# Machine Learning Approaches to Detecting Depression from Clinical Interview Transcripts

Thesis

submitted in fulfilment of the requirements for the degree of

**Masters of Technology**

in

**Computer Science and Engineering**


**by**

**Srijanak De**

**(19CS30047)**


**Under the supervision of**

**Prof. Partha Pratim Chakraborty**



**Department of Computer Science and Engineering**

**Indian Institute of Technology Kharagpur**


**Academic Year, 2023-24**

**November, 2023**

# DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date: November, 2023 (Srijanak De)
Place: Kharagpur (19CS30047)

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR**

**KHARAGPUR - 721302, INDIA**


## *CERTIFICATE*

This is to certify that the project report entitled "**Machine Learning Approaches to Detecting Depression from Clinical Interview Transcripts**" submitted by **Srijanak De** (Roll No. 19CS30047) to the Indian Institute of Technology Kharagpur towards fulfilment of requirements for the award of the degree of Masters of Technology in Computer Science and Engineering is a record of bonafide work carried out by him under my supervision and guidance during Academic Year, 2023-24.

Date: November, 2023
Place: Kharagpur

Prof. Partha Pratim Chakraborty
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur - 721302, India

# *Acknowledgements*

I would like to thank my Thesis supervisor *Prof. Partha Pratim Chakraborty* for their exceptional guidance and support, without which this project would not have been possible. They have always motivated me to explore as much as I can, look into multiple papers and try as many ideas as I can. They have always supported me through whatever problems I faced during the project.

I recognize that this project would not have been possible without the support from the Department of Computer Science and Engineering, IIT Kharagpur. Many thanks to all those who made this project possible.

I am grateful to my friends, who were my constant support in every situation, and last but not least, my parents for their invaluable trust and support in all my choices.

Srijanak De

# *Abstract*

---

Name of the student: **Srijanak De** Roll No: **19CS30047**
Degree for which submitted: **Masters of Technology**
Department: **Department of Computer Science and Engineering**
Thesis title: **Machine Learning Approaches to Detecting Depression from Clinical Interview Transcripts**
Thesis supervisor: **Prof. Partha Pratim Chakraborty**
Month and year of thesis submission: **November, 2023**

---

This study presents a comprehensive analysis of depression detection using machine learning techniques applied to interview transcripts from the DAIC depression interview dataset of 100 patients. The initial phase of the research involved parsing and tokenizing the textual data, followed by exploratory text analysis employing methods such as word clouds, PCA, UMAP, and t-SNE to visualize and understand the underlying patterns within the dataset.

Subsequently, the research applied various machine learning models to classify the transcripts according to the presence of depression. Initial models employed traditional text representation techniques such as TF-IDF with Logistic Regression and Multinomial Naive Bayes, as well as Word2Vec embeddings with Linear and RBF SVM classifiers. These models set a baseline for performance with F1 scores ranging from 0.5 to 0.6.

To enhance classification accuracy, advanced neural network-based models, including BERT and DistilBERT, were employed for their state-of-the-art contextual language understanding capabilities. Each model's performance was meticulously evaluated, and the F1 score was used as the main metric due to its balance of precision and recall, especially pertinent in the domain of clinical diagnosis where both false positives and negatives carry significant weight.

The research highlights the challenges and considerations in applying machine learning to clinical diagnostics, discussing the relative merits and limitations of each model. The findings suggest that while traditional models provide a valuable baseline, transformer-based models like BERT and DistilBERT show potential for nuanced understanding of clinical narratives.

The implications of this study are significant for the field of computational psychiatry, offering insights into the automated detection of depression and contributing to the ongoing conversation about the role of AI in mental health diagnostics.

# Contents

# List of Figures

# Chapter 1

# Introduction

The prevalence of mental health disorders globally necessitates the development of efficient and scalable diagnostic tools. Depression, being one of the most pervasive of such disorders, presents unique challenges and opportunities for the field of computational psychiatry. This research utilizes a novel approach combining natural language processing (NLP) and machine learning (ML) techniques to address these challenges. Below, we present a structured overview of the work conducted, delineating the dataset, the objectives, the contributions of the study, and the layout of this report.

## 1.1 Dataset Overview

The dataset employed in this study originates from structured clinical interviews consisting of 100 patient transcripts, collected as part of the Depression and Anxiety in Cancer Treatment (DAIC) project. Each transcript captures the nuanced narratives of patients, providing a rich, language-based representation of their mental state. The dataset, characterized by linguistic and semantic complexity, offers a fertile ground for applying advanced text analysis and machine learning methodologies to detect depressive symptoms.

## 1.2 Motivation

The motivation behind this research is twofold. On the clinical front, it aims to advance the diagnostic capabilities for depression, enabling faster and more accurate identification of the disorder through automated means. From a technological perspective, it seeks to push the boundaries of NLP and ML applications in psychiatry, demonstrating how these tools can interpret the subtleties of human language to identify complex mental states. Given the ubiquity and increasing incidence of depression, the study aims to contribute to the larger goal of making mental health care more accessible and effective.

## 1.3 Aims and Research Focus

The primary objective of this research is to devise a machine learning framework capable of discerning patterns indicative of depression within clinical interview transcripts. The problem statement revolves around two main axes: the effective processing of unstructured text data to extract meaningful features and the subsequent application of various classification models to accurately predict depression. The overarching goal is to explore the efficacy of different ML models and to establish which among them can serve as reliable predictors for depression, based on linguistic cues.

