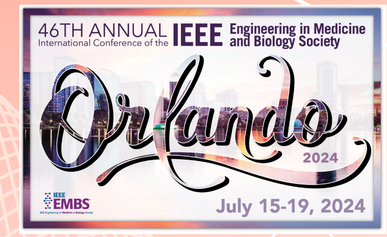
Data Quality Matters: Suicide Intention Detection on Social Media Posts Using RoBERTa-CNN

1st Emily Lin, 1st Jian Sun, 1st Hsingyu Chen, and 2nd Mohammad H. Mahoor

Ritchie School of Engineering and Computer Science, University of Denver



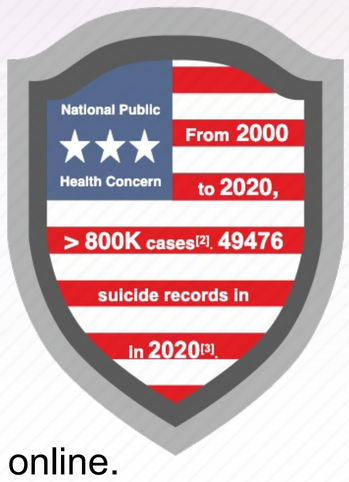
Poster No. B4P-24

OBJECTIVES

Detect suicide intention at the early stage by using RoBERTa-CNN to analyze social media posts.

INTRODUCTION

Motivation: Suicide is public concern, intention detection is important. EX: Worldwide **1.53 million** cases in 2020[1].



Why RoBERTa backbone?

- It extracts semantics and contextual features well.
- It succeeds in various NLP tasks.
- It performs sentence analysis.
- It is faster and better than other versions of BERT.

Why analyzing social media posts?

- Today, everybody express themselves freely online.

Why RoBERTa-CNN?

- It improves the performance
- It overcomes the constraints of RoBERTa.

DATASET

Data source: Reddit SuicideWatch and Depression channel.

We use **35,270** suicides and **74,770** non-suicides.

We clean data by human labors and **OpenAI API**, which is powerfully and efficiently.

Data Processing:

- Removing all noise symbols;
- Removing all non-English characters;
- Deleting all sentence fragments.

Table 1. Data Samples of the Suicide and Depression Detection Dataset

Suicide	1.	I feel worthless and useless to everyone. I'm a burden what's the point of being here?
	2.	I feel like overdosing again. I can't live like this.
	3.	I feel bad. I just wanna stop existing.
Non-suicide	1.	Would you ever trust your school therapist? I myself wouldn't tbh, there doesn't seem to be a lot of confidentiality.
	2.	Is it just me or is apple cider really good? it is one of the best drinks I think.
	3.	Why is everyone talking about us? I'm genuinely confused what exactly did we do?

