MoodLens: Machine Learning for Social Media-Based Depression Analysis

DSF PT07 - Group 11 Capstone Project Proposal

Table of Contents

Group 11 Members:

1. Cecily Wahome
2. Geoffrey Shikanda
3. Phoebe Wawire
4. John Isaac Mwangi
5. Nduku Kiteng'e
6. Keziah Gicheha
7. Joseph Matheri

Business Understanding

Mental health is an urgent issue globally, with depression affecting millions of individuals across all demographics. According to the World Health Organization (WHO), over 264 million people worldwide are affected by depression. The internet, and particularly social media, has become a place where people often express their struggles, including depression. A study published in The African Journal of Psychiatry found that the prevalence of depression among adult Kenyans was approximately 30%. The research noted that factors such as unemployment and poverty were significantly associated with higher rates of depression.

Early detection of depression symptons can be crucial in providing timely support or intervention. This project's goal is to leverage data science to better understand and detect depressive expressions in online platforms, potentially paving the way for more proactive mental health support.

Problem: Can we use data science to differentiate between depressive and non-depressive content based on text patterns on Reddit?

Using data from Reddit posts in subreddits related to depression (such as *Depression* and *SuicideWatch*) versus other subreddits, the project will test if there are distinct linguistic markers or features that can reliably indicate depressive tendencies in reddit users.

By analyzing patterns in language, sentiment, and topics, the project aims to pinpoint the specific elements of language that correlate with depression, offering insights into commonly associated words, phrases, and tones.In addition, the project would like to explore a way to trigger positive messaging to reddit users with a high tendency of depressive statement.

The goal of the project is to build and validate a predictive model that can classify a post as depressive or non-depressive. The model's performance will provide insights into how accurately this kind of data can support early detection.

The project seeks to explore the potential of such models to support real-world applications, like alert systems for moderators or insights for mental health outreach.

Hypothesis Statement

The project will test the hypothesis that an increased number of negative posts could mean that the person is likely to end up depressed.

Given the widespread use of platforms like Reddit, this research could benefit individuals by increasing awareness and intervention opportunities.

Objectives

1. Develop a predictive model for classifying depressive vs. non-depressive posts.
2. Create visual insights into depressive language trends.
3. Deploy a user-friendly tool to analyze reddit posts.

This project cuts across several industries mainly:

- **Mental Health and Psychology**: Provides insights into depressive language and behavior patterns based on past research.

- **Social Media**: Supports the growing interest from the public in making platforms safer and more supportive of its users, eg. Movies such as ‘The Social Network’ have shed light on the ill effects on social media on mental wellbeing of its users.
- **Data Science and Machine Learning**: Utilizes machine learning modelling techniques such as Natural Language Processing (NLP), recommender systems, and classification.
- **Public Health**: Assists in identifying and supporting at-risk individuals on digital platforms.

Target Audience:

The primary audiences for this project are:

1. **Social Media Platforms:** such as Reddit, Twitter, Facebook.

   Role: These platforms could invest in the tool to improve user safety and offer mental health support on their platforms. As shareholders, they would have a vested interest in improving content moderation, detecting high-risk posts, and potentially redirecting users to mental health resources.

2. **Investors in AI and Social Impact Ventures:** such as Venture capital firms, impact investors, and foundations focused on social good (e.g., Omidyar Network, Social Capital).

   Role: Investors focused on technology that drives positive social impact may be interested in supporting a project that uses AI to address mental health. They would provide the necessary capital to scale and maintain the project, seeking a return on investment or social impact metrics as the tool gains adoption.

3. **Academic and Research Institutions:** such as Universities with psychology, psychiatry, data science, and social sciences departments.

   Role: Researchers and academic institutions could benefit from the project for studies related to mental health trends, language analysis, and social media behavior. They could act as collaborators for research, providing grants, or be involved in testing and improving the predictive algorithms.

If implemented, this project could help identify individuals at risk of depression in real time, potentially leading to faster intervention and support. Social media platforms could incorporate the model as a tool to provide resources or reach out to users who may be struggling through shifting the content that they distribute for the participants. Moreover, the project could further mental health research by providing a scalable method of analyzing depressive language patterns, thus broadening the understanding of mental health trends over time.

This project draws from a growing body of research in **Natural Language Processing (NLP)** and **Mental Health Analysis**. Relevant studies include:

- Research on language patterns in depressed individuals, which identifies common indicators such as specific word choices and tone.
- Existing models for **sentiment analysis and emotional tone classification** in social media data.
- Previous projects on **mental health detection using NLP and machine learning** in platforms like Twitter and Facebook.

