PREDICTING MEDICAL CONDITIONS FROM REDDIT POSTS

Gaurav Srivastava, Arnab Roy, Om Chouhan, Priyamjeet Kumar

IIT Ropar

ABSTRACT

Modern day social media is filling up with discussions revolving around one's or a close one's health and such a platform can be of significant help when it comes to pre-diagnosis. This study applies machine learning approach to make a discovery about medical situation by analyzing text data from Reddits. Thus, the exploratory analysis of four ML models: Logistic Regression, LSTM, Decision Tree, and Random Forest was performed with the help of feature extraction using TF-IDF and Word2Vec. Logistic Regression had the highest accuracy score of 68.7/100; LSTM since it is a model that works best for sequential input data had an accuracy of 71.8/100. The analysis shows that more basic algorithms, such as Logistic Regression, and sequence-sensitive models like LSTM, are feasible for health data categorization from social media. Further research can focus on using transformer-based models as well as other domain embeddings in order to obtain even higher results with the presented methods of analysis The proposed approaches can be applied to developing real-time health monitoring and early warning systems in the future.

Index Terms— Machine Learning, Medical Condition Prediction, Social Media, Reddit, Text Classification

1. INTRODUCTION

In the last couple of years, social networks have emerged as one of the primary means by which people seek information and share experience of a medical nature, including symptoms, personal cases, and illnesses. Reddit being a forum-based social platform brings ample of user generated information content associated with different health problems. The privacy of Reddit engages the users to personal healthcare questions and answers enabling students and researchers in public health to gather sufficient data for analyzing epidemics and even may be used to identify potential ailments at their initial stage.

This work aims to leverage the machine learning models to predict diseases from Reddit post data. In other words, the models is intended to automatically identify and classify health conditions by considering the patterns within these posts within the context of their presence in social media platforms for real-time health monitoring and early surveillance. It can be very beneficial for healthcare managers, scientists, and legislators to know such things in order to potential future health problems and optimize preventative steps. The techniques employed, the analyses carried out and the findings generated should validate the possibility of this line of work and its potential application in the coming sections of this paper.

This work was supported by IIT Ropar.

2. METHODOLOGY

The methodology for this project involved several stages: data preprocessing, feature extraction, and model training. This section provides an overview of the techniques and tools used for each stage.

2.1. Dataset Description

In this study the dataset is formed of Reddit posts related to different medical conditions and the post had been labeled expressing a medical condition or a set of symptoms. As the basis of training a model to learn to classify patient conditions based on text, this dataset is unique in that it allows computation of unstructured social media text for healthcare purposes.

2.2. Data Preprocessing

Given the unstructured nature of Reddit posts, data preprocessing was crucial. Steps included cleaning the text to remove irrelevant characters, standardizing formatting, and filtering out stop words. Special attention was given to preserving medically relevant terms, as these play a critical role in classification.

2.3. Feature Extraction

For text representation, we employed two main vectorization techniques:

- **TF-IDF (Term Frequency-Inverse Document Frequency)**: This was applied to transform textual data into the numerical form, evaluating the importance of each term compared to the entire dataset. TF-IDF is a type of term frequency, which captures how many times each term appears, and emphasizes the words specifically to those posts.

- **Word2Vec**: We also used Word2Vec embeddings to capture contextual relationships. This permits the model to capture the meaning of the words, i.e., express the semantic similarities between different words, especially when analyzing medical text including novel terms.

2.4. Model Selection and Training

We experimented with multiple models to identify the best-performing approach for predicting medical conditions. These models included:

- **Logistic Regression**: Applied transformation on the text data in the form of TF-IDF vectorization. From the above vectorized data, trained a logistic regression model. Used to predict on the test set and calculated accuracy, precision, recall then portrayed with a confusion matrix.
- **Decision Tree**: Word2vec were utilized to convert the textual features preprocessed from the text data to the vectors. They also used these vectorized embeddings to train a decision tree classifier. Used the calculated model accuracy, precision, recall and visualized the performance of the model through confusion matrix.
- **Random Forest**: For inputto the model we used Reused Word2Vec for the input features. Trained a model where each tree is a decision tree and the estimator for the random forest used is 100. We obtained accuracy, precision, recall and have also built a confusion matrix.
- **LSTM (Long Short-Term Memory)**: Converted Word2Vec embeddings to the format to be fed into LSTM in as a sequence of vectors. Built and trained an LSTM model architecture using dropout layers at random time step and a full connection layer for classification. Doing an assessment of the accuracy, precision or recall and visualizing the results by using the confusion matrix.

