Utrecht University

"Why does the computer say I am depressed?" Towards a better understanding of concepts in automatic depression detection

by

Sytse Backx, student nr. 6532926

Submitted to the Artificial Intelligence Graduate Program
in partial fulfillment of the requirements for the degree of
Master of Science

Graduate Program in Game and Media Technology
Utrecht University
2025

"Why does the computer say I am depressed?" Towards a better understanding of concepts in automatic depression detection

APPROVED BY:

dr. Heysem Kaya .....
(Thesis Supervisor)

dr. Itır Önal Ertuğrul .....

DATE OF APPROVAL: DD.MM.YYYY

ABSTRACT

"Why does the computer say I am depressed?" Towards a better understanding of concepts in automatic depression detection

This thesis explores the interpretability of automatic depression detection. Specifically, it investigates whether humanly understandable concepts can be gained from a varied array of machine learning models. Using the clustering of Bag of Words (BoW) representations through Principal Component Analysis (PCA), Latent Semantic Analysis (LSA) and Bag of Concepts (BoC), words are grouped into concepts, which are then examined for their interpretability. Concept Relevance Propagation (CRP), a post-hoc analysis technique to uncover latent concepts in neural networks, is also applied to a Convolutional Neural Network (CNN) which is trained on word embeddings. Two data sets are used: The Extended Distress Analysis Interview Corpus (EDAIC), a collection of semi-structured interviews, and D-Vlog, a collection of personal video blogs. The results show that deep learning approaches have the $F1$ performance across the board, furthermore, on the D-Vlog dataset, PCA and LSA have a small negative effect on performance compared to a baseline BoW model, while dimensionality reduction through BoC causes a more significant decrease in performance. It is also shown that none of the simpler approaches are able to model the complexity of the EDAIC dataset. Moreover, the discovered concepts are compared to the symptoms of major depressive disorder as outlined in the DSM-V. The concepts discovered through CRP seem to be most meaningful as well as having some correlation to some symptoms of depression, with the concepts in both datasets, "Emotional Struggle" and "Being under Strain", having a close correlation to feelings of depression. Concepts discovered through PCA and LSA, while coherent, are not able to be linked to any specific symptoms of depression. The concepts discovered through BoC do not have enough coherence to be compared to the symptoms of depression in any meaningful way.

NEDERLANDSE ABSTRACT

Deze thesis verkent de interpreteerbaarheid van automatische depressiedetectie. Er wordt onderzocht of concepten die begrijpelijk zijn voor mensen worden geleerd door een breed scala aan machine learning-modellen. Woorden in Bag of Words (BoW) representaties worden door middel van Principal Component Analysis (PCA), Latent Semantic Analysis (LSA) en Bag of Concepts (BoC) gegroepeerd. Deze groepen worden vervolgens onderzocht voor hun begrijpbaarheid. Concept Relevance Propagation (CRP), een post-hoc analysemethode om onderliggende concepten in neurale netwerken bloot te leggen, wordt gebruikt op een Convolutional Neural Network (CNN) die getraind is op woordinbeddingen. Twee datasets worden gebruikt: Het Extended Distress Analysis Interview Corpus (EDAIC), een collectie semi-gestructureerde interviews, en D-Vlog, een collectie aan videoblogs. De resultaten laten zien dat, over alle datasets, deep-learing modellen de beste $F1$-score behalen. De resultaten laten ook zien dat, in de D-Vlog dataset, PCA en LSA een klein negatief op de prestaties van modellen ten opzichte van een BoW classificator als referentiepunt. Ook laten ze zien dat de dimentionaliteit verminderen door BoC een significant groter negatief effect heeft op een models prestatie. Verder wordt aangetoond dat de simpelere modellen de complixiteit van de EDAIC dataset niet goed kunnen modeleren. Ontdekte concepten worden vergeleken met de symptomen van een ernstige depressieve stoornis. Concepten blootgelegd door CRP lijken het meest betekenisvol te zijn, terwijl ze ook het best gelinkt kunnen worden aan symptomen van depressie. De meest belangerijke van CRP concepten “Emotional Struggle” en “Being under Strain” lijken gecorreleerd te zijn aan depressie. Concepten die ontdekt worden door PSA en LSA, hoewel samenhangend, kunnen niet gelinkt worden met specifieken symptomen van depressie. De concepten van BoC modellen zijn niet samenhangend genoeg om betekenisvol vergeleken te worden met symptomen van depressie.

ACKNOWLEDGEMENTS

First and foremost, I would like my first supervisor, dr. Heysem Kaya. Writing this thesis was not always a smooth process, but without his patience and understanding things would have been much worse. In a similar vein, I would like to thank my second supervisor dr. Itır Önal Ertuğrul for their patience with the the extended timeline required for this thesis. I also would like to thank Gizem Sogancioglu, whose guidance at the beginning of this research greatly shaped it. Finally, I would like to thank the study group: Mart, Lara and Stan, who were alongside me during the start of this thesis. Especially Stan for letting me use his version of the D-Vlog dataset.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

In recent times, machine learning has become more popular than ever. However, with the increasing complexity of neural networks, their interpretability decreased. In some domains, such as healthcare, understanding the reasoning behind a specific prediction is crucial for model auditing, validating decisions, or mitigating biases. This gave rise to a new field of research: explainable artificial intelligence (XAI) [28]. This field aims to shed light on the inner workings of artificial intelligence models. However, since XAI is still relatively new and different domains and applications have different requirements for explanations, it remains difficult to find a universal "best" explanation.

1.1. Problem Statement

This research focuses on automatic depression detection. According to the DSM-V [14], depressive disorders are characterized by the presence of a sad, empty, or irritable mood, accompanied by somatic and cognitive changes that significantly impact an individual's functioning. Some depressive disorders are also associated with suicidal ideation. With suicide responsible for approximately 700,000 deaths annually and ranking as the fourth leading cause of death among individuals aged 15 to 29 [69], improving early detection mechanisms could save many lives. Prior research has shown that artificial intelligence can assist in this endeavor [56].

However, it is critical that AI be used in conjunction with healthcare professionals to ensure ethical decision-making. Furthermore, diagnoses should not rely solely on a checklist of symptoms, but must consider an individual's broader context [14]]. Many existing AI models do not make their decision-making criteria transparent, making them difficult to interpret or validate. As a result, human experts remain necessary for evaluating predictions. To enable this, it is essential to understand on what basis these predictions are made. There remains a gap in the literature regarding interpretable AI models for depression detection with clearly defined decision criteria, which this research aimed to address.

1.2. Research Objective

The primary objective of this thesis is to investigate which criteria or concepts AI models might use in predicting depression, and to explore whether these criteria align with known symptoms of major depressive disorder as described in the DSM-V [14]. To achieve this, various models were trained on two distinct datasets: one comprising semi-structured interviews and another consisting of personal vlogs. These models were either inherently interpretable, or their learned concepts were analyzed using a post-hoc method.

For the inherently interpretable models, the number of dimensions in Bag of Words (BoW) representations is reduced by clustering individual words into broader concepts. This is carried out using three different methods: Latent Semantic Analysis (LSA), Principal Component Analysis (PCA), and Bag of Concepts (BoC), which are discussed in more detail in Sections 2.3.1, 2.3.2, and 2.3.3, respectively. The post-hoc analysis method used is Concept Relevance Propagation (CRP), which identifies concepts learned by a neural network. This method is further explained in Section 2.3.6.

A secondary novel contribution of this research was the application of CRP to the text modality, a use case that, to the best of our knowledge, has not been previously explored.

1.3. Research Questions

RQ 1: Using a Bag of Words (BoW) representation-based classifier as a baseline, can an interpretable depression detection model be formed using only humanly understandable concepts with $F1$ performance comparable to the baseline on the Extended Distress Analysis Interview Corpus (EDAIC) and D-Vlog datasets?

- SRQ 1.1: Does the use of Principal Component Analysis (PCA) to learn humanly understandable concepts provide a model with $F1$ performance comparable to the baseline model?
- SRQ 1.2: Does the use of Latent Semantic Analysis (LSA) to learn humanly understandable concepts provide a model with $F1$ performance comparable to the baseline model?

- SRQ 1.3: Does the use of the Bag of Concepts (BoC) method to learn humanly understandable concepts provide a model with $F1$ performance comparable to the baseline model?

RQ 2: How similar are the humanly understandable concepts used by depression detection models to the symptoms of major depressive disorder outlined by the DSM-V?

- SRQ 2.1: How similar are the concepts found using PCA to the symptoms of major depressive disorder in the DSM-V?
- SRQ 2.2: How similar are the concepts found using LSA to the symptoms of major depressive disorder in the DSM-V?
- SRQ 2.3: How similar are the concepts found using the BoC method to the symptoms of major depressive disorder in the DSM-V?
- SRQ 2.4: How similar are the globally important concepts found using Concept Relevance Propagation to the symptoms of major depressive disorder in the DSM-V?

1.4. Outline

This thesis offers a detailed background on the topic of this research and the methods used to conduct it in Chapter 2. Work related to automatic depression detection, including both datasets used, will be given in Chapter 3 before giving the methodology used to conduct research in Chapter 4. Chapter 5 presents the results of this research and this thesis will be concluded in Chapter 6.

2. BACKGROUND

This chapter provides an overview of some Natural Language Processing (NLP) techniques and state-of-the-art models for text classification before going into Explainable Artificial Intelligence (XAI). Here, a non-exhaustive list of intrinsically interpretable models will be given, as well as methods for post-hoc analysis for more complicated models and a short overview of concepts in XAI. Finally, the chapter will go deeper into the background of the tools and methods used in this research. The purpose of this background is to inform the reader enough about the methods used in this research.

2.1. Natural Language Processing

Natural language, while easily understandable to those who speak it, is a very complex system of grammar and vocabulary, not easily quantifiable for computers. Moreover, in most popular formats used for text encoding, such as Unicode [9] or ASCII [23], the value of a character has no inherent meaning. This will provide some basic background of common methods in NLP to overcome these difficulties.

2.1.1. Bag of Words

Bag of Words (BoW) representations are unorganized collections of the occurrence of a word in text documents, represented as a vector $v = (x_1, x_2, ...x_n)$ where x_i is the occurrence of a word in a list of words of length n. This occurrence can be represented in one of 3 ways: binary (0 means that the word is absent in the text, 1 means that it is present), frequency (how often the word occurs in the text) or using TF-IDF [77] (a measure where the amount of occurrences in a document is balanced with the total amount of occurrences in a corpus) [104]. When representations are made separately for two classes, the largest difference in these representations should yield words which are most representative of the difference between two different classes. These words can then be used to classify new text; this can be done with a simple "More words associated with class A means it belongs to class A" classifier or with more complicated classifiers such

as a Naive Bayes classifier as shown by Murphy [63] or a decision tree.

2.1.2. Word Tokenization

Since text is a sequence of characters, it first needs to be chopped into pieces, which are called tokens [57]. While tokenizing every sequence of letters separated by a white space seems natural and works in many cases, improvements can be made. These include making the singular and plural versions of nouns the same token or separating verb conjugations for a verb, for example splitting the word "walked" into "walk" and "ed", i.e. a verb stem and a verb conjugation indicating past tense. For a more exhaustive list of improving tokenization, see "Introduction to Information Retrieval" by D. Manning [57].

2.1.3. Word Embeddings

Since there is no relation between the value of a sequence of characters and its semantic meaning, word embeddings are used to model meaning. Word embeddings aim to create a vector space where semantically similar tokens/words are close to each other [41]. Several methods of initializing these spaces have been proposed, some prominent examples being word2vec [62] and GloVe [72]. Since words are embedded based on their meaning, it is possible for some relationships between words to be represented as translations between words. As shown in Figure 2.1, the translation between the male and female versions of words is roughly equal. Consequently, the distance between words can also be used as a measure of similarity.

2.1.4. Deep Learning Approaches for Text Classification

Currently, the most complex and most powerful types of AI are deep learning models. However, not all deep learning models are created equally. In this section, the advantages and disadvantages of the Convolutional Neural Network (CNN) and the transformer will be compared for the text classification. Traditionally, the Recurrent Neural Network (RNN) has also been used in Natural Language Processing (NLP), however due to its

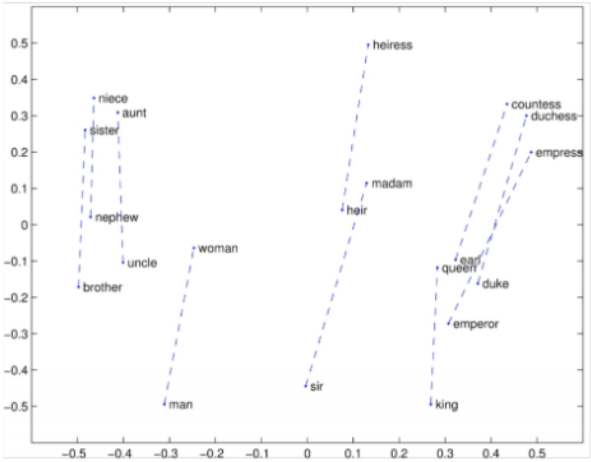

Figure 2.1: A projection of GloVe [72] done by Jiao et al. [39]

nature as a sequential model, it is not a good fit for all NLP tasks. Firstly, large inputs slow these models down a lot since all words have to be fed to the model sequentially. Secondly ,RNN's have the problem of vanishing or exploding gradients [37, 79], making them sub-optimal finding local or position invariant patterns. RNNs are best used for NLP tasks like part-of-speech or question answering [51].

2.1.4.1. Convolutional Neural Networks. In a Convolutional Neural Network (CNN), convolutional kernels are applied over word embeddings to detect a pattern, i.e., a specific phrase, learned by the kernel. The ability to detect phrases independently of their position in the text is useful for most text classification tasks. For example, the sentences "I went to the store and I am depressed" and "I am depressed and went to the store" functionally have the same meaning and should therefore be classified as the same. A simple implementation of this idea is proposed by Kim 2014 [46], more on this model in 2.3.4. This model still has the downside of not being able to find larger or more sparse patterns, both intra- and inter-sentence. The current state-of-the-art CNN architecture for larger inputs on text is TextConvoNet [89], which solves this problem, by having a higher-dimensional input space to allow both intra- and inter-sentence convolutions.

2.1.4.2. Transformers. First proposed by Vaswani et al. [100], transformers are the state-of-the-art deep learning model when it comes to NLP, the most well known examples being GPT [2] and BERT [26]. Transformers' improvement over earlier types of networks is its

ability to calculate self-attention in parallel. Self-attention is a way of contextualizing a word based on the real world meaning of surrounding words [41]. This attention is calculated with an input matrix X, a matrix where each row is the embedding of a word in the sentence. As shown in Equation 2.1, self-attention is computed using three different weight matrices—query (Q), key (K), and value (V)—each obtained by multiplying the input matrix X with its corresponding weight matrix, where d_k is the dimension of the query and key matrices.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

Transformers do not stop at just one self-attention calculation since words in a sentence can relate to each other in different ways, which is hard to do with a single self-attention calculation, i.e., with a single attention head [41]. To combine multiple attention heads, the results are concatenated and multiplied by a weighted matrix [100]. While the higher complexity these models offer is useful in finding complex relationships between words, it also has a drawback: with smaller training sets the risk of overfitting increases.

2.2. Background on Explainable Artificial Intelligence

A problem with many Artificial Intelligence (AI) models is that their internal logic is often not interpretable for humans. Although this is not a problem in all cases, this understanding is required in some domains for a model to be used. This can be due to a high cost of error requiring more trust, confirming some auxiliary requirement like a protected group not being discriminated against. Furthermore, being able to interpret a model also leads to a better scientific understanding and helps with debugging or improving a model.

Explainable Artificial Intelligence (XAI) is the field of creating a deeper understanding of the internals of AI models. Due to the broadness of the field, many important concepts within the field do not have very concrete definitions, let alone widely recognized methods of how to measure them. Luckily, there have been attempts to create axonomies within the field.

As mentioned before, there are several reasons for a desire to understand a model, Adadi et Berrada [4] define these motivations in 4 broad categories:

- Explain to justify, when an explanation is needed to justify some aspect of the model, usually to ensure fairness or to ensure some legal requirement is being met. For example, a government institution must ensure that they do not label possible fraud cases based on race.
- Explain to control, when an explanation is used to prevent error from occurring. Explanations could be used to discover flaws in an existing model. For example, discovering an image classifier is trained on the background of images instead of the subject will help in correcting errors made by the model.
- Explain to improve, when an explanation is used to help design a new iteration of a network. Neural network architectures are constantly changed to improve their performance; understanding what parts of a network are used most or what data is most important helps in this endeavor.
- Explain to discover, when explanations are used to learn new knowledge. In tasks where an AI model has surpassed human ability, new knowledge might be gained by understanding its actions. For example, a high-level chess model might discover general tactics unknown to humans.

These different motivations lead to different requirements for an explanation, including differing requirements for model performance or domain complexity. There are many factors that go into choosing the model best suited for a task. Where in some cases a decision tree, a highly interpretable model, might be sufficient for a task, in cases where a complex neural network is needed, methods like calculating feature importance are needed.
Even though the range of types of explanations is very wide, there are proposals for taxonomies. Doshi-Valez et Kim [28] for example, hypothesizes some axes along which explanations could be classified.

- Global vs. Local: Are explanations aimed at explaining a model's overall behavior (global), or are they aimed at explaining a specific decision (local)?

- Area, Severity of Incompleteness: What is the scope you want your explanation to tackle? One could, for example, imagine different scopes in the domain of self-driving cars. On the one hand, someone might be interested in the pure scientific knowledge of how self-driving cars function, while on the other hand, someone might need to be able to meet a specific requirement on a security checklist.
- Time constrained: How much time does a user need to understand an explanation? A long and thorough explanation might be good for scientists; however, there might be scenarios in which a decision needs to be made quickly and shorter simpler explanations are the norm.
- Nature of User Expertise: How much domain knowledge does the user have? Domain knowledge might influence how complex an explanation can be; Doctors for example, might prefer more technical/medical explanations than patients.

Adadi et Berrada [4] also makes the global vs local distinction and adds another axis, model-agnostic vs. model-specific. Model-agnostic models apply to all machine learning models, while model-specific models are usually limited to only working on a specific type of model.

Due to the wide range of requirements for explanations, evaluating their quality is also a hard task and no single metric which would properly evaluate all explanations exists [4, 12]. This necessitates that evaluations of explanations are properly aligned with the contribution of the research. Doing a user study on lay people instead of domain expertise might yield different results, for example, since domain experts might now have as high of a cognitive workload understanding the explanation which has been shown to affect interpretability [1, 49] and therefore the quality of an explanation. Although no unifying metric has been created yet, Doshi-Velez et Kim [28] has created a taxonomy of evaluation approaches in XAI.

- Application-grounded explanations, evaluating explanations for a specific task with an end user for the explanation in mind, e.g., can explanations be helpful to a doctor making a diagnosis
- Human-grounded explanations, evaluating explanations for qualities relating to ev-

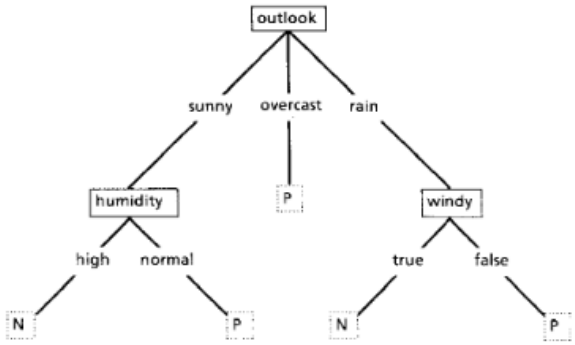

Figure 2.2: An example of a simple decision tree [76].

ery human. e.g., measuring how much time pressure affects the interpretability of an explanation

- Functionally-grounded explanations, evaluating explanations not with humans but with an objective measure as a proxy for interpretability.

2.2.1. Intrinsically Interpretable Models

Some models are simple enough that their logic can be comprehended without further explanation. For this to be possible, it is necessary for the input or the features to be intelligible for humans. An advantage of these types of models is that no post hoc explanations are required to understand their logic. Furthermore, because of their simplicity, they are often less computationally expensive to run. While their simplicity is an advantage in some cases, these models may not be able to model the complexity which some tasks require.

2.2.1.1. Decision Trees or Rule Based Models. One of the most intuitive models in artificial intelligence is the decision tree. They are a cascading set of conditions ending in a classification of a particular data point. An example can be found in Figure 2.2, where the weather is classified- either positive (P) or negative (N) for a particular activity. Once a decision tree has been constructed, the network can be used for local and global explanations. For local explanations, the specific path through the tree can be traced and the features used to make that particular decision are discovered. For global explanations, illustrations such as the one in Figure 2.2 can be created, these images can

give insight into a model's general behavior. In this specific case, the illustration leads to the insight that mild weather would be suitable for the particular activity. There are many optimization criteria to train the exact parameters of a decision tree which leads to different learning approaches. Some popular options are: Classification and Regression Trees (CART) [21], c4.5 [75] and Chi-Squared Automatic Interaction Detection (CHAID) [43].

2.2.1.2. Support Vector Machines and Support Vector Classifiers. A Support Vector classifier (SVC) is a classification model trained through the Support Vector Machine (SVM) learning method. These classifiers decide on a decision boundary between classes in a transformed feature space based on maximizing the margin between classes. In its most basic form, these classifiers linearly separate data into two classes, which leads to the decision function:

$$y(x) = w^T \phi(x) + b, \tag{2.2}$$

where $\phi(x)$ is a fixed feature-space transformation chosen before training, w^T is the weight factor learned during training, and b is bias (intercept) term.

One problem the basic form of an SVC encounters is that not all data is linearly separable, even in the transformed feature space. To address this, the kernel trick can be applied. With this trick $\phi(x)$ is substituted by a kernel function, which avoids constructing an explicit feature space. Furthermore, by using Lagrange multipliers and Karush-Kuhn-Tucker (KKT) conditions, we can reformulate the decision method as:

$$y(x) = \sum_{n=1}^{N} \alpha_n t_n k(x, x_n) + b, \tag{2.3}$$

where, α_n is weight assigned to training sample n during training, $t_n \in \{-1, 1\}$ is label of training sample n, $k(x, x_n)$ is the kernel function to calculate similarity x and x_n, and b denotes the bias term. In this expression, w is implied by α_n, t_n and $k(x, x_n)$ but is not explicitly calculated. As a result, this means the model is no longer intrinsically interpretable. Thus, post-hoc explainability methods must be used.

2.2.1.3. Methods for BoW Dimensionality Reduction. The BoW vectors generated from a large corpus often have a high dimensionality while being very sparse [45], since only small subsets of a full vocabulary usually occur in a specific text document. This causes computations with the vectors to become very expensive. A way to solve this problem is dimensionality reduction. Dimensionality reduction is a transformation in which a higher-dimensional space is reduced to a lower-dimensional space while maintaining as much of the meaningful information in the data as possible. There are many methods for dimensionality reduction, both linear and non-linear, van der Maarten et al. [99] has an extensive overview of methods and their performance, with the linear Principal Component Analysis (PCA) outperforming non-linear methods on real-world datasets. In addition to PCA, another popular method of dimensionality reduction is Linear Discriminant Analysis (LDA). Martinez et al. [60] and Borade et al. [20] both show that PCA outperforms LDA with the datasets where the amount of training data is small or when the training data are non-uniformly sampled from the underlying distribution. There also exists a method of reducing dimensionality in bag of word representations, specifically: Latent Semantic Analysis (LSA), more information on this method will be given in Section 2.3.1

Another method for reducing the dimensionality of a Bag of Words model is to first cluster the words into concepts before "bagging" them, creating a Bag of Concepts instead of a Bag of Words . As proposed by Kim et al. [45], words can first be transposed into word embeddings, after which they can be clustered through k-means clustering. This method is elaborated on further in Section 2.3.3.