The motivation for this project stems from recognizing the growing mental health crisis and the unique opportunity social media provides for understanding public health issues at scale. Given the role that platforms like Reddit play in individuals' daily lives, this project is motivated by the potential to use technology in a meaningful way—by improving the capacity to detect signs of depression and enabling more responsive support structures. This research aims to contribute toward making digital platforms safer, more supportive spaces, especially for users who might otherwise go unnoticed in their struggles.

Data Understanding

Datasource: https://www.kaggle.com/datasets/rishabhkausish/reddit-depression-dataset/data

The dataset was originally sourced from Kaggle and already includes several key features that can be used to analyze and predict depression indicators based on Reddit posts. Specifically, the data has 7 key columns:

- **subreddit**: The subreddit where each post was made, with posts from "Depression" and "SuicideWatch" labeled as 1 for depression and posts from other subreddits labeled as 0 (non-depression).
- **title**: The title of the Reddit post.
- **body**: The full text of the Reddit post, which may contain valuable information for understanding the context, tone, and possible indicators of depression.
- **upvotes**: Number of upvotes each post received, which may indicate the post's visibility or resonance with the community.
- **created_utc**: The timestamp of when the post was created in UTC, which can help in analyzing temporal trends.
- **num_comments**: The number of comments on each post, which could provide insights into community engagement.
- **label**: The target variable indicating depression (1) or non-depression (0) based on subreddit.

The raw data was collected from five Reddit subreddits (sub topics), categorized based on their content. These included: Teenagers, Depression, SuicideWatch, DeepThoughts, Happy

Since the data is already collected from Reddit, with over 6 million rows, further data acquisition may not be necessary. Infact, the team proposes reducing the dataset to about 100,000 rows for the purpose of this project. The team also managed to scrape Reddit for more recent posts using PRAW tool that we identified through recommendation from peers. To comply with Reddit's data collection policies and privacy standards we did not scrape any private user data.

In our exploration we identified that features in this dataset are clearly defined. Each provides specific information:

- **subreddit** serves as the source identifier, crucial for distinguishing between depressive and non-depressive content.
- **title** and **body** contain textual data, essential for natural language processing and sentiment analysis.
- **upvotes** and **num_comments** give insights into engagement, which may indirectly correlate with the severity or relatability of a post's content.
- **created_utc** enables time-based analysis for observing trends.
- **label** provides the classification for depression, necessary for supervised learning.

There has been prior research on identifying mental health indicators in social media platforms, particularly Reddit, using NLP and machine learning techniques. Other projects have worked on classifying depression using sentiment analysis, keyword detection, and predictive modeling.

Data Preparation

The data is stored in a CSV format. The preliminary data types are as follows:

- `subreddit`: String – indicates the subreddit from which the post originates.
- `title` and `body`: String – these contain unstructured text data.
- `upvotes` and `num_comments`: Integer/Float – count values that provide information on user engagement.
- `created_utc`: DateTime – Unix timestamp format that can be converted to a DateTime object.
- `label`:String (depression, non-depression).

Using a randomiser we reduced the main dataset to 98,826 posts The body column has some missing values (~20% missing), and the num_comments column also has some missing data (about 5% missing). All other columns are complete with no missing data, and the dataset uses a mix of numeric and object data types.With this view we will have to handle the missing values

The preprocessing steps for this dataset will include:

1. **Data Cleaning**: Handling any missing values (e.g., NaN in `num_comments`), outliers in `upvotes`, and any unexpected values.
2. **Text Preprocessing**: Tokenizing, removing stopwords, lemmatizing, and stemming for `title` and `body` columns. We may also apply techniques such as lowercasing, removing punctuation, and possibly handling slang or abbreviations specific to Reddit.
3. **Encoding Categorical Variables**: Encoding the `subreddit` variable if it's used as a feature in the model.
4. **Timestamp Conversion**: Converting `created_utc` from Unix timestamp to a readable DateTime format and extracting components like day, month, or hour to observe trends or patterns over time.
5. **Scaling Numeric Variables**: If using models sensitive to scale, we may standardize or normalize `upvotes` and `num_comments`.
6. **Balancing the Target Variable**: If the data is imbalanced, applying techniques like SMOTE or undersampling to balance the `label` distribution.

Some anticipated challenges include:

- **Text Complexity**: Reddit posts may contain slang, emojis, abbreviations, and misspellings that require more advanced preprocessing or specific handling.
- **Class Imbalance**: If there is a disproportionate number of posts labeled as 0 (non-depression) versus 1 (depression), this could skew model performance, requiring careful balancing.
- **Noisy Data in Text**: Posts may contain irrelevant information, such as URLs, usernames, or quotes from other users, which could add noise and reduce model accuracy.
- **Sparse Engagement Metrics**: Many posts may have minimal `upvotes` or `num_comments`, limiting the information available from these features.

We estimate to have about 100,000 rows from the dataset, based on the team's computing capacity.

