6-3-2022 11:00 AM

Psychological Understanding of Textual journals using Natural Language Processing approaches

Amirmohammad Kazemeinizadeh, *The University of Western Ontario*

Supervisor: Robert E. Mercer, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Computer Science

Recommended Citation

Kazemeinizadeh, Amirmohammad, "Psychological Understanding of Textual journals using Natural Language Processing approaches" (2022). *Electronic Thesis and Dissertation Repository*. 8854.
https://ir.lib.uwo.ca/etd/8854

Abstract

Recent NLP advancements have improved the state-of-the-art in well-known datasets and are appealing more attention day by day. However, as the models become more complicated, the ability to provide interpretable and understandable results is becoming harder so the trade-off between accuracy and interpretability is a concern that is yet to be addressed. In this project, the aim is to utilize state-of-the-art NLP models to provide meaningful insight from psychological real-world documents that contain complex structures. The project involves two main chapters each including a different dataset. The first chapter is related to binary classification on a personality detection dataset, while the second one is about sentiment analysis and Topic Modeling of sleep-related reports.

Keywords: personality detection, sleep health, topic modeling, text classification, sentiment analysis, computational psychology, interpretable artificial intelligence

Summary for Lay Audience

Recent advancements in artificial intelligence models that accept textual inputs are becoming more and more accurate. However, because of the differences between the nature of the artificial intelligence models and human functioning, understanding the AI outputs are becoming harder for humans. In this project, the aim is to utilize top AI models in the field of natural language processing to provide meaningful insight from psychological real-world documents that contain complex structures. The project involves two main chapters each including a different dataset. The first chapter is related to binary classification on a personality detection dataset, while the second one is about sentiment analysis and Topic Modeling of sleep-related reports.

Acknowledgements

At first, I would like to thank my thesis supervisor, Prof. Robert E. Mercer who supported me during this program. He was more than a mere supervisor to me and had my back during the challenges I faced during this year and a half, and his patience and guidance helped me to thrive in my academic and career skills. Next, I want to show appreciation to Prof. Erik Cambria who co-supervised me for the Personality Detection chapter of this thesis and Dr. Abhishek Pratap who co-supervised me in the Sleep Health project of this thesis. Also, I express my gratitude to my family for getting me to a point where my ego can lead the way from then on. Lastly, I sincerely appreciate my friends for their unforgettable helps in the hardships.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent advances in Artificial Intelligence (AI) have brought us to a point that feels that all the problems can be solved by the hands of AI. In this thesis, this claim is scrutinized. From one side, models such as GPT-3 which generates full articles [94] by just getting the topic and basic information try to convince us that the era of human centered tasks are reaching an end. But from the other side, the quote of Noam Chomsky comes into mind: “you do not get discoveries in the sciences by taking huge amounts of data, throwing them into a computer and doing statistical analysis of them: that’s not the way you understand things, you have to have theoretical insights.” [53].

Some tasks especially in the field of computer vision are straightforward and deterministic such as detecting and reading vehicle licence plates [62]. However, when it comes to Natural Language Processing, especially in the pragmatic layer of sentiment analysis [22], even different persons may have different ideas and judgements about a text after reading a piece of it. For example, suppose we are given this hypothetical tweet "I just dropped my iPhone and now it’s not working. Wonderful!". Is this person happy because he found an excuse to borrow money from his father to buy a new phone? Is he angry because he lost what he owned? Is he sarcastic? All of these interpretations could be correct. The complexity and the importance of understanding and interpreting the free text caused researchers to introduce and develop a domain named Natural Language Understanding [4]. Recent Machine Learning (ML) models, which incorporate Deep Neural Networks, try to understand the meaning of the text while also considering the context of the input. The performance of these models is usually evaluated on general datasets as a baseline for comparison. A well-known baseline for evaluation is the General Language Understanding Evaluation (GLUE) and the top performing models are known as state-of-the-art.

One further step to achieve in this field is to analyze the mental state of individuals using text. To connect these two, some pieces of text that try to convey the individual thought and mental health conditions are required. One of the most common approaches to providing this information is to utilize self-reported textual diaries or journals of each individual to learn about their introspection [10]. Considering state-of-the-art models which have achieved the best results on GLUE, the question becomes can we use these models to achieve similar results on mental health datasets? It is one of the main questions that comes to mind and the following two projects are going to find an answer for it. In order to do that, the first project utilizes a supervised learning method for the labeled dataset and the next project uses unsupervised learning methods on an unlabeled dataset thus covering both aspects of ML on the mental health area.

Following Chapter 2, which analyzes the previous endeavors, provides the required background about the project, and outlines the challenges for this thesis, the next two chapters focus on two projects concerning interpreting people's journals. The first one tries to classify individuals based on their personality traits. It utilizes one of the well-known personality trait detection datasets, known as Essays [89]. The Essays dataset contains 2468 anonymous self-reported journals, each from a different individual, labelled by the participant's personality traits [77]. Personality traits are defined as the set of relatively stable characteristics which describe our feelings and behaviour [71]. In this project, we use the Big-Five [48] which is more reliable than other tests for evaluating people's personality [93]. The Big-Five personality test describes people's personality traits in five independent metrics: openness, conscientiousness, extraversion, agreeableness, and neuroticism. To assess the participants' personality traits, we try to utilize psychology knowledge to evaluate available text featurization methods in the field of Natural Language Processing to suggest which ones perceive psychological concepts. We also introduce an approach to embed psychological insight into general text featurization methods. For further details regarding the project and the dataset, please refer to Section 3.1.

The next project focuses on users' diaries written before sleeping and after waking up. The participants completed daily surveys and answered the question "Before going to sleep, write a short description of your day, especially events that could affect your sleep" before sleeping and “Describe, in as much detail as you wish, how your sleep was last night.” after they wake up. The diaries are first fed into a polarity detection pipeline, and the polarity of each document is extracted. Then, according to each diary's polarity, a topic modeling approach is applied on the dataset and lastly, the extracted topics are analyzed using Linguistic Inquiry and Word Count [17]. More information regarding the dataset and the data collection approach is provided in Section 4.1.

This thesis attempts to take advantage of state-of-the-art Natural Language models, especially the pre-trained ones which are trained on massive amounts of data, to evaluate them on real-world datasets, people's journals, to see whether the models can extract useful information in the mental health field or not.

Chapter 3 focuses on extracting personality traits based on the state-of-the-art feature extraction and classification approaches in Natural Language Processing. Although the pretrained models have shown significant improvement in these tasks, the interpretability of state-of-the-art models in the mental health task has yet to be investigated. As the deep learning models become more sophisticated, they become less interpretable [31] and less understandable by humans. In the project, we investigate the interpretability of the results derived from the embedding of a state-of-the-art model in this area and improve that which also resulted in higher accuracy and transparency in the task.

In addition, after the enquiring investigation, the limitations faced and some approaches to tackle these limitations are discussed.

Chapter 4 tries to extract insight based on unlabeled sleep diaries using recent topic modeling, polarity detection, and linguistic analysis models. Polarity of the diaries are evaluated and then each document is given some scores based on the topic it belongs to, using Contextualized Topic Modeling [12]. After categorizing the journals as positive or negative, the positive and negative journals are each divided into some categories based on their context similarity using the topic modeling approach.

Overall, the contributions of this thesis can be listed as follows:

1. Finding out whether the current state-of-the-art models can understand psychological discoveries or not.

2. Investigating whether we can build a model to enhance the psychological understanding of the current pre-trained models or not.

3. Analyzing the polarity of sleep diaries of participants using pre-trained deep learning-based models which are multi-layer neural networks capable of understanding more complex structures.

4. Finding out what the topics are that people talk about in their sleep diaries by applying topic modeling models which also consider the context of documents.

5. Developing a new approach to interpret and present the topics extracted from the previous contribution.

The rest of the thesis is structured as follows: The Related work is discussed in Chapter 2. Chapter 3 explains the personality detection project which itself contains several sections. Section 3.1 provides the project introduction. Methodology, which elaborates on the datasets, the approach used in the project and the evaluation method, is explained in Section 3.2. After that, the details of experiments are provided in Section 3.3 and the results which also achieved state-of-the-art performance is provided in Section 3.4. The last section, Section 3.5 concludes the project. Chapter 4, the SleepHealth project, also consists of several sections. The first gives a brief introduction. Section 4.2 explains the methods used in this project and Section 4.3 provides the results after applying the models on the dataset. In the next section, the focus is on the limitation of the given approach and in the last section, Section 4.5, a conclusion is provided. In the last chapter, a summary of the thesis is presented along with outcomes accomplished.

Chapter 2

RelatedWork

The first challenge of the projects discussed in the following chapters is providing valid and reliable data. To extract people's personality traits, there are a variety of personality tests that are based on psychological discoveries [38]. The most accepted one in the field of psychology is the Big Five model, also called OCEAN [48]. This personality test is the one focussed on in this paper. OCEAN assesses five dimensions of personality (Openness to Experience, Conscientiousness, Agreeableness, Extraversion, and Neuroticism or when positively keyed, emotional stability). One other commonly used personality model, which is used in a comparison in Chapter 3, is Myers-Briggs, also known as MBTI [18]. MBTI categorizes personalities into 16 types; each one can be described as a combination of 4 binary categories (Extroversion/Introversion, Sensing/Intuition, Thinking/Feeling, Judging/Perceiving). Since the MBTI test has been questioned for its comprehensiveness, reliablity, and lack of independent categories, the OCEAN personality test is chosen as the main focus of this paper.

To evaluate sleep quality, Polysomnography (PSG) has long been known as the primary method since it captures sleep architecture as well as wake and sleep time [74]. However, due to its higher cost and constraints in availability, actigraphy has acted as an alternative to PSG for long-term sleep-wake cycle monitoring [74]. Usage of wrist actigraphy increased the generalizability of results due to its unobtrusive sleep measurement without disrupting sleep as PSG sometimes does and its mobility which provides measurement across a wide range of circumstances and locations. In addition to the above-mentioned differences, actigraphy can continuously capture records for longer periods with acceptable accuracy [5]. Besides previous methods which are all categorized as objective sleep assessment tools, subjective methods could also be an appealing cost-effective approach to assess metrics that contribute to objective sleep outcomes [73, 104]. Among them, sleep questionnaires are one of the most used tools for the preliminary subjective assessment of sleep. This approach is being used in novel experiments as well as in the traditional experiments [44]. The other method of subjective sleep evaluation is through sleep diaries.

Sleep diaries are records of an individual's sleeping and waking times during a period of several days [110]. They are often used in clinical setups as the primary aid for diagnosing and treating sleep disorders, as sleep patterns impact the symptoms a patient displays [109]. Thus, sleep diaries have been regarded as the gold standard for subjective sleep assessment [24].

The next problem is the kind of data and how it is provided and used. Regarding the Personality project, given the limited mental health service resources, there is a strong need for an automated assistant tool. AI models have proven to be good candidates as they perform more accurately than humans in personality judgment [122]. Some models used psycholinguistic features to identify personality [68]. In the field of deep learning-based automatic personality detection, the hierarchical CNN model [69] has attracted a lot of attention. A full comparison between previous proposed models has been given [77] and perspectives have been analyzed [108]. Although the deep models are improving the accuracy in this field and their approaches have built the foundations of our current work, they suffer from some issues that prevent them from serving as well as they ought to. For example, the results might be based on the studied socio-cultural group. Lewis [64] has analyzed this diversity and has shown that the results can vary depending on the observed cohort. In addition, due to the delicate nature of mental health tasks, trust is an important criterion that these black-box models cannot satisfy without using a post-hoc explainability approach [97].

Concerning the Sleep Health project, given the high adoption of mobile devices, electronic surveys and sleep diaries have been employed in digital health research to improve response accuracy and promote real-time direct data sharing [109]. App-based data collection reduces the time for data entry, automates the process of recording the data, and hence, decreases the cost of data gathering [44]. Although structured sleep diaries have been used before in subjective sleep assessment, few studies have investigated the utility of assessing sleep health using electronic sleep diaries [29, 33, 41]. This type of subjective data provides an opportunity for participants to freely write personalized journals which may be associated with triggers and risk factors for sleep health.

The main focus of this thesis is the usage of Natural Language Processing (NLP) models in the following psychological-related projects. For personality detection, current NLP models that understand human language are mostly proposed by large companies such as Facebook and Google, enabled by their high-spec infrastructure to create their high accuracy predictors [19, 30, 66]. Although they are not runnable on regular computers, their pre-trained versions can be used in personality detection with a small amount of fine-tuning to be adapted to this task [76, 112]. Considering that there is usually a trade-off between accuracy and simplicity, the task to obtain an optimal, yet simple model is non-trivial. Only a few papers, such as BB-SVM [55], have proposed high accuracy models in the field of presonality detection without sacrificing simplicity. BB-SVM also introduced a BERT-based personality model that can be used for longer sequences as well. However, even though this model is able to be run on ordinary computers, its interpretability, especially the justification for the choice of the pre-trained model, has yet to be addressed.

As well as the existing trade-off between complexity and accuracy, a trade-off also exists between performance and transparency (i.e., explainability of the outcomes). The higher performing models tend to be more opaque [31]. As the model becomes more opaque, the need for explainability increases. To alleviate this problem, post-hoc explainability is used. This type of explainability is divided into model-agnostic approaches, which can be used for any model, and model-specific ones. A full comparison of explainable AI methods is available [8].

Also, contemporary models learn from examples in specific datasets. This issue challenges the model when it faces new examples that are not the same as the previously observed ones since current models are not using experts' knowledge. So, even though the current models can do their best for their specific dataset, they cannot incorporate the socio-cultural diversity among groups of people, which results in the different ways they articulate their thoughts [64].

With the emergence of accurate AI models, theorists and researchers make normative claims based on the models' results [49]. Some of the previous experience has also shown how these models can be exploited for detrimental goals [37, 72].

To better understand the meaning of the text for further analysis of the Sleep Health data, based on the recent development of deep learning, AI methods for sentiment analysis have gained more attention, although they still suffer from issues such as requiring a large amount of data (dependency), obtaining different results based on different training datasets (consistency), and uninterpretablity of the reason behind the prediction (transparency) [22]. As Noam Chomsky indicates, we just cannot get to that understanding by throwing a complicated machine at it [53]; several manual inspection phases are still required to produce and understand AI-based results [8, 56]. To understand the meaning of a document, previous methods have tried to manually categorize the words into classes with the same meaning and then use statistical models for further analysis [25, 51, 80, 87, 88]. To automate the association of the words, researchers have tried to encode words into real-valued vectors in which words that are closer in the vector space are expected to be similar in meaning. This resulted in embeddings such as GloVe [90] and word2vec [78]. Although these representations were trained on a huge amount of data, extracting a fixed representation for each word without considering its surrounding words is not always accurate [79]. Disambiguation of the words that have several meanings in different contexts [75, 91] and extracting the sentence level meaning [3, 96] are examples of problems that were yet to be dealt with.