2.2.2. Post-Hoc Analysis

When simple models are insufficient to properly model a domain, more complicated models such as neural networks need to be used. Since they are not interpretable on their own, methods have been developed for explaining these models. In some cases it is possible to use intrinsically interpretable surrogate models to explain more complex models, while these models might have worse performance, they could help with understanding more complex models. Furthermore, one of the most intuitive ways of post-hoc analysis is feature importance. The basic idea is discovering what feature(s) a model (primar-

ily) uses to make a decision, which helps significantly in understanding the workings of a model, aiding both our understanding of the model as well as protecting against unwanted behavior. For example, feature importance protects against the classic cautionary tale where a model learns classification based on the background of images rather than the subject matter, as recreated in Ribeiro et al. [80]. Furthermore, feature importance can be used in other XAI methods such as Model Class Reliance (MCR). MCR provides the range of feature importance values across all well-performing models of a specified class rather than explaining the feature importance for a single model.
There are many different approaches to calculating feature importance. Permutation-error-based methods such as Permutation IMPortance (PIMP) [10] or SHapley Additive exPlanations (SHAP) [53] measure the increase in model error after permuting input features. Backpropagation-based methods such as Concept Relevance Propagation (CRP) [3], DeepLIFT [84] or saliency maps [85] propagate some sort of importance back through a model. Furthermore, surrogate models can also be used to explain a model's global behavior [48] or its local behavior [80]. A more extensive overview of post-hoc explainability methods can be found in Table 2.1.

2.2.3. Concepts in Explainable Artificial Intelligence

The term "concept" in XAI is very vague, it is a term that not only spans different domains but also different modalities. Broadly speaking, it is a unit, composed of input features, which is more understandable than those features individually. Despite the term's broadness, some criteria for the usefulness of a specific concept can also be defined. For example, Ghorbani et al.[34] defines three:

- *Meaningfulness*, a concept needs to be semantically meaningful on its own
- *Coherency*, individual examples of a concept need to have an overarching property
- *Importance*, a concept should be important in the true distinction between the classes, i.e., the color of a photographed object vs. the color of the background

In addition to these three inherently desirable criteria, there are also some other distinctions between existing approaches. One of these distinctions is to what extent

Method	Modal Range	Locality	Notes
CRP [3]	Specific	Local	Disentangles feature importance of a model into different concepts through conditional backpropagation.
LIME [80]	Agnostic	Local	Fits a simple, interpretable model locally around a predicted instance to explain model behavior.
Global Surrogate Models [48]	Agnostic	Global	May not model the complexity of the original predictive model well. Results may reflect the internal bias of the surrogate model.
SHAP [53]	Agnostic	Local	Provides consistent and theoretically grounded attributions through game theory, but computationally expensive with high-dimensional data.
PFI [22]	Agnostic	Global	Estimates feature importance by measuring the effect on model performance when feature values are permuted. Biased under feature correlation.
PIMP [10]	Agnostic	Global	Improves on PFI by generating p-values using permutation-based null distributions, addressing bias in feature importance scores.
Guided Backpropagation [90]	Specific	Local	Highlights input features via modified gradient flow. However, it fails to account for negative contributions and does not reliably reflect true model reasoning.
DeepLIFT [84]	Specific	Local	Tracks contribution scores by comparing neuron activations to a reference input. Efficient and preserves sign information.
Saliency Maps [85]	Specific	Global	Computes the gradient of output with respect to input features to visualize feature influence. Susceptible to noise and saturation.

Table 2.1: An overview of post hoc explainability methods.

Paper	Method	Near Complete	Pre-Trained Concepts
Ghorbani et al. [34]	ACE	✓	∅
Achtibat et al. [3]	CRP	∅	∅
Yeh et al. [107]	ConceptSHAP	✓	∅
Kim et al. [44]	TCAV	✓	∅
Zhou et al. [110]	IBD	∅	✓

Table 2.2: Methods of concept based explanation methods.

an approach tries to create concepts which completely cover a model's behavior. An approach like conceptSHAP [107] attempts to fully cover the model's behavior with its concepts while interpretable basic decomposition [110] has a "residual" concept for activations not covered by other concepts. Yeh et al [107] proposes a method to calculate the "completeness" of a given approach. So far, most existing research surrounding uncovering concepts that are used by a model has focused on extracting concepts from images. An overview of existing methods can be found in Table 2.2. While attempts of extracting concepts from text have been made [70] [101], these methods are attempting to find concepts inherent in the text, not finding concepts used by any model.

2.3. Background on Used Tools and Techniques

In this section, the methods used in this research are explained more thoroughly.

2.3.1. Latent Semantic Analysis

Latent Semantic Analysis (LSA) [25] aims to transform a large term by document matrix into a set of k (ca. 100) factors, which could be seen as coordinates in a k dimensional space where similarity is represented by the distance between points. First, a term by document matrix X is created, this matrix can be seen as stacking Bag of Words representations with a shared dictionary next to each other. X is then decomposed; this is done through a technique called Singular Value Decomposition (SVD), which produces 3 matrices $X = T_0 S_0 V_0^T$ so that T_0 and V_0^T are orthogonal matrices and S_0 is a

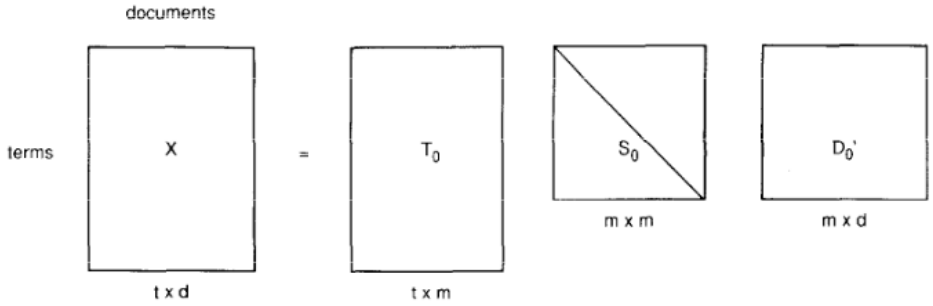

Figure 2.3: An example of SVD from [25].

diagonal matrix with singular values. A schematic approximation of this decomposition can be found in Figure 2.3. One of the properties of this decomposition is that it allows for an easy method of reducing the dimensionality of X into the best possible lower dimensionality approximation [25] [29]. By keeping the dimensions associated with the k highest singular values in S_0, the new matrices T_k, S_k, V_k can be created so that $T_k S_k V_k^T = X_k \approx X_o$. After this T_k and V_k can be seen as coordinates in a k-dimensional space where each axis is an underlying concept whose importance is determined by the corresponding value in S_k.

2.3.2. Principal Component Analysis

Principal Component Analysis (PCA) is a general method for dimensionality reduction, aiming to maximize variance in a projected space, spanned by (orthogonal) principal components. An intuitive way of understanding principal components is as a set of uncorrelated linear transformations transforming a set of data points in an existing coordinate system, X, into a new coordinate system X_2, where the largest variance in the data lies along the first axis, the second largest variance on the second axis, etc. [74]. This new coordinate system could have fewer dimensions than the original coordinate system; by drawing axes in a way that maximizes variance it is ensured that a maximum amount of variance is preserved in that dimensionality reduction. On can mathematically show that these axes, or principal directions, correspond to the eigenvectors of the covariance matrix of the data. The transformation of the nth principal component in a dataset of size p can be described through the formula: $\alpha_n^{'} x = \alpha_{n1} x_1 + \alpha_{n2} x_2 + ... + \alpha_{np} x_p$. By concatenating the values of α, a new weighing matrix W can be constructed and PCA subspace can be defined as $XW = X_2$. In practice, PCA can also be computed through

Singular Value Decomposition (SVD), which is often more stable and computationally efficient. To do so, the data matrix X is first mean-centered so that each feature has zero mean. The centered matrix X_c is then decomposed using SVD: $X_c = U\Sigma V^T$, where the rows of V^T correspond to the principal directions (eigenvectors of the covariance matrix). In this case, $U\Sigma$ is the projection of the data on the principal components. Note that this equation is the same as the one in Section 2.3.1, and just like in that section, we can take the top k dimensions to reduce the amount of dimensions in the data.
Although the principal components could be seen as underlying concepts in the data, they are not guaranteed to be semantically meaningful [106]. However, through the values (weights) of each column in W, the impact of each input feature on the principal component can be calculated, which might give insight into the concept it represents.

2.3.3. Bag of Concepts

Similar to a Bag of Words, a Bag of Concepts [45] is a histogram of how often words relating to specific concepts occur within a text. These concept representations are created through the following steps:

(i) transpose words in the corpus into an embedding space
(ii) use spherical k-means to find the location of concepts in the embedding space
(iii) associate each word in a document with a concept through its place in the embedding space
(iv) make a histogram for the occurrence of concepts within a document
(v) scale importance of each concept using CF-IDF

Words that occur frequently across all documents are not as useful for classification as words that occur more rarely [83], it is also highly likely that this also holds for concepts. Because of this, we use Concept Frequency-Inverse Document Frequency (CF-IDF), a method very similar to Term Frequency-Inverse Document Frequency. It balances the frequency of a concept in a document with its frequency in the entire corpus through

equation 2.4,

$$CF - IDF(c_i, d_j, D) = CF(c_i, d_j) \times \log \frac{|D|}{|d \in D; c_i \in d|} \tag{2.4}$$

where,

$$\begin{aligned}
(c_i, d_j, D) &= \text{concept i, document j, the corpus} \\
CF(c_i, d_j) &= \text{the frequency of a concept i in document j} \\
|D| &= \text{the number of documents in the corpus} \\
|d \in D; c_i \in d| &= \text{the number of documents in the corpus which contain concept i}
\end{aligned}$$

Note that since the term within the logarithm is not dependent on any specific document, it can be calculated as a scalar for each concept for the entire corpus at once.

2.3.4. Kim's text CNN

The model proposed by Yoon Kim [46] follows the basic principles described in Section 2.1.4.1. Its novel contribution is exploring different versions of an architecture (static, non-static, random and multi) where the only differences are the properties of the input channels. The static model only has a static channel, where words are initialized with word2vec [62] embeddings and are not changed during training. The non-static model has a non-static channel, initialized with word2vec but fine-tuned during training. The random version has a non-static channel but is initialized with random embeddings. Furthermore, it also explores using two channels at once, the static and non-static channels. An overview of this architecture is shown in Figure 2.4.

During experimentation, the static, non-static and multi-model seem to have similar performance, while the random model does seem to perform worse.

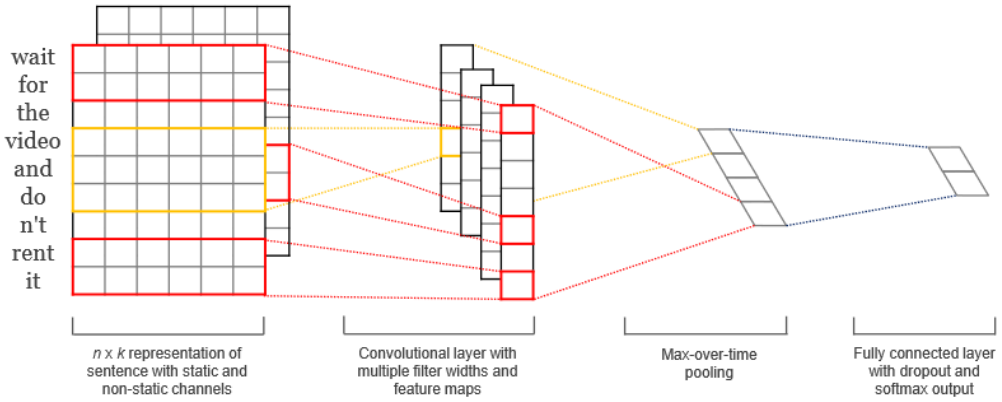

Figure 2.4: An overview of the multi-channel architecture of Kim's CNN [46].

2.3.5. Layer-Wise Relevance Propagation on Text

First proposed by Bach et al. [15], Layer-wise Relevance Propagation (LRP) is a feature importance backpropagation calculation method. This is done by propagating a relevance score R back through the network. Eventually, this propagation will reach the input layer, where you can then see how relevant each input is for any specific outcome. This relevance is not propagated from the entire output layer, but from a single neuron in the output layer. The relevance score of this neuron is initialized as the activation of the neuron from a specific example. The relevance of neuron i in layer l is calculated as the sum of the relevance contributions of i in layer $l+1$ (equation 2.5, where $R_{i\leftarrow j}^{(l,l+1)}$ is the relevance contribution of neuron i to neuron j).

$$R_i^l = \sum_j R_{i\leftarrow j}^{(l,l+1)} \tag{2.5}$$

These relevance contributions are calculated according to equation 2.6, where z_{ij} is the activation of neuron i multiplied with the weight W_{ij} and z_j is the activation of neuron j, which is equal to the sum of all individual contributions plus a bias, i.e. $\sum_i z_{ij} + b_j$

$$R_{i\leftarrow j}^{(l,l+1)} = \frac{z_{ij}}{z_j} \cdot R_j^{l+1} \tag{2.6}$$

One drawback of using the relevance contribution equation 2.6 is that for small z_j values, $R_{i\leftarrow j}$ can take unbounded values. Or, if $z_j = 0$, it would require a division by 0, which

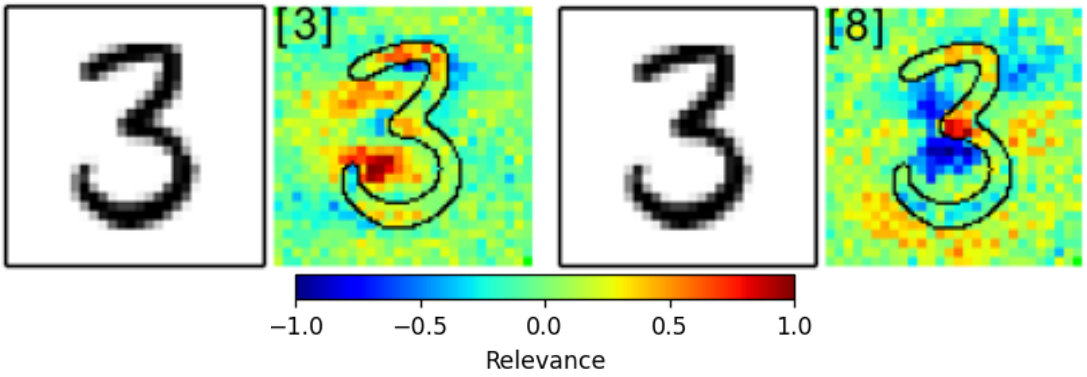

Figure 2.5: Relevance for classifying a handwritten 3 as a 3 and an 8 from [15].

is undefined. To combat this, a stabilizing variable, ϵ, can be introduced. This leads to to equation 2.7

$$R_{i\leftarrow\!\!-j}^{(l,l+1)} = \begin{cases} \frac{z_{ij}}{z_j+\epsilon} \cdot R_j^{l+1} & z_j \geq 0 \\ \frac{z_{ij}}{z_j-\epsilon} \cdot R_j^{l+1} & z_j < 0 \end{cases} \tag{2.7}$$

Two examples of what LRP could look like are given in Figure 2.5. In these examples, the heatmap is normalized to a range of [-1,1]. Figure 2.5 shows nicely that while the line segment in the middle is used as positive evidence in classifying the handwritten digit as both a 3 and an 8, the missing connections on the left side of the image were relevant as negative evidence in classifying this input as an 8 while being positive evidence of the digit being a 3.

For Layer-wise Relevance Propagation to work on CNN as described by Kim et al. [46] two adjustments need to be made. Arras et al. [11] proposes solutions to these problems. Firstly, the max-pooling layers should not be treated as a fully connected layer but in a winner-takes-all fashion, where the neuron with the highest activation in the previous layer receives all the relevance. Secondly, the relevance scores will be calculated for each coordinate in an embedding, these individual scores should be summed up to find the relevance of the word that is being embedded.

2.3.6. Concept Relevance Propagation

Proposed by Achtibat et al. [3] Concept Relevance Propagation (CRP) is a method for dissecting the flow of Layer-wise Relevance Propagation. Through this, it can be confirmed that some explicit concepts are learned and used by the network, helping us understand it better. This dissecting is based on making relevance scores conditional. Similarly to how in LRP only one specific output is backpropagated while the other outputs are zeroed out, only the relevance of a specific neuron or filter can also be back-propagated while all other neurons in the layer are zeroed out. It bases the usefulness of this method on the fact that the literature suggests that human-understandable concepts are encoded in filters or neurons in a network [18, 19]. Most of the time, a simple attribution map is insufficient to label a concept, it is hard to know whether an activation in an attribution map is caused by a shape, color, texture or something else. To combat this, Achtibat et al. also introduces RelMax, a method for finding images where a specific concept is maximally relevant, which are then masked to focus on the concept at hand. This allows a human to discover commonalities between the input and the discovered images, allowing them to label the concept if they can find any interpretable similarities. Note that these conditions can be applied consecutively, and therefore sub-concepts can be discovered. For example, a higher level concept "Police car" might consist of the concepts "car" and "black and white stripes", which themselves might break down into even lower concepts.
Since all neurons in a network are considered possible concepts, one could assume explanations could be overwhelming. However, since each concept is also associated with a relevance score, it is also very easy to discern how important each concept is.

2.3.7. Performance Measures

Accuracy by itself is not always sufficient to properly measure a model's performance. For example, if there is a 90%-10% class imbalance, predicting the majority class all the time would lead to a high accuracy even though this does not reflect a model's ability to model the domain. To better assess the capability of the model, it is important to separate different types of error a model can make, one way of doing this is described

	Predicted Positive	Predicted Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Table 2.3: Confusion matrix for types of classification.

in Table 2.3.

Using this taxonomy, the $F1$-score of a model can be calculated. This measure is the harmonic mean between precision ($\frac{TP}{TP+FP}$) and recall ($\frac{TP}{TP+FN}$) and can be calculated with equation 2.8.

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{2.8}$$

As one might notice, contrary to measures like accuracy, an $F1$-score is a measure for a models performance on a specific class. Depending on what task the model aims to achieve, this might not be desirable. In order to change $F1$ to make it not a measure for just specific for a class, one can calculate the macro averages of each classes' $F1$-score, the so-called macro-$F1$-score. This method of averaging emphasizes the models performance on more rare classes [94].

3. RELATED WORK

This chapter will give an overview of automatic depression detection in general before diving into work relating to the two datasets used more specifically.

3.1. Automatic Depression Detection

A substantial amount of research has been conducted on creating automated depression detection. Mao et al. [58] has compiled 264 studies of how different features were used to detect depression, including a list of 99 papers which use semantic input features. However, it should be noted that not all of these papers use a machine-learning model to detect depression. A more selective list of automated depression detection can be found in Squires et al. [91]. Even though some of these papers use methods which are decently interpretable on their own, such as logistical regression or SVMs, they do not focus on interpretability.

3.1.1. Interpretable Automatic Depression Detection

Even though they are not as numerable as the papers without explanations, there are papers investigating the explainability of deep learning models for automatic depression detection. The majority of research focuses on obtaining feature importance through different types of post-hoc analysis methods, such as SHAP [65] or integrated gradients [55]. Within this subsection of the research LIME seems to be the most popular method to gather feature importance, being used on models with varying modalities [54, 66, 98]. Other research, such as Zogan et al. [114], proposes a method where the attention of a model could be used as an interpretable feature of the model. Uban et al. [96] takes a different approach to explainability, performing error analysis and an ablation study to gain insight into the mistakes of their proposed model. An overview of these methods can be found in Table 3.1

Furthermore, Alghowinem et al.[6] presents a model-agnostic framework for feature

Paper	Network Type	Modalities	XAI Method
Zogan et al.[114]	MDHAN	tweets, user behaviour	hierarchical attention
Magboo et al.[54]	Multiple types	depression predictors	LIME
Uban et al.[96]	HAN + LSTM	text	error analysis, ablation study
Nguyen et al. [66]	DNN	health survey [42]	LIME
Nemesure et al. [65]	Custom pipeline	health survey	SHAP
Uddin et al. [98]	LSTM	text	LIME
Mahayossanunt et al. [55]	LSTM	video	IG

Table 3.1: Overview of explainable automatic depression detection.

selection for depression detection. This not only improves performance but also makes the depression detection model more interpretable it gives a better view of what features are primarily used by a model.

3.2. D-Vlog

The D-vlog dataset was first created by Yoon et al. [108]. It consists of features from YouTube videos posted from January 1st 2020 until December 31st 2021, selected through keywords from a depressed set (e.g. 'depression daily vlog', 'depression journey', 'depression vlog) or non-depressed set (e.g. 'daily vlog', 'grwm (get ready with me) vlog', 'haul vlog). This dataset only includes audio and video of these YouTube videos. There are 961 samples in this dataset, 555 samples are classified as depressed and 406 as non-depressed. Human annotators verified that the retrieved videos were indeed vlogs and labelled the data. The features extracted from the vlogs are 25 low-level audio features and the x and y coordinates of 68 facial landmarks. Characteristics of this dataset will be elaborated upon further in Section 4.1.1. Although relatively small improvements over the original paper have been made through changes in model architecture [111] [112] [95], the biggest gain in performance seems to stem from adding several datasets together during training [5] [92] or from adding additional modalities such as text [61] or eye information [35].

Paper	Model Type	Modalities	$F1$
Yoon et al. [108]	Multi-modal Transformer based architecture	A,F	0.635
Zhou et al. [112]	Time Aware Multi-modal Fusion Network (TAMFN)	A,F	0.658
Zhou et al. [111]	hybrid deep learning architecture (CAIINET)	A,F	0.666
Tao et al. [95]	transformer based architecture (DepMSTAT)	A,F	0.735
Ahmed et al. [5]	hybrid deep learning architecture	A,F, E	0.938
Sun et al. [92]	Graph Neural Network	A	0.951
Gimeno et al. [35]	Multi-modal Transformer based architecture	A,F, E	0.780
Meyberg [61]	Multi-modal Transformer based architecture	A,F, T	0.937

Table 3.2: Overview of studies conducted with D-vlog dataset. Modalities are low level audio features (A), facial landmarks (F), text (T) and eye information (E)

3.3. Distress Analysis Interview Corpus

The second dataset used in this research is the Extended Distress Analysis Interview Corpus (EDAIC) dataset[82]. This dataset is an extension of the DAIC-WoZ part of theDistress Analysis Interview Corpus (DAIC) dataset[36], which consists of semi-structured clinical interviews. Before these interviews the participants had to take several questionnaires measuring their psychological state, including the PHQ-8 [47] to indicate whether a participant was depressed. These interviews were conducted in four ways: face-to-face, teleconference, fully automated, and wizard-of-oz (WoZ) style. In both fully automated and WoZ styles, a digital avatar conducted the interviews, either automatically or human-controlled (WoZ). After the DAIC-WoZ was used in the Audio/Visual Emotion Challenge and Workshop (AVEC)[81] 2017, this part was extended into the Extended Distress Analysis Interview Corpus (EDAIC). A detailed overview of the characteristics of this dataset will be given in Section 4.1.2

With these datasets having a combined amount of more than 1000 citations on Google Scholar, a lot of research has been done on this dataset. DAIC-WoZ being in de datasets combined by Ahmed et al. [5] and Gimeno et al. [35] using both DAIC-WoZ and EDAIC in their research. Most of the research with this dataset keeps in line with the spirit of the original challenges, using different architectures and pipelines to improve performance. An overview of some notable papers using the DAIC-WoZ and EDAIC datasets can be found in Table 3.3 and Table 3.4 respectively.

Paper	Network Type	Modalities	$F1$	RMSE
Fang et al. [30]	(Bi-)LSTM Fused with Attention Network (MFM-Att)	T,V,A	-	3.68
Yadav et al. [103]	BGRU	T	0.928	
Marriwala et Chaudhary [59]	CNN, Bi-LSTM	T,A	0.780	-
Flores et al. [31]	Bi-LSTM with Self-Attention	T,V,A	0.930	-
Niu et al. [67]	HCAG	T,A	0.920	3.80
Sun et al. [92]	Graph Neural Network	A	0.922	-
Gimeno et al. [35]	Multi-modal Transformer Based Architecture	A,V	0.560	-
Ahmed et al. [5]	Hybrid Deep Learning Architecture	A,V	0.941	-
Yang et al. [105]	CNN	T,V,A	-	5.97

Table 3.3: Overview of automatic depression detection using DAIC-WoZ, modalities are text (T), video (V), audio (A).

Paper	Network Type	Modalities	$F1$	RMSE
Fang et al. [30]	(Bi-)LSTM Fused with Attention Network (MFM-Att)	T,V,A	-	5.17
Uddin et al. [97]	Spatio-Temporal Network with Attention and MFB (STN-Att-MFB)	T,V,A	-	5.78
Gimeno et al. [35]	Multi-modal Transformer Based Architecture	A,V	0.670	-
Zhang et al. [109]	Multi-Instance Learning with RoBERTa (Multi-MTRB)	T	0.860	-
Jin et al. [40]	GCN + LSTM with Spatiotemporal Attention	V, A	0.922	3.51
Sun et al. [93]	Transformer Based Architecture	V, A	-	3.78

Table 3.4: Overview of automatic depression detection using EDAIC, modalities are text (T), video (V) and audio (A).

