UNIVERSITI MALAYA

CYBERBULLYING DETECTION IN SOCIAL MEDIA

Supervisor: Associate Prof. Dr. Kasturi Dewi A/P Varathan
Student: Then Tsze Yen (S2194020)

feelings or actions re

Bullying /'

The actions, feeling
tasks combined in
cept or a theme, wit
mea or explaine

Introduction

RESEARCH BACKGROUND

Cyberbullying has become a prevalent issue, affecting individuals of all ages across various online platforms. The anonymity and ease of online communication have not only facilitated cyberbullying but also contributed to its harmful impact on mental health, with victims suffering from various mental, psychological and social issues, including depression, anxiety and even suicidal tendencies (Maurya et al., 2022).

Problem Statements

1.Challenges in extracting contextual features with higher performance due to lack of optimized hyperparameter

-Conventional techniques like TF-IDF and non-contextual word embeddings (GloVe, FastText, Word2Vec) face limitations in capturing context-aware features within text. Al-Hashedi et al. (2023) stated that transformer model can handle contextual information. As highlighted by Ajik et al. (2023), a significant number of existing studies have failed to optimize hyperparameters for the transformer models, leading to suboptimal performance and hindering the effectiveness in extraction of contextual features.

2. Lack of transparency and explainability in transformer models due to black-box nature

-While the detection of the offensive content includes hate speech (Bohra et al., 2018) and cyberbullying (Maity & Saha, 2021) are widely explored in the natural language processing (NLP) community, however in terms of the explainability remains underexplored.

Solutions for Problem Statements

Research Problem 1

Challenges in extracting contextual features with higher performance due to lack of optimized hyperparameter

Proposed Solution 1:

- Train transformer models due to its proficiency ability in capturing contextual information and compare with different hyperparameter settings by random search
- Find the optimized hyperparameters (which hyperparameters can achieve the best performance)

Research Problem 2

Lack of transparency and explainability in transformer models due to black-box nature

Proposed Solution 2:

- Apply Explainable Artificial Intelligence (XAI), Local Interpretable Model-Agnostic Explanations (LIME) to generate explanation for classification
- Identify the most influential words based on LIME weights through visualisation

Research Questions

What are the significant words that lead the model to classify a text as cyberbullying or not?

How to develop the cyberbullying detection model?

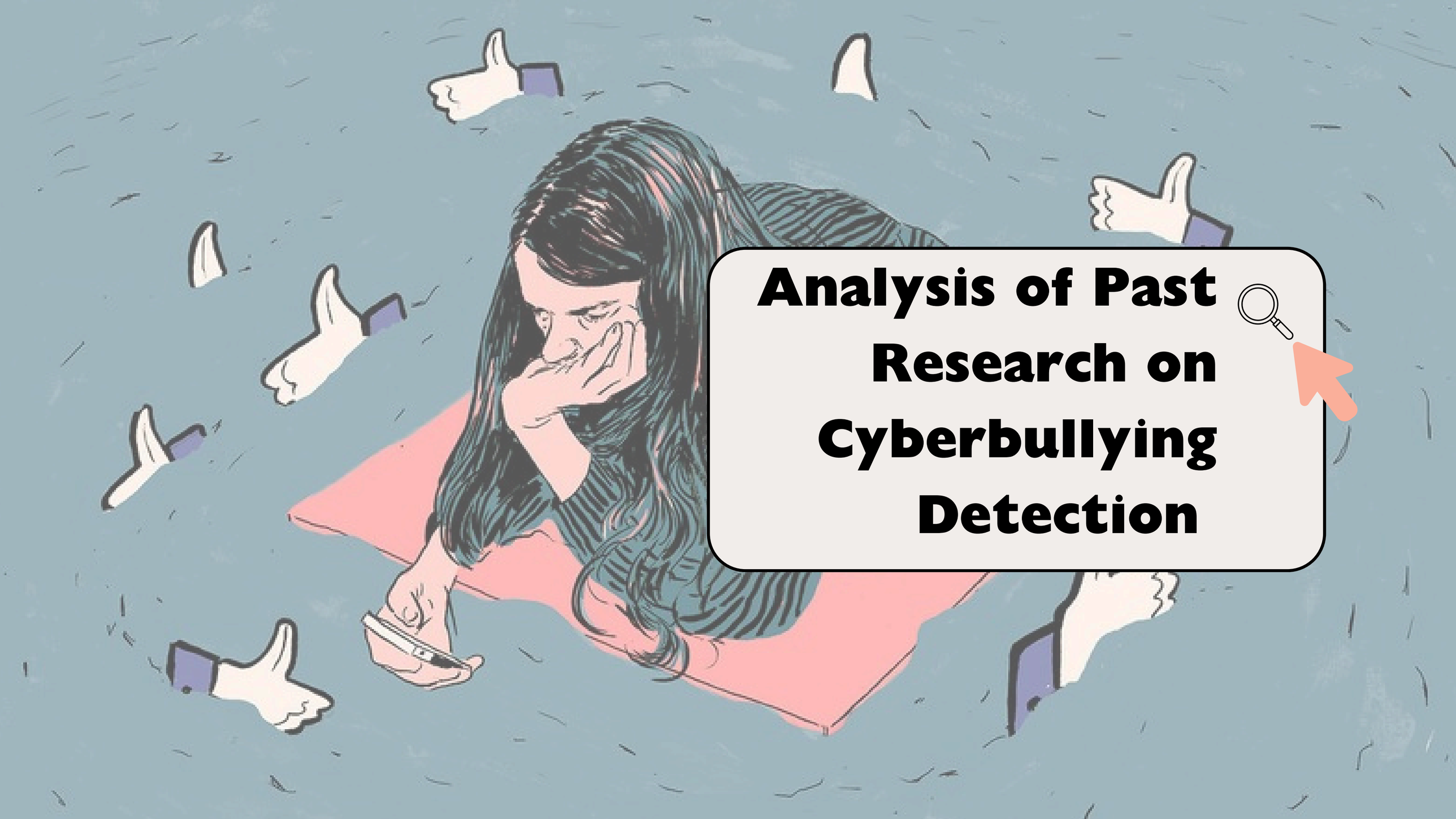
What is the performance of the cyberbullying detection model?

Research Objectives

To identify significant words contributing to cyberbullying.

