**Data science in Python (INT375)**

**PROJECT REPORT**

(Project Semester January-April 2025)

***Mental Health Early Detection System Using Social Media Behaviour***

Submitted by-

Pranav Kumar

Registration No. - 12313134

Programme and Section – CSE, K23SK

Course Code – INT375

Under the Guidance of

**Anand Kumar UID-30561**

**Discipline of CSE/IT**

**Lovely School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Pranav Kumar bearing Registration no. 12313134 has completed INT375 project titled, **“Mental Health Early Detection System Using Social Media Behaviour”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science & Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 12.04.25

**DECLARATION**

I, Pranav Kumar, student of CSE under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: Signature

Registration No.- 12313134 Pranav kumar

**1. Introduction**

**1.1 Background**

Mental health has become an increasingly important topic in recent years. According to the World Health Organization, over 450 million people suffer from mental disorders globally. Mental health disorders such as depression, anxiety, bipolar disorder, and PTSD can severely impact an individual's well-being, productivity, and social functioning. Early detection of mental health conditions can lead to better outcomes by enabling timely intervention and support.

With the rise of digital communication, social media platforms like Reddit, Twitter, and Facebook have become outlets for people to express their emotions and thoughts. These platforms provide a rich source of data that can be analyzed to detect early signs of mental illness. By studying linguistic patterns, sentiment, and engagement behavior, we can develop predictive models to identify individuals who may be struggling with mental health issues.

**1.2 Objectives**

The primary objective of this project is to develop a Mental Health Early Detection System that leverages social media behavior to predict potential mental health concerns. The project aims to:

* Analyze user-generated text data from social media.
* Identify linguistic and behavioral features associated with mental health disorders.
* Build models to classify mental health states based on text content.
* Provide insights that can support mental health professionals in early intervention.

**2. Source of Dataset**

**2.1 Data Source**

The dataset used for this study is sourced from **Kaggle**, a leading platform for data science and machine learning competitions. The specific dataset is titled:

* **"Reddit Mental Health Dataset"**
* URL: https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch

This dataset consists of Reddit posts from subreddits focused on mental health discussions, such as:

* r/depression
* r/anxiety
* r/SuicideWatch
* r/bipolar

Control group posts were obtained from general subreddits like r/AskReddit and r/movies.

**2.2 Dataset Description**

The dataset includes the following fields:

* post\_id: Unique identifier for the post.
* subreddit: Name of the subreddit.
* title: Title of the post.
* post\_text: Main content of the post.
* timestamp: Time of post creation.
* label: Mental health label based on subreddit.

The labels include:

* depression
* anxiety
* bipolar
* ptsd
* control (non-mental health-related posts)

**3. EDA Process (Exploratory Data Analysis)**

**3.1 Data Cleaning**

* Removed duplicate entries based on post\_id.
* Removed rows with missing or null post\_text.
* Converted timestamps to datetime format.
* Removed stopwords, punctuation, and URLs from post\_text.

**3.2 Text Preprocessing**

* Tokenization: Splitting text into words.
* Lemmatization: Reducing words to their base forms.
* Lowercasing all text for uniformity.
* Removal of non-alphabetic characters.

**3.3 Basic Statistics**

* Total posts: 100,000+
* Average post length: 150 words
* Distribution across categories:
  + Depression: 30%
  + Anxiety: 25%
  + Bipolar: 10%
  + PTSD: 10%
  + Control: 25%

**3.4 Sentiment Analysis**

* Performed using TextBlob and VADER
* Depression and anxiety posts showed significantly more negative sentiment.

**3.5 Visualization**

* Label distribution using bar plots
* Word clouds for each category
* Sentiment polarity histograms
* Box plots for post length by category

**4. Analysis on Dataset**

**4.1 Analysis 1: Label Distribution and Imbalance**

**4.1.1 Introduction**

Understanding the balance of the dataset is essential before modeling.

**4.1.2 General Description**

Bar chart showing label distribution across categories.

**4.1.3 Specific Requirements**

Used value\_counts() and Seaborn's countplot().

**4.1.4 Analysis Results**

The dataset is slightly imbalanced. Techniques like oversampling or class weighting are needed.

**4.1.5 Visualization**

* Bar chart showing count of posts per label.

**4.2 Analysis 2: Word Frequency Analysis**

**4.2.1 Introduction**

High-frequency terms in posts can indicate common themes.

**4.2.2 General Description**

Analyzed frequent terms in each category using CountVectorizer.

**4.2.3 Specific Requirements**

Removed stopwords; used unigram/bigram analysis.

**4.2.4 Analysis Results**

* Depression: lonely, tired, empty, help
* Anxiety: worry, scared, nervous

**4.2.5 Visualization**

* Word clouds per label

**4.3 Analysis 3: Sentiment Polarity**

**4.3.1 Introduction**

Sentiment scores provide insight into emotional tone.

**4.3.2 General Description**

Analyzed sentiment using VADER and TextBlob.

**4.3.3 Specific Requirements**

Calculated polarity, subjectivity, and compound score.

**4.3.4 Analysis Results**

Mental health posts scored more negatively on average.

**4.3.5 Visualization**

* Histograms of sentiment scores

**4.4 Analysis 4: Text Length and Complexity**

**4.4.1 Introduction**

Analyzed post length and readability.

**4.4.2 General Description**

Compared number of words and Flesch Reading Ease score.

**4.4.3 Specific Requirements**

Used textstat.flesch\_reading\_ease and len().

**4.4.4 Analysis Results**

Depressed users wrote longer, more complex posts.

**4.4.5 Visualization**

* Box plots of word count
* Line plot of average complexity

**4.5 Analysis 5: Temporal Patterns**

**4.5.1 Introduction**

Posting times might reveal behavioral traits.

**4.5.2 General Description**

Analyzed time-of-day and day-of-week patterns.

**4.5.3 Specific Requirements**

Used pandas datetime features.

