Depression detection in clinical interview transcripts using natural language processing and deep neural networks

by

Sudhir Mandarapu

Honours Thesis

Submitted by Sudhir Mandarapu

in partial fulfilment of the requirements for the degree

Bachelor of Software Engineering with Honours (2770)

Supervisors: Jojo Wong, Joanne Byrne, Marc Cheong, Yen Cheung & Eddie Robinson

Clayton School of Information Technology

Monash University

November 2018

Contents

List of Tables

List of Figures

Depression detection in clinical interview transcripts using natural language processing and deep neural networks

Sudhir Mandarapu
sman79@student.monash.edu
Monash University, 2018

Supervisors: Jojo Wong, Joanne Byrne, Marc Cheong, Yen Cheung & Eddie Robinson

Abstract

Depression is described as a continued lack of positive emotions that can have a detrimental effect on one's life. The sooner it is detected and addressed, the quicker the suffering individual can recover. This study aims to build a tool that applies natural language processing and machine learning techniques to predict whether someone is showing symptoms of depression based on textual features. This investigation is centered around the DAIC-WOZ depression dataset of transcripts and audio recordings from 189 clinical interviews with depressed and non-depressed people. Firstly, we established the linguistic features that are to be used in the machine learning classifier. These were extracted from clinical guidelines and literature about depression as well as from the dataset itself by identifying differences between the conversations with depressed and non-depressed individuals. Feature extraction from the dataset was performed using natural language processing techniques that have successfully been used for similar tasks in the recent past such as using the Linguistic Inquiry and Word Count (LIWC) tool and Latent Dirichlet Allocation (LDA). Once established, we used these features to train a deep neural network that can leverage linguistic characteristics when classifying texts. The model went through multiple iterations of modifying the input features and hyperparameters and was able to achieve an accuracy of 0.84375 and an F1 score of 0.80879. Furthermore, this study confirmed previous findings, among others, that negative emotion affect words, anxiety affect words, words about friends and leisure activities, as well as antidepressants are strong indicators of depression in clinical interview transcripts. We also found that LDA topics often perform better as input features when detecting depression than LIWC word frequencies. The production of the intelligent classifier model will complement the diagnosis process of clinicians and help them intervene and help individuals recover sooner.

Depression detection in clinical interview transcripts using natural language processing and deep neural networks

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Sudhir Mandarapu
November 8, 2018

Depression detection in clinical interview transcripts using natural language processing and deep neural networks

Acknowledgments

I would like to express my appreciation to everyone who supported me over the past year. I would like to thank Jojo Wong for providing me with resources and consistent feedback about my progress and results. I would like to express my very great appreciation to Eddie Robinson for helping me understand the field of mental health, about which I had minimal knowledge before starting this project. I would also like to thank Yen Cheung for making sure I was constantly on schedule. My grateful thanks are also extended to Marc Cheong for giving me technical advice and reviewing my work constantly. I would like to express my gratitude to Joanne Byrne for providing me with resources and suggestions which did end up being investigated as part of my study.

Finally, I would like to thank my parents, without whose emotional and financial support, I never could have been able to pursue this great opportunity to study at Monash University and work on an intriguing honours project.

Sudhir Mandarapu
November 2018

Table 1.1

List of abbreviations.

Abbreviation	*Meaning*
AVEC	Workshop on Audio/Visual Emotion Challenge
DSM-5	Diagnostic and Statistical Manual of Mental Disorders - 5
DNN	Deep neural network
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
NLP	Natural language processing
PHQ	Patient Health Questionnaire depression scale
ssToT	Semi-supervised topic modelling over time
SVM	Support vector machine
TF-IDF	Term Frequency Inverse Document Frequency

1 Introduction

1.1 Context and Motivation

Over one million adult Australians suffer from depression every year (beyondblue, 2018) and the mental illness is constantly growing at an alarming rate across the world (Whitaker, 2005). If not addressed, it can have detrimental effects on one's life and can give rise to other mental disorders or suicide (Rihmer, 2001). The sooner this condition is identified and addressed, the quicker a sufferer can recover from it (beyondblue, 2018). While the community has various support mechanisms such as health authorities and organisations like beyondblue, early detection and intervention enables individuals to access support earlier - increasing the likelihood of a cure and reducing the severity of depressive symptoms. Although medical professionals are the ultimate specialists and healthcare providers for the sufferers, they can benefit from an automated tool that can accurately identify individuals who are exhibiting the first signs of of mental illness. This study aims to build such a tool that is trained using transcripts of clinical interviews, which are often conducted as part of mental health assessments. A linguistic approach that uses natural language processing techniques was chosen for this study because, as explained by Blanken et al. (1993), language behaviour can be studied to reveal psychological symptoms. Furthermore, this research focused on the use of deep neural networks for building this model because they have not been used as much for the task of depression detection, and might be a better choice than more commonly used alternatives such as SVMs and logistic regression. The resulting tool can then be used to complement the diagnosis process of medical professionals so they can identify symptoms of depression sooner.

1.2 Aims and Subgoals

The main aim for this study is to develop an automated classifier that can detect symptoms of depression in clinical interview transcripts. This is achieved through meeting two subgoals. The first subgoal is to identify which textual linguistic features of the transcripts are good indicators of depression. The next subgoal is to figure out if deep neural networks are a better choice for the task of depression detection compared to other commonly used machine learning algorithms from previous studies such as SVMs, logistic regression and decision trees.

2 Literature Review

2.1 Depression

This section introduces depression and discusses how the mental illness is manifested in individuals based on current literature.

2.1.1 What is depression?

beyondblue (2018) describes depression as the lack of positive emotions. If these feelings continue for more than two weeks, individuals may be diagnosed with a range of conditions. beyondblue (2018) details several of these conditions in which depression is a significant symptom:

- Major depression: is characterised as lack of interest in usual activities and a continued negative mood.
- Melancholia: is a severe form of depression where physical forms of depression exist. Someone suffering from this can have psychomotor retardation and complete loss of joy (anhedonia).
- Psychotic depression involves hallucinations, delusions and paranoia.
- Bipolar affective disorder involves experiencing periods of depression, and also periods of mania. Mania involves feeling excessively happy and having lots of energy to the extent of being deprived of sleep over several days.
- Cyclothymic disorder can be described as a milder form of bipolar affective depression.
- Dysthymic disorder is a milder form of depression that lasts for more than two years.
- Lastly, seasonal affective disorder is a mood disorder that has a pattern of mood fluctuations based on the season.

The following section discusses what signs medical and mental health professionals look out for when diagnosing an individual with these conditions.

2.1.2 Markers of Depression

Diagnostic criteria

The DSM-5 (American Psychiatric Association, 2013) details diagnostic criteria for major depressive disorder which is used by medical professionals when diagnosing the mental illness in individuals. The following is a list of symptoms of the illness as outlined by the criteria:

- Depressed mood for the majority of the day on most days
- Lack of interest and pleasure in most activities
- Decrease in appetite on most days
- Insomnia
- Lack of energy on most days
- Psychomotor retardation (slowing of movements)
- Feelings of guilt and of being worthless
- Lack of concentration
- Repeated thoughts of death and suicidal ideation

As mentioned in section 2.1.1, these symptoms must be present in an individual for a minimum of two weeks to be considered as being depression.

Questionnaires

Medical and mental health professionals usually use questionnaires to help them diagnose patients with mental illness. One such questionnaire is the Patient Health Questionnaire depression scale (PHQ). The PHQ-8 is an eight question variant of the PHQ and is derived from the diagnostic algorithm of the Diagnostic and Statistical Manual of Mental Disorders fourth edition (American Psychiatric Association & American Psychiatric Association. Task Force on DSM-IV., 2000) and is a measure to assess depression (Kroenke et al., 2009). Each question in the questionnaire focuses on a different theme of depression which are:

- Decreased pleasure in activities
- Feeling down
- Sleep disorders
- Loss of energy
- Significant change in appetite
- Feeling worthless
- Concentration problems
- Hyperactivity or lower activity

The patient can choose an answer between 0 and 3 inclusive for each question, representing the frequency that they feel the way that is described in the question. The scores of the individual answers are then summed up, and if the score is greater than or equal to 10, the individual can be diagnosed with depression. If the score is greater than or equal to 20, they can be diagnosed as being severely depressed. The DAIC-WOZ dataset that is used for this study includes patient's classifications as per the PHQ-8 questionnaire (Gratch et al., 2014). The dataset is described in more detail in Section 3.1.

Absolutist language

Beck (1979) explains a difference between the language of anxious individuals and depressed individuals. Anxious individuals tend to include words that denote uncertainty when speaking about a situations they wish to avoid. An example that is provided is "If I make a mistake, my boss may fire me." The presence of the word "may" shows that the speaker believes there is a possibility of a positive outcome. On the contrary, Beck (1979) notes that depression sufferers' language tends to be unconditional because they do not have hope for positive outcomes. Hence, they use absolutist language, which is made up of words, phrases and ideas that represent totality of probability or magnitude and lack the use of nuances (Al-Mosaiwi & Johnstone, 2018). The following are some examples of absolutist and non-absolutist words:

- Absolutist words: "always", "totally", "entirely"
- Non-absolutist words: "rather", "somewhat", "likely"

Al-Mosaiwi & Johnstone (2018) found that absolutist thinking was strongly correlated with the presence of depression. This study used the LIWC[1] NLP tool (Section 2.2.2) to examine absolutism at a linguistic level in posts from a depression and suicide ideation internet forum and a control internet forum. It was found that the depression and suicide ideation forums contained a significantly greater number of absolutist words. In fact, absolutist words were more strongly correlated with the illness than negative emotion words. This presented an opportunity for this study to use absolutist words for depression detection.

Function words

Function words are considered to be a good way to understand thinking processes and therefore should be observed to identify depression in addition to just observing content words. Forgas, Vincze, & László (2013) explain that many studies have found the use of the "function words" category in LIWC as more indicative of relationships with depression than the other categories. One such function word group that has been identified to be commonly used by depression sufferers is the set of first person singular pronouns such as "I" and "mine". Brown & Weintraub (1984) had originally found that there was an elevated use of first person singular pronouns in texts written by depression sufferers. Rude, Gortner, & Pennebaker (2004) confirmed this finding by identifying the same trend when analysing linguistic patterns in essays of students who had depression. Therefore, function words are used as linguistic features to detect depression in this study. Stirman & Pennebaker (2001) found that suicidal poets have an increased use of first-person singular pronouns and a decreased use of first-person plural pronouns, and suggested that it was because the suicidal individuals were showing a lack of interest in social relationships. This can extend to depression, as suicidal individuals often deal with depression. Similarly, Pyszczynski & Greenberg (1987) discussed that depressed individuals tend to fall into a pattern of self-focus that results in intensified negative outcomes. This self-focus might

[1] http://liwc.wpengine.com/

be reflected in language as the use of first person personal pronouns. For these reasons, using the use of first person personal pronouns in this study could potentially help identify the presence of depression.

