Literature review

Depression Detection from textual data using Machine learning Models.

Abstract **:** Depression is one of those mental health problems that can have very serious consequences if untreated and causes the patient a lot of suffering. Early identification of depressive symptoms is a crucial step towards assessment, intervention and relapse prevention. Research suggests that the way people write can reflect mental well-being and mental health risks, individuals upbringing and emotions. In this literature review we will discuss the work done in this field and the machine learning techniques which were adopted.

Introduction

Mental health problems such as anxiety and depression affect millions of people especially after Covid-19 pandemic. Even then people might not seek medical attention in time due to the stigma around it or other reasons, leading to worsening of the condition and creating even more damaging results. These individuals often tend to substance abuse, or other various forms of self harm. Thus detection at an early stage becomes crucial.

Language is a tool we use to communicate with one another, oftentimes we not only communicate the intended message but also transmit information about ourselves : our upbringing, our mood, our emotional well-being etc. It is statistically proven that there is a correlation between differences in the way people talk and write, and having a mental health condition. In machine learning, use of Natural Language Processing to detect untreated health problems is seen quite often.

The social media sites such as X, Reddit, Instagram present a natural collection of user-generated texts. This data can result in excellent training of models and building good detection models.

SECTION 1 : UNED-MED at eRisk 2022: depression detection with TF-IDF, linguistic features and Embeddings

Method used

In this paper the authors proposed two approaches :

1. Feature-driven classifiers with features based on text data, TF-IDF terms, first-person pronouns use, sentimental analysis and depression terminology.
2. A Deep Learning classifier with pretrained Embeddings.

Dataset Description

The authors used two different datasets:

1) The official eRisk 2022 shared task 2 dataset.

This dataset comprises a collection of documents. Each document contains the post submission history of a user from Reddit. The users are labelled as either Positive (at risk of depression) and Negative (control group). On observing the dataset, the authors found some key notes. The dataset was such that it was deeply unbalanced and an observation proved that the negative users write shorter posts compared to the positive users even though the longest post was written by a negative user. This data also contained emojis, loose grammar, incorrect spellings and links to other posts or websites.

2) The UNED-MED depression Reddit dataset.

The authors used PRAW Python Reddit API Wrapper to extract the data, they applied search strategies which included targeting the specific words to get the best

messages.Specific search queries were used to identify users who had self-reported clinical depression diagnoses in various subreddits related to mental health. The resulting dataset comprised 235 users, with statistics resembling those of the positive users in the original eRisk dataset, enhancing the training data for depression detection.

Proposed Models

Traditional machine learning techniques, Random Forest and XGBoost while deep learning model CNN were suggested. Here the features for model 1 and 2 were a combination of TF-IDF and text-based and features for model 3 were Embeddings.Each model is bifurcated into three stages: data pre-processing, features, and classification. The models take one message by one user and predict if that user is at risk of depression (1) or not(0).

In the context of training data for depression detection, three distinct training sets were generated. The first, known as the "Original eRisk," combined the eRisk 2022 shared Task 2 training and test datasets. The second, termed "Augmented eRisk," incorporated the UNED-MED 2022 depression Reddit dataset into the Original eRisk training set. The third set, labelled "Relabeled eRisk," employed a relabelling approach based on sentiment analysis on the Original eRisk training set.

For the "Relabeled eRisk" dataset, it was observed that labels in the Original eRisk dataset were applied at the user level, signifying that all posts from a positive user were marked as positive and vice versa. To address the hypothesis that not all posts by users at risk contain relevant information for early risk detection and that training with all such posts marked as positive could decrease system performance, a strategy was devised. In this strategy, only posts labelled as positive were reclassified based on their sentiment analysis. Posts with a negative sentiment analysis above a particular threshold retained their positive classification, while others were reclassified as negative. This reclassification strategy was implemented using a twitter-XLM-roBERTa-base model trained on tweets and fine-tuned for Sentiment Analysis.