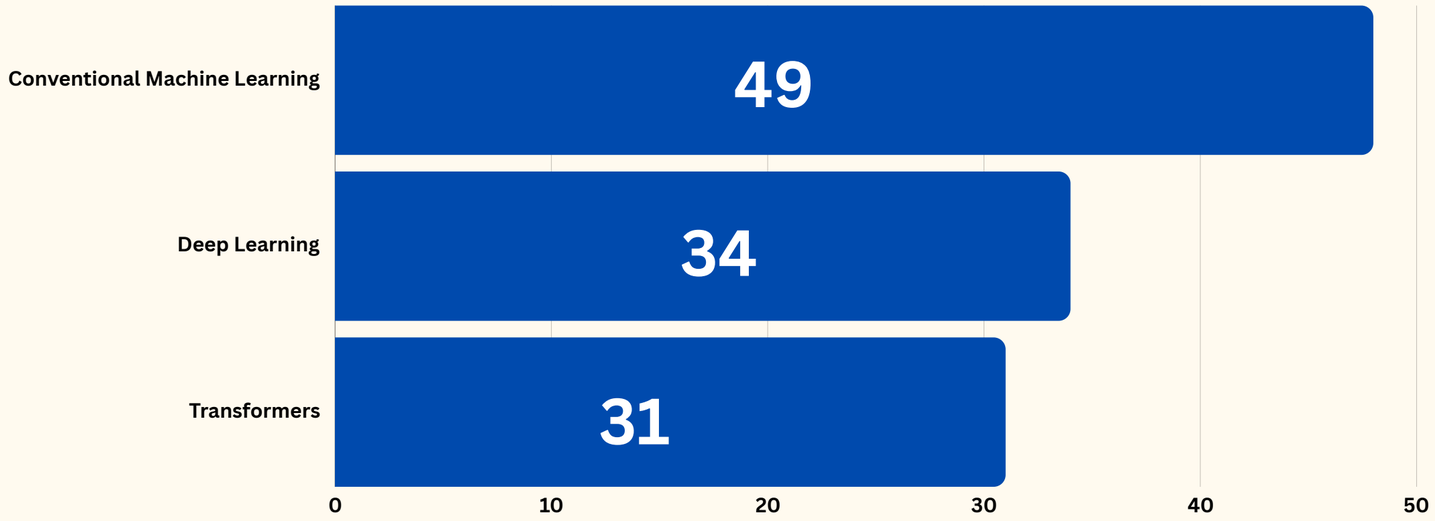
To develop a cyberbullying detection model.

To evaluate the cyberbullying detection model.