2.2 Natural Language Processing

This section of the literature review will investigate the application of natural language processing in the areas of mental illness and depression detection.

2.2.1 Background

Liddy (2001) describes Natural Language Processing as a range of techniques used to get computers to understand natural language writing or speech. Research in the field falls into the areas of generation of text and understanding of text. This study will be mainly concerned with understanding rather than generation. A natural language processing system can be described as having many different levels. These levels are listed below.

- Phonology involves understanding speech sounds and the analysis of sound waves.
- Morphology involves breaking words down into components such as suffixes and prefixes to derive the meaning of an unknown word.
- Lexical level involves understanding the meanings of individual words.
- Syntactic level involves analysing words to understand the grammatical structure of a sentence.
- Semantic level involves analysing interactions between words to understand the meaning of the sentence.
- Discourse level involves taking into account the entire conversation and making connections between sentences.
- Pragmatic level involves knowing the context to extract additional meaning from texts.

Depression is primarily diagnosed through studying an individual's emotions and behaviour. These aspects are generally retrieved through interviews and conversations. Consequently, natural language processing has great potential to detect symptoms of mental illnesses by analysing texts and speech on the different levels mentioned above. Some of these techniques are summarised in the next section.

2.2.2 Natural language processing techniques

This section will discuss some of the natural language processing tools and techniques that have been proven to work well in the field of depression detection and related areas.

Linguistic Inquiry and Word Count (LIWC)

Linguistic Inquiry and Word Count (LIWC) is a text analysis tool that counts words in grammatically-relevant and psychologically-relevant categories across multiple texts. This tool has two components: a processing component and a dictionary. The processing component scans through the input text and compares each word with the words in the dictionary. It then matches the word with the categories it falls into as defined by the dictionary. Finally, the text is given rankings for each of the categories. The purpose of doing this is to identify the most prevalent categories of words which can give insights into attributes such as the sentiment and style of thinking in the text (Tausczik & Pennebaker, 2009). LIWC also allows the creation of custom dictionaries that can be used to create categories that are tailored specifically for depression detection (Pennebaker, Booth, & E, 2007).

LIWC is a popular tool especially those that involve dealing with social media data (Coppersmith, Dredze, & Harman, 2014; de Choudhury, Gamon, Counts, & Horvitz, 2013). de Choudhury et al. (2013) used LIWC to measure the positive affect and negative affect of each user by analysing their twitter posts by comparing the texts in the posts with LIWC's positive and negative categories. Coppersmith et al. (2014) compared their texts to several LIWC categories and identified a strong correlation between the categories of "Swear", "Anger" and "NegEmo" in social media posts. This suggests that these specific categories can be used to identify symptoms of depression. Mowery et al. (2017) found that the existing "Sad" category was insufficient for identifying all possible signs of negative emotions in their data, and therefore augmented the category with additional keywords demonstrating the tool's flexibility. Pampouchidou et al. (2016) used LIWC to not only for detecting sentiment, but also identify the presence of certain topics related to social processes such as reference to family, friends, and social interaction. This once again shows the ability to use this tool for a range of purposes. These studies demonstrate LIWC's usefulness and flexibility when analysing texts for the presence of mental illness symptoms, and therefore was used similarly in this study.

Word Lexicons

Lexicons are collections of words that are associated with a particular idea that can be used with many of the natural language processing techniques that will be discussed in Section 2.3. Several studies have built lexicons from relevant literature.

Yazdavar et al. (2017) used the PHQ-9 variant of the PHQ described in Section 2.1.2 to build a lexicon of depressive words related to the themes of each question in the PHQ-9. Words from the lexicon were then used as seed terms for a semi-supervised version of a Latent Dirichlet Allocation model described in Section 2.3.2. Wang et al. (2013) carried out a study to determine depression inclination of different social media posts. This involved constructing a lexicon of words extracted from a knowledge database called HowNet[2] to build a Chinese text lexicon. This lexicon was then used for sentiment analysis to detect depressive mood. Similarly, Kang et al. (2016) built a mood lexicon of positive and negative words with polarities associated with them from the SentiStrength[3] and Visual Sentiment Ontology dictionaries. This mood lexicon was then used to calculate scores for sentences to use as input features for their machine learning model.

[2] http://www.keenage.com/

[3] http://sentistrength.wlv .ac.uk/

Additionally, de Choudhury et al. (2013) scraped Yahoo! questions and answers related to depression to identify what individuals talking about depression usually talk about.

As constructing lexicons has been a popular choice in previous studies about depression. Consequently, word lexicons were constructed for certain themes that were suggested by the domain expert of this study to be potential symptoms of depression. This is described further in Section 3.2.3.

Term Frequency Inverse Document Frequency (TF-IDF)

Term Frequency weighted by Inverse Document Frequency is a tool used in information retrieval as a variation to simple word frequency detection. This method assigns each word a weighting based on the amount of information represented by it. Setting a threshold weighting would allow us to ignore insignificant words that do not have much impact on the sentiment or meaning of the texts (Tausczik & Pennebaker, 2009). Both de Choudhury et al. (2013) and O'Dea et al. (2015) used this technique to remove stop words. Stop words are words like 'the' and 'a' that appear often but do not offer anything of importance while processing text. Consequently, TF-IDF is used to filter out insignificant words in this study.

Sentiment Analysis

Sentiment analysis in natural language processing is the study of emotions and attitudes towards something in texts (Medhat, Hassan, & Korashy, 2014). Wang (2013) considers sentiment analysis as the most important part of depression detection. Sentiment analysis involves two main aspects which are subject-independent analysis and subject-dependent analysis. Subject-independent analysis involves determining the polarity of text without taking into account the meaning or context of the text. It tends to focus on abstract features that suggest the sentiment of a sentence as being positive, negative or neutral. Subject-dependent analysis however, takes the subject into account and determines the polarity by also studying target-specific features of a sentence. Both these forms of sentiment analysis would have to be considered for the project as features of the model being developed.

Previous studies (Howes, Purver, & McCabe, 2014; Yazdavar et al., 2017) demonstrate that sentiment analysis scores are a popular choice when building depression-detecting models as they used them as well. For this study, sentiment analysis is done using TextBlob[4], a library for the Natural Language Toolkit[5], that was also used by Yazdavar et al. (2017). TextBlob outputs the polarity for a body of text between 0 and 1, 0 being negative sentiment, and 1 being positive sentiment. TextBlob does sentiment analysis by using a lexicon of words with pre-set polarities, and averaging the polarity across all the words in the text.

[4] https://textblob.readthedocs.io/en/dev/
[5] https://www.nltk.org/

2.3 Machine Learning

This section will discuss some of the machine learning tools and techniques that have been proven to work well in the field of depression detection and related areas.

2.3.1 Deep Neural Networks (DNN)

It can be seen from the summary in section 2.5 that several studies have applied machine learning algorithms in their investigations. However, it can be observed that there has been no significant use of neural networks. Neural networks are types of machine learning algorithms that was inspired by human cognition (Mnih & Hinton, 2009). They have layers of nodes that do some computation, whose weights are tweaked in a training process involving a forward propagation of inputs through the network, followed by a backward propagation to adjust the weights. Deng, Hinton, & Kingsbury (2013) mention that these algorithms are popular due to the following reasons:

- Making these networks deeper by adding more layers of nodes makes them more powerful
- Improvements to computing power has made it possible to train these deep networks well
- Using many output units improves the networks performance

Deep neural networks are a type of neural networks that have several layers of interconnected nodes and have gained traction due to the fact that they outperform several other machine learning algorithms (Schmidhuber, 2015). Deep neural networks are already used extensively in the medical field, especially in medical imaging. Ciresan, Giusti, Gambardella, & Schmidhuber (2012) use DNNs to build a pixel classifier that addresses the problem of the automatic segmentation of neuronal structures. This is done to help map a three-dimensional structure of a brain. Neural networks have also been successful in the area of natural language processing such as the project on Hierarchical Distributed Language Model (Mnih & Hinton, 2009). Yang et al. (2017) carried out a similar study to the one being done that made use of deep neural networks for the purpose of depression detection on a similar dataset. The model being built in this study was to predict the PHQ-8 scores which is described in section 3, instead of just classifying the presence of depression. This study used a multimodal approach where three deep convolutional neural networks were used to train three types of features (audio, video and textual) and the outputs of each of the three were then used as inputs to a final deep neural network that outputted the predicted PHQ-8 score.

Deep neural networks have been chosen as the machine learning algorithm because the majority of studies in the area do not use this machine learning algorithm. SVMs and logistic regression seem to be the most popular options. However as discussed by Karmen et al. (2015), similarly severe cases do not necessarily populated connected regions in a multidimensional space. Different symptoms, for example insomnia and lack of interest in activities, can both indicate an equally severe case, but because the magnitude of different features are greater, they wound not populate the same region. This invalidates the use of SVMs for the task of depression detection. There are some studies that also use decision trees for the purpose of depression detection. This also is not appropriate because there is not a hierarchical structure for the symptoms of depression that allows us to divide patients into categories based on one symptom, then further subdivide them based on the next symptom.

2.3.2 Topic modelling

Another common technique used in these NLP studies is called topic modelling. This strategy views a document as combinations of latent topics. A topic is described as a distribution of co-occurring words (Rumshisky et al., 2016; Yazdavar et al., 2017). Latent Dirichlet Allocation (LDA) is an example of one of the more popular topic models and is discussed in the following subsection.

Latent Dirichlet Allocation (LDA)
The LDA is a recently devised generative model for text and other data. This model assumes that there are a certain number of underlying latent topics according to which documents are generated. Each topic is represented as a multinomial distribution over all the words in the vocabulary. The document is generated by mixing the topics and taking a few topics from it and then taking a few words from that mixture (Blei, Ng, & Jordan, 2003). A predetermined number of topics is required in order for the model to create topics. There are two main approaches for using LDA as input features in the study. As described by (Zhang et al., 2015) the algorithm can be trained on the same dataset, and using it to infer topics on each of the documents in the dataset, or trained on a completely different dataset and infer topics in the study's dataset.

