***MindWatch**: A Smart Cloud-based AI solution for Suicide Ideation Detection using Large Language Models*

Introduction:

Suicide, a serious public health concern affecting millions of individuals worldwide, refers to the intentional act of ending one's own life. Suicide stands as a leading cause of mortality worldwide. According to CDC (CDC, 2023), suicide was responsible for 48,183 deaths in 2021. In addition, an estimated 12.3 million American adults seriously thought about suicide, 3.5 million planned a suicide attempt, and 1.7 million attempted suicides in 2021. Mental health issues such as depression, frustration, and hopelessness can directly or indirectly influence the emergence of suicidal thoughts. Early identification of these thoughts is crucial for timely diagnosis and intervention. Clinicians often provide a professional perspective and may not capture the firsthand experiences of patients. These notes typically document cross-sectional impressions taken during a specific point of care. In contrast, social media platforms have become a significant source of daily exchanges where individuals share their experiences, thoughts, and emotions in real-time. Due to the scarce availability of EHR data, suicide notes, or other verified sources, the automated detection of suicidal ideation in social media has gained significant attention in recent years leveraging advancements in artificial intelligence (AI). However, analyzing social-media texts can be a challenge due to the ambiguous and unstructured nature of data.

Supervised Machine Learning (ML) models such as Logistic Regression, Random Forest, Naïve Bayes, or advanced Natural Language Processing (NLP) models such as LSTM (Long Short-Term Memory) networks, can encounter challenges when capturing sentiments from large social media posts due to several reasons:

Noisy and Informal Language: Social media platforms are known for their informal language, including abbreviations, slang, and emoticons. This informal language can make it difficult to accurately interpret and extract meaningful insights from social media texts (Djuric et al., 2015).

Ambiguity and Contextual Understanding: Social media posts often lack context, which can lead to ambiguity in the intended meaning of the text. Understanding the contextual nuances and disambiguating the intended sentiment or message can be challenging in social media texts (Kouloumpis et al., 2011). Thus, care must be taken to avoid overgeneralization of findings to the broader population.

Sarcasm and Irony: Social media users frequently employ sarcasm, irony, or other forms of figurative language. Detecting and correctly interpreting such linguistic cues is a challenge for automated systems (Riloff et al., 2013).

Evolving Language and Neologisms: Social media platforms often give rise to new words, phrases, and expressions that may not be present in pre-existing language models or dictionaries. The rapid evolution of language on social media requires models to adapt quickly to capture these novel expressions (Eisenstein et al., 2014).

Privacy and Ethical Concerns: Analyzing social media texts for mental health detection raises privacy concerns. Balancing the need for identifying individuals at risk with the ethical considerations of privacy and data protection is a critical challenge (De Choudhury et al., 2016).

Labeling and Annotation: Creating accurate and reliable labeled datasets for training machine learning models is a challenging task. The subjective nature of annotating social media texts for suicidal ideation can lead to inter-annotator variability and bias, affecting the performance of the models (Burnap et al., 2015).

Class Imbalance: Social media platforms generate vast amounts of data, but instances of explicit suicidal ideation are relatively rare. This class imbalance can affect the performance of machine learning models, as they may be biased towards the majority class and struggle to generalize well to the minority class (Choudhury et al., 2017).

Our previous analysis showed that LSTM and other supervised ML models failed to capture the context of posts and in general, resulted in plenty of false positives. To overcome the above challenges and come up with a comprehensive and reliable AI tool, we made use of state-of-the-art models such as bi-LSTM and Large Language Transformer based models such as BERT-base, ALBERT and Bio-Clinical BERT along with OpenAI models such as GPT3.5 Turbo, GPT2 tokenizers and OpenAI Embeddings such as text-embedding-ada-002. In the sections below, we will explain the use of the above models and the results achieved so far to analyze the posts on the Reddit platform.