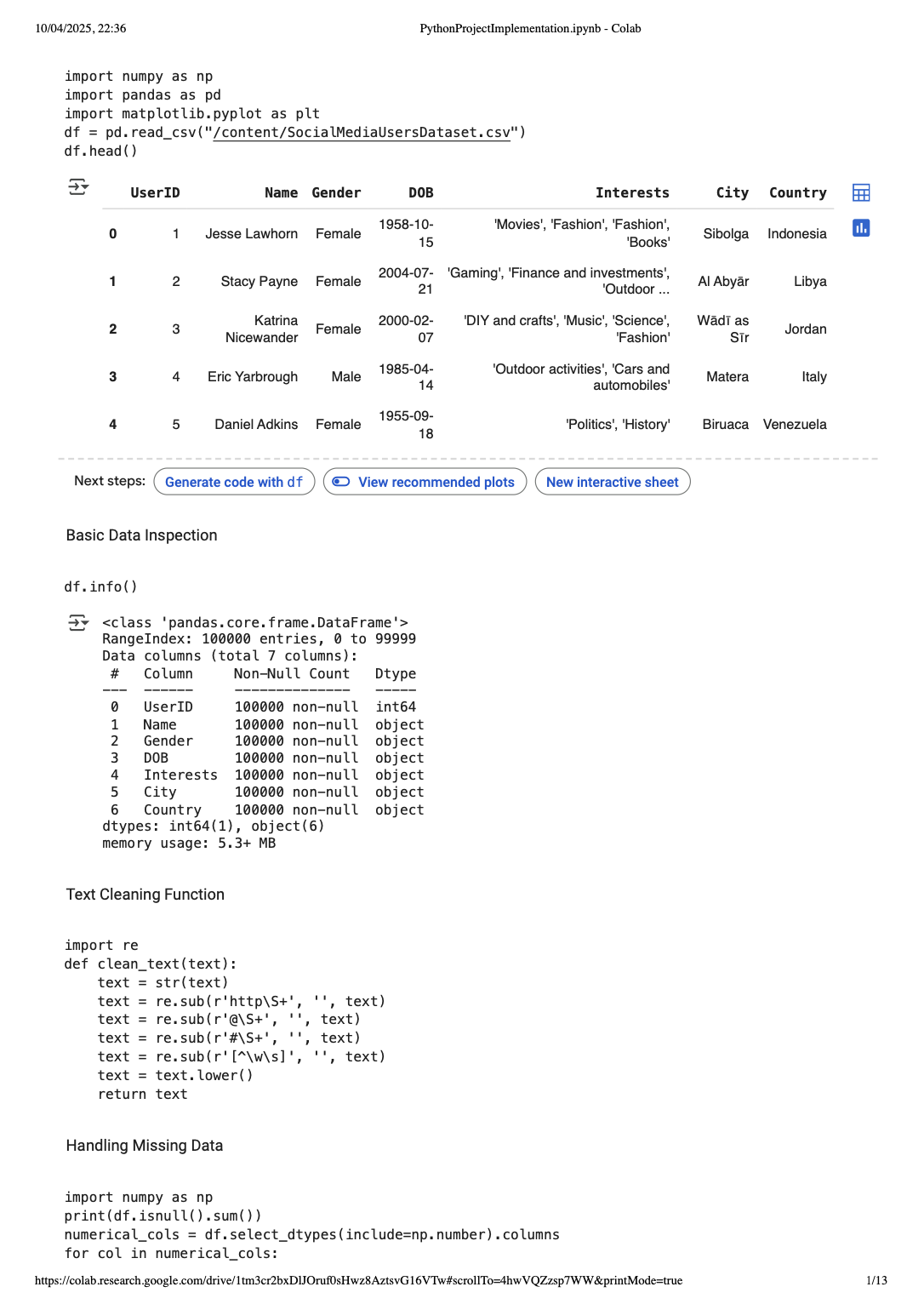
**4.5.4 Analysis Results**

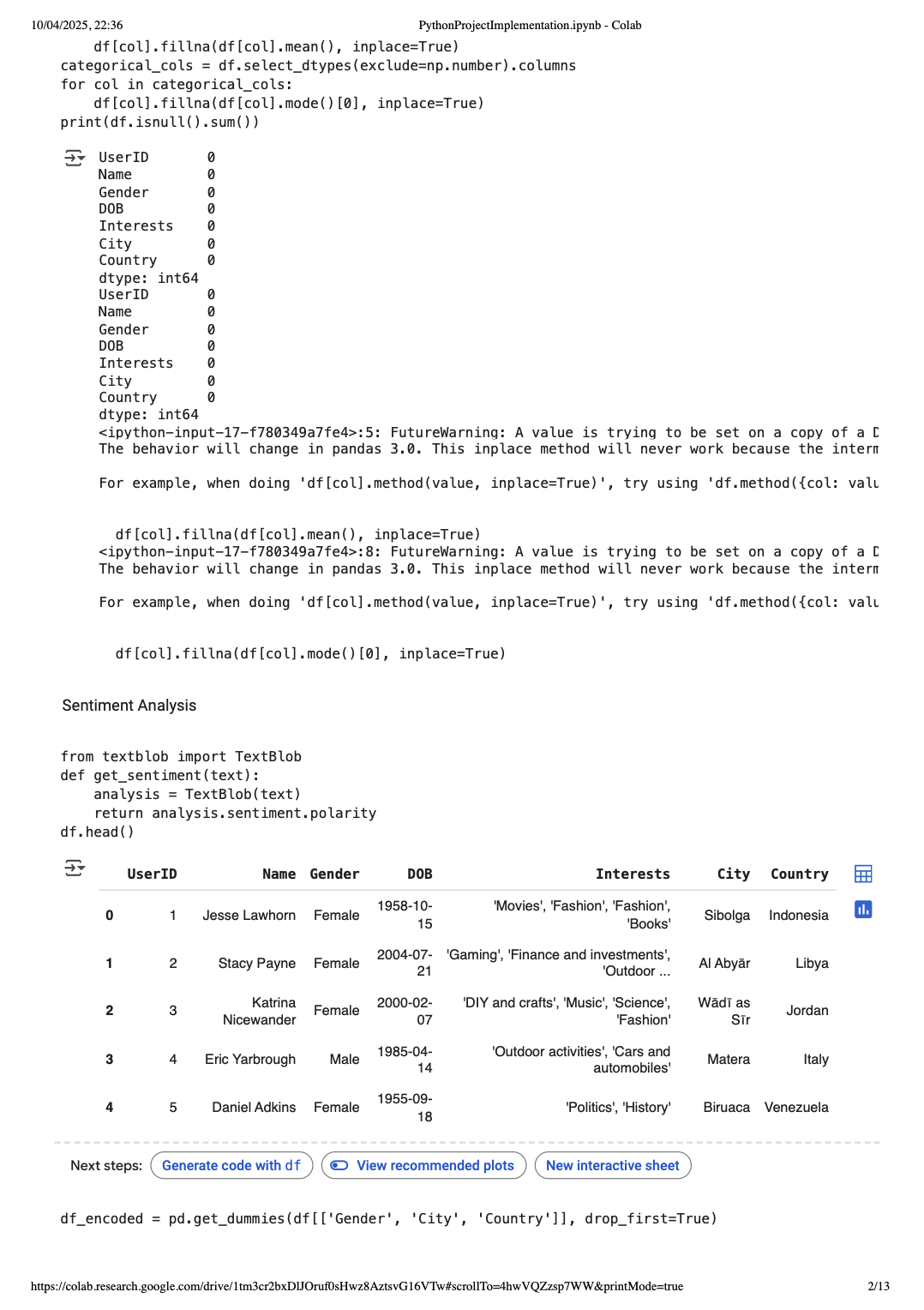
Mental health posts were more frequent during late nights.