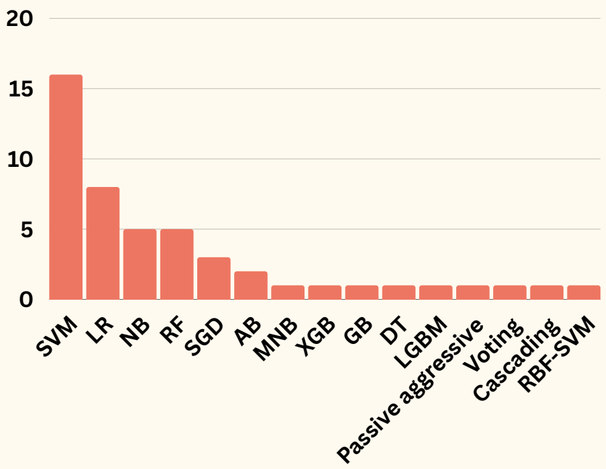


Analysis of Past Research on Cyberbullying Detection

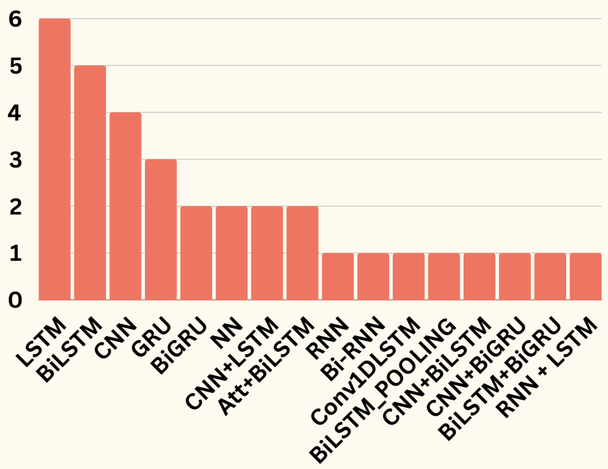
Techniques by Previous Research



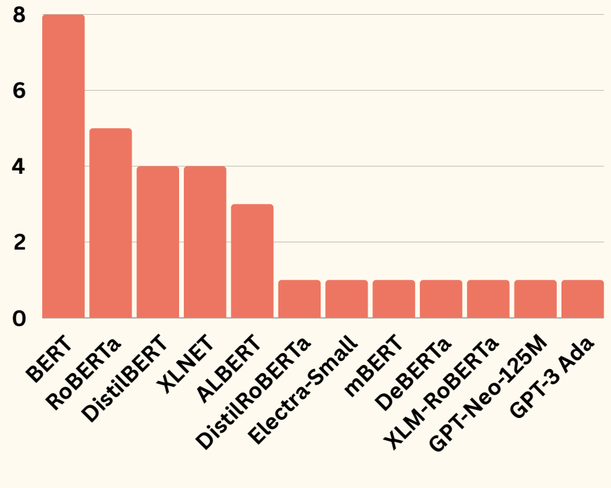
Conventional Machine Learning



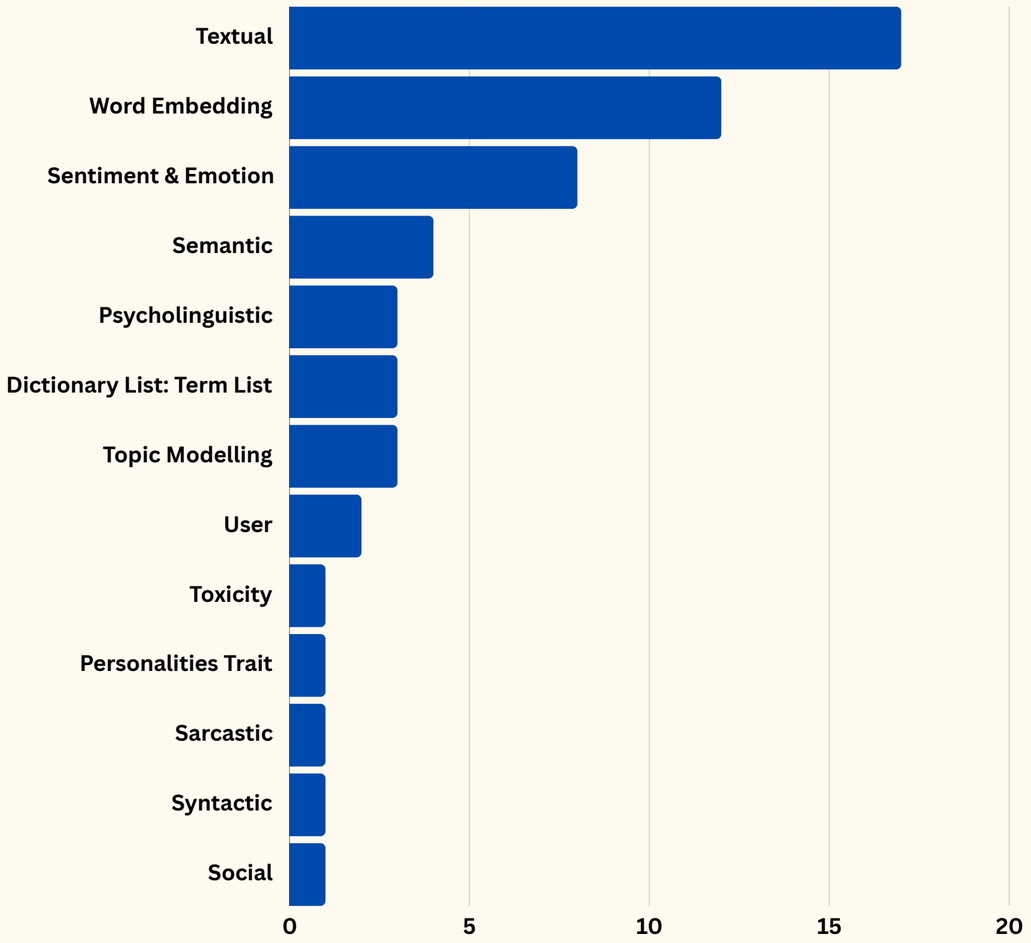
Deep Learning



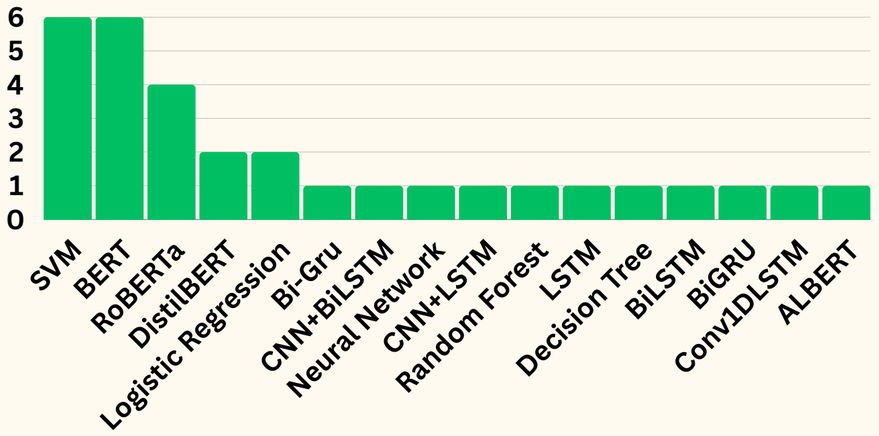
Transformer

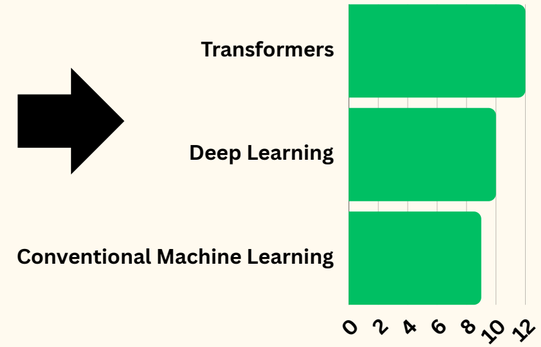


Features extracted by Previous Research



Best Models from Previous Research





Top models for other datasets

Research	Best Model Performance	Accuracy
Nithyashree et al. (2022)	BERT	98.00%
Raj et al. (2021)	Bi-Gru	96.98%
Raj et al. (2022)	CNN+BiLSTM	95.12%
Tripathy et al. (2020)	AlBERT	95.00%

Model Selection for Benchmarking

Top models for AMiCA dataset

Research	Text Preprocessing	Features	Models	Best Model Performance	Metrics
Van Hee et al. (2018)	1. Replace hyperlinks, @, spacings 2. Replace abbreviations 3. Tokenisation 4. Stemming	1.Textual Feature 2. Sentiment Feature 3. Dictionary List (Term List) 4. Topic Modelling	SVM	SVM	97.21% accuracy, 74.13% precision, 55.82% recall, 63.69% F-measure
Ali et al. (2021)	1. Remove URLs, numbers, spacings 2. Spelling corrections and replace abbreviations. 3. Tokenisation 4. Lemmatization	1. Textual Feature 2. Word Embedding 3. Dictionary List (Term List)	SVM	SVM	96.57% accuracy, 75.23% precision, 44.86% recall, 56.21% F-measure
Teng and Varathan (2023)	1. Replace emoji and emoticons to text 2. Remove punctuations, URLs, hashtags @, special characters, and spacings 3. Spelling corrections and replace abbreviations 4. Remove Stopwords 5. Lemmatization	1. Textual Feature 2. Sentiment & Emotion 4. Word Embedding 5. Psycholinguistic Feature 6. Dictionary List (Term List) 7. Toxicity Feature	Logistic Regression, Linear SVC	DistilBERT	97.41% accuracy, 73.89% precision, 71.00% recall, 72.42% F-measure
		NA	DistilBERT, DistilRoBERTs, Electra		


Methodology

Methodology

Data Collection

AMiCA

Text Pre-processing

1. Reduce elongated character, convert emoji and emoticon to text
2. Removal of punctuation, hash tags # and tags @, spacing, numeric characters
3. Lowercasing, spelling correction and replacing abbreviations
4. Tokenisation
5. Lemmatization