Each model was trained on vectorized data generated by TF-IDF and Word2Vec to learn patterns associated with each medical condition label. Using this diverse set of models allowed us to compare performance and identify the most effective approach for our classification task.

2.5. Evaluation Metrics

To assess the performance of each model, we used several evaluation metrics, including accuracy, precision, and recall. These metrics provided a comprehensive evaluation of the models' classification abilities, allowing us to gauge their potential utility in early health detection from social media text.

3. RESULTS

3.1. Classification Performance

3.1.1. Logistic Regression

The Logistic Regression model using TF-IDF has achieved accuracy of approxx 68.7%, with weighted precision, recall, and F1-score around 0.68 each. It performed well on distinct categories but encountered difficulties with classes that had overlapping symptoms.

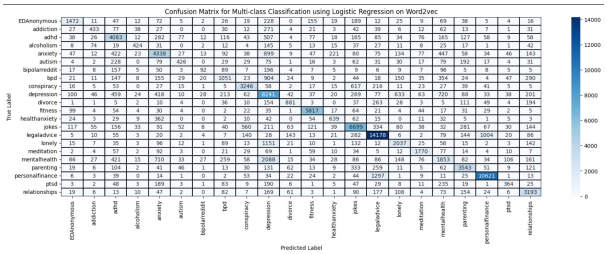


Fig. 1. Confusion Matrix for Logistic Regression on TF-IDF

```
Accuracy1_W2V: 0.6868122859433069
Weighted Precision1_W2V: 0.6808
Weighted Recall1_W2V: 0.6868
Classification Report 1_W2V:
                precision    recall  f1-score   support

    EDAnonymous       0.69      0.60      0.64      2435
       addiction       0.57      0.37      0.45      1156
            adhd       0.61      0.67      0.64      6072
      alcoholism       0.67      0.46      0.55       921
         anxiety       0.60      0.60      0.60      7282
          autism       0.56      0.31      0.40      1355
   bipolarreddit       0.43      0.12      0.18       790
             bpd       0.46      0.31      0.37      3370
      conspiracy       0.73      0.73      0.73      4444
      depression       0.52      0.70      0.60     11822
         divorce       0.64      0.49      0.56      1801
         fitness       0.89      0.92      0.90      6306
   healthanxiety       0.58      0.49      0.53      1308
           jokes       0.71      0.72      0.72      9251
     legaladvice       0.82      0.87      0.85     16252
          lonely       0.58      0.53      0.55      3880
      meditation       0.76      0.78      0.77      2282
    mentalhealth       0.40      0.29      0.34      6433
       parenting       0.71      0.73      0.72      4871
  personalfinance       0.88      0.87      0.87     12222
            ptsd       0.51      0.28      0.36      1282
   relationships       0.65      0.75      0.70      4249

        accuracy                           0.69    109784
       macro avg       0.64      0.57      0.59    109784
    weighted avg       0.68      0.69      0.68    109784
```

Fig. 2. Classification Report for Logistic Regression on TF-IDF

3.1.2. Decision Tree

With an accuracy of 44.3%, the Decision Tree model showed poor generalization on the test data. This highlights its limitations in handling high-dimensional, unstructured text data.

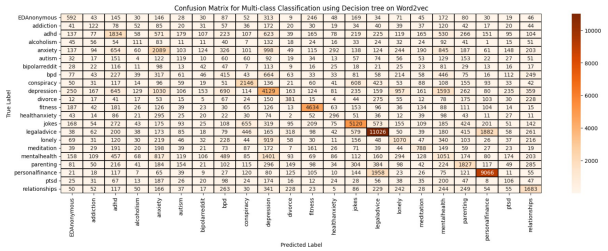


Fig. 3. Confusion Matrix for Decision Tree on Word2Vec

Accuracy2_W2V: 0.44311557239670624
Weighted Precision2_W2V: 0.4478
Weighted Recall2_W2V: 0.4431
Classification Report 2_W2V:

	precision	recall	f1-score	support
EDAnonymous	0.26	0.24	0.25	2435
addiction	0.10	0.11	0.10	1156
adhd	0.30	0.30	0.30	6072
alcoholism	0.13	0.12	0.12	921
anxiety	0.29	0.29	0.29	7282
autism	0.08	0.09	0.09	1355
bipolarreddit	0.05	0.05	0.05	790
bpd	0.12	0.12	0.12	3370
conspiracy	0.49	0.48	0.49	4444
depression	0.36	0.35	0.35	11822
divorce	0.20	0.21	0.21	1801
fitness	0.73	0.73	0.73	6306
healthanxiety	0.22	0.23	0.23	1308
jokes	0.61	0.55	0.58	9251
legaladvice	0.69	0.68	0.68	16252
lonely	0.27	0.28	0.27	3880
meditation	0.35	0.35	0.35	2282
mentalhealth	0.15	0.16	0.16	6433
parenting	0.36	0.38	0.37	4871
personalfinance	0.75	0.74	0.74	12222
ptsd	0.08	0.08	0.08	1282
relationships	0.39	0.40	0.39	4249
accuracy			0.44	109784
macro avg	0.32	0.32	0.32	109784
weighted avg	0.45	0.44	0.45	109784

Fig. 4. Classification Report for Decision Tree on Word2Vec

3.1.3. Random Forest

The Random Forest model achieved an accuracy of 64.3%, with some improvements over the Decision Tree model. However, it still struggled with complex, nuanced categories due to limitations in capturing word order and semantic dependencies.

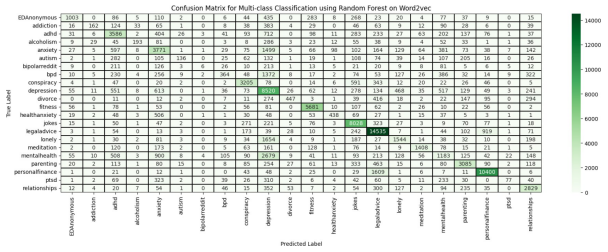


Fig. 5. Confusion Matrix for Random Forest on Word2Vec

Accuracy3_W2V: 0.6430900677694382
Weighted Precision3_W2V: 0.6428
Weighted Recall3_TFIDF: 0.6431
Classification Report 3_W2V:

	precision	recall	f1-score	support
EDAnonymous	0.74	0.41	0.53	2435
addiction	0.66	0.14	0.23	1156
adhd	0.52	0.59	0.55	6072
alcoholism	0.71	0.21	0.32	921
anxiety	0.48	0.52	0.50	7282
autism	0.62	0.10	0.17	1355
bipolarreddit	0.33	0.01	0.01	790
bpd	0.48	0.11	0.18	3370
conspiracy	0.70	0.72	0.71	4444
depression	0.43	0.72	0.54	11822
divorce	0.72	0.25	0.37	1801
fitness	0.84	0.90	0.87	6306
healthanxiety	0.68	0.33	0.45	1308
jokes	0.72	0.87	0.79	9251
legaladvice	0.76	0.89	0.82	16252
lonely	0.56	0.40	0.47	3880
meditation	0.78	0.62	0.69	2282
mentalhealth	0.32	0.18	0.23	6433
parenting	0.67	0.63	0.65	4871
personalfinance	0.87	0.85	0.86	12222
ptsd	0.59	0.06	0.11	1282
relationships	0.61	0.67	0.64	4249
accuracy			0.64	109784
macro avg	0.63	0.46	0.49	109784
weighted avg	0.64	0.64	0.62	109784

Fig. 6. Classification Report for Random Forest on Word2Vec

3.1.4. LSTM

The LSTM model reached an accuracy of 71.8%, effectively balancing performance across categories by capturing sequential dependencies. This model was particularly adept at handling context-rich data and showed promise in nuanced classification tasks.