Preprocessing & Windowfying

In the data preprocessing phase, a standard text preprocessing approach was employed on the text extracted from each post. This process involved several steps to ensure that the text was suitable for analysis. Specifically, the following preprocessing steps were applied:

1. Cleaning with reddit cleaner: The Python library "reddit cleaner" was utilised to clean the textual data. This cleaning process involved removing Markdown formatting, separating contractions, eliminating hyperlinks, HTML tags, numbers, and multiple spaces. Additionally, all text was converted to lowercase for consistency.

2. Retaining Stop Words: In contrast to traditional text preprocessing practices, stop words were retained as part of the text. This decision was made based on the belief that stop words might hold importance for the specific task of depression detection.

To address variations in text length across posts, a technique known as "windowfying" was applied. Some posts were considerably long, while others were exceptionally short. To mitigate this discrepancy and ensure that a substantial amount of text was processed in each step without compromising computational efficiency, a sliding window approach was employed. After the initial cleaning, the text from a post was combined with the text from its preceding messages within a configurable window size parameter

(denoted as "w"). Subsequently, the calculated features were computed on this message window, rather than solely on the text of the individual post. This windowfying process helped standardise the length of text input for analysis, ensuring that adequate information was considered while maintaining processing efficiency.

Features

In feature engineering, two distinct strategies were employed depending on whether the classifier algorithm belonged to traditional machine learning (models 1 and 2) or deep learning (model 3).

Traditional Features (for models 1 and 2):
For traditional machine learning algorithms, the features used were adapted from those utilized in the eRisk 2021 Task 2. These features were categorized as follows:

1. Text-Based Features: This category included two features: text length and the number of words in a message. These features were chosen because earlier in the study, it was observed that positive users tend to write longer texts. The information about text length and word count was retained with these features. Additionally, these features were normalized by text length and discretized into a fixed number of bins to make them suitable for modeling.

2. First-Person Pronouns: Research has shown that individuals with mental health issues, such as depression, tend to use more first-person pronouns in their speech. A feature was created to count the occurrences of first-person singular pronouns in a text.

3. Depression-Related Words: In previous iterations of the shared task, a word set containing self-harm-related terms was used as a feature. However, in this year's approach, a collection of words related to clinical depression, along with the moods and topics associated with it, were extracted from a source. This feature counted the number of depression-related words that appeared in a text.

These features were combined with TF-IDF-based features using Scipy Hsparse, enhancing the feature representation for traditional machine learning models.

TF-IDF Features (for models 1 and 2):
A TF-IDF featurizer was trained on the positive users' data and then applied to obtain TF-IDF features for each message window. Importantly, the featurizer was exclusively trained on positive data (or, in the case of the relabeled dataset, on messages that remained positive) to capture words frequently used by positive users.

Embeddings (for model 3):
In the case of the deep learning model (model 3), word embeddings were employed. Specifically, Stanford's pre-trained GloVe embeddings from Wikipedia 2014 with 100 dimensions were used. To prepare the data for embedding-based modeling, posts were windowfied and then padded to achieve a sufficiently long length that accommodated the longest messages before applying the embeddings. This approach allowed the deep learning model to capture semantic information and context within the text data, enhancing its ability to detect depression-related patterns.

Classifier Algorithms

In this study, three different classifiers were employed to predict whether a message window belonged to a user at risk of being "positive" or "negative," with the understanding that a positive decision was

considered final, whereas a negative decision could potentially be revised later. The classifiers used were as follows:

1. Random Forest: The Random Forest classifier was utilized, and the implementation was based on scikit-learn. Random Forest is an ensemble learning method known for its robustness and ability to handle complex data. It combines multiple decision trees to make predictions and is widely used in classification tasks. It was used to build predictive models for depression detection.

2. XGBoost: XGBoost, an ensemble learning technique, was employed as the second traditional machine learning model. XGBoost optimizes a distributed gradient boosting framework and is known for its high performance in various machine learning competitions. It was used to create predictive models for identifying users at risk of depression.

