COGNITIVE SCIENCE
RESEARCH GROUP
Queen Mary University of London

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

Peiling Yi and Arkaitz Zubiaga

About Me

- **Experience**
  - Over 10 years in industry as software engineer and project manager
  - Master in Software engineering at QMUL
  - **NOW-**

    A part-time **PhD student** at cognitive science research group, QMUL

    A part-time **teaching fellow** at QMUL

    A **Turing Enrichment student**

    A Candidate of indoor climbing instructor

- **Research**
  - Doing research in **Youth cyberbullying detection across different social medias**
  - Also interested in Novel deep **transfer learning** and **fair machine learning algorithms**.

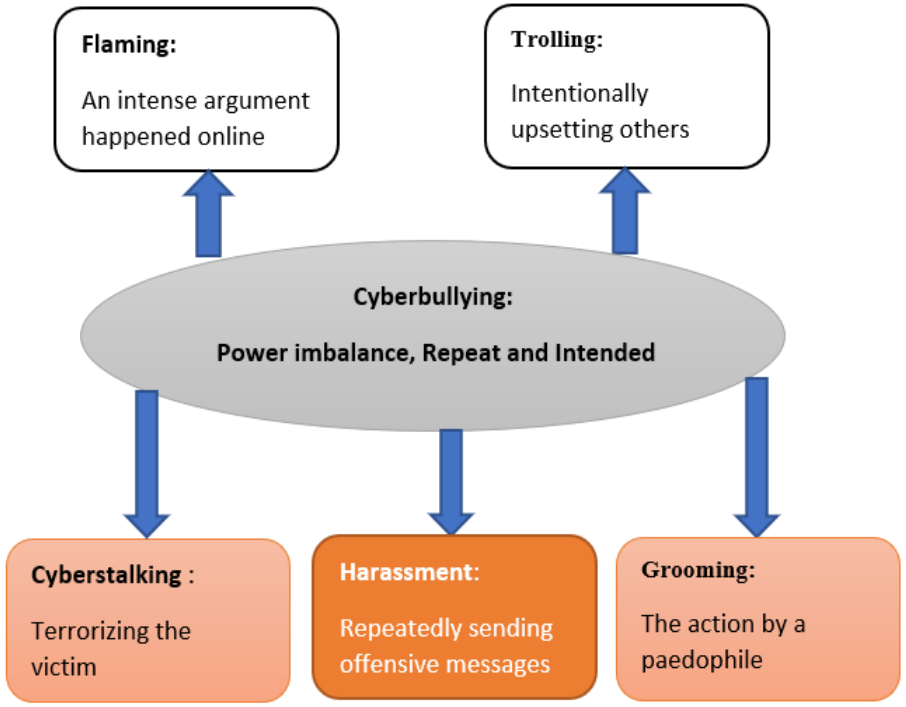
Contents

- What is Cyberbullying
- What is Cyberbullying detection
- Challenges and my studies
- Learning like human annotators:
  Cyberbullying detection in lengthy social media sessions
  - Motivation
  - Methods
  - Experiments
  - Results
  - Why is the method useful?

Contents

- **What is Cyberbullying**
- What is Cyberbullying detection
- Challenges and my studies
- Learning like human annotators:
  Cyberbullying detection in lengthy social media sessions
  - Motivation
  - Methods
  - Experiments
  - Results
  - Why is the method useful?

Backgrounds -**What is Cyberbullying?**



Orange indicates crime and **dark orange** indicates this is a crime and is the most common behaviour among youth cyberbullying

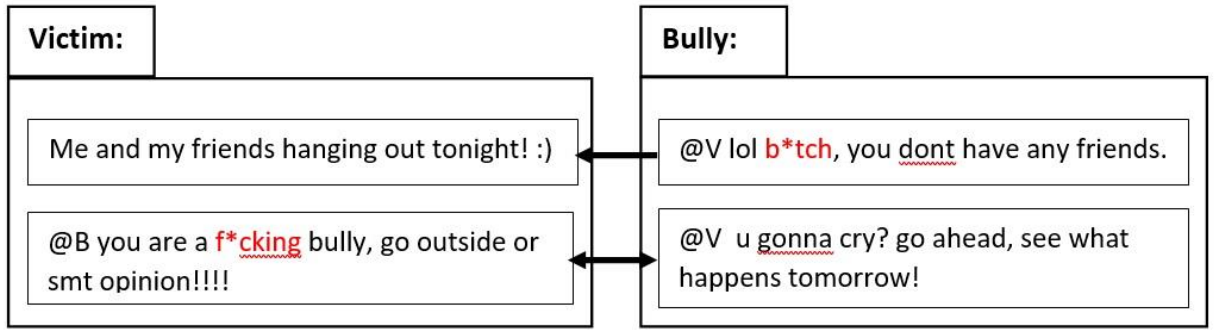
A form of bullying that is perpetrated through online devices and widely happened among **adolescents**.

The precise definition of cyberbullying **varies** slightly across studies and countries.

But **two characteristics** are widely accepted:

1.Repeated aggression - *A bully recurrently sends mocking message*.

2.Power imbalance - *Reveal personal or sensitive information of an indefensible victim.*