4. METHODS

This chapter will give an overview of the methods for this research. A general overview of the pipeline used in this research can be found in Figure 4.1. The rest of this chapter will elaborate on the contents of this pipeline.

4.1. Dataset Characteristics

4.1.1. D-Vlog

While the original data (v1) was provided by Yoon et al. [108], it had a total of 961 samples split into sets of 647, 102, 212 for the training, validation, and test sets, respectively. The data provided on its own is not very useful for this research since it did not include transcripts. However, links to the original YouTube videos were provided, making it possible to retrieve these videos and extract transcripts. Gimeno et al. [35] and Meyberg [61] attempted this, but both failed to retrieve some of the samples. This research uses the same subset of the dataset (v2) as Meyberg [61], where a total of 134 transcripts were missing (96, 11, and 27 from the training validation and test sets, respectively). Meyberg demonstrates that this subset is still representative of the original dataset. Unlike Meyberg, we did not stratify the dataset based on gender since this is not a factor in this research, which leads to slightly different training, validation, and test splits. The number of samples in each fold, as well as a breakdown of the class distribution, can be found in Table 4.1. The v2 dataset has a gender distribution of 542 women and 285 men. All vlogs combined have a mean duration of 583 seconds with a standard deviation of 449 seconds. On average, the depressed class has a slightly longer duration, with vlogs taking an average of 624 seconds with a standard deviation of 446 seconds and non-depressed vlogs taking an average of 534 seconds with a standard deviation of 449 seconds.

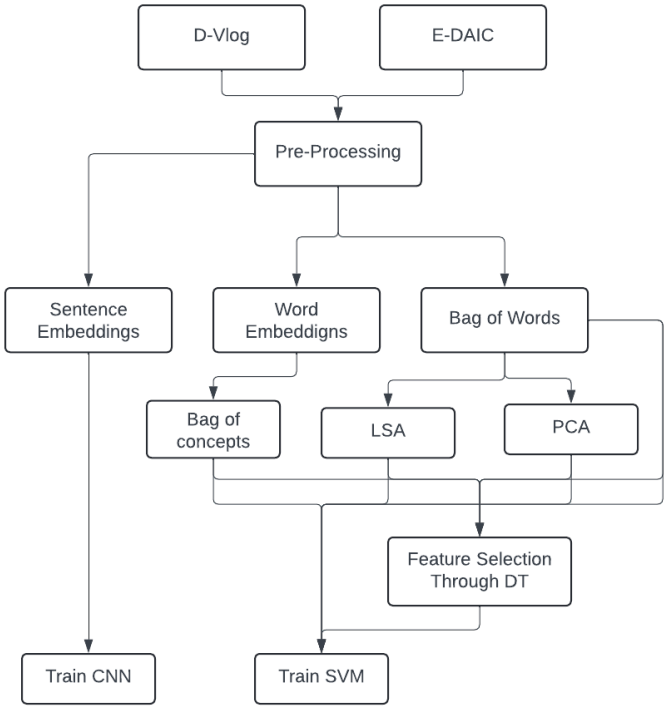

Figure 4.1: Overview of the pipeline for this research.

4.1.2. EDAIC

The EDAIC has a total of 275 samples, a detailed breakdown of the training, development, and test split and class distribution can be found in Table 4.1. The average duration of an interview is 942 seconds with a standard deviation of 305 seconds. The classes are similar in length: the depressed class has a mean duration of 976 seconds with a standard deviation of 317 seconds and the non-depressed class has a mean duration of 927 seconds with a standard deviation of 300 seconds. In total, 105 of the participants were women and 170 were men.

4.2. Pre-Processing

To prepare the data, different types of pre-processing need to be done to get the data in the right format. Our methods take 3 different formats: Bag of Words (BoW), Sentence Embeddings, and Word Embeddings.

Dataset	Class	Total	Train	Val	Test
Dvlog v2	Total	827	551	91	185
	Depressed	449	298	48	103
	Non-Depressed	378	253	43	82
EDAIC	Total	275	163	56	56
	Depressed	82	50	15	17
	Non-Depressed	193	113	41	39

Table 4.1: Breakdown of classes and splits in used datasets.

4.2.1. Bag of Words

For the BoW representation, the sentences first need to be tokenized. This is done using the BERT uncased tokenizer [27]. After tokenization, some token are removed. Firstly, some words which might cause leakage are removed; these words are: “depressed”, “depression”, “anti-depressants”, “depressant”. Secondly, a list of stopwords as defined by the NLTK[52] library (e.g., the, a, to) along with a list of punctuation is removed, since these words rarely provide useful information for a classification task. A conscious decision was made not to filter out filler words such as “uhm”, since these words might indicate a loss of concentration, one of the symptoms of depression. The vocabulary for the BoW is created by taking all unique tokens in both the training and validation set. To reduce noise and complexity, tokens which only occurred once in the training and validation corpus were removed.

4.2.2. Word Embeddings

For the word embeddings, the “bert-base-uncased” model [27], which was introduced by the original BERT paper [26], was used. This model was originally trained for masked language modeling and next-sentence prediction. It was trained on data from Wikipedia [33] and a book corpus [113] and generates an embedding vector with 768 dimensions for each token it receives as input. Since these embeddings are context-aware, simply removing the leaky words mentioned in the previous paragraph is not ideal, since that would alter the sentence structure. Instead, the attention of the model for these words

is set to 0, making it so that the model ignores them without damaging the sentence structure. Furthermore, these words are simply not counted when creating the K-Means model needed for BoC or during the creation of the bag of concepts themselves.

4.2.3. Sentence Embeddings

Since EDAIC is already split into sentences, this data does not need any preprocessing before being fed to a sentence embedding model. For D-VLog, Spacy [38] was used to split the transcripts into sentences. To embed these sentences, two different embedding models were used, both Sentence-BERTs [78]. The specific models were "all-MiniLM-L6-v2" [7] and "all-mpnet-base-v2" [8]. All-MiniLM-L6-V2 is an augmented version of the MiniLM model [102] trained on a variety of datasets. It was specifically trained for the task of identifying, given one sentence from a pair, which sentence among a set of randomly sampled candidates was originally paired with it in the dataset. The model generates an embedding vector with 768 dimensions. The MPNet model [88] is trained for the same task and fine-tuned on the same data as the MiniLM model; however, it generates an embedding vector with 384 dimensions, half that of the MiniLM model. Much like the word embeddings, the attention of the depression related leaky words was set to 0 during sentence embedding, making sure the model did not take them into account when embedding while maintaining sentence structure.

4.3. Interpretable Models

The general idea of this part of the research is to see if it is possible to reduce the amount of dimensions in BoW representations to see if the words will be clustered into humanly-interpretable concepts. Since these representations are task-agnostic, they represent all variation in the data, not necessarily the variation useful for our classification task. In order to not have to explore irrelevant features, feature selection is performed. The impact of this selection on the models performance is also measured.

4.3.1. Principal Component Analysis and Latent Sentiment Analysis

For both of these approaches, the implementation of Scikit [71] is used. Both the training and validation folds are used to train these models. The Scikit implementations both use singular value decomposition to reduce the dimensionality of the BoW representations and therefore cannot create models with a dimensionality larger than the number of samples in the dataset. Since our smallest dataset (EDAIC) has a combined 219 samples in its training and validation fold, the number of dimensions tested will be $n_{dim} \in \{100, 200\}$.
Another consequence of both implementations using singular value decomposition is that the exploration of the concepts generated is also very similar. For each feature which is explored, the 10 words with the highest absolute value in the right singular value matrix are taken, they are chosen as representative for that feature.

4.3.2. Bag-of-Concepts

The Bag of Concepts model as described in Section 2.3.3 will be implemented. Unlike the original paper, the word embeddings are first normalized to a unit length; this means the Euclidean distance function used by the K-means clustering model will more closely resemble a cosine distance function since it is standard to use this as a similarity measure for word embeddings. The K-means model is created using the same vocabulary as in the BoW representations. 5 models are created, with k being $k \in \{100, 200, 500, 1000\}$. These models use the greedy Mmeans++ seeding method [13], 12 different models are trained before the best model is used. During this clustering, the words closest to each cluster centroid are kept track of, so these words can be used to explore the concept of the cluster.

4.3.3. Support Vector Classifier

The eventual classifier of the interpretable methods will be a Support Vector Classifier. All data will be standardized by the scaling transformation implemented by Scikit [71]. The impact of L2-normalization and min-max normalization is also inves-

tigated. Models are trained using the training data using both, one, or neither of these normalization techniques. To identify the optimal hyperparameters for the SVC, a grid search is performed with 5-fold cross validation. This search evaluates the following:

- Regularization parameter $C \in \{0.001, 0.01, 0.1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$
- Kernel coefficient $\gamma \in \{\texttt{scale}, \texttt{auto}\}$
- Kernel type $\in \{\texttt{rbf}, \texttt{linear}, \texttt{sigmoid}, \texttt{poly}\}$

The normalization techniques and hyperparameters with the best validation set performance will be chosen for the final model, which are trained on a combinaion of the training and validation data before being evaluated with the test set performance. The SVCs have macro $F1$-performance as their performance metric.

4.3.4. Feature Selection

All the mentioned methods of dimensionality reduction generate large vectors (100-1000 dimensions) for each transcript where each dimension represents some concept. Since these concepts represent all semantic meaning in the transcripts, many of these dimensions might not be useful for our classification task. Furthermore, in terms of explainability, having more concepts to explain and examine will not help in the models' interpretability. Therefore, some feature selection might prove useful. This selection will be made with the automatic generation of a decision tree, using a similar method as was used during the generation of the SVC. This time, 3 fold cross-validation is used in the training set for hyperparameter tuning and the hyperparameters tuned are:

- Splitting criterion: {`gini`, `entropy`, `loss_log`}
- Minimum fraction of data required to split a node: {0.30, 0.25, 0.20, 0.15, 0.10, 0.05}

After this, the features/dimensions used in the decision tree will be selected as features which will be used to train the SVC. In order to ensure that performance does not suffer too much, the performance of the models with feature selection will be compared to

models without feature selection. Like the SVC, the decision trees have macro $F1$-score as their performance metric.

4.4. Deep Learning

The model architecture described in Section 2.3.4 is used, specifically the static architecture. Since the datasets are not very large, overfitting would be very likely if the hyperparameter of the original model (100 filters per filter size) were directly applied. Therefore, the number of filters is fine-tuned before training. 4, 8, 12 and 24 filters per filter size are tested with 5-fold cross-validation, and the hyperparameter with the lowest loss will be selected for the final training. During cross-validation, early stopping was applied. If the models' loss did not improve for 3 epochs, the training was stopped. The model is trained with an Adam optimizer and with the cross-entropy loss function. To mitigate the effect of class imbalance on the performance of the models, a scaling weight inverse to the class frequency is applied to each class.

After this CRP was applied to the test fold of the dataset. Although CRP is class-specific, in our binary classification problem the relevances are inverted for our 2 classes. Meaning that if a concept has a relevance of 5 for the depressed class, it will have a relevance of -5 for the non-depressed class. For each sample, the relevance of each neuron in the second to last layer will be saved, as well as the heatmap this neuron casts on the sentence embeddings. Since a heatmap on the embeddings itself is not interpretable, the heat on the embeddings is summed per word, resulting in a heatmap that indicates the relevance of each sentence. To find the neuron with the highest relevance, and therefore the most relevant concepts, all absolute relevances of all neurons are summed and all neurons with more than 5% of total relevance are investigated. For each of these neurons, the five examples with the highest absolute relevance for that neuron are retrieved. No quantitative analysis of how these concepts relate to the symptoms in the DSM-V[14] is performed, however the aid of the GPT-4o model accessed through ChatGPT [68] was used to label the concepts. The retrieved examples are given to ChatGPT with the question whether it could find a unifying concept. This was done without telling it the classification task these concepts were used for, since that might color its response.

Library	Sci-Kit Learn	Torch	NLTK	Transformers	Zennit-Crp	Tensorflow	Spacy
Version	1.5.2	2.6.0	3.9.1	4.44.2	0.6.0	2.18.0	3.8.5

Table 4.2: Version of libraries used.

4.5. Experimental Setup

All experiments were performed on an Acer Nitro 5 laptop, which runs Windows, equipped with an Intel CPU and NVIDIA GPU. An overview of the version of the most important libraries used can be found in Table 4.2.

5. EXPERIMENTAL RESULTS

This chapter is divided into two main sections. The first presents a quantitative analysis of the performance of the proposed models, addressing the results needed to answer Research Question 1. The second provides a qualitative analysis of the identified concepts, addressing the results needed to answer Research Question 2.

5.1. Quantitative Experimental Results

In general, the deep learning methods performed the best, outperforming both the BoW model, which was used as a baseline for the dimensionality reduction methods, as well as all the dimensionality reduction methods themselves. The BoC models seem to perform the worst across the board. The LSA model seems to be the top performing dimensionality reduction model, only being outperformed by PCA in the D-Vlog dataset when performing feature selection. Further discussion of these results can be found in Section 6.1

5.1.1. D-Vlog

The $F1$ performance of the previously described methods is presented in Table 5.1. The best hyperparameters and normalization steps for the support vector machines and decision trees can be found in Appendix A.1, the ideal number of filters per filter size for both CNN_{mini} and CNN_{mpnet} was 8, however, the impact of this hyperparameter was insignificant. Confusion matrices of the top two approaches, CNN_{mini} and BoW, can be found in Table 5.2 and in Table 5.3.

5.1.2. EDAIC

The results of the proposed methods on the EDAIC dataset ares presented in Table 5.4. The hyperparameters leading to the best performance for the support vector machines and decision trees, as well as the best performing normalization, can be found

	BoW	BoW_{fs}	PCA	PCA_{fs}	LSA	LSA_{fs}	BoC	BoC_{fs}	CNN_{mini}	CNN_{mpnet}
$F1_{depres}$	<u>0.941</u>	0.822	0.931	0.925	0.938	0.904	0.730	0.804	**0.947**	0.929
$F1_{macro}$	<u>0.934</u>	0.810	0.924	0.912	0.928	0.890	0.768	0.781	**0.939**	0.917

Table 5.1: The $F1$ performances of tested models on the D-Vlog test set. A subscript of fs means feature selection was applied and t, the subscript on the CNN indicates what kind of word embedding was used. The best performing results are in bold and the second best performances are underlined.

		Predicted	
		Healthy	Depressed
Actual	Healthy	75	7
	Depressed	4	99

Table 5.2: Confusion Matrix of CNN_{mini} model trained on D-Vlog.

		Predicted	
		Healthy	Depressed
Actual	Healthy	77	5
	Depressed	7	96

Table 5.3: Confusion Matrix of BoW model trained on D-Vlog.

in Table A.2. For CNN_{mini}, the ideal number of filters per filter size was 12, while for CNN_{mpnet} it is 8. Since there is a significant gap between the depression $F1$ and the macro $F1$ it might be useful to explore the types of error the model makes. Across all dimensionality reduction methods, a consistent pattern emerges: the models tend to favor the majority class (non-depressed), resulting in a low recall. Additionally, because the number of depressed predictions is low, the precision estimates are based on limited data and should therefore be interpreted with caution. As for the deep learning methods, the confusion matrices of both models can be found in Tables 5.5 and 5.6.

	BoW	BoW_{fs}	PCA	PCA_{fs}	LSA	LSA_{fs}	BoC	BoC_{fs}	CNN_{mini}	CNN_{mpnet}
$F1_{depres}$	0.174	0.378	0.222	0.267	0.313	0.323	0.182	0.000	<u>0.467</u>	**0.571**
$F1_{macro}$	0.480	0.536	0.488	0.499	0.519	0.532	0.491	0.404	<u>0.636</u>	**0.714**

Table 5.4: The $F1$ performances of tested models on the EDAIC test set. A subscript of fs means feature selection was applied and t, the subscript on the CNN indicates what kind of word embedding was used. The best performing results are in bold and the second best performances are underlined.

		Predicted	
		Healthy	Depressed
Actual	Healthy	33	6
	Depressed	10	7

Table 5.5: Confusion Matrix of CNN_{mini} model trained on EDAIC.

		Predicted	
		Healthy	Depressed
Actual	Healthy	36	3
	Depressed	9	8

Table 5.6: Confusion Matrix of CNN_{mpnet} model trained on EDAIC.

5.2. Qualitative Analysis of Concepts

5.2.1. D-Vlog

5.2.1.1. Dimensionality Reduction. The concepts and their representative tokens of the PCA, LSA and BoC models can be found in Table 5.7, Table 5.8 and Table 5.9. Due to the fact that all these classifiers used a linear kernel, the value of the coefficient for each weight can naively be interpreted as feature importance and no post-hoc analysis needs to be done to calculate the feature importances.

In order to better understand the feature importances, it is also important to know the bias term of these classifiers. For example, the PCA classifier has a bias term of -0.626, which means that it considered a participant depressed unless proven otherwise. This might explain why the "non-depressed" indicating features on average have a higher absolute importance. The LSA classifier has a bias term of 0.046 and the BoC classifier has a bias term of 0.268. When comparing the concepts used by the models when PCA or LSA is applied, there are similarities and differences. For example, the "mental illness" concept is present in both. This concept contains tokens which together make adhd, and the token for bipolar. The PCA concept "student dialogue" and the LSA concept "talking" also have significant overlap. This is not surprising due to the similar methods in which PCA and LSA are calculated. A full list of all concepts discovered by both these approaches can be found in Appendix B. Besides these similarities, there are also some significant differences; the concepts of "foreign place" in PCA and "fishing" in LSA seem to be exclusive to their respective methods. The concepts uncovered through BoC are seemingly different, both in coherency and subject, from those discovered by PCA and LSA. Since there is some coherency in the discovered concepts, such as the words

"french" and "france" in the "urban fantasy drama" concepts. However, many groups of words cannot be unified under single concepts. This also means that it is hard to compare these concepts with those found by other methods, the only overlap is the "Teen Drama & Fashion Montage" from BoC and the "clothing" concept from LSA.

concept	Weight	Representative tokens
student dialogue	**8.210**	dialogue, hair, um, group, discussion, students, peer, student, mum, jordan
foreign place	5.840	spanish, portuguese, china, panda, language, tea, chinese, temple, shanghai, native
being mad	**-5.735**	fucking, ##hd, shit, experience, understand, fuck, ad, near, bait, death
student work	4.479	dialogue, paper, sheets, group, discussion, student, scrap, peer, ##dc, pieces
metal illness	-3.114	um, ##hd, ad, bipolar, doctor, bait, ##ination, uh, ##rus, bed
gaming	2.467	uh, ##cum, sugar, english, noble, creed, assassin, games, dislike, rai
european education	-1.349	paris, canterbury, university, hinges, sony, ##mute, great, anxiety, christchurch, trainers
buddism	1.126	nah, meditation, ##itate, buddhist, med, teach, become, wisdom, believe, quan
party	-0.765	age, barbie, sorority, sort, countries, country, alcohol, sugar, 21, ##ros
maths/ thinking	0.668	context, math, ##away, vulnerability, whoa, otis, pseudo, tan, winning, wheels

Table 5.7: Concepts discovered in the D-Vlog dataset through PCA. The most important weights for the concepts most important to non-depression (highest positive value) and depression (lowest negative value) are shown in bold.

concepts	Weight	Representative Tokens
fishing	**1.840**	bait, jerk, fish, fishing, spawn, water, rod, ##fish, shallow, pointer
talking	1.569	dialogue, hair, um, mum, jordan, ian, discussion, lip, group, peer
mental illness	**-1.174**	um, ##hd, ad, bait, doctor, bipolar, ##ination, uh, ##rus, bed
lingquistics	1.160	um, sentence, passive, verb, lip, bal, smack, paper, $, legitimate
clothing	-0.152	flea, market, ##rone, gonna, brand, $, winter, gym, warm, potion

Table 5.8: Concepts discovered in the D-Vlog dataset through LSA. The most important weights for the concepts most important to non-depression (highest positive value) and depression (lowest negative value) are shown in bold.

concepts	Weight	Representative Tokens
Urban Fantasy Drama	**1.453**	hairy obe witch softly edmonton kicking retail trenches coke mile long french equals oakland france drafted role bulldogs kicked
Glamorous Setbacks	1.297	blonde sided wedge landed successes miserable redhead champagne entertained
Design Identity	**-1.233**	evie patterns
Eccentric High Society	-1.034	gentlemen barton realising et dimension charting cute upon pale billionaire deux occupy mice lord gentleman lady underside corresponding craved shall starbucks makeup quo sings flying cells walt uber sphere henrik esq elvis geo petite weeds astronaut geometry calculus measure daring mathematics surpassed sparkle outlet shooter
British Academia & Governance	-1.006	salute wales maids complaint newcastle statement stool scholarly cambridge speak behalf free trust guards afternoons tackle coloured statements marry stoke speaks linen scoring officials resign
Historical Violence & Isolation	-0.751	menace passed onward 405 holden taunting 1893 70s 18th fleeting secluded ignored stairs rare cb
Morphemic Suffix	0.427	##li ##ic ##ar ##ni ##hi ##don ##ga ##ht ##um ##ks ##el ##nan ##or ##gy ##no ##der ##w ##sion ##na ##get ##ator ##ky ##gle ##ma ##gio ##lo ##go
Multilingual Function Words	0.308	##d so ve un che ar de nate lie ##gan we ku elle do she
Pop Science & Celebrity Culture	-0.248	voiced equations cliffs vinci maxwell republic waitress salty pope shea crashed flash sweet bye bug authors presley updates best
Creative Violence & Media Tropes	-0.229	v rite pour verse kill bag lash ang cade version card worthy hale king © eta ae end cards irs ami
Teen Drama & Fashion Montage	-0.097	rows velvet filmed lot crowd glitter veins finals undo wardrobe cocoa bbc spit clothes cups jingle bookstore towels grab refer dance cafeteria demi crouch earrings formal silk sacrifice break logo heads jeans perfume loaf alps numbers chairs

Table 5.9: Concepts discovered in the D-Vlog dataset through BoC. The most important weights for the concepts most important to non-depression (highest positive value) and depression (lowest negative value) are shown in bold.

5.2.1.2. CRP. From all concepts encoded in the second to last layer of the model, six concepts have relevance over 5%; these concepts account for 84.723% of all relevance in the model. 3 other concepts, each between 5% and 0.5% of the total relevance, make up another 12.682% of the total relevance. The amount of relevance generated by each concept throughout the test fold of the dataset can be found in Figure 5.1. A swarm plot of the relevance of all samples can be found in Appendix C.1. All but one of these concepts are indicative of the non-depressed class. In 66 of the 185 test samples, this concept was the only concept with non-zero relevance. The following is a list of all discovered concepts, ranked from most absolute relevance to lowest, a full list of all snippets which contributed to the classification of each concept can be found in Appendix D.1.

- The "Emotional Struggle" concept is the most relevant concept used by the model; besides being the most relevant concept, it is also the only concept that contributes towards a "depressed" classification. The snippets that were indicative of this concept all related to the topic of severe emotional struggle, with mentions of depression, anxiety, antidepressants, and suicide.
- The "Resilience & Determination" concept is characterized by the mention of some struggle, be it academic (ID 750), historic (ID 267), or interpersonal (ID 898), with a hopeful or determined tone.
- The "Collaborative Learning" concept is characterized by snippets where groups (or experiences in groups) in academic settings are mentioned.
- The "Dialogue Appreciation" concept is similar to the "Collaborative Learning" concept in the sense that the snippets talk about experiences in group settings, but whereas the "Collaborative Learning" focuses on the context of these groups, i.e, the academic setting; this concept is more about the way these groups are, with the main focus being appreciation of these groups having an open and tolerant atmosphere.
- The "Skepticism Toward Structure" concept indicates some rejection or critique of established structures or norms, for example, a critique of an academic institution (ID 794) or using a nontraditional makeup technique (ID 776)

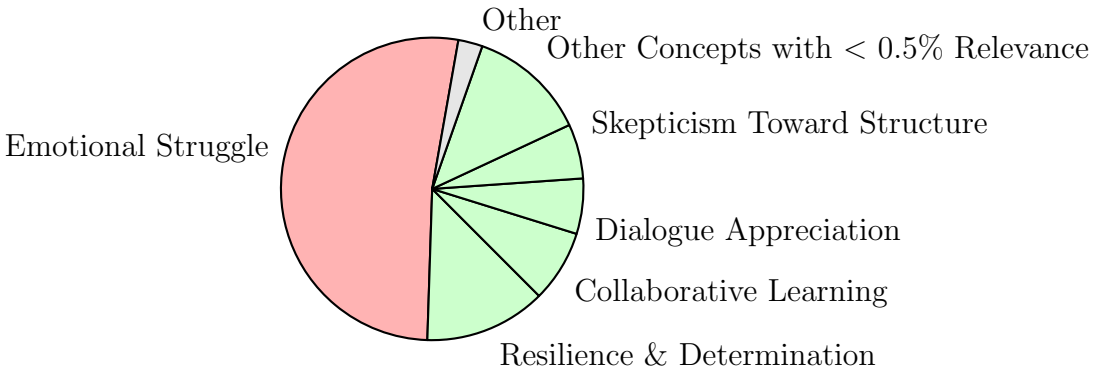

Figure 5.1: Distribution of concept relevance in D-Vlog. If a conept is colored red it contributes to a depressed classification, green to a non-depressed classification.

