Date: 04-10-2024

PROJ-CS781: Mid-term Review

Project Title: Analyzing Social Media Posts for Mental Health Disorder Detection	**Group#: 29**
Team (Roll / Name): 1. SOUMYADEEP NANDY **(13000121033)** 2. PRITHWISH SARKAR **(13000121037)** 3. SAGNIK MUKHOPADHYAY **(13000121040)** 4. ARKAPRATIM GHOSH **(13000121058)**	**Guide: Nairanjana Chowdhury**

Questions:

1. Justify the uniqueness of your project.

Ans. The project stands out because it leverages machine learning (ML) and natural language processing (NLP) techniques to analyze social media data for mental health disorder detection. Unlike traditional methods of diagnosis, this project offers an automated, scalable, and cost-effective approach using real-time analysis, contributing to early detection and intervention. Moreover, the focus on multiple social media platforms like Reddit and Twitter makes it versatile in detecting conditions like depression, anxiety, and stress.

2. What are the expected benefits from your project?

Ans. The expected benefits are as follows

- **Early Detection**: It aids in identifying mental health issues early, facilitating timely intervention.
- **Scalability**: Handles and processes large volumes of social media data efficiently.
- **Cost-Effectiveness**: Reduces the need for traditional mental health screening.
- **Increased Awareness**: Raises awareness about mental health by highlighting potential indicators.
- **Research Contribution**: Offers valuable insights that can inform future research.
- **Accessibility**: Integrates with healthcare systems to enhance mental health diagnosis and monitoring.

3. a) State the findings from your analysis till date. Show your analysis documentation in PRD till date following the template .

Ans. Initial data collection is completed using Reddit and Twitter Sentiment Analysis Dataset from Kaggle. Sentiment analysis and feature extraction using a Bag of Words (BoW) model have been implemented to identify indicators of mental health disorders.

3. b) What is the % completion progress of analysis? Explain the calculation logic.

Ans. Analysis progress is **30%** complete. This calculation is based on completing data collection, sentiment analysis, and the first round of model training and validation.

4. a) State the findings from your design till date. Show your design documentation in PRD till date following the template .

Ans. The architecture integrates NLP techniques for feature extraction and ML models like Naive Bayes, Support Vector Machines (SVM) and Random Forest for classification. The web scraping module for data extraction from Reddit is functional.

4. b) What is the % completion progress of design? Explain the calculation logic.

Ans. Design progress is **30%** complete, calculated based on preliminary data extraction and model design, with optimization pending.

5. State the tools you are using for analysis and design; apply Software Engineering concepts.

Ans. **Analysis**: Python libraries such as NLTK for text preprocessing, TextBlob for sentiment analysis, Scikit-learn for classification, Kaggle for Twitter Sentiment Analysis Dataset and API tools like PRAW (Reddit).

Design: UML for system modeling, and LaTeX for documentation.

6. Are you foreseeing any risk in completing your project?

Ans. Potential risks include challenges in increasing model accuracy and ethical concerns regarding data privacy, particularly when handling sensitive user-generated content from social media.

7. State the study references you have used in this semester (after Synopsis preparation)

Ans. The references are as follows :

- Munmun De Choudhury, Michael Gamon, Scott Counts, Eric Horvitz. (2013). Predicting Depression via Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*. https://api.semanticscholar.org/CorpusID:13626864

- Sharath Chandra Guntuku, David Bryce Yaden, Margaret L. Kern, Lyle H. Ungar, Johannes C. Eichstaedt. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18, 43-49. https://api.semanticscholar.org/CorpusID:53273218

- Priya Mathur, Amit Kumar Gupta, Abhishek Dadhich. (2022). Mental Health Classification on Social-Media: Systematic Review. *Proceedings of the 4th International Conference on Information Management & Machine Intelligence*. https://api.semanticscholar.org/CorpusID:258970659

- Hasan, Mohammad. (2023, November). The Impact of Social Media on Mental Health and Well-Being on Students.

- Moin Nadeem. (2016). Identifying depression on Twitter. *arXiv preprint arXiv:1607.07384*.

- Hatoon S AlSagri, Mourad Ykhlef. (2020). Machine learning-based approach for depression detection in Twitter using content and activity features. *IEICE Transactions on Information and Systems*, 103(8), 1825-1832. The Institute of Electronics, Information and Communication Engineers.

- Konda Vaishnavi, U Nikhitha Kamath, B Ashwath Rao, N V Subba Reddy. (2022). Predicting Mental Health Illness using Machine Learning Algorithms. *Journal of Physics: Conference Series*, 2161(1), 012021. https://dx.doi.org/10.1088/1742-6596/2161/1/012021

- Ramin Safa, S. A. Edalatpanah, Ali Sorourkhah. (2023). Predicting mental health using social media: A roadmap for future development. *arXiv preprint arXiv:2301.10453*. https://arxiv.org/abs/2301.10453

8. Submission of RM (excel file) and PP (Microsoft Project Plan) as separate attachments following the given templates. Show relevant parts of PRD prepared till date following the template .

Ans. RM (Excel file) and PP (Project Plan in Microsoft Project) are required as separate attachments, detailing resource management and timeline adherence.

9. What is the % completion progress of the prototype (refer target for 7th Semester as set by Guide)? Explain the calculation logic.

Ans. Prototype progress is **30%** complete. This calculation is based on completing data collection, sentiment analysis, and the first round of model training and validation.

10. Additionally, Guides should ask questions (what's, how-to's) on understanding of the target system and expected functions.

Ans. **What's**: What are the specific mental health conditions targeted by your classification system? - Normal, Depression, Bipolar Disorder, PTSD (Post Traumatic Stress Disorder)

How-to's: How does your system ensure data privacy when collecting and analyzing sensitive social media posts? -

- **Reddit API**: We use the publicly accessible Reddit API, which is free and adheres to Reddit's terms of service. Only publicly available data (such as posts and comments) is extracted, ensuring that no private user information is accessed. This helps maintain the privacy of individuals while still enabling valuable insights into mental health discussions.
- **Twitter Data**: Instead of directly scraping or collecting data from Twitter, we utilize an open-source dataset from Kaggle, which has been publicly shared under appropriate licenses. This ensures that the data has already been anonymized and shared with public consent, further safeguarding user privacy.