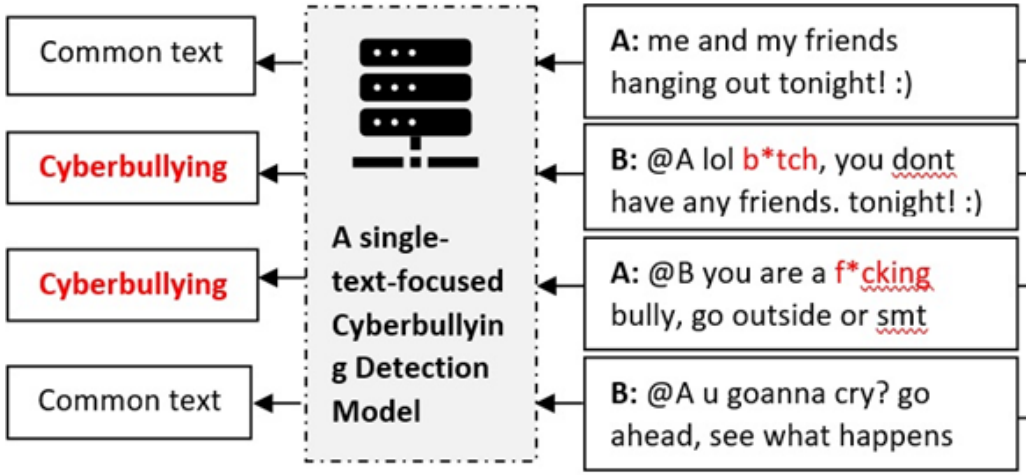


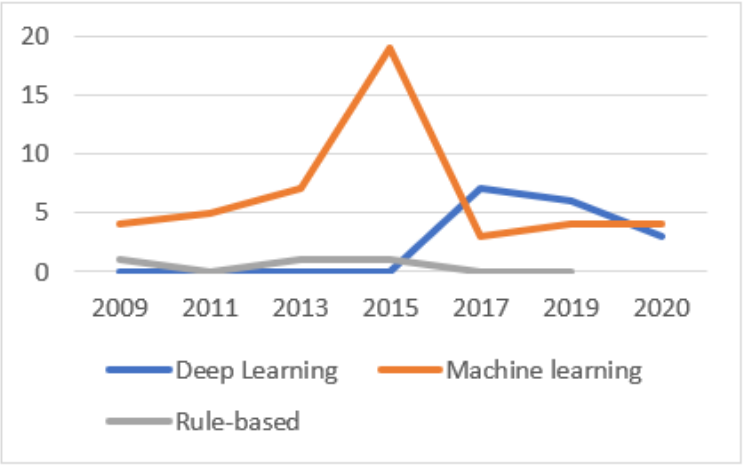
Contents

- What is Cyberbullying
- **What is Cyberbullying detection**
- Challenges and my studies
- Learning like human annotators:
  Cyberbullying detection in lengthy social media sessions
  - Motivation
  - Methods
  - Experiments
  - Results
  - Why is the method useful?

Backgrounds **-What is Cyberbullying Detection ?**

The most common task in cyberbullying detection is to distinguish bullying posts from other general social media posts.

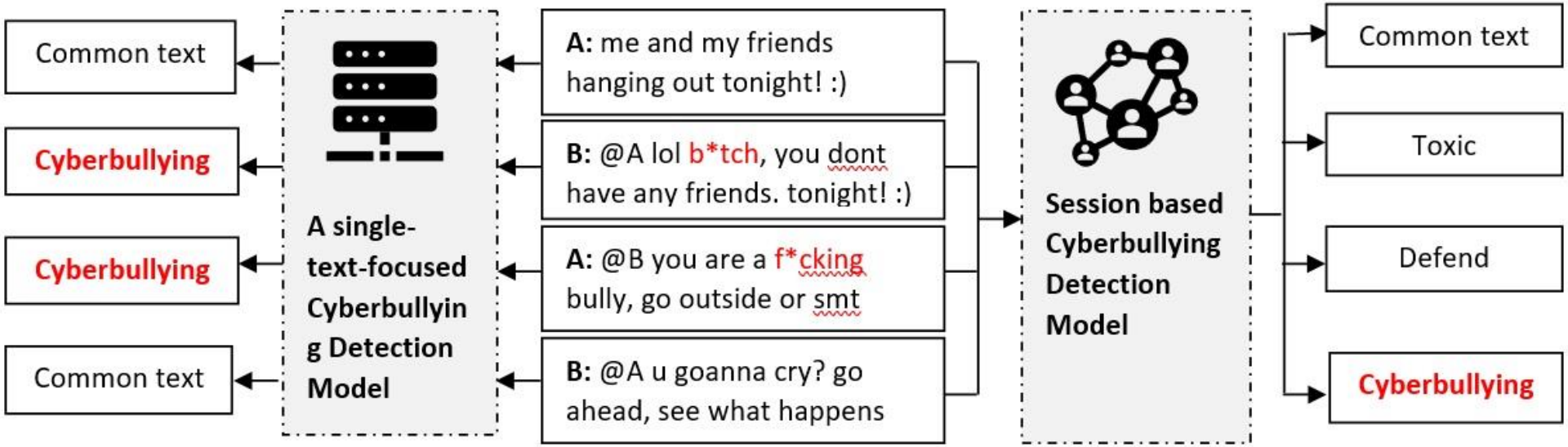




Backgrounds -What is Cyberbullying Detection ?

Past: Existing research in cyberbullying detection has predominantly focused on methods that analyse **isolated social media posts**.

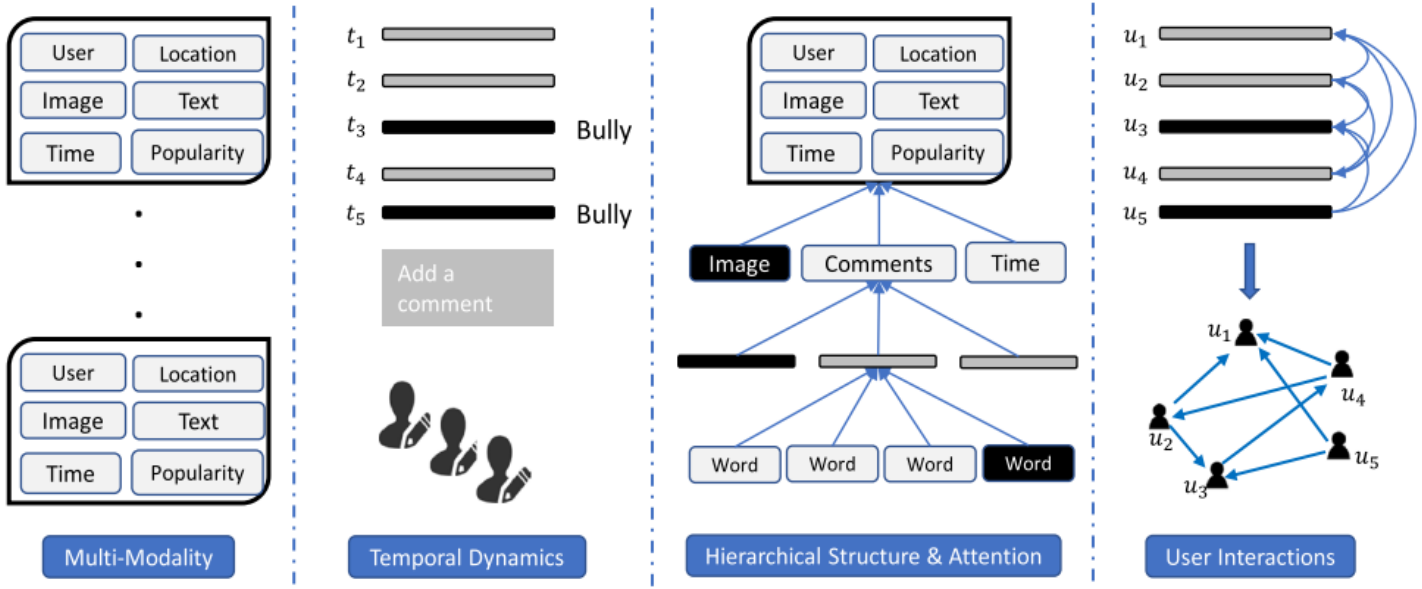
Now: Research on **a sequence of posts** and associated multimedia content–a holistic view of how the abuse develops. **Different** from other abusive langue detection model.