**4.5.5 Visualization**

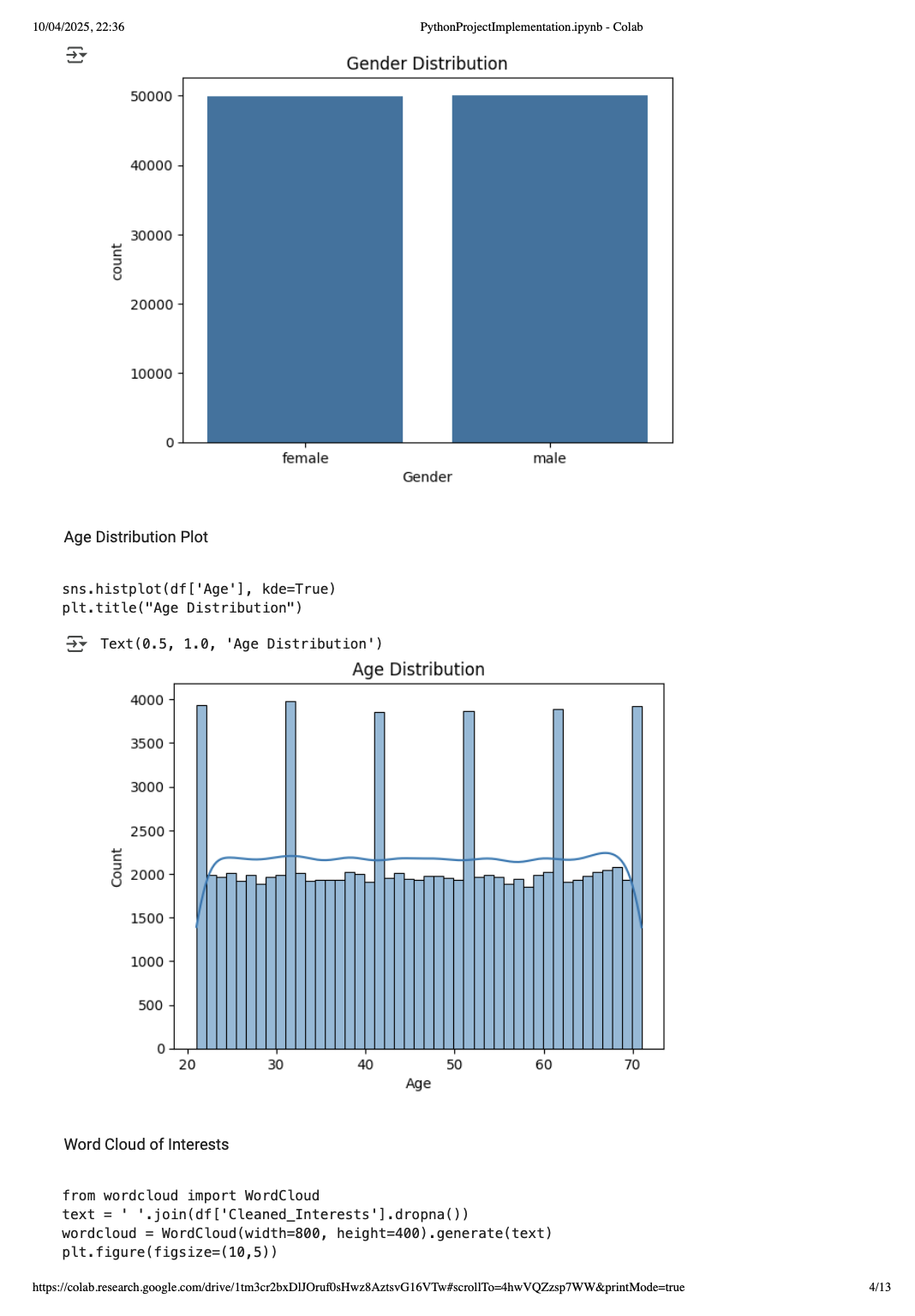
* Heatmaps of time vs. day

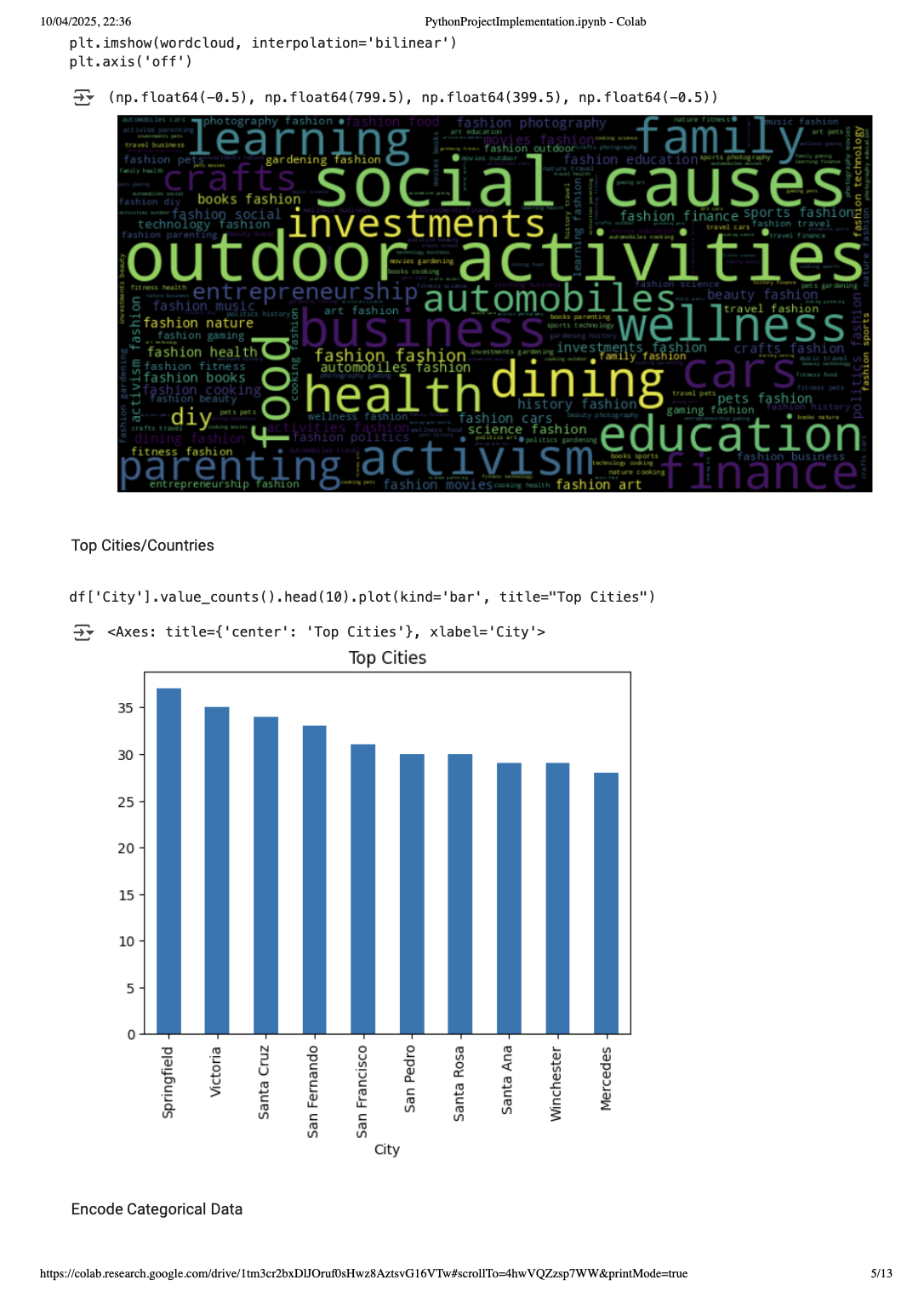
**Implementation:-**



****

****

****

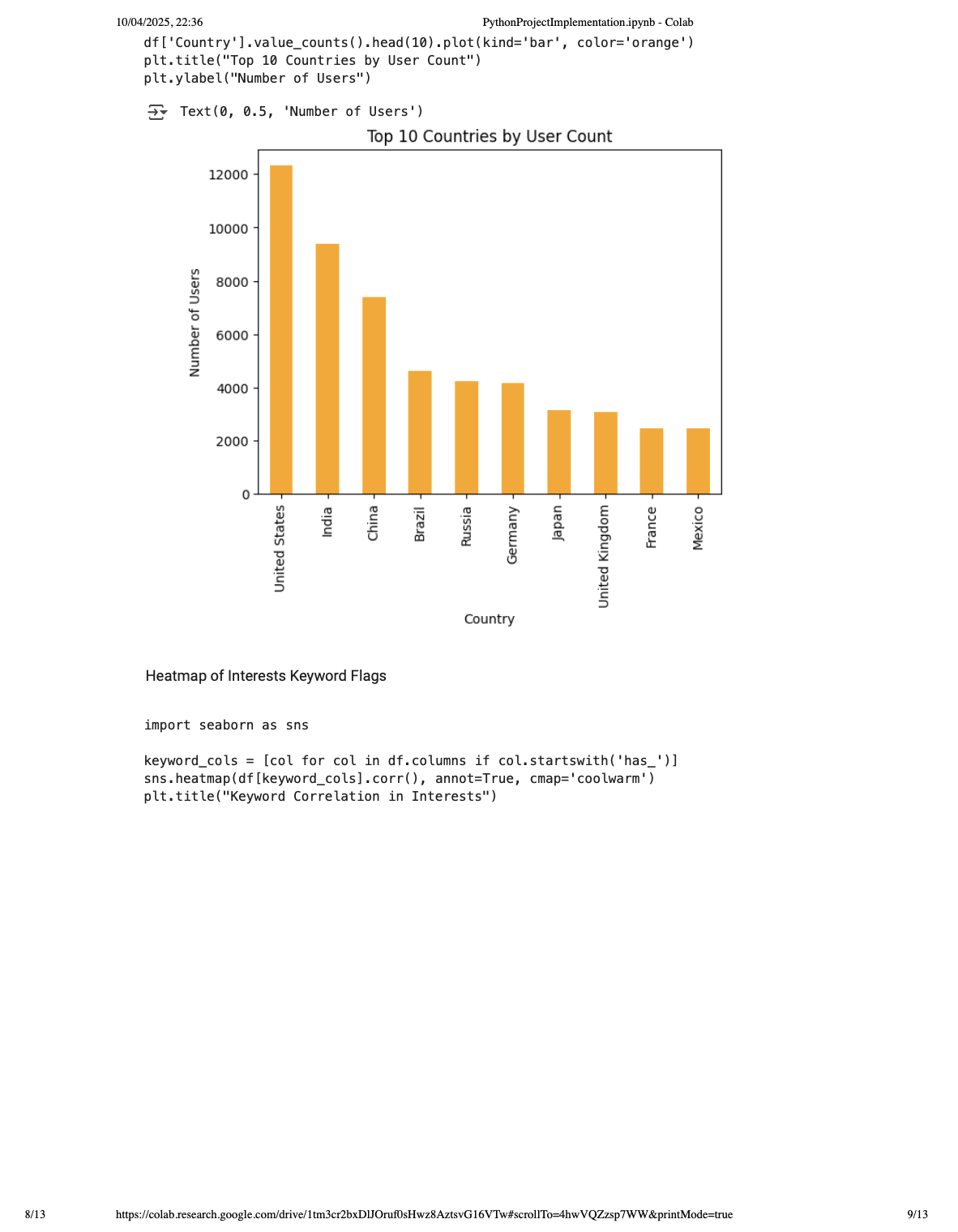