To consider the context of a word, ELMo [92] was developed. It basically looks at the entire sentence before assigning an embedding to each word. Since the trained models were highly biased, based on their structure and their training data, the results produced for a sentence whose domain was outside the scope of the training data were not satisfactory. This issue yielded to the introduction of ULM-FiT [43], which proposed a process to fine-tune the model for various tasks in addition to its contextualized embeddings. In the current state of Natural Language Understanding, newer models have arrived. They consider the context of the text, and can be fine-tuned on different tasks such as sentiment analysis. Many of these models have been developed by refining Bidirectional Encoder Representations from Transformers (BERT) [30, 43], such as RoBERTa [66], the optimized version of BERT; DistilBERT [101], a distilled version of BERT; BigBird [123], which improves BERT with the ability to input longer sequences, and other models [60, 63, 107, 119]. Although BERT and its subsequent models are state-of-the-art in many NLP tasks [30], and further analysis of their architectures and performance has been done by other researchers [59, 121], they still lack in some tasks such as commonsense reasoning [28] and negation [42, 52]. This resulted in impeding researchers' efforts to focus on purely AI-based outcomes [21, 76]

Because of the above-mentioned challenges, although recent digital health models have produced valuable outcomes, especially in dividing documents into categories based on their semantic context (Topic modeling) [39, 45, 124], and analyzing the polarity of text (Polarity detection) [61, 114, 124], they are still not yet able to perceive the correct meaning of complex documents and sometimes misinterpret the meaning of generated topics [46].

In the next chapters, we focus on resolving the above mentioned challenges. For the personality project, by making the AI models more interpretable, more descriptive facts can be obtained based on their results. Ethical concerns can be slightly alleviated because of the insight which the model provides. One of the few works that address both improving personality detection accuracy using deep learning models and providing understandable insight using post-hoc explanablity approaches [76] is used as the baseline for the personality project.

Regarding the Sleep Health project, we analyze the individualized factors that are linked to sleep quality and daily functioning based on the data collected remotely from a longitudinal observational study called SleepHealth Mobile App Study (SHMAS). We aim to utilize state-of-the-art NLP models to detect the polarity of documents, assign a topic to each journal, and then interpret the outcome in a novel approach to see whether there is any future prospective in using AI in analyzing sleep health-related free-text.

Chapter 3

Project 1: Interpretable Representation Learning for Personality Detection

3.1 Introduction

AI [54] has the potential to assist health experts in dealing with the increasing rate of mental health issues and disorders. This increasing trend has been the subject of recent investigations such as the recent trends in mental ill health and health-related behaviors in two cohorts of UK adolescents that show depressive symptoms and self-harm were higher in 2015 compared with 2005 [86]. How social media impacts mental health (including the mental health of adolescents and rising teen suicide rates) has also been studied [84]. This increasing rate of mental issues has accelerated due to the COVID-19 pandemic. According to a Kaiser Family Foundation poll, people have become more socially isolated and stressed. Nearly half of Americans report the coronavirus crisis is harming their mental health [2, 32].

According to a 2020 Harris Poll, between 46% and 51% of US adults were using social media more since the outbreak began [100]. Increased social media use means more digital footprints, and since people's personality and private traits can be identified based on them [58], this pandemic challenge can be turned into an advantage to provide more support for people based on their needs. A WHO survey showed that COVID-19 further burdened the already limited mental health services in many countries [115]. Since mental health service resources are limited and mental health issues have increased, the increase in social media use provides an opportunity for AI researchers to utilize the produced digital footprints to help diagnose people's mental health issues.

Personality traits are defined as the set of relatively stable characteristics which describe our feelings and behaviour. These traits play important roles in individuals' futures and life

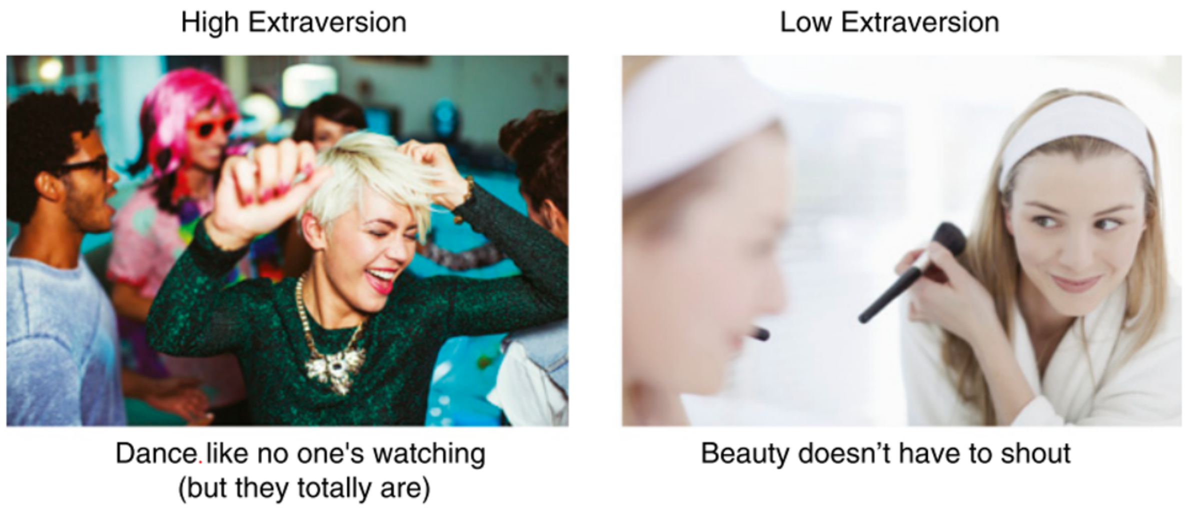


Figure 3.1: An example of a personalized recommender system promoting two types of advertisement for different people in terms of extroversion.

outcomes [85, 98]. Among the various personality tests, the Big-Five, which is also called OCEAN, is known to be the most reliable test for assessing people's personality [48]. The OCEAN test describes personality in five measures: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Previous work has investigated the relationship between personality and mental disorders. Studies have shown that neuroticism plays a vital role in depressive and anxiety disorders [34].

Regarding the other traits, resilience demonstrates a strong inverse relationship with neuroticism and strong positive relationships with extraversion and conscientiousness and a small but statistically significant positive relationship with openness [23]. Hence, understanding a person's personality can provide a better insight for detecting mental illnesses.

In addition to psychological motivation, personality traits are also useful in recommender-systems [72, 120], product and service personalization [102, 113], job screenings [65], social network analysis [70], and sentiment analysis [22].

For example, using personalized advertisements shown in Figure 3.1 can increase profits in selling cosmetics [72]. In this example, the left image with its caption is designed for an extrovert person, and the right image along with its caption is designed for an introverted individual. Each appeals to its corresponding personality type.

There is a common belief that as the deep learning models get more complex, they become less interpretable. This divergence occurs since current state-of-the-art models in all Natural Language Processing fields are using sophisticated architectures. Personality detection is no exception. In the personality detection state-of-the-art model [76], the BERT-base model performed better than its combination with psycholinguistic features, which is not expected. So, the question that comes to mind is whether the mentioned models capture proper psychological information or not.

In this work, we address the following two questions: Does the embedding, which is used for the current state-of-the-art model, capture psychological information? If not, how can it be improved? In order to answer these questions, we first introduce an approach for evaluating embeddings in personality detection. Following that, with metric learning in mind [54], we apply two different approaches using two Siamese architectures for generating the embeddings from the psychological statements. The first approach produces sentence embeddings by means of computing semantic similarities between psychological statements representing different traits. In the second approach, different variants of another Siamese sentence encoder, Sentence-BERT, for producing sentence embeddings for classifying psychological traits are investigated. Both of these approaches surpass the previous state-of-the-art models used in this task with the BFI statement data [11, 48, 47]. The second approach outperforms the previous state-of-the-art models with the Essays dataset [89] and the Kaggle personality dataset [50]. Extensive experiments with the Essays dataset and the BFI statements are performed and discussed. These experiments have focussed on these two datasets since the MBTI test (used in the Kaggle personality dataset) has been questioned for its comprehensiveness, dependability, and lack of independent categories [93], whereas the OCEAN personality test (the Essays dataset) is considered as more reliable. These approaches[1] not only outperform the previous state-of-the-art model but also reduce the computational overhead.

3.2 Methodology

This section discusses the interpretable sentence representation generation approaches using the Siamese architectures, the dataset we use for training the model, and the datasets used for evaluating the performance of the models. The sentence representation is generated by means of computing the semantic similarities between psychological statements. The reason behind choosing this approach is to preserve enriched semantics in the vector representations. Finally, the approach to interpret the output of the model is discussed along with the evaluation of the model. The interpretability of our approach is evaluated using the feature relevance and visual explanation methods of the post-hoc explainability category (see the taxonomy in Fig. 6 of [8]), by computing the cosine similarity between the input and baseline sentences and using PCA visualization, respectively.

[1] Code: `https://github.com/amirmohammadkz/interpretable_personality`

Table 3.1: The baseline sentences for each trait of the Big Five personality test

Text	Trait	Label
I am extrovert	Ext	1
I am introvert	Ext	0
I am agreeable	Agr	1
I am disagreeable	Agr	0
I am neurotic	Neu	1
I am stable	Neu	0
I am an open person	Opn	1
I am not an open person	Opn	0
I am conscientious	Con	1
I am casual	Con	0

Datasets

We used the following publicly available personality datasets in our analyses:

Essays This well-known stream-of-consciousness dataset consists of 2468 essays written by students and annotated with the binary labels of the Big Five personality traits which were found by a standardized self-report questionnaire [89].

Kaggle MBTI This data was collected through the PersonalityCafe forum providing a diverse selection of people interacting in an informal online social setting. The dataset comprises 8675 records of each person's last 50 posts on the website along with their MBTI binary personality type [50].

Evaluating the Embeddings

In order to evaluate the pretrained BERT-base model for meaningful personality representations, we have used a simplified version of the Big Five Inventory (BFI) [11, 48, 47]. BFI is a self-report questionnaire that consists of 44 short phrases. Participants rate each of these statements based on their situation. Each statement focuses on assessing one of the five traits. We have simplified this version to make it easier for language models to extract meaningful representations from them. For example, the statement “I am someone who is talkative”, which assesses the extraversion rate of a person, is converted to “I am talkative”. In addition, to increase the dataset size, we have also added the adapted version of BFI [27, 36] to the original one. The final simplified statement set consists of 85 sentences, 44 of which belong to the original BFI statements and the rest are obtained from the adapted version. We have also used two baseline sentences for each trait. These sentences are listed in Table 3.1. We then use the

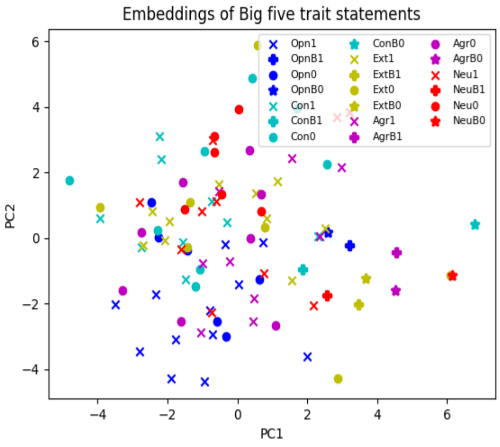


Figure 3.2: Visualization of the personality statements after applying PCA on the average of the output of layer 11 of Bert-base [76] for both BFI and baseline (indicated with 'B') sentences. 1 and 0 mean "High" and "Low" rate of a specific trait, respectively, and "B" is for baseline sentences.

pretrained version of BERT-base to extract the representations of the tokens. We have followed the best representation of [76] which is averaging the output of the second to last layer to get the final representation of each statement. Next, we transform the embeddings using a PCA [1] with 2 principal components. The result of the PCA is illustrated in Fig. 3.2. The B-points are clustered in the upper half of the bottom right quadrant, whereas the 0- and 1-points are almost all in the left or upper quadrants. The representations of the baseline sentences are very close to each other and the distance between them and the corresponding trait statements are much larger. Hence, we can conclude that even when [76] gets high accuracy using these representations, it will not be generalizable since the extracted embeddings do not manifest the related personalities. Considering that this current state-of-the-art representation uses a rich corpus and state-of-the-art language models, we can infer that older ones probably also suffer from this issue. Furthermore, even if the baseline representations obtained from the previous methods maintain sufficient distance, their classification performance is worse compared to [76] which is also not acceptable. This motivates our investigating a model which cannot only improve the classification performance but also enhance explainability.

Interpretable Representation for Personality Detection

This chapter investigates two different approaches for producing vector representations from psychological statements. The core idea behind both approaches is to use the extracted embeddings from the baseline sentences and BFI statements in order to evaluate the performance of the model. The output embedding can be explainable using this comparison.

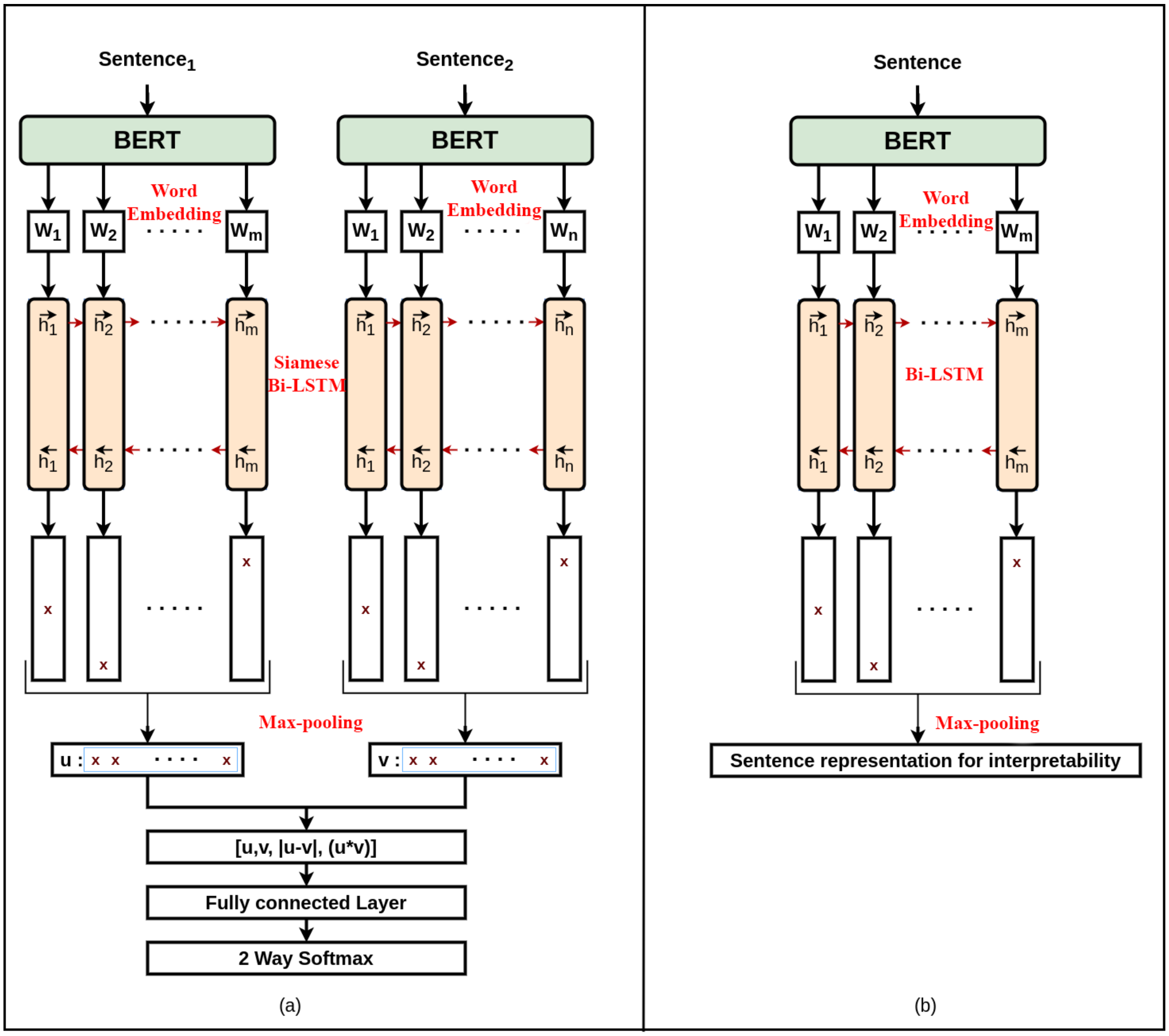


Figure 3.3: Architecture of the model with Siamese Bi-LSTM and max-pooling for the interpretable tool for personality detection. (a) The training of the model, (b) After training, the Bi-LSTM followed by the max-pooling layer act as the sentence encoder.

