reason, this method is not only computationally expensive, but also leads to a shift in original attention due to the use of intermediate algorithms to merge segment embeddings belonging to the same session. Thereby running the risk of degrading model performance.

3 PROBLEM DEFINITION

In this section, we formally present the problem of cyberbullying detection in lengthy social media sessions. While tackling this problem, we define the following two hypotheses which we aim to assess through this research:

Hypothesis 1: Cyberbullying events can be detected through aggregated analyses of smaller blocks of text that form the lengthy social media sessions.

Hypothesis 2: Cyberbullying incidents can occur at different points of a session, hence requiring holistic analyses of sessions.

Based on the above two hypotheses and in line with previous work, we define key items that will be used in the problem definition:

Definition 1: Cyberbullying detection We define cyberbullying detection as a binary classification task. A binary cyberbullying classification task consists in determining if each social media session of unrestricted length in $S \in \{S_1, \ldots, S_n\}$ contains a cyberbullying incident, i.e. $Y_i \in \{0, 1\}$, where $Y_i = 1$ means at some point within the session there is an incident of cyberbullying, and $Y_i = 0$ means that no cyberbullying of any kind occurs.

Definition 2: Social media session: A social media session S_i is a sequence of posts $C_i^1, \ldots, C_i^m$, where two or more users interact with one another. In particular, we denote $S_i \in \{C_i^1, \ldots, C_i^m\}$, where C_i^m is the m^{th} post in S_i.

Definition 3: Sliding window with overlap We use a fixed window size n with a step size of m (where $n > m$ and, in our experiments, $m = 0.1 * n$) to run through each S_i with x fixed-length chunks $\{K_i^1, \ldots, K_i^x\}$. There is therefore an overlap of $n - m$ between adjacent windows. A text chunk K_i^x is not equal to C_i^x. During processing, each of these chunks is associated with an index.

Definition 4: Annotator An Annotator is a binary cyberbullying classifier A_k that can indicate if there is an incident of cyberbullying in a text K_i^x, i.e. each text chunk is associated a binary label $y_i \in \{0, 1\}$ (1 = cyberbullying, 0 = no cyberbullying).

Definition 5: Judger A judger J_k is a rule-based binary classifier that judges whether cyberbullying happened somewhere in the entire social media session according to the aggregate of the annotator outputs.

Problem definition: Our core objective is to demonstrate the potential of aggregating the analysis of smaller text blocks within social media sessions to inform the ultimate, aggregate judgement on whether a cyberbullying incident is found in the session. As an input to the problem, we have a given social media session S_i, a set of underlying comments $C_i \in \{C_i^1, \ldots, C_i^m\}$, which are grouped into sliding windows $K_i \in \{K_i^1, \ldots, K_i^x\}$. Our learning goal is to train a transformer-based annotator A_k to maximise prediction accuracy. We set the rules for the judger J_k as $Y(S_i) = J_k(A_k(K_i^1, ..., K_i^x)$, such that the outcome of the analysis of each text block can be aggregated.

4 METHODOLOGY

4.1 Theoretical Analysis

While the proposed approach in this study is intuitively inspired by human annotators, the theoretical support behind it is based on the nature of neural networks. Hence, there are a number of aspects to be considered when effectively training a neural network that implements this approach, such as effective updating of parameters and avoiding the impact of catastrophic forgetting, which are important to consider when we need to deal with sessions of different size. We discuss three core aspects related to our research objective, which inform the design of our LS-CB framework.

1) Parameter sharing: The transformer is a sequence transduction model based on an encoder-decoder configuration [37]. One of the essential characteristics of its structure is that any shift in the input data will be reflected into the output feature map [14]. Therefore, by passing the parameters of the first annotator model that has learned short social media sessions, the acquired knowledge can be transferred to the second annotator learner who learns long social media sessions. Additionally, the risk of overfitting in quadratic training can be reduced and convergence accelerated [28].

2) Batch normalisation: Batch Normalisation (BN) is a surprisingly simple yet effective method for deep domain adaptation tasks [18]. In this study, we aim to normalise the distribution differences between sessions of different size by modulating the statistics of all BN layers throughout the annotator classifier layers. Thereby improving the generalisability of the model by reducing the generalisation error.

