Factor Detection Task of Cyberbullying Using the Deep Learning Model

Yu-Hsuan Wu[1], Sheng-Wei Huang[1], Wei-Yi Chung[2], Chen-Chia Yu[3], Jheng-Long Wu[1]

[1]Dept. of Data Science,Soochow University,
[2]School of Big Data Managemen,Soochow University,
[3]Dept. of Language Science,The University of Edinburgh

IEEE BigData 2022 Workshops
6th International Workshop on
Big Data Analytics for Cyber Intelligence and Defense

SCU NLP Lab

Introduction

- The large amount of data posted on social media offers much flexibility and opportunities to be discovered in exploring various social issues.
- Cyberbullying is highly lethal to individuals and is difficult to effectively punish or prevent due to its vague definition.
- In this study, textual cyberbullying is the main research object. The main purpose of this study is to provide a set of analysis methods for cyberbullying.

Introduction

Main contributions:

1. Multiple cyberbullying factor corpus can be created by the study because there are no benchmark dataset about the cyberbullying factors to a classification task.
2. The factor detection of cyberbullying classification can be solved by the powerful learning models such as the BERT classifier.
3. The proposed cyberbullying detection model can train from the proposed corpus, and can apply to the big data analysis on the Internet for cyberbullying.

Related work – Definitions and Factors of Cyberbullying

- Bullying can be defined as the bully's intention to cause harm to the victim, repeated bullying behaviors, and power imbalance between the bully and the victim. (Olweus 1994)
- However, cyberbullying is more specific than traditional bullying, and the repetition and power imbalance are more difficult to judge, due to the anonymity mechanism of online discourse.
- Menin et al (2021) made important findings on the definition of bullying based on adolescents' perceptions. They found that "intent" and "perceived harm" are important factors that constitute cyberbullying.
- Define cyberbullying as the commenters have intent to cause harm via words or other media online, and the victim can feel themselves got harm by the comments.

Related work – Extracting Cyberbullying Factor from Definition

- The concept of "intent to harm" can be categorized as aggressive or offensive language and other directly harmful language such as insults, intimidate, harassment. (Rosenthal et al 2021 ; Kalraa et al 2021)
- Zhong et al (2022) found that the language used in cyberbullying contains direct and indirect hurtful sentences. Based on the example of indirectly harmful sentences in their study, this type of sentence can be specifically defined as sarcasm.
- Arousal refers to the degree of intensity or aggressiveness of the words used and whether the commenter has strong emotions.
- Valence is based on the emotion of reading the sentences while being the victim's perspective.

Related work – Existing Corpus for Bullying Detection

- Most of existing corpus for bullying detection are only classified articles or sentences as bullying/non-bullying or only add toxic labels. (Ask.fm, Formspring, Myspace, Twitter)
- Some studies have labeled some factors of bullying such as inflammatory words, offensive language, toxicity, etc., but not enough to be comprehensive. (Maskat et al 2020; Zhang et al 2020; Tang et al 2020)

