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
© Amirmohammad Kazemeinizadeh 2022

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Recommended Citation

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.
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

Abstract .............................................. i

Summary for Lay Audience ................... ii

Acknowledgements ................................. iii

List of Figures .................................. vi

List of Tables .................................... vii

1 Introduction ................................... 1

2 RelatedWork .................................... 5

3 Project 1: Interpretable Representation Learning for Personality Detection 9
   3.1 Introduction .................................. 9
   3.2 Methodology .................................. 11
   3.3 Experiments .................................. 17
   3.4 Results ...................................... 18
   3.5 Conclusion .................................. 23

4 Project 2: SleepHealth Data Analysis 24
   4.1 Introduction .................................. 24
   4.2 Methods ...................................... 25
      4.2.1 Study Details ............................. 25
      4.2.2 Data Collection ......................... 26
      4.2.3 Data featurization ..................... 26
   4.3 Results ...................................... 27
      4.3.1 Sentiment Analysis ..................... 27
      4.3.2 Topic Models .................. 28
   4.4 Limitations .................................. 35
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>An example of a personalized recommender system</td>
<td>10</td>
</tr>
<tr>
<td>3.2</td>
<td>Visualization of the personality statements after applying PCA on the average</td>
<td>13</td>
</tr>
<tr>
<td>3.3</td>
<td>Architecture of the model with Siamese Bi-LSTM and max-pooling for the</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>interpretable tool for personality detection</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>Architecture of Sentence-BERT</td>
<td>16</td>
</tr>
<tr>
<td>3.5</td>
<td>Visualization of the personality statements after applying PCA on the feature</td>
<td>16</td>
</tr>
<tr>
<td>3.6</td>
<td>Visualization of the personality statements after applying PCA on the feature</td>
<td>16</td>
</tr>
<tr>
<td>4.1</td>
<td>Flow of the sleep health from data extraction to topic interpretation</td>
<td>25</td>
</tr>
<tr>
<td>4.2</td>
<td>KDE plot of comparison of polarity scores of AM vs PM participants</td>
<td>28</td>
</tr>
<tr>
<td>4.3</td>
<td>Histogram of journal distribution across different topics</td>
<td>29</td>
</tr>
<tr>
<td>4.4</td>
<td>Polarity box plot of Positive topics of PM journals</td>
<td>30</td>
</tr>
<tr>
<td>4.5</td>
<td>Word cloud of Positive topics of PM journals</td>
<td>30</td>
</tr>
<tr>
<td>4.6</td>
<td>Polarity box plot of Negative topics of PM journals</td>
<td>31</td>
</tr>
<tr>
<td>4.7</td>
<td>Word cloud of Negative topics of PM journals</td>
<td>31</td>
</tr>
<tr>
<td>4.8</td>
<td>Polarity box plot of Positive topics of AM journals</td>
<td>33</td>
</tr>
<tr>
<td>4.9</td>
<td>Word cloud of Positive topics of AM journals</td>
<td>33</td>
</tr>
<tr>
<td>4.10</td>
<td>Polarity box plot of Negative topics of AM journals</td>
<td>34</td>
</tr>
<tr>
<td>4.11</td>
<td>Word cloud of Negative topics of AM journals</td>
<td>34</td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>The baseline sentences for each trait of the Big Five personality test</td>
<td>12</td>
</tr>
<tr>
<td>3.2</td>
<td>Comparison of accuracies of <em>PredLabels</em> of different representations</td>
<td>19</td>
</tr>
<tr>
<td>3.3</td>
<td>Comparison of <em>SimScores</em> of different representations</td>
<td>20</td>
</tr>
<tr>
<td>3.4</td>
<td>Accuracy of Bi-LSTM with max-pooling and Sentence BERT models on the</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Essays and Kaggle datasets</td>
<td>21</td>
</tr>
<tr>
<td>3.5</td>
<td>The Pearson correlation between the Predlabel accuracy and the Essays</td>
<td></td>
</tr>
<tr>
<td></td>
<td>accuracy for all Sentence-BERT embeddings</td>
<td>21</td>
</tr>
<tr>
<td>4.1</td>
<td>Summary of PM journal topics</td>
<td>29</td>
</tr>
<tr>
<td>4.2</td>
<td>Summary of AM journal topics</td>
<td>32</td>
</tr>
</tbody>
</table>
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

Related Work

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, reliability, 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 \cite{55}, have proposed high accuracy models in the field of personality 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 \cite{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 \cite{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 \cite{64}.