Studies in depression detection have used this technique in the past. Rumshisky et al. (2016) utilised a 75-topic LDA model that was trained on the same dataset as was used in the study and was able to uncover topics linked to psychiatric symptoms of suicide and severe depression. The study was able to get some insights that the topics that were associated with people who were readmitted included eating disorders, infection and dementia. Similarly, Yazdavar et al. (2017) applied LDA on social media data. The study had hypothesized that analysing a user's topic preferences and word usage would allow them to monitor depression symptoms. They used an unsupervised version of LDA and extracted latent topics discussed by the users. It was found that the topics identified by LDA were not specific enough to be associated with depression, so the study had to apply a semi-supervised approach where some domain knowledge about depression was applied to the LDA by defining some rules to constrain the occurrence of some terms together. They were pretty much able to provide some depression-related seed words which then had words of semantically related terms cluster around them and form a topic. The seed words were extracted from literature about depression. (Zhang et al., 2015) used an LDA algorithm provided in the Mallet toolkit (Mccallum, 2002) to assess the depression-prediction ability of the two different LDA approaches. They also compared the performance of a machine learning model built by them using LDA topic features and LIWC features. The study revealed that using topics inferred from the same dataset performed just as well, and sometimes even better than a model trained on another dataset. This could because when they are inferred from the same dataset, the topics are relevant to the domain of the dataset. Therefore, topics are inferred from the same dataset in this study. Furthermore, the model trained with LDA features performed better than one trained with LIWC features, making it a good choice to use it as input features for this study.

2.3.3 Oversampling

There are times when the dataset provided is imbalanced, which can lead to issues such as over-fitting while training (Yang et al., 2016). The dataset being used in the study, the DAIC-WOZ depression dataset (described further in section 3.1), contains an imbalance with 72% of the data examples being from non-depressed individuals and 28% being from depressed individuals. In order to deal with this issue, random oversampling (He & Garcia, 2008) can be used. This method involves randomly selecting examples from the minority class, replicated them, and adding them to the total set of examples. This results in an equal number of examples in each class.

2.4 Social media data

Several studies in the area of depression detection and mental illness favour using social media data (Coppersmith et al., 2014; de Choudhury et al., 2013; Mowery et al., 2017; O'Dea et al., 2015; Sadeque, Xu, & Bethard, 2018). Yazdavar et al. (2017) claim that collecting social media data is favorable because unlike some other methods, such as recruiting participants and using questionnaires, it does not suffer from sampling bias and underrepresentation. Social media data on the other hand is able to capture users' behaviour and emotions that reflect their mental health in real-time and participants would not have experienced the cognitive bias that researchers are using their texts to identify signs of mental illness.

Social media is also suitable to identify changes in behaviour over time. Yazdavar et al. (2017) analysed the posts of self-reported depressed people by studying the change of the content that was posted by those particular people over the course of some time. Similarly, de Choudhury et al. (2013) also uses social media to diagnose major depression by identifying the behavioural attributes of an individual from their data from the duration of a year before the onset of their depression. The study measured certain depressive symptoms from the data about the social media posts. Depressive mood was measured by analysing the text using LIWC. The symptom of insomnia was measured by looking at metadata about what times the posts were made. Another symptom measured was decrease in engagement. This was achieved by analysing changes in the user's replies to posts, retweets, and links. This study was able to identify patterns of increased concern about medication, and increase in expression of religious thoughts in depressed individuals.

These findings show that social media does indeed offer insights into the behaviour and emotions of individuals, that would help researchers identify potential symptoms of depression. However, there are certain disadvantages of using social media. First of all, many of the studies that use social media depend on manual tagging of posts by the researchers (O'Dea et al., 2015; Zuorba, Olan, & Cantara, 2017) and using these tags to train their model. This is not ideal because depression is only officially diagnosed if symptoms persist for at least 2 weeks (American Psychiatric Association, 2013). This cannot be determined by observing individual social media posts. Other studies depend on self-reporting and specific posts of individuals (Coppersmith et al., 2014; de Choudhury et al., 2013; Yazdavar et al., 2017). This is also problematic as there is no guarantee of the honesty of the individual. These imperfections may have skewed the features used to train the models in each of these studies. These issues are avoided by using the DAIC-WOZ depression dataset. Each participant in that dataset was classified as being depressed or not through the use of clinical methods. This dataset is further discussed in Section 3.2.

2.5 Analysis of previous studies

Table 2.1

Summaries of previous studies grouped by the aspect of mental health they are primarily concerned with.

Study	*Dataset*	*NLP techniques*	*Features*	*Machine learning algorithms*	*Results*
Depression					
(Yazdavar et al., 2017)	23 million Twitter tweets	LDA, ssToT, sentiment analysis	Topics learned by the ssToT	ssToT	Average accuracy: 0.68; precision: 0.72
(Pampouchidou et al., 2016)	Depression Sub-Challenge of AVEC 2016	Depressive lexicon	Visual features, audio features, ratio of laughter, rate of speech, depression lexicon	Decision trees	F1 score: 0.52

(Rumshisky et al., 2016)	Psychiatric discharge notes	LDA	Age, gender, use of public health insurance, top-N bag of words features, most informative words from each record, topics learned by LDA	SVM	Area under the curve: 0.784
(Yang et al., 2016)	Depression Sub-Challenge of AVEC 2016	Word frequencies	Sleep status, emotions, PHQ score, audio and visual features	SVM, decision tree	F1 score: 0.857
(Coppersmith et al., 2014)	Social media data	N-grams, LIWC, sentiment analysis	N-grams, LIWC scores	Log-linear classifier	Language use, as measured by LIWC, is different between control and diagnosed users.
(de Choudhury et al., 2013)	Social media data	LIWC, TF-IDF, Depression lexicon built from Yahoo!	4 emotion categories and 22 linguistic style categories from LIWC, depression lexicon, and mentions of antidepressants	Logistic regression	Average accuracy: 0.7; precision: 0.74
(Wang et al., 2013)	Social media data	Depression lexicon built from HowNet	Polarity, use of pronouns, emoticons, interaction with other users interaction with others	Logistic regression	Precision: 0.8

(Yang et al., 2017)	Depression Sub-Challenge of AVEC 2017 (Similar to this study)	Paragraph vectors (vector representation of a paragraph)	Visual, audio and textual features, separate paragraph vector for each question of the transcript	Convolutional neural network, deep neural network	Root mean square error (RMSE): 5.974
Psychosis					
(Corcoran et al., 2018)	Transcripts from narrative-based protocol interviews and prompt-based protocol interviews	Latent semantic analysis, POS tagging	Semantic coherence, use of possessive pronouns	Logistic regression	Accuracy: 0.79
Suicidality					
(Fernandes et al., 2018)	Psychiatric patient records	Rule-based approach and machine learning approach, POS tagging, lexicon	Bag of words using POS tags, lexicon words frequency	SVM, random forest, naive bayes	Precision: 0.667; recall: 0.983; F1 score: 0.795
(O'Dea et al., 2015)	Twitter data	TF-IDF, word frequencies	Basic features of filtered word frequencies	SVM and logistic regression	Accuracy: 0.76; F1 score for classifying as strongly concerning: 0.64; F1 score for classifying as possibly concerning: 0.83; F1 score for classifying

					as safe to ignore: 0.62
(Zhang et al., 2015)	Social media data	LIWC, LDA	LDA topics inferred from the same dataset	Linear regression	RMSE: around 11

Table 2.1 contains several studies similar to this one and includes summaries of each one. Not all these studies are directly comparable due to use of different datasets and performance metrics, but certain gaps can be identified that are addressed by this study.

Firstly, it can also be seen that out of these 13 fairly recent studies on the detection of psychological conditions, only the one conducted by Yang et al. (2017) has utilised deep neural networks. This demonstrates that deep neural networks have not been used substantially for the task of depression. As this study focuses on the use of DNNs, it is able to address this gap.

Furthermore, the suitability of deep neural networks can be seen by comparing the models of Yang et al. (2017) and Zhang et al. (2015). Both these studies used RMSE as their metric. Although they used different input features and datasets, Yang et al. (2017) performed considerably better with an root mean square error (RMSE) almost half of the other study that used linear regression. This suggests that deep neural networks may be more suitable for the task of depression detection.

Additionally, some of these studies (Pampouchidou et al., 2016; Yang et al., 2016, 2017) trained their models on datasets from the Audio/Visual Emotion Challenge Workshop (AVEC) depression sub challenge. These datasets are similar to the one used in this study as they are recordings and sometimes transcripts of clinical interviews conducted for the diagnosis of depression. As can be seen from the table, none of these AVEC studies applied the LIWC tool or Latent Dirichlet Allocation to the dataset. This presents yet another opportunity to assess whether these tools will perform well on that particular dataset.

3 Methodology

This section describes the research approach employed to produce a tool that applies natural language processing and deep neural networks to identify symptoms of depression specifically in clinical interview transcripts. Figure 3.1 is a high-level diagram of this approach that will be discussed in further detail in the following sections.

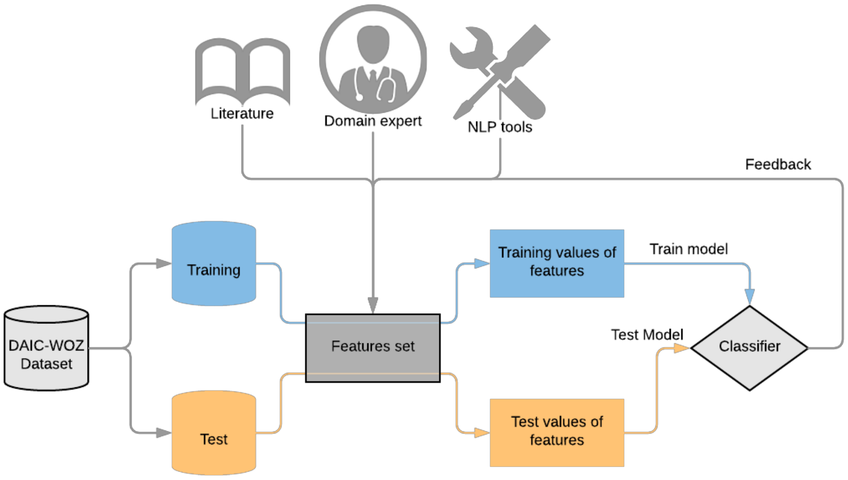


Figure 3.1: Diagram of the methodology used.