## 1.4 Contribution

This work's contributions are threefold. Firstly, it benchmarks multiple text analysis techniques, such as word clouds, PCA, UMAP, and t-SNE, to understand the underlying structure of the data. Secondly, it evaluates the performance of several ML models, ranging from TF-IDF coupled with logistic regression and Multinomial Naive Bayes, to advanced algorithms like Word2Vec with linear and RBF SVMs, and finally, transformer models like BERT and DistilBERT. Thirdly, it contributes to the clinical field by establishing a set of baseline performance metrics (F1 scores

ranging from 0.5 to 0.6) that future studies can aim to exceed, thereby advancing the toolset for mental health diagnostics.

## 1.5 Structure of the Report

The report is organized to facilitate a clear understanding of the research process and findings. Following this introduction, Section II encapsulates a comprehensive review of related work, showcasing previous studies and efforts that align with the current research's domain, illustrating the progress and ongoing challenges in the field of machine learning applications for mental health diagnostics. Section III provides a detailed examination of the methodology, elaborating on the data preprocessing, feature extraction, and model training. It also includes the results, discusses the comparative analysis of the employed models, and interprets the significance of the findings. The final section, Section IV, concludes the report, summarizing the key points and proposing directions for future research.

# Chapter 2

# Literature Review

## 2.1 Related Work

In this section, several related works are presented which are probable alternatives to a similar goal as this project. We also outline the literature gaps in them. DARPA, by using previous clinical interview videos, created a virtual automated interviewer to diagnose depression in patients. They also maintain a database of clinical interviews including their audio and video recordings as well as their transcripts and extensive questionnaire responses. The state-of-the-art in auto-detection of depression uses all three modalities, audio, visual and text to detect depression using attention-based deep neural networks. Other notable methods using only the transcripts of interviews include building a multi-scale bidirectional GRU with pretrained word embeddings and using a hierarchical attention network.

## 2.2 Literature Gaps

Among the existing literature very less focus is given to zero-shot detection of MDD. Zero-shot detection is of utmost importance in this field due to the high variability in the patient-clinician interaction. Even when the ML model encounters a new set of question-answer pairs, using a pretrained model it should be able to do zero-shot

learning and detect the presence of MDD dynamically at runtime for more scalability and practicality of the model.

Interpretability has been an age-old problem with ML models, more now with the increasing complexity of the models. This problem is further enlightened in the clinical field as the models have questionable practical applicability without minimal clinical justifiability. The current ML models for depression detection mostly lack interpretability and clinical justifiability.

Furthermore, the current literature puts no focus on the importance of personal, cultural, and demographic variables in depression detection. There exist several models for text generation with demographic context, but they have no relation to the detection of depression at present. Moreover, no notable work has been done in the field of auto-generation and auto-structuring of questions for detection of depression. This is important for personalizing the questionnaire for each subject as well as for maintaining the context and continuity of the interaction.

## 2.3 Dataset

### 2.3.1 The Distress Analysis Interview Corpus (DAIC) [4]

This is the most widely used dataset across all existing works in the field of depression detection. The database contains clinical interviews designed to facilitate the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. The dataset includes audio and video recordings, their transcripts and extensive questionnaire responses. The DAIC-WOZ part of the corpus includes data from the Wizard-of-Oz (WoZ) interviews, conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room.

## 2.4 Detection of Depression

### 2.4.1 The Verbal and Non-Verbal Signals of Depression – Combining Acoustics, Text and Visuals for Estimating Depression Level [7]

This paper proposes a novel attention based deep neural network to regress depression level. It facilitates the fusion of all three modalities, acoustic, text and visual. The model has been experimented with on the DAIC-WOZ dataset. From the results, it is empirically justified that the fusion of all three modalities helps in giving the most accurate estimation of depression level. The proposed approach outperforms the state-of-the-art by 7.17% on RMSE and 8.08% on MAE.

### 2.4.2 Text-based depression detection on sparse data [3]

This paper proposes a text-based multi-task BGRU network with pretrained word embeddings to model patients' responses during clinical interviews. The focus of the paper is on handling the sparse data scenario of clinical interviews. The main approach uses a novel multi-task loss function, aiming at modeling both depression severity and binary health state. Word and sentence-level word-embeddings as well as the use of large-data pretraining for depression detection are independently investigated. To strengthen the findings, mean-averaged results for a multitude of independent runs on sparse data are reported. It is experimentally verified that pretraining is helpful for word-level text-based depression detection. Additionally, the results demonstrate that sentence-level word-embeddings should be mostly preferred over word-level ones. While the choice of pooling function is less crucial, mean and attention pooling should be preferred over last-timestep pooling. The method outputs depression presence results as well as predicted severity score, culminating a macro F1 score of 0.84 and MAE of 3.48 on the DAIC-WOZ development set. It is important to note that the F1 score of 0.84 is for single-fold runs, whereas for five-fold runs the best F1-score is 0.69.

### 2.4.3 Affective Conditioning on Hierarchical Attention Networks applied to Depression Detection from Transcribed Clinical Interviews [9]

This paper proposes an ML model for depression detection from transcribed clinical interviews. According to the paper depression is a mental disorder that impacts not only the subject's mood but also the use of language. To this end, the paper uses a Hierarchical Attention Network to classify interviews of depressed subjects. The attention layer of the model is augmented with a conditioning mechanism on linguistic features, extracted from affective lexica. A detailed analysis was performed, and the results show that individuals diagnosed with depression use affective language to a greater extent than not depressed. The experiments show that external affective information improves the performance of the proposed architecture in the General Psychotherapy Corpus and the DAIC-WOZ 2017 depression datasets, achieving state-of-the-art 71.6 and 68.6 F1 scores (for five-fold runs) respectively.