Backgrounds **-What is session based Cyberbullying Detection ?**

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.



[Session-Based Cyberbullying Detection: Problems and Challenges]

Contents

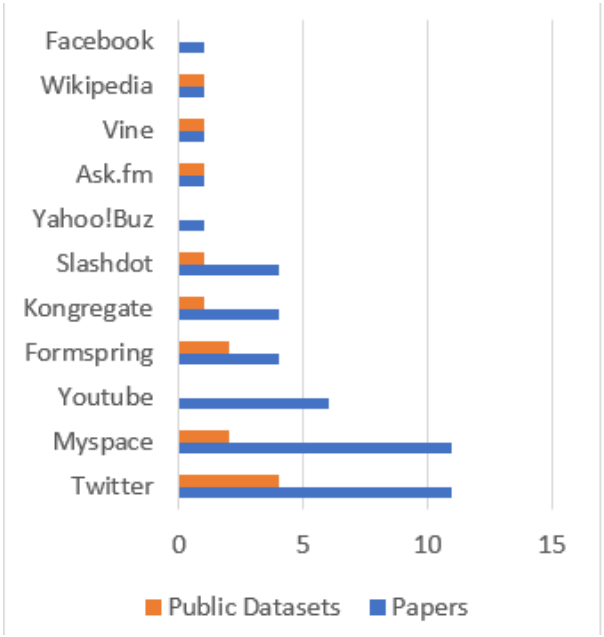
- What is Cyberbullying
- What is Cyberbullying detection
- **Some challenges and my studies**
- Learning like human annotators:
  Cyberbullying detection in lengthy social media sessions
  - Motivation
  - Idea
  - Methods
  - Experiments
  - Results
  - Why is the method useful?

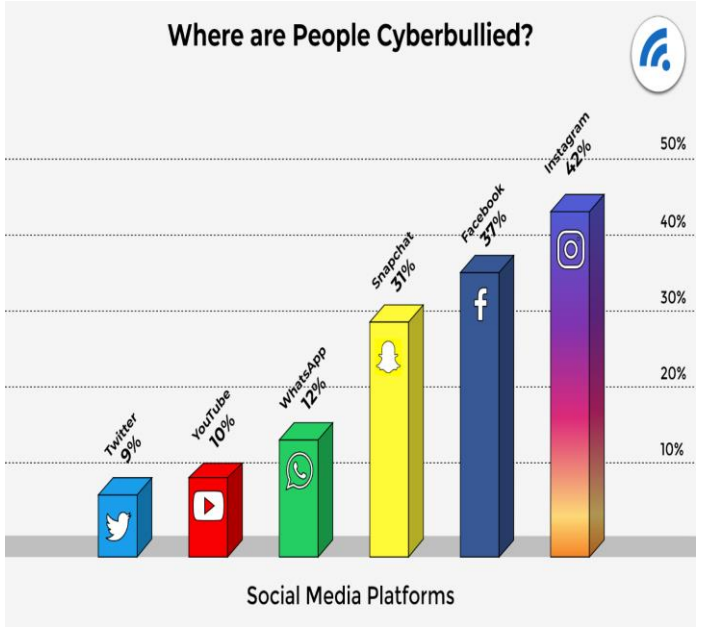
Youth cyberbullying detection across different social medias

SOCIETY ISSUES:

- The **fears** of cyberbullying and its impact on children and teens are still considerably increasing. *(Macaulay et al.,)*
- In the US, Pew Research Centre found **59%** of teens have been bullied online, **90%** believe online harassment is a problem and **63%** says it is a major problem*(Monica Anderson)*
- In the UK, A survey from Young minds& the children's society indicates **74% of 12 to 15 years olds** have a profile on social media,**39%** of young people said they have experienced cyberbullying and **60%** have seen somebody be harassed online *(The Children's Society and YoungMinds*)
- The most troublesome are most teenagers believe that teachers, social media companies **have failed** to address the issue *(The Children's Society and YoungMinds).*

RESEARCH ISSUES:





Youth cyberbullying detection across different social medias

Challenges:

1. Adolescent cyberbullying is prevalent but under-researched on adolescent cyberbullying.

Lack of age and cyberbullying datasets →

Studies:

Transferring knowledge from the source to the target social media :*Weakly Supervised Cross-platform Teenager Detection with Adversarial BERT.*

2. Lower performance on models' generalisability across different social media

Struggling a situation where there is a big data drift from source to target. →

Combining a Multi-Transformer embedding Alignment Strategies into Adversarial Networks, forced Target Transformer encoder to map the target input to source Transformer latent representation space :*The Cyberbullying detection across social media platforms via platform-aware adversarial encoding*

3. How cyberbullying detection research within a broader social media session from a data and methodological perspective.