5.2.2. EDAIC

The concepts used by the dimensionality reduction methods are not very coherent. Since these models also do not have good performance, there is no indication that these concepts model something with meaning. Therefore, the only concepts explored in this section will be those used by the best performing model, the CNN model trained on mpnet embeddings.

From all concepts encoded in the second to last layer of the model, 6 concepts have relevance over 5%, these concepts account for 94.072% of all relevance in the model; the amount of relevance generated by each concept throughout the test fold of the dataset can be found in Figure 5.2, where a red color indicates the concepts contributes to a depressed classification and a green color indicates it contributes to a non-depressed classification. A swarm plot of the relevance of all samples can be found in Appendix C.2. Notable from this swarm plot is that in nearly all samples, all concepts have some relevance. This means that the model does not look at whether a concept is present or not, but that it balances the prevalence of these concepts to make a decision. The following is a list of all concepts, ranked from highest absolute relevance to lowest, and an explanation of their label. The top 5 segments for each concept can be found in Appendix D.2.

- The most relevant concept, “Being under Strain”, unifies enduring different types of challenges, be that financial struggle as seen in the transcripts of participants test_35 and test_21 or dealing with tiredness as in the transcripts of participants

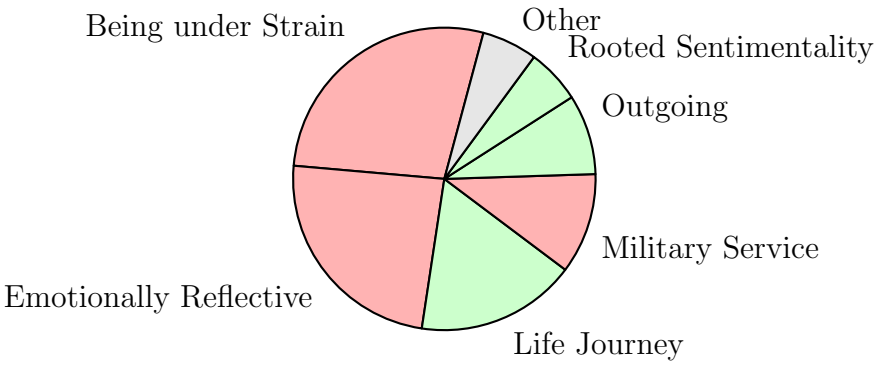

Figure 5.2: Distribution of concept relevance in EDAIC. If a conept is colored red it contributes to a depressed classification, green to a non-depressed classification.

test_21 and test_27.

- The "Emotionally Reflective" concept indicates some form of introspection or emotional reflection. For this concept, it is important to note that emotional reflection does not always indicate emotional expression. The very relevant "wow" present in the interviews for participants test_14 and test_02 is a good example of this. Often, this phrase is an indication of some internal readjustment. In addition to what is being said, the fact that responses relevant to this concept are often shorter could also be an indication of this concept, since people are often not very talkative when introspecting.
- The "Life Journey" concept indicates the subject of long-term life developments such as changes in where the participant lived or career changes.
- The "Military Service" concept is unified by the mentioning of the military or things that are often associated with the military or military experience such as PTSD.
- The "Outgoing" concept indicates extroversion in the participant, be that the subject directly calling themselves outgoing or the subject mentioning more extroverted hobbies such as traveling.
- The "Rooted Sentimentality" concept is unified by the fond memories with a sentimental tone.

6. CONCLUSION AND DISCUSSION

6.1. Performance Comparison (RQ 1)

6.1.1. Performance Comparison on D-Vlog

When feature selection was not applied, the BoW baseline model demonstrated the best performance among the SVM classifiers. However, the decline in $F1$ performance for the LSA and PCA models was less than 0.1; with this test set, that means that the models made one and two additional misclassifications respectively. This is not a significant change. The BoC model did perform significantly worse. Although the BoW model yielded the highest performance, it should be noted that it required substantially longer training times due to the necessity of training a separate parameter for each word in the vocabulary. When feature selection was applied, the BoW model experienced a substantial decrease in performance, whereas the PCA and LSA models showed only a marginal decline. Notably, the BoC model was the only one to exhibit improved performance following feature selection. In addition to the comparison of the SVM classifiers, it is worth mentioning that the best performance achieved in this study was from a deep learning model—specifically, a CNN trained on MiniLM word embeddings.

6.1.2. Performance Comparison on EDAIC

For the EDAIC dataset, all SVM models performed poorly. Despite being trained with $F1$ score as the optimization metric, the classifiers were still affected by class imbalance. These models demonstrated a strong tendency to predict the majority class (non-depressed). Consequently, none of the dimensionality reduction techniques had a significant effect on model performance. Feature selection also had minimal impact, except in the case of the baseline model, where performance improved significantly. Although this increase in performance may appear promising, the simplicity of such models limits their capacity to capture the complexity of this domain. Deep learning approaches, by contrast, show greater potential. For instance, the CNN_{mpnet} model correctly classi-

fied 36 out of 39 healthy individuals. On the surface, this might suggest that individuals classified as depressed by the model are likely to benefit from prioritized access to care. However, it is important to be aware of the base-rate fallacy [16], which this way of providing prioritized care would be an example of. If the class imbalance in the dataset reflects the actual prevalence of depression among individuals being screened, then prioritizing care based solely on model predictions would result in a substantial number of false-positive non-depressed individuals receiving priority, while approximately half of the truly depressed individuals would not be identified for expedited care.

6.2. Learned and Discovered Concepts (RQ 2)

This section will compare the various concepts discovered with the symptoms of major depressive disorder found in the DSM-V [14], specifically those listed under diagnostic criteria A. While none of the discovered concepts are direct one-to-one translations of symptoms this section will go into to what extent and how the discovered concepts relate to symptoms of major depressive disorder. Out of the 9 symptoms outlined in the DSM-V, the following 6 symptoms could be linked to a specific concept discovered in either dataset: a depressed mood (1), diminished interest or pleasure (2), insomnia and hypersomnia (4), fatigue (6), feelings of worthlessness (7) and diminished ability to think. How exactly these concepts are related to discovered concepts is outlined in the rest of this section. While suicidal ideation (9) was able to be linked to "bad mental health" concept in the D-Vlog dataset, this concept is so broad that it would be incorrect to say that these were specifically linked. Significant weight loss or gain (3) and psychomotor agitation or retardation (5) were never linked to any concept found in either dataset. The exact reasons for why these symptoms were not learned as concepts can only be speculated on. However, one possible explanation would be that these symptoms are physical characteristics, which may not be as obvious when only looking at text.

6.2.1. Concepts Discovered in D-Vlog

The concepts discovered through PCA, LSA and deep learning seem to keep to a general pattern, where 1 (or 2 in the case of PCA) general concept for "bad mental

health" is used, while all the other concepts are indications of non-depression. This one "bad mental health" concept can be linked to the symptoms of depression in several ways. However, its broadness also means that it could include normative emotions of extreme sadness like grief and therefore might not be very useful in validating a depression prediction. Although some of the concepts discovered through the BoC method have some aspects in common with those found by PCA and LSA, they are not nearly as coherent as those discovered through other methods and also do not bear any resemblance to the symptoms of major depressive disorder.

A lot of the non-depressed indicating concepts in PCA and LSA, there is one interesting link that can be made. The "student dialogue" and "talking" concepts both indicate social interaction. Links have been made between depression and a low sense of belonging [24] or loneliness [32], so while being in groups and talking do not guarantee someone feels less a sense of belong or less loneliness, a valid correlation can be made between depression and a lack of social interaction. The inverted (lots of social interaction is correlated to non-depression) could also be a valid link, making it a useful concept for depression classification. The rest of the concepts seem to be typical vlog topics: school, travel, hobbies, etc. While this is a valid distinction in this dataset, these concepts do not relate adequately to the symptoms of major depressive disorder. Besides the same link between non-depression and the social interaction from the "Dialogue Appreciation" and "Collaborative Learning" outlined above, the deep learning concepts do have some links to the DSM-V symptoms, the "resilience & determination" concept and the "Skepticism toward structure" are both incongruent with some of the symptoms in the DSM-V. Resilience and determination are at odds with the hopelessness mentioned in symptom 1, the fatigue or loss of energy mentioned in symptom 6 and the feeling of worthlessness mentioned in symptom 7. Furthermore, skepticism towards structure requires a form of critical thinking and steadfastness which would be the opposite of symptom 8, a diminished ability to think or concentrate, or indecisiveness.

6.2.2. Concepts Discovered in EDAIC

Similarly to the concepts found in the D-Vlog dataset, although no discovered concepts are one-to-one translations of symptoms of major depressive disorder, some cor-

relations can be made. In the “Being under Strain” concept, there are several snippets highly indicative of symptoms. Symptom 4, hypersomnia and insomnia, seems to be the subject of the snippet from participant test_21, while the mentions of tiredness in the snippet from participant test_27 are an example of symptom 6, fatigue or loss of energy. However, this concept is not exclusive to symptoms of depression, participant test_35 mentions financial struggle. Therefore, it would not be valid to say this concept indicates some symptom of depression. The DSM itself notes that “Responses to a significant loss ... may resemble a depressive episode.”, on how this response to a loss or struggle should be judged it says “This decision inevitably requires the exercise of clinical judgment”. [14]
The “emotionally reflective” concept does seem to have a higher correlation to a symptom; if a participant takes more time to reflect on their emotions, this could be an indication of a diminished ability to think or concentrate, or indecisiveness, symptom 8.
The “Life journey” does not have any direct correlation with any symptom, unless a connection between a diminished capacity for big-picture or long-term thinking and depressive symptoms.
The “military” concept has no clear connection to the symptoms of major depressive disorder, however, military service has been linked to an increased chance of major depressive disorder [64][17], which could explain why the model has learned this concept.
The “outgoing” concept is contraindicative of symptom 2, a diminished interest or pleasure in activities, it is a reasonable assumption that people who partake in a lot of social activities enjoy those activities.
If we take into account that the snippets from the “rooted sentimentality” concept were joyfully looking back at the past, this concept is opposed the feeling of sadness and emptiness described in symptom 1.

6.3. Limitations and Future Directions

The most significant limitation of this research is the absence of a formal evaluation of the discovered concepts. In particular, a user study on the interpretability of these concepts would have been valuable, especially given that the application of CRP to the text modality is novel. Furthermore, the labeling of these concepts was done in an ad

hoc manner, a more formal approach where different large language models would have been evaluated also might have proven useful. Research into concept-based methods for text modalities is still in its early stages, and there are still strides to be made in the research area.

When comparing the discovered concepts to symptoms outlined in the DSM-V specifically, another issue arises. Proper interpretation of the DSM-V requires clinical training, and as such, any interpretation made by a layperson, such as the author of this thesis, should be approached with appropriate caution and skepticism.

This thesis has also kept the datasets used separated; however, combining these datasets might increase performance significantly. Sun et al. [92] and Ahmend et al. [5] have combined the D-Vlog dataset and the DAIC-WoZ dataset (the precursor to EDAIC) with promising results. Besides combining datasets, possible performance increases might also be found in the usage of other models, different types of word embedding or data augmentation. Furthermore, during the training of the Support Vector classifier models, samples from each class were equally weighted. This made the models more susceptible to guessing majority class, changing the weights of samples based on their class frequency might improve performance significantly.

6.4. Ethical Concerns

Although it is highly unlikely that this research will cause any physical harm to anyone, there are ethical considerations that must be addressed. By their nature, explanations should give insight and inspire confidence in a neural network. However, this trust might be misplaced and overconfidence could be harmful. Smith et al. [86] note in their study on explanations and feedback in interactive ML that it is important to manage expectations of machine learning capabilities.

Furthermore, the consequences of training a neural network to detect depression should be noted. Currently, these networks are mainly intended to be used as a supportive tool for medical professionals, who should already be bound by ethical guidelines which prevent unethical use of this technology. However, it is still important to consider the consequences of possible unintended use as well. On the one hand, the dangers of mistakes could be grave. False negatives might prevent a depressed person from getting the

help they need and exacerbate their situation. False positives might cause someone to develop more depressive symptoms, similar to the nocebo effect [73]. The latter is also at odds with the primun-non-nocere (first do not harm) principle, one of the core tenets of medicine. On the other hand, there are also upsides to this technology being available to the general public. Firstly, a true positive answer might motivate someone to get treated for depression. Secondly, medical care might not be accessible due to, for example, an economic situation or long waiting times. In these cases, an AI might be preferable to no medical care.

Another topic one should consider when engaging with automatic depression detection is fairness. Models might be biased to some group due to a designer's or a society's biases. For example, Lai et al. [50] shows that some current diagnostic criteria exhibit sex-based biases. An overview of challenges faced by automatic depression detection, collected from physicians, can be found in Sogancioglu et al. [87]

References

[1] Ashraf Abdul et al. "COGAM: measuring and moderating cognitive load in machine learning model explanations". In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–14.

[2] Josh Achiam et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[3] Reduan Achtibat et al. "From attribution maps to human-understandable explanations through concept relevance propagation". In: *Nature Machine Intelligence* 5.9 (2023), pp. 1006–1019.

[4] Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160.

[5] Sabbir Ahmed et al. "Taking all the factors we need: A multimodal depression classification with uncertainty approximation". In: *IEEE Access* 11 (2023), pp. 99847–99861.

[6] Sharifa Alghowinem et al. "Interpretation of depression detection models via feature selection methods". In: *IEEE Transactions on Affective Computing* 14.1 (2020), pp. 133–152.

[7] *all-MiniLM-L6-v2*. URL: `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2` (visited on 05/20/2025).

[8] *all-mpnet-base-v2*. URL: `https://huggingface.co/sentence-transformers/all-mpnet-base-v2` (visited on 05/20/2025).

[9] Julie D Allen et al. "The unicode standard". In: *Mountain view, CA* (2012), pp. 660–664.

[10] André Altmann et al. "Permutation importance: a corrected feature importance measure". In: *Bioinformatics* 26.10 (Apr. 2010), pp. 1340–1347. ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btq134`.

[11] Leila Arras et al. "What is relevant in a text document?": An interpretable machine learning approach". In: *PloS one* 12.8 (2017), e0181142.

[12] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115.

[13] David Arthur and Sergei Vassilvitskii. *k-means++: The advantages of careful seeding*. Tech. rep. 2006.

[14] American Psychiatric Association. *The Diagnostic and Statistical Manual of Mental Disorders (5th ed.)* 2013, pp. 155–189. DOI: https://doi.org/10.1176/appi.books.9780890425596.

[15] Sebastian Bach et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140.

[16] Maya Bar-Hillel. "The base-rate fallacy in probability judgments". In: *Acta Psychologica* 44.3 (1980), pp. 211–233.

[17] Natalie Bareis and Briana Mezuk. "The relationship between childhood poverty, military service, and later life depression among men: evidence from the Health and Retirement Study". In: *Journal of affective disorders* 206 (2016), pp. 1–7.

[18] David Bau et al. "Understanding the role of individual units in a deep neural network". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30071–30078.

[19] Zhou Bolei et al. "Object detectors emerge in deep scene cnns". In: *International Conference on Learning Representations* (2015).

[20] Sushma Niket Borade and Ramesh P Adgaonkar. "Comparative analysis of PCA and LDA". In: *2011 International Conference on Business, Engineering and Industrial Applications*. IEEE. 2011, pp. 203–206.

[21] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[22] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.

[23] Vinton Gray Cerf. *ASCII format for network interchange*. Tech. rep. 1969.

[24] Chanokruthai Choenarom, Reg Arthur Williams, and Bonnie M Hagerty. "The role of sense of belonging and social support on stress and depression in individuals with depression". In: *Archives of psychiatric nursing* 19.1 (2005), pp. 18–29.

[25] Scott Deerwester et al. “Indexing by latent semantic analysis”. In: *Journal of the American society for information science* 41.6 (1990), pp. 391–407.

[26] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[27] Jacob Devlin et al. *google-bert/bert-base-uncased*. 2018. URL: https://huggingface.co/google-bert/bert-base-uncased (visited on 05/13/2025).

[28] Finale Doshi-Velez and Been Kim. “Considerations for Evaluation and Generalization in Interpretable Machine Learning”. In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Ed. by Hugo Jair Escalante et al. Cham: Springer International Publishing, 2018, pp. 3–17. ISBN: 978-3-319-98131-4. DOI: 10.1007/978-3-319-98131-4_1.

[29] Nicholas E Evangelopoulos. “Latent semantic analysis”. In: *Wiley Interdisciplinary Reviews: Cognitive Science* 4.6 (2013), pp. 683–692.

[30] Ming Fang et al. “A multimodal fusion model with multi-level attention mechanism for depression detection”. In: *Biomedical Signal Processing and Control* 82 (2023), p. 104561.

[31] Ricardo Flores et al. “Audiface: Multimodal deep learning for depression screening”. In: *Machine Learning for Healthcare Conference*. PMLR. 2022, pp. 609–630.

[32] Reid Griffith Fontaine et al. “Loneliness as a partial mediator of the relation between low social preference in childhood and anxious/depressed symptoms in adolescence”. In: *Development and Psychopathology* 21.2 (2009), pp. 479–491.

[33] Wikimedia Foundation. *Wikimedia Downloads*. URL: https://dumps.wikimedia.org.

[34] Amirata Ghorbani et al. “Towards automatic concept-based explanations”. In: *Advances in neural information processing systems* 32 (2019).

[35] David Gimeno-Gómez et al. “Reading between the frames: Multi-modal depression detection in videos from non-verbal cues”. In: *European Conference on Information Retrieval*. Springer. 2024, pp. 191–209.

[36] Jonathan Gratch et al. “The distress analysis interview corpus of human and computer interviews.” In: *LREC*. Reykjavik. 2014, pp. 3123–3128.

[37] Sepp Hochreiter. “The vanishing gradient problem during learning recurrent neural nets and problem solutions”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6.02 (1998), pp. 107–116.

[38] Matthew Honnibal et al. “spaCy: Industrial-strength Natural Language Processing in Python”. In: *Zenodo* (2020). DOI: 10.5281/zenodo.1212303.

[39] Qilu Jiao and Shunyao Zhang. “A brief survey of word embedding and its recent development”. In: *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. Vol. 5. IEEE. 2021, pp. 1697–1701.

[40] Nani Jin, Renjia Ye, and Peng Li. “Diagnosis of depression based on facial multi-modal data”. In: *Frontiers in Psychiatry* 16 (2025), p. 1508772.

[41] Dan Jurafsky and James H. Martin. *Speech and Language Processing (3rd ed. draft)*. https://web.stanford.edu/~jurafsky/slp3/(visited 2024-06-10). 2024.

[42] Yang Wha Kang et al. “Korea community health survey data profiles”. In: *Osong public health and research perspectives* 6.3 (2015), pp. 211–217.

[43] Gordon V Kass. “An exploratory technique for investigating large quantities of categorical data”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 29.2 (1980), pp. 119–127.

[44] Been Kim et al. “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)”. In: *International conference on machine learning*. PMLR. 2018, pp. 2668–2677.

[45] Han Kyul Kim, Hyunjoong Kim, and Sungzoon Cho. “Bag-of-concepts: Comprehending document representation through clustering words in distributed representation”. In: *Neurocomputing* 266 (2017), pp. 336–352.

[46] Yoon Kim. "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. DOI: 10.3115/v1/D14-1181.

[47] Kurt Kroenke et al. "The PHQ-8 as a measure of current depression in the general population". In: *Journal of affective disorders* 114.1-3 (2009), pp. 163–173.

[48] Deepthi Praveenlal Kuttichira et al. "Explaining black-box models using interpretable surrogates". In: *PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26–30, 2019, Proceedings, Part I 16*. Springer. 2019, pp. 3–15.

[49] Isaac Lage et al. "Human evaluation of models built for interpretability". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 7. 2019, pp. 59–67.

[50] Meng-Chuan Lai et al. "Sex/gender differences and autism: setting the scene for future research". In: *Journal of the American Academy of Child & Adolescent Psychiatry* 54.1 (2015), pp. 11–24.

[51] Qian Li et al. "A survey on text classification: From traditional to deep learning". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 13.2 (2022), pp. 1–41.

[52] Edward Loper and Steven Bird. *NLTK: The Natural Language Toolkit*. https://arxiv.org/abs/cs/0205028. Accessed: 2025-05-13. 2002.

[53] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[54] Vincent Peter C Magboo and Ma Sheila A Magboo. "Important Features Associated with Depression Prediction and Explainable AI". In: *International Conference on Well-Being in the Information Society*. Springer. 2022, pp. 23–36.

[55] Yanisa Mahayossanunt et al. "Explainable Depression Detection Based on Facial Expression Using LSTM on Attentional Intermediate Feature Fusion with Label Smoothing". In: *Sensors* 23.23 (2023), p. 9402.

[56] Anshu Malhotra and Rajni Jindal. "Deep learning techniques for suicide and depression detection from online social media: A scoping review". In: *Applied Soft Computing* 130 (2022), p. 109713.

[57] Christopher D Manning. *Introduction to information retrieval*. Syngress Publishing, 2008.

[58] Kaining Mao, Yuqi Wu, and Jie Chen. "A systematic review on automated clinical depression diagnosis". In: *npj Mental Health Research* 2.1 (2023), p. 20.

[59] Nikhil Marriwala, Deepti Chaudhary, et al. "A hybrid model for depression detection using deep learning". In: *Measurement: Sensors* 25 (2023), p. 100587.

[60] Aleix M Martinez and Avinash C Kak. "Pca versus lda". In: *IEEE Transactions on Pattern Analysis and Machine intelligence* 23.2 (2001), pp. 228–233.

[61] Stan Meyberg. "Use your words": Towards Gender Fairness for Multimodal Depression Detection". MA thesis. Utrecht University, 2024. URL: `https://studenttheses.uu.nl/handle/20.500.12932/46981`.

[62] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations". In: *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. 2013, pp. 746–751.

[63] Kevin P Murphy et al. "Naive bayes classifiers". In: *University of British Columbia* 18.60 (2006), pp. 1–8.

[64] Dimitrios Nasioudis et al. "Depression in military medicine cadets: a cross-sectional study". In: *Military Medical Research* 2 (2015), pp. 1–5.

[65] Matthew D Nemesure et al. "Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence". In: *Scientific reports* 11.1 (2021), p. 1980.

[66] Hung Viet Nguyen and Haewon Byeon. "Explainable Deep-Learning-Based Depression Modeling of Elderly Community after COVID-19 Pandemic". In: *Mathematics* 10.23 (2022), p. 4408.

[67] Meng Niu et al. “Hcag: A hierarchical context-aware graph attention model for depression detection”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 4235–4239.

[68] OpenAI. *GPT-4o: OpenAI’s GPT-4 optimized model*. Accessed: 2025-05-28. 2024. URL: https://openai.com/.

[69] World Health Organization. *Depressive disorder (depression)*. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/depression (visited on 05/28/2024).

[70] Aditya Parameswaran, Hector Garcia-Molina, and Anand Rajaraman. “Towards the web of concepts: Extracting concepts from large datasets”. In: *Proceedings of the VLDB Endowment* 3.1-2 (2010), pp. 566–577.