3) Memory consolidation: Intuitively, if a model can access old training data when retraining, catastrophic forgetting would be reduced. Memory replay studies in continual learning solve catastrophic forgetting by selecting a small portion of old data or by generating synthetic samples [20]. Inspired by this, in order to consolidate the pre-trained memory, we select a portion of short-session samples combined with all long-session samples for continuous training.

4.2 Proposed Framework: LS-CB

In this section, we introduce our end-to-end novel framework LS-CB. LS-CB aims to leverage the power of transformer fine-tuning while circumventing the impact of the transformer's limited input length. This enables the cyberbullying detection model to learn from the holistic view of a lengthy input text to help improve performance and reduce computational complexity. As shown in Figure 3, LS-CB has four main components: Annotator Training Samples Selector, Annotator, Annotator training corrector and Judger. All test samples are split into sliding windows and input to the trained Annotator. The Annotator in turn will determine which sliding windows contain signs of cyberbullying behaviour. The set of outputs from the Annotator will then be fed to the Judges to determine whether a particular social media session should be deemed as cyberbullying by aggregating the outputs. In the following subsections, we will explain these four components in detail.

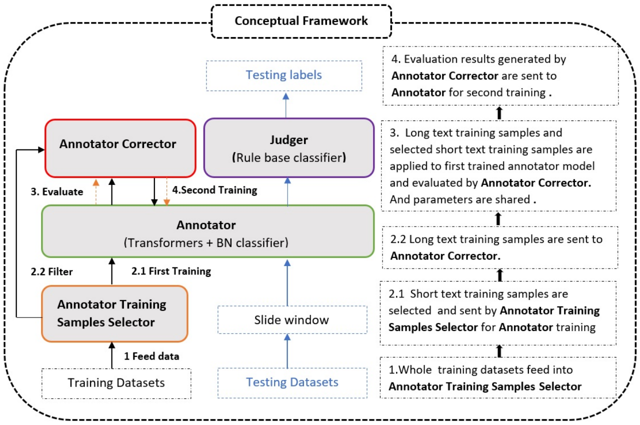

Conceptual Framework
Testing labels
4. Evaluation results generated by Annotator Corrector are sent to Annotator for second training .
Annotator Corrector
Judger
(Rule base classifier)
3. Long text training samples and selected short text training samples are applied to first trained annotator model and evaluated by Annotator Corrector. And parameters are shared .
3. Evaluate
4.Second Training
Annotator
(Transformers + BN classifier)
2.2 Long text training samples are sent to Annotator Corrector.
2.2 Filter
2.1 First Training
Annotator Training Samples Selector
Slide window
2.1 Short text training samples are selected and sent by Annotator Training Samples Selector for Annotator training
1 Feed data
Training Datasets
Testing Datasets
1.Whole training datasets feed into Annotator Training Samples Selector

Figure 3: Conceptual Framework. The dark line represents the training data flow. The blue line shows the test data flow. The orange dashed line represents the annotator's initial training parameters, which are used to initialise the annotator's second training.

4.2.1 *Annotator training samples selector.*

To adapt transformers as base models for training annotators, the first factor we need to consider is the selection of input data to facilitate the first annotator training. Instinctively, a short social media session within input limits (< 512 tokens) is expected to contain initially sufficient, complete information to help models understanding typical cyberbullying patterns. The function of Annotator Training Samples Selector is to select a suitable size of short text sessions for the annotator, and send the rest of the long text sessions to the Annotator training corrector for preparing the annotator to learn with long texts. Furthermore, considering that in some datasets there may not be a sufficient number of short texts to provide learning, the Annotator training sample selector does not have to obey the input length limit of the transformers to select short texts. Instead, the selector can iteratively extend the length of the short text selection until finding the appropriately sized training material, and then apply an arbitrary truncation method for the annotator [33].

4.2.2 *Annotator and Annotator corrector.*

The annotator is composed of a transformer encoder and two fully connected layers on top. The two-layer feed forward network is designed with ReLU activation and 512 hidden size for the first layer and Softmax activation for the output layer. The annotator corrector attempts to predict labels for all sliding windows through a series of rules related to the ground truth of the relevant session. The output of the annotator corrector is the training data for secondary annotator supervised learning. These two components are combined to train a transformer-based machine annotator that can identify cyberbullying incidents based on partial information from the sliding windows rather than the complete information from the whole session. The Annotator learning process is illustrated in Figure 4 and described below.