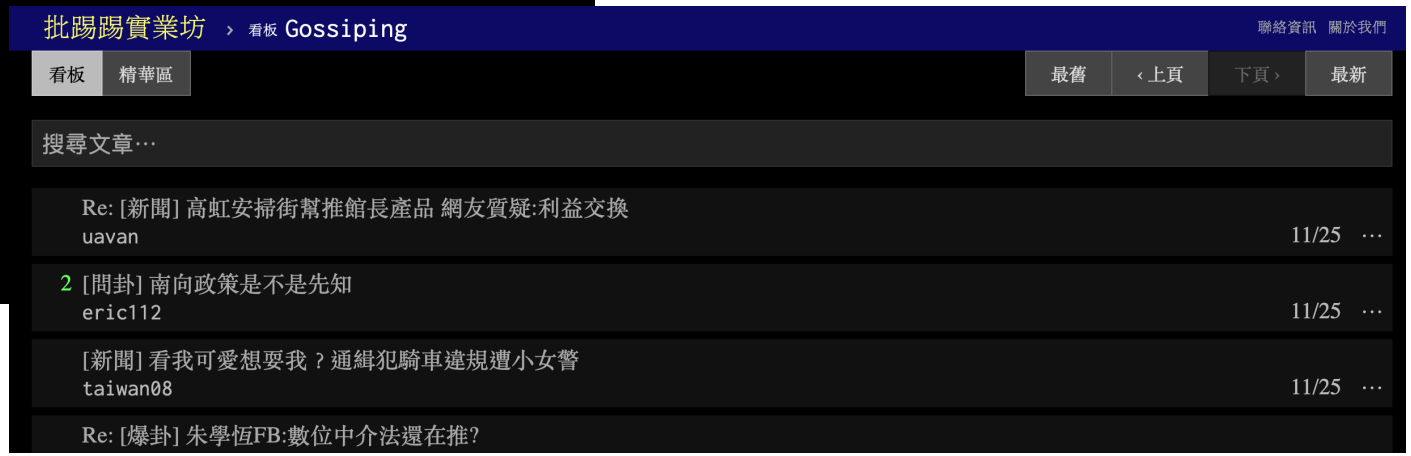
Related work – Bullying Detection

- Classification method such as SVM, Randomforest, etc., are used to detect bullying articles or sentences. (Kumar et al 2019; Kanan et al 2020; Abarna et al 2022)
- However recent studies have shown that deep neural network-based approaches are more effective than traditional machine learning techniques on detecting bullying. (Emon et al 2022; Ahmed et al 2022; Raj et al 2022)
- Neural network methods have advantages in classifying a wide range of cyberbullying features. Therefore, this study will use a neural network approach in combination with a pre-trained model to build the model.

Model – Data

- Social media platform：PTT
- Post：125
- Comment：5926
- Time：Board opening date – January 2022

← PTT board list



PTT post list →

Model – Data

Author
Post title
Post time

Post content

Comment content

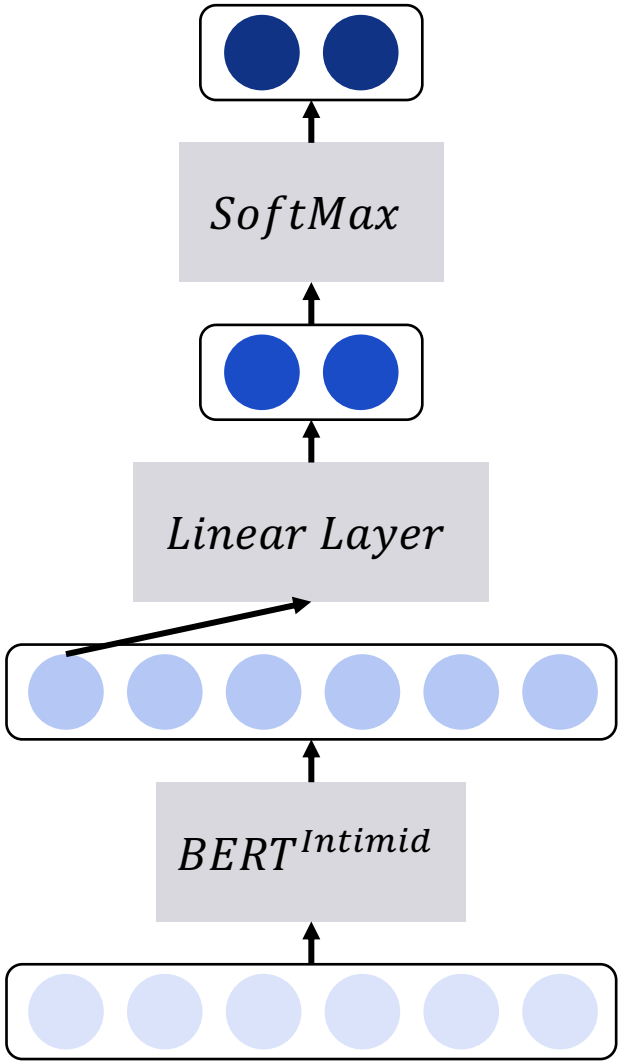
Board

Model – Data Annotation

Factors	Labeled as	Define
Intimidation	intimidation/non-intimidation	The use of words to intimidate the main character of the article.
Harassment	harassment/non-harassment	A variety of offensive behavior that is repetitive and unwilling.
Insults	insult/non-insult	Words that contain insulting or demeaning words.
Sarcasm	sarcasm/non-sarcasm	Sarcasm intended to expose the contradiction or shortcomings of the article's main character.
Arousal	arousal/non-arousal	The commenter has clear emotional fluctuation from the sentence.
Valence	positive, neutral, and negative	The emotion that will result from reading a statement from the perspective of the person being bullied.