With the emergence of accurate AI models, theorists and researchers make normative claims based on the models’ results \cite{49}. Some of the previous experience has also shown how these models can be exploited for detrimental goals \cite{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) \cite{22}. As Noam Chomsky indicates, we just cannot get to that understanding by throwing a complicated machine at it \cite{53}; several manual inspection phases are still required to produce and understand AI-based results \cite{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 \cite{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 \cite{90} and word2vec \cite{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 \cite{79}. Disambiguation of the words that have several meanings in different contexts \cite{75, 91} and extracting the sentence level meaning \cite{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 explainability 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. Studies have shown that neuroticism plays a vital role in depressive and anxiety disorders.

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 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.

1Code: [https://github.com/amirmohammadkz/interpretable_personality](https://github.com/amirmohammadkz/interpretable_personality)
Table 3.1: The baseline sentences for each trait of the Big Five personality test

<table>
<thead>
<tr>
<th>Text</th>
<th>Trait</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am extrovert</td>
<td>Ext</td>
<td>1</td>
</tr>
<tr>
<td>I am introvert</td>
<td>Ext</td>
<td>0</td>
</tr>
<tr>
<td>I am agreeable</td>
<td>Agr</td>
<td>1</td>
</tr>
<tr>
<td>I am disagreeable</td>
<td>Agr</td>
<td>0</td>
</tr>
<tr>
<td>I am neurotic</td>
<td>Neu</td>
<td>1</td>
</tr>
<tr>
<td>I am stable</td>
<td>Neu</td>
<td>0</td>
</tr>
<tr>
<td>I am an open person</td>
<td>Opn</td>
<td>1</td>
</tr>
<tr>
<td>I am not an open person</td>
<td>Opn</td>
<td>0</td>
</tr>
<tr>
<td>I am conscientious</td>
<td>Con</td>
<td>1</td>
</tr>
<tr>
<td>I am casual</td>
<td>Con</td>
<td>0</td>
</tr>
</tbody>
</table>

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
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 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.
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 ($\hat{h}_i$) and backward LSTMs ($\hat{h}_i$) [103]. For each time step, max-pooling is applied over these concatenated hidden representations ($[\hat{h}_i, \hat{h}_i]$) 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 straightforward 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:

\[ \forall s_i \in S_t : \text{SimScore}(s_i) = \]
\[ (-1)^{l_i-1} C(s_i, b_{t,1}) + (-1)^l C(s_i, b_{t,0}) \]
and
\[ \text{PredLabel}(s_i) = \begin{cases} 
1, & \text{if } C(s_i, b_{t,1}) > C(s_i, b_{t,0}) \\
0, & \text{otherwise}
\end{cases} \]