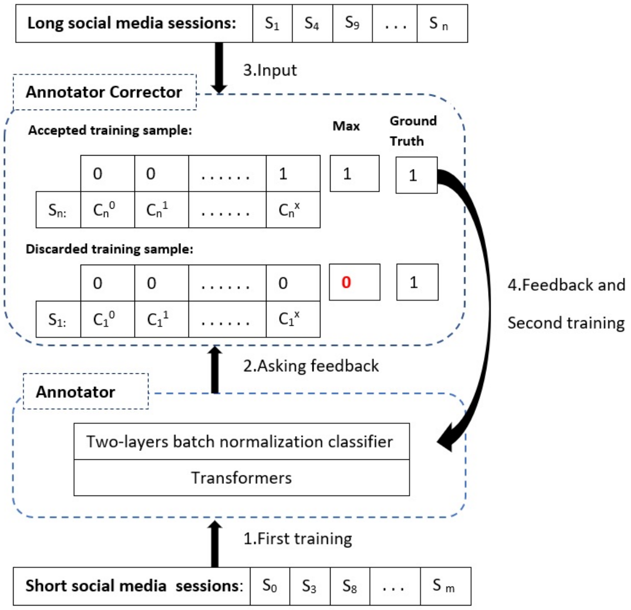


Figure 4: The Annotator learning process

Annotator first training: Annotator's first supervised training by using the selected short text social media sessions.

Input long text social media sessions and ask for feedback: The first trained annotator and long-text social media sessions are input to the Annotator Corrector for feedback. All long-text social media sessions are divided into sliding windows, which are arranged in its original order as they occur within the session. The Annotator will give a soft label for each window, and the Annotator Corrector will use the following three rules to select the new training datasets as input data for the Annotator's second training.

1) If a session's ground truth is False, then label all windows of the session as False, ignoring the soft labels predicted by the Annotator.

2) If a session's ground truth is True, and at least one chunk's soft label predicted by the Annotator is True, all windows keep the soft labels. Otherwise, the session will be taken from second training datasets.

3) All short text sessions keep their ground truth.

Annotator's second training: Each item of second training datasets is a window from a session with labels generated by the Annotator Corrector. The initialised parameters of the second training come from the weights of the first training Annotator for transferring the knowledge acquired from the short text session cyberbullying detection.

4.2.3 Judger. The Judger is a rule-based classifier that makes decisions following predefined rules. These rules can vary depending on the objective. In our case, focusing on session-based cyberbullying detection, we are looking at whether social media sessions meet the following criteria inherent to cyberbullying: **1)** Cyberbullying is a form of cyberaggression, **2)** there is a power imbalance among the individuals involved, and **3)** there is a repetition of the aggression. This means that cyberbullying events within sessions will meet the conditions:

$Max(A_k(K_i^0, \ldots, K_i^x)) \approx Y_i$,

where A_k is the Annotator, K_i is a text block and Y_i is the cyberbullying label.

Based on this intuition and considering the criteria for a session to be deemed cyberbullying, we synthesise the Judger's rules as follows: **Rule 1:** If there are no incidents of cyberbullying in any of the windows that belong to the social media session, the judge determines that there are no incidents of cyberbullying in the social media session; and **Rule 2:** A judge determines that there is an incident of cyberbullying in a social media session if there is at least one incident of cyberbullying in the windows pertaining to that social media session.

5 EXPERIMENT SETTINGS

Through our experiments, we aim to answer the following research questions: **RQ1:** How does LS-CB perform in the long session based cyberbullying detection task as compared to state-of-the-art cyberbullying detection methods? **RQ2:** Do all of the components of LS-CB positively contribute to the performance? **RQ3:** Based on qualitative analyses, do both Hypotheses 1 & Hypothesis 2 hold true?

5.1 Datasets

To conduct informative and objective experiments in this study, we use two different session-based cyberbullying datasets from two different types of social media platforms (Vine and Instagram), selected based on the three criteria: **1)** a dataset was collected with social media sessions as the collection unit, **2)** the data collection followed a strict definition of cyberbullying, and **3)** the dataset has been widely used and evaluated in existing session-based cyberbully detection research.

Both the selected datasets are manually annotated and publicly available session-based datasets. Each session in the datasets includes multi-interactive conversations. They both suffer from class imbalance, with sessions flagged as cyberbullying being a minority compared to neutral sessions. Although in the data collection the authors intentionally limited the number of comments per session to reduce the session length, we can still observe that the length per session far exceeds the maximum input limit of most transformers. See Table 1 for statistics of both datasets.