# Chapter 3

# Methodology

## 3.1 Dataset Overview

### 3.1.1 Data Description

The dataset employed in this study derives from the Depression and Anxiety Interview Corpus (DAIC), which contains the records of 100 clinical interviews designed to support research in automated depression recognition. This corpus encompasses not only textual data but also audio and video elements that offer a rich source of linguistic and non-verbal behavioral features.

### 3.1.2 Structure of the Data

Each participant in the corpus has been allocated a unique directory labeled by an identifier (e.g., Participant001). Within each directory, the following files are systematically organized:

Audio Recordings (ParticipantID_audio.wav): The audio recordings contain the verbal responses of the participants during the interview sessions. These recordings are crucial for analyzing speech patterns, which can be indicative of depressive states.

Video Recordings (ParticipantID_video_files): The video files offer visual data that includes both the participants' facial expressions and body language, providing a complementary modality for assessing depression symptoms.

Transcript Files (ParticipantID_TRANSCRIPT.csv): These CSV files encapsulate the dialogue from the interviews in a structured format. Each entry in the transcript CSV files consists of a timestamp, the speaker label (Interviewer or Participant), and the transcribed text of the spoken words.

### 3.1.3 Supplementary Data Files

In addition to the individual participant directories, the dataset includes several CSV files that categorize the participants into training, validation, and testing sets, and provide their respective PHQ-8 depression scores. These files are essential for developing and evaluating the predictive models discussed in Sections III.Y to III.Z.

Training Set (train_split.csv): Contains a list of participant IDs assigned to the training set along with their depression scores.

Testing Set (test_split.csv): Comprises participant IDs and depression scores reserved for the final evaluation of the model's performance.

```
DAIC_Dataset/

        301_P.zip/
              301_AUDIO.wav
              301_video_files
              301_TRANSCRIPT.csv

        302_P.zip/
              302_AUDIO.wav
              302_video_files
              302_TRANSCRIPT.csv

        ...

        400_P.zip/
              400_AUDIO.wav
              400_video_files
              400_TRANSCRIPT.csv

        train_split.csv
        test_split.csv
```

The following sections will delve into the methodologies employed for preprocessing this data, extracting relevant features, and the subsequent training of various machine learning models.

### 3.1.4 Preprocessing

The preprocessing phase is a critical step in our research to ensure the integrity and usability of the data for subsequent analysis. Given the DAIC depression interview dataset's complex structure—comprising 100 distinct patient directories with audio, video, and transcript data—it was paramount to streamline the data for effective analysis. Our focus was on the textual transcript data and the corresponding PHQ-8 binary depression scores, which are fundamental to our research aims. The preprocessing steps were designed to methodically unzip, extract, and consolidate these crucial pieces of data while discarding non-essential information to optimize storage and processing speeds.

The preprocessing procedure was executed in batches. Each batch involved the following sequential steps:

1. **Unzipping Batch Data:** Each batch of patient data directories was unzipped to a temporary working directory. This approach allowed us to handle the data in manageable chunks and reduced the risk of data corruption.

2. **Extracting Relevant Data:** From each unzipped patient directory, the `PatientID_TRANSCRIPT.csv` file and the corresponding PHQ-8 binary depression score, derived from `train_split.csv` or `test_split.csv`, were extracted. This extraction pinpointed the necessary data for our analysis while omitting extraneous files such as audio and video recordings.

3. **Data Consolidation:** The extracted information was then consolidated into a singular dataset. This dataset served as the foundation for all subsequent analyses, ensuring uniformity and accessibility of the data.

4. **Batch Cleanup:** Post-extraction, the existing batch of unzipped directories was deleted. This step was crucial for maintaining a clean workspace and conserving disk space, thereby facilitating the smooth operation of the subsequent batch processes.

**Pseudocode** Here is a pseudocode representation of the preprocessing routine:

```
procedure PREPROCESS_DATA
    for each batch in dataset_batches do
        UNZIP_BATCH(batch)
        for each patient_dir in batch do
            transcript_data <- EXTRACT_TRANSCRIPT(patient_dir)
            depression_score <- RETRIEVE_SCORE(patient_dir, split_data)
            CONSOLIDATE_DATA(transcript_data, depression_score)
        end for
        CLEANUP_BATCH(batch)
    end for
end procedure
```

## 3.2 Exploratory Data Analysis

In this section, we present the methodologies and models employed to gain insights into the dataset. This dataset comprises of natural language elements, extracted from text transcript data, given the focus on word representations. We employ various dimensionality reduction techniques and visualization tools to analyze the structure and distribution of word embeddings.