Solution Overview:

The Amazon SageMaker Studio is the integrated development environment (IDE) within SageMaker that provides us with all the ML features that we need in a single pane of glass. Training and fine-tuning models like bi-LSTM, BERT, GPT, and other advanced deep-learning architectures typically require substantial computing power. These models often have a large number of parameters and require significant computational resources for efficient training and optimization. We utilized the SageMaker kernels for this purpose with ml.m5.16x large compute instance. We used a **Reddit dataset** with the size of **2,32,000** labeled records marked as suicidal or non-suicidal. We trained the bi-LSTM model and fine-tuned the pre-trained BERT models such as ALBERT and Bio-Clinical BERT on the above dataset using SageMaker IDE. The motivation behind using these models is their ability to capture important contextual information and perform well when fine-tuned for specific datasets.

• The bi-LSTM model is a type of recurrent neural network (RNN) that can effectively capture sequential information in text data. It is used to learn the patterns and dependencies within social media posts.

• ALBERT, which stands for "A Lite BERT," is a variant of the BERT model that is pre-trained on a large corpus of publicly available data, including web portals and social media texts. ALBERT performs well when fine-tuned for specific datasets and is easier to train as its lightweight. We used AlBERT-Base-V1 for our solution.

• Bio- Clinical BERT is another variant of the BERT model that is pretrained on medical and electronic health records (EHRs) data. Its pretraining on medical contexts makes it suitable for capturing important contextual information from social media posts related to suicide, depression, and mental health.

Below is the architecture of the AWS ecosystem we used for training and plan to further use for deployment of the tool.

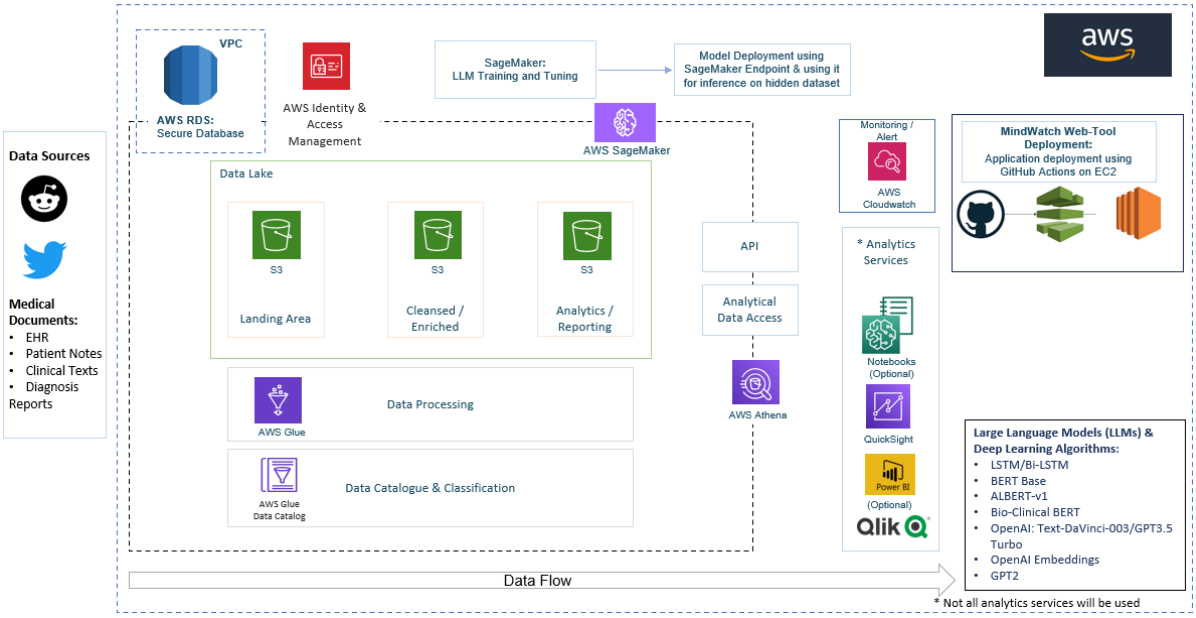


Figure-1 : AWS Architecture for MindWatch Models Training & Deployment

The workflow includes the following steps:

1. We designed a Data Lake Architecture using AWS S3 buckets. The raw training data was initially ingested in the bronze S3 bucket.
2. AWS Glue services such as Glue Crawlers, Glue Jobs and Glue Data Catalog was used for converting the raw csv data to parquet format and the same was stored in silver S3 bucket, followed by pre-processing & cleaning of raw texts/posts. The pre-processed and cleaned texts/dataset was stored in S3 Gold bucket- enriched data, ready for consumption for model training/visualization.
3. The AWS Athena was used to perform SQL Queries on cleaned Glue DB and the same was used by AWS QuickSight Service for visualizations and exploring word-counts.
4. Finally, the state-of-art models were trained/fine-tuned on SageMaker studio by consuming the S3 Gold bucket data.
5. The model artifacts, after training/fine-tuning, such as model weights, tokenizers, config files, etc. was saved in another S3 bucket to make use for inference on hidden dataset.
6. The fine-tuned BERT models- ALBERT and Bio-Clinical BERT artifacts are also uploaded on hugging face portal which makes it easier to use, especially when developing a web-tool like MindWatch.
7. Finally, as mentioned above, we developed a web-based tool which can encompass above models and OpenAI GPT series models such as GPT3.5 Turbo, GPT2 and OpenAI Embeddings which will not only give help doctors to compare the custom LLMs (ALBERT/Clinical-BERT) with OpenAI models but will make the tool more powerful and smart as we continue to train them on more datasets.
8. Below is the current working model of the MindWatch application we plan to deploy on AWS ecosystem, which will make it more scalable and responsive.