Model – Classifier

- The input of model is a comment.
- Use BERT to extract hidden features vector of input comment.
- Using the first hidden feature vector to estimate the classification probability of cyberbullying factor though Linear Layer and SoftMax.



Experiment – Annotation Consistency

- Each comment were annotated by three annotators. Since each annotator's judgment of the sentences could be biased, the gold standard was determined by the results of the three annotators' votes.
- Cronbach's alpha was used to evaluate the quality and consistency of the annotations. Each score has been maintained at above 0.5.

TABLE I. DATA DISTRIBUTION ON THREE ANNOTATORS

Factor	Yes	No
Intimidation	48	5878
Harassment	2175	3751
Insult	1603	4323
Sarcasm	3641	2285
Arousal	1618	4308

	Positive	Neutral	Negative
Valence	37	2598	2349

TABLE II. ANNOTATION CONSISTENCY ON THREE ANNOTATORS

Factor	Cronbach's Alpha
Intimidation	0.56
Harassment	0.65
Insult	0.69
Sarcasm	0.61
Arousal	0.53
Valence	0.63

Experiment – model

The classifiers compared in the experiments:

- BERT_α: The pre-trained model distilbert-base-multilingual-cased is used.
- BERT_β: The pre-trained model bert-base-multilingual-uncased is used in this BERT model.
- RFC: Considered as a baseline. The TF-IDF method is used to obtain the feature vectors and the random forest classification algorithm is used.
- SVM: Considered as the baseline. The TF-IDF method is used to obtain the feature vectors.

Experiment –
All models for experiments

- The results of BERT_β is better than another model. The possible reason for this result is that the task of learning whether a comment sentence has cyberbullying factors are relatively complex and difficult, so more hidden features need to be learned to obtain better results.

TABLE III. MACRO AVERAGE PERFORMANCE COMPARISON OF MODELS

	Arousal			*Harassment*			*Insult*			*Intimidate*			*Sarcasm*			*Valence*		
	P	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***
BERT_α	0.60	0.60	0.60	0.63	**0.63**	**0.63**	0.66	0.66	0.66	0.70	0.70	0.69	0.58	0.58	0.58	0.45	0.42	0.43
BERT_β	0.61	**0.61**	**0.61**	0.65	0.62	0.63	0.66	**0.66**	**0.66**	**0.80**	**0.78**	**0.77**	**0.59**	**0.59**	**0.59**	**0.49**	**0.46**	**0.47**
SVM	0.61	0.54	0.40	**0.68**	0.52	0.43	**0.79**	0.52	0.47	0.60	0.51	0.52	0.71	0.51	0.44	0.43	0.26	0.20
RFC	**0.62**	0.54	0.38	0.67	0.52	0.43	0.77	0.52	0.46	0.60	0.51	0.52	0.67	0.51	0.44	0.41	0.26	0.20

Case study

- Using big data analysis of two cases to describes the cyberbullying pattern of PTT.
- The time period was from the beginning of the case until January 2022.
- A total of 224 posts and 39,541 comments related to two events on PTT