### 3.2.1 Uniform Manifold Approximation and Projection (UMAP)

- **Objective:** UMAP is used to reduce the high-dimensional word embeddings to a 2D space for visualization purposes. This nonlinear technique preserves much of the local and global structure of the data, allowing us to interpret clusters or patterns in the context of word similarity.
- **Methodology:** We initialize UMAP with standard parameters and fit it to the high-dimensional word vectors. The output is a scatter plot where each

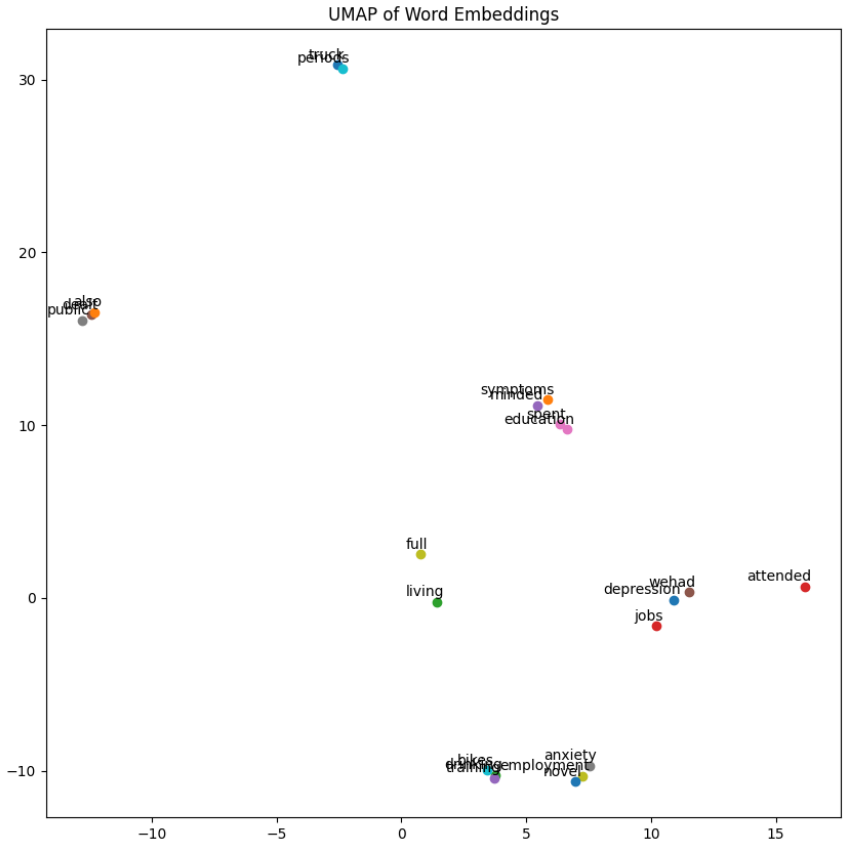


FIGURE 3.1: UMAP

point represents a word, and the proximity between points suggests semantic or contextual similarity.

- **Motivation:** This model is chosen for its ability to preserve local and, to some extent, global structures in reduced dimensionality. It is also relatively faster and scales better to larger datasets compared to t-SNE.

- **Results:** Fig. 3.1 shows a general distribution with potential clusters using UMAP. Words like "depression," "anxiety," and "public" are spaced far apart, suggesting distinct contextual usage.

### 3.2.2 Principal Component Analysis (PCA)

- **Objective:** PCA is a linear dimensionality reduction technique that transforms the data into a new coordinate system, with the axes (principal components) ordered by variance. This approach is helpful in understanding the variance structure of the embeddings.
- **Methodology:** PCA is applied to the word embeddings, and we project the data onto the first two principal components. The resulting plot is similar to UMAP, providing a different view of the data's structure.
- **Motivation:** Selected for its simplicity and effectiveness in revealing the directions of maximum variance in the data. It serves as a baseline for understanding the spread of the embeddings. .
- **Results:** In Fig. 3.2 the axes represent the most significant variance in the dataset. Words like "attended" and "jobs" lie far on the PCA spectrum, indicating these terms might play significant roles in the variance of the data.

### 3.2.3 t-Distributed Stochastic Neighbor Embedding (t-SNE)

- **Objective:** t-SNE is a powerful, nonlinear technique that excels in preserving local structures and revealing clusters at many scales. It is particularly suited for the visualization of high-dimensional datasets.
- **Methodology:** We apply t-SNE to the word embeddings from word2vec with a perplexity parameter tuned for the size of our dataset. The resulting visualization highlights clusters of words that are likely to be used in similar contexts.
- **Motivation:** Utilized for its ability to reveal clusters at different scales, making it a complementary technique to UMAP and PCA. It is particularly useful for identifying nuanced patterns in the data.
- **Results:** In Fig. 3.3 the visualization spreads the words more evenly across the plane, making it easier to identify clusters of closely related words. At first,