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 PredLabels and the average of the SimScores. For the Sentence-BERT models, the BFI statements and baseline statements are fed to the pretrained encoders and then the accuracy of the PredLabels and the average of the SimScores 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, PredLabels and SimScores 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 SimScores 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 PredLabels 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>O</th>
<th>C</th>
<th>E</th>
<th>A</th>
<th>N</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT (average) [76]</td>
<td>61.11</td>
<td>52.94</td>
<td>41.18</td>
<td>64.71</td>
<td>56.25</td>
<td>55.24</td>
</tr>
<tr>
<td>BERT (CLS)</td>
<td>33.33</td>
<td>58.82</td>
<td>41.18</td>
<td>47.06</td>
<td>62.5</td>
<td>48.58</td>
</tr>
<tr>
<td>Bi-LSTM with max-pooling</td>
<td>94.44</td>
<td>100.00</td>
<td>32.35</td>
<td>100.00</td>
<td>53.13</td>
<td>75.98</td>
</tr>
<tr>
<td>average_word_embeddings_glove.6B.300d</td>
<td>33.33</td>
<td>58.82</td>
<td>70.59</td>
<td>76.47</td>
<td>43.75</td>
<td>56.59</td>
</tr>
<tr>
<td>average_word_embeddings_glove.840B.300d</td>
<td>33.33</td>
<td>64.71</td>
<td>88.24</td>
<td>70.59</td>
<td>62.50</td>
<td>63.87</td>
</tr>
<tr>
<td>average_word_embeddings_komninos</td>
<td>33.33</td>
<td>70.59</td>
<td>76.47</td>
<td>70.59</td>
<td>75.00</td>
<td>65.20</td>
</tr>
<tr>
<td>average_word_embeddings_levy_dependency</td>
<td>33.33</td>
<td>41.18</td>
<td>47.06</td>
<td>64.71</td>
<td>62.50</td>
<td>49.76</td>
</tr>
<tr>
<td>nli-bert-base</td>
<td>66.67</td>
<td>76.47</td>
<td>70.59</td>
<td>88.24</td>
<td>100.00</td>
<td>80.39</td>
</tr>
<tr>
<td>nli-bert-base-cls-pooling</td>
<td>77.78</td>
<td>76.47</td>
<td>70.59</td>
<td>88.24</td>
<td>93.75</td>
<td>81.36</td>
</tr>
<tr>
<td>nli-bert-base-max-pooling</td>
<td>77.78</td>
<td>88.24</td>
<td>70.59</td>
<td>88.24</td>
<td>93.75</td>
<td>83.72</td>
</tr>
<tr>
<td>nli-bert-large</td>
<td>94.44</td>
<td>94.12</td>
<td>100.00</td>
<td>88.24</td>
<td>93.75</td>
<td>94.11</td>
</tr>
<tr>
<td>nli-bert-large-cls-pooling</td>
<td>88.89</td>
<td>88.24</td>
<td>100.00</td>
<td>88.24</td>
<td>100.00</td>
<td>91.90</td>
</tr>
<tr>
<td>nli-bert-large-max-pooling</td>
<td>88.89</td>
<td>82.35</td>
<td>100.00</td>
<td>88.24</td>
<td>100.00</td>
<td>93.07</td>
</tr>
<tr>
<td>nli-distilbert-base-max-pooling</td>
<td>72.22</td>
<td>88.24</td>
<td>17.65</td>
<td>88.24</td>
<td>93.75</td>
<td>72.02</td>
</tr>
<tr>
<td>nli-distilroberta-base-v2</td>
<td>77.78</td>
<td>82.35</td>
<td>11.77</td>
<td>88.24</td>
<td>87.50</td>
<td>69.53</td>
</tr>
<tr>
<td>nli-distilroberta-base-v2</td>
<td>72.22</td>
<td>94.12</td>
<td>70.59</td>
<td>88.24</td>
<td>100.00</td>
<td>85.03</td>
</tr>
<tr>
<td>nli-mpnet-base-v2</td>
<td>100.00</td>
<td>88.24</td>
<td>94.12</td>
<td>94.12</td>
<td>93.75</td>
<td>94.04</td>
</tr>
<tr>
<td>nli-roberta-base</td>
<td>94.44</td>
<td>82.35</td>
<td>100.00</td>
<td>88.24</td>
<td>93.75</td>
<td>91.76</td>
</tr>
<tr>
<td>nli-roberta-base-v2</td>
<td>83.33</td>
<td>94.12</td>
<td>100.00</td>
<td>88.24</td>
<td>100.00</td>
<td>93.14</td>
</tr>
<tr>
<td>nli-roberta-large</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>88.24</td>
<td>100.00</td>
<td>97.65</td>
</tr>
<tr>
<td>paraphrase-distilroberta-base-v1</td>
<td>33.33</td>
<td>70.59</td>
<td>47.06</td>
<td>70.59</td>
<td>87.50</td>
<td>61.81</td>
</tr>
<tr>
<td>paraphrase-xlm-r-multilingual-v1</td>
<td>83.33</td>
<td>70.59</td>
<td>47.06</td>
<td>76.47</td>
<td>93.75</td>
<td>74.24</td>
</tr>
<tr>
<td>stsb-bert-base</td>
<td>72.22</td>
<td>76.47</td>
<td>76.47</td>
<td>76.47</td>
<td>87.50</td>
<td>77.83</td>
</tr>
<tr>
<td>stsb-bert-large</td>
<td>88.89</td>
<td>88.24</td>
<td>100.00</td>
<td>82.35</td>
<td>68.75</td>
<td>85.65</td>
</tr>
<tr>
<td>stsb-distilbert-base</td>
<td>72.22</td>
<td>88.24</td>
<td>29.41</td>
<td>82.35</td>
<td>93.75</td>
<td>73.19</td>
</tr>
<tr>
<td>stsb-distilroberta-base-v2</td>
<td>72.22</td>
<td>82.35</td>
<td>70.59</td>
<td>82.35</td>
<td>100.00</td>
<td>81.50</td>
</tr>
<tr>
<td>stsb-mpnet-base-v2</td>
<td>94.44</td>
<td>94.12</td>
<td>94.12</td>
<td>100.00</td>
<td>93.75</td>
<td>95.29</td>
</tr>
<tr>
<td>stsb-roberta-base</td>
<td>100.00</td>
<td>70.59</td>
<td>76.47</td>
<td>82.35</td>
<td>100.00</td>
<td>85.88</td>
</tr>
<tr>
<td>stsb-roberta-base-v2</td>
<td>88.89</td>
<td>70.59</td>
<td>88.24</td>
<td>88.24</td>
<td>100.00</td>
<td>87.19</td>
</tr>
<tr>
<td>stsb-roberta-large</td>
<td>100.00</td>
<td>94.12</td>
<td>76.47</td>
<td>88.24</td>
<td>100.00</td>
<td>91.77</td>
</tr>
</tbody>
</table>
Table 3.3: Comparison of SimScores of different representations. The two first models are driven from state-of-the-art paper\cite{76}, and the third one is our fine-tuned model. The next ones are different Sentence-BERT models listed alphabetically.