METHODOLOGY

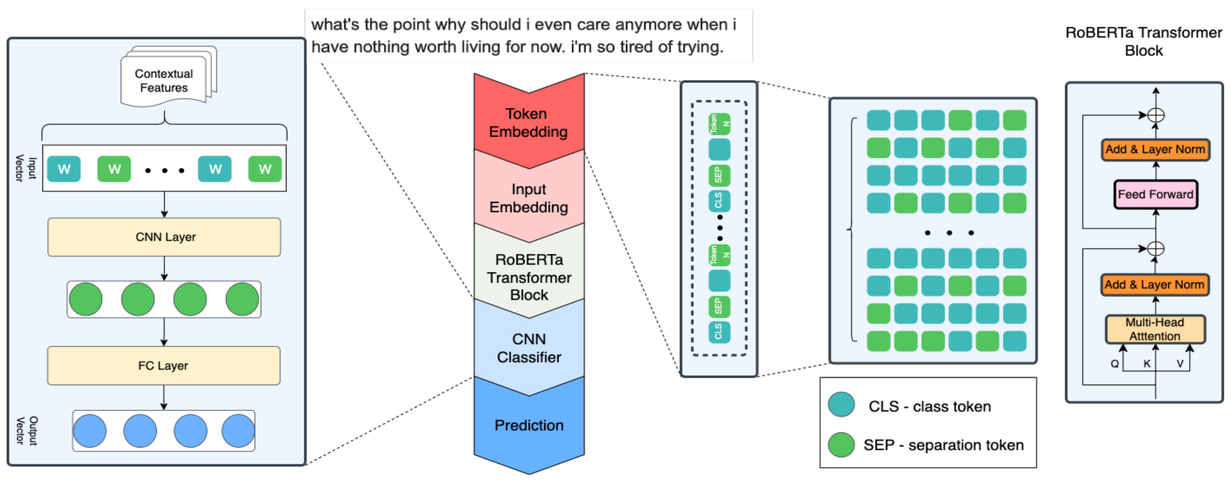


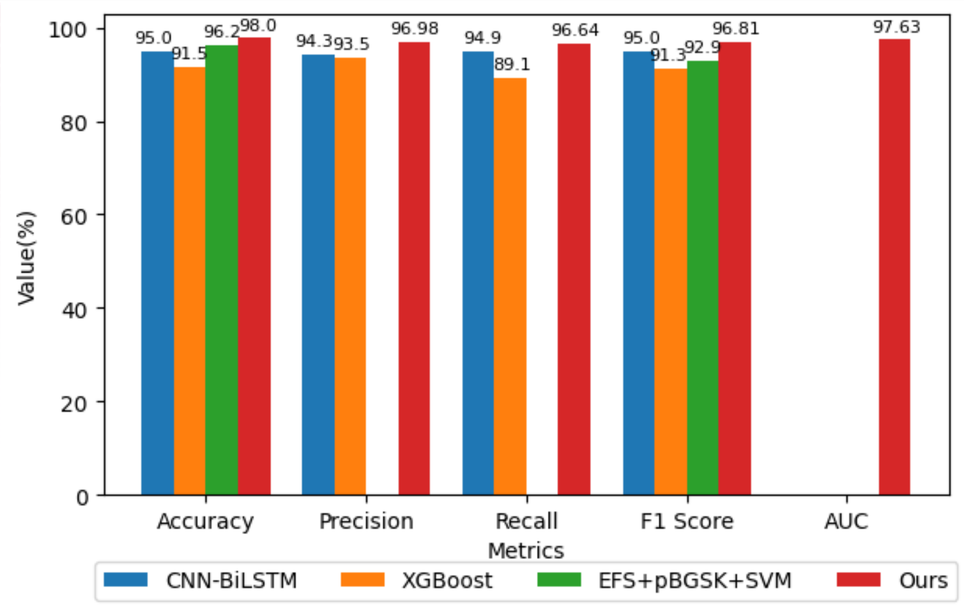
Fig. 1. shows that RoBERTa-CNN contains the RoBERTa backbone and CNN head. RoBERTa gets global features, while CNN focuses on the local ones.

- RoBERTa extracts textual features (complex contextual information + sophisticated linguistic patterns).
- CNN head empowers the model to obtain linguistic patterns and extract speech-related features better.

EXPERIMENT & RESULTS

THE EXPERIMENTAL RESULTS ON SDD DATASET.