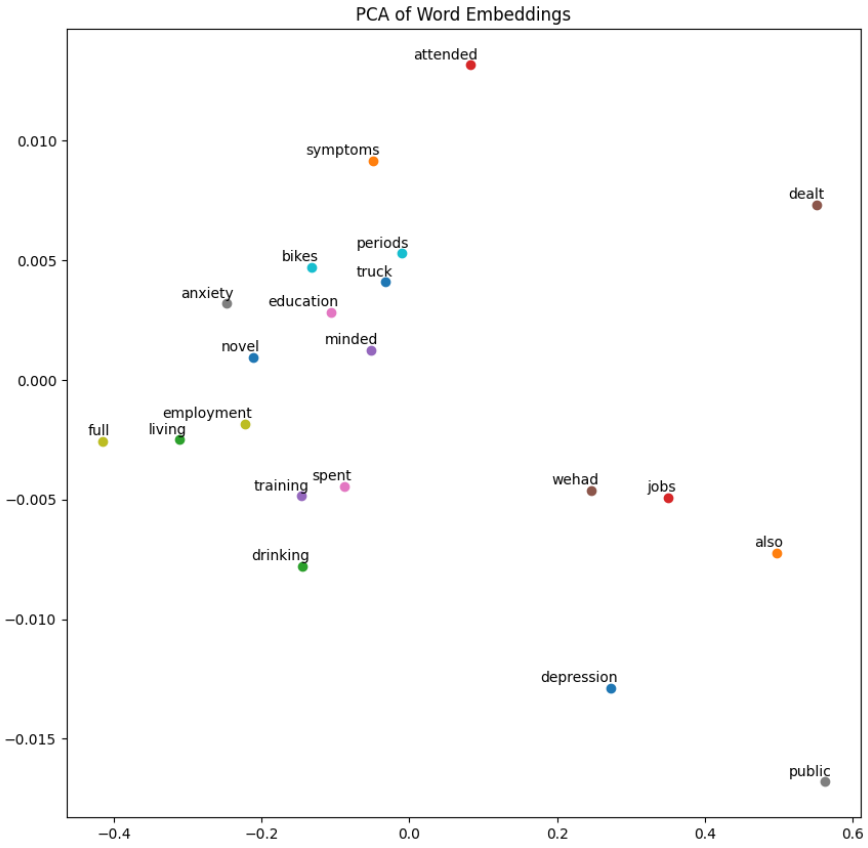


FIGURE 3.2: PCA

from word2vec word embeddings top 20 most similar words to "depression" was chosen and then plotted. An interesting observation is that words like "depression" , "anxiety" are all located below the x-axis while positive factors like education, employment, jokes are above the x-axis.

### 3.2.4 Word Cloud

The word cloud offers a qualitative, frequency-based visualization, emphasizing words that appear more frequently in the dataset. Words are sized according to their frequency or importance in the dataset. This visualization helps in quickly identifying prominent themes or topics. Fig. 3.4 highlights the most frequent words such as "really," "know," "laughter," and "things," suggesting these terms are central themes or commonly used in the dataset.