[71] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[72] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[73] Ivan Požgain, Zrinka Požgain, and Dunja Degmečić. “Placebo and nocebo effect: a mini-review”. In: *Psychiatria Danubina* 26.2 (2014), pp. 0–107.

[74] “Principal Component Analysis for Special Types of Data”. In: *Principal Component Analysis*. New York, NY: Springer New York, 2002, pp. 338–372. ISBN: 978-0-387-22440-4. DOI: 10.1007/0-387-22440-8_13.

[75] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[76] J. Ross Quinlan. “Induction of decision trees”. In: *Machine learning* 1 (1986), pp. 81–106.

[77] Anand Rajaraman and Jeffrey D Ullman. *Mining of massive datasets*. Autoedicion, 2011, pp. 1–17.

[78] Nils Reimers and Iryna Gurevych. “Sentence-bert: Sentence embeddings using siamese bert-networks”. In: *arXiv preprint arXiv:1908.10084* (2019).

[79] Antônio H Ribeiro et al. “Beyond exploding and vanishing gradients: analysing RNN training using attractors and smoothness”. In: *International conference on artificial intelligence and statistics*. PMLR. 2020, pp. 2370–2380.

[80] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “” Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[81] Fabien Ringeval et al. “Avec 2017: Real-life depression, and affect recognition workshop and challenge”. In: *Proceedings of the 7th annual workshop on audio/visual emotion challenge*. 2017, pp. 3–9.

[82] Fabien Ringeval et al. “AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition”. In: *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*. 2019, pp. 3–12.

[83] Stephen Robertson. “Understanding inverse document frequency: on theoretical arguments for IDF”. In: *Journal of documentation* 60.5 (2004), pp. 503–520.

[84] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. “Learning important features through propagating activation differences”. In: *International conference on machine learning*. PMLR. 2017, pp. 3145–3153.

[85] K Simonyan, A Vedaldi, and A Zisserman. “Deep inside convolutional networks: visualising image classification models and saliency maps”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR. 2014.

[86] Alison Smith-Renner et al. “No explainability without accountability: An empirical study of explanations and feedback in interactive ml”. In: *Proceedings of the 2020 chi conference on human factors in computing systems*. 2020, pp. 1–13.

[87] Gizem Sogancioglu et al. “Fairness in AI-based mental health: Clinician perspectives and bias mitigation”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Vol. 7. 2024, pp. 1390–1400.

[88] Kaitao Song et al. “Mpnet: Masked and permuted pre-training for language understanding”. In: *Advances in neural information processing systems* 33 (2020), pp. 16857–16867.

[89] Sanskar Soni, Satyendra Singh Chouhan, and Santosh Singh Rathore. "TextConvoNet: A convolutional neural network based architecture for text classification". In: *Applied Intelligence* 53.11 (2023), pp. 14249–14268.

[90] J Springenberg et al. "Striving for Simplicity: The All Convolutional Net". In: *ICLR (workshop track)*. 2015.

[91] Matthew Squires et al. "Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment". In: *Brain Informatics* 10.1 (2023), p. 10.

[92] Chenjian Sun et al. "A novel study for depression detecting using audio signals based on graph neural network". In: *Biomedical Signal Processing and Control* 88 (2024), p. 105675.

[93] Hao Sun et al. "Multi-modal adaptive fusion transformer network for the estimation of depression level". In: *Sensors* 21.14 (2021), p. 4764.

[94] Songbo Tan. "Neighbor-weighted k-nearest neighbor for unbalanced text corpus". In: *Expert Systems with Applications* 28.4 (2005), pp. 667–671.

[95] Yongfeng Tao et al. "DepMSTAT: Multimodal spatio-temporal attentional transformer for depression detection". In: *IEEE Transactions on Knowledge and Data Engineering* 36.7 (2024), pp. 2956–2966.

[96] Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. "Explainability of depression detection on social media: From deep learning models to psychological interpretations and multimodality". In: *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project*. Springer, 2022, pp. 289–320.

[97] Md Azher Uddin, Joolekha Bibi Joolee, and Kyung-Ah Sohn. "Deep multi-modal network based automated depression severity estimation". In: *IEEE transactions on affective computing* 14.3 (2022), pp. 2153–2167.

[98] Md Zia Uddin et al. "Deep learning for prediction of depressive symptoms in a large textual dataset". In: *Neural Computing and Applications* 34.1 (2022), pp. 721–744.

[99] Laurens Van Der Maaten, Eric O Postma, H Jaap van den Herik, et al. "Dimensionality reduction: A comparative review". In: *Journal of Machine Learning Research* 10.66-71 (2009), p. 13.

[100] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[101] Andreas Waldis, Luca Mazzola, and Alexander Denzler. "Towards eXplainable AI in Text Features Engineering for Concept Recognition". In: *Statistical Language and Speech Processing: 8th International Conference, SLSP 2020, Cardiff, UK, October 14–16, 2020, Proceedings 8*. Springer. 2020, pp. 122–133.

[102] Wenhui Wang et al. "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers". In: *Advances in neural information processing systems* 33 (2020), pp. 5776–5788.

[103] Uma Yadav and Ashish K Sharma. "A novel automated depression detection technique using text transcript". In: *International Journal of Imaging Systems and Technology* 33.1 (2023), pp. 108–122.

[104] Dongyang Yan et al. "Network-based bag-of-words model for text classification". In: *IEEE Access* 8 (2020), pp. 82641–82652.

[105] Le Yang et al. "Multimodal measurement of depression using deep learning models". In: *Proceedings of the 7th annual workshop on audio/visual emotion challenge*. 2017, pp. 53–59.

[106] Chih-Kuan Yeh et al. "On completeness-aware concept-based explanations in deep neural networks, Appendix A". In: *Advances in neural information processing systems* 33 (2020), pp. 20554–20565.

[107] Chih-Kuan Yeh et al. "On concept-based explanations in deep neural networks". In: (2019).

[108] Jeewoo Yoon et al. "D-vlog: Multimodal vlog dataset for depression detection". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 11. 2022, pp. 12226–12234.

[109] Xu Zhang et al. “Optimizing depression detection in clinical doctor-patient interviews using a multi-instance learning framework”. In: *Scientific Reports* 15.1 (2025), p. 6637.

[110] Bolei Zhou et al. “Interpretable basis decomposition for visual explanation”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 119–134.

[111] Li Zhou et al. “CAIINET: Neural network based on contextual attention and information interaction mechanism for depression detection”. In: *Digital Signal Processing* 137 (2023), p. 103986.

[112] Li Zhou et al. “TAMFN: Time-aware attention multimodal fusion network for depression detection”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2022), pp. 669–679.

[113] Yukun Zhu et al. “Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books”. In: *The IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015.

[114] Hamad Zogan et al. “Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media”. In: *World Wide Web* 25.1 (2022), pp. 281–304.

Glossary

AI Artificial Intelligence. 5, 7

AVEC Audio/Visual Emotion Challenge and Workshop. 25

BoC Bag of Concepts. iii, iv, 2, 3, 12, 17, 30, 31, 35, 37, 43

BoW Bag of Words. iii, iv, 2, 4, 12, 15, 17, 28–31, 35, 43

CF-IDF Concept Frequency-Inverse Document Frequency. 17

CNN Convolutional Neural Network. iii, iv, 5, 6, 43

CRP Concept Relevance Propagation. iii, iv, 2, 3, 13, 14, 21, 33, 46

DAIC Distress Analysis Interview Corpus. 25

DAIC-WoZ Distress Analysis Interview Corpus Wizard-of-Oz. x, 25, 26, 47

EDAIC Extended Distress Analysis Interview Corpus. iii, iv, x, 2, 25, 26, 30, 31, 35

LDA Linear Discriminant Analysis. 12

LIME Local Interpretable Model-Agnostic Explanations. 14, 23

LRP Layer-wise Relevance Propagation. 19–21

LSA Latent Semantic Analysis. iii, iv, 2, 3, 12, 15, 35, 37, 43, 44

MCR Model Class Reliance. 13

ML Machine Learning. 47

NLP Natural Language Processing. 4–6

PCA Principal Component Analysis. iii, iv, 2, 3, 12, 16, 35, 37, 43, 44

PFI Permutation Feature Importance. 14

PIMP Permutation IMPortance. 13, 14

RNN Recurrent Neural Network. 5, 6

SHAP SHapley Additive exPlanations. 13, 14

SVC Support Vector classifier. 11, 32, 33, 47

SVD Singular Value Decomposition. 15, 17

SVM Support Vector Machine. 11, 23, 43

TF-IDF Term Frequency-Inverse Document Frequency. 4, 17

XAI Explainable Artificial Intelligence. 4, 7, 9, 13, 24

Appendices

A. Hyperparameters

A.1. D-Vlog

Method	BoW	BoW_{fs}	PCA	PCA_{fs}	LSA	LSA_{fs}	BoC	BoC_{fs}
Dimensionality	-	-	200	200	200	100	500	1000
L2 normalization	✓		✓	✓	✓		✓	
Minmax normalize			✓	✓				
C	4	32	64	32	128	0.1	64	8
Gamma	scale	scale	auto	scale	auto	scale	auto	scale
Kernel	linear	linear	sigmoid	linear	rbf	linear	rbf	linear
Decision Tree Crit.	-	entropy	-	entropy	-	gini	-	gini
Min Weight Frac.	-	0.05	-	0.05	-	0.1	-	0.05

Table A.1: Best performing hyperparemters for D-Vlog dataset.

A.2. EDAIC

Method	BoW	BoW_{fs}	PCA	PCA_{fs}	LSA	LSA_{fs}	BoC	BoC_{fs}
Dimensionality	-	-	100	100	100	100	500	1000
L2 normalization	✓						✓	✓
Minmax normalize		✓			✓	✓	✓	✓
C	128	2	16	32	8	4	2	2
Gamma	auto	scale	scale	scale	scale	scale	scale	scale
Kernel	rbf	sigmoid	sigmoid	sigmoid	rbf	sigmoid	rbf	sigmoid
Decision Tree Crit.	-	gini	-	gini	-	entropy	-	gini
Min Weight Frac.	-	0.2	-	0.05	-	0.1	-	0.05

Table A.2: Best performing hyperparemeters for EDAIC dataset.

B. All concepts discovered during PCA

B.1. All concepts in D-Vlog dataset discovered during PCA

Table B.1: List of all Concepts discovered by PCA.

0	['##less', 'job', 'welcome', 'channel', 'able', 'city', 'say', 'many', 'technology', 'important']
1	['nah', 'meditation', '##itate', 'buddhist', 'med', 'teach', 'become', 'wisdom', 'believe', 'quan']
2	['um', 'women', 'uh', 'sugar', '##cum', 'children', 'gonna', 'put', 'cup', 'husband']
3	['tammy', 'gates', 'mr', 'detention', 'tape', 'taxi', 'mother', 'vc', 'backpack', 'ter']
4	['women', 'children', 'workplace', 'universities', 'academics', 'men', 'academic', 'choices', 'university', 'leadership']
5	['chain', '##oche', 'double', 'cr', 'row', 'segment', 'stitch', 'space', 'granny', '##t']
6	['uh', 'um', 'english', 'noble', 'chain', 'creed', 'assassin', '##oche', 'games', 'dislike']
7	['uh', '##cum', 'sugar', 'english', 'noble', 'creed', 'assassin', 'games', 'dislike', 'rai']
8	['##ination', '##ast', 'thesis', '##cr', 'pro', 'phd', '##hd', 'motivation', 'fear', 'environment']
9	['bait', 'jerk', 'fish', 'fishing', 'water', 'spawn', 'rod', '##fish', 'shallow', 'pointer']
10	['fucking', '##hd', 'shit', 'experience', 'understand', 'fuck', 'ad', 'near', 'bait', 'death']
11	['um', '##hd', 'ad', 'bipolar', 'doctor', 'bait', '##ination', 'uh', '##rus', 'bed']
12	['##hd', 'ad', 'dash', 'adult', 'adults', '##owe', 'number', 'coach', 'hall', '##ll']
13	['sentence', 'passive', 'verb', 'cake', 'object', 'subject', '##ici', 'tense', '##ple', '##eng']
14	['um', 'sentence', 'passive', 'verb', 'lip', 'bal', 'smack', 'paper', '$', 'legitimate']
15	['calendar', 'template', '##oj', 'tribe', '##da', 'um', 'ss', 'vibe', '##vy', '##xy']
16	['cheese', 'hedges', 'shrimp', 'toast', 'crab', 'plants', 'marina', 'olive', 'fence', 'lobby']
17	['mum', 'jordan', 'ian', '##dc', 'blah', 'cheese', 'vi', 'sister', 'folk', 'baby']
18	['china', 'tea', 'panda', 'temple', 'chinese', 'shanghai', '##dc', 'skulls', 'vi', '##bba']
19	['spanish', 'portuguese', 'china', 'panda', 'language', 'tea', 'chinese', 'temple', 'shanghai', 'native']
20	['lip', 'spanish', 'portuguese', 'smack', 'bal', 'legitimate', '##m', 'zoom', 'fucking', '$']
21	['spanish', 'portuguese', 'um', 'language', 'native', '##dc', 'learning', 'vi', 'learn', 'brazilian']
22	['slot', 'goodness', '##h', 'dialogue', 'planners', 'billy', 'students', 'scents', 'lip', 'planner']
23	['string', 'rack', 'pattern', 'strings', 'spin', 'ball', 'dense', '##et', 'cross', 'patterns']
24	['dialogue', 'paper', 'sheets', 'group', 'discussion', 'student', 'scrap', 'peer', '##dc', 'pieces']
25	['dialogue', 'hair', 'um', 'group', 'discussion', 'students', 'peer', 'student', 'mum', 'jordan']
26	['dialogue', 'affiliate', 'pregnancy', 'pregnant', 'ob', '##ses', '$', 'students', 'song', '##ester']
27	['song', 'affiliate', 'album', 'songs', 'records', '$', 'gem', 'pearl', 'record', 'band']
28	['affiliate', '##ense', 'ads', 'money', 'sponsors', 'content', 'product', 'sponsor', 'paper', 'links']

Continued on next page

Table B.1 – continued from previous page

29	['salmon', 'glory', 'gonna', 'christmas', 'wreath', 'garland', 'books', 'tri', 'decorate', 'rice']
30	['glory', '##ses', 'ob', 'anxiety', '##sive', 'highest', 'year', 'santa', 'kings', 'halloween']
31	['glory', 'song', 'paper', 'album', 'chain', 'sheets', 'christmas', 'highest', 'dialogue', 'songs']
32	['glory', 'chemotherapy', 'cancer', 'highest', 'santa', 'sun', 'kings', '##cre', 'um', 'christmas']
33	['folk', 'waving', 'mercy', 'chemotherapy', 'drowning', 'dogs', 'business', 'et', '##sy', 'folks']
34	['korean', 'japanese', 'folk', '##nam', 'bus', 'students', 'chemotherapy', 'korea', 'gang', '##lk']
35	['chemotherapy', '##cre', 'cancer', 'pan', '##atic', 'nausea', '##as', '##mo', 'med', '$']
36	['$', 'folk', 'glory', '##ssel', 'tyler', 'stud', 'fucking', 'bracelet', 'japanese', 'size']
37	['paralysis', 'dialogue', 'sleep', 'dream', 'dreaming', 'luc', 'scream', 'door', 'anxiety', 'hallway']
38	['students', 'classroom', 'school', 'teaching', 'year', 'teacher', 'kids', 'space', 'storage', 'student']
39	['hair', 'paralysis', 'extensions', 'afro', 'sleep', 'curly', 'colour', 'dream', 'dreaming', 'curl']
40	['hair', 'tickets', 'pen', 'teller', 'timbers', 'elton', 'concert', 'extensions', 'match', 'japanese']
41	['tickets', 'teller', 'pen', 'paralysis', 'students', 'elton', 'concert', 'classroom', 'dream', 'ill']
42	['fucking', 'pen', 'tickets', 'teller', 'hair', 'shit', 'dialogue', 'elton', 'gonna', 'dissertation']
43	['timbers', 'fucking', 'match', 'boating', 'miller', 'paralysis', 'fantastic', 'students', 'lead', 'galaxy']
44	['er', 'anxiety', '##m', 'panic', 'hair', 'timbers', 'love', 'te', '##sco', 'attack']
45	['er', '##m', 'paralysis', 'te', '##sco', 'cute', 'wa', '##yst', 'vinci', 'patio']
46	['patio', 'plant', 'er', 'storage', '##aks', 'kay', 'fishing', '##m', 'toy', 'gonna']
47	['er', '##m', 'patio', 'plant', 'wa', 'bath', 'te', '##sco', 'pot', 'timbers']
48	['patio', 'plant', 'pot', 'plants', 'gravel', 'roots', 'height', 'blooms', 'bottom', 'pain']
49	['uh', 'fucking', 'fuck', 'shit', 'pain', 'scan', 'classroom', 'chemotherapy', '##ination', '##ast']
50	['dude', '##iate', '##um', '##one', 'uh', '##ani', 'love', 'op', 'cr', 'freaking']
51	['phd', 'stage', 'er', 'motivation', 'task', 'fucking', 'thesis', 'pain', '##m', 'uh']
52	['dude', '##iate', 'pain', '##um', '##one', 'daughter', '##ani', 'op', 'son', 'bus']
53	['uh', '##ester', 'pieces', 'love', 'disability', 'coffee', 'historical', 'trim', '##fe', 'caf']
54	['love', '##mart', 'wal', 'husband', 'pain', 'pieces', 'car', 'water', 'tattoo', 'gonna']
55	['uh', 'son', 'pain', 'disability', 'tattoo', 'theater', 'sheets', 'historical', 'kristen', 'quarter']
56	['pieces', '##ection', '##oll', 'dude', 'tease', 'colour', 'uh', 'pink', 'purple', '##um']
57	['$', 'anxiety', 'panic', 'nike', 'bought', 'cousins', 'flea', 'routine', 'blue', 'colour']
58	['love', 'design', 'colored', 'drew', 'paint', '##ester', 'marker', '##ayo', 'toby', '2008']
59	['scan', 'love', 'baby', 'anxiety', 'panic', 'scans', 'pregnancy', '##hora', 'sep', 'celebrity']
60	['nursing', 'peer', 'clinical', 'pressure', 'book', 'dvd', 'love', 'ste', '##cope', 'math']
61	['uganda', 'army', 'resistance', 'love', 'lord', 'legislation', 'husband', 'sheets', 'classroom', '##gold']
62	['tattoo', 'love', 'piercing', 'tattoos', 'panther', 'ear', 'stretching', '$', 'brush', 'ribs']
63	['uganda', 'nursing', 'army', 'resistance', 'tattoo', 'peer', 'lord', 'clinical', 'piercing', 'pressure']

Continued on next page

Table B.1 – continued from previous page

64	['##min', 'uganda', 'pepper', 'names', 'diego', 'tattoo', 'moist', 'pre', 'mandarin', 'celebrity']
65	['uganda', 'tattoo', 'piercing', 'hearth', '##stone', 'tattoos', 'apply', 'scan', 'ear', 'potion']
66	['governing', '##oint', 'miracles', 'cease', 'prop', '##tower', 'evidence', 'claims', '##hes', 'god']
67	['governing', '##oint', 'classroom', 'miracles', 'tattoo', 'cease', 'piercing', '##tower', 'evidence', 'prop']
68	['piccolo', 'hearth', '##stone', '##nation', 'sharp', '##ggio', '##uch', 'tune', '##min', 'playing']
69	['peer', 'pressure', '##min', 'pepper', 'allowed', 'ski', 'diego', 'mandarin', '##zer', 'sep']
70	['peer', 'goodwill', 'velvet', 'tori', 'pressure', 'mom', 'gonna', 'gloss', 'lip', '##ging']
71	['piccolo', 'goodwill', 'scan', 'velvet', 'tori', 'potion', 'gloss', '##nation', '##uch', '##ggio']
72	['potion', 'peer', 'potter', 'piccolo', 'harry', 'merlin', 'included', 'sparkle', 'pressure', 'polish']
73	['piccolo', 'names', '##nation', '##ggio', '##uch', 'sharp', 'ultrasound', 'tune', '##ure', 'brush']
74	['flea', 'market', 'gonna', '##rone', 'brand', 'winter', '$', 'tun', 'target', 'cars']
75	['hearth', '##stone', 'flea', 'market', '##rone', 'names', 'winter', 'brand', 'peer', 'debt']
76	['vera', '##oe', 'al', 'juice', 'gonna', 'bacon', 'weight', 'plug', 'eating', 'jerk']
77	['jerk', 'fishing', 'pointer', 'slender', 'diver', 'game', 'deep', '##9', 'j', 'apply']
78	['hearth', '##stone', 'cars', 'videos', 'content', 'nike', 'cousins', 'versus', 'enjoy', '##not']
79	['hearth', 'debt', '##stone', 'gun', '##fish', 'sweat', 'beaver', 'towels', 'saskatchewan', 'troll']
80	['hearth', '##stone', 'versus', 'juice', 'vera', 'gym', 'series', '##oe', 'weaver', 'content']
81	['lights', 'shannon', 'debt', 'vera', '##oe', 'eating', 'agency', 'juice', 'lips', 'umbrella']
82	['cars', 'car', 'channel', '##not', 'coffee', 'polish', '##ma', 'routine', 'dealers', 'nail']
83	['wear', 'tan', 'bi', 'weekly', 'tonight', 'mental', 'disorder', '##cd', 'clock', 'panic']
84	['pee', 'tests', 'jerk', 'tt', '##c', '##ive', 'spawn', 'test', '##vu', 'lights']
85	['jerk', 'spawn', 'tan', 'bi', 'tonight', 'weekly', 'pointer', 'wear', 'gym', 'depot']
86	['collect', 'howl', 'dolls', 'buttons', 'festival', 'clean', 'fuse', 'tan', 'hood', '##rift']
87	['happiness', 'nail', 'polish', 'university', 'happy', 'color', 'nails', '##ester', 'spawn', '12']
88	['jerk', 'tun', 'spawn', 'debt', 'pointer', 'whoa', 'polish', 'nail', 'target', '##ging']
89	['cars', 'debt', 'gonna', '##not', '##press', 'car', 'university', 'canterbury', 'hearth', '##l']
90	['feeling', 'cinnamon', '##ants', '##de', 'gym', '##press', 'integration', 'anti', 'coach', 'children']
91	['anxiety', 'cars', 'nail', 'panic', 'polish', 'channel', 'color', '##not', '##ty', 'gonna']
92	['cinnamon', 'integration', 'coach', 'subway', 'aroma', '##therapy', '##fusion', 'research', 'breath', '##tam']
93	['paris', 'download', 'pee', 'bray', 'banjo', 'tests', 'pot', '##ty', 'florence', 'tt']
94	['depot', 'coffee', 'dutch', 'christy', 'beer', 'cars', 'brothers', 'roger', 'paris', 'controller']
95	['hinges', 'sony', 'cable', '##7', 'upside', 'screw', 'rotate', 'rip', 'ribbon', 'flip']