Less focus on social media sessions →

Session-based Cyberbullying Detection in Social Media: A Survey

4. the lengthy nature of social media sessions challenges the applicability and performance of session-based cyberbullying detection models

Models struggle to adopt lengthy input →

Aggregate the predictions made by transformer models on smaller sliding windows extracted from lengthy social media sessions, leading to an overall improved performance *Learning like human annotators: Cyberbullying detection in lengthy social media sessions*

Contents

- What is Cyberbullying
- What is Cyberbullying detection
- Challenges and my studies
- **Learning like human annotators: Cyberbullying detection in lengthy social media sessions**
  - **Motivation**
  - Idea
  - Methods
  - Experiments
  - Results
  - Why is the method useful?

Motivation-Current modelling methods

- **Hierarchical networks with attention:** This approach leverages the hierarchical network to reflect the structure of a social media session
- **Multimodal learning:** Considers representing the joint representations of different multimodal data
- **User interaction modelling:** Relies on the assumption that cyberbullying events often take place in the form a series of interactions. Therefore, approaches incorporating sequences of user interactions have also been studied.
- **Transformer-based pre-trained language models**

Truncate is common!

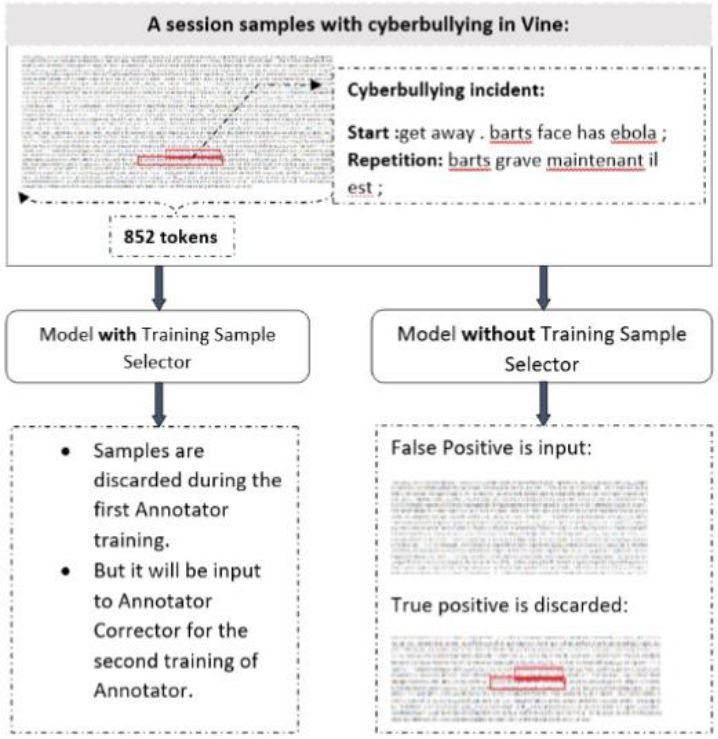
Motivation- Cyberbullying detection in lengthy social media sessions

- Lengthy is the nature!

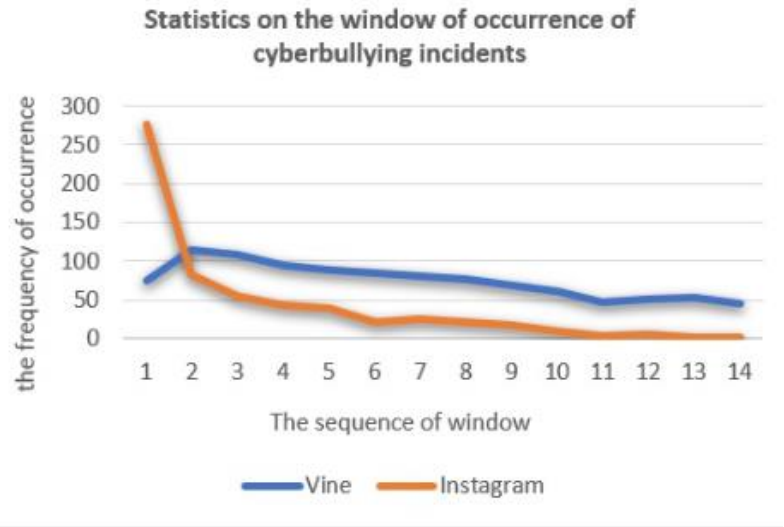
Dataset statistics.

	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

- False positive as input!



- Cyberbullying event can occur different points in lengthy session!



Motivation-Long text classification

- **Text Truncation:** 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.

- Information loss

- **Selecting relevant sentences:** Highly dependent on downstream tasks and selection methods.

- No study for cyberbullying detection.

- **Hierarchical transformers:** Reserved all inputs and processing embeddings through the pipeline