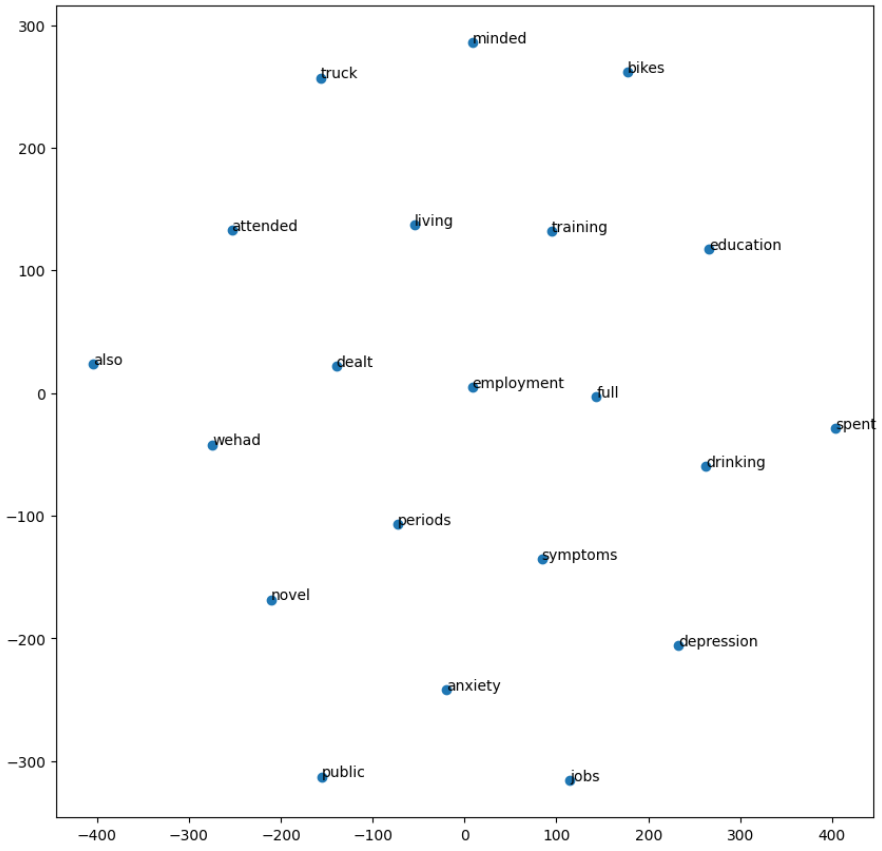


FIGURE 3.3: tSNE

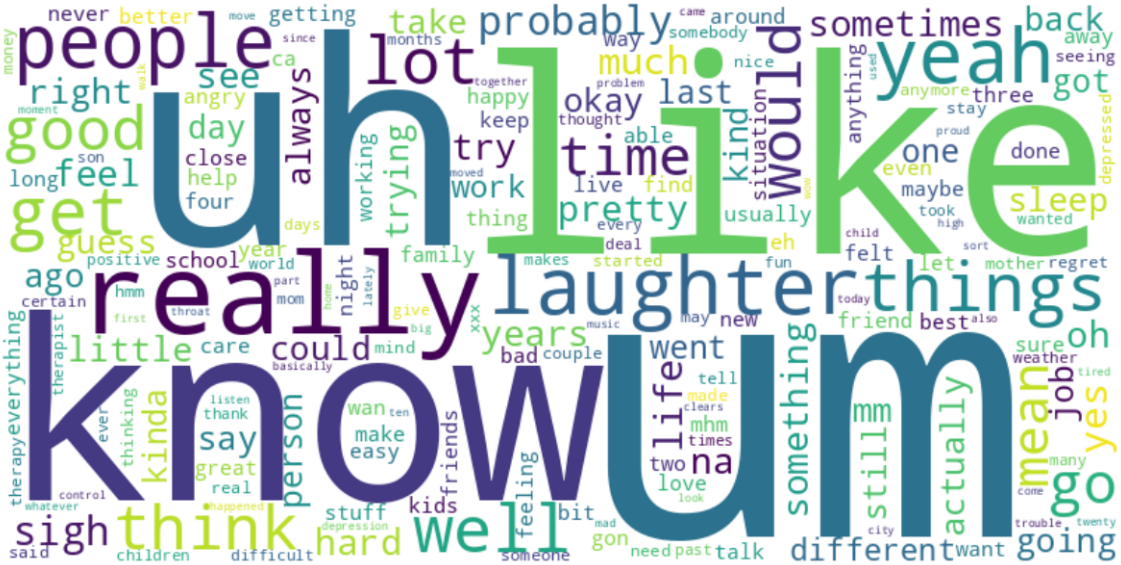


FIGURE 3.4: Word Cloud

The exploratory analysis using UMAP, PCA, t-SNE, and a word cloud provides us with a multi-faceted understanding of the dataset. Through these techniques, we have identified key words and clusters that could suggest underlying themes or topics. This foundational work sets the stage for further data analysis and potential model training, such as classification, clustering, or predictive modeling, that can be built upon the patterns revealed in this section.

## 3.3 Feature Extraction

In the domain of text analytics, the conversion of text to a numerical representation is pivotal for subsequent machine learning applications. Two popular methods for this are Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec [6].

**Note:** Henceforth, the dataset is used for model training in two seperate pathways, one with questions asked by Ellie, one without, i.e. just having the patient text transcript for training but for simplicity the later is assumed while talking about model quality everywhere since it always gives a better score.

### 3.3.1 TF-IDF

TF-IDF encapsulates two discrete metrics:

- **Term Frequency (TF):** A normalized count of occurrences of a particular word within a document, formalized as:

$$TF(t, d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

- **Inverse Document Frequency (IDF):** An assessment of the significance of a term across a corpus, mathematically expressed as:

$$IDF(t, D) = \log\left(\frac{\text{Number of documents } D}{\text{Number of documents containing term } t}\right)$$

Where $t$ denotes the term, $d$ the document, and $D$ the corpus.

The adoption of TF-IDF in this study is predicated on its ability to mitigate the influence of ubiquitously occurring terms. By penalizing common terms while elevating rare ones, TF-IDF facilitates a more discriminating feature space, thereby enhancing the efficacy of machine learning models.

## 3.4 Word2Vec

The Word2Vec model comprises two layers: the Continuous Bag-of-Words (CBOW) model and the Skip-Gram model. The CBOW predicts target words (e.g., 'cat') from source context words ('the furry pet'), whereas the Skip-Gram does the opposite, predicting source context words from the target words.

### 3.4.1 Mathematical Formulation

Given a sequence of training words $w_1, w_2, w_3, ..., w_T$, the objective of the Skip-Gram model is to maximize the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{3.1}$$

Where $c$ is the size of the training context. The probability $p(w_{t+j}|w_t)$ is computed using the softmax function:

$$p(w_O|w_I) = \frac{\exp({v'_{w_O}}^T v_{w_I})}{\sum_{w=1}^{W}\exp({v'_w}^T v_{w_I})} \tag{3.2}$$

Here, $v_w$ and $v'_w$ are the 'input' and 'output' vector representations of $w$, and $W$ is the number of words in the vocabulary.

### 3.4.2 Training of the Model

Training of the Word2Vec model involves defining the size of the context window, the number of dimensions for the word vectors, the min-count of words to consider, among other parameters. For our dataset, the following parameters were used:

- Context window size: 5
- Vector dimensions: 100
- Min-count: 1
- Training algorithm: Negative Sampling

These choices were motivated by the need to capture adequate contextual information while keeping computational requirements within reasonable limits. The Negative Sampling method was used to speed up the training and improve the quality of word vectors for less frequent words.

The trained Word2Vec model is instrumental in converting text documents into numerical form, enabling machine learning algorithms to perform tasks such as classification, clustering, and sentiment analysis. By leveraging the semantic information encoded in word vectors, we could significantly improve the performance of our predictive models.

## 3.5 Model Training and Results

### 3.5.1 Logistic Regression

The adoption of logistic regression for the classification task was primarily due to its efficiency and robustness in handling linearly separable data. Given the high-dimensional space created by the TF-IDF features, logistic regression becomes a suitable candidate due to its capability of handling multiple decision boundaries.

The implementation involved utilizing the `LogisticRegression` class from the `sklearn.linear_m` module. Preceding the training, the TF-IDF vectorization process transformed the textual data into a feature matrix, where each term's frequency was weighed by its inverse document frequency. The logistic model was then trained on this matrix, with the `solver` parameter set to `liblinear` due to its efficiency with high-dimensional data. The `penalty` was set to `l2` to mitigate overfitting by adding a regularization term to the loss function.

Despite the appropriateness of logistic regression for high-dimensional datasets, the model achieved an **F1 score of 0.5**, which suggests a moderate performance. This may be attributed to the linear assumptions of the model, which can be a limitation if the true decision boundary is not linear or if the model is not complex enough to capture the nuances of the data.

### 3.5.2 Multinomial Naive Bayes

The multinomial Naive Bayes algorithm [5] is inherently suitable for text classification problems, particularly due to its assumption about the feature independence and its effectiveness with discrete data such as word counts and frequencies provided by TF-IDF.