Instagram [15]. Instagram is a social media platform where users can post images associated with comments, that others can like or reply to. In this dataset, authors collect each media object and its associated comments, which altogether make a social media session. Each media object contains the following information: media URL, media content, post time, caption and the number of likes/followed/shared.

Table 1: Dataset statistics.

	Instagram	Vine
Cyberbullying Ratio	0.29	0.30
# Sessions	2218	970
# Comments	159,277	70,385
# Users	72,176	25,699
Average length per session	900	698
Maximum length per session	10678	4511
Average # users per session	33	26

Vine [27]. Vine is a video-based online social network, where users can post videos and others can comment on them. This dataset was created by the same authors as the Instagram dataset and therefore employed a very similar approach to collect and annotate this dataset.

Note that both datasets provide labels for cyberbullying or neutral at the session level. In our research making predictions at both session and sliding window level, we therefore focus the evaluation at the session level. Predictions made at the sliding window level are used as pseudo labels to help improve the session level predictions.

5.2 Experiment Setup

5.2.1 Pre-processing. To set up the experiments, we mainly follow the pre-processing method of Ge et al. [13] when it comes to aggregating session data and cleaning the data. Due to our different objectives from the original authors, we do not perform oversampling of the data and we do not truncate the sessions to limit their length. Since real-world cyberbullying is expected to be imbalanced, we test our framework in a more realistic, imbalanced scenario. In Ge et al. [13], the session length is set to 140 and the sentence length is set to 30, which is contrary to our goal of studying long sessions.

5.2.2 Hyperparameter settings. For a fair comparison, we adopt training hyperparameters that are recommended by Sun et al. [33], which is widely accepted for fine-tuning classification models; Batch size: 16; Learning rate (Adam): 2e-5; The number of epochs: 4.

5.2.3 Baselines. We compare our framework with three types of baseline models: State-of-the-art cyberbullying detection models, six Transformer-based pre-trained language models (PLM) and The Long-Document Transformer (Longformer).

In the case of state-of-the-art models, due to limited reproducibility, we report performance scores as shown in their papers, which are comparable given the same experiment settings of the original authors.

State-of-the-art cyberbullying detection models. We include three competitive baselines:

- **HANCD [4]:** HANCD consists of two levels of Hierarchical Attention Network(HAN), one at the word level and the other at the comment level. These two HANs can capture the differential importance of words and comments in different contexts. Then the bidirectional GRU is employed to capture the sequence of contents.
- **HENIN[3]:** HENIN focuses more on learning various interactions between heterogeneous objects displayed in social media sessions. A comment encoder is created to learn the representations of user comments through a hierarchical self-attention neural network so that the semantic and syntactic cues of cyberbullying can be captured. A post-comment co-attention mechanism learns the interactions between a posted text and its comments. Moreover, two graph convolutional networks are leveraged to learn the latent representations depicting how users interact with each other in sessions, and how posts resemble each other in terms of content.
- **TGBully[13]:** TGBully builds a unified temporal graph for each social media session, thereby modeling temporal dynamics and topic coherence throughout user interactions.

***Transformer-based pre-trained language models (PLM).*.** We test six transformer models, namely BERT [10], Roberta [19], MPNET [32], Electra [7], Distilbert [31] and XLnet [42]. As these models inherently need the input data to be truncated, we follow the truncation strategy defined by Sun et al. [33].

LongFormer, the long document transformer. LongFormer [2] optimises the transformer's self-attention mechanism and combines local windowed attention and task-driven global attention.

5.2.4 Evaluation. We use three widely used evaluation metrics for the cyberbullying detection task and imbalanced datasets, namely recall, precision and micro-F1. We randomly choose 80% of media sessions for training and the remaining 20% for testing. Each model is run 5 times to report the average performance.

6 RESULTS

We present and discuss the results of our experiments, answering in turn to the three research questions we set forth.

6.1 RQ1: Overall Performance

Table 2 shows a comparison of all LS-CB variants with six transformer models against all the baselines. We observe that our proposed LS-CB model outperforms the wide range of baselines under study, particularly using the RoBERTa (LS-CB_Roberta) and the Mpnet transformer models (LS-CB_Mpnet); indeed, they consistently outperform all baselines.