Both of these approaches use Siamese architectures using deep learning models. The first approach utilizes Siamese Bi-LSTM with max-pooling over time of the output vectors. This model is trained on the simplified BFI statement pairs for computing the similarity between them. The second approach evaluates the Sentence-BERT variants [96]. The reason behind choosing the Siamese models here is that we try to detect the personality traits not by applying direct classification approaches but rather by preserving the semantics of the statements where statements reflecting similar traits remain close to each other in the embedding space. This objective is achieved by leveraging psychological datasets (the BFI statements and the baseline sentences).

Bi-LSTM with Max-pooling To extract the feature vectors of both the BFI statements and the baseline sentences, we have used the Siamese architecture of Bi-LSTM over the BERT word embeddings from layer 11 of BERT-base. The architecture is inspired by the InferSent model [26]. The basic idea of this model is to generate a sentence embedding by means of computing the semantic similarity between two sentences. This semantics attempts to preserve the personality trait from the BFI statement.

For the word embeddings we have chosen the output of layer 11 of the pre-trained BERT-base. For any given sentence pair, word embeddings are fed to two identical Bi-LSTMs. These Bi-LSTMs share the same parameters and weights. For a sequence of N words, Bi-LSTM produces a set of N vectors. The final hidden state representation for each time step is generated by concatenating the hidden representation of the forward $(\overrightarrow{h_i})$ and backward LSTMs $(\overleftarrow{h_i})$ [103]. For each time step, max-pooling is applied over these concatenated hidden representations $([\overrightarrow{h_i}, \overleftarrow{h_i}])$ to generate an intermediate sentence representation. In the next step, three operations, concatenation, point-wise difference and point-wise multiplication, are performed on the representations obtained for both of the sentences from the sentence pair. Finally, the outcome of these three matching operations are concatenated and fed to a feed-forward neural network for classification like [26]. Suppose, u and v are the intermediate representations for the sentences after max-pooling. Then $[u, v, |u - v|, (u * v)]$ would be the final feature representation to be fed to the following classifier. The classifier outputs either 0 or 1 where 1 indicates the sentences offer semantically similar traits and 0 otherwise. Fig. 3.3 portrays the overall architecture of the model. After the training is done, the Bi-LSTM together with the max-pooling layer acts as the encoder for generating the sentence representation. This representation is a 768 dimensional vector.

Sentence-BERT Sentence-BERT [96] is a refinement of the pretrained BERT using Siamese and triplet structures. It can derive sentence representations preserving the semantics of the sentences. Unlike BERT, which outputs rich token embeddings and [CLS] with poor semantics for the sentence, Sentence-BERT produces semantically richer sentence embeddings. It is trained on the sentence pairs from the SNLI dataset [16] and multi-genre NLI dataset [116]. It has been shown that sentence embedding models trained on natural language inference datasets have better semantic preserving abilities [26]. For this reason, Sentence-BERT outputs semantically richer sentence embeddings.

Sentence-BERT incorporates a mean-pooling operation over the output of each BERT embedding to generate two sentence embeddings for the sentence pair. Then two matching operations, concatenation and point-wise difference, are performed on them. Finally, this feature is fed to the softmax classifier. After the fine tuning is complete, the fine-tuned BERT with

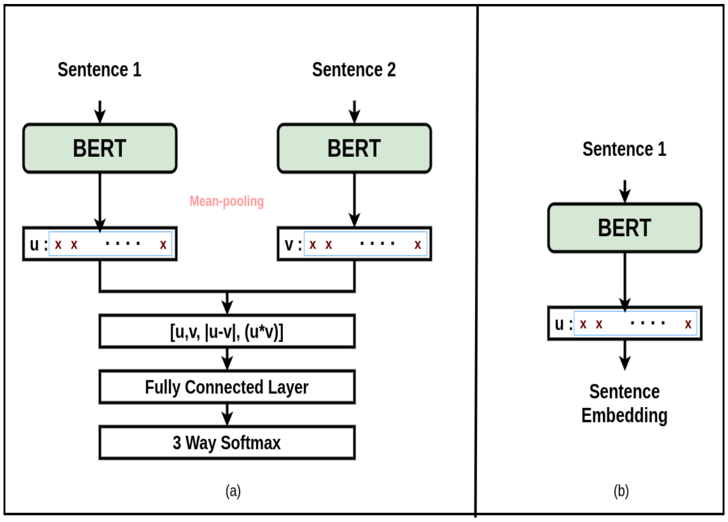


Figure 3.4: Architecture of Sentence-BERT. (a) Training of the model on the natural language inference datasets. (b) Sentence encoder.

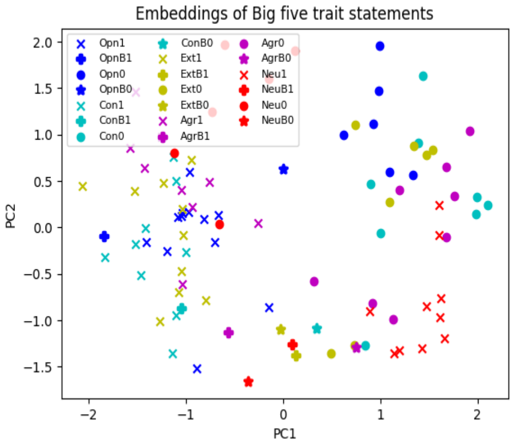


Figure 3.5: Visualization of the personality statements after applying PCA on the feature vectors of Bi-LSTM and max-pooling. 1 and 0 mean “High” and “Low” rate of a specific trait, respectively, and “B” is for baseline sentences.

the mean-pooling act as the sentence encoder. Using this pretrained Sentence-BERT is then a straight-forward approach. After being given a sentence, it directly outputs the corresponding 768 dimensional vector sentence embedding. The architecture of Sentence-BERT is shown in Fig. 3.4. We have conducted experiments on the Essays, the BFI statements, and the Kaggle datasets using different variants of Sentence-BERT [95]. In all cases the overall architecture remains the same, only the BERT encoder is varied. Some prominent variants are RoBERTa [66] and MPNet [107].

3.3 Experiments

To analyze the effectiveness of our Siamese Bi-LSTM model, for each personality trait t, we create all possible corresponding BFI statement pairs together with the appropriate label, $(s_i, s_j, l_{i,j})$, where $l_{i,j}$ is 1 if the statements s_i and s_j have the same label and 0 if s_i and s_j have different labels. Then, we feed the statement pairs as inputs to the model and use $l_{i,j}$ as the label which the model tries to predict. Applying this approach over the BFI statements, the data set has 681 sentence pairs. Among these, 600 samples are used for training and the remaining 81 are used for validation. This small dataset was sufficient for training the Siamese LSTM model with some good training and validation accuracies. While testing this model on the BFI statements, it achieved a better result compared to the previous models [76]. This comparison is performed using the *PredLabel* and *SimScore* metrics. In addition, the finetuned embedding are also assessed by replacing the embedding part of the model in [76] for classifying the Kaggle and Essays datasets. However, the model trained on this data did not achieve state of the art accuracies as the training data was comparably small.

We have trained the Siamese Bi-LSTM model for only 25 epochs where the best result was found at the 21st epoch. While training, the batch size was set to 10 with 10% dropout. Standard gradient descent was used for optimization with a learning rate $1e^{-5}$. The forward and backward LSTMs' hidden representations are 384 dimensional vectors.

After the training phase, we use the feature vectors extracted from the Bi-LSTM for evaluation as we did in Section 3.2 for the BFI statements. After extracting the feature vectors of both the BFI statements and the baseline sentences, for each statement that belongs to trait t we assign a similarity score and prediction label based on the closeness to the corresponding baseline sentences as following:

$$\forall s_i \in S_t : SimScore(s_i) = \\ (-1)^{l_i - 1} C(s_i, b_{t,1}) + (-1)^{l_i} C(s_i, b_{t,0})$$

and

$$PredLabel(s_i) = \begin{cases} 1, & \text{if } C(s_i, b_{t,1}) > C(s_i, b_{t,0}) \\ 0, & \text{otherwise} \end{cases}$$

where l_i is the label of s_i, C is cosine similarity, and $b_{t,0}$, $b_{t,1}$ are the baseline feature vectors of trait t. To report the result of a specific model, we use accuracy for the *PredLabel*s and the average of the *SimScore*s. For the Sentence-BERT models, the BFI statements and baseline statements are fed to the pretrained encoders and then the accuracy of the *PredLabel*s and the average of the *SimScore*s are computed. While testing, we aggregated both the simplified and non-simplified versions of the BFI statements to generate a more generalized model. The embeddings of the BFI and the baseline statements are extracted from the encoder portion of the Siamese Bi-LSTM as previously described and finally, *PredLabel*s and *SimScore*s are measured.

In the case of experimenting with the Essays dataset, no further training is performed. The statements are fed to the models (both the Bi-LSTM with max-pooling and the Sentence-BERTs). Then they are tested against the baseline statements to compute the performance metrics. The Kaggle dataset is tested with the Sentence-BERTs only.

3.4 Results

The accuracies of the *PredLabels* are shown in Table 3.2, and the *SimScore*s for the BFI statements, in Table 3.3. For three traits, Bi-LSTM with max-pooling outperforms the CLS and average methods of BERT which were used in [76]'s state-of-the-art model for this task and outperforms on the average result as well. For each of the personality traits, the 0- and 1-statements form distinguishable and well-separated clusters except for the Neuroticism and Extroversion baseline sentences, which are so close to each other. The PCA result is illustrated in Fig. 3.5. The evaluation also tries to identify whether the model is able to assign the correct binary trait label to the statements. For Openness, Conscientiousness, and Agreeableness, as it is shown in Fig. 3.5, the model can almost completely understand which statement belongs to which baseline trait. Regarding Neuroticism, although the *SimScore* is better than both the CLS and the average methods, the classification metric was not satisfactory. Extraversion also seems to be the most difficult trait to be identified by baseline sentences. Although the statements are separated, the embeddings of "I am extrovert" and "I am introvert" are still too close, resulting in the poor result. We believe this issue happens because of the dataset which is used for training BERT.

Overall, since we have not used the baseline sentences in any phase of the training process, and they are used only in the evaluation, we believe that Bi-LSTM with max-pooling has used the general language model knowledge enriched with knowledge from the psychological statements to distinguish between traits. Average results have shown that this model is successful in learning the personality trait-specific representations while retaining its knowledge from the pre-trained BERT.

Even though the Bi-LSTM with max pooling outperforms the previous state-of-the-art when compared by performance metrics as well as richer personality trait-specific representation generation, the Sentence-BERT based model outperforms this one. We have experimented with different variants of Sentence-BERT. Among them, the most prominent results are found when RoBERTa-large or MPNet are used as the encoder in the Sentence-BERT architecture.

Table 3.2: Comparison of accuracies of *PredLabel*s of different representations. The two first models are driven from state-of-the-art paper[76], and the third one is our fine-tuned model. The next ones are different Sentence-BERT models listed alphabetically.

Model	O	C	E	A	N	Average
BERT (average) [76]	61.11	52.94	41.18	64.71	56.25	55.24
BERT (CLS)	33.33	58.82	41.18	47.06	62.5	48.58
Bi-LSTM with max-pooling	94.44	100.00	32.35	100.00	53.13	75.98
average_word_embeddings_glove.6B.300d	33.33	58.82	70.59	76.47	43.75	56.59
average_word_embeddings_glove.840B.300d	33.33	64.71	88.24	70.59	62.50	63.87
average_word_embeddings_komninos	33.33	70.59	76.47	70.59	75.00	65.20
average_word_embeddings_levy_dependency	33.33	41.18	47.06	64.71	62.50	49.76
nli-bert-base	66.67	76.47	70.59	88.24	100.00	80.39
nli-bert-base-cls-pooling	77.78	76.47	70.59	88.24	93.75	81.36
nli-bert-base-max-pooling	77.78	88.24	70.59	88.24	93.75	83.72
nli-bert-large	94.44	94.12	100.00	88.24	93.75	94.11
nli-bert-large-cls-pooling	88.89	88.24	100.00	88.24	100.00	93.07
nli-bert-large-max-pooling	88.89	82.35	100.00	88.24	100.00	91.90
nli-distilbert-base	72.22	88.24	17.65	88.24	93.75	72.02
nli-distilbert-base-max-pooling	77.78	82.35	11.77	88.24	87.50	69.53
nli-distilroberta-base-v2	72.22	94.12	70.59	88.24	100.00	85.03
nli-mpnet-base-v2	100.00	88.24	94.12	94.12	93.75	94.04
nli-roberta-base	94.44	82.35	100.00	88.24	93.75	91.76
nli-roberta-base-v2	83.33	94.12	100.00	88.24	100.00	93.14
nli-roberta-large	100.00	100.00	100.00	88.24	100.00	**97.65**
paraphrase-distilroberta-base-v1	33.33	70.59	47.06	70.59	87.50	61.81
paraphrase-xlm-r-multilingual-v1	83.33	70.59	47.06	76.47	93.75	74.24
stsb-bert-base	72.22	76.47	76.47	76.47	87.50	77.83
stsb-bert-large	88.89	88.24	100.00	82.35	68.75	85.65
stsb-distilbert-base	72.22	88.24	29.41	82.35	93.75	73.19
stsb-distilroberta-base-v2	72.22	82.35	70.59	82.35	100.00	81.50
stsb-mpnet-base-v2	94.44	94.12	94.12	100.00	93.75	95.29
stsb-roberta-base	100.00	70.59	76.47	82.35	100.00	85.88
stsb-roberta-base-v2	88.89	70.59	88.24	88.24	100.00	87.19
stsb-roberta-large	100.00	94.12	76.47	88.24	100.00	91.77

Table 3.3: Comparison of *SimScore*s of different representations. The two first models are driven from state-of-the-art paper[76], and the third one is our fine-tuned model. The next ones are different Sentence-BERT models listed alphabetically.