As this study involves building a tool that detects symptoms of depression in clinical interview transcripts, the DAIC-WOZ dataset was chosen to train and test the model.

3.1 DAIC-WOZ Depression Dataset

The DAIC-WOZ Depression Database is part of the larger Distress Analysis Interview Corpus (DAIC) (Gratch et al., 2014) and contains transcripts of clinical interviews with depressed and non-depressed individuals. Each interview involved questions designed to support the diagnosis of psychological distress. Each interviewee spoke to an computer animated interviewer, Ellie, who was controlled by a human operator in a different room. The transcripts are stored in Comma-separated Values (CSV) files, where each row is a phrase that was spoken. Usually, one row in the CSV file is the prompt from Ellie, followed by the response of the interviewee. The transcript CSV file includes the following attributes: start time, stop time, speaker and what was actually spoken. An example of a transcript from the dataset has been included in Appendix A. The dataset also includes audio recordings of the interviews as well as the facial expressions data of the interviewees. However, because the scope of the project is limited to textual linguistic features, only the transcripts will be utilised for this study.

The dataset was divided into three sets, a test set, training set and development set. However the classifications of only the training and development sets were accessible to this study. The training set is used to train the model with cross-validation, and the development set was used at the end to do a final evaluation. The training set has a total of 105 transcripts, 76 of which are from non-depressed individuals and 29 from depressed individuals whereas the development set has a total of 32 transcripts 20 of which are from non-depressed individuals and 12 from depressed individuals. It was planned to only use the training set while developing the model, and use the development set at the end of development to assess the resulting model. The DAIC-WOZ Depression Database will be referred to as the DAIC dataset hereinafter for the sake of convenience.

3.2 Feature set

The study begins with establishing an initial list of features that will be extracted from each transcript in the dataset and used as inputs whilst training the machine learning model. As shown in Figure 3.1, these features were established based on three sources:

- Research done on literature about depression and are described in Section 2.
- Discussions with one of the project's supervisors, who is a lecturer from the Faculty of Medicine and Health Sciences at Monash University who specialises in mental health and is the domain expert (E. Robinson, personal communication, August, 2018).
- Outputs of certain natural language processing tools.

Features from literature and NLP tools are chosen based on how popular and successful certain techniques have been in previous studies. Additional features are also determined following discussions of some interview transcripts of depressed individuals with E. Robinson (August, 2018). The following subsections describe each of these features.

3.2.1 Percentages of default LIWC categories

The first set of features are the percentage scores of the default LIWC categories. As discussed in Section 2.2, LIWC is a NLP tool that identifies the prevalence of certain categories of words in a given body of text (Pennebaker, Booth, & E, 2007). It outputs the percentage of words in the text that appear in each category for each transcript into a CSV file. In the first iteration of the model, all of the default LIWC categories except for punctuation categories will be considered as inputs. Punctuation categories are not included because the DAIC-WOZ dataset transcripts do not include any punctuation. A full list of these categories are available in Appendix C.

3.2.2 Sentiment scores

Sentiment analysis scores were also chosen to be part of the initial set of features. As discussed in Section 2.1, depression is characterised as the presence of a prolonged negative mood. Sentiment analysis is able to identify the polarity of the text, which may suggest if the speaker's mood is positive or negative. As described in Section 2.2.3.5, TextBlob[6] is used sentiment analysis. The polarity score is then used as a feature.

3.2.3 Word lexicons

The next group of features are word lexicons built for this study. As described in Section 2.1, increased usage of absolutist words was found among depressed

[6] https://textblob.readthedocs.io/en/dev/

individuals. Furthermore, LIWC allows for adding user-created categories. Therefore, a category of absolutist words was added to the tool. A list of these words has been included in Appendix B. Furthermore, other potential linguistic features of depressed texts were identified through discussions of the transcripts with E. Robinson (August, 2018). The mentioning of antidepressants was found to be a good indicator of a presence of depression. This was also supported and used in the study by de Choudhury, Gamon, Counts, & Horvitz (2013). As conducted by Shen et al. (2017), the Wikipedia page for "List of antidepressants" was scraped to get a list of antidepressant brands and names to determine if a transcript includes mentions of any of them. Furthermore, E. Robinson (August, 2018) also suggested to include psychotherapies, which were also extracted through web scraping from the Wikipedia page for "List of psychotherapies". A complete list of the antidepressants and psychotherapies scraped for this study have been included in Appendix D and Appendix E. An issue was that antidepressants and psychotherapies were mentioned so sparsely that its percentage score was always rounded down to zero. Consequently, the feature was modified from a percentage to a binary value where one represented the fact that it was used at least once in the transcript, and zero represented that it was never used.

3.2.4 LDA topics

Another set of features are the presence of topics extracted using a topic model. An LDA model from the Mallet package (Mccallum, 2002) described in Section 2.3.2 is used to learn topics. The model is trained on all the transcripts in the training set of the DAIC-WOZ dataset. The number of topics tried are between 10 and 200, incrementing by 10 topics each time. These numbers of topics are chosen by following on from Zhang et al. (2015) who used topic sizes between 10 and 300, also incrementing by 10 each time. The performance of the model when using each of these topic sets are then assessed to identify the best number of topics.

3.3 Feature analysis

Before starting, it is useful to investigate if there actually are any trends for the values of the features in the dataset. This was done by retrieving the values for each of the features for each transcript and identifying if there is a statistically significant difference between the depressed group and non-depressed group. A Student's t-test (du Prel, Röhrig, Hommel, & Blettner, 2010; Kim, 2015) is used for this because the data fits its criteria. The feature values are extracted from the transcripts that are from interviews that were conducted independent of each other. Furthermore, all participants in the dataset are randomly sampled from a population in California as described by Gratch et al. (2014), so it can be assumed that the feature values of both the depressed and non-depression groups are normally distributed and have the same variance. A threshold of 0.05 was chosen to identify if a feature was statistically different.

3.4 Preprocessing

The data has to be preprocessed in different ways depending on which feature is being extracted and which natural language processing tool is being used. To input the transcripts into LIWC, only the participant's parts of a transcript are of interest. Consequently, the "value" column of each row of the transcript that were spoken by the participant are extracted and compiled into a text file. Each of these text files are then inputted into LIWC. LIWC outputs a CSV file with percentages of each of its default dictionaries (Pennebaker et al., 2007). Ignoring the interviewer's prompts has the disadvantage of losing the context of each response, however it is simple to perform in the earlier iterations of the project. Furthermore, the NLP tools being used in the initial iteration do not have any use for the context, so it is acceptable to discard it for now. The same text file that is also processed using TextBlob, as discussed in section 3.2.2 to output the sentiment polarity.

The LIWC scores, sentiment score, LDA topic scores, and custom word presence features are then all compiled into a single feature array that are used as the input features for a particular transcript. Following this step, it was found that there was a large imbalance of positive and negative examples. 70.1% of the examples were classified as not depressed and only 29.9% were classified as depressed. This lead to results that suggested overfitting, and the model would often classify all transcripts as not depressed and still achieve a high accuracy. This issue was solved using the machine learning technique of oversampling that is described in Section 2.3.3. By replicating some data from the minority class (depressed), the performance was improved in subsequent iterations. A similar approach was taken by Yang et al. (2017) who used re-sampled audio files to achieve a balanced training set for their depression detecting model.

3.5 Model training and evaluation

The part of the DAIC-WOZ depression dataset that is accessible to this study only includes the training and development sets. Therefore, the development set is used as the test set, and the training set is used to train a deep neural network. Some applications and advantages of DNNs have been detailed in Section 2.3.1. The architecture of the deep neural network involves three-hidden-layers, with 300 nodes in each hidden layer that use Rectified Linear Unit (ReLu) activation functions. A ReLu activation function is chosen because of its success and popularity in various different applications (Arora, Basu, Mianjy, & Mukherjee, 2016). Furthermore, the network has a single output node that uses a sigmoid activation function. This activation function was chosen because this is a binary classification problem and sigmoid is bound between 0 and 1 (Weisstein, 2018), allowing us to assign 0 to mean a negative result and 1 to mean a positive result. The batch size was set to 5, and the number of epochs was adjusted based on the performance of the model and the features used. Using three-fold cross validation, the training set was divided into an appropriate number of subsets, where one of these subsets was used as an intermediary test set for the other sets. This will be iterated for each of the subsets. This cross-validation approach is appropriate for this dataset as it is commonly used on smaller datasets such as the DAIC dataset. The machine learning model will then be given the test data with the expected outcomes. The metrics that will be recorded for this study are the average accuracy and the F1 score of the model over each of the folds. The performance of the model will be assessed incrementally, by first training it just using LIWC features, then just using LDA features, and so on. The model will then be trained with combinations of feature groups, and finally with all the features. This approach allowed us to pinpoint which features and combinations are the best for training the model.

4 Results

4.1 Feature distributions

As part of the feature analysis stage of the project, the distributions of feature values for depressed and non-depressed groups of transcripts in the DAIC-WOZ training set were determined. A Student's t-test (Kim, 2015) was applied with the null hypothesis H_0 being that there is no statistically significant difference between means of the samples, and with a threshold of 0.05.

Table 4.1 contains the features that had the lowest p-values after applying the test, as well as some additional features that were believed to be symptoms of depression as described in Section 2.1. The first column is the name of the feature, the second column specifies which group the feature had a greater mean in, the second column contains the p-value as calculated from the statistical test, and the last column states whether the null hypothesis can be rejected or not. Furthermore, 10 LDA topics out of the 170 were found to also have significantly different means between the depressed and non-depressed groups. Some of the words in each of these topics are presented in Table 4.2.

Table 4.1

Student t-test statistical difference in different input features. Box-and-whiskers charts of the significantly different features are available at Appendix F.

Feature	*More prominent in*	*p-value*	*Reject H_0?*
LIWC (Pennebaker et al., 2007) categories			
negemo (Negative affect words)	Depressed	0.01834	yes
anx (Anxiety affect words)	Depressed	0.01243	yes
health (Health/illness biological processes)	Depressed	0.03024	yes
power (Core drives about power)	Depressed	0.02223	yes
leisure (Concerns about leisure)	Non-depressed	0.03985	yes
i (1st person personal pronouns)	Depressed	0.22624	no
absolutist	Non-depressed	0.65776	no
friend (Friends social words)	Non-depressed	0.06827	no
Sentiment			
Polarity	Non-depressed (positive sentiment more prominent in non-depressed)	0.11847	no
Custom word lexicons			
antidepressants	Depressed	0.00414	yes
psychotherapies	Depressed	0.00582	yes

Table 4.2

Words from some significantly different topics.