Objective 1

Analysis of Words in Cyberbullying Corpus

Word Cloud → Significant words

Explainable Artificial Intelligence (XAI): Local Interpretable Model-Agnostic Explanations (LIME)

Objective 2

Fine-Tuning Transformer Models

DistilBERT

AIBERT

RoBERTa Tiny

- Split train and test set in ratio of 90:10
- 5-fold Cross Validation
- Hold Out Testing

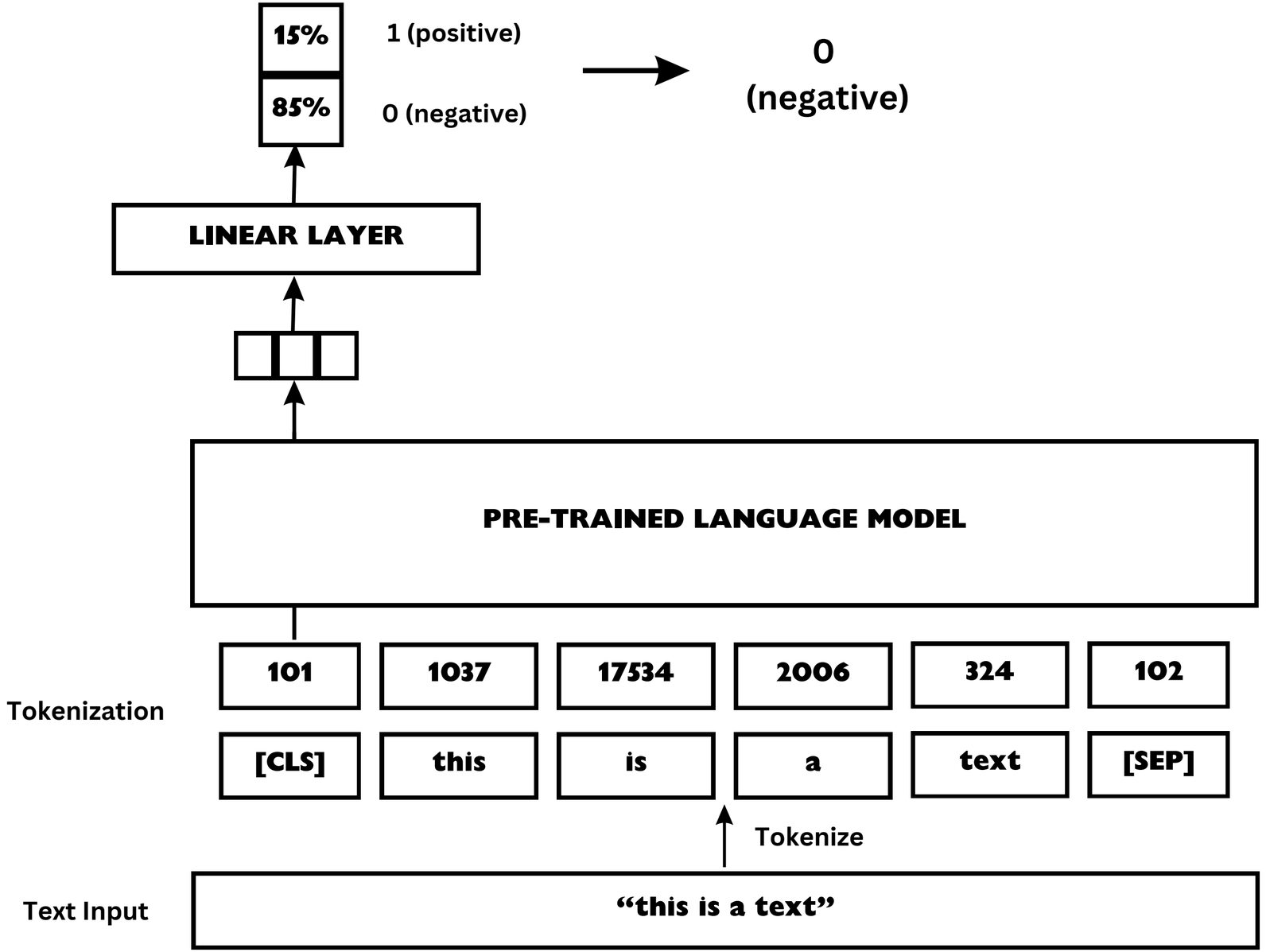
- Hyperparameter 1: batch size 8, learning rate 0.00005
- Hyperparameter 2: batch size 16, learning rate 0.00002

Objective 3

Evaluation

- Accuracy
- Precision
- Recall
- F1-score

Processing flow of Pretrained Language Models



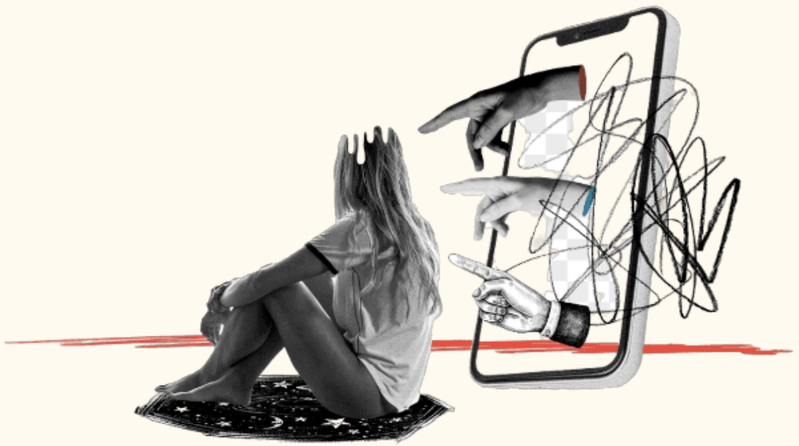
Hyperparameters applied in fine-tuning PLMs

Problem Statement 1

Hyperparameter	DistilBERT		AlBERT		RoBERTa Tiny	
Repository path	distilbert-base-uncased		albert-base-v2		haisongzhang/roberta-tiny-cased	
Batch size for training	8	16	8	16	8	16
Batch size for testing	8	16	8	16	8	16
Maximum sequence length	512		512		512	
Activation function	gelu		gelu-new		gelu	
Learning rate	0.00005	0.00002	0.00005	0.00002	0.00005	0.00002
Dropout probability	0.1		0.1		0.1	

Local Interpretable Model-Agnostic Explanation (LIME)

Problem Statement 2



What is XAI, LIME?