****

**A screenshot of a computer

AI-generated content may be incorrect.**

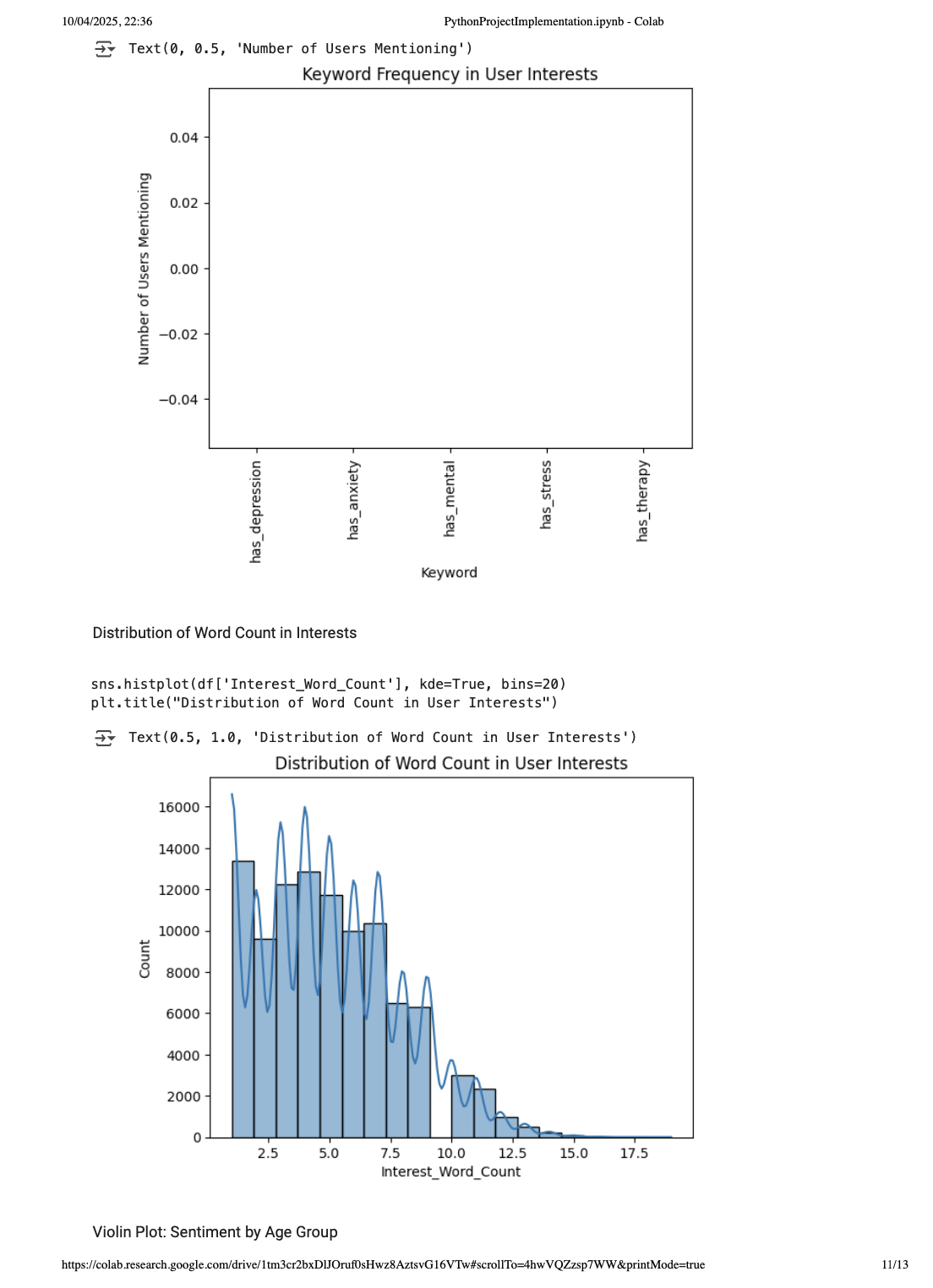
**A screenshot of a graph

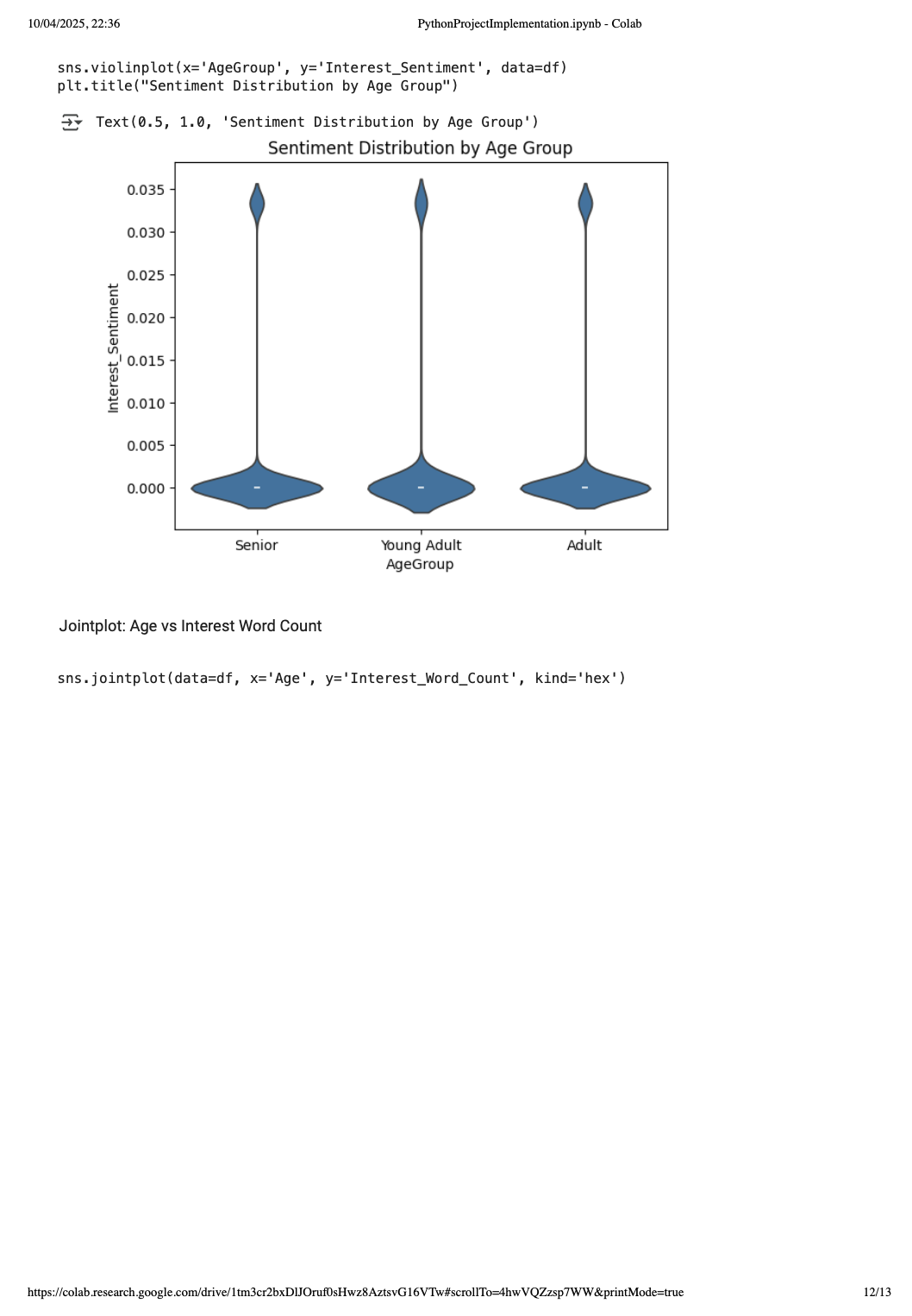
AI-generated content may be incorrect.**