3. Convolutional Neural Network (CNN): A deep learning model, specifically a Convolutional Neural Network (CNN), was implemented using Keras. The CNN architecture consisted of a convolutional layer with 64 units, followed by a GlobalMaxPooling layer to reduce dimensionality, a Dense layer with ReLU activation for feature processing, and an output Dense layer with sigmoid activation to make binary classification decisions. CNNs are well-suited for capturing spatial patterns in data, making them appropriate for text classification tasks, including depression detection based on textual data.

These classifiers were trained and evaluated using the features and data preprocessed as described earlier, with the goal of accurately classifying message windows as either "positive" or "negative" with respect to the risk of depression. The use of both traditional machine learning models and a deep learning model allowed for a comprehensive exploration of different approaches to address the depression detection task.

Training strategy was such that descending training weights were given to positive posts. This was in order for the system to prioritise earlier messages and detect the positive users (at risk of depression) as fast as possible.Messages created by negative users were all assigned the same training weight (1). While messages created by positive users were assigned descending weights with a fixed rate (2 to 1) from oldest to most recent message.

Results

The experimental setup was such that there were 5 different runs constructed. Each run had different combination of training data, classifier and training and test window size. All runs used weighted training. The authors say, their best-performing model achieved a latency-weighted F1 score of 0.233, whereas the top-performing group, NLPGroup-IISERB, achieved 0.690.

Comparing the runs, it was found that run 1 and run 3 performed the best across various metrics. Run 1 employed a Random Forest model trained on the augmented dataset with a feature window size of 10, while run 3 used an XGBoost model trained on the relabeled dataset with a feature window size of 100.

Their findings suggest that strategies to enhance the training dataset, such as using the augmented dataset with more positive users, were effective. Smaller feature window sizes seemed to yield better results, as indicated by the ranking of the runs. However, using smaller window sizes during training may not be necessary, and exploring further combinations was challenging due to constraints.

The Deep Learning model gave the worst results in run 4 which makes one wonder why that happened, since usually Deep Learning

models perform better than the traditional models in similar circumstances. The sliding window size was 10 which was the same as run 1 and run 3, and the latency value was also very high comparatively. Thus error during implementation can be concluded.

Conclusion

The authors of this study developed multiple classifier models, including those based on TF-IDF, text, and specially-tailored features, as well as a Deep Learning classifier model utilizing embeddings. To address the imbalance in the training data, additional data was obtained from Reddit, and the original training data was relabeled. The test results indicate that the systems achieved modest performance, suggesting the need for further efforts to reach state-of-the-art results.

Future research directions may involve continued exploration of data relabeling strategies and potential experimentation with zero-shot learning techniques. This approach could enhance the system's portability across different disease detection tasks with minimal adjustments.

SECTION 2 : Deep learning for prediction of depressive symptoms in a large textual Dataset

The study introduces an efficient method utilizing Long Short-Term Memory (LSTM)-based Recurrent Neural Networks (RNN) to identify text messages that describe self-reported symptoms of depression. This approach is applied to a substantial dataset gathered from a public online resource for young individuals in Norway, comprising text-based questions posed by youth. Instead of relying solely on word frequencies, the research extracts robust features from reflections of potential depression symptoms predefined by medical and psychological experts, offering a more effective feature set. The RNN-based deep learning technique is then employed to learn sequential features that distinguish texts describing depression symptoms from those without such descriptions. The resulting model demonstrates superior performance when compared to conventional methods.

Additionally, the study emphasizes the robustness of the extracted features by showcasing their effectiveness in clustering. These features, rooted in potential depression symptoms, enable the model to provide meaningful explanations for its decisions using an explainable Artificial Intelligence (XAI) algorithm called Local Interpretable Model-Agnostic Explanations (LIME). The proposed feature-based approach centred on depression symptoms outperforms traditional methods that prioritize word frequency, indicating that focusing on specific depression symptoms yields better results. While the study was conducted on a Norwegian dataset, it suggests that a similar approach could be applied to datasets in other languages with appropriate annotations and symptom-based feature extraction, contributing to the development of more advanced mental health technologies like intelligent chatbots.