Topic	*More prominent in*	*p-value*	*Words*
4	Depressed	0.02171	reading book dogs relationship suppose schedule prefer dog knew type peak lots asked older requires dry socially awkward breaking professionally
18	Depressed	0.02952	**crying cry** books who's boundaries single benadryl stepmom spontaneous written **sniffle** laughing true signings custody apartment letting benefits isn't skills
26	Depressed	0.03219	**therapy health** vegetables raw wholesale eating basically opinion **brain nutrition** choices short i'd blame internet clarity experts vegan educators **mental**
46	Depressed	0.02527	quit face snow engaged evening steam physically terms eighty loner worker specific tells reintroduction exhaling portrays pluses ninetee blow nineties
122	Non-depressed	0.0393	**love** wanna great **film** mad **daughter** learned m's lamb **fiancee** sharp putting **festival** gut ugh **movies** bear encouraging seafood **zealand**
136	Depressed	0.01298	music business movies ago computers mouth regret yesterday sound you've scenery house ignoring yep administration he'd tossing engineering store qualities
141	Depressed	0.04751	saint louis pyramids guys you'd **surreal** unsure missouri calmed imagine town history playing

			guilty euphoria exhilarating unmanageable mind's **outweighed**
156	Depressed	0.03562	grown desert officer foster saturday stressed plan backgrounds actual quick grandmother schools rock born fall dad controlling kidnapped customs gully

4.2 Model Performance

Throughout the development of the model, the performance was measured. There are three groups of features: the LIWC features, LDA features, and the independent features (absolutist words, polarity, psychotherapies and antidepressants). As mentioned in Section 3.5, the model was trained incrementally, first using separate feature groups, and then using combinations of features. This allowed the feature set with the highest test F1 score and accuracy to be identified. The final performance metrics of each model are presented in Table 4.3. Similarly, in order to choose the best number of topics to use for the LDA model, various number of topics were extracted. The proportions of each of those topics were identified in the transcripts and models that detect depression were trained for them. The number of topics that were tried range from 10 to 200 topics. The performances of the models for each of these topic numbers are shown in Table 4.4.

Table 4.3

Classification accuracy and F1 scores of different set of features.

Features	*Cross-validation accuracy*	*Test F1 score*	*Test accuracy*
Baseline (random)	0.50000	0.50000	0.50000
LIWC	0.59841	0.26733	0.53125
LDA (170 topics)	0.82857	0.70186	0.77083
Antidepressants + absolutist words + polarity	0.65674	0.41742	0.61458
LIWC + LDA	0.86865	0.62488	0.68750
LIWC + antidepressants + absolutist words + polarity	0.68016	0.46118	0.61458
LDA + antidepressants + absolutist words + polarity	0.85238	0.48990	0.70833
LIWC + LDA + antidepressants + absolutist words + polarity	0.82024	0.49258	0.69792
LDA features + antidepressants + polarity + psychotherapies	0.89285	0.73205	0.8125
LDA + antidepressants + polarity + psychotherapies + statistically different LIWC	**0.86865**	**0.80879**	**0.84375**

Table 4.4

Classification accuracy and F1 scores for different number of topics in LDA.

Number of LDA topics	*Cross-validation Accuracy*	*Test F1 score*	*Test accuracy*
10	0.72103	0.37264	0.53125
20	0.75556	0.19001	0.51042
30	0.82857	0.40725	0.614583
40	0.84484	0.25733	0.53125
50	0.77857	0.26982	0.5625
60	0.86111	0.48883	0.67708
70	0.86071	0.29762	0.66667
80	0.84365	0.23617	0.60417
90	0.84484	0.18611	0.57292
100	0.80277	0.37554	0.57291
110	0.85317	0.41005	0.70833
120	0.86190	0.49897	0.72917
130	0.88571	0.27941	0.62500
140	0.80357	0.36970	0.58333
150	0.86071	0.04761	0.53125
160	0.78651	0.37323	0.64583
170	**0.82857**	**0.70186**	**0.77083**
180	0.84484	0.34524	0.63542
190	0.80397	0.10305	0.45833
200	0.83531	0.39889	0.59375

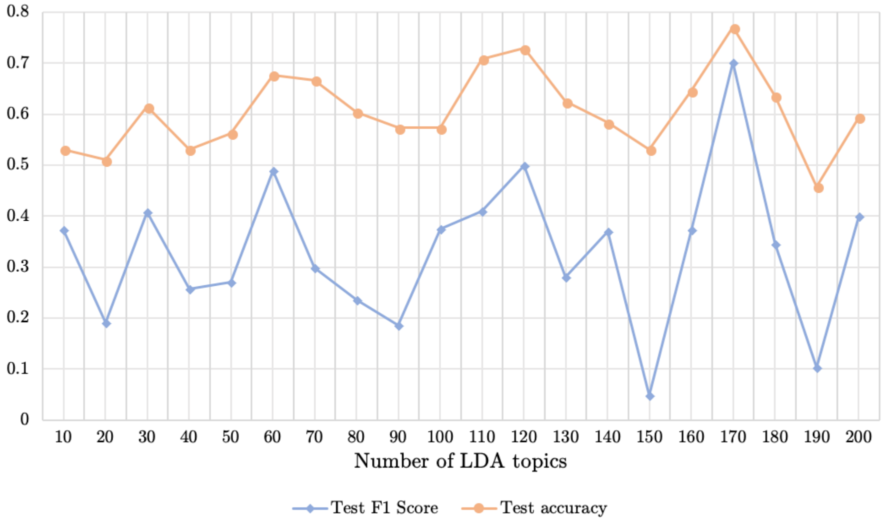


Figure 4.1: Graph of the test accuracy and test F1 score for different number of topics in LDA.

5 Discussion

5.1 Feature distributions

The feature distributions can be assessed using Table 4.1. There were several features that were significantly different between the depressed and non-depressed groups. Firstly, negative affect words were used significantly more by depressed participants than non-depressed participants. This confirms the findings of Morales & Levitan (2016) who hypothesised that depressed individuals perceive themselves and the world negatively, leading to an increased used of negative terms. Furthermore, a significantly greater use of anxious affect words was found in the transcripts of depressed individuals. This is consistent with a study by Sartorius et al. (1996) which found that the symptoms of anxiety and depression sometimes co-occur together.

A non-LIWC feature that was included in this model was the use of antidepressants. This feature was included due to the fact that Shen et al. (2017) had found a 165% greater use of them in conversations with depressed individuals. This trend was confirmed in this study as the use of antidepressants was also significantly higher in depressed transcripts. Psychotherapies were also found to be used much more by the depressed group, however, out of all the psychotherapy-related words and phrases that were compiled, only three in particular were mentioned in the transcripts: "psychotherapist", "psychotherapy" and "group therapy". Using a psychotherapy dictionary may therefore not be appropriate for this dataset.

There was also one other feature that appeared significantly more in non-depressed transcripts, which are words about personal concerns relating to leisure activities. The fact that depressed individuals mention leisure activities far less is consistent with the diagnostic criteria presented by the American Psychiatric Association (2013), which explains that a symptom of depression is a loss of interest in activities one would normally enjoy. Another feature that appeared more in the non-depressed group is friend-related social words, meaning that depressed individuals mentioned their social interactions with friends far less. This is consistent with beyondblue (2018) which stated that a major theme of depression is isolation and feeling like a burden to others, which leads to introverted and isolated behaviour. It also suggests that depressed individuals are speaking less about others, and more about themselves, and therefore thinking a lot more about themselves, which is consistent with findings of Pyszczynski and Greenberg (1987).

Furthermore, there were some features that were expected to have a significantly different distribution, but actually did not. As mentioned in Section 2.1.2, it was expected to see a significantly greater use of first person personal pronouns in the depressed group; however, this was not the case. This could just be because of the context of the transcripts and the questions asked were not able to reveal this underlying linguistic feature of depressed individuals. Additionally, Al-Mosaiwi and Johnstone (2018) had described that depressed individuals tend to use absolutist words far more than non-depressed individuals, but this trend was not supported by this study, and there was a very small difference between the two groups. This could once again just be because of the context and type of questions asked in the clinical interviews were unable to uncover this trend.

As stated in Section 4.1, there were also 10 LDA topics that showed significant differences between the depressed and non-depressed groups. Some of these topics are shown in Table 4.2. Topic 26 contains words related to health and was more prevalent in the depressed transcripts. This reflects the trend of the health LIWC category described earlier, which has words related to health and illness and appeared more in depressed transcripts. Topic 141 is also interesting as the words in it tend to reflect intense emotions. The greater use of them in depressed transcripts can suggest the presence of intense emotions in depressed individuals. Topic 122 is an example of an LDA topic that was more prevalent in non-depressed transcripts. The words in this topic seem to be representative of family and romantic relationships as well as some leisure activities. The presence of the word “zealand” could refer to New Zealand. Since these interviews were conducted with participants around the greater Los Angeles area, mentioning New Zealand could refer to travelling. This once again reflects the greater use of leisure activity words by non-depressed individuals.

The other topics that were found to be significantly different do not seem to have themes as the ones that have been discussed. This is expected as previous studies (Andrzejewski & Zhu, 2009; Ramage, Manning, & Dumais, 2011) found that LDA topics sometimes do not correlate with human reasoning.

5.2 Model performance

5.2.1 Overall performance

Out of the performances of the different number of LDA topics shown in Table 4.4, it is apparent that a model that used 170 topics as the input features performed the best, with a test accuracy of 0.77083 and a test F1 score of 0.70186. Consequently, it was chosen as the number of topics to be used as one of the input features of the model. A greater number of topics than 170 appears to consistently do poorly. As described by Zhang et al. (2015), this is probably because when the granularity of the topics is small, they are unable to capture symptoms of depression. Similarly, when the number of topics is too small, the granularity of the topics is too large and the topics are probably too broad to capture specific individual symptoms of depression.