Continued on next page

Table B.1 – continued from previous page

96	['sony', 'hinges', 'student', 'alternative', 'cable', 'camera', 'upside', 'nail', '##7', 'polish']
97	['hinges', 'sony', 'cable', 'scan', 'upside', '##7', '##oche', 'cricket', 'stitch', 'single']
98	['paris', 'blah', 'student', 'alternative', '##how', 'whoa', 'mike', 'tan', '##aru', 'test']
99	['collect', 'howl', 'testosterone', 'canterbury', 'sort', 'download', 'anxiety', 'buttons', 'canal', 'fuse']
100	['paris', 'canterbury', 'university', 'hinges', 'sony', '##mute', 'great', 'anxiety', 'christchurch', 'trainers']
101	['paris', '##ants', '##press', 'stitch', 'anti', '##de', 'canterbury', '##oche', 'fat', 'tree']
102	['canterbury', 'panic', 'christchurch', 'cricket', 'university', 'gym', '##mute', 'sport', 'train', 'fat']
103	['student', 'alternative', 'escape', 'gloss', '##eral', 'mascara', 'fusion', 'jeans', 'drugs', 'smart']
104	['gloss', 'fusion', 'lip', 'stitch', '##oche', 'appointment', 'tree', 'xl', 'bryant', 'nine']
105	['stitch', '##oche', 'single', 'yoga', 'tree', 'pot', 'train', 'se', 'nine', 'cricket']
106	['workout', 'beats', 'panic', 'marketing', 'sean', 'hinges', 'sony', 'hearth', 'eating', 'jobs']
107	['canterbury', 'cricket', 'university', 'christchurch', '##mute', 'sport', 'gym', 'societies', 'testosterone', 'belly']
108	['canterbury', 'gloss', 'train', 'christchurch', 'appointment', 'hinges', 'fusion', '##mute', 'sony', 'paris']
109	['figures', 'gym', 'paris', 'workout', 'card', 'trade', 'energy', 'levels', 'jones', 'figure']
110	['cricket', 'canterbury', 'sport', '!', 'chain', 'appointment', 'dogg', 'football', 'clock', 'christchurch']
111	['cricket', 'sport', 'gloss', 'football', 'web', 'beats', 'marketing', 'yoga', 'fusion', 'pot']
112	['elton', '##ella', 'pennant', 'concert', 'fishing', 'tomorrow', 'walk', 'workout', 'positive', 'paris']
113	['cricket', 'vulnerability', 'stereotypes', 'sport', 'football', 'sean', 'pride', 'roles', 'therapist', 'student']
114	['sean', 'jobs', 'marketing', 'cricket', 'beats', '##grad', 'assistants', 'tuition', 'graduate', 'resume']
115	['cricket', 'figures', 'workout', 'eating', 'whoa', 'tan', 'figure', 'sport', 'chewing', 'trade']
116	['train', 'excuses', 'belief', 'figures', 'bus', 'subway', 'context', 'black', 'training', 'illness']
117	['gag', 'cricket', 'workout', '##ty', 'pot', 'figures', 'vomiting', 'nightmare', 'testosterone', 'bus']
118	['beats', 'canterbury', 'marketing', 'figures', 'christchurch', '##mute', 'vegas', 'pot', 'money', 'university']
119	['whoa', 'belief', 'vulnerability', 'escape', 'gag', '##ging', 'destiny', 'cricket', 'newcastle', 'gym']
120	['testosterone', '##pressive', 'subway', 'bus', 'sean', 'de', 'generation', 'degree', 'brain', '##grad']
121	['pot', '##ty', 'spotting', 'trained', 'elton', 'tan', 'pennant', '##ella', 'son', 'canterbury']
122	['gag', 'vomiting', 'nightmare', '##ging', 'fishing', 'fear', 'ph', 'positive', 'cricket', 'silly']
123	['whoa', 'workout', 'figures', 'newcastle', 'gloss', 'loose', 'function', 'fusion', 'episode', 'cricket']
124	['whoa', 'testosterone', 'bus', 'gym', 'pot', 'newcastle', 'belief', 'kids', '##ros', 'quran']
125	['mike', 'figures', 'whoa', 'context', 'math', 'sean', '##ty', 'michelle', 'trade', 'quran']

Continued on next page

Table B.1 – continued from previous page

126	['cricket', 'mike', 'beats', 'marketing', 'bus', 'paris', 'sport', 'tonight', 'michelle', 'choose']
127	['whoa', 'gym', 'newcastle', 'workout', 'record', 'journey', 'mike', 'weights', 'mountain', '108']
128	['whoa', 'gag', 'vulnerability', 'figures', 'stereotypes', 'nightmare', 'vomiting', 'newcastle', 'train', 'pot']
129	['testosterone', 'mike', 'vulnerability', 'stereotypes', 'yoga', 'southern', 'pe', 'filmmaking', 'northern', 'south']
130	['testosterone', 'bus', 'recording', '##0', 'record', 'subway', 'constantly', '108', 'keyboard', '##mist']
131	['whoa', 'panic', 'filmmaking', '##cd', 'therapist', 'newcastle', 'weights', 'brain', 'attacks', 'job']
132	['testosterone', 'fishing', 'ski', 'disorder', 'tomorrow', 'belief', 'mike', 'skiing', 'vi', 'nano']
133	['whoa', 'filmmaking', 'ski', 'northern', 'skiing', 'weights', 'newcastle', '##eem', 'snow', 'southern']
134	['belief', 'testosterone', 'destiny', 'escape', 'southern', 'energy', 'expression', 'blog', 'stereotypes', 'mike']
135	['weights', 'context', 'mountain', 'math', 'yoga', 'southern', 'filmmaking', 'question', 'bus', 'canterbury']
136	['whoa', 'testosterone', 'excuses', 'fishing', 'train', 'training', 'tan', 'pussy', 'walk', 'newcastle']
137	['filmmaking', 'northern', 'production', 'script', 'subway', 'montana', 'cinematographer', 'director', 'films', 'practical']
138	['context', 'math', '##away', 'vulnerability', 'whoa', 'otis', 'pseudo', 'tan', 'winning', 'wheels']
139	['accommodation', 'weight', 'filmmaking', 'chewing', 'building', 'energy', 'rooms', 'weights', 'generation', 'vitamin']
140	['climax', 'weights', 'mountain', 'esteem', 'disability', 'nora', '##£', 'surgery', 'pack', 'hoping']
141	['subway', 'tram', 'stereotypes', 'barbie', 'quran', 'southern', 'underground', 'fishing', 'cheese', 'eight']
142	['yoga', 'chewing', '##lvis', 'spitting', 'ski', 'bus', 'subway', '##rien', 'quran', 'illness']
143	['ski', 'subway', 'skiing', '##eem', '##pon', 'us', 'tram', 'illness', 'loser', 'whoa']
144	['bus', 'context', 'size', 'subway', 'chewing', 'weights', 'washing', 'social', 'bracelet', 'tired']
145	['otis', 'generation', 'discussion', '##si', 'challenging', 'weights', '##£', 'trainer', 'center', 'guys']
146	['generation', 'fishing', 'constantly', 'nice', 'whoa', 'context', 'bull', 'beats', '##away', 'tomorrow']
147	['southern', 'subway', 'weights', 'south', 'spirit', 'ski', 'north', 'skiing', 'flag', 'mountain']
148	['subway', 'nora', 'tram', 'energy', 'kuwait', 'autism', 'ski', 'constantly', '##ied', 'bull']
149	['subway', 'chewing', 'spitting', 'tram', 'vulnerability', 'disorder', 'age', 'food', '##mist', '##ssi']
150	['yoga', 'testosterone', '##lvis', 'bangs', 'nora', '##rien', 'um', 'subway', 'question', 'pe']
151	['chewing', 'spitting', 'cute', 'eating', 'whoa', 'isabella', 'bangs', 'bows', 'brazil', 'food']
152	['southern', 'south', 'bangs', 'north', 'confederate', 'paper', 'advisor', 'regional', 'nails', 'flag']

Continued on next page

Table B.1 – continued from previous page

153	['weights', '##away', 'spirit', 'gag', 'industry', 'customers', 'context', 'mountain', 'appointment', 'winning']
154	['generation', 'emotion', 'suffer', 'vulnerability', 'weights', 'yoga', 'montana', 'quran', 'subway', 'snow']
155	['weights', 'race', 'mountain', 'african', 'snow', 'panic', 'american', 'energy', 'birth', 'otis']
156	['bangs', 'workout', 'customers', 'hair', 'social', 'birth', 'excited', 'customer', '2', 'cut']
157	['chewing', 'spitting', 'bangs', 'eating', 'disorder', 'digest', 'southern', 'food', 'age', '##ive']
158	['bangs', 'barbie', 'mental', 'temple', 'sorority', 'anxiety', '##z', 'context', 'health', 'hair']
159	['yoga', 'bangs', 'article', 'subway', 'isabella', 'barbie', 'bows', 'natalie', 'gotta', 'gonna']
160	['weights', 'workout', 'age', 'mountain', 'spirit', 'college', 'social', '2', 'depot', 'privileged']
161	['bangs', 'weights', 'workout', 'customer', 'customers', 'coffee', 'exam', 'lips', 'door', 'manager']
162	['barbie', 'sorority', 'age', 'vulnerability', 'yoga', 'love', 'spirit', 'year', 'snow', 'country']
163	['age', 'generation', 'bangs', 'drinking', 'exam', '##p', 'college', 'guys', 'opinions', 'weights']
164	['weights', 'vulnerability', 'mountain', 'african', 'race', 'discussion', 'bangs', 'industry', 'yoga', 'struggle']
165	['bangs', 'snow', '##ius', 'montana', 'thanksgiving', 'pine', 'vulnerability', 'post', '##£', 'minus']
166	['exam', 'weights', 'yoga', 'barbie', 'lips', 'ry', 'un', '##i', 'fingers', 'napoleon']
167	['bangs', 'bind', 'vulnerability', 'daughter', 'dry', 'climax', 'black', 'american', 'african', 'ry']
168	['weights', 'generation', 'suffer', 'race', 'attacks', 'african', 'panic', 'peppers', 'accommodation', 'mike']
169	['bangs', 'barbie', 'posts', 'sorority', 'sunday', 'te', '69', '##sco', 'traumatic', 'exam']
170	['bangs', 'weights', 'temple', 'traumatic', 'cu', '##z', 'grief', 'daughter', 'journal', 'mountain']
171	['bangs', 'surgery', 'barbie', 'damn', 'constantly', 'discussion', 'exam', 'african', 'quarters', 'spirit']
172	['barbie', 'weights', 'sorority', 'dresses', 'cooks', '##tum', '##par', 'shout', 'son', 'chewing']
173	['bangs', 'temple', '##z', 'wounds', 'fail', 'spirit', 'earth', 'eight', 'cu', 'kuwait']
174	['customer', 'customers', 'social', 'makeup', 'bows', 'pay', 'space', '##£', 'isabella', 'natalie']
175	['traumatic', 'spirit', 'barbie', 'te', 'journal', '##sco', 'advisor', 'faith', 'grief', 'paper']
176	['te', '##sco', 'coach', 'february', 'march', 'mental', 'vulnerability', 'children', 'hollis', 'group']
177	['exam', 'bangs', 'mom', 'clemson', 'tiger', 'channel', '##i', 'exams', '##away', 'favored']
178	['bangs', 'advisor', 'paper', 'graduation', 'sugar', 'weights', 'mental', 'bar', 'health', 'weekend']
179	['age', 'barbie', 'sorority', 'sort', 'countries', 'country', 'alcohol', 'sugar', '21', '##ros']
180	['frequency', 'blind', 'image', 'means', 'sensory', 'substitution', 'demonstration', 'blessed', 'horizontal', 'tone']
181	['discussion', 'te', 'group', 'fighting', '##sco', 'son', 'emotions', 'felt', 'jug', 'twitching']
182	['article', 'twitter', 'shout', 'makeup', 'teeth', 'age', 'attacks', 'vulnerability', 'recruiting', 'youtube']

Continued on next page

Table B.1 – continued from previous page

183	['sugar', 'sensitivity', 'cha', 'candy', 'sort', 'blush', 'vi', 'teeth', 'horrible', '##rly']
184	['image', 'frequency', 'blind', 'barbie', 'article', 'pedestal', 'otis', 'sensory', 'guys', 'sorority']
185	['bind', 'blind', 'frequency', 'interview', 'stream', 'sensory', 'weights', 'pts', 'substitution', 'image']
186	['sugar', 'temple', 'discussion', 'southern', 'students', 'exposure', 'favored', 'guys', '##c', 'pit']
187	['sugar', 'lips', 'add', '##eral', 'sensitivity', 'conceal', 'health', 'mike', 'group', 'uk']
188	['interview', '##away', 'super', '##pse', '##ram', 'scary', '##la', '##ig', 'mill', 'affects']
189	['presentation', 'final', 'kill', 'legend', 'article', 'lastly', 'crying', 'self', 'screaming', 'story']
190	['stereotypes', 'sort', 'asian', 'faith', 'sticky', 'beliefs', 'stereo', 'cute', '##type', 'coach']
191	['sugar', '##l', 'rather', 'shout', 'customer', 'media', '##eral', 'motivation', 'ke', 'photography']
192	['sugar', 'rather', 'article', '##tum', '##par', 'crossed', 'exam', 'fingers', 'download', '##ig']
193	['stereotypes', 'sticky', 'asian', 'stereo', 'lips', '##type', 'cut', '11', 'discussion', 'bind']
194	['bind', 'racism', 'racist', 'presentation', '##ers', 'final', 'favored', 'thesis', 'guys', 'legend']
195	['age', 'panic', 'health', 'teeth', 'attacks', 'barbie', 'article', 'podcast', 'snow', 'sick']
196	['host', '##par', '##tum', 'germany', 'sabbath', 'privileged', 'country', 'hotels', 'relationship', 'customer']
197	['sugar', 'interview', 'beliefs', 'manager', 'snow', 'office', 'candy', 'meetings', 'sensitivity', 'faith']
198	['##pse', '##la', 'police', '##ram', 'uh', 'yu', 'bind', 'sleep', 'group', 'funeral']
199	['interview', 'sugar', 'sunglasses', 'collaboration', 'blah', 'josh', 'quit', 'police', 'stress', 'website']

B.2. All concepts discovered in D-Vlog dataset during LSA

Table B.2: List of all concepts discovered by LSA.

0	['##less', 'job', 'welcome', 'channel', 'able', 'city', 'say', 'many', 'technology', 'important']
1	['nah', 'um', 'meditation', 'uh', 'gonna', 'okay', 'women', 'got', '##itate', 'put']
2	['nah', 'meditation', '##itate', 'buddhist', 'med', 'teach', 'wisdom', 'become', 'quan', '##ji']
3	['tammy', 'gates', 'mr', 'detention', 'tape', 'taxi', 'mother', 'vc', 'backpack', 'ter']
4	['women', 'children', 'workplace', 'universities', 'academics', 'men', 'academic', 'choices', 'university', 'leadership']
5	['chain', '##oche', 'double', 'cr', 'row', 'segment', 'stitch', 'space', 'granny', '##t']
6	['uh', 'chain', '##oche', 'um', 'english', 'noble', 'creed', 'double', 'assassin', 'row']
7	['uh', 'english', 'noble', 'creed', 'assassin', 'games', 'dislike', '##cum', 'sugar', 'precisely']
8	['##ination', '##ast', 'thesis', '##cr', 'pro', 'phd', '##hd', 'motivation', 'fear', 'environment']
9	['bait', 'jerk', 'fish', 'fishing', 'spawn', 'water', 'rod', '##fish', 'shallow', 'pointer']
10	['fucking', '##hd', 'shit', 'experience', 'understand', 'fuck', 'ad', 'near', 'death', 'grasp']

Continued on next page

Table B.2 – continued from previous page

11	['um', '##hd', 'ad', 'bait', 'doctor', 'bipolar', '##ination', 'uh', '##rus', 'bed']
12	['##hd', 'ad', 'dash', 'adult', 'adults', '##owe', 'number', 'coach', 'hall', '##ll']
13	['sentence', 'passive', 'verb', 'cake', 'object', 'subject', '##ici', 'tense', '##ple', '##eng']
14	['um', 'sentence', 'passive', 'verb', 'lip', 'bal', 'smack', 'paper', '$', 'legitimate']
15	['calendar', 'template', 'tribe', '##oj', '##da', 'um', 'ss', 'vibe', '##vy', '##xy']
16	['cheese', 'hedges', 'shrimp', 'toast', 'crab', 'calendar', 'plants', 'um', 'marina', 'olive']
17	['mum', 'jordan', 'ian', 'cheese', 'blah', 'dialogue', 'spanish', 'portuguese', 'folk', 'hedges']
18	['spanish', 'portuguese', 'slot', 'language', 'paper', 'cute', 'students', 'native', 'learning', 'dialogue']
19	['china', 'panda', 'tea', 'spanish', 'portuguese', 'temple', 'chinese', 'shanghai', 'skulls', 'language']
20	['lip', 'smack', 'bal', 'legitimate', 'spanish', 'portuguese', '##m', 'zoom', 'fucking', '$']
21	['spanish', 'portuguese', 'um', 'language', '##dc', 'native', 'vi', 'learning', 'learn', 'brazilian']
22	['dialogue', 'group', 'student', 'discussion', 'students', 'peer', 'string', 'anxiety', '##ses', 'ob']
23	['slot', 'goodness', '##h', 'dialogue', 'planners', 'billy', 'students', 'scents', 'lip', 'planner']
24	['string', 'rack', 'pattern', 'strings', 'spin', 'ball', 'dense', '##et', 'cross', 'patterns']
25	['dialogue', 'paper', 'sheets', 'group', '##dc', 'scrap', 'discussion', 'pieces', 'student', 'peer']
26	['dialogue', 'hair', 'um', 'mum', 'jordan', 'ian', 'discussion', 'lip', 'group', 'peer']
27	['song', 'affiliate', 'album', 'songs', 'records', '$', 'gem', 'pearl', 'record', 'band']
28	['affiliate', '##ense', 'ads', 'money', 'sponsors', 'content', 'product', 'sponsor', 'links', 'google']
29	['salmon', 'glory', 'gonna', 'christmas', 'wreath', 'garland', 'books', 'tri', 'decorate', 'rice']
30	['pregnancy', 'salmon', 'pregnant', '##ester', 'affiliate', 'scan', 'baby', 'um', 'pee', 'trim']
31	['halloween', 'maker', 'fabric', 'print', 'zipper', 'graph', 'fabrics', 'bags', 'yarn', 'taller']
32	['glory', 'chemotherapy', 'cancer', 'highest', 'santa', 'sun', 'kings', 'christmas', '##cre', 'exercise']
33	['##ses', 'ob', 'glory', '##sive', 'dialogue', '##sing', 'highest', 'christmas', 'santa', 'spanish']
34	['folk', 'waving', 'mercy', 'chemotherapy', 'dogs', 'drowning', 'business', 'et', '##cre', 'folks']
35	['korean', 'japanese', 'folk', '##nam', 'bus', 'students', 'gang', 'korea', 'ob', '##lk']
36	['chemotherapy', '##cre', 'cancer', 'pan', '##atic', 'nausea', '##as', '##mo', '$', 'med']
37	['hair', 'chemotherapy', 'students', '##cre', 'classroom', 'cancer', '##atic', 'pan', 'colour', '##ses']
38	['students', 'classroom', 'school', 'teaching', 'year', 'teacher', 'kids', 'space', 'storage', 'timbers']
39	['hair', 'paralysis', 'extensions', 'afro', 'curly', 'colour', 'curl', 'clips', 'clip', 'sleep']
40	['paralysis', 'sleep', 'dreaming', 'luc', 'dream', 'experience', 'scream', '##id', 'lip', 'hallway']
41	['fucking', 'students', 'classroom', 'shit', 'brush', 'fuck', 'love', 'school', 'glow', 'kids']
42	['tickets', 'teller', 'pen', 'elton', 'students', 'paralysis', 'concert', 'classroom', 'ill', 'row']
43	['timbers', 'match', 'boating', 'fucking', 'paralysis', 'miller', 'fantastic', 'galaxy', 'lead', 'preview']
44	['fucking', 'dissertation', 'paralysis', '1940s', 'sort', 'space', 'architecture', 'studio', '##lage', 'col']

Continued on next page

Table B.2 – continued from previous page

45	['er', '##m', 'te', '##sco', 'cute', '##yst', 'vinci', 'jeans', 'paralysis', '##igan']
46	['patio', 'er', 'plant', '##m', 'storage', '##aks', 'kay', 'toy', 'fishing', 'pot']
47	['er', '##m', 'patio', 'wa', 'bath', 'plant', 'te', '##sco', 'uh', 'dissertation']
48	['patio', 'plant', 'pot', 'plants', 'gravel', 'roots', 'height', 'blooms', 'bottom', 'apartment']
49	['uh', 'fucking', 'fuck', 'shit', 'classroom', 'pain', 'scan', '##ination', '##ast', 'chemotherapy']
50	['dude', 'phd', '##iate', 'uh', '##um', 'stage', '##one', '##ani', 'op', 'motivation']
51	['dude', 'love', '##iate', '##um', '##one', '##ani', 'op', 'uh', 'patio', 'cr']
52	['uh', 'phd', 'stage', 'fucking', 'theater', 'relationship', 'love', 'california', 'moved', 'task']
53	['uh', '##ester', 'pieces', 'love', 'disability', 'coffee', 'historical', 'trim', '##fe', 'anxiety']
54	['love', '##mart', 'wal', 'husband', 'pain', 'pieces', 'water', 'tattoo', 'car', 'gonna']
55	['uh', 'pain', 'son', 'theater', 'disability', 'historical', 'tattoo', 'size', 'kristen', 'quarter']
56	['pieces', 'dude', '##ection', '##oll', 'tease', 'fucking', '##um', 'miscellaneous', '##iate', 'colour']
57	['phd', 'gonna', 'love', '##ester', 'sheets', 'pee', 'patio', 'thesis', 'motivation', 'trim']
58	['love', 'design', 'colored', '##ester', 'paint', 'drew', 'marker', '##ayo', 'toby', '2008']
59	['love', 'scan', 'anxiety', '##hora', 'panic', 'sep', 'baby', 'brushes', 'scans', 'pregnancy']
60	['nursing', 'clinical', 'peer', 'pressure', 'book', 'scan', 'dvd', 'math', 'anxiety', 'ste']
61	['love', 'peer', 'pressure', 'classroom', 'sheets', 'nursing', 'kids', 'moist', 'husband', '##min']
62	['uganda', 'army', 'resistance', 'lord', 'legislation', '##gold', 'fein', 'africa', 'scan', 'soldiers']
63	['uganda', 'army', 'nursing', 'tattoo', 'resistance', 'lord', 'peer', 'piercing', 'clinical', 'pressure']
64	['peer', 'pressure', 'nursing', '##hora', 'sep', 'clinical', 'pieces', '##ester', 'brushes', 'color']
65	['brush', 'apply', 'piccolo', 'hearth', 'shade', 'potion', '##stone', 'peer', 'game', '$']
66	['governing', '##oint', 'miracles', 'cease', 'prop', '##tower', 'evidence', 'claims', 'god', '##hes']
67	['governing', '##oint', 'classroom', 'miracles', 'tattoo', 'cease', 'piercing', '##tower', 'evidence', 'prop']
68	['peer', 'pressure', 'potion', 'potter', 'harry', 'classroom', 'allowed', '##min', 'tattoo', 'included']
69	['peer', 'piccolo', 'pressure', '##min', 'hearth', 'pepper', '##stone', '##nation', 'sharp', 'diego']
70	['peer', 'goodwill', 'velvet', 'tori', 'pressure', 'mom', 'gonna', 'gloss', 'lip', '##ging']
71	['piccolo', 'goodwill', 'scan', 'velvet', 'tori', 'flea', 'gloss', 'baby', 'potion', '##nation']
72	['potion', 'piccolo', 'potter', 'harry', 'peer', 'merlin', 'included', 'governing', 'wax', 'sparkle']
73	['piccolo', 'names', '##nation', '##ggio', '##uch', 'sharp', 'ultrasound', 'tune', '##ure', 'flute']
74	['flea', 'market', '##rone', 'gonna', 'brand', '$', 'winter', 'gym', 'warm', 'potion']
75	['hearth', '##stone', 'peer', 'flea', 'market', 'names', 'love', 'winter', '##rone', 'brand']
76	['vera', 'al', '##oe', 'juice', 'gonna', 'bacon', 'eating', 'weight', 'plug', '##cd']
77	['jerk', 'fishing', 'pointer', 'game', 'signing', 'apply', 'slender', '##aki', 'diver', 'mixer']