XAI collectively refers to techniques or methods, which help explain a given model's decision-making process.

LIME is a technique used to explain the predictions of any model.

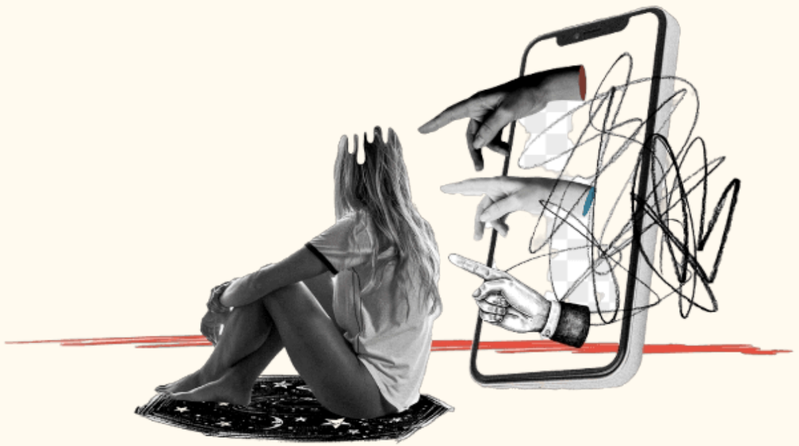
Model Agnosticism: LIME treats the model it explains as a 'black box', meaning it doesn't need to know the internal workings of the model. This allows LIME to work with any type of model, regardless of its complexity or structure.

Local Explanations: LIME provides explanations that are accurate and interpretable within the vicinity of a specific prediction. This means it focuses on understanding why a particular prediction was made for a given instance, rather than explaining the model's behavior across all possible instances.

Local Interpretable Model-Agnostic Explanation (LIME)

Problem Statement 2

How LIME works?



Text Input: Choose a specific text instance (e.g., a tweet or comment) to understand the model's prediction.

Generating Perturbations: LIME creates several modified versions of the selected text by altering words.

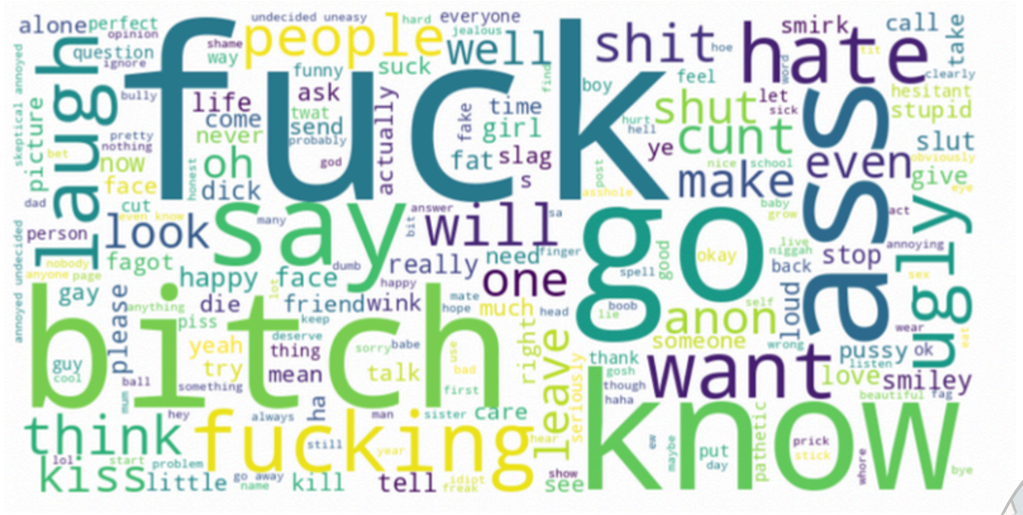
Getting Predictions for Perturbations: The cyberbullying detection model predicts the probability of cyberbullying for each perturbed text.

Fitting an Interpretable Model: LIME fits our best model to these instances, using the complex model's predictions as the target.

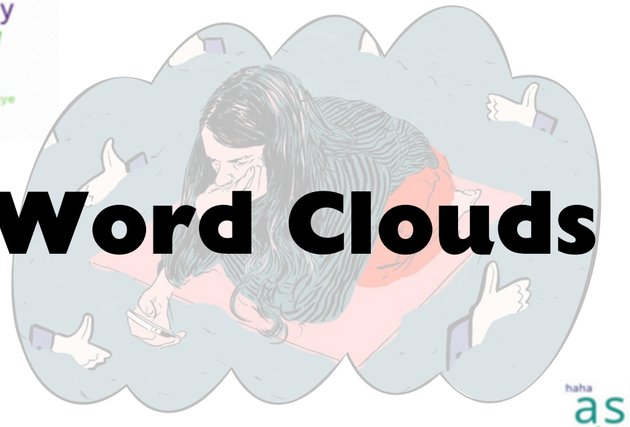
Extracting Important Words: The interpretable model identifies words or phrases that most influenced the prediction, such as abusive terms.

Presenting the Explanation: LIME presents the key words in a human-understandable way, clarifying why the model flagged the text as cyberbullying.

Results and Discussions



Word Cloud of Cyberbullying (Class 1)



Word Clouds

Word Cloud of Non-Cyberbullying (Class 0)

Performances evaluation metrics under Cyberbullying Class (Class 1)

		Cyberbullying Class (Class 1)											
		Batch size = 8, Learning rate = 0.00005						Batch size = 16, Learning rate = 0.00002					
		5-fold Cross Validation			Hold Out			5-fold Cross Validation			Hold Out		
Models	**Epoch**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
DistilBERT	**1**	65.53	53.20	57.48	71.69	64.50	**67.91**	85.56	67.81	74.67	95.31	86.80	90.86
	2	70.56	51.68	58.94	76.51	58.74	66.46	80.00	73.14	76.54	92.79	90.89	91.83
	3	73.92	53.16	61.54	73.71	58.36	65.15	83.46	77.86	**80.61**	93.35	88.66	90.94
	4	73.89	54.83	**62.87**	74.13	59.67	66.12	83.99	77.06	79.90	94.51	89.59	**91.98**
AlBERT	**1**	0.00	0.00	0.00	0.00	0.00	0.00	72.56	54.24	61.07	87.65	54.09	66.89
	2	0.00	0.00	0.00	0.00	0.00	0.00	73.96	55.20	61.99	76.08	75.09	75.58
	3	0.00	0.00	0.00	0.00	0.00	0.00	72.57	58.84	64.51	80.29	71.93	**75.88**
	4	0.00	0.00	0.00	0.00	0.00	0.00	71.84	59.37	**64.77**	75.80	75.09	75.44
RoBERTa Tiny	**1**	82.40	59.92	68.44	86.29	71.34	78.13	79.49	66.06	72.09	88.61	79.55	83.84
	2	78.32	63.51	69.95	84.15	76.95	**80.39**	77.02	70.03	73.39	86.15	84.39	**85.26**
	3	79.89	62.84	70.29	85.59	73.97	79.36	78.20	69.74	73.59	86.49	83.27	84.85
	4	80.09	64.40	**70.52**	87.87	72.68	79.55	76.78	71.08	**73.71**	87.23	82.53	84.81