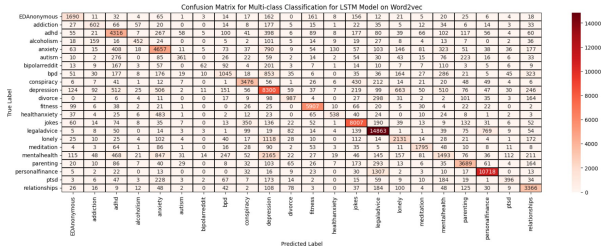


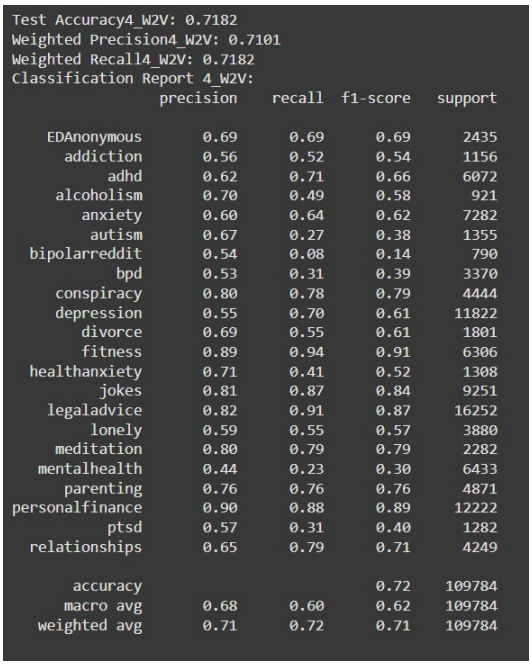
Fig. 7. Confusion Matrix for LSTM on Word2Vec

3.2. Comparison of Model Performance

Table 1 summarizes the performance metrics across all models. Logistic Regression and LSTM demonstrated superior performance, with Logistic Regression providing strong baseline accuracy and LSTM excelling in capturing sequential patterns in text.

4. CONCLUSION

This paper also shows that it is possible to use machine learning algorithms to predict medical conditions from Reddit post



```
Test Accuracy4_W2V: 0.7182
Weighted Precision4_W2V: 0.7101
Weighted Recall4_W2V: 0.7182
Classification Report 4_W2V:
                precision    recall  f1-score   support

    EDAnonymous       0.69      0.69      0.69      2435
      addiction       0.56      0.52      0.54      1156
           adhd       0.62      0.71      0.66      6072
     alcoholism       0.70      0.49      0.58       921
        anxiety       0.60      0.64      0.62      7282
         autism       0.67      0.27      0.38      1355
  bipolarreddit       0.54      0.08      0.14       790
            bpd       0.53      0.31      0.39      3370
     conspiracy       0.80      0.78      0.79      4444
     depression       0.55      0.70      0.61     11822
        divorce       0.69      0.55      0.61      1801
        fitness       0.89      0.94      0.91      6306
  healthanxiety       0.71      0.41      0.52      1308
          jokes       0.81      0.87      0.84      9251
   legaladvice        0.82      0.91      0.87     16252
         lonely       0.59      0.55      0.57      3880
     meditation       0.80      0.79      0.79      2282
   mentalhealth       0.44      0.23      0.30      6433
      parenting       0.76      0.76      0.76      4871
personalfinance       0.90      0.88      0.89     12222
           ptsd       0.57      0.31      0.40      1282
  relationships       0.65      0.79      0.71      4249

       accuracy                           0.72    109784
      macro avg       0.68      0.60      0.62    109784
   weighted avg       0.71      0.72      0.71    109784
```

Fig. 8. Classification Report for LSTM on Word2Vec

Table 1. Comparison of Model Performance Metrics

Model	Accuracy	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1-Score
Logistic Regression	0.687	0.68	0.69	0.68
Decision Tree	0.443	0.45	0.44	0.45
Random Forest	0.643	0.64	0.64	0.62
LSTM	0.718	0.71	0.72	0.71

using Logistic Regression and LSTM as the most promising models. Logistic Regression with an accuracy of 68.7% functioned well as a baseline model along with LSTM because of its sequence learning abilities and achieved an accuracy of 71.8 % in the experiment. However, the problem of distinguishing between the fine categories is crucial to Decision Tree and Random Forest models which seek a simpler approach, but these results show that the simpler models can perfectly work and further, more complex models like sequence-based LSTM have their merits when it comes to dealing with the language sequence. Further research can be conducted based on transformer models and specialty-based content vectors for generalization with an avenue for application in real-time health assessment and prevention systems.