TABLE IV. DATA DISTRIBUTION ON EACH EVENT

Event	Post	Comment
Johny Depp	17	2,499
Alisasa	207	37,042

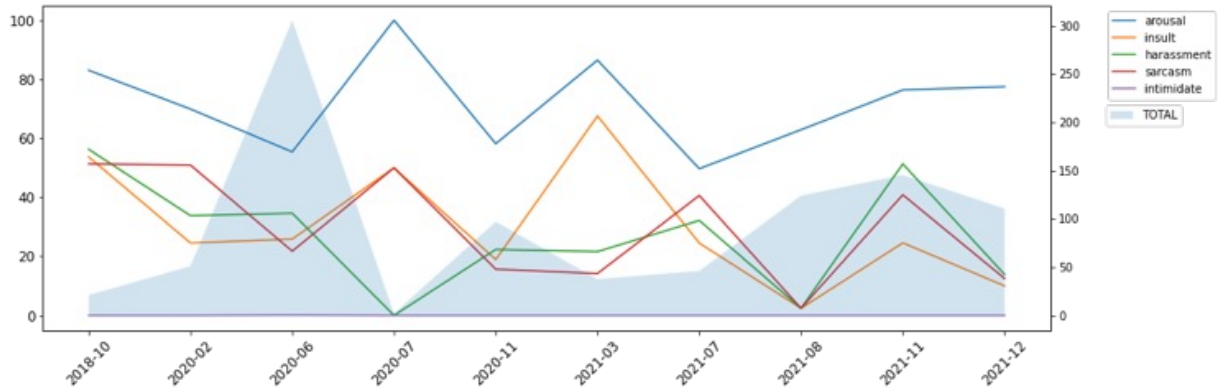
Two kind of graph to analyze the the cyberbullying factor:

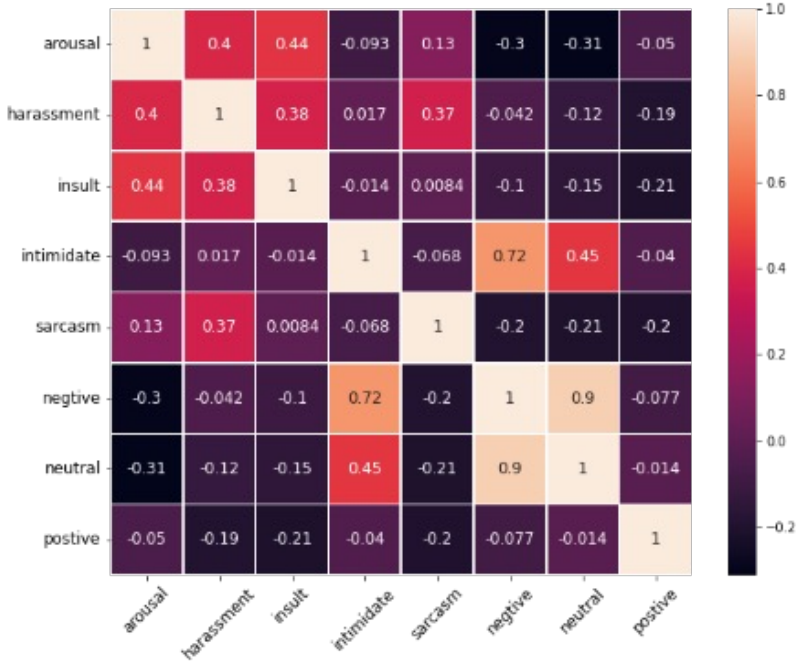
1. Line Graph: with the X-axis being the time and the Y-axis being the percentage of comments that match a particular cyberbullying factor to the total number of comments for that time period.
2. Heatmap: show the relevance of each factor in a specific case.

Case study – (Celebrity)
The Johnny Depp and Amber Heard Event

SCU NLP Lab

- Time: October 2018 to December 2021.
- Each month has a high percentage of comments with arousal and most of the comment are with emotions.
- A similar trend for insults, sarcasm and harassment in the case, with a weak to moderate correlation between these three factors.