Continued on next page

Table B.2 – continued from previous page

78	['hearth', '##stone', 'jerk', 'videos', 'nike', 'content', 'cousins', 'cars', 'fishing', 'enjoy']
79	['debt', 'gun', 'hearth', '##stone', '##fish', 'sweat', 'beaver', 'towels', 'saskatchewan', 'troll']
80	['hearth', '##stone', 'versus', 'series', 'weaver', 'vera', 'juice', '##oe', 'content', 'al']
81	['lights', 'shannon', 'agency', 'chloe', 'umbrella', 'lips', 'sort', 'vera', 'eating', '##oe']
82	['cars', 'car', 'channel', '##not', '##ma', 'dealers', 'juice', 'vera', 'al', '##oe']
83	['gonna', 'dolls', 'howl', 'collect', 'gun', 'debt', 'buttons', 'happiness', 'er', 'fuse']
84	['jerk', 'spawn', 'tung', 'tests', 'pee', 'rig', '##carbon', '##sten', 'tt', '##oro']
85	['pee', 'tests', '##c', 'lights', 'shannon', 'tt', 'bipolar', 'agency', '##fish', 'background']
86	['collect', 'howl', 'buttons', 'festival', 'dolls', 'fuse', 'clean', 'tan', 'hood', '##rift']
87	['happiness', 'polish', 'nail', 'university', 'color', 'canterbury', 'nails', 'coat', 'coats', '##laze']
88	['jerk', 'spawn', 'tun', 'debt', 'polish', 'pointer', 'nail', 'whoa', 'lights', '##log']
89	['cars', 'gonna', 'debt', '##not', 'university', 'happiness', 'car', 'canterbury', 'hearth', '##ma']
90	['##press', '##ants', '##de', 'anti', 'children', 'feeling', 'gym', 'mum', 'illness', 'tests']
91	['cinnamon', 'nail', 'coach', 'integration', 'polish', 'anxiety', 'feeling', 'cars', 'felt', 'color']
92	['cinnamon', 'integration', 'coach', 'subway', 'aroma', '##therapy', '##fusion', 'research', 'breath', '##tam']
93	['paris', 'tests', 'pee', 'pot', '##ty', 'cars', 'tt', 'panic', 'spotting', '##vu']
94	['download', 'banjo', 'bray', 'song', 'music', 'week', 'black', '##hear', 'brian', 'canal']
95	['student', 'polish', 'nail', 'alternative', 'canterbury', 'coat', 'color', '##l', 'mike', 'bus']
96	['hinges', 'sony', 'cable', '##7', 'upside', 'screw', 'rotate', 'camera', 'rip', 'ribbon']
97	['hinges', 'sony', 'cable', 'upside', '##7', 'screw', 'rotate', 'screen', 'scan', 'cricket']
98	['paris', 'student', 'blah', 'alternative', '##how', 'whoa', 'mike', 'tan', '##aru', 'test']
99	['collect', 'howl', 'testosterone', 'sort', 'canterbury', 'buttons', 'download', 'canal', 'fuse', 'university']
100	['paris', 'canterbury', 'hinges', 'university', 'sony', '##mute', 'trainers', 'great', 'anxiety', '##how']
101	['paris', 'canterbury', 'stitch', '##oche', '##ants', 'anti', '##de', '##press', 'tree', 'fat']
102	['canterbury', 'cricket', 'train', 'sport', 'christchurch', 'gym', 'panic', '##mute', 'figures', 'university']
103	['student', 'alternative', 'gloss', '##eral', 'fusion', 'escape', 'mascara', 'jeans', 'lip', 'drugs']
104	['gloss', 'fusion', 'lip', 'appointment', 'stitch', '##oche', 'color', 'xl', 'bryant', 'friday']
105	['stitch', '##oche', 'yoga', 'single', 'tree', 'cricket', 'train', 'testosterone', 'pot', 'nine']
106	['anxiety', 'cricket', 'sport', 'football', 'train', 'love', 'videos', 'subway', 'vulnerability', 'player']
107	['workout', 'canterbury', 'energy', 'gloss', 'cricket', 'hinges', 'paris', 'sony', 'chain', 'marketing']

Continued on next page

Table B.2 – continued from previous page

108	['canterbury', 'cricket', 'university', 'christchurch', 'sport', '##mute', 'stitch', '##oche', 'testosterone', 'gym']
109	['canterbury', 'christchurch', 'university', '##mute', 'whoa', 'societies', 'com', 'gloss', 'sean', 'train']
110	['cricket', 'sport', '!', 'football', 'drink', 'dogg', 'jeans', 'chain', 'vera', 'canterbury']
111	['canterbury', 'university', 'christchurch', '##mute', 'clock', 'testosterone', 'weekly', 'row', 'ill', '!']
112	['cricket', 'gloss', 'testosterone', 'sport', 'fusion', 'pot', '##ty', 'football', 'lip', 'ill']
113	['cricket', 'vulnerability', 'sean', 'sport', 'stereotypes', 'pride', 'football', 'therapist', '##grad', 'assistants']
114	['sean', 'marketing', 'beats', 'jobs', '##grad', 'belief', 'assistants', 'tuition', 'graduate', 'testosterone']
115	['figures', 'workout', 'tan', 'cricket', 'bus', 'eating', 'figure', 'card', 'chewing', 'trade']
116	['context', 'train', 'math', 'ski', 'bus', 'excuses', 'belief', 'care', 'illness', 'subway']
117	['gag', 'pot', '##ty', 'spotting', 'testosterone', 'workout', 'trained', 'vomiting', 'jeans', 'nightmare']
118	['beats', 'canterbury', 'marketing', 'figures', 'christchurch', 'vegas', '##mute', 'money', 'stereotypes', 'university']
119	['whoa', 'belief', 'escape', 'vulnerability', 'gag', 'cricket', '##ging', 'destiny', 'newcastle', 'gym']
120	['cricket', 'figures', 'sport', 'gag', 'tests', 'football', 'panic', 'tt', 'fishing', 'mental']
121	['context', 'cricket', 'beats', 'math', 'marketing', 'dealt', 'anxiety', 'self', 'ended', 'night']
122	['whoa', 'train', 'testosterone', 'context', 'excuses', 'beats', 'math', 'marketing', 'newcastle', 'training']
123	['whoa', 'newcastle', 'figures', 'workout', 'cricket', 'tan', 'medical', 'gloss', 'fusion', 'wear']
124	['whoa', 'bus', 'quran', 'sean', 'mike', 'kids', 'positive', 'testosterone', 'jobs', '##grad']
125	['testosterone', '##l', 'canterbury', '##eral', 'mountain', 'add', 'bus', 'weights', 'alternative', 'student']
126	['whoa', 'gym', 'figures', 'clock', '##mist', '##ssi', 'pe', 'temple', 'sean', 'loose']
127	['whoa', 'newcastle', 'workout', 'gym', 'record', '108', '##0', 'mike', 'cam', 'recording']
128	['gag', 'whoa', 'vulnerability', 'stereotypes', 'testosterone', 'nightmare', 'vomiting', 'train', 'figures', 'ph']
129	['testosterone', 'tt', 'anxiety', 'pee', 'southern', 'vulnerability', 'tests', 'stereotypes', '##away', 'yoga']
130	['testosterone', 'record', '##0', 'vulnerability', 'pe', 'recording', '##mist', '108', '##ssi', 'constantly']
131	['whoa', 'filmmaking', 'cricket', 'panic', 'mike', 'context', 'production', 'quite', '##cd', 'northern']

Continued on next page

Table B.2 – continued from previous page

132	['testosterone', 'mike', 'vulnerability', 'ski', 'belief', 'workout', 'fishing', 'stereotypes', 'tomorrow', 'escape']
133	['whoa', 'ski', 'filmmaking', 'skiing', '##eem', 'northern', 'illness', 'weights', 'mountain', '##pon']
134	['belief', 'testosterone', 'fishing', 'destiny', 'escape', 'southern', 'expression', 'sort', 'walk', 'blog']
135	['mike', 'vulnerability', 'belief', 'stereotypes', 'michelle', 'facilities', '##rricular', 'basically', 'generation', '##cu']
136	['whoa', 'excuses', 'testosterone', 'fishing', 'training', 'mike', 'train', 'tan', 'gag', 'newcastle']
137	['filmmaking', 'production', 'northern', 'montana', 'subway', 'script', 'thanksgiving', 'cinematographer', 'otis', 'director']
138	['context', 'math', 'vulnerability', 'pseudo', '##away', 'wheels', 'whoa', 'otis', 'winning', 'tan']
139	['yoga', 'school', 'workout', 'college', 'sick', 'week', 'gag', '2', '##lvis', '!']
140	['weights', 'nora', 'climax', 'mountain', 'pack', 'hoping', '##£', 'context', 'esteem', 'disability']
141	['yoga', 'ski', 'filmmaking', 'skiing', '##eem', '##lvis', 'daughter', 'illness', 'loser', 'mental']
142	['quran', 'barbie', 'allah', 'subway', 'race', 'elisa', 'yoga', 'cheese', 'african', 'clips']
143	['subway', 'ski', 'tram', 'skiing', 'chewing', 'spitting', '##eem', 'eating', 'us', 'disorder']
144	['bus', 'subway', 'weights', 'context', 'mountain', 'southern', 'size', 'bracelet', 'social', 'chewing']
145	['otis', 'generation', '##£', 'constantly', 'pack', '##si', 'challenging', 'workout', 'rolls', 'age']
146	['otis', 'training', 'excuses', '##si', 'co', 'train', 'weights', 'challenging', 'pussy', 'trainer']
147	['chewing', 'filmmaking', 'spitting', 'vulnerability', 'nora', 'dealing', 'production', 'issues', 'testosterone', 'kuwait']
148	['subway', 'nora', 'ski', 'tram', 'energy', 'whoa', '##ied', 'bull', 'skiing', 'underground']
149	['african', 'race', 'women', 'black', 'american', 'loving', 'man', 'belief', 'subway', 'gonna']
150	['chewing', 'spitting', 'food', 'disorder', 'subway', 'eating', 'water', 'pack', 'hot', 'clicking']
151	['yoga', 'testosterone', '##lvis', 'bangs', 'blog', '##rien', 'nora', 'posts', 'phone', 'photography']
152	['vulnerability', 'mom', 'isabella', 'subway', 'bows', 'natalie', 'roof', 'mommy', 'cry', '##away']
153	['weights', 'subway', 'barbie', 'mountain', 'spirit', 'snow', 'website', 'article', 'tram', 'thanksgiving']
154	['##away', 'nails', 'tan', 'winning', 'mom', 'southern', 'nora', 'appointment', 'beauty', 'paper']
155	['generation', 'weights', 'suffer', 'subway', 'emotion', 'vulnerability', 'yoga', 'mountain', 'guys', 'panic']
156	['southern', 'yoga', 'weights', 'chewing', 'spitting', 'temple', '##z', 'south', 'exam', '##lvis']
157	['chewing', 'spitting', 'bangs', 'eating', 'generation', 'disorder', 'digest', 'food', '##ive', 'montana']
158	['chewing', 'panic', 'workout', 'spitting', 'disorder', 'customer', 'yoga', 'spirit', 'african', 'race']
159	['bangs', 'workout', 'temple', 'hair', 'mental', 'videos', 'subway', 'makeup', 'barbie', '##z']
160	['bangs', 'discussion', 'attendance', '9', '11', 'twitter', 'dialogue', 'religion', 'weights', 'greatly']
161	['weights', 'mountain', 'exam', 'workout', 'clemson', 'affects', 'privileged', 'context', 'social', 'tan']

Continued on next page

Table B.2 – continued from previous page

162	['bangs', 'age', 'dry', 'drinking', 'sort', 'bus', 'college', 'customers', 'alcohol', 'love']
163	['bangs', 'barbie', 'sorority', 'yoga', '##z', '##away', 'surgery', 'dollars', 'coffee', 'dry']
164	['generation', 'southern', 'exam', 'subway', 'north', 'suffer', 'fishing', 'racism', 'favored', 'husband']
165	['weights', 'mountain', 'vulnerability', 'brazil', 'age', 'uk', 'us', 'bangs', '##rda', 'feedback']
166	['bangs', 'race', 'african', 'women', 'pine', 'american', 'black', 'depot', '##lea', '##sco']
167	['bangs', 'generation', 'barbie', 'weights', 'sorority', 'medication', 'posts', 'race', 'suffer', 'sugar']
168	['bangs', 'snow', 'exam', 'thanksgiving', '##ius', 'montana', 'daughter', 'minus', 'grief', 'christmas']
169	['bangs', 'workout', 'temple', '2', 'accommodation', 'daughter', 'car', 'mental', 'health', 'damn']
170	['bangs', 'discussion', 'spirit', 'exam', 'surgery', 'ski', 'temple', 'vulnerability', 'funeral', 'sort']
171	['temple', 'birth', 'traumatic', 'cu', '##z', 'zoe', 'grief', 'yoga', 'husband', 'earth']
172	['barbie', 'spirit', 'cute', 'sorority', 'exam', 'coach', 'hollis', '$', 'sandals', 'mom']
173	['wounds', 'temple', 'exam', '##z', 'battle', 'size', 'weight', '##c', 'attacks', 'interview']
174	['traumatic', 'advisor', 'grief', 'journal', 'paper', 'spirit', 'weights', 'faith', 'graduation', '##bir']
175	['barbie', 'panic', 'sorority', 'exam', 'bangs', 'attacks', 'traumatic', 'student', 'article', 'internet']
176	['bangs', 'age', 'panic', 'dresses', 'yu', 'cooks', 'cook', 'countries', 'style', '##amps']
177	['te', '##sco', 'customer', 'social', 'february', 'march', 'privileged', 'police', 'vulnerability', 'sandwich']
178	['age', 'barbie', 'temple', 'sorority', 'discussion', 'country', 'countries', '##z', 'alcohol', 'african']
179	['barbie', 'sorority', 'sugar', 'discussion', 'exam', 'blind', 'frequency', 'son', 'felt', 'interview']
180	['age', 'article', 'interview', 'medication', 'blind', 'traumatic', 'sugar', 'frequency', 'stress', 'bind']
181	['weights', 'interview', 'exam', 'health', '##par', '##tum', 'channel', 'temple', 'issues', 'rather']
182	['customers', 'industry', 'est', 'clients', 'customer', 'dressed', '##hetic', 'temple', 'professionals', 'interview']
183	['frequency', 'blind', 'image', 'sensory', 'te', '##sco', 'substitution', 'police', 'bangs', 'panic']
184	['blind', 'frequency', 'sensory', 'image', 'substitution', 'mike', 'cute', 'sound', 'group', 'subway']
185	['cute', 'bangs', 'makeup', 'exam', 'hollis', '$', 'coach', 'sandals', 'blush', 'pts']
186	['sugar', 'host', 'interview', '##par', 'class', 'germany', '##tum', 'hotels', 'hotel', 'france']
187	['bony', '##cap', 'knee', 'gotta', 'customers', 'frequency', 'image', 'blind', 'montana', 'stream']
188	['posts', '69', 'customer', 'article', 'social', 'twitter', 'pit', 'exam', 'rather', 'favored']
189	['exam', 'interview', 'funeral', 'blah', 'uh', 'emotions', 'issues', 'leopard', 'host', 'bony']
190	['shout', '##tum', 'bind', '##par', 'makeup', 'lady', 'medication', '##ram', 'medicine', '##ig']
191	['##tum', '##par', 'makeup', 'article', 'generation', 'customer', 'attacks', 'temple', 'phd', 'shout']
192	['country', 'racism', 'softball', 'cha', 'vi', 'naked', 'teeth', '##rly', 'sugar', '##b']
193	['sugar', 'bind', 'sensitivity', 'racism', 'presentation', 'te', 'final', 'blessed', '##sco', 'favored']

Continued on next page

Table B.2 – continued from previous page

194	['sugar', 'faith', 'rather', 'age', 'group', 'sensitivity', '1st', 'interview', 'gum', 'ke']
195	['bind', 'yu', 'photography', 'tar', 'bangs', 'customers', '##amps', '##p', 'barbie', 'damn']
196	['sugar', 'attacks', 'customers', 'medication', 'class', 'sensitivity', 'sick', '##ig', 'rather', '##la']
197	['racism', '##pse', 'large', 'bony', 'necklace', '##cap', 'society', 'bind', 'phd', '##la']
198	['article', 'bind', 'twitter', 'door', 'pocket', 'pit', 'shirt', 'host', 'health', 'stream']
199	['interview', '##c', '##tum', '##par', '1st', 'police', 'stress', 'posts', 'teeth', 'collaboration']

C. Relevance Swarm Plots

For these swarm plots, first the amount of relevance per instance was normalized so that the sum of the absolute amount of relevance summed up to 1. For the D-Vlog swarm plot, samples where a concept had less than 5% of total absolute relevance were omitted, this was done to prevent cluttering.

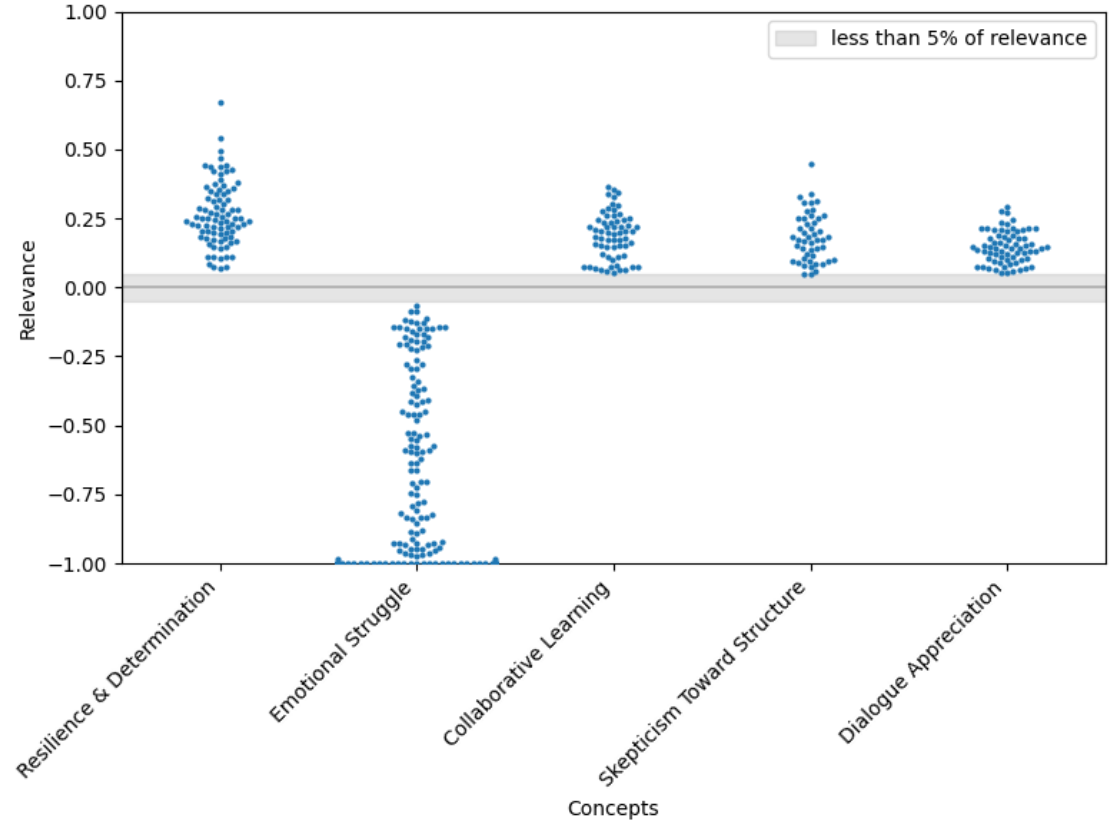

Figure C.1: Swarm plot of relevances in D-Vlog test set.

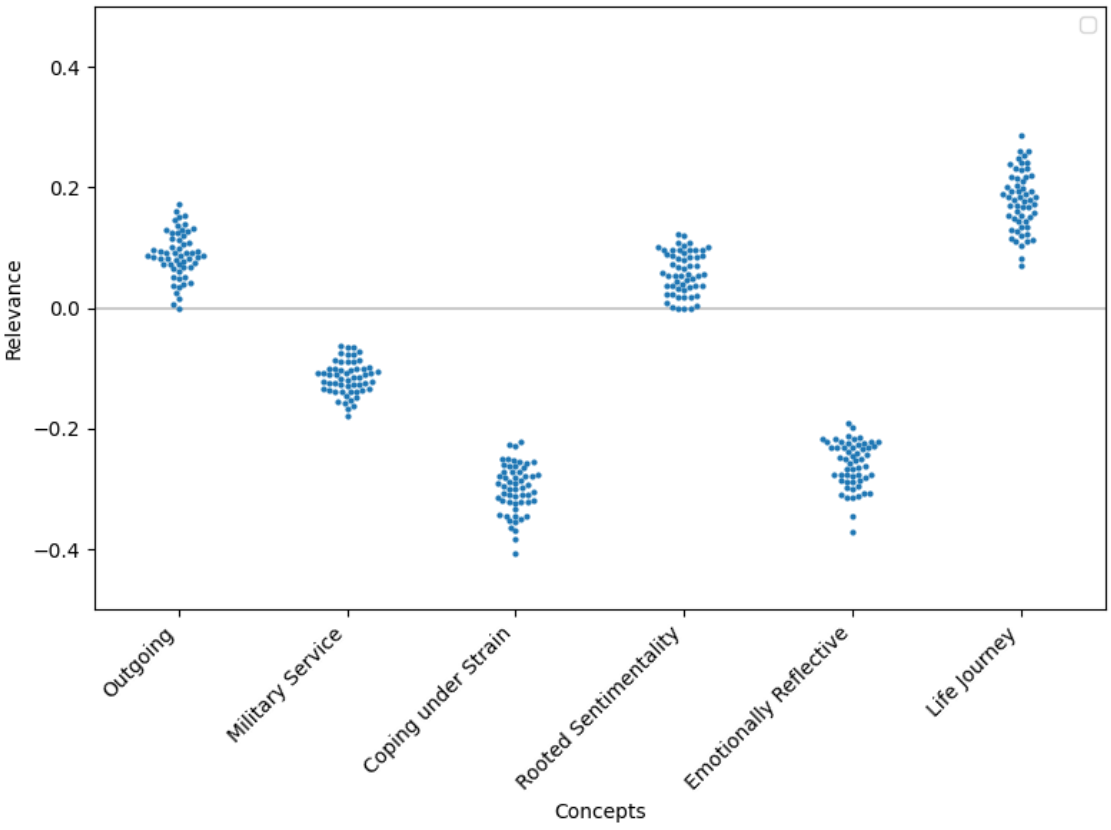

Figure C.2: Swarm plot of relevances in EDAIC test set.

D. Pieces of transcripts relevant for concepts discovered through CRP

The following is a collection of snippets that are indicative of each concept, each concept is represented by the 5 snippets with the highest absolute relevance for that concepts. This relevance is displayed below the participant ID. A positive relevance indicates that the snippet indicated non-depression, a negative relevance that the snippet indicated depression. Heat is the value of the heatmap at that specific sentence.

D.1. Pieces relevant for D-Vlog

D.1.1. "Emotional Struggle" concept

Table D.1: Most relevant samples for "Emotional Struggle" concept.

ID	sentence	heat
410	and I just feel like crying and I'm like heightened up in my anxiety and for no reason and like the things I'm feeling, I feel like nobody would feel unless it's a withdrawal.	-240.171
-81.339	It's specific to drug withdrawal because I don't think your brain chemistry can do that by itself and it really fucking sucks.	-154.434
	Sorry about using bad language but it really, really sucks and even if I'm going like following the doctor's orders the withdrawal is just horrible	-132.562
	and I don't really know how I'm surviving.	-146.256
	I'm just really tired and really, there's so much to do.	-144.627
403	In November, I considered killing myself, and since then I think I've come a long way.	-167.211
-73.081	I don't think I've got better in the slightest, but I think my mentality towards suicide and what I was considering has changed.	-159.979
	I self-harm, but I'm trying to overcome that.	-143.903
	And I'm currently on certainly 50mg antidepressants.	-123.027
	And I was originally on Citalopram at 10mg, but I just didn't do anything for me, so I've really been on this for a couple of weeks.	-136.159

Continued on next page

ID	sentence	heat
170	Like, that way is specific to when I'm really hurting and really, like, it's painful.	-174.615
-72.624	Something is really hurting me, and I don't know what it was.	-151.109
	And it was, like, very emotional cry, and it happened twice in a row and two nights, and I was missing someone dearly.	-119.489
	And possibly the fact that I was thinking about, like, basically, it's like, I'm dead to that person, and just sucks, and I don't know, it's very, I don't know, I just got really emotional, and it was really, really hard to get past that, because, like I said, it hurt so much, and it wasn't a physical pain.	-163.006
	It was, like, completely mentally overwhelming, and I didn't know what to do.	-117.979
20	I just felt really, really more than depressed, it was almost like despair and I don't know if you've ever felt such despair where you just don't want to be here.	-200.009
-71.878	And so at age 15 I attempted suicide and was put on psychiatry for three months, they wouldn't let me out because they didn't trust me, didn't think that I would not try again.	-141.315
	So the depression got really, really bad, so bad that I just didn't want to be here anymore.	-138.103
	And that is something that I wouldn't wish on my worst enemy that feeling and it's very hard to describe if you've never felt it.	-132.653
	Now most people have about highs and lows that are normal, that are normal going through stressful events in life or things happening as there's loss, there's divorce, death and these are all natural.	-107.048
498	But I have suppressed so many emotions that I cannot continue to run from everything that I'm afraid of.	-218.554
-71.042	Because I am depressed and I have been diagnosed for four months now.	-118.648
	And I am falling apart.	-105.533
	I have to face my fear of abandonment, being alone.	-139.368
	My face of being, I have to face the fear of being hurt again.	-126.895

D.1.2. "Resilience & Determination" concept

Table D.2: Most relevant samples for "Resilience & Determination" concept.