Continue

Performances evaluation metrics under Non-Cyberbullying Class (Class 0)

		Non-cyberbullying Class (Class 0)											
		Batch size = 8, Learning rate = 0.00005						Batch size = 16, Learning rate = 0.00002					
		5-fold Cross Validation			Hold Out			5-fold Cross Validation			Hold Out		
Models	**Epoch**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
DistilBERT	**1**	97.87	98.46	98.16	98.22	98.72	98.47	98.29	99.25	98.29	99.34	99.78	99.56
	2	97.62	99.06	98.33	**97.95**	**99.09**	**98.52**	**98.73**	**99.03**	**98.88**	99.54	99.64	99.59
	3	97.72	99.09	98.40	97.93	98.95	98.44	98.76	99.19	98.97	99.43	99.68	99.56
	4	**97.75**	**99.10**	**98.42**	97.99	98.95	98.47	98.82	99.26	99.04	**99.48**	**99.74**	**99.61**
AlBERT	**1**	95.21	1.00	97.54	95.21	1.00	97.54	98.10	98.92	98.51	97.73	99.61	98.67
	2	95.21	1.00	97.54	95.21	1.00	97.54	**98.22**	**99.13**	**98.67**	98.75	98.81	98.78
	3	95.21	1.00	97.54	95.21	1.00	97.54	98.26	98.73	98.50	**98.59**	**99.11**	**98.85**
	4	95.21	1.00	97.54	95.21	1.00	97.54	98.34	98.71	98.53	98.75	98.79	98.77
RoBERTa Tiny	**1**	98.01	99.29	98.64	98.57	99.43	98.99	98.53	99.10	98.81	98.98	99.49	99.23
	2	98.18	99.12	98.65	98.84	99.27	99.06	98.70	99.19	98.94	**99.21**	**99.32**	**99.27**
	3	**98.18**	**99.31**	**98.74**	**98.70**	99.37	99.03	98.82	99.25	99.03	99.16	99.35	99.25
	4	**98.22**	**99.27**	**98.74**	**98.63**	**99.49**	**99.06**	**98.95**	**99.31**	**99.13**	99.12	99.39	99.26

Training time of fine-tuning Pretrained Language Models each epoch

Models	Time (mins)	
	Batch size = 8, Learning rate = 0.00005	Batch size = 16, Learning rate = 0.00002
DistilBERT	22.0	20.5
AlBERT	46.0	41.5
RoBERTa Tiny	8.25	7.6

Discussion

1.Result of model can be different in different environment.

Same hyperparameter used by Teng and Varathan (2023) applied on DistilBERT, with default hyperparameter setting which batch size for training and testing is 8, learning rate of 0.00005, but under different environment, Python version 3.10, and GPU of A100 achieved worse performance with dropping in 4.51% of F1-score.

2. Importance of Hyperparameter Tuning

Default hyperparameters were not suitable for AlBERT, leading to undefined metrics.

A larger batch size (16) combined with a smaller learning rate (0.00002) outperformed the configuration with a batch size of 8 and learning rate of 0.00005.

Performance evaluation metrics of AlBERT using default hyperparameter

Models	Cyberbullying			Non-cyberbullying		
	P	R	F	P	R	F
Default hyperparameter	0.00	0.00	0.00	95.21	1.00	97.54

Discussion

3.Effectiveness of Small-Sized Transformer Models

The research found that smaller transformer models can be effective while requiring fewer computational resources.

- DistilBERT showed the best overall performance among the tested models.
- RoBERTa Tiny had the shortest training time (7.6 minutes per epoch, 30 minutes total for one fold).
- Model Selection Based on Resources
  - For limited resources (RAM, GPU, time): RoBERTa Tiny is recommended.
  - For optimal performance: DistilBERT is recommended.

4.Integration of XAI: LIME to enhance transformer model transparency and interpretability

- Example: "you" weighted 0.76 for cyberbullying classification
- Such behavior would suggest that the model learned to associate the mere presence of second-person pronouns, such as "you", with a higher likelihood of cyberbullying.
- LIME which provides justification allows the users to be more understanding why the decision is made that decide a text to be cyberbullying and non-cyberbullying.

Example outputs of LIME

(i)

Enter Text to Analyze

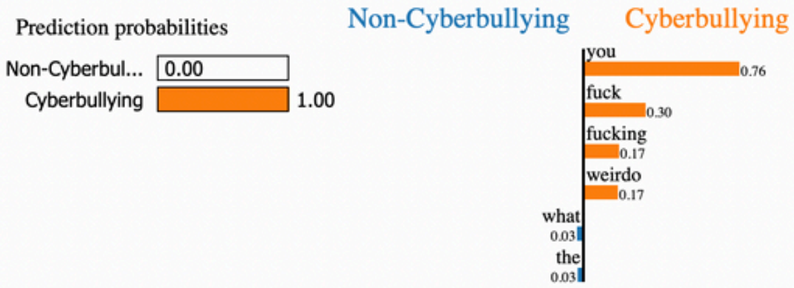
What the fuck!?! No! U fucking weirdo

Analyze

⚠️ Warning!! Our model says this is a Cyberbullying Post!