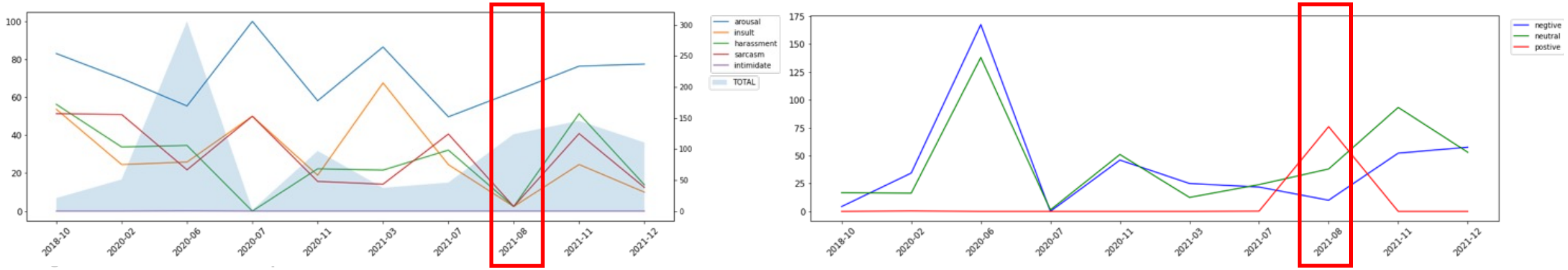




Case study – (Celebrity)
The Johnny Depp and Amber Heard Event

SCU NLP Lab

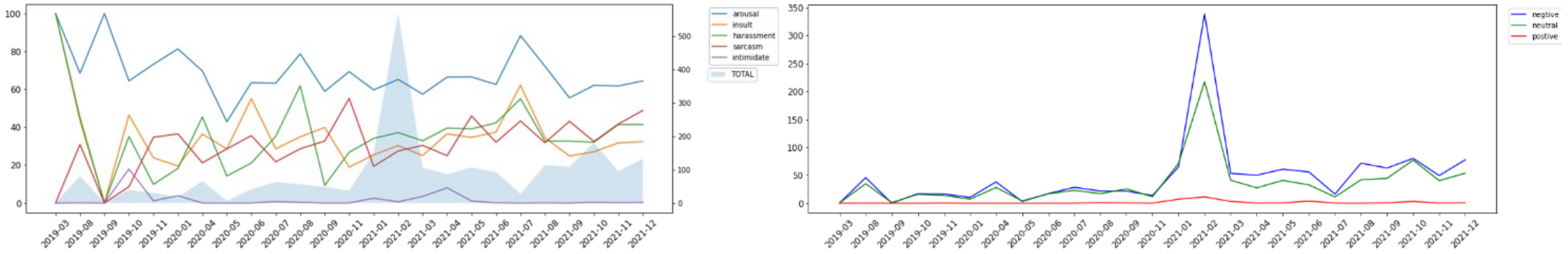
- The condemnation of Johnny in the comments at the outbreak of the case was one-sided due to Amber's accusations against him.
- In August 2021, more neutral tone of the discussion in the forum later in the case, when Johnny fought back against Amber's false allegations.
- Valence reflects the victim's sentiment when reading the comments, and a significant increase in the total number of positive sentiments can be observed, which reflects the beginning of solidarity with Johnny.