When training the models with all of the feature sets, there were three feature sets that performed the best and had similar metrics. Each of these three had LDA features. The addition of LIWC features, the inclusion of polarity and the custom dictionaries (antidepressants and psychotherapies words) seem to have increased the performance, but not to a significant degree. Consequently, LDA can be assumed as having the highest impact on the learning. LIWC categories have little impact on the model performance, suggesting that latent topics may be better indicators of topics than word frequencies. This is consistent with the findings of Zhang et al. (2015) which proposed that the topics extracted from the texts can contain more information than LIWC lexicons. The best classification accuracy of 0.84375 and best F1 score of 0.80879 was achieved when the input features were made up of LDA topics, polarity score, mentions of antidepressants, mentions of psychotherapies and LIWC scores for statistically different categories. This accomplished the first subgoal of this project, which was to determine a good set of features for depression detection.

5.2.2 Comparison with related studies

Due to time constants, different machine learning algorithms were not trained in order to compare their performances with the deep neural network. However, the second subgoal of the study can still be met by comparing the results of this study to other studies that trained different machine learning algorithms using similar datasets or similar features. This section will perform this comparison to verify whether deep neural networks are indeed more suited for the task of depression detection than other algorithms. The DAIC-WOZ depression dataset and similar datasets have actually been used for one of the AVEC sub challenges over the years. However, one must realise that the challenges involve using a training, development and test set. Our study was only able to access the training and dev set, resulting a smaller number of training examples which is important to consider when comparing with the other models.

The study by Yang et al. (2017) was similar to our study, but built a model that predicted the PHQ-8 score of a clinical interview instead of just classifying it as showing symptoms of depression. This means the model would have to predict a score between 0 and 24, which is far greater number of classes than simple binary classification as done in our study. Predicting the PHQ score is useful it can let us know how severe the depressive symptoms are, as a score higher than 20 is considered severely depressed as described in section 2.1.2. It performed well and achieved a test set RMSE of 5.974 which was better than the competition's baseline of 6.970. Furthermore, that study used a multimodal approach, where they took into account audio, video and textual features and gender-specific models, compared to just textual features in our study. This presents opportunities for extensions, further described in Section 6. They also achieved high performance on just using textual features, however the reason for this is that their features were specific to the questions asked based on the PHQ-8. As the study illustrates, their best performing feature was the n-gram of a specific answer of the participant, where the participant answered yes or no to a question of if they were diagnosed with depression. This means the same model would not be able to be generalised to other datasets, or even other questions in the same dataset. Our model instead does not focus on such specific answers, and retrieves features over the entire transcript. This does lose the context of each of the answers of the user which can have great semantic value, but makes the model more generalisable.

The study by Zhang et al. (2015) uses LDA and LIWC when training their models. Our study had slightly different results when assessing the performance using different number of LDA topics learned from our dataset. Our model has best performance for 170 topics, with the performance consistently lower for numbers under 100 and over 100. In the study by Zhang et al. (2015), the performance of the model that used topics learned from their dataset was consistent for any number of topics, with a slight decrease in performance when a relatively large number of topics is used. Furthermore, both our study that used clinical interview transcripts and this study that used social media found that LIWC percentages do not contribute much to the performance of the model. LIWC has been used heavily for psychological analysis, but these two studies show that they are not as useful to train depression-detecting machine learning models in both written (social media posts) and spoken (interview transcripts) texts.

Another study that used social media data was the one by Kang et al. (2016) which used textual features to identify depressive users using an SVM. Each training example for their study was a sentence, and the feature vector was made up of different scores for interjections, negativity, adjectives, nouns, verb, adverbs and punctuation in the sentence. Their textual features were therefore far fewer and simpler than what was used in this study. They were able to achieve an accuracy of 0.84547 and an F1 score of 0.84388 across the datasets they used, which is slightly better than the results of this study. However, the difference between written social media posts and transcripts of spoken clinical interviews must be taken into consideration. Their model may not work well on our dataset, and our model in turn may not work well on their dataset.

Resnik et al. (2013) conducted a similar study to ours, where LIWC and LDA features were used in a linear regression model for the purpose of depression detection. The study's model was able to produce a best F1 score of 0.5 when utilising a 50-topic LDA model features along with LIWC scores. The model built in our study was able to achieve an F1 score of 0.62488 when using just LDA model features and LIWC scores. As the deep neural network from this study performed better, it suggests that deep neural networks are better machine algorithms to use for this task. However, it is important to realise that Resnik et al. (2013) used a dataset of essays from students, that were probably have a greater variability of content than a structured as a clinical interview in which the interviewers will attempt to ask similar questions to each participant.

Pampouchidou et al. (2016) conducted a depression classification study on a similar dataset. Unlike our study, they focused on audio and visual features and used a decision tree for classification. They were able to achieve a F1 score 0.52 to identify depressed individuals. This lower than the F1 score achieved by our study, once again suggesting that deep neural networks are superior.

Through this comparison, it can be observed that this study's model will be more generalisable than other depression detection studies as it does not depend on the structure of the text. Furthermore, this comparison also serves as evidence that deep neural networks might be better algorithms to use for the purpose of depression detection.

6 Conclusion & Future Work

The aim of this study was to develop an automated classifier that applies deep neural networks and natural language processing to detect symptoms of depression in clinical interview transcripts. The study was therefore centered around the DAIC-WOZ depression dataset of clinical interviews. By studying the existing literature, consulting with a domain expert and using natural language processing tools, a set of features that could be used to represent each transcript in the dataset is established. Out of this set, this research uncovered that certain features such as the use of negative emotion words, use of health and illness-related words, mentioning of antidepressants, and mentioning of psychotherapies appeared significantly more in depressed transcripts, whereas the use of leisure activities related words appeared significantly more in non-depressed transcripts. After several iterations, an accuracy of 0.84375 and an F1 score of 0.80879 was achieved.

The best set of features to achieve these metrics was made up of the following: LDA topics, polarity score, mentions of antidepressants, mentions of psychotherapies and LIWC scores for statistically different categories. As the accuracy and F1 score are far higher than what would have achieved from randomly classifying, it can be concluded that the model does indeed detect symptoms of depression successfully. Despite this, there are certain opportunities and issues that can be addressed in order to take full advantage of the data and potentially improve the performance of the model.

First of all, this study does not take into account the gender of the participants. Stratou et al. (2013) assessed a gender-dependent model against a gender-independent model to detect depression using non-verbal features, and found that the gender-dependent models performed with a 5 to 25% improved accuracy. Yang et al. (2016) followed on from that study and generated decision trees based on answers participants gave to different questions asked in the DAIC-WOZ dataset (Gratch et al., 2014). The result showed that different decision trees were created for males and females, which suggests that there is indeed some difference in the symptoms of depression shown by each gender. The DAIC-WOZ dataset classifies each participant as either a male or female, so it is possible to train separate gender-specific models. This provides an opportunity for a study that uses similar input features as this research, but trains them separately on a dataset of just male participants' transcripts and another of just female participants' transcripts. However, the amount of data in the current DAIC-WOZ dataset is relatively low to train a deep neural network. Dividing the dataset based on gender would further reduce the training set size, and the number of data examples would be too few to train two separate gender dependent neural networks.

This progresses to the next limitation of the study, which is the small amount of data being used to train the model. It is possible that there was some overfitting due to the fact that the size of the dataset was small. Using a larger dataset of clinical interviews may yield better performance results for the model and more representative distributions of the features discussed in Section 5.1.

Furthermore, this study does not make use of the audio files and the facial features provided in the DAIC-WOZ depression dataset. These additional data could potentially improve the performance of the model if used along with the textual features that were determined in this study. Yang et al. (2017) built a model that predicted PHQ-8 scores on the AVEC 2017 depression sub challenge dataset, which is similar to the DAIC-WOZ dataset. They used a multimodal approach, where they built three separate deep convolutional neural networks; one for video features, one for audio features and one of textual features. They then combined them to predict a PHQ-8 score, performing far better than the baselines. Similarly, Morales and Levitan (2016) combined both the audio and textual features to build a SVM model. They then used a feature selection algorithm to select the optimal subset of features, and found that this best selection was a combination of textual and audio features. These studies present an opportunity to improve the model built in this study by incorporating audio and video features.

Another limitation of this study is that it does not take into account the temporal nature of depression. As described in the DSM-5 (American Psychiatric Association, 2013), depression can be diagnosed once symptoms persist for at least two weeks. In order to properly diagnose depression, this temporal aspect must be considered. Yazdavar et al. (2017) address this by using a modified LDA algorithm, called *semi-supervised topic modelling over time* (ssToT), that is semi-supervised as it allows adding seed words so that topics are formed around them, and allows monitoring the change to these topics over time. Karmen et al. (2015) takes a different approach to this problem, and instead looks for certain words and phrases that represent symptoms of depression, and searches for frequency terms associated with them. These frequency terms such as “never”, “sometimes”, “often” and “always” give a relative idea about how long the individual has had the symptoms and whether they qualify as being depressed. Both approaches are potential extensions to the current study to incorporate the temporal nature of depression into the model.

Once these limitations have been satisfied, the model will be able to capture depression sufficiently in order to predict its presence at a satisfactory level. This model should be used primarily for screening patients and complementing the diagnosis process rather than replacing mental health professionals entirely. It is also important to realise that this model will only be able to notify the medical professional of potential symptoms, not assess the severity of the mental illness or suggest what the next steps to take are. These would require the expertise of the medical professional. Overall, the study was able to uncover some linguistic features that are good signs of depression and bring us one step closer to saving time and resources for medical professional, and more importantly, potentially saving the lives of individuals suffering from depression.

References

Al-Mosaiwi, M., & Johnstone, T. (2018). In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. *Clinical Psychological Science*, 216770261774707.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5ff).* American Psychiatric Pub.

American Psychiatric Association, & American Psychiatric Association. Task Force on DSM-IV. (2000). *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition: DSM-IV-TRff.* American Psychiatric Association.

Andrzejewski, D., & Zhu, X. (2009). Latent dirichlet allocation with topic-in-set knowledge. *Proceedings of the NAACL HLT 2009*. Retrieved from https://dl.acm.org/citation.cfm?id=1621835

Arora, R., Basu, A., Mianjy, P., & Mukherjee, A. (2016). *Understanding Deep Neural Networks with Rectified Linear Units*. *arXiv [cs.LG]*. Retrieved from http://arxiv.org/abs/1611.01491

Beck, A. T. (1979). *Cognitive Therapy and the Emotional Disorders*. Penguin.

beyondblue. (2018). *Anxiety and depression: An information booklet*.