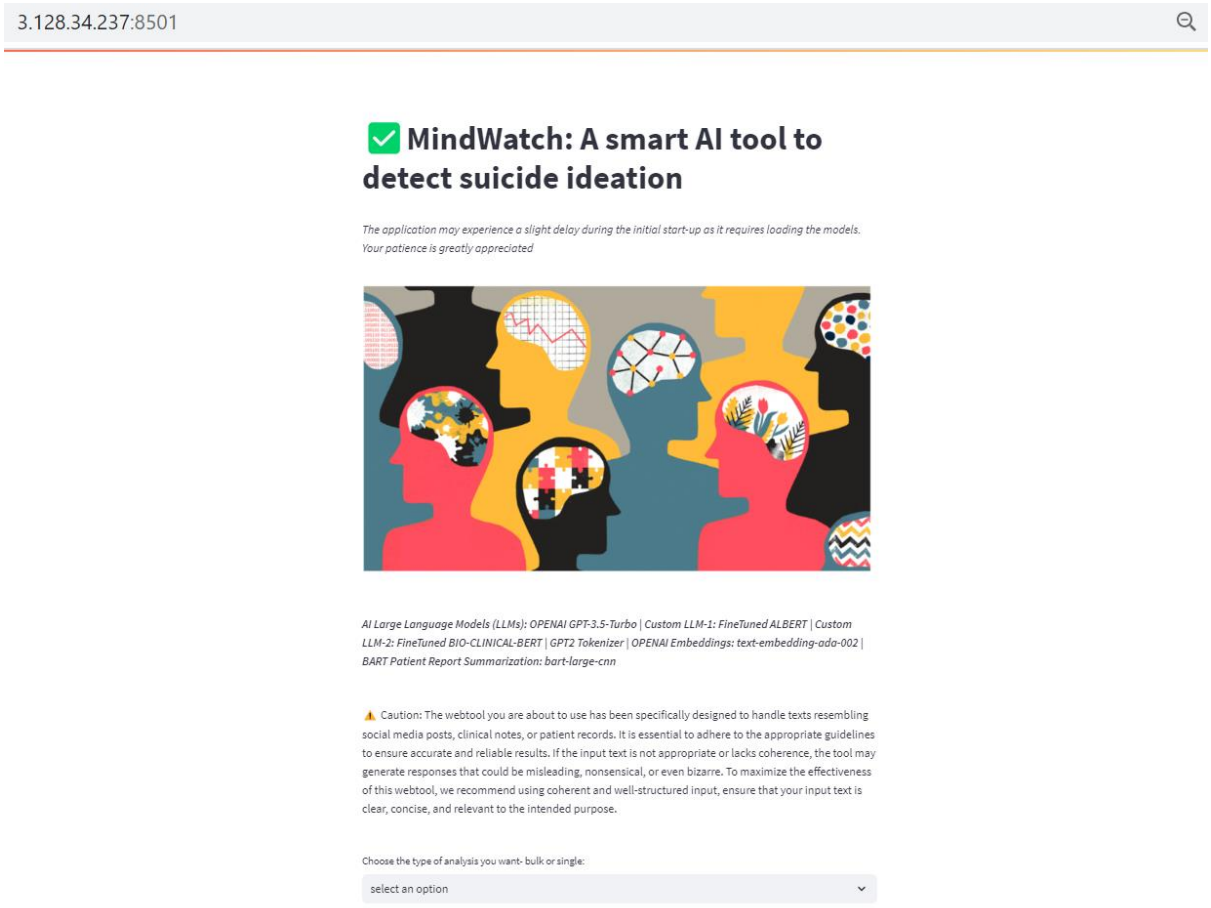


Figure-2 : MindWatch AI Tool

9. We integrate a secure database using AWS RDS service to store and retrieve Patient's information.
10. The above application will be deployed on AWS EC2 using GitHub Actions as self-hosted runner.

Results:

AI Language Model	Accuracy	Precision	Recall
Bi-LSTM	0.9404437	0.938488	0.941558
AlBERT	0.9750848	0.981456	0.968023
ClinicalBERT	0.9488788	0.978481	0.91703
Ensembled (Voting Classifier)	0.9701618	0.981686	0.957667

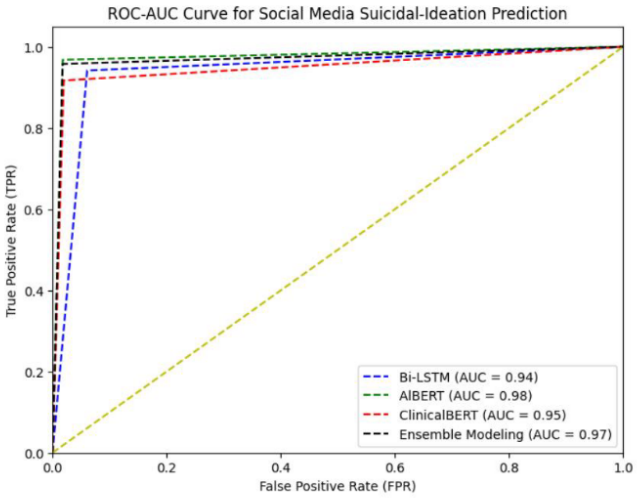


Figure-3: Results for Custom LLMs

All the three models perform exceptionally well, with accuracy and precision/recall greater than 92%. However, as of now ALBERT has been performing better than all the other custom trained/fine-tuned models and even gives better results than zero-shot classification accuracies obtained from OpenAI GPT3.5 Turbo (ChatGPT) on hidden datasets.