Model	O	C	E	A	N	Average
BERT (average) [76]	0.011	0.007	-0.003	0.026	0.002	0.009
BERT (CLS)	0.001	0.001	-0.011	0.012	0.009	0.002
Bi-LSTM with max-pooling	0.082	0.064	-0.01565	0.079	0.008	0.044
average_word_embeddings_glove.6B.300d	0.000	0.039	0.038	0.066	0.011	0.031
average_word_embeddings_glove.840B.300d	0.000	0.036	0.040	0.082	0.077	0.047
average_word_embeddings_komninos	0.000	0.036	0.031	0.039	0.077	0.036
average_word_embeddings_levy_dependency	0.000	-0.007	0.002	0.020	0.075	0.018
nli-bert-base	0.124	0.148	0.073	0.253	0.321	0.184
nli-bert-base-cls-pooling	0.145	0.134	0.063	0.277	0.330	0.190
nli-bert-base-max-pooling	0.116	0.141	0.035	0.187	0.224	0.141
nli-bert-large	0.231	0.211	0.160	0.270	0.211	0.217
nli-bert-large-cls-pooling	0.224	0.166	0.159	0.281	0.304	0.227
nli-bert-large-max-pooling	0.163	0.169	0.246	0.283	0.264	0.225
nli-distilbert-base	0.068	0.149	-0.088	0.194	0.224	0.109
nli-distilbert-base-max-pooling	0.088	0.147	-0.082	0.162	0.166	0.096
nli-distilroberta-base-v2	0.037	0.119	0.046	0.180	0.181	0.112
nli-mpnet-base-v2	0.148	0.086	0.209	0.253	0.223	0.184
nli-roberta-base	0.194	0.158	0.142	0.228	0.356	0.215
nli-roberta-base-v2	0.160	0.117	0.138	0.206	0.226	0.169
nli-roberta-large	0.248	0.278	0.245	0.274	0.415	**0.292**
paraphrase-distilroberta-base-v1	0.020	0.025	0.002	0.060	0.080	0.037
paraphrase-xlm-r-multilingual-v1	0.032	0.030	-0.004	0.074	0.117	0.050
stsb-bert-base	0.158	0.129	0.150	0.200	0.212	0.170
stsb-bert-large	0.251	0.174	0.145	0.261	0.140	0.194
stsb-distilbert-base	0.119	0.163	-0.041	0.221	0.272	0.147
stsb-distilroberta-base-v2	0.045	0.131	0.039	0.196	0.227	0.128
stsb-mpnet-base-v2	0.174	0.081	0.206	0.191	0.179	0.166
stsb-roberta-base	0.259	0.095	0.152	0.305	0.352	0.233
stsb-roberta-base-v2	0.107	0.097	0.122	0.190	0.243	0.152
stsb-roberta-large	0.218	0.262	0.077	0.226	0.315	0.219

Table 3.4: Accuracy of Bi-LSTM with max-pooling and Sentence BERT models on the Essays and Kaggle datasets. The two first models are driven from state-of-the-art paper[76], and the third one is our fine-tuned model. The next ones are different Sentence-BERT models listed alphabetically.

MODEL	Essays						Kaggle MBTI				
	O	C	E	A	N	Average	I/E	N/S	T/F	P/J	Average
Majority Baseline	51.5	50.8	51.7	53.1	50.0	51.4	77.0	85.3	54.1	60.4	69.2
BERT-base [76]	64.6	59.2	60.0	58.8	60.5	60.6	78.3	86.4	74.4	64.4	75.9
BERT-large [76]	63.4	58.9	59.2	58.3	58.9	59.7	78.8	86.3	76.1	67.2	77.1
Bi-LSTM max-pooling_combined	61.7	54.6	55.0	56.7	55.9	56.8	-	-	-	-	-
average_word_embeddings_glove.6B.300d	63.2	58.5	56.3	57.2	58.5	58.7	77.2	86.5	76.9	66.2	76.7
average_word_embeddings_glove.840B.300d	63.0	58.0	57.2	57.5	57.7	58.7	78.6	87.1	**79.6**	68.6	78.5
average_word_embeddings_komninos	62.5	57.9	55.3	56.6	58.5	58.1	77.0	86.2	74.3	63.0	75.1
average_word_embeddings_levy_dependency	61.4	55.9	54.0	53.3	56.7	56.3	77.0	86.2	70.2	60.5	73.5
nli-bert-base	64.0	**60.0**	58.7	58.2	60.4	60.2	77.6	86.4	70.8	62.5	74.3
nli-bert-base-cls-pooling	63.8	59.7	57.7	59.1	60.1	60.1	77.6	86.3	71.1	62.2	74.3
nli-bert-base-max-pooling	63.0	58.0	56.7	57.4	58.4	58.7	77.5	86.2	69.7	61.8	73.8
nli-bert-large	63.5	59.8	57.1	58.7	60.8	60.0	77.6	86.3	71.2	62.2	74.3
nli-bert-large-cls-pooling	63.6	59.2	57.9	58.7	60.1	59.9	77.5	86.3	71.3	62.7	74.4
nli-bert-large-max-pooling	63.0	58.1	58.3	58.5	59.1	59.4	77.5	86.2	70.8	61.9	74.1
nli-distilbert-base	62.5	58.8	58.5	57.8	59.4	59.4	77.6	86.2	71.4	62.3	74.4
nli-distilbert-base-max-pooling	62.4	57.0	57.5	57.5	60.2	58.9	77.5	86.2	68.8	61.7	73.6
nli-distilroberta-base-v2	63.2	58.5	59.5	58.7	**61.5**	60.3	81.0	87.3	77.9	**71.5**	**79.4**
nli-mpnet-base-v2	64.2	58.8	59.7	59.1	60.6	60.5	81.0	87.2	78.1	69.3	78.9
nli-roberta-base	62.0	59.1	58.9	59.2	59.0	59.6	77.7	86.3	72.0	62.4	74.6
nli-roberta-large	63.9	59.5	**60.2**	**59.5**	61.3	**60.9**	80.7	87.2	77.7	70.9	79.1
nli-roberta-base-v2	62.8	59.7	58.9	59.3	60.8	60.3	77.9	86.5	72.0	63.1	74.9
paraphrase-distilroberta-base-v1	**65.0**	57.8	59.3	59.0	59.7	60.2	80.1	87.1	76.2	70.7	78.5
paraphrase-xlm-r-multilingual-v1	63.6	58.1	58.8	57.3	59.8	59.5	79.1	86.6	74.2	67.8	77.0
stsb-bert-base	64.0	59.1	57.7	58.1	60.6	59.9	78.1	86.5	72.4	63.4	75.1
stsb-bert-large	62.4	56.9	58.0	58.1	61.4	59.4	77.5	86.5	71.3	62.4	74.4
stsb-distilbert-base	62.8	58.0	58.0	57.1	59.3	59.1	78.5	86.5	73.1	64.6	75.7
stsb-distilroberta-base-v2	63.8	58.9	58.5	58.9	59.8	60.0	**81.1**	87.2	77.3	71.0	79.2
stsb-mpnet-base-v2	64.2	58.6	58.7	59.0	61.1	60.3	**81.1**	**87.5**	78.0	69.1	78.9
stsb-roberta-base	63.4	58.2	57.4	57.8	59.5	59.3	80.3	86.8	76.1	65.8	77.2
stsb-roberta-base-v2	63.4	58.7	59.7	58.9	60.6	60.3	81.0	87.3	77.5	70.3	79.0
stsb-roberta-large	62.7	58.4	57.6	58.0	59.7	59.3	80.1	86.6	74.2	65.4	76.6

Table 3.5: The Pearson correlation between the Predlabel accuracy and the Essays accuracy for all Sentence-BERT embeddings. *p <.05. **p <.001, two-tailed.

O	C	E	A	N	Ave.
0.086	0.488*	0.208	0.662**	0.533**	0.700**

In terms of accuracy of the *PredLabel*s and *SimScore*s, overall, RoBERTa-large performs the best. It achieves an accuracy for *PredLabel* of 97.65% which is almost double the previous state-of-the-art model's accuracy [76]. Apart from Agreeableness, its *PredLabel* accuracy is 100%, whereas for Agreeableness, it's 88.24%. MPNet achieves 100% *PredLabel* accuracy for Agreeableness. On average MPNet achieves 95.29% *PredLabel* accuracy. In terms of *SimScore*s, RoBERTa-large performs the best in all cases apart from Agreeableness. Still, its average value, 0.292, is more than three times that of [76]'s result. For Agreeableness, the encoder with MPNet performs the best for *SimScore*, 0.305, and on average it achieves 0.233.

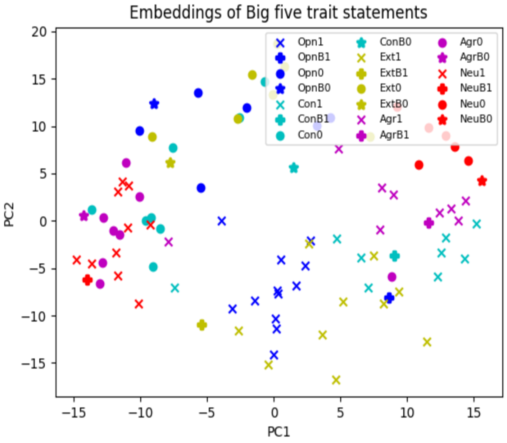


Figure 3.6: Visualization of the personality statements after applying PCA on the nli-roberta-large version of Sentence BERT. 1 and 0 mean “High” and “Low” rate of a specific trait, respectively, and “B” is for baseline sentences.

Fig. 3.6 portrays in a 2D projection the representations generated by the RoBERTa-large version of Sentence-BERT, showing that the closeness of each statement to any particular trait is very clear. For each of the personality traits, the 0-, 1- and B-statements form distinguishable and well-separated clusters (with a couple of exceptions) as demonstrated. One issue of note, two metrics, *PredLabel* and *SimScore*, are used to measure the performance of the model. PCA has been used only to provide a visualization of the embeddings to show how close the representations of the similar trait samples are. We have also used other visualization techniques like t-SNE, UMAP, and LDA. As all the visualization results are very similar, we have reported only the PCA visualization.

To evaluate the generalizability of the model, we tested these models on the Essays and the Kaggle personality datasets. This time the Bi-LSTM with max-pooling performs worse than [76]’s work. The overall accuracy is almost 2% lower for the Essay dataset. But this is justifiable as this Siamese model was trained on very short sentences from the BFI statements, whereas the Essays dataset comes with long paragraphs. Additionally, LSTM based models face shortcomings while working with very long sequences. But the Sentence-BERT models, without any kind of additional operations, outperform the BERT-based averaging technique [76]. This time, RoBERTa-large achieves 60.9% accuracy which is an almost 1 percentage point boost compared to the previous works. In the case of the Kaggle personality dataset, RoBERTa-large gains almost 2 percentage points more accuracy (79.1%). However, DistilRoBERTa performs the best for this dataset and achieves 79.4% accuracy. In both cases, MPNet shows prominent results with accuracies 60.3% and 78.9%, respectively.

We also computed the Pearson correlation of the accuracy of PredLabel and Essays to see if the PredLabel accuracy gives any insight into how an encoder works for real world datasets. As demonstrated in Table 3.5, although the experimented encoders are not specifically designed for long sequence datasets such as Essays, for most traits, especially the average of the traits, there is a significant positive correlation between these two accuracies. Hence, we can conclude that using PredLabel is a good approach for picking the best encoder for real-life datasets.

One notable significance of these models is that none of them have been enhanced with any kind of additional psychological features, unlike [76]. While training, the models are simply trained with sentence pairs. Thus it reduces the computational overhead as well. And as RoBERTa-large was initially trained over larger sequences and then fine-tuned again over natural language inference data, Sentence-BERT with RoBERTa-large earns the capability to produce sentence embeddings preserving richer semantics than the others. Furthermore, as the Sentence-BERT models are trained on a very large corpus of real life inference data compared to the Siamese LSTM model, which is trained on the small BFI statement pairs dataset, they have achieved the ability to provide better representations of the statements.

3.5 Conclusion

In this chapter, we analyze the weakness of the state-of-the-art personality detection model. In addition, with computationally less overhead our model delivers sentence embeddings for psychological statements with rich semantics. Our results show that our enriched representations distinguish the personality traits better than the CLS and average methods which are common in the field. Furthermore, we have used the enriched representations in addition to Sentence-BERT models to classify traits based on their closeness to the baseline psychological statements so the result can be regarded as interpretable. Our experiments improved the Kaggle state-of-the-art accuracy by 2.3 percentage points and Essays by 0.3 percentage points. This work restricts the statements at the sentence level. In future it can be extended to the paragraph level using hierarchical models like SMITH [118] so that better representations from the paragraphs can also be captured. Besides, BFI statements can be used within the prediction model to identify the closeness of each of the samples in the dataset with each of the BFI statements. We believe this method will help psychologists to get better insights into the prediction.

Chapter 4

Project 2: SleepHealth Data Analysis

4.1 Introduction

Sleep-related disorders affect up to 70 million Americans [9, 67, 105] and more than 30% of the U.S. adult population is estimated to being sleep deprived. The CDC has described sleep deprivation as a public health epidemic in the U.S. [67]. Studies have shown that there is a strong association between sleep deficiency and mortality [111], resulting in 5 of the top 15 causes of death in the U.S. [57]. Insufficient sleep also impacts negatively on decision-making and reaction time, making judgments, and workplace productivity [81, 99]. The most common comorbidities of sleep deprivation are anxiety, depression, bipolar disorder, and ADHD [40, 6].

In recent decades, understanding and assessing "sleep health" has emerged as an important area of research in sleep medicine. According to WHO, health is defined as "A state of complete physical, mental and social well-being and not merely the absence of disease or infirmity." [82]. So, in order to assess health in sleep, the focus should be both on the well-being and the absence of disease in different aspects: physical, mental, and social [20]. This goal can be achieved by monitoring several indicators that contribute in sleep disorders or disturbances such as sleep latency, number and duration of nocturnal awakenings, etc. [83].