Among the baseline models, we observe the very competitive performance of Transformer PLMs, consistently outperforming both state-of-the-art cyberbullying detection models as well as LongFormer, which are still behind LS-CB. Figure 5 enables a visual comparison of the performance of LS-CB with their Transformer-only counterparts. We see that our LS-CB variants generally improve the performance of their Transformer-based counterparts, with the exception of XLnet on the Vine dataset, where LS-CB_XLnet underperforms slightly.

6.2 RQ2: Ablation Analysis

To address RQ2 by assessing whether the different components of LS-CB are positively contributing to the final performance, we conduct ablation studies to analyse the impact of the Annotator Training Sample Selector and the Annotator Corrector. In the interest of focus, we use LS-CB_Roberta and LS-CB_MPNET as the base models to conduct the ablative experiments.

Datasets		Vine			Instagram		
Approach	Model	F1	Recall	Precision	F1	Recall	Precision
Cyberbullying detection models	HANCD	0.70	0.75	N /A	0.79	0.81	0.77
	HENIN	0.68	0.64	0.82	0.84	0.83	0.90
	TGBully	0.71	0.77	N /A	0.81	0.83	N /A
Long text transformers	LongFormer	0.62	0.67	0.60	0.72	0.72	0.72
Transformer-based pre-trained language models	BERT	0.79	0.79	0.78	0.83	0.83	0.83
	Roberta	0.82	0.79	0.81	0.86	0.85	0.86
	MPNET	0.81	0.80	0.83	0.83	0.82	0.84
	Electra	0.81	0.80	0.80	0.84	0.83	0.89
	XLnet	0.81	0.79	0.82	0.83	0.83	0.82
	Distilbert	0.78	0.79	0.78	0.84	0.81	0.88
Transforms with Our Framework	LS-CB_BERT	0.81	**0.83**	0.80	0.86	**0.86**	0.87
	LS-CB_Roberta	**0.84**	**0.85**	**0.85**	**0.87**	**0.86**	0.89
	LS-CB_MPNET	**0.87**	**0.87**	**0.88**	**0.86**	**0.87**	0.86
	LS-CB_Electra	0.82	**0.83**	**0.84**	**0.87**	**0.86**	0.88
	LS-CB_XLnet	0.79	0.77	**0.84**	0.82	0.83	0.84
	LS-CB_Distilbert	0.79	0.81	0.79	0.84	0.83	0.89

Table 2: A comparison of baselines with LS-CB on Vine and Instagram.

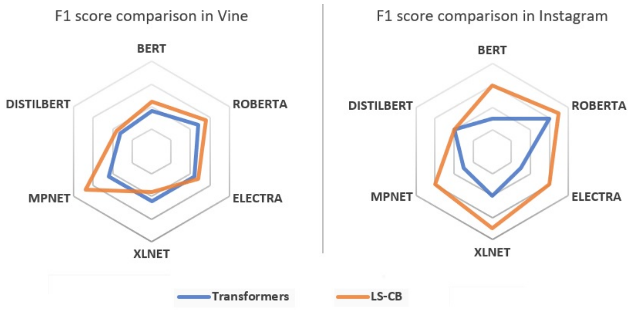


Figure 5: The impact of LS-CB on six transformers

Datasets	Vine			Instagram		
Model	F1	R	P	F1	R	P
LS-CB_Roberta	0.84	0.85	0.85	0.87	0.87	0.98
Without Selector	0.80	0.79	0.82	0.83	0.82	0.83
Without Corrector	0.76	0.75	0.80	0.80	0.79	0.82
LS-CB_MPNET	0.84	0.83	0.86	0.86	0.87	0.86
Without Selector	0.81	0.84	0.80	0.83	0.83	0.83
Without Corrector	0.76	0.78	0.75	0.82	0.82	0.81

Table 3: Performance comparison of CD_LS models on Vine.

6.2.1 Ablating the Annotator Training Sample Selector. We disable the Annotator Training Sample Selector by feeding all sessions to the annotator for the first training. We apply the truncation method for long text sessions. The results of the framework with and without Annotator Training Sample Selector are summarised in Table 3. We see that the performance decreases across all transformers when it is disabled, after truncation, some important information is lost, causing some data to become noisy during model training. We select a 852 token length session in Vine to illustrate it in Figure 6. Where cyberbullying incidents occurred outside the scope of 512 tokens and there is no sample selector, the incidents will be filtered out after truncation.