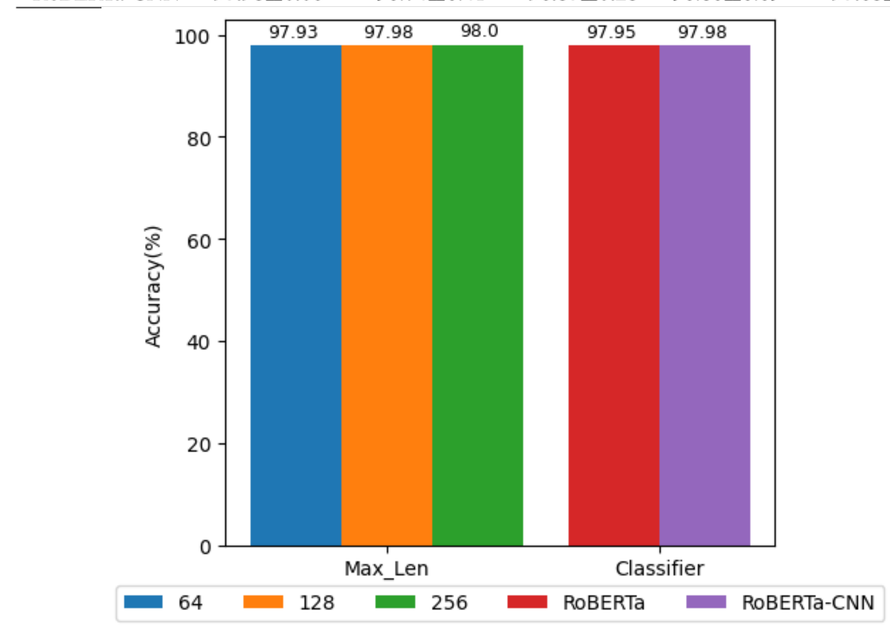
Models	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)	AUC (%)
CNN-BiLSTM [25]	95.00	94.30	94.90	95.00	-
XGBoost [25]	91.50	93.50	89.10	91.30	-
EFS + pBGSK + SVM [26]	96.20	-	-	92.90	-
Ours	**98.00±0.09**	**96.98±0.22**	**96.64±0.30**	**96.81±0.14**	**97.63±0.13**



EXPERIMENT & RESULTS

THE ABLATION STUDY RESULTS. THE UPPER TABLE IS TO TUNE MAX_LEN, AND THE BOTTOM ONE IS TO COMPARE WITH THE ORIGIN ROBERTA. THE VALUE FORMAT IS MEAN±STD.

Max_Len	Accuracy (%)	Precision(%)	Recall (%)	F1 Score (%)	AUC(%)
64	97.93±0.10	**96.99±0.45**	96.46±0.63	96.72±0.15	97.54±0.22
128	97.98±0.06	96.74±0.41	**96.87±0.28**	96.80±0.09	**97.68±0.06**
256	**98.00±0.09**	96.98±0.22	96.64±0.30	**96.81±0.14**	97.63±0.13
Classifier	Accuracy (%)	Precision(%)	Recall (%)	F1 Score (%)	AUC(%)
RoBERTa	97.95±0.06	**97.03±0.39**	96.70±0.39	**96.86±0.05**	97.66±0.11
RoBERTa-CNN	**97.98±0.06**	96.74±0.41	**96.87±0.28**	96.80±0.09	**97.68±0.06**



CONCLUSION

- RoBERTa-CNN can detect suicide intention accurately by analyzing social media posts.
- Removing the noise from the dataset improves the data quality significantly.
- High-quality data empowers the neural network to get high accuracy simply.
- Longer text embedding benefits in the model.

REFERENCES

[1] Jose Manoel Bertolote and Alexandra Fleischmann. Suicide and psychiatric diagnosis: a worldwide perspective. *World psychiatry*, 1(3):181, 2002.

[2] "*NCHS Data Brief, Number 433, March 2022*". CDC. Retrieved March 3, 2022.

[3] "*Suicides in the U.S. reached all-time high in 2022, CDC data shows*". NBC News. August 10, 2023. Retrieved August 11, 2023.

ACKNOWLEDGEMENTS

UNIVERSITY OF DENVER

Dr. Mohammad H. Mahoor, IEEE Senior Member,
Dr. Michelle B. Sabick, ASB Fellow, Dr. David Wenzhong Gao, IEEE Fellow
James Harrer, IEEE Sr. Life Member | Partners in Scholarship (PINs) summer research support | Miss. Zhuo Chen, Mr. Jiahan Li