Detecting Suicide Risk Patterns using Hierarchical Attention Networks with Large Language Models

L. Koushik[1], Vishruth M[2], Anand Kumar M[3]
Department of Information Technology,
National Institute of Technology Karnataka, Surathkal
koushik.201it131@nitk.edu.in[1],
vish.201it167@nitk.edu.in[2],
m_anandkumar@nitk.edu.in[3]

Abstract

Suicide has become a major public health and social concern in the world . This Paper looks into a method through use of LLMs (Large Language Model) to extract the likely reason for a person to attempt suicide , through analysis of their social media text posts detailing about the event , using this data we can extract the reason for the cause such mental state which can provide support for suicide prevention. This submission presents our approach for CLPsych Shared Task 2024. Our model uses Hierarchical Attention Networks (HAN) and Llama2 for finding supporting evidence about an individual's suicide risk level.

1 Introduction

Suicide is a common and very serious concern that affects many lives globally (Picardo et al., 2020). Detecting signs of suicide early is important to provide timely help and support to those at risk. Even though many current methods for identifying suicide risk exist, they have their limitations in terms of efficiency and accuracy. Traditional methods focus on factors such as psychiatric diagnoses, agitation, past suicidal behavior or even self-reported questionnaire surveys (Maclean et al., 2023). However, these methods sometimes struggle to predict suicidal thoughts accurately, and there's a need for more effective tools.

Social media platforms have become a valuable source for understanding and identifying suicide risk. People often share their thoughts and emotions on platforms like Twitter, Facebook, and Reddit. Diagnosing these posts on these platforms can be very helpful to get insights into the lives of the individuals who might be struggling with suicidal thoughts. In recent years, there's been exciting progress in the use of technology to enhance suicide detection, particularly the use artificial intelligence. Many researches have used various machine learning (Lekkas et al., 2021) and deep learning algorithms (Sourirajan et al., 2020) to detect signs of suicide risk with varying levels of success. Improving the current models could end up being very helpful to prevent many suicide cases.

Based on reddit data, in this paper we use LLMs for extracting suicidal thoughts from the user. LLMs use advanced natural language processing algorithms to analyze vast amounts of textual data, including social media posts, to identify patterns and linguistic cues associated with suicidal thoughts. By leveraging the power of LLMs, researchers and mental health professionals can develop more sophisticated and accurate tools for detecting and understanding the individual.

2 Related Work

There is a rising number of research being done in suicide risk detection. This has been a key focus in the field of Natural Language Processing. This task has been done using a lot of methods, but the key focus keeps evolving over time as the field of Machine Learning and Artificial Intelligence keeps expanding. Various Machine Learning algorithms have been used for this task like Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNN) in (Chatterjee et al., 2019) or ensemble learning methods using Convolutional Neural Networks (CNN) and XGBoost used by (Kim et al., 2020). There has also been an increase in the use of transformer based models (Poświata and Perełkiewicz, 2022).

Recent work on large language models (LLMs) suggest that they can perform well on NLP tasks such as information extraction (Agrawal et al., 2022) and question answering (Singhal et al., 2022) which could help us in identifying evidence supporting individual's suicide risk level from a given social media post. We used a hierarchical attention network (HAN) (Yang et al., 2016) in our model to capture the importance of the words and the sen-

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 227–231
March 21, 2024 ©2024 Association for Computational Linguistics

tences of the post to find the portions of the text which indicate the presence of suicide risk.

3 Dataset

This paper discusses our involvement in CLPsych 2024 Shared Task (Chim et al., 2024). The problem statement was to use an open source LLM to provide evidence for the assigned suicide risk level of a person on the basis of their linguistic content. Our task was to highlight the parts of the text which indicate evidence of suicide risk and explain the assignment of a particular suicide risk level using our model.

The dataset we used is from the 2019 CLPsych Shared Task A (Shing et al., 2018),(Zirikly et al., 2019) (University of Maryland Reddit Suicidality Dataset, Version 2). This includes a collection of data from Reddit posts within the r/SuicideWatch community. A careful selection process is employed to focus exclusively on posts where individuals openly share personal experiences related to suicide attempts. The dataset includes Reddit users and their r/SuicideWatch posts, alongside their suicide risk levels in four classes: No, Low, Moderate and Severe risk. However we were asked to exclude posts and users labeled as no risk. The task participants were required to sign data sharing agreements and abide by ethical practice during the competition.