6.2.2 Ablating the Annotator Corrector. When we disable the Annotator Corrector, the secondary annotator training will not take place. The sliding window technique is directly applied to segment all test sessions, then applying the first trained annotator to predict a label on each sliding window, which will then be fed to the Judger. From Table 3, we see that the performance decreases across all transformers when the Annotator corrector is disabled. The reason is that the Annotator only learned the data distribution of short sessions after the first training, and did not learn the data distribution of long sessions. So the acquired knowledge is not comprehensive. After the long text in the test data is segmented, the distribution of the lengths observed is altered, which the Annotator struggles with leading to performance drop.

6.3 RQ3: Validation of hypotheses

6.3.1 Hypothesis 1: Cyberbullying events can be detected through aggregated analyses of smaller blocks of text that form the lengthy social media sessions. The learning goal of the annotators in LS-CB is to independently decide whether smaller blocks of information extracted from long sessions contain cyberbullying events. Our results indicate that splitting sessions into smaller blocks to then aggregate the predictions leads to improved performance. Our objective here is to further validate that this is not coincidental and that we can achieve the same through sliding windows of different size. Hence, we experiment on LS-CB_Roberta model with different

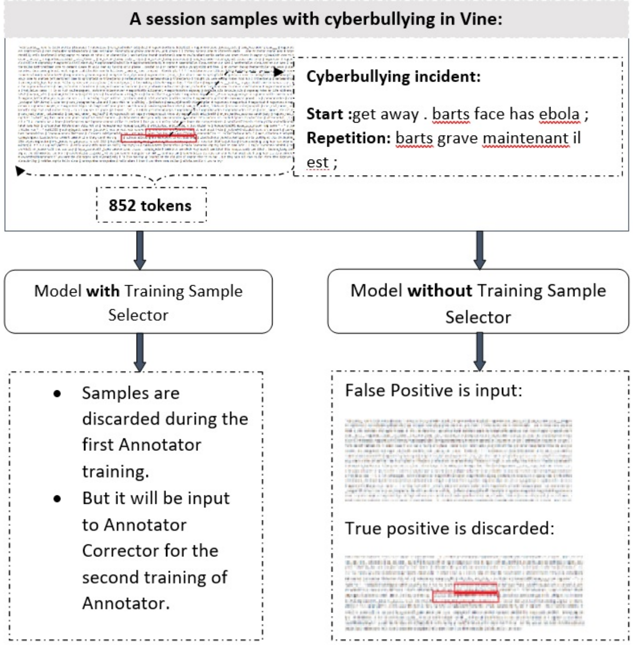
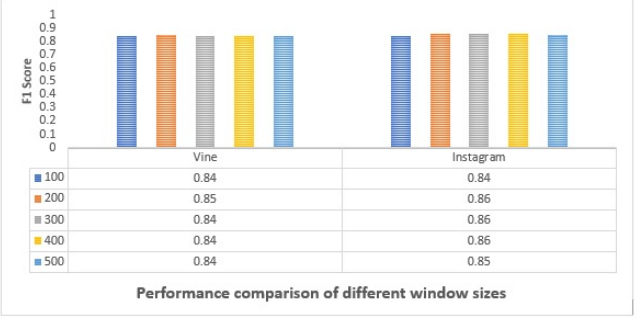

A session samples with cyberbullying in Vine:
Cyberbullying incident:
Start :get away . barts face has ebola ;
Repetition: barts grave maintenant il est ;
852 tokens
Model with Training Sample Selector
Model without Training Sample Selector
Samples are discarded during the first Annotator training.
But it will be input to Annotator Corrector for the second training of Annotator.
False Positive is input:
True positive is discarded:

Figure 6: A sample of Annotator Training Sample Selector's ablation analysis.

window sizes [100, 200, 300, 400, 500] and step sizes [90, 180, 270, 360, 450]. From Figure 7 we see that even with a much smaller window size of 100 tokens, the annotator can still support the detection of cyberbullying events. There is in fact little variation of performance with different window sizes.

We therefore confirm hypothesis 1, proving the potential of aggregating predictions on smaller portions of each session.



	Vine	Instagram
▮ 100	0.84	0.84
▮ 200	0.85	0.86
▮ 300	0.84	0.86
▮ 400	0.84	0.86
▮ 500	0.84	0.85

Figure 7: Performance comparison of different window size on Vine and Instagram.