- Attention shift

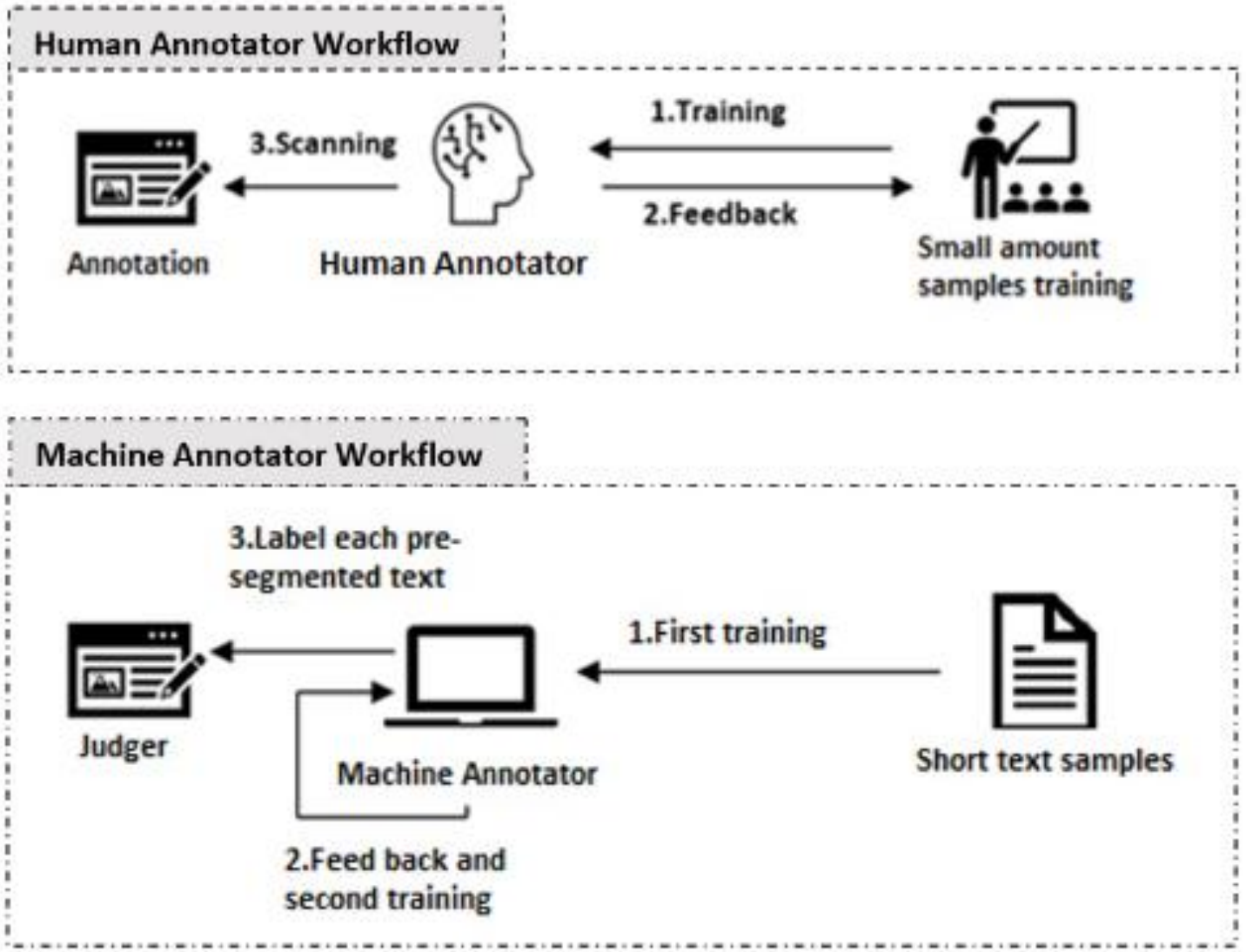
Contents

- What is Cyberbullying
- What is Cyberbullying detection
- Challenges and my studies
- **Learning like human annotators: Cyberbullying detection in lengthy social media sessions**
  - Motivation
  - **Methods**
  - Experiments
  - Results
  - Why is the method useful?

Methods - **Idea**

- **The solution is inspired by a human annotator workflow.**

- When dealing with cognitive tasks, humans retain a small amount of key information referred to as "**work memory**" , which emphasizes the importance of carefully crafting the description and key terms when explaining an annotation task .



Methods - **Theoretical Analysis**

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.

1) Parameter sharing: 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 session.

2) Batch normalisation: Normalise the distribution differences between sessions of different size by modulating the statistics of all BN layers throughout the annotator classifier layers.

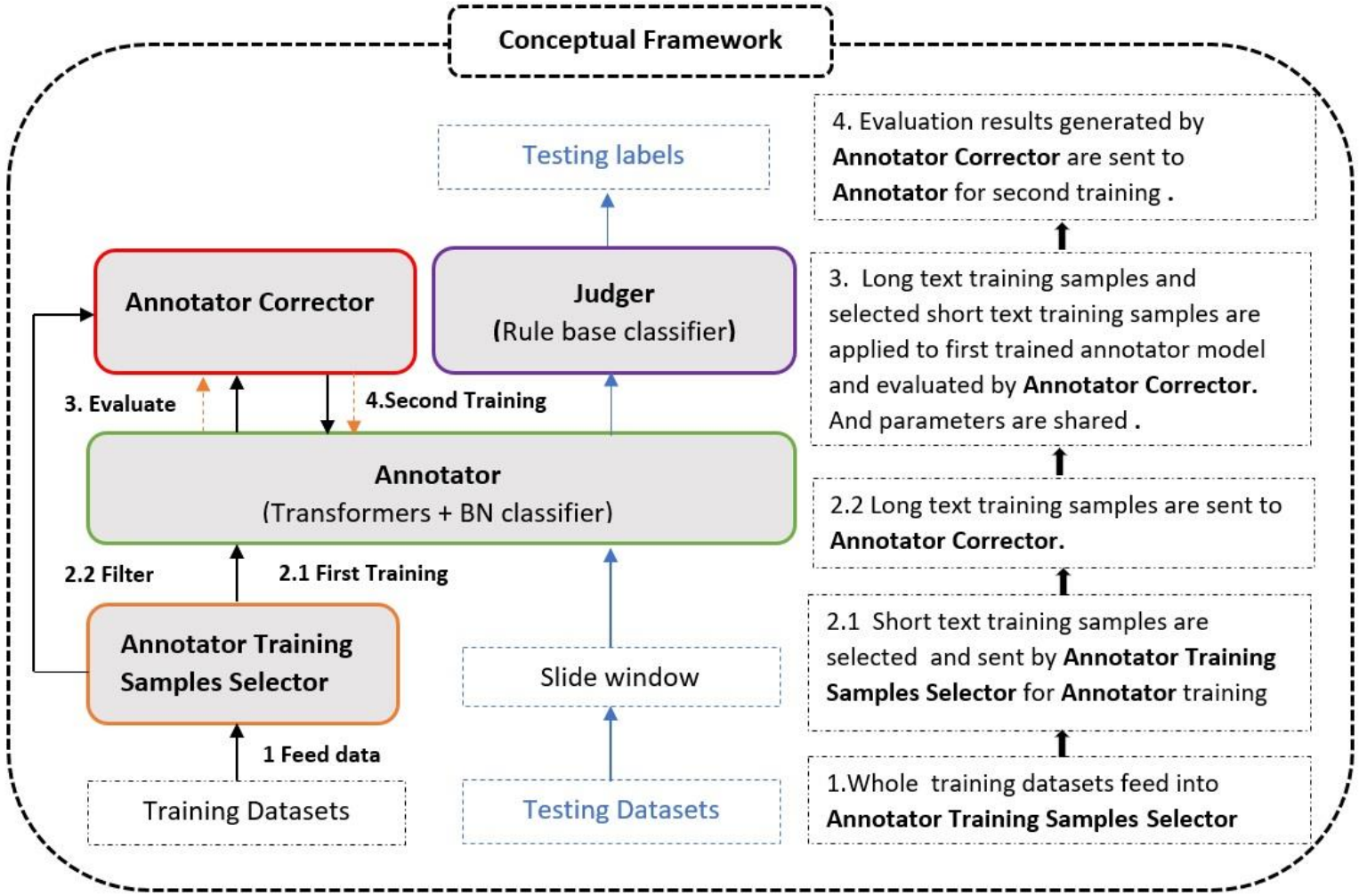
3) Memory consolidation: Memory replay studies in continual learning solve catastrophic forgetting by selecting a small portion short-session samples combined with all long-session samples for continuous training.