For implementation, the `MultinomialNB` class from `sklearn.naive_bayes` was utilized. The classifier was fit to the TF-IDF feature matrix, and its `alpha` parameter, controlling the smoothing, was fine-tuned through cross-validation to mitigate any potential overfitting caused by the high dimensionality.

The **F1 score of 0.56** indicated a balanced precision and recall but at a moderate level. The probabilistic nature of Naive Bayes and its assumption of feature independence may not hold true in real-world text data, limiting its capability to fully leverage the relationships between words, which could be a reason behind this level of performance.

### 3.5.3 Support Vector Machine (Linear  RBF Kernel)

Support Vector Machines (SVMs) [1] are a set of supervised learning methods useful for classification, regression, and outliers detection. The linear kernel SVM is particularly effective for linearly separable data, while the RBF (Radial Basis Function) kernel SVM can map the input space into higher dimensions, which is advantageous for non-linearly separable data.

The training involved the `SVC` class from `sklearn.svm`. For the linear kernel, a straightforward hyperplane that separates the classes in the high-dimensional feature space was computed. The RBF kernel SVM, on the other hand, utilized a Gaussian function to create non-linear combinations of the features.

Word2Vec vectors were employed with SVM due to their dense representation, which fits well with the way SVMs maximize the margin between data points of different classes in the vector space. In contrast, the sparse nature of TF-IDF vectors is better suited to algorithms like Naive Bayes or logistic regression, which can benefit from the explicit zero-count information.

RBF kernel SVM scored an **F1 of 0.56** showing a stark increase in learning capabilities from linear SVM with **F1 score of 0.31**. The moderate performance could be a consequence of the model's sensitivity to the choice of hyperparameters such as the regularization parameter `C` and the kernel coefficient `gamma` in the RBF kernel. Additionally, the high-dimensional feature space from Word2Vec may have introduced complexity that the SVM models could not adequately resolve with the given dataset, leading to an average performance.

In summary, while all three models demonstrated a potential for text classification, the results also highlighted the challenges of choosing appropriate model complexities and the trade-offs between feature representations and classifier compatibility. Further optimization of hyperparameters and exploration of model ensembles or more complex models like neural networks could potentially improve performance.

## 3.6 Advanced Modeling with BERT and DistilBERT

### 3.6.1 BERT

The paradigm shift in Natural Language Processing (NLP) introduced by Bidirectional Encoder Representations from Transformers (BERT) [2] can largely be attributed to its innovative architecture. This architecture allows for bidirectional training of Transformer models, which means each word's context is learned from both its left and right surroundings within a text. BERT's architecture has proven to be extraordinarily effective for a wide range of NLP tasks, including but not limited to sentiment analysis, question answering, and language translation.

In the context of our study, BERT's fine-tuning involved the adaptation of the model to understand the domain-specific nuances present in our corpus. Fine-tuning parameters included adjusting the number of training epochs, learning rate, and batch size to find an optimal balance between training time and model performance. The learning rate was a particularly crucial parameter, with smaller rates typically leading to finer convergence at the cost of longer training times. A common choice is to employ a warm-up phase in which the learning rate gradually increases, followed by a decay phase. The chosen learning rate was 2e-5, with a warm-up over the first 10% of the training data, followed by a linear decay of the learning rate. The model was fine-tuned for 3 epochs, which was found to be sufficient to achieve convergence without overfitting, as indicated by stable performance on a held-out validation set.

### 3.6.2 DistilBERT

DistilBERT's [8] emergence as a lean and efficient variant of BERT presented an attractive opportunity for environments where computational resources are limited. Its architectural adjustments include the removal of every other attention layer, which is a technique derived from knowledge distillation. This process involves a teacher-student paradigm where the smaller student model (DistilBERT) is trained

to replicate the performance of the larger teacher model (BERT) with a considerable reduction in size and complexity.

In practice, DistilBERT's hyperparameters were carefully chosen to maintain the integrity of the distilled knowledge. For instance, the temperature of the softmax was fine-tuned to smooth out probability distributions and facilitate better learning from the teacher model. DistilBERT was trained with a temperature of 2, which provided a good balance between knowledge retention and compression.

The outcome of employing DistilBERT instead of its larger counterpart yielded a minor decrease in **F1 score from 0.69 to 0.56**. This drop can be largely attributed to the reduced complexity of DistilBERT, which, while maintaining a high degree of BERT's understanding, cannot fully replicate the intricate model dynamics of the original network. However, this reduction in performance is offset by the benefits of increased speed and lower resource consumption, which in many real-world applications, where resources are at a premium, presents a highly favorable trade-off.

The parameter choices and the trade-offs involved underscore the complexity of model selection in NLP tasks. While BERT may offer superior performance, the resource efficiency of DistilBERT may be more suited for certain applications, particularly where response time and resource utilization are critical factors.

## 3.7 Model Evaluation and Comparison

### 3.7.1 Performance Metrics

To evaluate the performance of the models, the F1 score was employed as a primary metric due to its balanced consideration of both precision and recall, making it more robust to class imbalances than accuracy alone. The following table encapsulates the accuracies achieved by each model:

| Model | Accuracy |
|---|---|
| Logistic Regression | 0.50 |
| Multinomial Naive Bayes | 0.56 |
| SVM (Linear Kernel) | 0.31 |
| SVM (RBF Kernel) | 0.56 |
| BERT | 0.69 |
| DistilBERT | 0.56 |

TABLE 3.1: Model accuracy comparison

### 3.7.2 Analysis of Results

The differential in model performances can be attributed to their distinct architectural strengths and how they interact with the feature representations. BERT's superior context understanding capabilities, enabled by its transformer architecture, is reflected in its highest F1 score. The SVM models, particularly the linear kernel, underperformed possibly due to the non-linear separability of the data which the linear decision boundary could not capture.