6.3.2 Hypothesis 2: Cyberbullying incidents can occur at different points of a session, hence requiring holistic analyses of sessions. Here we analyse the pseudo-labels generated by the annotator for the different windows across sessions, as an approximation to analyse where the cyberbullying incidents are observed. We look at the results with the best-performing LS-CB_Roberta model with a window size of 400 and a step size of 360. Figure 8 depicts the aggregate statistics of positions (i.e. indices of windows) where cyberbullying incidents occur. We observe remarkable differences between the two datasets. Where the majority of cyberbullying incidents occur at the beginning in the Instagram dataset, these are more uniformly spread throughout the entire session for the Vine dataset. These results indicate that (1) as hypothesised, cyberbullying incidents can occur anywhere within a session, highlighting the importance of avoiding session truncation, and (2) LS-CB can handle the different cyberbullying patterns observed across datasets.

The differences in patterns across both datasets may be largely due to the data collection strategies. This may also explain the lower performance scores (see Table 2) achieved overall by the different models on the Vine dataset, due to its more challenging nature.

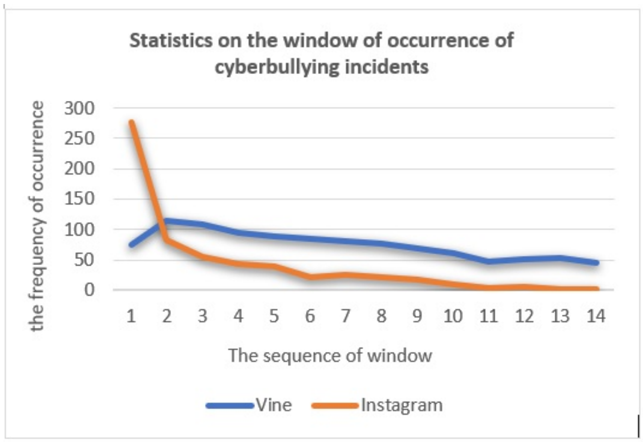


Figure 8: Frequency of cyberbullying across windows.

7 CONCLUSION

With LS-CB, we have introduced the first attempt at handling lengthy social media sessions for cyberbullying detection through the use of Transformer model. Beyond the frequent limit of 512 tokens, LS-CB enables handling sessions of unrestricted length. Through experiments on two datasets using six different Transformer models, we have shown the effectiveness of our framework, outperforming a wide range of competitive baselines.

Our experiments also enable validating our hypotheses that lengthy social media sessions can be processed as an aggregate of smaller fragments for cyberbullying detection, and that the cyberbullying incidents can happen at different points of a session. Our work does however have some limitations, as our evaluation is limited to the performance on the entire sessions, and our predicted pseudo-labels for the smaller blocks within sessions cannot be evaluated with existing datasets and annotations. This however calls for the collection and finer-grained annotation of cyberbullying datasets to enable research in this direction.

Our work opens up other avenues for future work, including testing the effectiveness of LS-CB for other classification tasks involving social media sessions, e.g. rumour stance classification in long social media conversations [47].

REFERENCES

[1] Alan Baddeley. 1992. Working memory. *Science* 255, 5044 (1992), 556–559.

[2] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).

[3] Hsin Yu Chen and Cheng Te Li. 2020. HENIN: Learning heterogeneous neural interaction networks for explainable cyberbullying detection on social media. In *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*. Association for Computational Linguistics (ACL), 2543–2552.

[4] Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, and Huan Liu. 2019. Hierarchical attention networks for cyberbullying detection on the instagram social network. In *Proceedings of the 2019 SIAM international conference on data mining*. SIAM, 235–243.

[5] Lu Cheng, Ruocheng Guo, Yasin N Silva, Deborah Hall, and Huan Liu. 2021. Modeling temporal patterns of cyberbullying detection with hierarchical attention networks. *ACM/IMS Transactions on Data Science* 2, 2 (2021), 1–23.

[6] Lu Cheng, Jundong Li, Yasin N Silva, Deborah L Hall, and Huan Liu. 2019. Xbully: Cyberbullying detection within a multi-modal context. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 339–347.

[7] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*. 1–18.

[8] Maral Dadvar, FMG de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.

[9] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*. Springer, 693–696.

[10] 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)*. 4171–4186.

[11] Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying bert to long texts. *Advances in Neural Information Processing Systems* 33 (2020), 12792–12804.