Methods – LS-CB Framework

Two hypotheses:

H1: Cyberbullying events can be detected through aggregated analyses of smaller blocks of text that form the lengthy social media sessions.

H2: Cyberbullying incidents can occur at different points of a session, hence requiring holistic analyses of session.



Methods-Annotator Training

- **Annotator:**

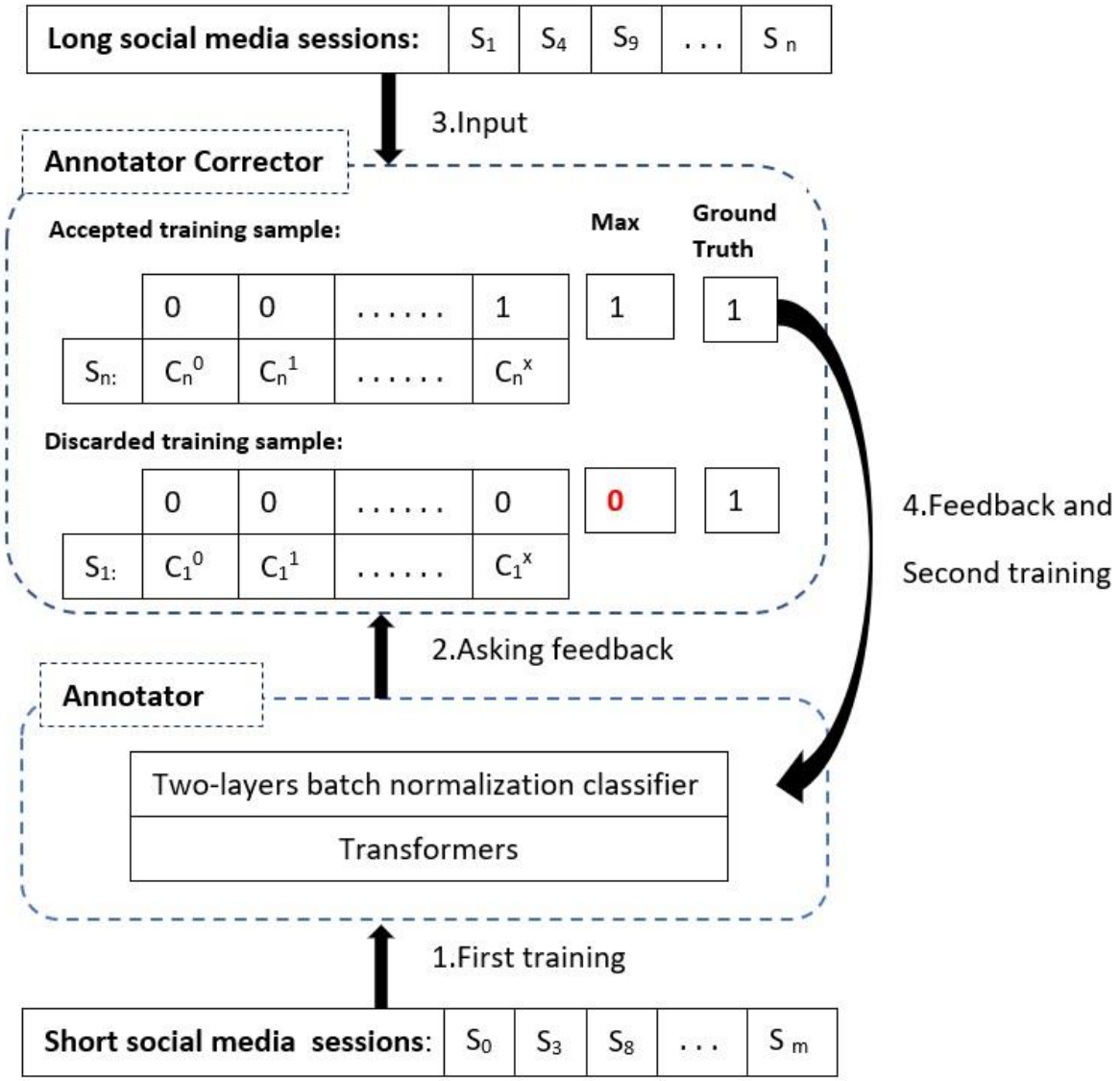
The two-layer feed forward network is designed with ReLU activation and 512 hidden size for the first layer and Softmax activation for the output layer.

- **Annotator corrector:**

A functional select component to decide what samples can been sent to second annotator training

- **Training targets:**

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.



Contents

- What is Cyberbullying
- What is Cyberbullying detection
- Challenges and my studies
- **Learning like human annotators: Cyberbullying detection in lengthy social media sessions**
  - Motivation
  - Methods
  - **Experiments**
  - Results
  - Why is the method useful?