Methods

The authors expressed their views on various models :
Deep Neural Networks (DNNs): The text mentions that deep neural networks have been instrumental in various research fields, including pattern recognition and AI. While it highlights their robustness, it also mentions two main disadvantages: overfitting and lengthy model training.

Deep Belief Network (DBN): The text briefly references the deep belief network as one of the first successful deep learning algorithms. DBN consists of Restricted Boltzmann Machines (RBMs) and is known for faster training compared to some earlier approaches.

Convolutional Neural Networks (CNNs): CNNs are discussed as popular models, especially in image processing, due to their ability to extract features alongside training data. CNNs are effective for image and video pattern analysis but are noted as less commonly applied to temporal data analysis.

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM): RNNs, particularly those with LSTM units, are highlighted as suitable choices for sequential data and pattern analysis. LSTM is mentioned as a solution to address the vanishing gradient problem in handling high-dimensional and time-sequential data.

Machine Learning for Analyzing Human States: The text mentions the widespread use of machine learning for analyzing physical and mental states of human beings from various data sources, citing references [35–41]. This suggests that the paper's context may involve the application of machine learning in analyzing human states.

Explainable Artificial Intelligence (XAI): The importance of explainable AI (XAI) is emphasized in the text to ensure transparency and trust in machine learning-based decisions. Several state-of-the-art explanation algorithms, including LIME, SHAP, and LRP, are mentioned as methods for explaining the behavior of black-box models.

Local Interpretable Model-Agnostic Explanations (LIME): LIME is specifically mentioned as a lightweight and quick post-hoc explanation algorithm that attempts to generate satisfactory explanations for model decisions.

Data Collection and processing

The data processing in this study involved obtaining a large text-based dataset from a public Norwegian information website called "ung.no." This website provides a platform for youth to post questions anonymously in Norwegian about various challenges and problems they face in their everyday lives. In response, professional experts such as doctors, psychologists, and nurses provide answers and advice. These questions and answers are publicly available online.

The specific steps for data processing are as follows:

1. Data Collection: The researchers obtained text data from the "Mental health and emotions" category on the ung.no website. This category focuses on topics related to mental health and emotional well-being. Users of the website, typically young people, pre-define and categorize the topics of their posts within this category.

2. Data Categorization: The text data collected from the "Mental health and emotions" category were categorized into several subgroups:

a. Depressive Conditions: Some of the texts describe depressive conditions that have already been diagnosed by health professionals. These texts likely contain descriptions of individuals' experiences with diagnosed depression.

b. Narratives and Symptoms: Many texts describe narratives and the ensuing symptoms experienced by individuals. These texts may either inquire if the described symptoms could represent depression or suggest depression as a possible diagnosis. The researchers consider these texts to represent self-perceived depressive symptoms.

c. Narratives without Mention of Depression: Some texts describe narratives and the resulting mental states without explicitly mentioning depression. These texts provide insights into individuals' experiences without a specific focus on depression.

3. Data Preprocessing: While the text data collected from the website may be relatively short, they typically contain information about the factors that trigger mental states and the resulting symptoms and behaviors. The data may also include information related to clinical diagnoses and self-perceived symptoms.

Overall, the data processing involved categorizing the text data into different groups based on the nature of the content, with a specific focus on identifying self-perceived depressive symptoms. This categorized dataset likely served as the foundation for further analysis and modeling in the study to develop a reliable model for detecting mental health issues, particularly symptoms related to depression.

Accordingly, the data is classified into categories, depression being one of them. Then, a trained GP went through the posts, confirming descriptions of depressive symptoms. A list of sentences and words are summarized analyzing the messages in the database where they may indicate the person having depression. A medical practitioner validated the sentences and words.