We plan to visualise the data using the following tools:

1. **Target Distribution**: A bar chart or pie chart to visualize the proportion of depressive (label=1) vs. non-depressive (label=0) posts.
2. **Subreddit Distribution**: A bar chart showing the number of posts per subreddit to understand data representation across subreddits.
3. **Engagement Metrics**: Histograms or box plots for `upvotes` and `num_comments` to observe their distribution and identify any outliers.
4. **Time-based Trends**: Line plots to analyze posting frequency over time, which might reveal patterns related to specific days or times.
5. **Text Analysis**: Word clouds or bar charts showing the most frequent words in posts labeled as depressive vs. non-depressive.

6. **Sentiment Analysis**: If sentiment scores are calculated, visualizing the sentiment distribution within each label could provide insights into common emotional expressions in depressive posts.

Modeling

Given the nature of the data and the task of classifying Reddit posts as depressive or non-depressive, **Natural Language Processing (NLP)** models are appropriate for this problem. Some suitable techniques include:

1. **Logistic Regression**: As a simple baseline model that can be used after converting text data to numerical features using TF-IDF or CountVectorizer.
2. **Naive Bayes**: Effective for text classification and can handle high-dimensional sparse data well.
3. **Support Vector Machine (SVM)**: Known for its performance in text classification tasks, especially in cases where the dataset may be imbalanced.
4. **Random Forest or Gradient Boosting**: These ensemble methods can provide robust predictions and handle nonlinear relationships well.

The target variable for our analysis is `label`, where:

- 1 = depressive posts (from Depression and SuicideWatch subreddits)
- 0 = non-depressive or neutral posts (from Teenagers, DeepThoughts, and Happy subreddits)

The baseline model will likely be **Logistic Regression**. These models are relatively simple and quick to train, for establishing initial performance before moving to more complex models if needed. This is a **classification problem**, as the task is to predict a binary outcome (depressive or non-depressive) based on the text data in each Reddit post.

To evaluate the model's performance, the following metrics will be utilized:

- **Accuracy**: The proportion of correctly predicted posts (both depressive and non-depressive).
- **Precision**: The ratio of true positive predictions to the total predicted positives, indicating how many of the predicted depressive posts were actually depressive.
- **Recall (Sensitivity)**: The ratio of true positive predictions to the actual number of positive cases, indicating how many actual depressive posts were correctly identified.
- **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two.

- **ROC-AUC Score**: To evaluate the model's ability to distinguish between the two classes across different thresholds.

The MVP will involve building a basic classification model that can predict whether a Reddit post is depressive or non-depressive based on text content. This smaller project can be accomplished within a week by:

- Collecting and preprocessing the dataset.
- Implementing a simple model (e.g. Logistic Regression).
- Evaluating the model using accuracy and confusion matrix.

The steps that will be used to enhance the model from MVP include:

- **Enhance the Model**: Experiment with more models (e.g., SVM and Random Forest) for improved accuracy.
- **Feature Engineering**: Implement further feature engineering depending on results form initiail MVP model.
- **Visualizations**: Create additional visualizations to present data insights, model performance, and classification results.
- **User Interface**: Develop a more sophisticated web application for user interaction and display of results, allowing users to input text and see real-time predictions.

Deployment

The final results will be reported through a combination of visual dashboards showcasing the model's performance metrics, confusion matrix, and accuracy metrics. Additionally, a summary report will detail the analysis, findings, and insights gained from the project.

The team will deploy a **web application** that allows users to enter Reddit post text and receive predictions on whether the post is likely to indicate depression. This application will be built using Streamlit for Python.

The web application will have the following functionalities:

- A user-friendly text input box for users to paste or type Reddit post content.
- A "Predict" button that triggers the model to analyze the input.
- Display of the prediction result (depressive or non-depressive) along with confidence scores.
- Optional visualizations to show the distribution of predictions and relevant statistics about the model's performance.

Tools/Methodologies

The team proposes to use the following tools to carry out the project:

- **Data Gathering and Cleaning**: `pandas`, `numpy`,for text preprocessing.
- **Exploratory Data Analysis**: `matplotlib`, `seaborn` for visualizations.
- **NLP Processing**: `nltk`, `scikit-learn` for feature extraction and model training.
- **Sentiment analysis:** VADER sentiment analysis
- **Modelling algorithm**
    - Logistic Regression
    - Naive Bayes
    - Support Vector Machine (SVM)
    - Random Forest
    - BERT

The analysis will initially be performed on a local machine, using Jupyter Notebook for ease of experimentation. As the project develops, it may transition to a cloud environment like Google Colab for enhanced computational resources

The data will initially be stored locally on the machine for development and testing. If the project scales or requires collaboration, it will be stored in the cloud (e.g. or Google Cloud Storage-through Colab) for easier access and sharing among team members.
The Jupyter notebooks will be hosted on a github repository which will be updated on a regular basis.