Experiments - **Design**

- **RQ1: Superiors** to current research?
- **RQ2:** Do all of the components of LS-CB positively contribute to the performance and **Why**?
- **RQ3:** Based on qualitative analyses, do both Hypotheses 1 & Hypothesis 2 **hold true**?

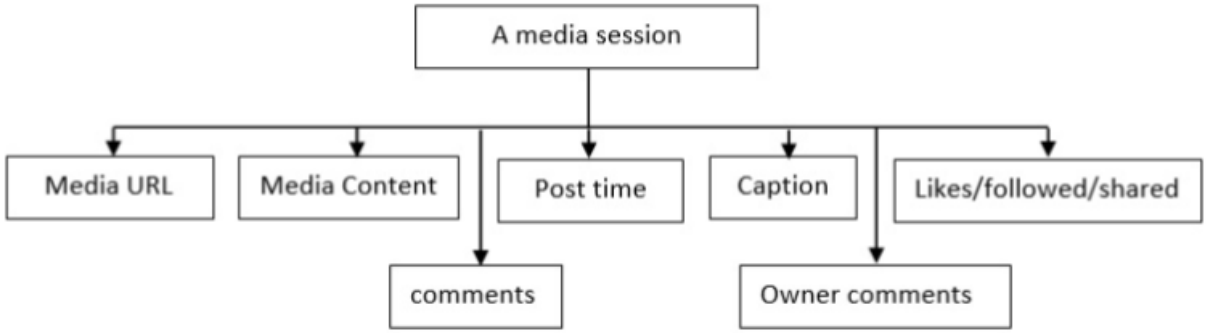
Experiments – Datasets selection

A dataset was collected with **social media sessions** as the collection unit.

The data collection followed a strict **definition** of cyberbullying.

The dataset has been **widely used** and evaluated in existing session based cyberbully detection research.

	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



A media session structure of Instagram and Vine

Experiments – Experiments setup

Follow the previous research experiments setup and try to fair comparison:

- **Pre-processing**- *[Suyu Ge, Lu Cheng, and Huan Liu. 2021. Improving cyberbullying detection with user interaction]*

But we do not :
1.Oversampling of the data
2.Truncate the sessions to limit their length.

- **Hyperparameter settings**- *[Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification]*

1.Batch size: 16;
2.Learning rate (Adam): 2e-5;
3.The number of epochs: 4

- **Baselines**

1.Session based cyberbullying detection models:
HANCD& HENIN& TGBully

2.Transformer-based pre-trained language models:
BERT & Roberta &MPNET & Electra & Distilbert & and XLnet

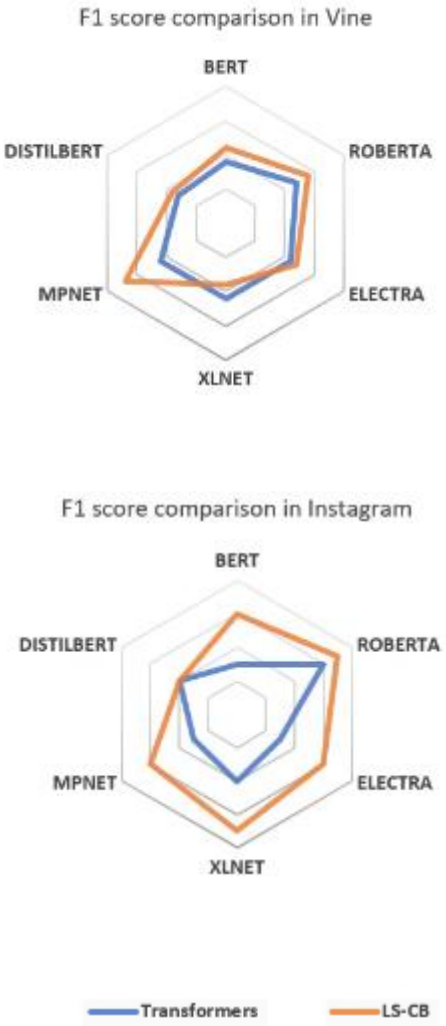
3. the long document transformer:
LongFormer

Contents

- What is Cyberbullying
- What is Cyberbullying detection
- Challenges and my studies
- **Learning like human annotators: Cyberbullying detection in lengthy social media sessions**
  - Motivation
  - Methods
  - Experiments
  - **Results**
  - Why is the method useful?

Results-**RQ1:** Superiors to current research?