In this study, first, we used the extracted data which are individualized factors that are linked to sleep quality and daily functioning which is gathered from the SleepHealth Mobile App Study (SHMAS). The obtained data is divided into two categories based on their structure. Structured surveys which contain numerical results for each question can easily be transferred to cross-section analysis without any further preprocessing or analysis because of the nature of that type of data. The latter category contains unstructured surveys which are the main focus of this project.

Since unstructured journals are textual information, numerical information should be extracted from them. Hence, it passes through an NLP pipeline which starts with a polarity

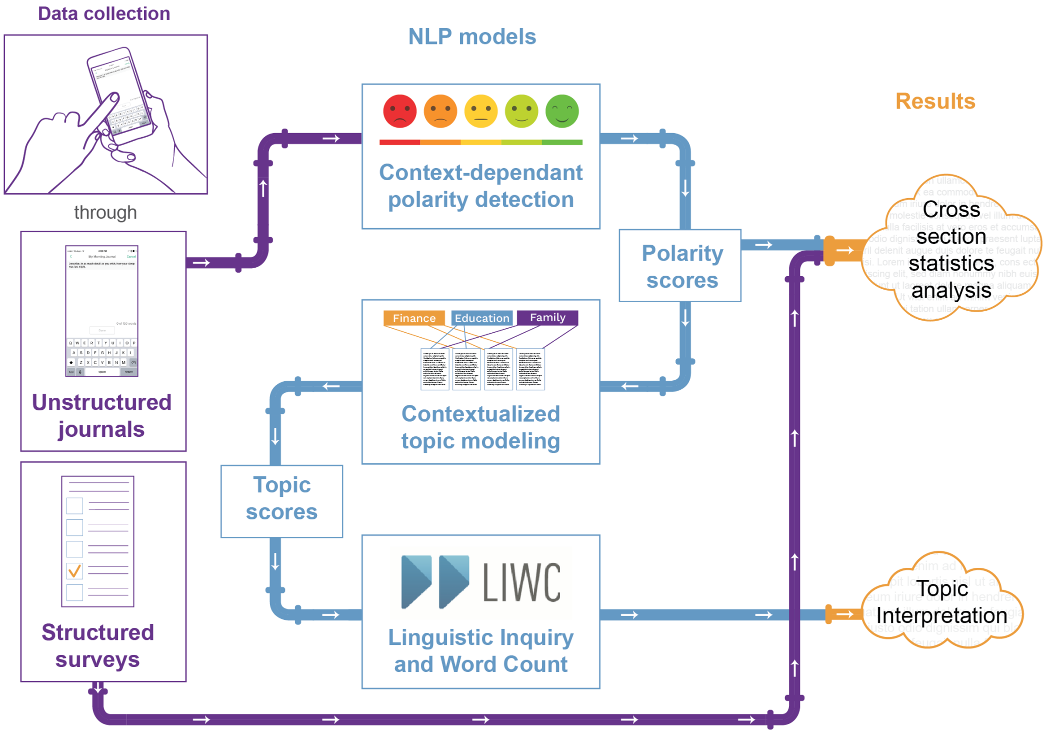


Figure 4.1: Flow of the sleep health from data extraction to topic interpretation

detection model, then topic modeling and the last one, Linguistic Inquiry and Word Count (LIWC), to extract all the information required to interpret the journals.

The first step in the NLP pipeline assigns a polarity score between 0 and 1 to each journal based on the sentiment it conveys. This polarity score can be used further for cross section analysis to find its association with the structured data. The second step divides positive and negative journals, independently, into clusters each encompassing a specific topic. After finding the topic and the polarity to which each journal belongs, LIWC is utilized to analyze the meaning of each topic, thereby providing more information about the context that the topic is representing by investigating its subsumed journals. After all of these phases, the topics of the unstructured data are provided along with their polarity and the top examples in each topic.

4.2 Methods

4.2.1 Study Details

The SHMAS study was developed using Apple's ResearchKit (RK) and was launched on the Apple App Store on March 2, 2016 [29]. Eligible participants were 18 years or older, lived in the United States, had reading and writing proficiency in English, and had an iPhone with the application installed [29].

4.2.2 Data Collection

The study app collected a combination of real-world data about sleep ranging from self-reported surveys to optional sharing of health-kit data passively.

Baseline Surveys Consented participants completed various surveys related to their sociodemographics, family health, and sleep status. Specifically, the onboarding survey collected basic participant demographics including weight, sex, height, age, and family information. Additionally, participants were asked lifestyle and quality of life-related questions such as alcohol, caffeine, and tobacco consumption. Participants' health conditions and beliefs about how they relate to their sleep were captured in the My Health survey. These individual surveys were gradually administered over the first three days of the study and were intended to be completed once during onboarding and subsequently every quarter.

Baseline sleep-related surveys This type of survey consists of two surveys about their sleep duration and quality (the Sleep Habits survey) as well as sleep disorder symptoms and daytime function (the Sleep Assessment survey), respectively [13, 29].

Self-reported sleep journals In the second form of data, the daily activities which were to be completed for a minimum of 5 out of 7 days each quarter, included AM and PM check-in activities such as Consensus Sleep Diary (CSD) [24] and feedback from their previous night's sleep and current day's activities. Participants had the option to complete the quarterly surveys and daily activities as many times as they wanted.

For complete details on study protocol and data collection refer to the SHMAS study data descriptor paper [29].

4.2.3 Data featurization

To better understand the text and its underlying meaning, we utilized two approaches in the field of NLP named polarity detection and topic modeling. In each approach, we benefited from state-of-the-art models evaluated on general NLP benchmarks.

Polarity detection

To detect the polarity, i.e., the sentiment (positive/neutral/negative), from participants' self-reported daily sleep diaries, we used an NLP-based approach. First, we split the samples into sentences using the Natural Language Toolkit (NLTK) [14] platform. Each sentence is then fed into the Huggingface sentiment analysis pipeline [117] with the softmax function, so the polarity output for each sentence would be a continuous number between 0 and 1 which are the negative and positive ends of the spectrum, respectively. The polarity score of the input journal would also be the average of its sentences' polarity scores.

We used the fine-tuned version of DistilBERT [101] on the Stanford Sentiment Treebank (SST-2) [106] dataset. This model is a lightweight Neural Network that also considers the context of the input text in comparison to its classic counterparts. For example, the sentence "I am not happy" will correctly be classified as a negative sentence despite the positive word "happy".

Topic Modeling

The task of topic modeling is designed to assign a topic to each document that summarizes its semantic context. The generated topics can be described using the words that are incorporated the most in building the semantic/co-occurrence within that topic. We used the Combined Topic Modeling (CTM) version of Contextualized Topic Modeling [12] which is an improved version of Latent Dirichlet Allocation (LDA) [15] that also considers the context of the given input. The input of the model aggregates the bag-of-words representation of the pre-processed dataset along with its SentenceBERT [96] embeddings. To improve the quality of the output topics, we first divide the journals into positive and negative subsets using our polarity detection pipeline and apply Topic Modeling afterward. We also benefited from coherence value to optimize the number of topics along with manual inspection of topics' meanings. After topics are extracted and each document is labeled with its corresponding topic, the corpus of each topic is interpreted using LIWC, which is a gold standard software that analyzes different psycholinguistic aspects of a given text [87].

4.3 Results

4.3.1 Sentiment Analysis

Of the 5644 participants who submitted at least one sleep journal, 3744 (66.3%) of them submitted both AM and PM journals. 5201 (92.2%) participants contributed 39,103 AM journals

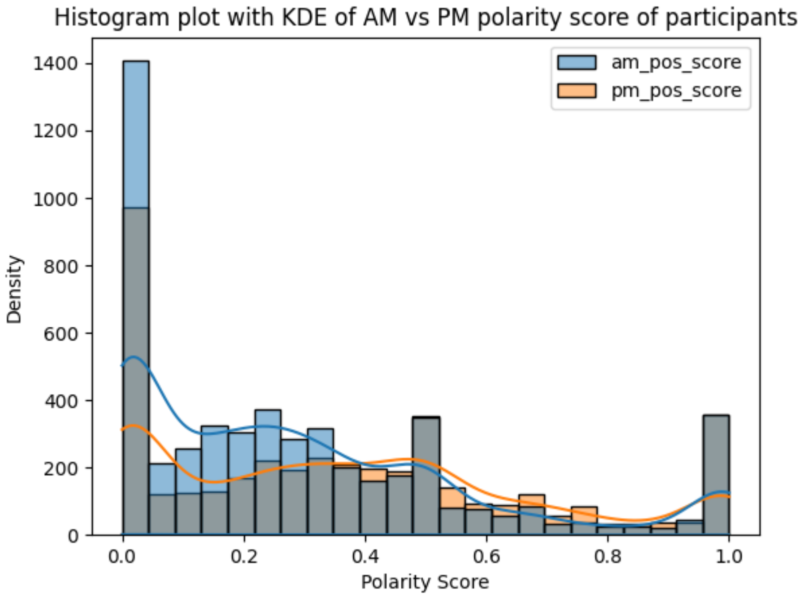


Figure 4.2: KDE plot of comparison of polarity scores of AM vs PM participants

whereas 4187 (74.2%) participants contributed 24,663 PM journals. The median (IQR) of AM and PM journals submitted per each participant were 2 (IQR=1-6) and 2 (IQR= 1-4), respectively. We defined the cutoff value of negative and positive sentiment was 0.5. Within the 39,103 AM journals, 27,666 (70.8%) were categorized into negative sentiment and 11,336 (29%) were positive sentiment whereas 14,391 (58.4%) were categorized into negative sentiment and 10,189 (41.3%) were positive sentiment for the 24,663 PM journals.

4.3.2 Topic Models

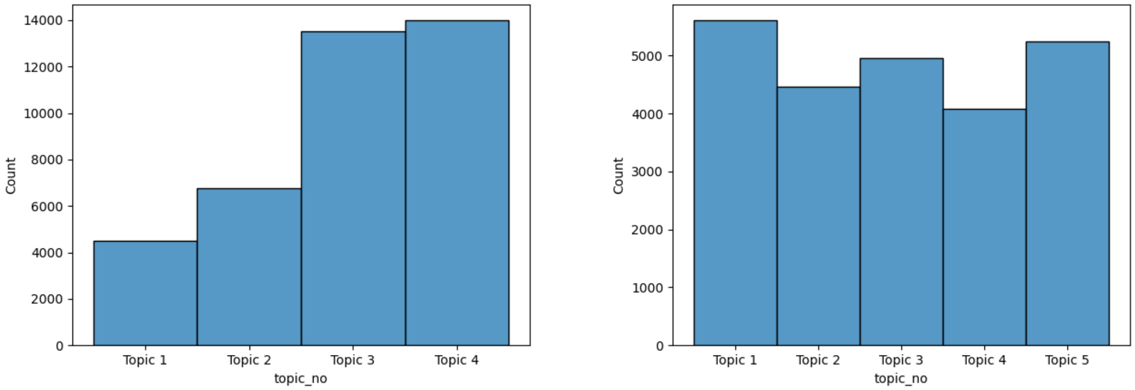
We extracted 3 positive topics and 2 negative topics for the PM journals, and 2 positive topics and 2 negative topics for the AM ones. The distribution of journals across topics can be observed in Figure 4.3

LIWC results used for interpretation are summarized across different topics by the weighted average of the LIWC scores with the weights of Topic scores.

PM journals:

Users answered the question “Before going to sleep, write a short description of your day, especially events that could affect your sleep”

In the positive topics, the first topic contains brief (6.08 WC (average word count provided by LIWC)) satisfactory reports about the day. This claim is also verified with highest scores



(a) AM journals (b) PM journals

Figure 4.3: Histogram of journal distribution across different topics

Topic	Theme	Highest LIWC Scores	Example
1	Concise satisfactory day report	‘sexual’, ‘assent’, ‘Lifestyle’, ‘attention’, ‘Affect’, ‘tone_pos’, ‘leisure’, ‘fulfill’, ‘Conversation’, ‘emotion’, ‘emo_pos’, ‘risk’, ‘relig’, ‘wellness’, ‘Period’, ‘Exclam’, ‘Clout’, ‘allure’	No problems tofay(99%)/ Very active outdoors. Napped afterward.(99%)/ No issues(99%)/ Thank God its Friday(99%)
2	Detailed positive productive day, Hopes, Interaction	‘family’, ‘AllPunc’, ‘food’, ‘motion’, ‘friend’, ‘affiliation’, ‘visual’, ‘ethnicity’, ‘Social’, ‘reward’, ‘article’, ‘want’, ‘male’, ‘prosocial’, ‘det’, ‘we’, ‘socrefs’, ‘Comma’, ‘female’, ‘Tone’, ‘Dic’, ‘polite’	No trouble staying awake. I hope I will do well in computer science. I hope I will be able to do better in math. I hope the tutor tomorrow will be able to help me understand what we’re learning in Calculus. I hope I will be able to finish chemistry lab and computer science on time. I hope my dad and I will be able to keep our plans for fall break and that mom will be able to take me to the airport.(99.7%)
3	Longest journals, CPAP need, Home	‘WC’, ‘they’, ‘focusfuture’, ‘focuspast’, ‘space’, ‘acquire’, ‘need’, ‘home’, ‘tech’, ‘Culture’, ‘cause’, ‘verb’, ‘conj’, ‘auxverb’, ‘prep’, ‘ipron’, ‘Apostro’, ‘shehe’, ‘i’, ‘function’, ‘pronoun’, ‘Linguistic’, ‘ppron’, ‘WPS’, ‘Authentic’, ‘you’	Went to work, shopping, got home at 8:15pm. Made a sandwhich and laid back on the sofa..fell asleepfor aboht 20 min, watched an episode of the Kennedys.. felt low energy. Turned TV off at 10:30 pm..when i remembered my agreement to turn off such things one hour before sleep. Now I will clean my cpap machine, complete my nightly routine, go to bed. Its 10:41 pm now(93%)
4	Most negative, Brief report, Mental health	‘moral’, ‘allnone’, ‘auditory’, ‘negate’, ‘swear’, ‘adj’, ‘quantity’, ‘Drives’, ‘achieve’, ‘comm’, ‘netspeak’, ‘curiosity’, ‘Analytic’, ‘memory’, ‘death’, ‘socbehav’, ‘emo_anx’, ‘emo_anger’, ‘time’, ‘emo_sad’, ‘money’, ‘BigWords’, ‘politic’, ‘mental’, ‘work’	Nothing out of the ordinary(86%)/Anxious day dealing woth attorney(83%)/ Slow boring depressed(65%)/ Boring day, panic attack took vistaril(63.5%)
5	Physical illness	‘health’, ‘nonflu’, ‘power’, ‘Cognition’, ‘focuspresent’, ‘adverb’, ‘cogproc’, ‘number’, ‘insight’, ‘Physical’, ‘discrep’, ‘tentat’, ‘certitude’, ‘differ’, ‘tone_neg’, ‘feeling’, ‘emo_neg’, ‘conflict’, ‘Perception’, ‘fatigue’, ‘lack’, ‘QMark’, ‘substances’, ‘illness’, ‘filler’	CPAP used setting 7 cmH2O Night meds. Norvasc 10 mg (blood pressure) Praminpexole .5 mg (leg movement during sleep) Altorvastain 10 mg (cholesterol) Celecoxib 200 mg (inflammation) Spinal stimulator: Off Bed room temp. Conditions: Dry Bulb Temp: 73 Deg f Dew point temp: 50 Deg f Care taker for spouse # times up during night: 3 # times up for bathroom: 1 Media on: yes(99%)