4 Methodology

The architecture of the process followed in this paper can be mainly divided into two parts ,first being use of Hierarchical Attention Modeling to get highlights from the posts from the user and then using these highlights with the post , using this as LLAMA-2 LLM for generating the summarized reason for suicide attempt .

4.1 Hierarchical Attention Modeling for highlights

Hierarchical Attention Networks (HAN) are basically a type of neural network architecture that are used to capture the importance of sequential data present in various hierarchical levels. These work especially well for tasks involving text or document classification. The aim behind using HAN in this paper is to address the challenge of understanding context at different levels such as words, sentences, and entire documents. This architecture is especially very useful for tasks where the meaning of a document is affected not only by individual words but also by the hierarchical structure of sentences and paragraphs.

There can be multiple levels to this attention mechanism: word-level attention mechanism consists of an attention mechanism which assigns different weights to words based on their relevance to the overall document. This helps the model to focus on important words. We use this representation with its respective attention weights to get a context vector representing the document-level information for each word. Similarly sentence-level attention mechanism consists of an attention mechanism which assigns different weights to each sentence based on their importance to the document. We use this representation with its respective attention weights to get a context vector representing the document-level information for each sentence.

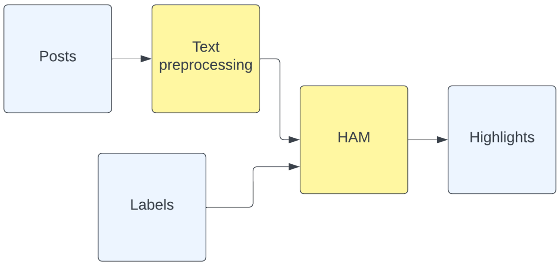

Figure 1: Architecture of the HAN for highlights

Our model architecture uses a Hierarchical Attention Network (HAN) for text extraction. The input sequences undergo embedding, converting them into fixed-size vectors. A bidirectional LSTM layer processes the embedded sequences, capturing word-level contextual information. The attention mechanism then focuses on specific words, forming a context vector. Afterwards, another bidirectional LSTM layer processes these word representations to capture sentence-level context. This hierarchical approach that integrates word and sentence attention mechanisms, allows the model to find out important features at varying levels of detail, improving its capacity for accurate text classification.

4.2 Using Llama2 for summarization

Here in this work we utilize an Large Language Model (LLM) named LLAMA2 as a key component, particularly for the pivotal task of generating concise and summarized insights into the likely reasons behind suicide attempts. The selection of LLAMA2 as it's an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align to human preferences for helpfulness and safety with this it gives it remarkable proficiency in processing and comprehending extensive textual data, making it a well-suited candidate for the task to be done in the paper.

Data from 2 trillion tokens publicly accessible sources were used to pretrain Llama 2. The fine-tuning data consists of more than a million newly annotated human cases in addition to publicly accessible instruction datasets.

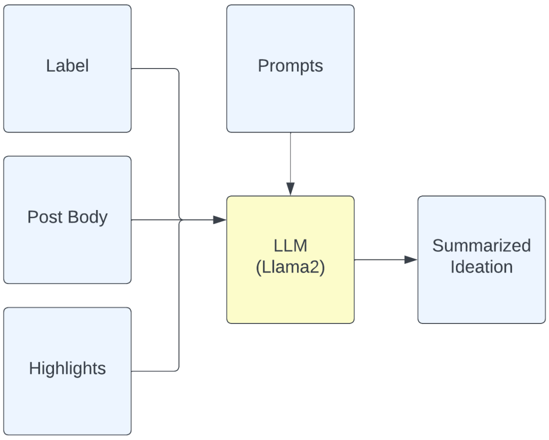

Figure 2: Architecture with LLM model

In this paper, we employ a 4-bit quantization technique to load the Llama 2 7b chat hf version of the model. This approach lowers the computational and memory expenses associated with inference by representing weights using low-precision data types such as 8-bit integer, rather than the customary 32-bit floating point.By lowering the bit count, the resulting model uses less energy and requires less memory, which enables us to make better use of the huge LLM model.