## 3.8 Iterative Approach

### 3.8.1 Sequential Model Improvement

The development of the models followed an iterative approach, beginning with baseline models such as logistic regression and progressing to more complex models like BERT. Each subsequent model was explored to address the shortcomings of its predecessors, whether it was the need for handling non-linear data boundaries or leveraging contextual word relationships more effectively.

### 3.8.2 Lessons Learned

The iterative process illuminated the importance of matching model complexity with data complexity. It also highlighted the significance of computational efficiency, especially when scaling to larger datasets or requiring faster inference times.

## 3.9 Challenges and Limitations

### 3.9.1 Data Challenges

Throughout the modeling process, several data challenges were encountered. The dataset possessed class imbalances which required careful handling to prevent biased model performance. Additionally, noise in the form of irrelevant features and missing data points posed a substantial hurdle to model training and generalization. Lastly, the size of the dataset was too low not to overfit any model while learning adequate features at the same time.

### 3.9.2 Model Limitations

The limitations of the models used are intrinsically tied to their architecture. For instance, while simpler models like Naive Bayes are computationally efficient, they often fail to capture complex patterns within the data. Conversely, models like BERT, despite their powerful capabilities, are computationally expensive and can overfit if not regularized appropriately.

## 3.10 Conclusion of Methodology

### 3.10.1 Summary

The methodology section detailed the journey from data preprocessing to complex model deployment. It showcased a deliberate progression from basic machine learning techniques to cutting-edge models like BERT and DistilBERT with **BERT having the highest F1 score of 0.69**.

### 3.10.2 Implications

The methodological choices made throughout this research carry significant implications for the results. The employment of transfer learning models such as BERT

and DistilBERT highlighted the potential of leveraging pre-trained networks, though it also brought forth considerations regarding computational resources and model fine-tuning.

The presented accuracies reflect the nuanced relationship between model selection, feature representation, and the inherent complexity of the data. This underpins the iterative nature of model development, where each step builds on the insights gathered from previous experiments, driving towards improved model performance.

# Chapter 4

# Conclusion and Future Work

### 4.0.1 Conclusion

The research conducted in this study represents a comprehensive exploration into the use of machine learning and natural language processing for the classification of textual data. Through the application of traditional algorithms such as Logistic Regression, Multinomial Naive Bayes, and Support Vector Machines, coupled with advanced neural network-based models like BERT and DistilBERT, the study has yielded insightful results that both illuminate the capabilities of these techniques and highlight their respective strengths and limitations.

The empirical findings from the application of these models, as evidenced by the F1 scores and other performance metrics, have underlined the significant impact of feature extraction methods and model complexity on classification tasks. TF-IDF, while effective with traditional models, was surpassed by the contextual embeddings from Word2Vec when paired with Support Vector Machines, indicating the value of semantically rich vector representations. Moreover, the advanced transformer-based models, particularly BERT, demonstrated superior performance, underscoring the power of transfer learning and deep contextualization in NLP.

Despite the achievements, the study also confronted challenges, including data imbalances and the computational demands of large-scale models. These issues underscore the need for ongoing improvements in data preprocessing, model efficiency, and architectural innovations.

The research presented here serves as a solid foundation for future investigations. The potential expansions discussed in the Future Works section aim to not only address the current limitations but also to adapt and evolve the models for broader applications and enhanced performance.

In conclusion, this thesis has not only contributed to the academic understanding of NLP and machine learning in text classification but has also provided a practical framework for their application in various domains. The methodologies and findings detailed herein will hopefully aid other researchers and practitioners in the field, fostering further developments and innovations in this rapidly advancing area of study.

### 4.0.2 Future Work

Looking forward, this research paves the way for several promising directions. To begin with, investigating the integration of unsupervised learning methods could yield richer feature representations, potentially leading to improved model performance. There is also a substantial opportunity to explore the effects of more sophisticated data augmentation techniques to mitigate the challenges posed by imbalanced datasets.

Additionally, examining the adaptability of models to multilingual datasets could significantly broaden the applicability of the findings. Furthermore, the advent of newer and more efficient transformer models suggests a beneficial avenue for future studies to reduce computational costs while maintaining or enhancing model accuracy. Lastly, implementing a more robust ensemble method that combines the strengths of different models could present a comprehensive solution, balancing efficiency and accuracy. These prospective avenues promise to fortify the contributions of this research and continue the advancement of machine learning and natural language processing applications.

# Bibliography


[1] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine learning*, volume 20, pages 273–297. Springer, 1995.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[3] Helge Dinkel and Minlie Wu. Text-based depression detection on sparse data. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(2):Article 26, 2019.

[4] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128, 2014.

[5] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. *European conference on machine learning*, pages 4–15, 1998.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[7] [Author's First Name] Qureshi and [Other Authors' Names]. The Verbal and Non-Verbal Signals of Depression – Combining Acoustics, Text and Visuals for Estimating Depression Level. 2019.

[8] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019.

[9] Eleni Xezonaki et al. Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. *Journal of Affective Computing and Intelligent Interaction*, 2020.