Case study – (Internet Celebrity) Alisasa Event

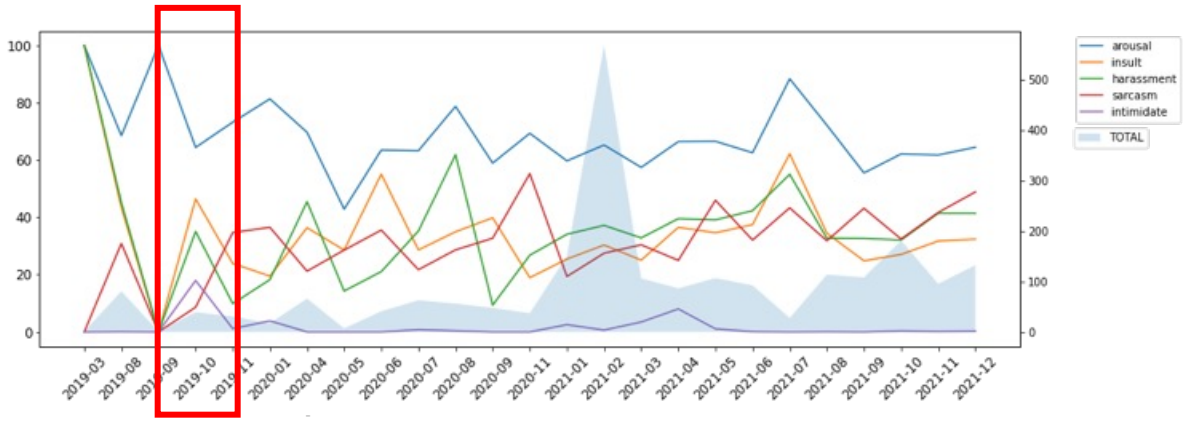
SCU NLP Lab

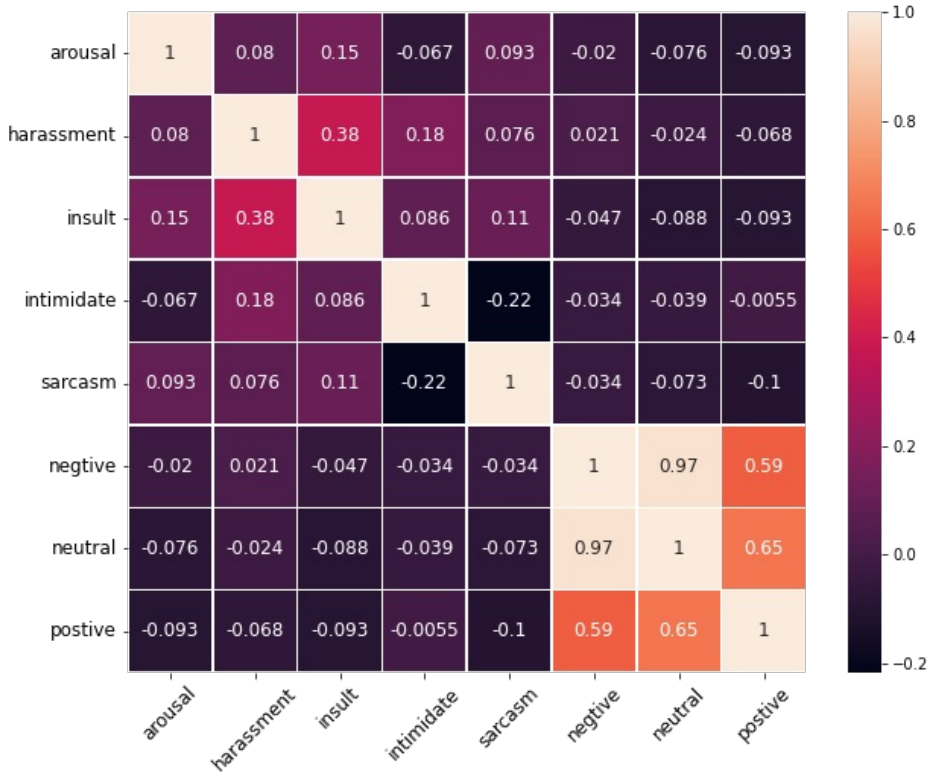
- Time: 2019 to the end of 2021.
- Comments discussing her controversial statements had significant mood waves, which is reflected in the high percentage of arousal.
- Valence reflects the fact that most of the comments made her feel negative emotions after reading them, which means that most of the comments directed at her controversial statements were offensive.



Case study – (Internet Celebrity) Alisasa Event

- The two factors of harassment and insult show moderate correlations, while sarcasm is less correlated with them, but the trends are partially similar.
- While there are few comments that match intimidation, they are often accompanied by explicit harassment and insults, as seen in the October 2019 comments.





Discussion

- Based on the case studies, the proposed cyberbullying detection model constructed in this research can detect cyberbullying patterns in real cases.
- The weak prediction performance of arousal and sarcasm by the proposed model reflected the useful hidden features which are difficult to learn.
- The most difficult aspect to the model was the valence factor. Since the definition of valence in this study was judged from the perspective of the victim.
- Most of the valence factors in the dataset were negative sentences, which may have made the model difficult for the model to judge positive sentences.

Conclusion

- This study successfully created the Chinese cyberbullying factors corpus, built the BERT classifier, and analyzed the big data of two real cyberbullying cases.
- The predictions of the model can be consistent with the temporal development.
- Victims of cyberbullying comments will be considered, as well as the role each responder takes in cyberbullying.

Thank you for listening.

Thank you for listening.