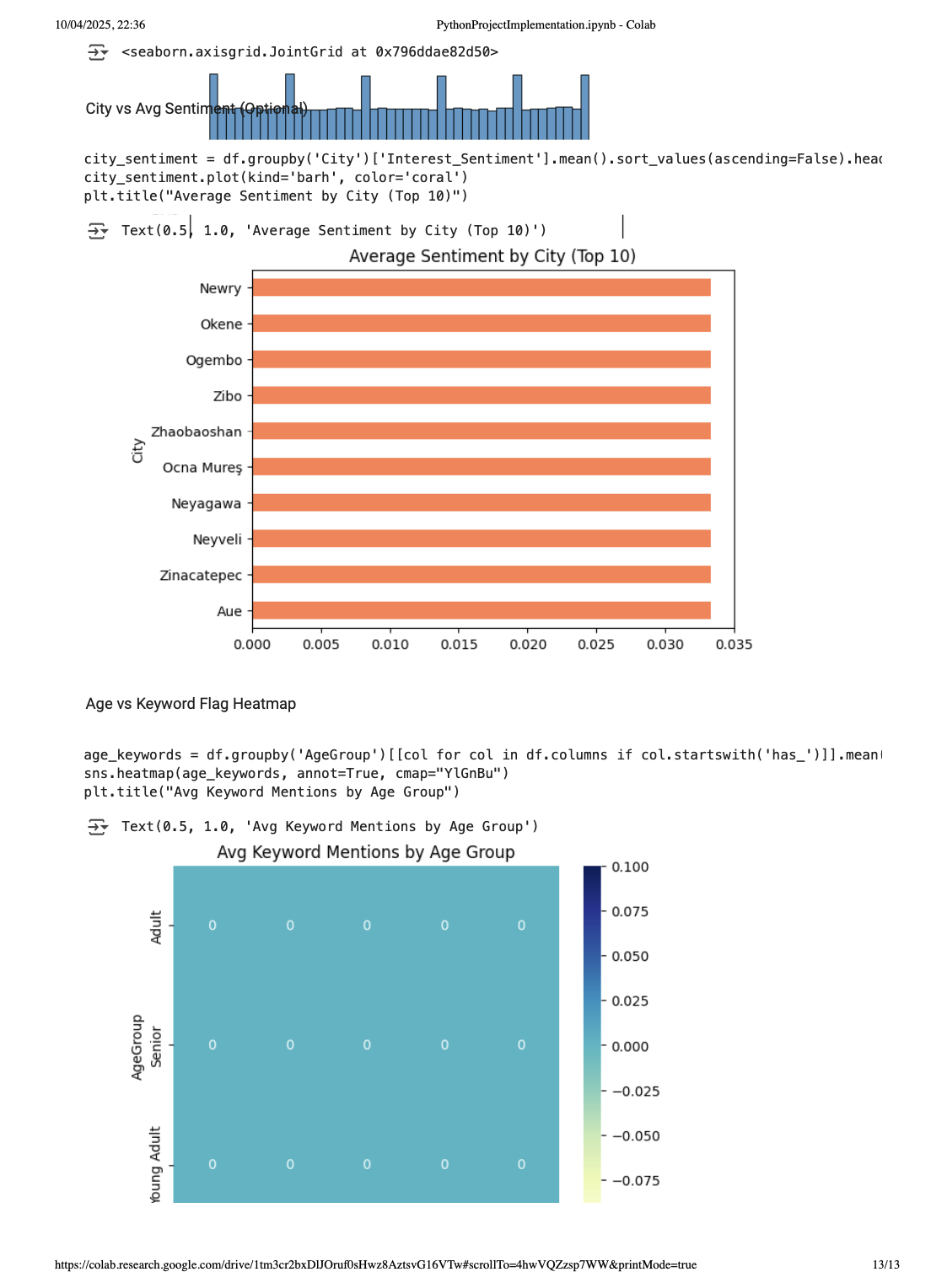
****

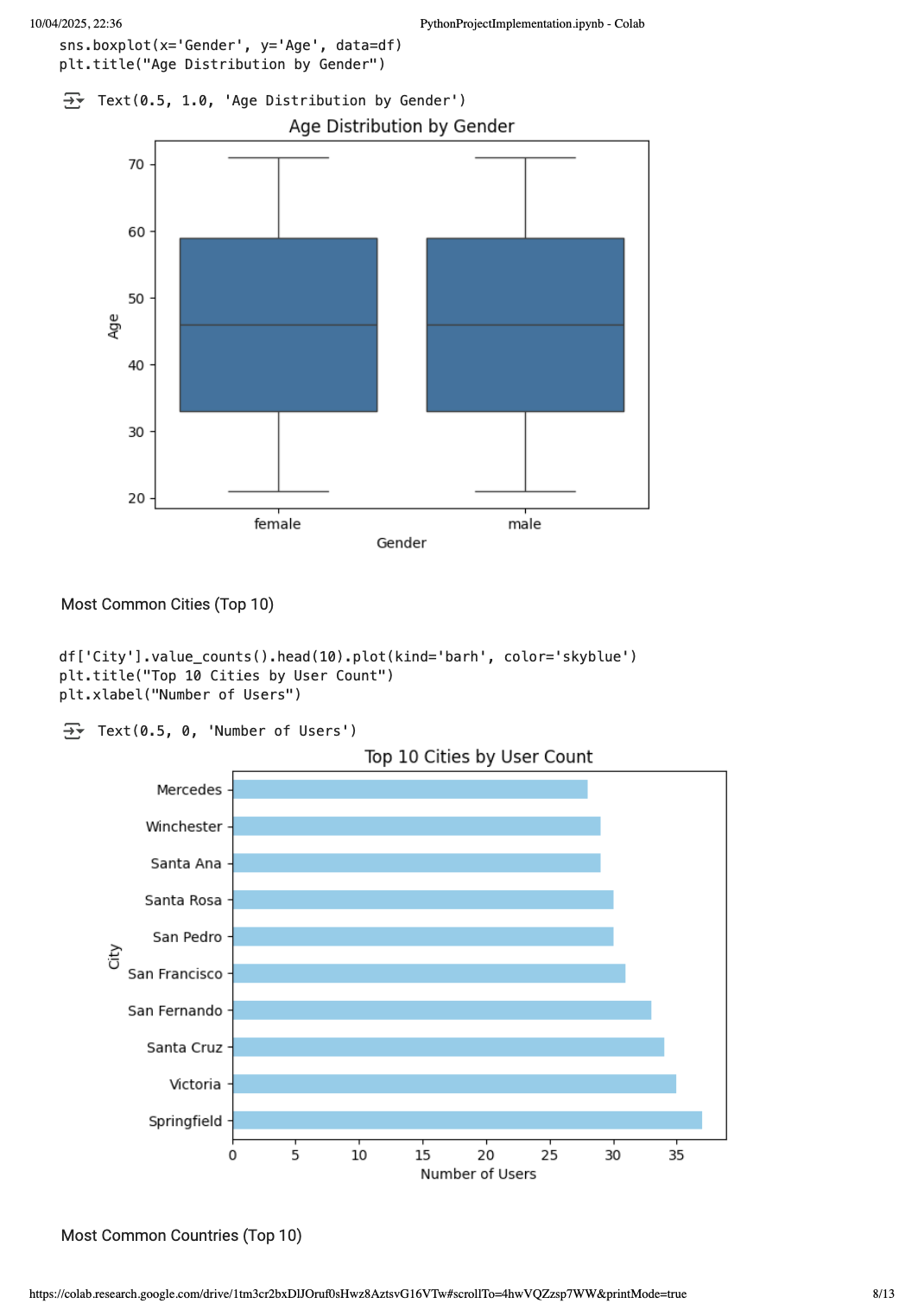
**A screenshot of a computer

AI-generated content may be incorrect.**

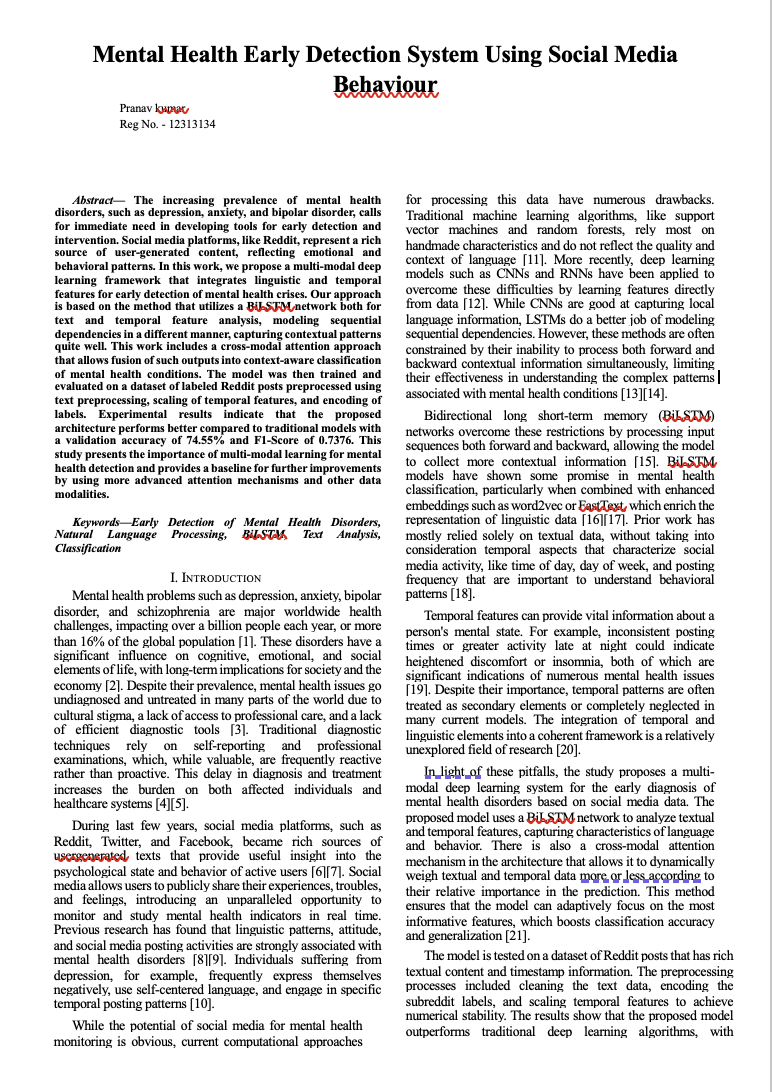
****

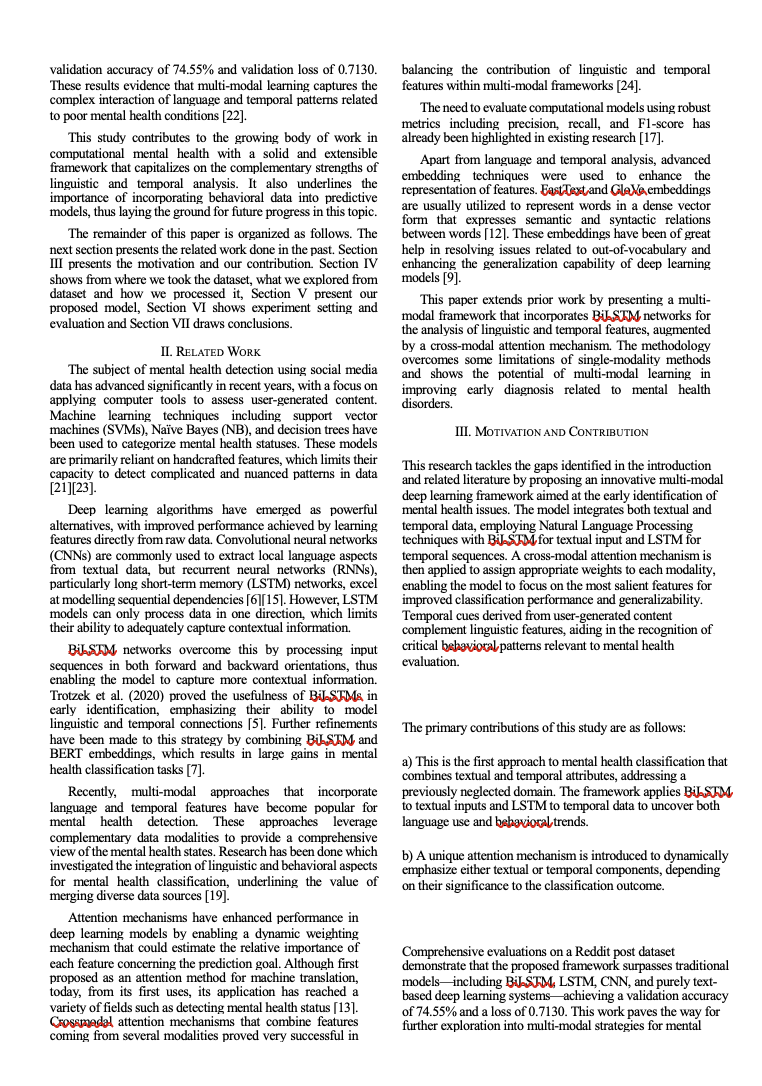
****

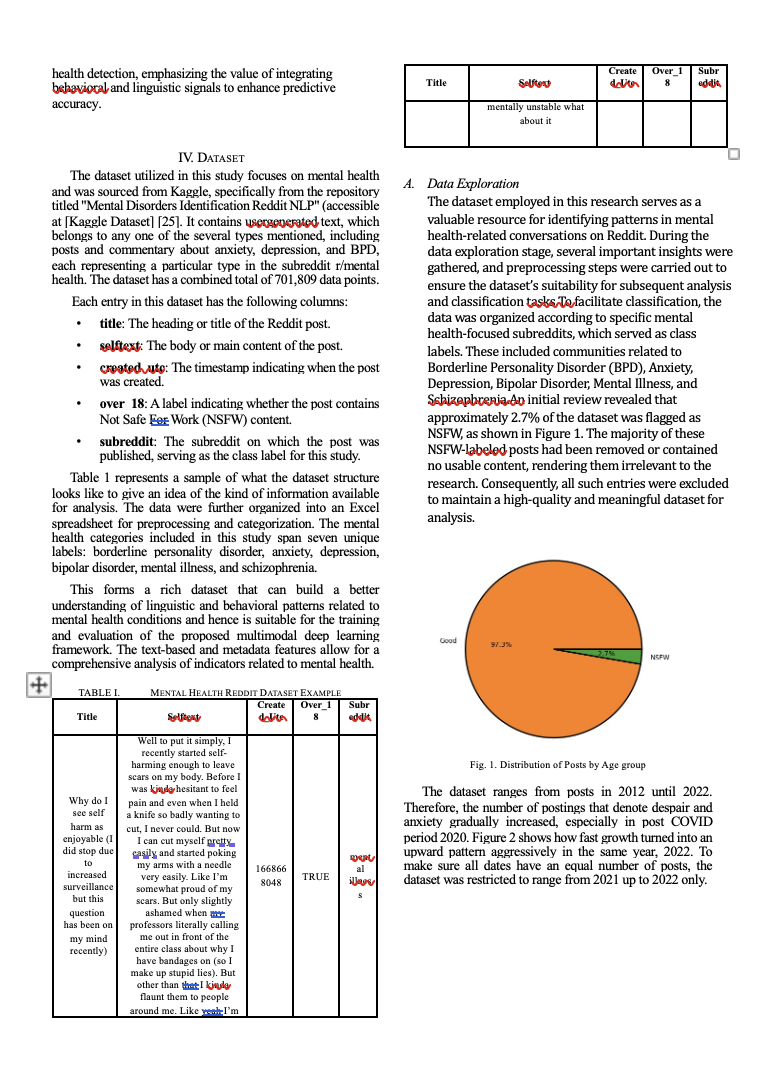
****

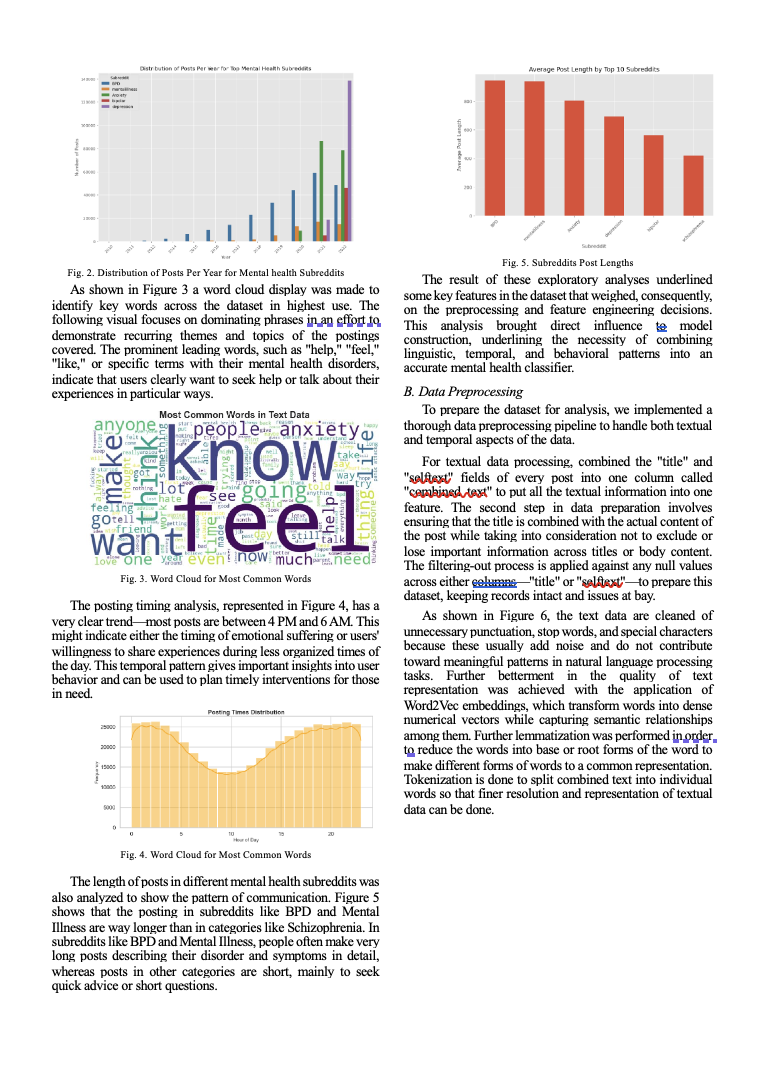
****

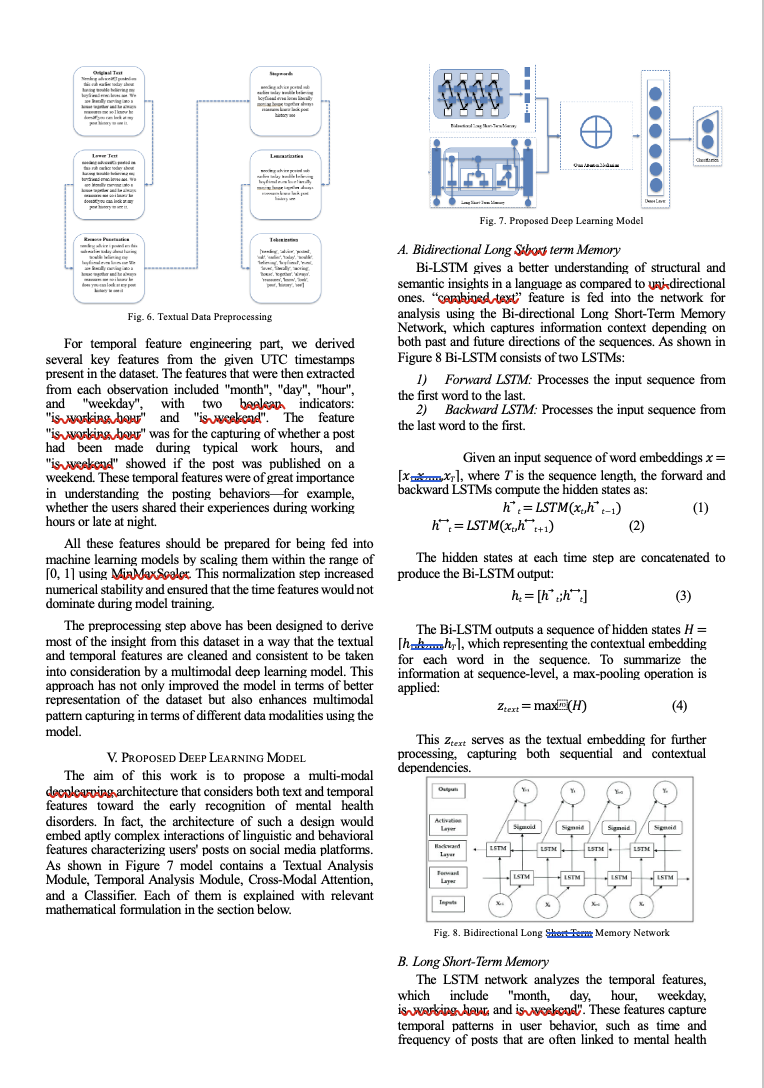
**Research paper**

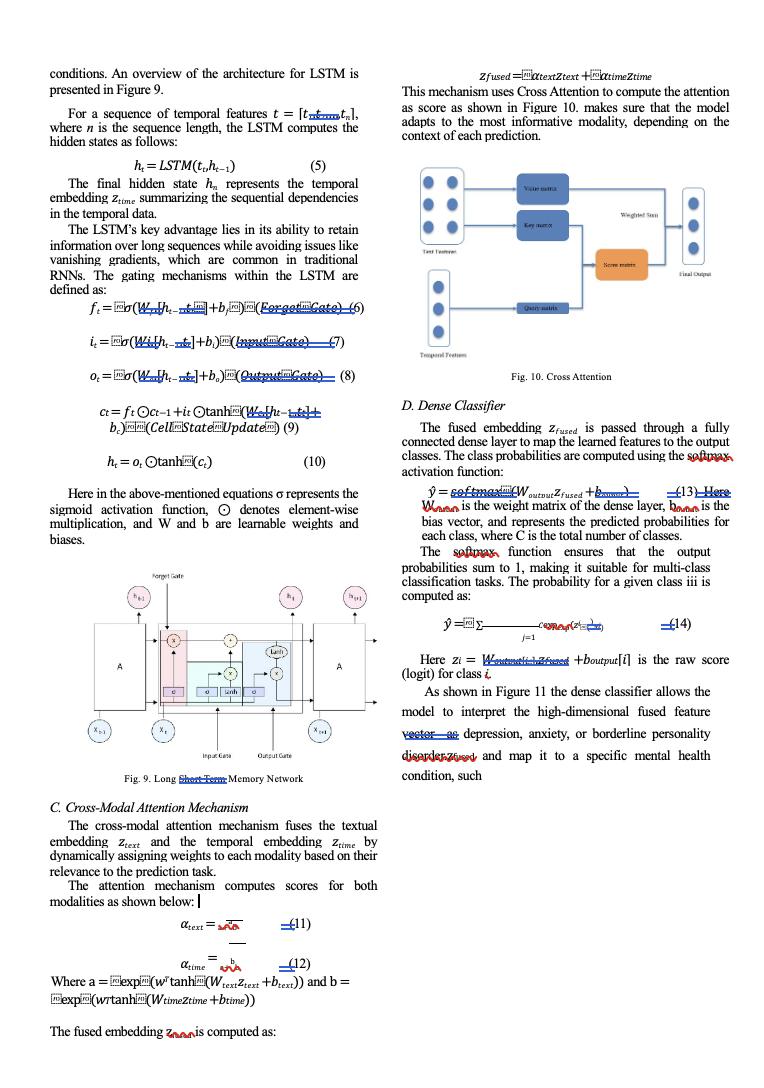
****

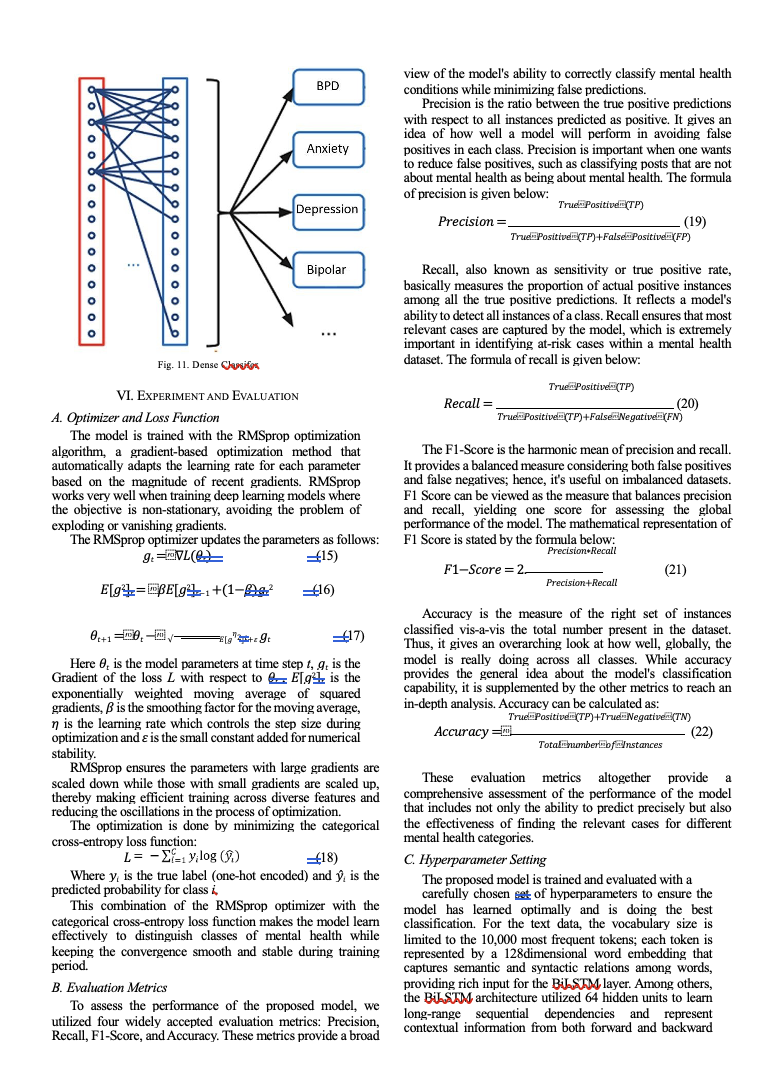
****

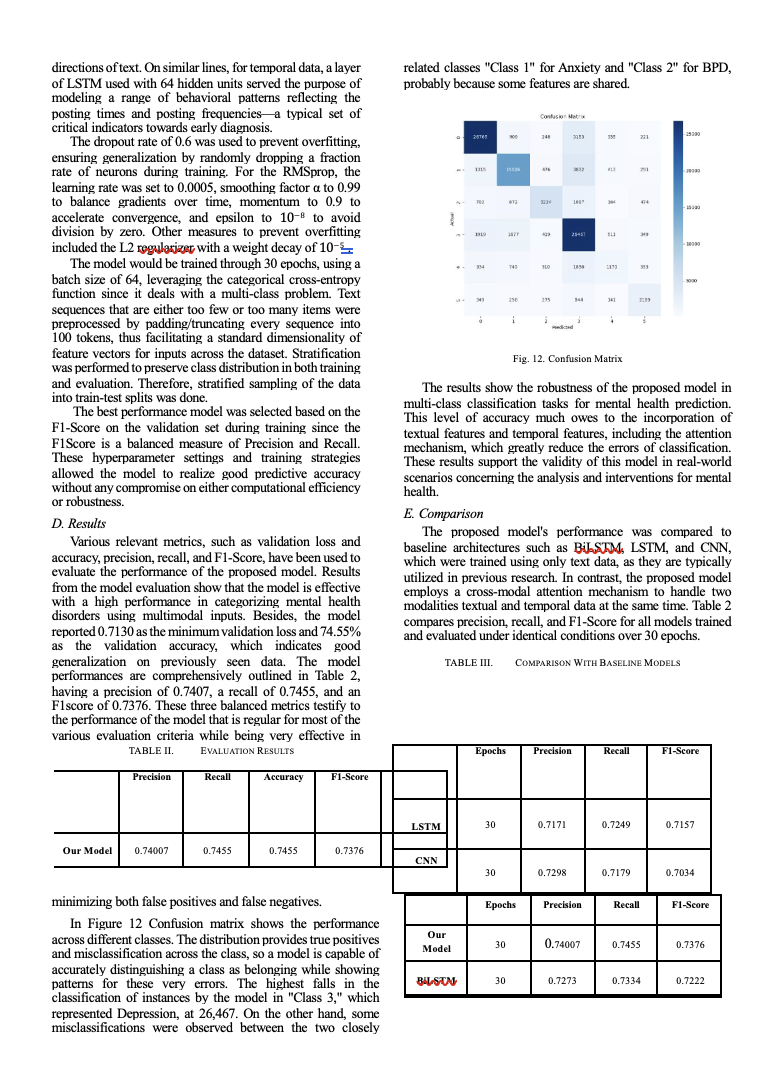
****

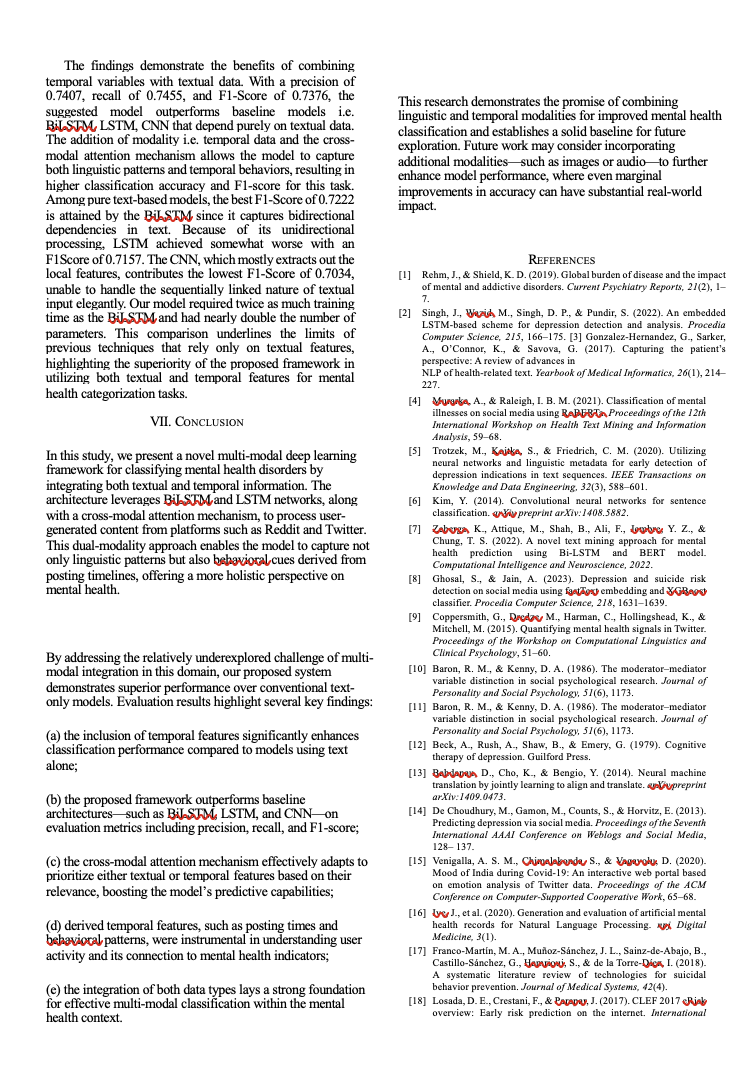