Explanation

Lime Explanation



Text with highlighted words

what the fuck no you fucking weirdo

(ii)

Enter Text to Analyze

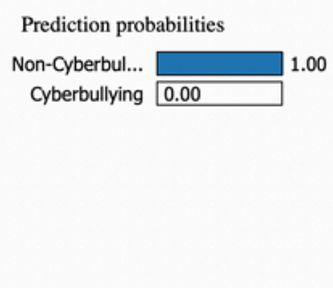
What the fuck!?! No! fucking weirdo

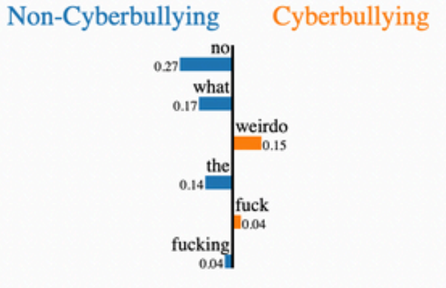
Analyze

✅ No worry! Our model says this is a Non-cyberbullying Post!

Explanation

Lime Explanation





Text with highlighted words

what the fuck no fucking weirdo

Comparison with benchmarking

Research	Model	Fold	Cross-validation				Hold Out			
			A	P	R	F	A	P	R	F
Van Hee et al. (2018)	SVM	10	96.97	73.32	57.19	64.26	97.21	74.13	55.82	63.69
Ali et al. (2021)	SVM	NA	NA	NA	NA	NA	96.57	75.23	44.86	56.21
Teng and Varathan (2023)	DistilBERT	5	97.22	75.60	61.71	67.90	97.41	73.89	71.00	72.42
This research	DistilBERT	5	**98.07**	**83.99**	**77.06**	**79.90**	**99.25**	**94.51**	**89.59**	**91.98**
	RoBERTa Tiny	5	**98.37**	**76.78**	**71.08**	**73.71**	**98.60**	**86.15**	**84.39**	**85.26**
	AlBERT	5	**96.95**	72.57	58.84	64.51	**97.81**	**80.29**	**71.93**	**75.88**

Conclusion

Conclusion

Objective 1:

- Word cloud visualizations identify prevalent cyberbullying language
- LIME provides explanations for output by showing the significant words that is influential for the result.

Objective 2:

- Fine-tuned for cyberbullying detection models with different hyperparameters.
- Batch size 16, learning rate 0.00002 has caused transformer models to have better performance.

Objective 3:

- DistilBERT outperformed:
- F1-score 91.98% (hold-out testing) on AMiCA dataset
- New benchmarks established
- If resource limited, RoBERTa Tiny can be used

Limitations:

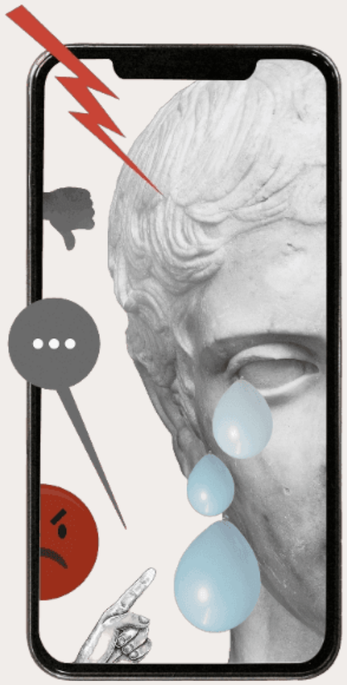
- English language data only
- Text-based inputs only (no images, videos)

Future Directions:

- Explore other XAI algorithms like SHAP
- Multimodal learning (text, images, videos)
- Cross-lingual and multilingual models

Test Our Application!

REFERENCES



Agrawal, S., & Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. arXiv (Cornell University). https://arxiv.org/pdf/1801.06482

Ali, W. N. H. W., Mohd, M., & Fauzi, F. (2021). Cyberbullying Predictive Model: Implementation of Machine Learning Approach. 2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP). https://doi.org/10.1109/camp51653.2021.9497932

Ali, W. N. H. W., Mohd, M., Fauzi, F., Shirai, K., & Noor, M. J. M. (2021). Implementation Of Hyperparameter Optimisation And Over-Sampling In Detecting Cyberbullying Using Machine Learning Approach. Malaysian Journal of Computer Science, 78–100. https://doi.org/10.22452/mjcs.sp2021no2.6

Al-Hashedi, M., Soon, L., Goh, H., Lim, A. H. L., & Siew, E. (2023). Cyberbullying detection based on emotion. IEEE Access, 11, 53907–53918. https://doi.org/10.1109/access.2023.3280556

Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. Computers & Security, 90, 101710. https://doi.org/10.1016/j.cose.2019.101710

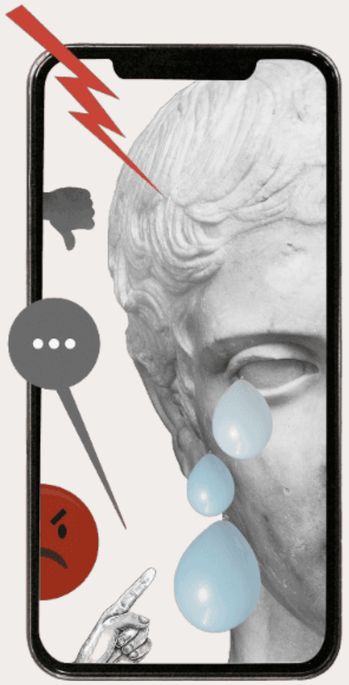
Dadvar, M., Trieschnigg, D., Ordelman, R., & De Jong, F. (2013). Improving Cyberbullying Detection with User Context. In Lecture Notes in Computer Science (pp. 693–696). https://doi.org/10.1007/978-3-642-36973-5_62
Dalvi, R. R., Chavan, S. B., & Halbe, A. (2020). Detecting A Twitter Cyberbullying Using Machine Learning. Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020). https://doi.org/10.1109/iciccs48265.2020.9120893

Desai, A., Kalaskar, S., Kumbhar, O., & Dhumal, R. (2021). Cyber Bullying Detection on Social Media using Machine Learning. ITM Web of Conferences, 40, 03038. https://doi.org/10.1051/itmconf/20214003038

Goldfeder, B., & Griva, I. (2023). Explaining Cyberbullying Trait Detection Through High Accuracy Transformer Ensemble. 2023 IEEE Conference on Artificial Intelligence (CAI). https://doi.org/10.1109/cai54212.2023.00116

Hani, J., Mohamed, N., Mostafa, A. E., Emad, Z., Amer, E., & Mohammed, A. (2019). Social Media Cyberbullying Detection using Machine Learning. International Journal of Advanced Computer Science and Applications, 10(5). https://doi.org/10.14569/ijacsa.2019.0100587

REFERENCES



Islam, M., Uddin, M. A., Islam, L., Akter, A., Sharmin, S., & Acharjee, U. K. (2020). Cyberbullying detection on social networks using machine learning approaches. 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). https://doi.org/10.1109/csde50874.2020.9411601

Jacobs, G., Van Hee, C., & Hoste, V. (2020). Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text? Natural Language Engineering, 28(2), 141–166. https://doi.org/10.1017/s135132492000056x

Jadhav, A., Punekar, R., Sayyed, T., & Vora, D. (2021). Cyberbullying Detection In Social Media Text. Vidyabharati International Interdisciplinary Research Journal, 1527–1533.