Table 4.1: Summary of PM journal topics

in “assent”, “Affect”, “tone_pos”, “fulfill”, “emotion”, “emo_pos”, and the top topics were “wellness”, “leisure”, “Lifestyle”, and “Sexual”. This topic also has the highest positive score

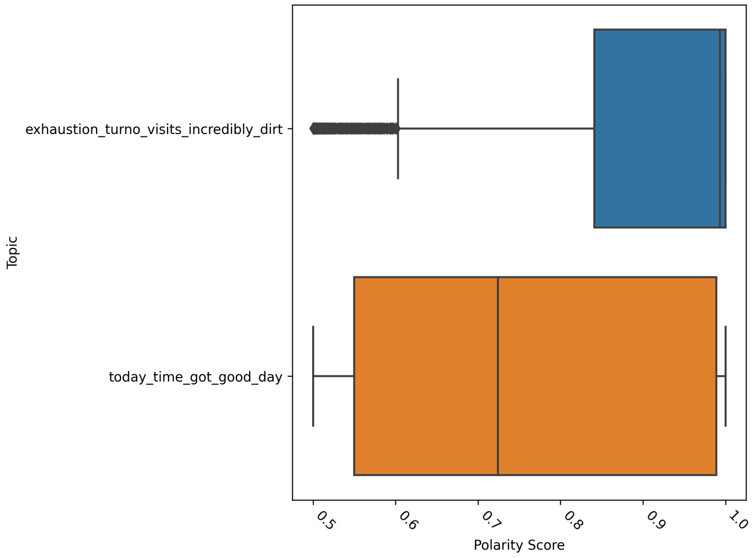
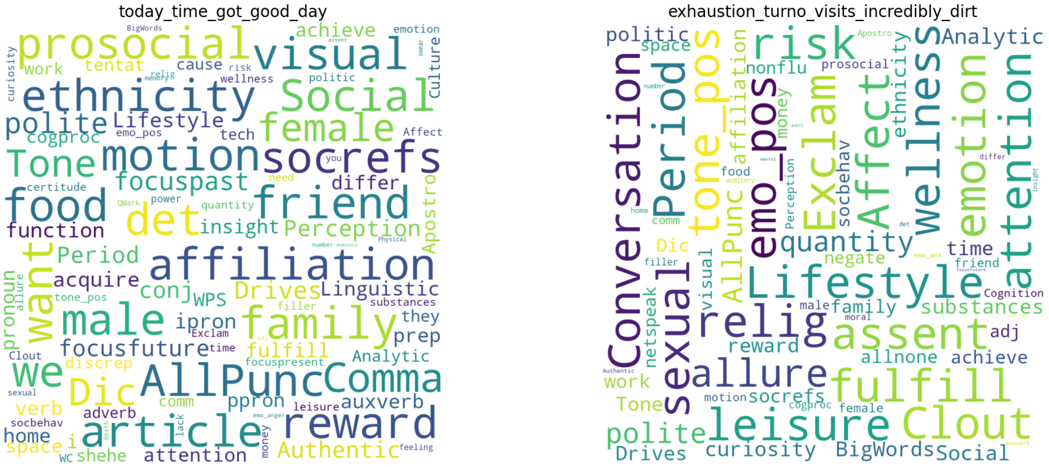


Figure 4.4: Polarity box plot of Positive topics of PM journals



(a) Topic 1

(b) Topic 2

Figure 4.5: Word cloud of Positive topics of PM journals

median (Figure 4.4, quartiles between 0.841 to 1 and a median of 0.993) among the 5 topics.

In the second topic, people reported a productive (highest scores in “motion” and “reward”) positive day with more detail (25.10 WC) and also some hopes for the future (“focusfuture” is about 4 times more than the previous topic). Also, the highest scores in “family”, “friend”, “Social”, “prosocial”, “male”, and “female”, show that they were also talking about their relatives and the interaction between them. The sentiment plot (Figure 4.4, quartiles between 0.55

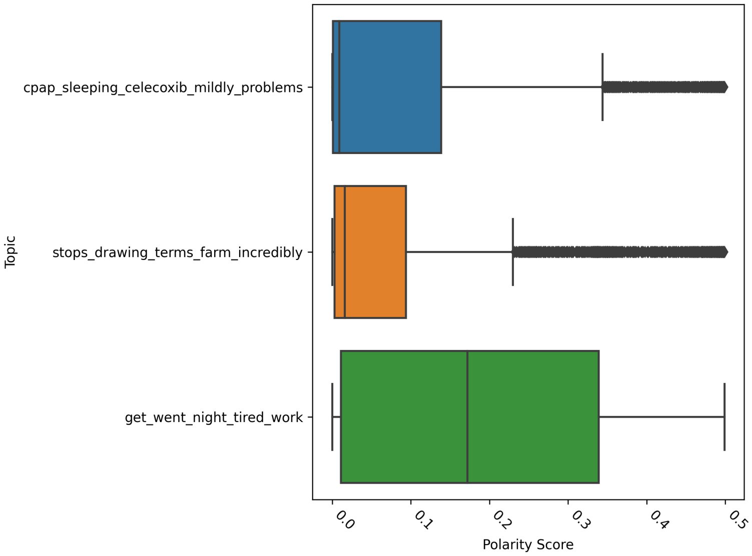
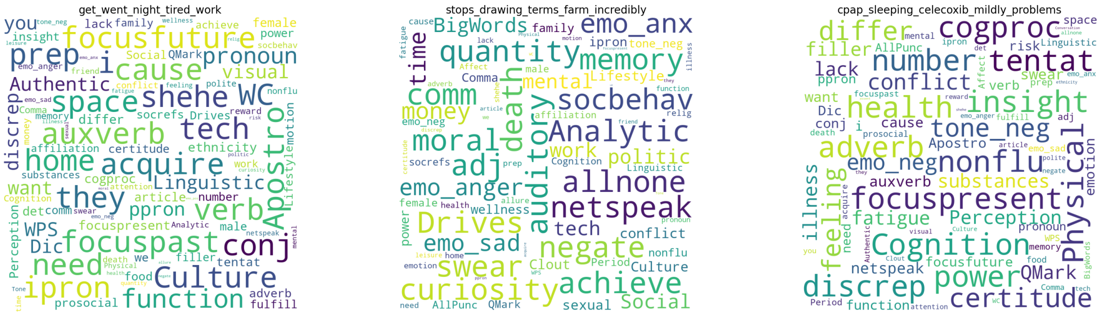


Figure 4.6: Polarity box plot of Negative topics of PM journals



(a) Topic 3

(b) Topic 4

(c) Topic 5

Figure 4.7: Word cloud of Negative topics of PM journals

to 0.989 and median 0.724) shows a wide range of positive journals which may also consist of negative sentences.

Topic 3 consists of the longest journals (43.59 WC) which report about the routine, with the highest focus on “future” and “past”. The highest score in “tech” may also refer to the CPAP machine, and the highest “need” score magnifies the tasks that have to be done. The focus is mostly in the “home” category. The wide sentiment plot (Figure 4.6, quartiles between 0.011 to 0.339 and a median of 0.172) shows that the journals’ polarities are more diverse than the other PM negative topics.

The 4th topic, which is mostly a negative report, describes the quality of the day in a few words (7.47 WC). The conciseness of the reports can also be perceived by the highest “An-

Topic	Theme	Highest LIWC Scores	Example
1	Detailed positive last night sleep	’Apostro’, ‘certitude’, ‘time’, ‘Drives’, ‘affiliation’, ‘Cognition’, ‘cogproc’, ‘insight’, ‘tentat’, ‘risk’, ‘focusfuture’, ‘reward’, ‘want’, ‘money’, ‘Social’, ‘socbehav’, ‘polite’, ‘ethnicity’, ‘focuspresent’, ‘achieve’, ‘you’, ‘ipron’, ‘Period’, ‘Dic’, ‘AllPunc’, ‘we’	Fell asleep right away and slept fine. Woke up ince during thr night (don\’t know why) but immediately went back to bed. It is Columbus day so I\’m off work. Went to bed at 11:15 and naturally woke up at 4:45. Feel fine and not tired, though I\’m sure I\’ll do my usual “I don\’t have to go to work this morning routine” of waking up, having my coffee, checking email/internet time, then going back to bed for a bit. (99.8%)
2	Brief report about Satisfactory Sleep	’moral’, ‘Conversation’, ‘emo_pos’, ‘emotion’, ‘tone_pos’, ‘Affect’, ‘fulfill’, ‘allure’, ‘adj’, ‘verb’, ‘adverb’, ‘Physical’, ‘focuspast’, ‘politic’, ‘family’, ‘netspeak’, ‘filler’, ‘Clout’, ‘Exclam’, ‘Tone’, ‘sexual’, ‘assent’, ‘nonflu’	Slept ok(96.5%)/ Slept fine.(95.8%)/ I slept great(95.1%)
3	Longest reports, several woke ups and some because of the CPAP	’WC’, ‘cause’, ‘auxverb’, ‘attention’, ‘Perception’, ‘acquire’, ‘need’, ‘wellness’, ‘home’, ‘discrep’, ‘number’, ‘tech’, ‘differ’, ‘memory’, ‘Culture’, ‘male’, ‘female’, ‘socrefs’, ‘prep’, ‘conj’, ‘article’, ‘WPS’, ‘det’, ‘Linguistic’, ‘ppron’, ‘i’, ‘function’, ‘space’, ‘motion’, ‘Authentic’, ‘shehe’, ‘Comma’, ‘they’, ‘pronoun’	Slept Good overall except didnt get to bed until midnight due to cleaning the house and getting home late. Made myself wear the cpap mask and i woke up 2:15am due to not being used to it. Made myself put it back. usually that is when i take it off for the remainder of night. I fell back asleep well and woke again around 4am because my cat woke me up, went to bathroom and put mask back on, fell asleep well. My Daughter woke me up at 7am and i got up then. Was very tired due to not enough sleep, but could tell i had a littlr clearer thinking from using cpap machine. (99.1%)
4	exhausted day with sleepiness or illness signs	’curiosity’, ‘quantity’, ‘power’, ‘allnone’, ‘substances’, ‘mental’, ‘illness’, ‘feeling’, ‘tone_neg’, ‘auditory’, ‘emo_neg’, ‘emo_anx’, ‘emo_anger’, ‘emo_sad’, ‘swear’, ‘visual’, ‘prosocial’, ‘health’, ‘conflict’, ‘Analytic’, ‘comm’, ‘friend’, ‘Lifestyle’, ‘leisure’, ‘work’, ‘food’, ‘relig’ , ‘negate’, ‘QMark’, ‘BigWords’, ‘fatigue’, ‘lack’, ‘death’	Headache at 330am!(97.5%)/I passed out from my pain medication(97.4%)/ Super sleepy and passed out.(97.3%) /Sleepy(97.3%)/ Exhausted(97.2%)

Table 4.2: Summary of AM journal topics

alytic” score. The highest “negate” score verifies that in many of this topic’s samples, users tend to use phrases such as “Nothing to report”, “Nothing particular”, etc. The highest scores in ‘’swear”, “moral” , “curiosity”, “emo_anx”, “emo_anger”, “emo_sad” “memory”, and especially “mental” demonstrate that in these journals people are negatively obsessed with topics that disrupt their mental stability. The topics that they are obsessed with can also be understood by the highest scores in “achieve”, “death”, “time”, “money”, “politic”, and “work”. The narrowest negative sentiment plot (Figure 4.6, quartiles between 0.003 to 0.094 and a median of 0.016) also signifies the negativity of this topic.

In the last topic, topic 5, people talk about sleeping problems or the way they dealt with that (14.35 WC). This topic is mostly focused on physical health problems which can be perceived with the highest scores in “health”, “Physical”, “fatigue”, and “illness”. Their negative feedback can also be observed by the highest scores in “tone_neg”, “feeling”, “lack”, “conflict”, and “emo_neg”. The sentiment box plot (Figure 4.6, quartiles between 0.001 to 0.1385 and a median of 0.009) has the lowest median among the 5 topics.

AM journals:

Users answered this question: “Describe, in as much detail as you wish, how your sleep was last night.”

The first topic reported positively about the last night sleep with detail (19.10 WC). This

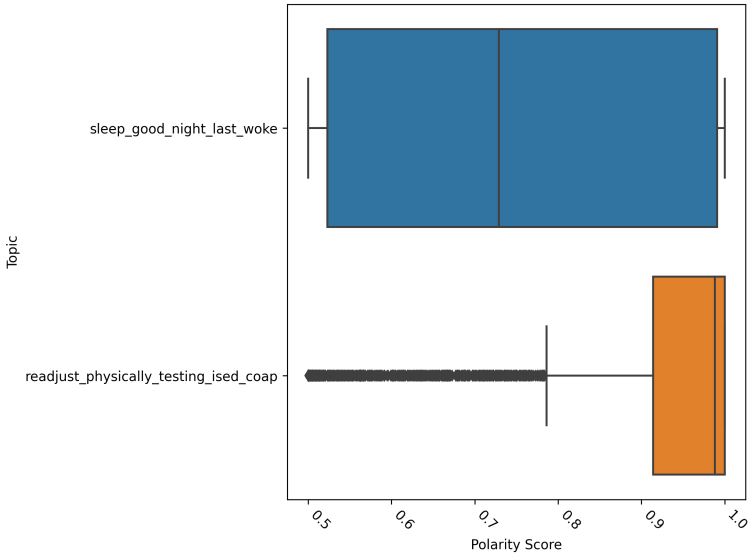
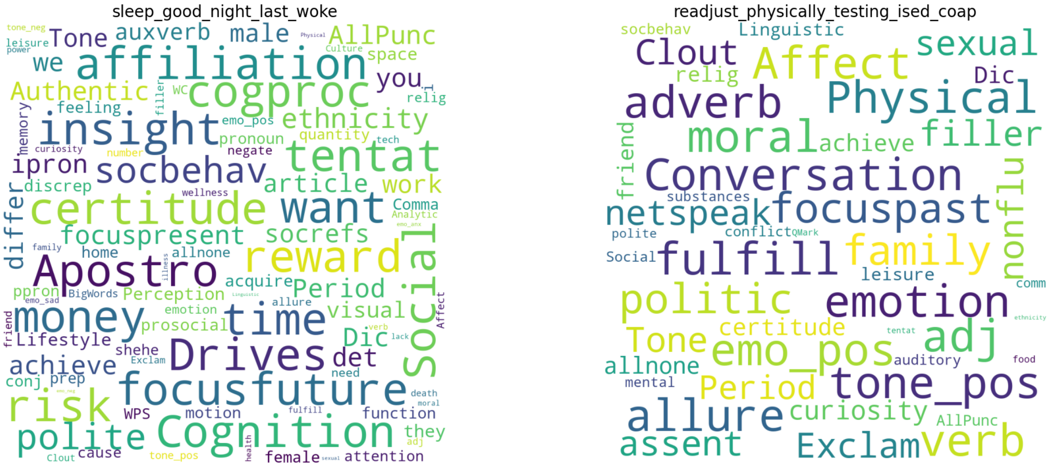


Figure 4.8: Polarity box plot of Positive topics of AM journals



(a) Topic 1

(b) Topic 2

Figure 4.9: Word cloud of Positive topics of AM journals

detail can also be perceived by the high score in “time” of the LIWC results. We notice that they also report about “future” and “present” when the journal was referring to “social”, “achieve”, “reward”, “want”, and “money”. Also, the high score in “Cognition” reflects different ways people think or refer to their thinking about their sleep quality. The wide sentiment box plot in the positive range (Figure 4.8, 0.523 to 0.991 quartiles with a median of 0.729) shows that the details also consists of negative sentences.