ID	sentence	heat
750	no no go area, another important factor is that the admission percentage that has been allocated to for an incident is just so small, it's just so so little, like less enough for example in area of medicine, a university has just 15% for foreigners, whereas the other 85% are just for Germans, and this I think is also one of the main factors why some students like some friends today don't get to study their dream courses, you waste all these years for you to study what you really want, and then at the end of the day, you're not able to study it because this proprietary course was hard, and let me tell you guys, this proprietary course is no tight split, like the bombardie with so many stuff, like sometimes it's just like you find yourself starting on a daily basis, I know that I was only sleeping in the library in this virtual course, and then at the end of the day, you just be marked down and your scores would not be fine, I think it's just not fair how everything is being rated, like for example now, in order for you to study medicine here, your grades have to be very very high, and the Germans in their second year school, like they do so many extracurricular courses like sports and whatnot, so let's just say that this person now is very good in sport and made a very high mark in sport, this mark is going to be counted and what would it mean if he's going to study medicine, whereas there's this foreigner here who didn't do sports, like they do the main thing, like the part subjects, but because he didn't get like a very high score like this German in sport, he would not be taken, so at the end of the day, you just see yourself so frustrated, I applied to 15 universities guys, 15, I'm not even one, and all the people I did was very due course with, it was just only two people who ended up studying medicine, who we are like roughly, roughly 50, and only two people ended up studying medicine, mind you, 98% of us wanted to study medicine, and the thing is that they'll keep on telling you,	101.018

Continued on next page

ID	sentence	heat
52.133	yeah you're going to do it, you're going to make it, you can't do it, they won't even tell you the truth, like in deep preparation course, I could remember all of us, it was just like, yeah, you're better, stop shaffing, it's you're probably going to do it, just read hard, but it's out there in the real world, it's not like that, you just get frustrated and your dreams shattered, all my friends from this repeat course, they just ended up like me, doing one stupid course, they never intended in their life to do, and my own case, as I said before, I had nothing else in mind, like nothing else came to me, I was like it's either medicine or nothing, and that is very very bad guys, we should always have like a plan B too, like always fall back on if something doesn't happen, I just ended up studying this narcissist course that I'm not even enjoying, my main message in this video for students who are aspiring to study in general are, number one, you have to know that you just don't come and start immediately, that was what I told, I thought I would just come here and start in here, but it didn't happen, I came here	82.739
	and I had to do another nine months of doge course, that is approximately one year, when you count the holiday, then you do this one year, that is two years, after three years of waiting at home, doing the doge course at home and waiting for your visa, and then you come here, you're not even studying what you want to study, so I would strongly advise you to make your inquiries very very well, and also the language, the language, the language guys, if you ever ever have the opportunity to study in a country that speaks the same language as you, please, please, please, please, please, take this offer, do not comment, stress yourself, and kill yourself, and just get frustrated for nothing, okay, that is what I have to say, and thirdly, I'm still not the first point, I don't know how I jumped to this second point, so regarding the time waste, just know that you have to invest time into all this process before you even start your main course, if you're lucky enough to have it, and the most frustrating part of it is that all my classmates in secondary school, they are all done with their bachelor degree, like most of them are even working, I could remember that I call my mother crying, I'm like, what am I doing here, why did I even choose this in the first place,	96.514

Continued on next page

ID	sentence	heat
	oh I regret doing this, if I was adjusted back in Nigeria, gone to university there, and I would have been doing it by now, you know, and um, six years later, I'm just starting out, it's, it's not so fair, for me, it's just a time, a very, very, very big waste of time, but then my mommy has sweet nature, always tell me, people at the fence, like destiny is at the fence, it is not how far, but how well, do not get me wrong guys, I am really, really grateful for this opportunity, I am grateful to go and I am grateful to my family, to my parents, for, for, for making this happen, and then sometimes when I want to quit, I know that so many things have been sacrificed for me to be here and whatnot, and then the third factor to consider is that the number of admissions given to foreign students is very limited, like it's small, like this little title of this video is that I regret studying in Germany, but do not get me wrong guys, I am grateful for this opportunity, like I said before, studying in Germany has told me more good than harm, I just feel that there are certain sets back, like the time and everything that I said, if I got to be in Nigeria, I would have been doing my school right now, and all that, so please, always try to make your inquiries very, very well, do you know that sometimes, really people ask me for how they could come study in Germany, maybe for themselves, or they are sibling, or cousin, I don't know, or relation, when I tell them everything it takes, like everything that I did up until this moment, they're like, wow, I'm not going to do that, that is too much, certain from the money you have to have and every other thing, and I know that this is not compared to the tuition fees that is being paid in other countries, it is not at all, so I am so grateful for this opportunity, but then they are also some disadvantages to it and all that, so make sure you know what you're getting yourself into, because really guys, like I said, this wasn't something we sat down and planned very well, so just this spontaneous something and and I'm here three years after, I didn't know it was going to take this long before I finally start up a course in the university, if I knew it was going to take this long, I probably wouldn't have studied here, yes I would say that now, but then this is not to scare anyone who is aspiring to study in Germany, Germany is a very, very nice country, their educational system is top not, like the way they teach you to think for yourself and solve problems, not being able to like just reproduce everything you thought, it's just really nice, so that is it guys, I just felt like sharing with you all my experience,	82.09

Continued on next page

ID	sentence	heat
	so just let you know that it is not as easy as it seems, or you think there are we students studying abroad, we just go to McDonald's and KFC, it's now, it's chicken and chips with ice cream, it is not like that, so please, please make your inquiries and know everything you're supposed to know before choosing a study course, thank you guys for watching	
	and I will see you in my next video, please don't forget to subscribe, like, comment, are you also a foreign student, what are your experiences, what are your setbacks, what are your advisors, to other students who are aspiring to study abroad, also do not forget to tell me a notification bill to be notified every time I post a video, I would definitely be doing an updated version of this video, like I said, there are so many areas of this topic that I can just fit into one video, also do not forget to check out my other videos,	85.178
267	Nice.	64.787
45.686	A really good sound.	75.794
	Nice, I have a Jerusalem's Rose.	67.854
	Now, the reason why it's both a Jerusalem's Rose is because it has the Jerusalem cross and the Rose on it.	89.949
	And they use these in the crusades I'm assuming side arms because I'm assuming they didn't use swords the length of a small child.	93.81

Continued on next page

ID	sentence	heat
559	yeah, this session was, it was pretty good, it was in a terminal auditorium and it was, it wasn't really a dialogue but it was more of a big lecture about dialogue, not, we didn't have a dialogue but like we talked about like different types of dialogue and like how people perceived dialogue and the first video we watched like two videos which was a big part of the lecture and one video was about the speaker who talked about how stereotypes don't tell the whole story, the stereotypes aren't necessarily not true but how they don't tell the whole story and that happens a lot in dialogue where people just don't, they don't know the whole story and they don't get to know people and they just make assumptions and like it's just assuming that like all people from Africa are like poor and boring and like at war and stuff like that or just, it's just making like blatant like generalizations which is really like disrespectful and it's not something that should be done and she talked about how you really have to sit down and like, it was like people should strive to learn about other cultures so it can be more understanding and not seem as ignorant and things like that.	85.032
44.998	I thought the video was good, it was pretty interesting as you told her story about going up in Africa and how we were made like just assume like things and that like weren't true at all and because they aren't true and like a lot of people in Africa aren't like they want to strive to be better and it's not just all what's portrayed in the news	85.54
	and she said the media has a lot to do with that too because like since we live in America we just see the viewpoint from American standpoints and whatever the news tells us that's what we're going to see unless we actually go to these cultures and see what it's actually like and try and understand.	69.413
	The second video we watched was it talked a little bit about evolution and I thought that video was really cool	69.241
	because yeah the guy who was talking I guess that he was just drawing the whole time during it which I mean the drawings were really good and that kept me like interested and he was more of like a biological type guy so we talked a little bit about the science	77.064

Continued on next page

ID	sentence	heat
898	We started off by playing a game to get to know one another and introduce ourselves, and the leader, the person who's leading the dialogue, got up and was saying his name and his major where he came from and something about him, and then the next person would get up that had something in common with him, and so forth until everyone had done it, and I guess it lightened the mood, and some made everyone more comfortable and whatnot to dialogue, and then they passed out a worksheet about dialogue, and it was to compare debating discussion and dialogue and see the differences and all of them and whatnot.	83.854
44.862	And basically, we went over that debate is when two people are going back and forth, or multiple people are going back and forth, and to argue their side or whatever, and then counter arguing, and then in debate winning is the goal, and debate implies a conclusion, and then there's discussion, and there is an equal playing field in discussion, and in discussion, the more perspectives voice the better, and it can be an open or close-ended, and then we got to dialogue, and it was more about listening to one another's sides in order to understand each other and our own points of view and whatnot, and dialogue remains open-ended, and in dialogue finding common ground is the goal.	85.928
	So that's where we went over first to get a, just so we could understand what dialogue was where we started, and then we went over some guidelines about our new student dialogue, pretty much	80.356
	so we wouldn't offend anyone, or, you know, we just had to acknowledge that, you know, racism, sexism, and all that exists, and it may surface from time to time, and that we have to trust people are always doing the best they can in the dialogue, so that we can learn from them and be productive and whatnot, and it was saying that we had to listen to one another, and that everyone had an equal and valid voice in the dialogue, and that we had to challenge the person's ideas instead of the actual persons that we wouldn't offend people, obviously, and stuff like that.	63.071

Continued on next page

ID	sentence	heat
	And then the leader handed out another piece of paper called the social identity wheel, like that, and we had to go over a few different topics, like our race, ethnicity, our socio-economic class, our gender, sexual orientation, age, national origin, first language, and physical, emotional developmental ability, and our religion, or spiritual affiliation, and we had to fill out each of those around the wheel, and then inside the wheel we had to order them in order of what we think about the most, I guess, on a daily basis, which I don't really think of many of these things, daily basis, but that's all right, maybe some people do.	71.914
896	You know I'm doing something for this new student dialogue anyway, I just kind of walked in I've wait listed	76.066
43.539	I was about to not graduate my freshman year because of this new student dialogue, which was actually pointless So well done again cuz that's why we're having to some people Clums and got a lot of it and	80.173
	mystery right you only have two and a half more minutes only two and a half more minutes	76.851
	So let's see what did we talk about Those pubes on your face, I know I'm looking at this	71.044
	and I'm like maybe I should shave soon I'm probably not gonna happen though.	69.627

D.1.3. "Collaborative Learning" concept

Table D.3: Most relevant samples for "Collaborative Learning" concept.

ID	sentence	heat
598	and we had a group project where we had slips of paper of different types of people like male female Muslim Christian atheist and different things like that like poor middle-class hot glass	81.581
43.859	and then we had to figure out where we thought that the people would fit and of what is who are more privileged than the others and so that kind of got us to got us talking and got a good dialogue going within the group	82.841

Continued on next page

ID	sentence	heat
	and so it was very good it was a good experience like I feel like the whole new student dialogue for Clemson	88.102
	and it's a part of library 100 because it does help people get connected and help people be able to learn what kind of norms are good for dialogue and how people can take those out into Clemson with and talk to people who have different views on things and use those norms so that they won't bring judgment on those people and that they will make them feel comfortable and that they can just talk to people without getting in arguments and also I believe that like the the dialogue and just library 100 in general is good for like just to meet people and to learn how to accept people better	84.44
	and I think it is a very useful tool for the university	88.737
595	There were more of a reserved voices in our group as like with any group, there were definitely people who stood out and had something to say basically about every topic that we chose, which is nothing wrong with that, that's great, but there was definitely people who was more reserved, who may have only spoken once, just to maybe say they spoke, but that's just where all groups, so that's expected.	70.41
41.028	Some people just, you just can't get everyone to feel comfortable in a two hour period time, so there was definitely reserved voices in our dialogue, but I don't make that a hindered dialogue experience at all because the majority of people were active voices and I would like to think that I play the active role in my group because everybody knows that in groups you always have those awkward moments where everybody, when a facilitator acts a question, everyone kind of just looks at the next person, they're not really sure what to say or should they share or should they shouldn't, so I felt like those are like the awkward moments in dialogue and I felt like that's when I kind of came in as a group member to kind of like help relieve those awkward moments by just sharing my opinion and I felt like by doing that I helped other people feel open to share their opinion as well and I feel like other people helped me as well when I seen them open to share	72.014

Continued on next page

ID	sentence	heat
	and I like to think like our whole group was very open minded, everyone might not have agreed with each other's opinion, which is completely normal, but I feel like everyone was respectful and they were open minded to at least hear what the person had to say whether they agreed with or not and I felt like that was just a really good environment to be in, especially when you're with new people and you're trying to basically get a group of like 20 people to dialogue	85.861
	so I mean I felt like it was pretty good, I was kind of nervous about my experience, well I was kind of nervous about attending a new student dialogue but once I got there we did some of the iceberg names, it just felt like, it just kind of felt like no big deal with that, I kind of knew the people before and I still think it was great, they were very honest with us, they helped keep us on track of things, they definitely kept an interest and so we weren't bored for like two hours, so that was really good, so I think mom was on a Monday night, so I felt like overall my experience or dialogue in was pretty well, it went pretty well and at the end of my dialogue experience at first I was unsure like why would Clemson or more specifically library 100, once you do a dialogue before now I just thought like it was weird, I didn't see the need for it, but actually when I sat in, when I sat down after the dialogue and I actually reflect me, the dialogue is completely necessary for library 100, most people take a library 100, our freshmen and a lot of us come from high schools that we attended, that we attended to all four years and we probably just moved on from the same middle school, so we knew we basically all the same people, so and if we didn't know all the same people, we really did get diversity like we get on this campus, so I felt like that is like why library 100 wants to make sure that you do a dialogue component because if at this university where we have so many different people, so many different ideas, so many different beliefs, I feel like it's very important to know how to dialogue and to have effective communication skills with people who not only believe in the same thing you do or look like you do, but all people on the campus and I feel like that was very important	83.283

Continued on next page

ID	sentence	heat
	and I do see dialogue as a useful tool in our university space because I've definitely used the tips that I've got from this dialogue in my communications class that I'm taking	86.659
623	We also have a special course here, it's called like rats, we call it rats lab, veterinary technician students here, call it rats lab, but it stands for like intro to research animal techniques, you know, you study how to, what it means to be a laboratory animal, like technologists, or technician.	63.14
34.084	And there are different routes you have to understand to being, for veterinary technicians, you don't have to just stay in a clinic field or whatnot.	48.584
	You can branch out to others, and the thing about the school is they really do show you those options, they really do push you to being, trying more fields than just stick with being in a clinic, and that's not for everyone.	70.21
	So, second semester, I guess second semester, it's a bit more hectic in terms of labs that we have.	75.773
	Since there was overcrowding of students last semester, I'm taking like small animal care and larger animal care together, and the way they do at this school was that you're supposed to take one or the other like each semester, so that you don't get as crowded as much as possible.	73.115
225	And I know all of this because I studied Latin when I was in high school for an entire year I took Latin and that took the National Latin Exam and aced that mofo.	73.413
33.734	Nobody thought I could do it.	52.887
	Actually almost didn't want me to take the class because I had low English grades	66.315
	so they figured that my foreign language skills would be even worse but it turns out that I'm actually better with Latin than I am with English.	71.343
	At least on the level that I was being taught of Latin	63.467
758	Yes, there were about five guys in the group, the majority of people in our group were guys.	58.498
31.376	There were about five guys in the group that spoke up every time one of our instructors asked a question.	62.6

Continued on next page

ID	sentence	heat
	They kind of led the group, our instructor said afterwards, I stated a little bit afterwards, and he said it was the hardest dialogue for people that he'd ever done because he said it was pointy to get people to talk.	65.588
	The most reserved voices in our dialogue were the 30 other people who weren't really talking.	55.566
	Again, I don't think it wasn't the instructor's fault.	62.285

D.1.4. "Dialogue Appreciation" concept

Table D.4: Most relevant samples for "Dialogue Appreciation" concept.

ID	sentence	heat
595	and I like to think like our whole group was very open minded, everyone might not have agreed with each other's opinion, which is completely normal, but I feel like everyone was respectful and they were open minded to at least hear what the person had to say whether they agreed with or not and I felt like that was just a really good environment to be in, especially when you're with new people and you're trying to basically get a group of like 20 people to dialogue	52.722

Continued on next page

ID	sentence	heat
33.038	so I mean I felt like it was pretty good, I was kind of nervous about my experience, well I was kind of nervous about attending a new student dialogue but once I got there we did some of the iceberg names, it just felt like, it just kind of felt like no big deal with that, I kind of knew the people before and I still think it was great, they were very honest with us, they helped keep us on track of things, they definitely kept an interest and so we weren't bored for like two hours, so that was really good, so I think mom was on a Monday night, so I felt like overall my experience or dialogue in was pretty well, it went pretty well and at the end of my dialogue experience at first I was unsure like why would Clemson or more specifically library 100, once you do a dialogue before now I just thought like it was weird, I didn't see the need for it, but actually when I sat in, when I sat down after the dialogue and I actually reflect me, the dialogue is completely necessary for library 100, most people take a library 100, our freshmen and a lot of us come from high schools that we attended, that we attended to all four years and we probably just moved on from the same middle school, so we knew we basically all the same people, so and if we didn't know all the same people, we really did get diversity like we get on this campus, so I felt like that is like why library 100 wants to make sure that you do a dialogue component because if at this university where we have so many different people, so many different ideas, so many different beliefs, I feel like it's very important to know how to dialogue and to have effective communication skills with people who not only believe in the same thing you do or look like you do, but all people on the campus and I feel like that was very important	74.308
	and I do see dialogue as a useful tool in our university space because I've definitely used the tips that I've got from this dialogue in my communications class that I'm taking	53.854

Continued on next page

ID	sentence	heat
	and I've definitely seen results in it, just by being the dialogue just made me an overall better communicator and when you can communicate better with people on the university's campus, then you're networking which is very awesome because you never know whose parents and who knows who you will need in the future, so I feel like it's extremely useful to any university experience and I just hope that other students from different universities get to experience the same experience that we experience with dialogue because I feel like it is something that you just use more than you think you use and when you actually get tips on how to use it effectively like you can do so much with it and go see your advantage, so that's my um summary and my blog on my new student dialogue I hope you enjoy and that's it!	57.2
958	So if you're looking to apply, or you're just interested in like what's it like being an art student in college, if you have any questions, leave a comment, and I will hopefully be able to answer them.	33.395
23.29	Word of caution, I'm not a professional teller about her of the school, so just so you know, I'll be really unprofessional and biased about what I say.	44.053
	Also, a little about me, I am a transfer student, so I'm going to have no general education other than art history to talk about, so if you're looking into English or math, and what's it like, or is it hard or easy, I won't be able to answer your questions, but if you have questions about art classes, fundamentals, and major classes, I'll let you know right now I'm going to be taking fundamental classes, and hopefully eventually sequential art classes.	33.7
	I'm also a double major, so eventually I'll be taking painting classes also.	26.72
	I'm living in the dorms, just for this year, hopefully I'll be able to find some friends to move into like an apartment next year.	29.968
794	Hi, I'm Chris Darr, I'm a freshman here at Clemson and this is my video blog for the student dialogue portion of library 100.	33.786
22.789	My specific topic for the student dialogue was race and how it impacts and affects our society.	41.979
	We discuss discrimination, stereotypes, several misconceptions that are had about race and its role.	25.471

Continued on next page

ID	sentence	heat
	We even discussed the role race has on college, college acceptance, college life.	28.768
	Just kind of make a connection between college students and the topic.	34.223
758	Yes, there were about five guys in the group, the majority of people in our group were guys.	28.449
20.369	There were about five guys in the group that spoke up every time one of our instructors asked a question.	35.537
	They kind of led the group, our instructor said afterwards, I stated a little bit afterwards, and he said it was the hardest dialogue for people that he'd ever done because he said it was pointy to get people to talk.	29.604
	The most reserved voices in our dialogue were the 30 other people who weren't really talking.	26.357
	Again, I don't think it wasn't the instructor's fault.	26.837
821	Hello, and welcome to my video log on my new student dialogue.	30.264
20.145	My new student dialogue was on socio-economic status, and I thought it was pretty interesting, not, and it wasn't too boring, as I thought it could have been, but there are many active voices in my dialogue, a lot of people who like to talk, like to have their voice heard, but what I found is that some of the more reserved voices were actually where more of the important points came from the people who really thought about what they were going to say, didn't over-top, and just allowed their thoughts to be heard and thought on by the group.	37.179
	So, well, I thought my group at the end, by the end of the time, did achieve a true dialogue, it, we, we came talking more freely after we got more comfortable with each other, and we actually had a pretty solid conversation on socio-economic status that I thought was not going to happen, just judging by the first five or ten minutes of our group.	25.205
	The active voice has really dominated the first part of the discussion, there's just a couple of like two or three kids just talking to, talking at each other, just trying to out-talk one another, but then as it got on, people got more comfortable talking.	24.83

Continued on next page

ID	sentence	heat
	The quieter kids started making good points, and they really enhanced the discussion with their thoughts.	27.695

D.1.5. "Skepticism Toward Structure" concept

Table D.5: Most relevant samples for "Skepticism Toward Structure" concept.

ID	sentence	heat
794	And it was I guess successful.	72.051
39.815	But I don't really see the purpose of it being required for incoming freshmen and trying to transfer students.	104.957
	I mean, people get together and talk about topics on their own.	69.849
	So forcing them to do outside a class.	62.072
	I didn't really see why it was necessary component.	82.922
935	I've also created um colors.	82.065
39.784	So like Color groups so that it would be like green red blue Um, there's like five or six different colors So that you don't have to have tinkertown, you know reading rainbows	80.988
	Um, I kind of customize	78.76
	that for my classroom But if you do colors or even numbers there's	76.509
	numbers in the packet as well So like number group one group two group three Then you can use this same method give them the choices here These cards are sized for the target clear Polypockets those um clear adhesive packets, but you can easily laminate and just either use a magnet If you've had a refrigerator or a magnetic board or you can just put these on a wall somewhere and as long as you're laminated So, um, there's this is pretty versatile.	73.225
854	It's a little more personal in terms of your classes are smaller.	69.981
37.655	Now your second year students, like there are not that many that returned professors are able to focus on you more because your classes are smaller.	89.077
	I think the most we had in one class was maybe 50, if even 50, maybe even less than 50.	73.403

Continued on next page

ID	sentence	heat
	So having the professors be able to focus on you because it's a smaller group.	64.903
	It's kind of cool because you get to have your questions answered more, you don't get lost in a big crowd of students.	73.227
566	I was very like I don't care who you work with maybe I need to like write that in the instructions versus just saying it maybe some of the students need to see that in case they weren't clear	59.333
30.633	so definitely something for me to consider for next school year your great teacher it's us the ones that need to improve this one goes back to the early comment that I made	69.133
	so	56.875
	yeah again it goes back to yes students need to take accountability for their actions but as a teacher sitting now with them one-on-one and helping them see whether or not it's a goal that I want to achieve for next school year	58.31
	and I think I can help them learn how to improve and help them with that skill because that's going to be a skill that they'll need beyond my class when they have a job or a relationship you know self-reflecting on you know their behaviors I feel like when we got to work at our own pace I can get more done faster I don't like the new one through four grading system I like the stations	57.83
776	I always loved her blush and I just kept examining it	64.769
30.186	and then I was thinking of how much I love the blonde brunette redhead you know MSF's the way they go from the darker to the lighter	56.17
	and I would use that angled brush and I would do the angled brush and then I would put it directly like that on my cheeks	68.248
	so then I was getting the dark all the great you know going all the way up to the highlight	50.386
	so that's what I've been doing different as I've been just using my bronzer to contour	57.506

D.2. Pieces relevant for EDAIC

The following section of the Appendix is removed in compliance with the EULA of the EDAIC dataset, which disallows the publication, full or in part, of the dataset. A confidential copy containing the appendices in question is shared with the thesis supervisors for grading purposes.