Blanken, G., Dittmann, J., Grimm, H., Marshall, J. C., & Wallesch, C.-W. (1993). *Linguistic Disorders and Pathologies: An International Handbook*. Walter de Gruyter.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research: JMLR*, (3), 993–1022.

Brown, S. R., & Weintraub, W. (1984). Verbal Behavior: Adaptation and Psychopathology. *Political Psychology*, *5*(1), 107.

Ciresan, D., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in Neural Information Processing Systems*, 2843–2851.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational*

Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. https://doi.org/10.3115/v1/w14-3207

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., … Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry: Official Journal of the World Psychiatric Association* , *17*(1), 67–75.

de Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. Presented at the Association for the Advancement of Artificial Intelligence.

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. https://doi.org/10.1109/icassp.2013.6639344

du Prel, J.-B., Röhrig, B., Hommel, G., & Blettner, M. (2010). Choosing Statistical Tests. *Deutsches Aerzteblatt Online*. https://doi.org/10.3238/arztebl.2010.0343

Fernandes, A. C., Dutta, R., Velupillai, S., Sanyal, J., Stewart, R., & Chandran, D. (2018). Identifying Suicide Ideation and Suicidal Attempts in a Psychiatric Clinical Research Database using Natural Language Processing. *Scientific Reports*, *8*(1), 7426.

Forgas, J. P., Vincze, O., & László, J. (2013). *Social Cognition and Communication*. Psychology Press.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherere, S., Nazarian, A., … Morency, L. P. (2014). The Distress Analysis Interview Corpus of human and computer interviews. In *Language Resources and Evaluation Conference* (pp. 3123–3128).

He, H., & Garcia, E. A. (2008). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, (9), 1263–1284.

Howes, C., Purver, M., & McCabe, R. (2014). Linguistic indicators of severity and progress in online text-based therapy for depression. In *Proceedings* *of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 7–16).

Kang, K., Yoon, C., & Kim, E. Y. (2016). Identifying depressive users in Twitter using multimodal analysis. In *2016 International Conference on Big Data and Smart Computing (BigComp)* (pp. 231–238).

Karmen, C., Hsiung, R. C., & Wetter, T. (2015). Screening Internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods. *Computer Methods and Programs in Biomedicine*, *120*(1), 27–36.

Kim, T. K. (2015). T test as a parametric statistic. *Korean Journal of Anesthesiology*, *68*(6), 540–546.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1-3), 163–173.

Liddy, E. D. (2001). Natural Language Processing. In *Encyclopedia of Library and Information Science, 2nd Ed*. Marcel Decker, Inc.

Mccallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. Retrieved from http://www.citeulike.org/group/3030/article/1062263

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113.

Mnih, A., & Hinton, G. E. (2009). A scalable hierarchical distributed language model. *Advances in Neural Information Processing Systems*, 1081–1088.

Morales, M. R., & Levitan, R. (2016). Speech vs. text: A comparative analysis of features for depression detection systems. In *2016 IEEE Spoken Language Technology Workshop (SLT)* (pp. 136–143).

Mowery, D., Smith, H., Cheney, T., Stoddard, G., Coppersmith, G., Bryan, C., & Conway, M. (2017). Understanding Depressive Symptoms and Psychosocial Stressors on Twitter: A Corpus-Based Study. *Journal of Medical Internet Research*, *19*(2), e48.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions*, *2*(2), 183–188.

Pampouchidou, A., Marias, K., Yang, F., Tsiknakis, M., Simantiraki, O., Fazlollahi, A., … Simos, P. (2016). Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*. https://doi.org/10.1145/2988257.2988266

Pennebaker, J. W., Booth, R. J., & E, F. M. (2007). *Operator's Manual Linguistic Inquiry and Word Count: LIWC2007*.

Pyszczynski, T., & Greenberg, J. (1987). Self-regulatory perseveration and the depressive self-focusing style: a self-awareness theory of reactive depression. *Psychological Bulletin*, *102*(1), 122–138.

Ramage, D., Manning, C. D., & Dumais, S. (2011). Partially Labeled Topic Models for Interpretable Text Mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 457–465). New York, NY, USA: ACM.

Resnik, P., Garron, A., & Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1348–1353).

Rihmer, Z. (2001). Can better recognition and treatment of depression reduce suicide rates? A brief review. *European Psychiatry: The Journal of the Association of European Psychiatrists*, *16*(7), 406–409.

Robinson, E. (2018, August). Personal interview.

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121–1133.

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. M., McCoy, T. H., & Perlis, R. H. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational Psychiatry*, *6*(10), e921.

Sadeque, F., Xu, D., & Bethard, S. (2018). Measuring the Latency of Depression Detection in Social Media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18*. https://doi.org/10.1145/3159652.3159725

Sartorius, N., Ustün, T. B., Lecrubier, Y., & Wittchen, H. U. (1996). Depression comorbid with anxiety: results from the WHO study on psychological disorders in primary health care. *The British Journal of Psychiatry. Supplement*, (30), 38–43.

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks: The Official Journal of the International Neural Network Society*, *61*, 85–117.

Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., … Zhu, W. (2017). Depression detection via harvesting social media: A multimodal dictionary learning solution. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)* (pp. 3838–3844).

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, *63*(4), 517–522.

Stratou, G., Scherer, S., Gratch, J., & Morency, L. (2013). Automatic Nonverbal Behavior Indicators of Depression and PTSD: Exploring Gender Differences. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 147–152).

Tausczik, Y. R., & Pennebaker, J. W. (2009). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., & Bao, Z. (2013). A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. In *Lecture Notes in Computer Science* (pp. 201–213).

Weisstein, E. W. (2018). Sigmoid Function. *Http://mathworld.wolfram.com/*. Retrieved from http://mathworld.wolfram.com/SigmoidFunction.html

Whitaker, R. (2005). Anatomy of an Epidemic: Psychiatric Drugs and the Astonishing Rise of Mental Illness in America. *Ethical Human Sciences and Services: An International Journal of Critical Inquiry*, *7*(1), 23–35.

Yang, L., Jiang, D., He, L., Pei, E., Oveneke, M. C., & Sahli, H. (2016). Decision Tree Based Depression Classification from Audio Video and Language Information. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*. https://doi.org/10.1145/2988257.2988269.

Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M. C., & Sahli, H. (2017). Multimodal Measurement of Depression Using Deep Learning Models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC '17*. https://doi.org/10.1145/3133944.3133948

Yazdavar, A. H., Al-Olimat, H. S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., … Sheth, A. (2017). Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media. *Proceedings of the ... IEEE/ACM International Conference on Advances in Social Network Analysis and Mining. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, *2017*, 1191–1198.

Zhang, L., Huang, X., Liu, T., Li, A., Chen, Z., & Zhu, T. (2015). Using Linguistic Features to Estimate Suicide Probability of Chinese Microblog Users. In *Human Centered Computing* (pp. 549–559). Springer International Publishing.

Zuorba, H. D., Olan, C. L. O., & Cantara, A. D. (2017). A Framework for Identifying Excessive Sadness in Students through Twitter and Facebook in the Philippines. In *Proceedings of the International Conference on Bioinformatics Research and Applications 2017 - ICBRA 2017*. https://doi.org/10.1145/3175587.3175600

Appendix

Appendix A

Portion of one of the conversation transcripts from the DAIC-WOZ Depression Database.

Start time	*Stop time*	*Speaker*	*Value*
32.475	42.343	Ellie	hi i'm ellie thanks for coming in today i...
44.341	48.102	Ellie	how are you doing today
50.132	52.004	Participant	i'm doing well how about you
52.776	54.095	Ellie	i'm great thanks

Appendix B

List of absolutist words used as the custom LIWC dictionary for this study. The Asterix (*) represents a wild-card character sequence. For example, "definite*" includes "definite" and "definitely".

absolut*
all
always
complet*
constant*
definite*
entire
ever*
full*
must
never
nothing
total*
whole

Appendix C

List of default LIWC categories except the punctuation categories as retrieved from http://liwc.wpengine.com/compare-dictionaries/.

Analytical Thinking
Clout
Authentic
Emotional Tone
Words per sentence
Words greater than 6 letters
Dictionary words
Function Words
Total pronouns
Personal pronouns
1st pers singular
1st pers plural
2nd person
3rd pers singular
3rd pers plural
Impersonal pronouns
Articles
Prepositions
Auxiliary verbs
Common adverbs
Conjunctions
Negations
Regular verbs
Adjectives
Comparatives
Interrogatives
Numbers
Quantifiers
Affect Words
Positive emotion
Negative emotion
Anxiety
Anger
Sadness
Social Words
Family
Friends
Female referents
Male referents
Cognitive Processes
Insight
Cause
Discrepancies
Tentativeness
Certainty
Differentiation
Perpetual Processes
Seeing
Hearing
Feeling
Biological Processes
Body
Health/illness
Sexuality
Ingesting
Core Drives and Needs
Affiliation
Achievement
Power
Reward focus
Risk/prevention focus
Past focus
Present focus
Future focus
Relativity
Motion
Space
Time
Work
Leisure
Home
Money
Religion
Death
Informal Speech
Swear words
Netspeak
Assent
Nonfluencies
Fillers

Appendix D

List of psychotherapies as retrieved from https://en.wikipedia.org/wiki/List_of_psychotherapies.