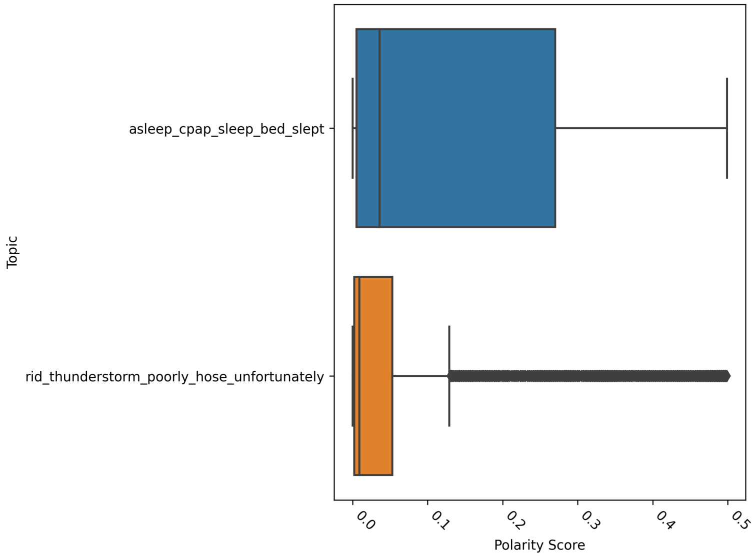
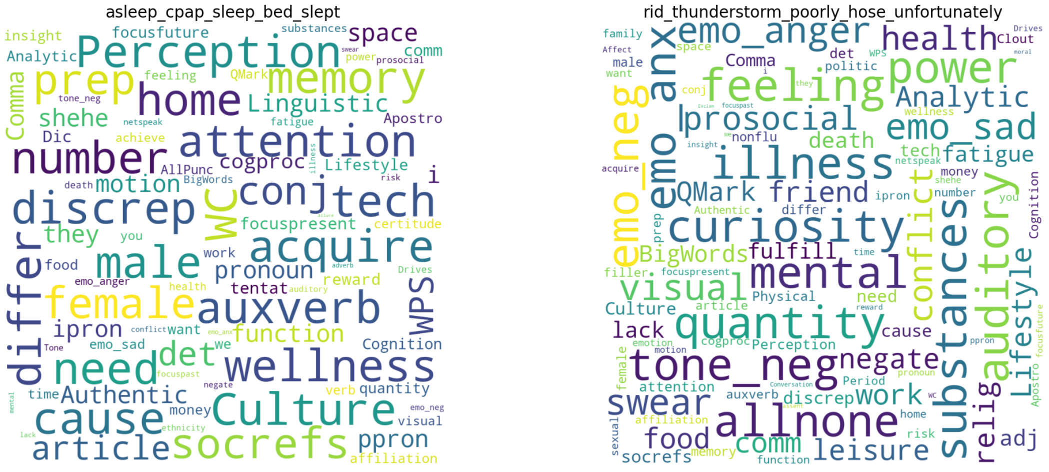


Figure 4.10: Polarity box plot of Negative topics of AM journals



(a) Topic 3

(b) Topic 4

Figure 4.11: Word cloud of Negative topics of AM journals

The second topic briefly reported a satisfactory sleep (4.41 WC). This can be verified by high amounts in “Tone”: “tone_pos”, “emo_pos”, “emotion”, “affect”, “assent” and “fulfill”. The high score in “focuspast” also intensifies their conciseness in their answer to the question. The narrow positive sentiment box plot (Figure 4.8, 0.914 to 1.0 quartiles with a median of 0.988) also magnifies their focus on answering the question without deviation from the main topic. Reports with top scores are as short as “Slept ok”.

Regarding the negative topics, in topic 3, which consists of the longest reports (36.23 WC), participants reported several wake ups, some of which were because of the CPAP machine. The reason can also derived from the high score in “Causation”, and “Tech” which may refer to the CPAP machine. The high scores in “Linguistic”, “home”, and “motion” can also be interpreted as the times the participant returned to bed and fell back to sleep. The sentiment box plot (Figure 4.10, 0.005 to 0.27 quartiles with a median of 0.036) illustrates that although the negative sentences of the journals are more numerous, the positive ones also exist.

The last negative topic, topic 4, briefly (9.24 WC) reports an exhausted day with signs of sleepiness or illness. High scores in “emo_neg”, “emo_anger”, “emo_sad”, and “fatigue” shows the negative sentiments and “mental”, “illness”, “health”, “death” shows the topics that they are expressing the feelings about. As expected, the narrow negative box plot (Figure 4.10, 0.002 to 0.053 quartiles with a median of 0.009) verifies that the most negative comments are focused on this topic.

4.4 Limitations

Although the approaches used in the experiments are quite novel and achieved state-of-the-art in their area, they still have some disadvantages, especially when it comes to real-life messy data.

The first limitation is regarding the sentiment analysis pipeline. The DistilBERT model used was fine-tuned on SST-2, which is the binary classification version of the Stanford Sentiment Treebank. It tries to put all sentences into one of the positive or negative classes. Hence, when a neutral sentence is passed through the pipeline, it tends to be placed on one of the two ends of the positive/negative spectrum. For example, “Got a haircut”, “Back to work tomorrow”, and “Mailed a controller today” got 0.095, 0.995, and 0.017 polarity scores while they sound like neutral sentences.

The second disadvantage was the misclassification problem, which is always expected from an ML model. Understanding the sentences, especially the ones with vague meaning, is difficult even for humans. Suppose the sentence “Too warm even with blanket off and window open”. The word “warm” plays the most important role here. It can be perceived both as ‘friendly’ which is a positive word and ‘high temperature’ which is a negative word. In the model we used, this sentence is classified as a positive sentence with a 0.822 polarity score. If we clarify the sentence by adding “The weather” at its beginning, The result becomes 0.117 which is a complete negative score. It seems that the model cannot disambiguate the “warm” with the given amount of information. We also tried to test “Too warm weather” and “Too warm person” which resulted in positive and negative sentences, respectively.

The next two challenges that we faced can be explained with this example: “Not good heart irregular headache used coap but had osa of 39 seconds and csa too. 12 total”. This report contains two sentences which got one negative and one positive polarity score, respectively, and positive in total. The reason “12 total” is classified as a positive sentence is that it is considered as an independent sentence while the meaning of the previous sentence is crucial for figuring out the meaning of the latter one. The other problem concerning this and many other examples was the high number of typo mistakes which will result in increasing the number of unknown tokens for LIWC, Topic Modeling, and also the Polarity detection pipeline. In this case, “COAP” should be corrected to “cpap” and “ised” is likely to be “used”. In future work, we will try to spell check and do typo correction before applying the models.

The last drawback concerns the Topic Modeling pipeline. Although the model used outperforms the LDA model on benchmark datasets, it still has much to do to get sensible, coherent topics which capture a broad area and be unrelated to the other topics. In our experiments, we found that the pattern of writing and the length of the reports had a remarkable impact on gathering reports in a topic though the intention of using Topic Modeling was to cluster topics based on the subject they are talking about more than the syntactic patterns.

4.5 Conclusion

Psychosocial features generated from sleep diaries show notable differences in participants with varying sleep and daily functioning. Future work will include understanding key sleep-health topics represented in sleep diaries that determine personalized long-term outcomes in sleep health.

Chapter 5

Conclusions and Future Work

This thesis looks at the capability of state-of-the-art models in the field of Natural Language Processing to understand and extract information from people's self-reporting journals. Pre-trained deep learning models are used to benefit from the information learned based on the huge amount of available training data and to utilize it on journal datasets with limited size.

The main contributions of this thesis are:

1. Investigating the capability of current state-of-the-art models in capturing psychological information of personality detection journals.

2. Building a novel approach which captures psychological information from personality detection journals and improves the state-of-the-art accuracy in document classification in this field.

3. Investigating the sentiment of people-provided sleep diaries using a pre-trained deep learning-based sentiment analysis model. The deep learning model is basically a multi-layer neural network which can extract more complex patterns. The deep learning model used in this thesis is previously trained on a huge amount of sentiment classification data which tries to predict the polarity (positive/negative) of input texts based on their semantic structure.

4. Applying contextualized topic modeling to classify sleep journals into categories based on the topic they are talking about.

5. Using a novel approach to analyze and interpret extracted topics and to provide insight for each journal category.

With respect to the first contribution, we introduced an approach to evaluate embeddings in personality detection by including psychological information in pure AI driven models. We conclude that the current state-of-the-art model, though it achieved the highest accuracy in the field, is not using a psychology-aware embedding.

Regarding the second contribution, we benefited from InferSent, which is inspired from the Siamese architecture [26], to improve the psychological insight embedded in the features extracted from participants' journals. We also evaluated SentenceBERT [96] using the evaluation metric mentioned in the first contribution. Results showed the metric we used for evaluation has a significant correlation with the accuracy of the model on the Essays dataset.

Concerning the SleepHealth project, we realized that although participants tend to write negative journals more than positive ones, they wrote more positively in the PM journals than in the AM journals, which refers to our third contribution.

For the fourth contribution, we extracted 3 positive topics and 2 negative topics for the PM journals, and 2 positive topics and 2 negative topics for the AM ones. The number of topics were optimized both by manual inspection and using evaluation using coherence value.

In order to analyze the theme of each topic and to provide more insight, LIWC software was used [87]. Different psycholinguistics aspects of the journals within each topic were investigated and the interpretations were obtained based on the top examples within each topic and the top LIWC scores achieved in each topic.

5.1 Future Work

In a nutshell, this thesis tries to provide insight based on applying state-of-the-art pre-trained NLP models on real-world psychological journals. This insight can be obtained in different forms, from improving the quality of features extracted from psychological texts to obtain higher accuracy in predicting people's personality to automatically extracting journals' polarities and investigating the topics each one is talking about. Although the improvements and results have shown considerable capability of these models, their limitations brings us the idea that maybe more manual inspection and involvement of psychological ideas are required in order to achieve better results.

Since the results provided in Table 3.2 showed that recent general pre-trained models achieved higher accuracies than even the fine-tuned previous state-of-the-art model, we can assume that considering the small amount of data in this field, using recent pre-trained NLP models will bring us better results, hence, one direction of future work will be investigating more pre-trained models and how they are performing on the personality detection task. The other prospective direction regarding the personality detection could be applying and evaluating the same approach on other domains of psychology to see whether we can achieve significant correlation between the accuracy of those domains' datasets and the PredLabel accuracies we introduced using domain specific self-report questionnaires. Newer models can be leveraged to evaluate their capability in understanding the psychological and cognitive abilities of humans. This evaluation can be assessed by using the same approach as it is discussed in this thesis such as introducing two evaluation metrics or following a similar method to psychological approaches in designing linguistic questionnaires to understand individuals' personality features such as lexical hypothesis [35].

Focusing on the Sleep Health project, a prospective approach is to use a better fine-grained polarity detection model which at least considers the neutral polarity class too. It is also realized that typo errors disrupt the performance of deep-learning based models and hence trying to use a typo correction method in further investigations would be an opportunistic future work. Also regarding topic modeling, taking advantage of different recent models such as Top2Vec [7] may result in a better performance outcome in terms of extracting more meaningful topics and circumventing the extra intermediate steps such as finding the optimum number of topics, etc. The other prospective approach can be the comparison between the top words represented by the topic modeling model and the LIWC outcome to see how they are related to each other. In this study, the top words of the topic modeling could not bring us enough information to understand the context of each topic.

Bibliography

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[2] Joel Achenbach. Coronavirus is harming the mental health of tens of millions of people in U.S., new poll finds. https://www.washingtonpost.com/health/coronavirus-is-harming-the-mental-health-of-tens-of-millions-of-people-in-us-new-poll-finds/2020/04/02/565e6744-74ee-11ea-85cb-8670579b863d_story.html, 2020.

[3] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. DocBERT: BERT for document classification. *arXiv preprint arXiv:1904.08398v3*, 2019.

[4] James Allen. *Natural Language Understanding*. Benjamin-Cummings Publishing Co., Inc., 1988.

[5] Sonia Ancoli-Israel, Roger Cole, Cathy Alessi, Mark Chambers, William Moorcroft, and Charles P Pollak. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*, 26(3):342–392, 2003.

[6] Kirstie N Anderson and Andrew J Bradley. Sleep disturbance in mental health problems and neurodegenerative disease. *NSS News*, 5:61–75, 2013.

[7] Dimo Angelov. Top2Vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.

[8] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[9] American Sleep Association. Sleep statistics: Data about sleep disorders. https://www.sleepassociation.org/about-sleep/sleep-statistics/, 2016.

[10] Lisa Feldman Barrett. Feelings or words? understanding the content in self-report ratings of experienced emotion. *Journal of personality and social psychology*, 87(2):266, 2004.

[11] Verónica Benet-Martínez and Oliver P John. Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, 75(3):729, 1998.

[12] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, 2021.

[13] Sage Bionetworks. Synapse. https://www.synapse.org/#!Synapse:syn18492837/wiki/590797. Accessed: 2022-4-5.

[14] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[15] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[16] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[17] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. The development and psychometric properties of LIWC-22. Austin, TX: University of Texas at Austin, https://www.liwc.app, 2022.

[18] I. Briggs Myers. *Introduction to Type: A Guide to Understanding Your Results on the Myers-Briggs Type Indicator (revised by L K Kirby & K D Myers)*. CA: Consulting Psychologists Press, 1993.