****

****

****

****

****

****

**5. Conclusion**

The analysis of social media behavior provides significant insights into the mental states of users. By leveraging NLP and sentiment analysis techniques, we were able to detect linguistic and temporal patterns that correlate with different mental health conditions. These findings can be instrumental in designing tools that assist healthcare providers in early identification of mental health issues.

Key takeaways:

* Social media content can reveal early symptoms of mental illness.
* Depressed and anxious users exhibit distinct language and sentiment profiles.
* Posting behavior varies by mental health state.

**6. Future Scope**

The potential of this work can be expanded in several directions:

**6.1 Multi-Modal Data Integration**

Include image, video, and interaction data (likes, comments) for richer context.

**6.2 Real-Time Monitoring**

Develop real-time monitoring systems to detect high-risk posts.

**6.3 Personalized Recommendations**

Offer personalized mental health content, support groups, or helpline links.

**6.4 Collaboration with Experts**

Work with psychologists to validate the model's predictions.

**6.5 Ethical and Privacy Considerations**

Establish ethical guidelines for data usage, ensuring user anonymity and consent.

**7. References**

1. Kaggle Datasets: https://www.kaggle.com
2. Nikhileswar Komati, Reddit Mental Health Dataset: https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch
3. TextBlob Documentation: https://textblob.readthedocs.io
4. VADER Sentiment Analysis: https://github.com/cjhutto/vaderSentiment
5. Scikit-learn Documentation: https://scikit-learn.org
6. SpaCy NLP Toolkit: https://spacy.io
7. World Health Organization (WHO) Mental Health Reports
8. Python Text Analysis Tools: NLTK, textstat
9. Research Paper: Resnik et al. (2015), "Beyond Labeled Data: Detecting Anxious Depression on Reddit"