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



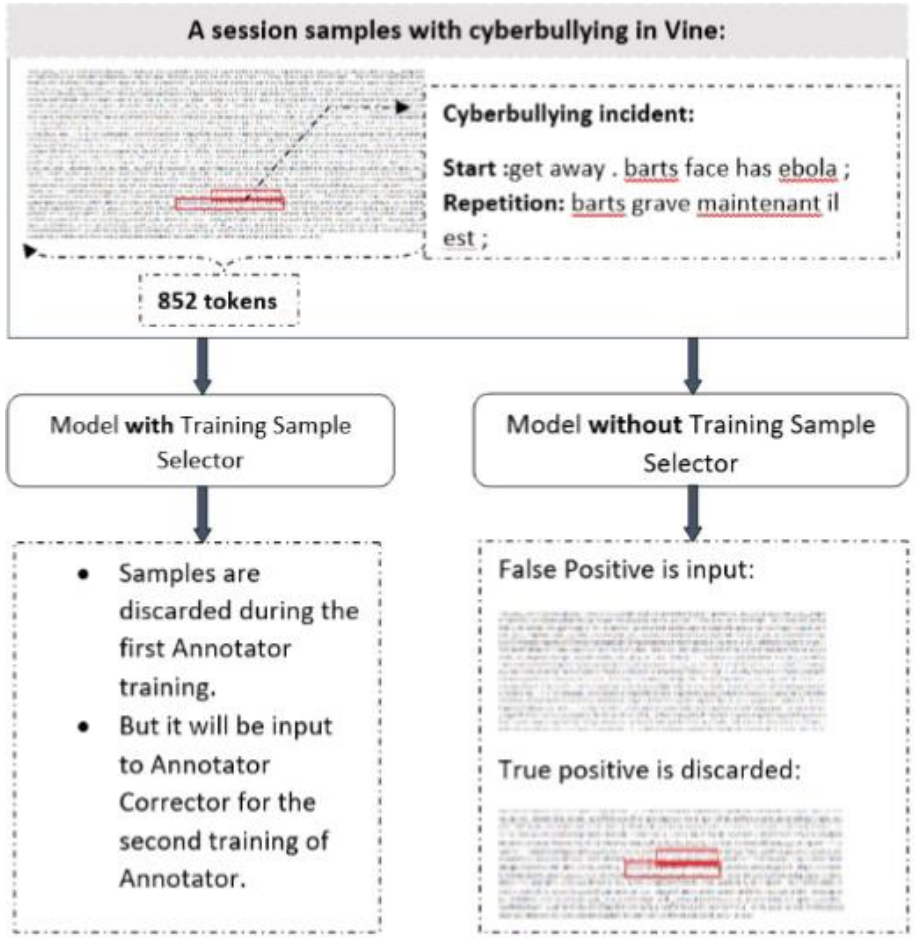
Results-**RQ2**: Ablation Analysis

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

- **Without Corrector**:

The reason is that the Annotator **only learned the data distribution of short sessions**. 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.

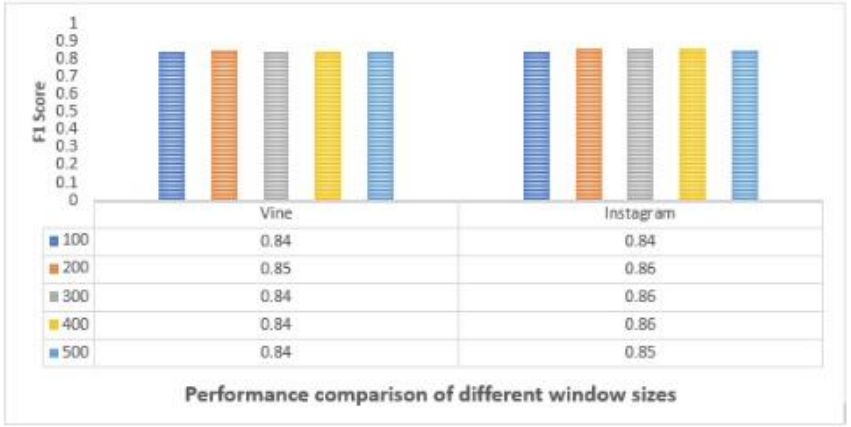
- **Without selector**:



Results-RQ3: Validating hypothesis

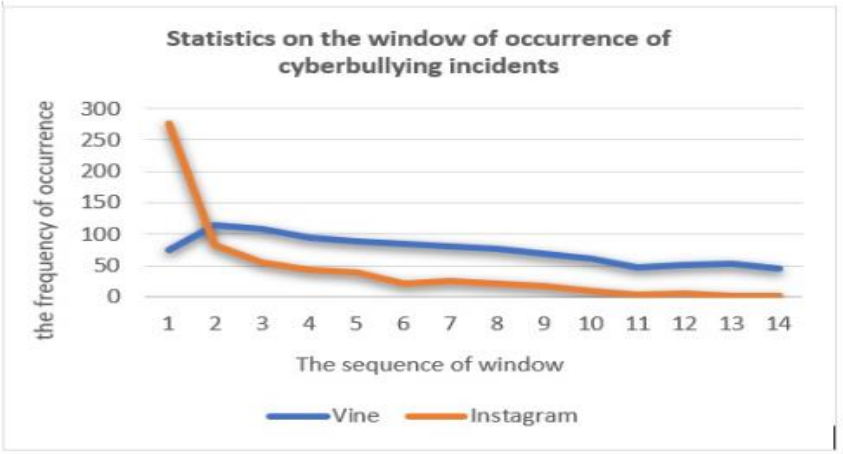
Validate the hypothese1:

Cyberbullying incidents can be detected through aggregated analysis of smaller chunks derived from lengthy social media sessions **.**



Validate the hypothese2:

Cyberbullying incidents can occur at different points of the session**.**

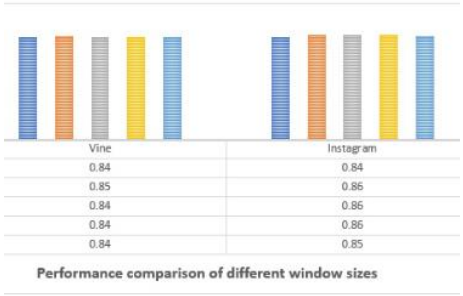


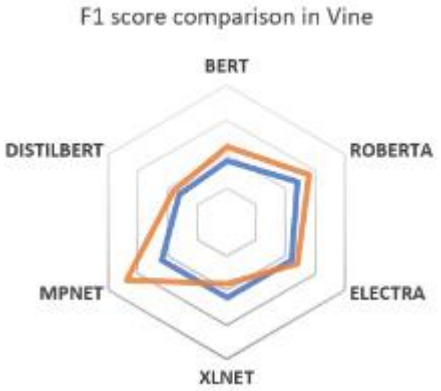
Contents

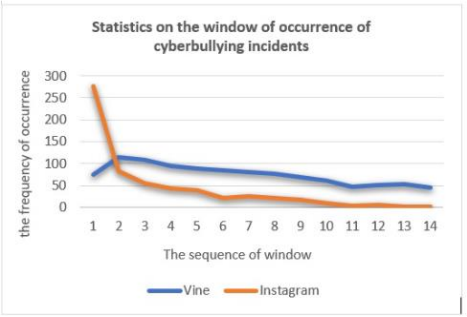
- What is Cyberbullying
- What is Cyberbullying detection
- Challenges and my studies
- **Learning like human annotators: Cyberbullying detection in lengthy social media sessions**
  - Motivation
  - Methods
  - Experiments
  - Results
  - **Why is the research useful?**

Why is the research useful?

- 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 cyberbullying detection competitive baselines.
- Calls for the collection and **finer-grained annotation** of cyberbullying datasets to enable research in this direction.







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

Thank you for your time