Mathur, S. D., Isarka, S., Dharmasivam, B., & Jaidhar, C. D. (2023). Analysis of Tweets for Cyberbullying Detection. In 2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC). https://doi.org/10.1109/icsccc58608.2023.10176416

Maurya, C., Muhammad, T., Dhillon, P., & Maurya, P. (2022). The effects of cyberbullying victimization on depression and suicidal ideation among adolescents and young adults: a three year cohort study from India. BMC Psychiatry, 22(1). https://doi.org/10.1186/s12888-022-04238-x

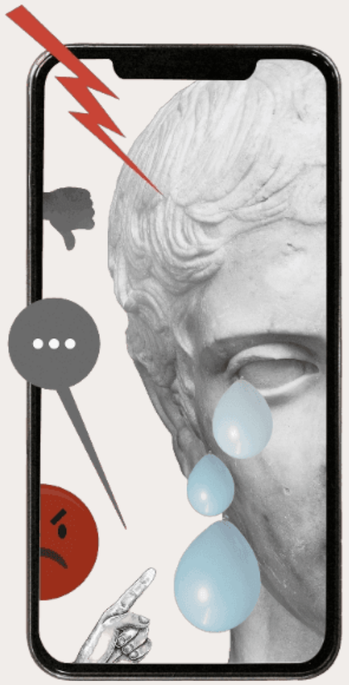
Muneer, A., & Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on Twitter. Future Internet, 12(11), 187. https://doi.org/10.3390/fi12110187

Nithyashree, V. Hiremath, B. N. , Vanishree, L. , Duvvuri, A. , Madival, D. A. and Vidyashree, G. (2022). Identification of Toxicity in MultimediaMessages for Controlling Cyberbullying on Social Media by Natural Language Processing. 2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), Shivamogga, India, 2022, pp. 12-18, doi: 10.1109/DISCOVER55800.2022.9974631.

Obaid, M. H., Guirguis, S. K., & Mesbah, S. (2023). Cyberbullying Detection and Severity Determination Model. IEEE Access, 11, 97391–97399. https://doi.org/10.1109/access.2023.3313113

Ogunleye, B., & Dharmaraj, B. (2023). The use of a large language model for cyberbullying detection. Analytics, 2(3), 694–707. https://doi.org/10.3390/analytics2030038

REFERENCES



Ottoson, D. (2023). Detection on social platforms using Large Language Models. https://www.diva-portal.org/smash/get/diva2:1786271/FULLTEXT01.pdf

Paul, S., & Saha, S. (2020). CyberBERT: BERT for cyberbullying identification. Multimedia Systems, 28(6), 1897–1904. https://doi.org/10.1007/s00530-020-00710-4

Pawar, R., & Raje, R. R. (2019). Multilingual Cyberbullying Detection System. IEEE Explore. https://doi.org/10.1109/eit.2019.8833846

Perera, A., & Fernando, P. (2021). Accurate cyberbullying detection and prevention on social media. Procedia Computer Science, 181, 605–611. https://doi.org/10.1016/j.procs.2021.01.207

Raj, C., Agarwal, A., Bharathy, G., Narayan, B., & Prasad, M. (2021). Cyberbullying Detection: Hybrid models based on machine learning and natural language processing techniques. Electronics, 10(22), 2810. https://doi.org/10.3390/electronics10222810

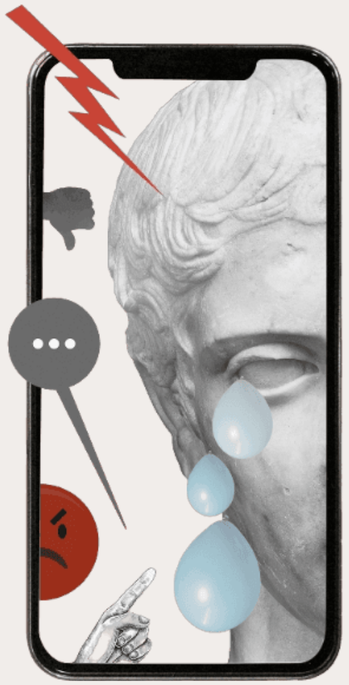
Raj, M., Singh, S., Solanki, K., & Selvanambi, R. (2022). An application to detect cyberbullying using machine learning and deep learning techniques. SN Computer Science, 3(5). https://doi.org/10.1007/s42979-022-01308-5

Rathnayake, G., Atapattu, T., Herath, M., Zhang, G., & Falkner, K. (2020). Enhancing the Identification of Cyberbullying through Participant Roles. Proceedings of the Fourth Workshop on Online Abuse and Harms. https://doi.org/10.18653/v1/2020.alw-1.11

Saker, J., Sultana, S., Wilson, S. R., & Bosu, A. (2023). ToxiSpanSE: An Explainable Toxicity Detection in Code Review Comments. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.03386

Teng, T., & Varathan, K. D. (2023). Cyberbullying Detection in Social Networks: A comparison between machine learning and transfer learning approaches. IEEE Access, 11, 55533–55560. Knowing how to interact with others in an online environment, including appropriate use of tone and lingo.

REFERENCES



Tripathy, J. K., Chakkaravarthy, S. S., Satapathy, S. C., Sahoo, M., & Vaidehi, V. (2020). ALBERT-based fine-tuning model for cyberbullying analysis. Multimedia Systems, 28(6), 1941–1949.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. PLOS ONE, 13(10), e0203794. https://doi.org/10.1371/journal.pone.0203794

Vijayakumar, & Adolf, H. P. D. (2021). Multimodal Cyberbullying Detection using Hybrid Deep Learning Algorithms. International Journal of Applied Engineering Research, 16(7), 568.
Ottoson, D. (2023). Detection on social platforms using Large Language Models. https://www.diva-portal.org/smash/get/diva2:1786271/FULLTEXT01.pdf


THANK
YOU!