[12] Jatin Ganhotra and Sachindra Joshi. 2021. Does Dialog Length matter for Next Response Selection task? An Empirical Study. *arXiv preprint arXiv:2101.09647* (2021).

[13] Suyu Ge, Lu Cheng, and Huan Liu. 2021. Improving cyberbullying detection with user interaction. In *Proceedings of the Web Conference 2021*. 496–506.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[15] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*. Springer, 49–66.

[16] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*. 3–6.

[17] Aiqi Jiang and Arkaitz Zubiaga Zubiaga. 2023. SexWEs: Domain-Aware Word Embeddings via Cross-lingual Semantic Specialisation for Chinese Sexism Detection in Social Media. In *Proc. of ICWSM*.

[18] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. 2016. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779* (2016).

[19] 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 preprint arXiv:1907.11692* (2019).

[20] David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems* 30 (2017).

[21] Damiano Menin, Annalisa Guarini, Consuelo Mameli, Grace Skrzypiec, and Antonella Brighi. 2021. Was that (cyber) bullying? Investigating the operational definitions of bullying and cyberbullying from adolescents' perspective. *International journal of clinical and health psychology* 21, 2 (2021), 100221.

[22] Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 2018. Efficient and Robust Question Answering from Minimal Context over Documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1725–1735.

[23] Dan Olweus. 1994. Bullying at school: basic facts and effects of a school based intervention program. *Journal of child psychology and psychiatry* 35, 7 (1994), 1171–1190.

[24] Dan Olweus. 2001. Bullying at school: Tackling the problem. *OECD observer* (2001), 24–24.

[25] Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 838–844.

[26] Sayanta Paul and Sriparna Saha. 2020. CyberBERT: BERT for cyberbullying identification. *Multimedia Systems* (2020), 1–8.

[27] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in Vine. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 617–622.

[28] Oscar Reyes and Sebastian Ventura. 2019. Performing multi-target regression via a parameter sharing-based deep network. *International journal of neural systems* 29, 09 (2019), 1950014.

[29] Hugo Rosa, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior* 93 (2019), 333–345.

[30] Semiu Salawu, Yulan He, and Joanna Lumsden. 2017. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing* 11, 1 (2017), 3–24.

[31] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[32] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems* 33 (2020), 16857–16867.

[33] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification?. In *China national conference on Chinese computational linguistics*. Springer, 194–206.

[34] Jatin Karthik Tripathy, S Sibi Chakkaravarthy, Suresh Chandra Satapathy, Madhulika Sahoo, and V Vaidehi. 2020. ALBERT-based fine-tuning model for cyberbullying analysis. *Multimedia Systems* (2020), 1–9.

[35] Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PloS one* 13, 10 (2018), e0203794.

[36] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the international conference recent advances in natural language processing*. 672–680.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[38] Filippos Karolos Ventirozos, Iraklis Varlamis, and George Tsatsaronis. 2017. Detecting aggressive behavior in discussion threads using text mining. In *International Conference on Computational Linguistics and Intelligent Text Processing*. Springer, 420–431.

[39] Kai Wang, Jiahui Huang, Yuqi Liu, Bin Cao, and Jing Fan. 2020. Combining feature selection methods with bert: An in-depth experimental study of long text classification. In *International Conference on Collaborative Computing: Networking, Applications and Worksharing*. Springer, 567–582.

[40] Kaige Wang, Qingyu Xiong, Chao Wu, Min Gao, and Yang Yu. 2020. Multi-modal cyberbullying detection on social networks. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[41] Jaideep Yadav, Devesh Kumar, and Dheeraj Chauhan. 2020. Cyberbullying detection using pre-trained bert model. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 1096–1100.

[42] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[43] Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. 2019. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *The World Wide Web Conference*. 3427–3433.

[44] Peiling Yi and Arkaitz Zubiaga. 2022. Cyberbullying detection across social media platforms via platform-aware adversarial encoding. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1430–1434.

[45] Peiling Yi and Arkaitz Zubiaga. 2022. Session-based Cyberbullying Detection in Social Media: A Survey. *arXiv preprint arXiv:2207.10639* (2022).

[46] Ruixuan Zhang, Zhuoyu Wei, Yu Shi, and Yining Chen. 2019. BERT-AL: BERT for Arbitrarily Long Document Understanding. (2019).

[47] Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management* 54, 2 (2018), 273–290.