Linear Discriminant analysis for visualization : To visualize different features, linear discrimination analysis was adopted by the authors. LDA is nothing but an eigen value decomposition problem trying to maximize the inter-class scatterings of them.
The graphs authors visualized likely show visualizations of feature spaces obtained using different techniques, including LDA and Principal Component Analysis (PCA). These visualizations help illustrate the effectiveness of the proposed features in terms of sample clustering and class separation.

Deep recurrent neural network (RNN) for modeling emotional states

Emotional states in text data, as expressed during conversations, can be seen as a sequence of words that unfold over time. Therefore, a machine learning model capable of handling time-sequential data is well-suited for this task. In this context, the study opts for Recurrent Neural Networks (RNNs), a popular deep learning technique known for effectively modeling sequential information. RNNs are characterized by recurrent connections that link past to present states and hidden states, emphasizing the importance of memory in neural networks.

However, traditional RNNs often encounter a challenge known as the "vanishing gradient problem," particularly when processing long-term data with extensive dependencies over time. To address this issue, the Long Short-Term Memory (LSTM) architecture was introduced. Figure 10 provides an illustration of a sample deep neural network consisting of 50 LSTM units. Each LSTM memory block includes a cell state and three gates: the input gate, the forget gate, and the output gate.

In summary, the study utilizes RNNs, particularly LSTM units, to model emotional states expressed in text data over time. LSTM's design, with its cell state and gates, helps address the vanishing gradient problem commonly encountered when dealing with long-term dependencies in sequential data.

Conclusion and Comparison with Traditional models

This study presents a robust depression prediction system based on symptom-based features and a time-sequential LSTM-based machine learning model. The research utilizes a dataset of text samples from a public Norwegian information website, distinguishing between depression and non-depression texts. Through tenfold experiments, the proposed approach demonstrates superior accuracy, achieving 98% mean accuracy over traditional

machine learning models and achieving remarkable performance.

In addition to comparing favorably against traditional approaches, the study explores a rule-based classification method based on symptom presence, achieving an 84.20% accuracy. However, the proposed approach, leveraging LSTM-based RNN, outperforms other models in distinguishing depression and non-depression texts. The second dataset experiments further validate the system's efficiency, showcasing robustness and high recall rates for both depression and non-depression categories. The proposed system surpasses existing methods, providing a valuable tool for depression prediction.

This study explores a multimodal approach for automatically detecting depression symptoms in text data, with a focus on providing decision support in healthcare. The research employs a combination of one-hot encoding applied to robust features describing depression symptoms and a deep learning method known as Recurrent Neural Network (RNN), specifically Long Short-Term Memory (LSTM).

The text data is sourced from ung.no, a public information channel aimed at young individuals in Norway. Through one-hot encoding, the study extracts words representing depression symptoms from different sentences in the text data. Subsequently, a deep RNN, based on LSTM, is trained to model two emotional states: depression and non-depression. The trained RNN is then utilized to predict the emotional state within unknown text data.

The proposed approach demonstrates notable performance, achieving mean prediction accuracies of 98% and 99% on two datasets containing approximately 11,807 and 21,807 texts, respectively. In contrast, traditional approaches achieve a maximum mean recognition performance of only 91%, underscoring the robustness of the proposed method. Furthermore, the study incorporates an eXplainable Artificial Intelligence (XAI) algorithm, LIME, to assess the system's ability to provide meaningful explanations for its decisions.

The features developed in this research have the potential to support machine learning-based decisions and facilitate the design of user interfaces for more effective mental health care. This deep learning system could be further explored with comprehensive datasets and applied in real-time mental health care services, such as smart chatbot systems, to offer informational support for depression-related concerns among both healthcare professionals and young individuals.

References

Ageitos, E. C., Romo, J. M., & Araujo, L. (2022). UNED-MED at eRisk 2022: depression detection with TF-IDF, linguistic features and Embeddings. *CEUR-WS*, *3180*(-), 1-11. https://ceur-ws.org/Vol-3180/paper-68.pdf

Uddin, M. Z., Dysthe, K. K., & Folstad, A. (2022). Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applications*, *34*(-), 721-744. https://doi.org/10.1007/s00521-021-06426-4