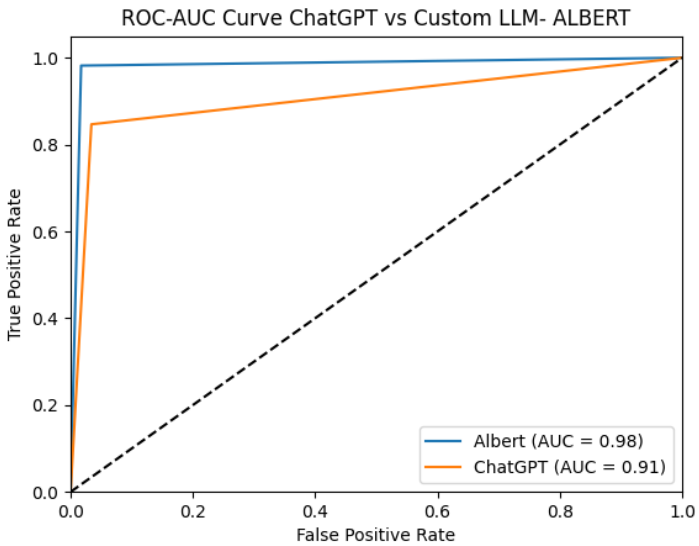
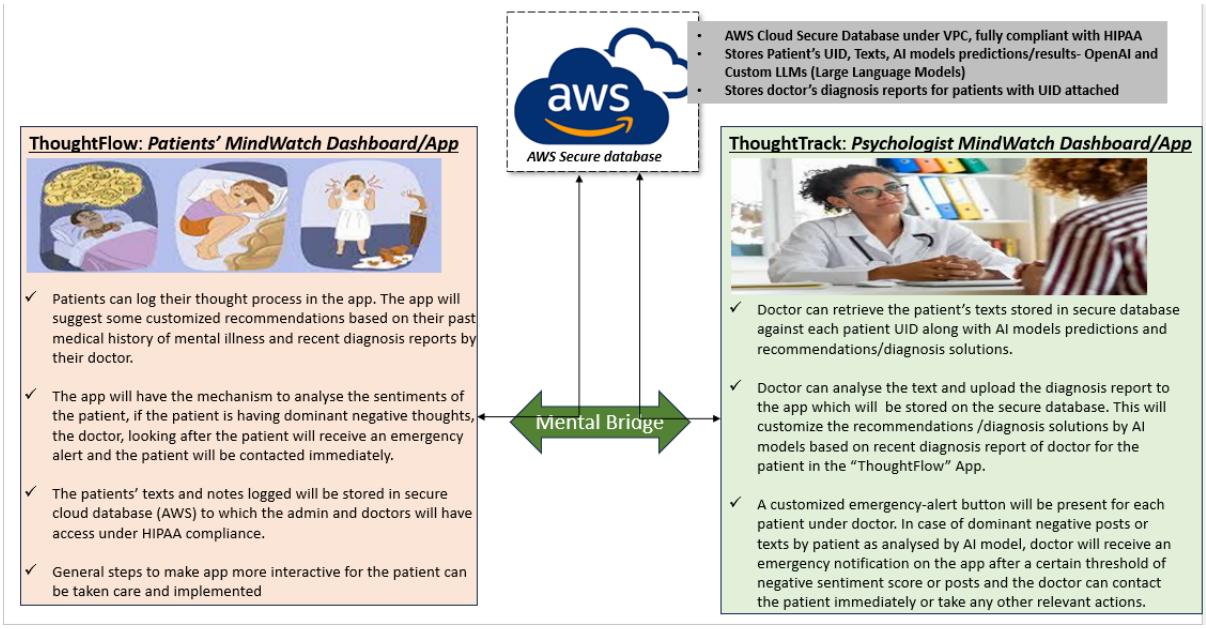


Figure-4: OpenAI ChatGPT vs Custom LLM- ALBERT on hiddent 5000 records

Key Features of MindWatch Tool:

1. Integration of OpenAI models with Custom LLMs for comprehensive and reliable solution with prescriptions or possible diagnosis solutions based on the symptoms analysed from texts/posts.
2. Doctors can customize the prescriptions or diganosis reports suggested by Custom LLMs (ALBERT/Bio-Clinical BERT) based on their specific documents by uploading the same on the web-tool.
3. Single and Bulk file analysis/prediction capabilities.
4. Dignosis report summarization and suggestions.

MindWatch- Bigger Picture (tentative workflow) :



Conclusion:

In conclusion, the MindWatch AI tool, powered by cutting-edge AI language models and an AWS infrastructure, offers a groundbreaking solution for detecting suicidal posts on social media. By accurately identifying individuals at risk of suicide, we can intervene promptly and provide timely support, potentially saving lives. The integration of Custom LLMs combined with OpenAI GPT models and embeddings, ensures high-performance and comprehensive detection capabilities. Through continuous refinement and evaluation, MindWatch can contribute to a safer and more supportive online environment, fostering mental well-being in our communities.

References:

Djuric, N., Zhou, D., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In Proceedings of the 24th International Conference on World Wide Web (pp. 29-30).

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the OMG! In Proceedings of the International Conference on Weblogs and Social Media (pp. 538-541).

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 704-714).

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. PloS one, 9(11), e113114.

De Choudhury, M., Counts, S., & Horvitz, E. (2016). Predicting postpartum changes in emotion and behavior via social media. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (pp. 1460-1474).

Burnap, P., Colombo, W., Scourfield, J., & Hodorog, A. (2015). Machine classification and analysis of suicide-related communication on Twitter. In Proceedings of the 26th ACM Conference on Hypertext & Social Media (pp. 75-84).

Choudhury, M. D., Gamon, M., Counts, S., & Horvitz, E. (2017). Predicting depression via social media. In Proceedings of the 7th International Conference on Weblogs and Social Media (pp. 128-137).