- abreaction therapy
- accelerated experiential dynamic psychotherapy
- acceptance and commitment therapy
- addiction psychiatry
- adlerian therapy
- adventure therapy
- analytical psychology
- animal-assisted therapy
- art therapy
- attachment therapy
- attachment-based psychotherapy
- attachment-based therapy
- attack therapy
- autogenic training
- aversion therapy
- behavior modification
- behavior therapy
- behavioral activation
- bibliotherapy
- biodynamic psychotherapy
- bioenergetic analysis
- biofeedback
- biological psychiatry
- body psychotherapy
- brief psychotherapy
- chess therapy
- child and adolescent psychiatry
- child psychotherapy
- classical adlerian psychotherapy
- client-centered psychotherapy
- co-counselling
- cognitive analytic therapy
- cognitive behavior therapy
- eye movement desensitization and reprocessing
- family constellations
- family therapy
- feminist therapy
- focusing (psychotherapy)
- forensic psychiatry
- freudian psychotherapy
- functional analytic psychotherapy
- future-oriented therapy
- geriatric psychiatry
- gestalt theoretical psychotherapy
- gestalt therapy
- grief counseling
- group analysis
- group therapy
- guided affective imagery
- hakomi
- holding therapy
- holotropic breathwork
- human givens
- humanistic psychology
- hypnotherapy
- immuno-psychiatry
- inner relationship focusing
- integral psychotherapy
- integrative body psychotherapy
- integrative psychotherapy
- intensive short-term dynamic psychotherapy
- internal family systems model
- interpersonal psychoanalysis
- interpersonal psychotherapy
- palliative medicine
- parent management training
- parent–child interaction therapy
- pastoral counseling
- person-centered therapy
- play therapy
- poetry therapy
- positive psychology
- positive psychotherapy
- postural integration
- primal integration
- primal therapy
- process oriented psychology
- process psychology
- progressive counting
- prolonged exposure therapy
- provocative therapy
- psychedelic therapy
- psychiatrist
- psychoanalysis
- psychodrama
- psychodynamic psychotherapy
- psychopharmacology
- psychosomatic
- psychosurgery
- psychosynthesis
- psychotherapy
- pulsing
- rational emotive behavior therapy
- rational living therapy
- reality therapy
- rebirthing-breathwork
- recovered-memory therapy
- reichian therapy
- relational-cultural therapy
- relationship counseling

cognitive
neuropsychiatry
cognitive therapy
coherence therapy
collaborative therapy
compassion focused
therapy
concentrative
movement therapy
contemplative
psychotherapy
contextual therapy
conversational model
conversion therapy
cross-cultural
psychiatry
dance therapy
daseinsanalysis
depth psychology
descriptive psychiatry
developmental
disability
developmental needs
meeting strategy
dialectical behavior
therapy
drama therapy
dreamwork
dyadic developmental
psychotherapy
eating disorders
eclectic psychotherapy
ecological counseling
emergency psychiatry
emotional freedom
expressive therapies
journal therapy
jungian psychotherapy
liaison psychiatry
logic-based therapy
logotherapy
marriage counseling
mentalization-based
treatment
metacognitive therapy
method of levels
milieu therapy
military psychiatry
mindfulness-based
cognitive therapy
mindfulness-based
stress reduction
mode deactivation
therapy
morita therapy
motivational
interviewing
multimodal therapy
multisystemic therapy
multitheoretical
psychotherapy
music therapy
narrative therapy
neuropsychiatry
nonviolent
communication
nouthetic counseling
nude psychotherapy
object relations
psychotherapy
orthodox psychotherapy
pain medicine
remote therapy
rogerian psychotherapy
sandplay therapy
schema therapy
sensorimotor
psychotherapy
sex therapy
sexual identity therapy
sleep medicine
social therapy
solution focused brief
therapy
somatic experiencing
somatic psychology
status dynamic
psychotherapy
structural family
therapy
supportive
psychotherapy
systematic
desensitization
systemic therapy
t-groups
therapeutic community
thought field therapy
transactional analysis
transference focused
psychotherapy
transpersonal
psychology
transtheoretical model
twelve-step programs
vegetotherapy
wilderness therapy
writing therapy

Appendix E

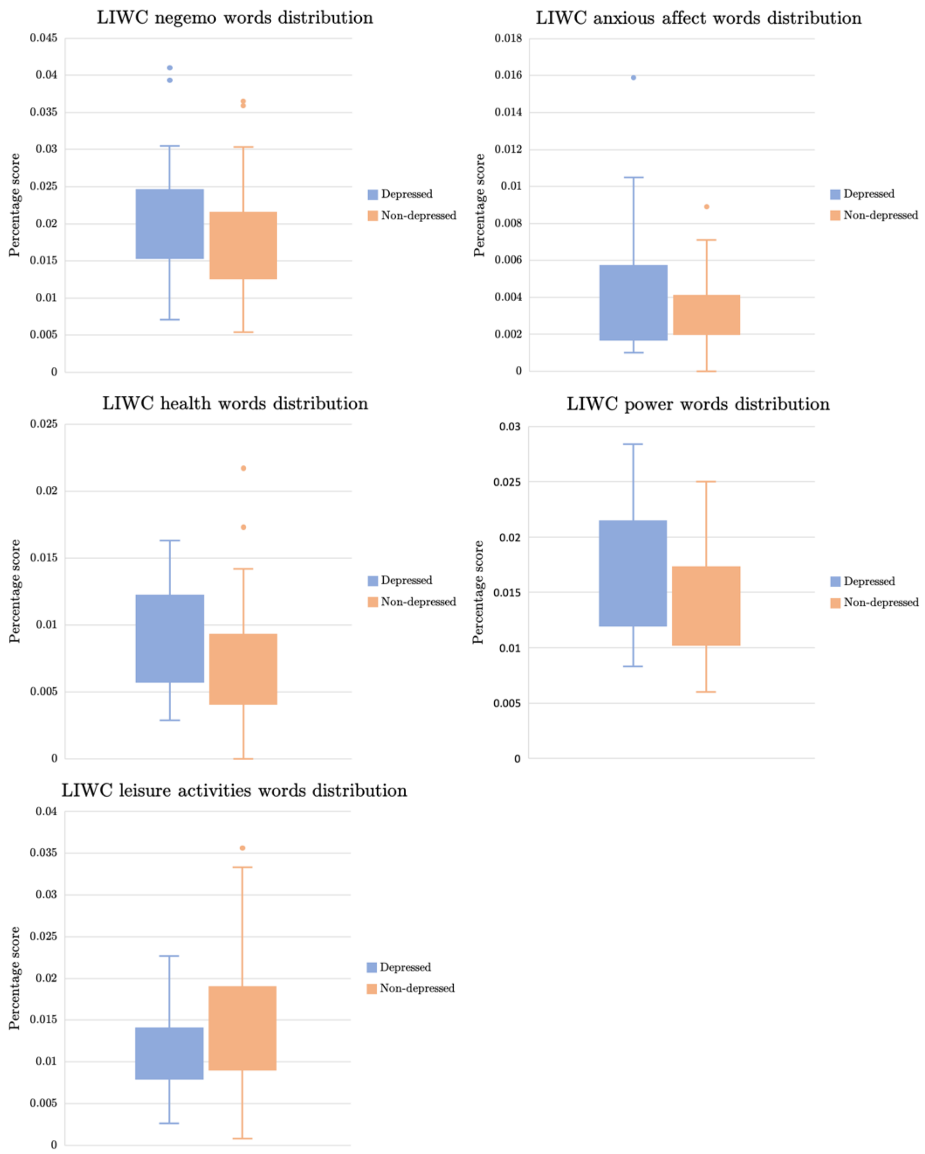
List of antidepressants as retrieved from https://en.wikipedia.org/wiki/List_of_antidepressants.

abilify
adapin
ademetionine
agedal
agomelatine
alnert
ambivalon
amineptine
aminomine
amioxid
amisulpride
amitriptyline
amitriptylinoxide
amoxapine
anafranil
aripiprazole
asendin
atomoxetine
aurorix
aventyl
azafen/azaphen
benmoxin
bifemelane
bolvidon
brexpiprazole
bupropion
buspar
buspirone
butriptyline
carbamazepine
cariprazine
caroxazone
celeport
celexa
chloride
cipralex
cipramil
citalopram
clomipramine
clozapine
clédial
coaxil
conflictan
cymbalta
phenelzine
pheniprazine
hypericum
imipramine
imipraminoxide
indalpine
indeloxazine
indopan
inkazan
insidon
iprindole
iproclozide
iproniazid
isocarboxazid
isoniazid
istonil
ixel
jatrosom
ketalar
ketamine
lamotrigine
latuda
levomilnacipran
lexapro
linezolid
lithium
lithobid
lofepramine
lomont
lucelan
ludiomil
lurasidone
lustral
luvox
manerix
maprotiline
marplan
mebanazine
medifoxamine
melitracen
melixeran
metapramine
metatone
methyltryptamine
pertofrane
phenoxypropazine
pipofezine
pirazidol
pirlindole
pivagabine
pivhydrazine
pivoxil
pristiq
propizepine
prothiaden
protriptyline
prozac
quetiapine
quinupramine
reboxetine
remeron
rexulti
risperdal
risperidone
rubidium
rubinorm
safrazine
savella
sediel
selegiline
seroquel
seroxat
sertraline
serzone
setiptiline
sinequan
sintamil
sjw
solian
stablon
stelapar
surmontil
symbyax
tandospirone
tecipul
tedizolid
teniloxazine
tetrindole

deanxit
demexiptiline
desipramine
desvenlafaxine
desyrel
dibenzepin
dimetacrine
dixeran
dosulepin
doxepin
duloxetine
dutonin
edronax
effexor
elavil
eldepryl
elen
elronon
emsam
endep
eprobemide
equilibrin
escitalopram
eskalith
esketamine
etafron
etoperidone
etryptamine
faverin
fetzima
fluoxetine
flupentixol
fluvoxamine
gamanil
gerdaxyl
humoryl
metralindole
metryptamine
mianserin
milnacipran
minaprine
mirtazapine
moclobemide
nardil
nefadar
nefazodone
nialamide
nitroxazepine
nogedal
noin
nomifensine
norpramin
nortriptyline
norval
noveril
noxiptiline
octamoxin
olanzapine
opipramol
optimax
oxaflozane
oxcarbazepine
oxitriptan
paliperidone
pamelor
parmodalin
parnate
paroxetine
parstelin
paxil
perforatum
perphenazine
thyroxine
tianeptine
tiazesim
tofenacin
tofranil
toloxatone
tolvon
tonerg
tranylcypromine
trausabun
trazadone
trazodone
trifluoperazine
triiodothyronine
trimipramine
trintellix
tryptan
tryptophan
valdoxan
valnoctamide
valproate
valpromide
venlafaxine
victoril
viibryd
vilazodone
viloxazine
vivactil
vivalan
vortioxetine
wellbutrin
zelapar
zimelidine
ziprasidone
zoloft
zyprexa

Appendix F

Box-and-whiskers charts of some of the significantly different features.


LIWC negemo words distribution
Percentage score
🟦 Depressed
🟧 Non-depressed
LIWC anxious affect words distribution
Percentage score
🟦 Depressed
🟧 Non-depressed
LIWC health words distribution
Percentage score
🟦 Depressed
🟧 Non-depressed
LIWC power words distribution
Percentage score
🟦 Depressed
🟧 Non-depressed
LIWC leisure activities words distribution
Percentage score
🟦 Depressed
🟧 Non-depressed