The highlight extractions from the Hierarchical Attention Networks (HAN) procedure and the original post bodies are the two essential components that are concatenated and fed into the model in the next phase. The LLM is able to obtain a thorough representation of all the pertinent information included in the Reddit postings thanks to collective participation. With the help of its pre-trained knowledge gained from exposure to a variety of language patterns and the development of suicidal content expertise, the LLM demonstrates a remarkable capacity to extract the finer features contained in the concatenated input.

The LLM using it's trained architecture helps to create brief summaries that capture the most likely causes of the reported suicide attempts after it receives the concatenated input. The interpretability and accessibility of the research findings are improved by the model's ability to condense complex information into brief and insightful outputs. The produced summaries function as combined representation that provide insightful information about the underlying causes of people's experiences with suicidal thoughts.

Finally we can see that, Llama 2 here is a potent model to help in the study for psychiatrists and researchers, helping to provide complex and educational summaries that facilitate comprehension of the multidimensional character of suicide-related narratives posted users in social media sites.

5 Results

The results of our methodology demonstrate strong performance, with high precision and recall values indicating accurate and comprehensive summarization of suicide-related content within the dataset. The generated summaries not only provide meaningful insights into likely reasons behind suicide attempts but also maintain interpretability. Overall, these results suggest the potential of our method in extracting and understanding sensitive content within online communities, contributing to both research and mental health support systems.

article multirow booktabs

Table 1: Highlights Results-1

Recall	Precision
0.886	0.893

Table 2: Highlights Results-2

Mean Consistency	Max Contradiction
0.784	0.889

Table 3: Summarized text Results

Mean Consistency	Max Contradiction
0.901	0.233

6 Conclusion

In conclusion, our research endeavors to address the intricate challenges of understanding and summarizing suicide-related narratives within the r/SuicideWatch community. The utilization of advanced techniques, including Hierarchical Attention Networks (HAN) and the Large Language Model (LLM) LLAMA2, has yielded promising results. The application of HAN facilitates the extraction of critical information, while LLAMA2's proficiency in processing extensive textual data ensures the generation of concise and insightful summaries.

7 Ethics and Limitations

We obtained our dataset from the University of Maryland and adhered to ethical standards throughout the research process. The dataset, comprising sensitive information, has been handled with utmost confidentiality, and no sharing has occurred to maintain participant privacy and comply with ethical guidelines. Time constraints posed challenges in conducting a more extensive research. This constraint affected the depth of our analysis and the ability to explore additional variables. Despite these limitations, we believe our study provides valuable insights within the given constraints.

Acknowledgements

We would like to express our sincere gratitude to the American Association of Suicidology for their invaluable assistance in making the dataset available for this study. We appreciate the Institute's commitment to promoting scientific research in the field of suicidology.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the clpsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific reports*, 10(1):1–6.

Damien Lekkas, Robert J Klein, and Nicholas C Jacobson. 2021. Predicting acute suicidal ideation on instagram using ensemble machine learning models. *Internet interventions*, 25:100424.

Brant R Maclean, Tahni Forrester, Jacinta Hawgood, John O'Gorman, and Jurgita Rimkeviciene. 2023. The personal suicide stigma questionnaire (pssq): relation to self-esteem, well-being, and help-seeking. *International journal of environmental research and public health*, 20(5):3816.

Jacobo Picardo, Sarah K McKenzie, Sunny Collings, and Gabrielle Jenkin. 2020. Suicide and self-harm content on instagram: A systematic scoping review. *PloS one*, 15(9):e0238603.

Rafał Poświata and Michał Perełkiewicz. 2022. Opi@lt-edi-acl2022: Detecting signs of depression from social media text using roberta pre-trained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 276–282.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Vaibhav Sourirajan, Anas Belouali, Mary Ann Dutton, Matthew Reinhard, and Jyotishman Pathak. 2020. A machine learning approach to detect suicidal ideation in us veterans based on acoustic and linguistic features of speech. *arXiv preprint arXiv:2009.09069*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.