[19] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[20] Daniel J Buysse. Sleep health: Can we define it? Does it matter? *Sleep*, 37(1):9–17, 2014.

[21] Erik Cambria, Yang Li, Frank Z Xing, Soujanya Poria, and Kenneth Kwok. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM'20)*, pages 105–114, 2020.

[22] Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80, 2017.

[23] Laura Campbell-Sills, Sharon L Cohan, and Murray B Stein. Relationship of resilience to personality, coping, and psychiatric symptoms in young adults. *Behaviour Research and Therapy*, 44(4):585–599, 2006.

[24] Colleen E Carney, Daniel J Buysse, Sonia Ancoli-Israel, Jack D Edinger, Andrew D Krystal, Kenneth L Lichstein, and Charles M Morin. The consensus sleep diary: standardizing prospective sleep self-monitoring. *Sleep*, 35(2):287–302, 2012.

[25] Michael A Cohn, Matthias R Mehl, and James W Pennebaker. Linguistic markers of psychological change surrounding september 11, 2001. *Psychol. Sci.*, 15(10):687–693, 2004.

[26] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017.

[27] Paul T Costa Jr and Robert R McCrae. Domains and facets: Hierarchical personality assessment using the revised neo personality inventory. *Journal of Personality Assessment*, 64(1):21–50, 1995.

[28] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, 2015.

[29] Sean Deering, Abhishek Pratap, Christine Suver, A Joseph Borelli, Jr, Adam Amdur, Will Headapohl, and Carl J Stepnowsky. Real-world longitudinal data collected from the SleepHealth mobile app study. *Sci Data*, 7(1):418, 2020.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, 2019.

[31] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, 2018.

[32] Alyssa Fowers and William Wan. A third of Americans now show signs of clinical anxiety or depression, Census Bureau finds amid coronavirus pandemic. https://www.washingtonpost.com/health/2020/05/26/americans-with-depression-anxiety-pandemic/?arc404=true, 2020.

[33] Tim Gawley. Using solicited written qualitative diaries to develop conceptual understandings of sleep. *International Journal of Qualitative Methods*, 17(1):160940691879425, 2018.

[34] David P Goldberg and Peter Huxley. *Common Mental Disorders: A Bio-social Model*. Tavistock/Routledge, 1992.

[35] Lewis R Goldberg. The structure of phenotypic personality traits. *American Psychologist*, 48(1):26, 1993.

[36] Lewis R Goldberg. International personality item pool: A scientific collaboratory for the development of advanced measures of personality and other individual differences. https://ipip.ori.org, 1999.

[37] Hannes Grassegger and Mikael Krogerus. The data that turned the world upside down. *Vice Motherboard*, 28, 2017.

[38] Robert J. Gregory. The history of psychological testing. In *Psychological Testing: History, principles, and applications*, chapter 2, pages 32–58. Pearson, 7th edition, 2013.

[39] Akash Gupta, Shrey Aeron, Anjali Agrawal, and Himanshu Gupta. Trends in COVID-19 publications: Streamlining research using NLP and LDA. *Front. Digit. Health*, 3, 2021.

[40] Harvard Health. Sleep and mental health - sleep deprivation can affect your mental health. https://www.health.harvard.edu/newsletter_article/sleep-and-mental-health, 2021.

[41] Jenny Hislop, Sara Arber, Rob Meadows, and Sue Venn. Narratives of the night: The use of audio diaries in researching sleep. *Sociological Research Online*, 10(4):13–25, 2005.

[42] Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312, 2021.

[43] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, 2018.

[44] Vanessa Ibáñez, Josep Silva, and Omar Cauli. A survey on sleep questionnaires and diaries. *Sleep medicine*, 42:90–96, 2018.

[45] Hamed Jelodar, Rita Orji, Stan Matwin, Swarna Weerasinghe, Oladapo Oyebode, and Yongli Wang. Artificial intelligence for Emotion-Semantic trending and people emotion detection during COVID-19 social isolation. *arXiv preprint arXiv:2101.06484*, 2021.

[46] Hamed Jelodar, Yongli Wang, Mahdi Rabbani, Gang Xiao, and Ruxin Zhao. A collaborative framework based for semantic Patients-Behavior analysis and highlight topics discovery of alcoholic beverages in online healthcare forums. *Journal of Medical Systems*, 44(5):101, 2020.

[47] Oliver P John, Eileen M Donahue, and Robert L Kentle. Big Five Inventory. *Journal of Personality and Social Psychology*, 1991.

[48] Oliver P. John, Laura P. Naumann, and Cristopher J. Soto. Paradigm shift to the integrative Big Five Trait taxonomy. In *Handbook of Personality: Theory and Research*, pages 114–158. Guilford Press, 2008.

[49] Deborah G Johnson. *Computer Ethics*. Wiley Online Library, Englewood Cliffs (NJ), 1985.

[50] Mitchell Jolly. (MBTI) Myers-Briggs personality type dataset, Sep 2017.

[51] Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143, 2014.

[52] Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, 2020.

[53] Yarden Katz. Noam Chomsky on Where Artificial Intelligence went wrong. https://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/, 2017. Accessed: 2022-4-5.

[54] Mahmut Kaya and Hasan Sakir Bilge. Deep metric learning: A survey. *Symmetry*, 11:1066, 2019.

[55] Amirmohammad Kazameini, Samin Fatehi, Yash Mehta, Sauleh Eetemadi, and Erik Cambria. Personality trait detection using bagged svm over bert word embedding ensembles. *arXiv preprint arXiv:2010.01309*, 2020.

[56] Amirmohammad Kazemeini, Sudipta Singha Roy, Robert E Mercer, and Erik Cambria. Interpretable representation learning for personality detection. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 158–165, 2021.

[57] Kenneth D Kochanek, Sherry L Murphy, Jiaquan Xu, and Elizabeth Arias. Mortality in the United States, 2013. *NCHS Data Brief*, 178:1–8, 2014.

[58] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[59] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, 2019.

[60] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, (ICLR)*, 2020.

[61] Katherine Lanyi, Rhiannon Green, Dawn Craig, and Christopher Marshall. COVID-19 vaccine hesitancy: Analysing twitter to identify barriers to vaccination in a low uptake region of the UK. *Front. Digit. Health*, 3, 2022.

[62] Rayson Laroca, Evair Severo, Luiz A Zanlorensi, Luiz S Oliveira, Gabriel Resende Gonçalves, William Robson Schwartz, and David Menotti. A robust real-time automatic license plate recognition based on the yolo detector. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, 2018.

[63] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[64] Richard Lewis. *When Cultures Collide*. Nicholas Brealey Publishing, Boston, MA, 2010.

[65] Cynthia CS Liem, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, and Cornelius J König. Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 197–253. Springer, 2018.

[66] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[67] Yong Liu, Anne G Wheaton, Daniel P Chapman, Timothy J Cunningham, Hua Lu, and Janet B Croft. Prevalence of healthy sleep duration among adults — United States, 2014. *MMWR. Morbidity and Mortality Weekly Report*, 65(6):137–141, 2016.

[68] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007.

[69] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79, 2017.

[70] Jose Maria Balmaceda, Silvia Schiaffino, and Daniela Godoy. How do personality traits affect communication among users in online social networks? *Online Information Review*, 38(1):136–153, 2014.

[71] Gerald Matthews, Ian J Deary, and Martha C Whiteman. *Personality Traits*. Cambridge University Press, 2003.

[72] Sandra C Matz, Michal Kosinski, Gideon Nave, and David J Stillwell. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48):12714–12719, 2017.

[73] Stéphanie Mazza, Hélène Bastuji, and Amandine E Rey. Objective and subjective assessments of sleep in children: Comparison of actigraphy, sleep diary completed by children and parents' estimation. *Front. Psychiatry*, 11, 2020.

[74] Catherine McCall and W Vaughn McCall. Comparison of actigraphy with polysomnography and sleep logs in depressed insomniacs. *J. Sleep Res.*, 21(1):122–127, February 2012.

[75] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation. https://dl.acm.org/doi/abs/10.5555/3295222.3295377. Accessed: 2022-4-5.

[76] Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. Bottom-Up and Top-Down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189, 2020.

[77] Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, pages 1–27, 2019.

[78] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.*, 26, 2013.

[79] Roberto Navigli. Word sense disambiguation. *ACM Computing Surveys*, 41(2):1–69, 2009.

[80] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. Lying words: predicting deception from linguistic styles. *Pers. Soc. Psychol. Bull.*, 29(5):665–675, 2003.

[81] Teryl K Nuckols, Jay Bhattacharya, Dianne Miller Wolman, Cheryl Ulmer, and José J Escarce. Cost implications of reduced work hours and workloads for resident physicians. *N. Engl. J. Med.*, 360(21):2202–2215, 2009.

[82] World Health Organization. Constitution of the World Health Organization. https://www.who.int/about/governance/constitution, 2022.

[83] World Health Organization et al. WHO technical meeting on sleep and health: Bonn Germany, 22–24 January 2004. Technical Report WHO/EURO:2004-4242-44001-62044, World Health Organization. Regional Office for Europe, 2004.

[84] Jeff Orlowski. The social dilemma. https://www.thesocialdilemma.com/, 2021.

[85] Daniel J. Ozer and Verónica Benet-Martínez. Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, 57(1):401–421, 2006.

[86] Praveetha Patalay and Suzanne H Gage. Changes in millennial adolescent mental health and health-related behaviours over 10 years: a population cohort comparison study. *International Journal of Epidemiology*, 48(5):1650–1664, 2019.

[87] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin, 2015.

[88] James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. When small words foretell academic success: the case of college admissions essays. *PLoS One*, 9(12):e115844, 2014.

[89] James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296, 1999.

[90] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[91] Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, 2017.

[92] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

[93] David J Pittenger. Cautionary comments regarding the Myers-Briggs type indicator. *Consulting Psychology Journal: Practice and Research*, 57(3):210, 2005.

[94] Liam Porr. A robot wrote this entire article. are you scared yet, human? | gpt-3. https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3, 2020.

[95] Nils Reimers. Pretrained models. SBERT.net, https://www.sbert.net/docs/pretrained_models.html, 2020.

[96] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[97] 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*, pages 1135–1144, 2016.

[98] Brent W Roberts, Nathan R Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R Goldberg. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, 2(4):313–45, 2007.

[99] Mark R Rosekind, Kevin B Gregory, Melissa M Mallis, Summer L Brandt, Brian Seal, and Debra Lerner. The cost of poor sleep: Workplace productivity loss and associated costs. *J. Occup. Environ. Med.*, 52(1):91–98, 2010.

[100] Alexandra Samet. 2020 US social media usage: How the coronavirus is changing consumer behavior. https://www.businessinsider.com/2020-us-social-media-usage-report, 2020.

[101] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *EMC2: 5th Edition Co-located with NeurIPS'19*, 2019.

[102] Hanna Schneider, Katrin Schauer, Clemens Stachl, and Andreas Butz. Your data, your vis: Personalizing personal data visualizations. In *Proceedings of IFIP Conference on Human-Computer Interaction – INTERACT 2017*, pages 374–392, 2017.

[103] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[104] M A Short, M Gradisar, L C Lack, H Wright, and M A Carskadon. The discrepancy between actigraphic and sleep diary measures of sleep in adolescents. *Sleep Med.*, 13(4):378–384, 2012.

[105] Danica C Slavish, Daniel J Taylor, and Kenneth L Lichstein. Intraindividual variability in sleep and comorbid medical and mental health conditions. *Sleep*, 42(6), 2019.

[106] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, October 2013.

[107] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. *Adv. Neural Inf. Process. Syst.*, 33:16857–16867, 2020.

[108] Clemens Stachl, Ryan L Boyd, Kai T Horstmann, Poruz Khambatta, Sandra Matz, and Gabriella M Harari. Computational personality assessment-an overview and perspective. *PsyArXiv*, 2021.

[109] Lorenzo Tonetti, Roberto Mingozzi, and Vincenzo Natale. Comparison between paper and electronic sleep diary. *Biological Rhythm Research*, 47(5):743–753, 2016.

[110] Tudor Văcărețu, Nikolaos Batalas, Begum Erten-Uyumaz, Merel van Gilst, Sebastiaan Overeem, and Panos Markopoulos. Subjective sleep quality monitoring with the hypnos digital sleep diary: Evaluation of usability and user experience. In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 113–122, 2019.

[111] Alexandros N Vgontzas, Duanping Liao, Slobodanka Pejovic, Susan Calhoun, Maria Karataraki, Maria Basta, Julio Fernández-Mendoza, and Edward O Bixler. Insomnia with short sleep duration and mortality: the penn state cohort. *Sleep*, 33(9):1159–1164, 2010.

[112] Prashanth Vijayaraghavan, Eric Chu, and Deb Roy. DAPPER: Learning domain-adapted persona representation using pretrained BERT and external memory. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 643–652, 2020.

[113] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Quay Au, Bernd Bischl, Markus B"uhner, and Heinrich Hussmann. Opportunities and Challenges of Utilizing Personality Traits for Personalization in HCI: Towards a shared perspective

from HCI and Psychology. In *Personalized Human-Computer Interaction*, pages 31–64. De Gruyter, Oldenbourg, Germany, 2019.

[114] Julia Walsh, Jonathan Cave, and Frances Griffiths. Spontaneously generated online patient experience of modafinil: A qualitative and NLP analysis. *Front. Digit. Health*, 3, 2021.

[115] COVID-19 disrupting mental health services in most countries, WHO survey. https://www.who.int/news/item/05-10-2020-covid-19-disrupting-mental-health-services-in-most-countries-who-survey, 2020.

[116] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[117] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[118] Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1725–1734, 2020.

[119] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.*, 32, 2019.

[120] Han Yin, Yue Wang, Qian Li, Wei Xu, Ying Yu, and Tao Zhang. A network-enhanced prediction method for automobile purchase classification using deep learning. In *Proceedings of Pacific Asia Conference on Information Systems (PACIS)*, page 111, 2018.

[121] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.

[122] Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015.

[123] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. *Adv. Neural Inf. Process. Syst.*, 33:17283–17297, 2020.

[124] Jingwen Zhang, Haoning Xue, Christopher Calabrese, Huiling Chen, and Julie H T Dang. Understanding human papillomavirus vaccine promotions and hesitancy in northern California through examining public facebook pages and groups. *Front. Digit. Health*, 3, 2021.

Curriculum Vitae

Name:	Amirmohammad Kazemeinizadeh
Post-Secondary Education and Degrees:	Iran University of Science and Technology Tehran, Iran B.Sc. Computer Engineering, 2015-2020
	University of Western Ontario London, ON, Canada Master’s Studies in Computer Science, 2020-2022 Supervisor: Dr. Robert E. Mercer
Honours and Awards:	Western Graduate Research Scholarship 2020-2022
	Mitacs Globalink Research Award 2021
Related Work Experience:	Teaching Assistant The University of Western Ontario 2020-2022
	Intern Vector Institute 2021-2021
	Trainee Center for Addiction and Mental Health 2021-2022