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

<table>
<thead>
<tr>
<th>MODEL</th>
<th>Essays</th>
<th>Kaggle MBTI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>O/C/E/A/N/Average</td>
<td>I/E/N/S/T/F/P/J/Average</td>
</tr>
<tr>
<td>Majority Baseline</td>
<td>51.5/50.8/51.7/53.1/50.0/51.4</td>
<td>77.0/85.3/54.1/60.4/69.2</td>
</tr>
<tr>
<td>BERT-base [76]</td>
<td>64.6/59.2/60.0/58.8/60.5/60.6</td>
<td>78.3/86.4/74.4/64.4/75.9</td>
</tr>
<tr>
<td>BERT-large [76]</td>
<td>63.4/58.9/59.2/58.3/58.9/59.7</td>
<td>78.8/86.3/76.1/67.2/77.1</td>
</tr>
<tr>
<td>Bi-LSTM max-pooling_combined</td>
<td>61.7/54.6/55.0/56.7/55.9/56.8</td>
<td>-/-/-/-/-/-</td>
</tr>
<tr>
<td>average_word_embeddings_glove.6B.300d</td>
<td>63.2/58.5/56.3/57.2/58.5/58.7</td>
<td>77.2/86.5/76.9/66.2/76.7</td>
</tr>
<tr>
<td>average_word_embeddings_glove.840B.300d</td>
<td>63.0/58.0/57.2/57.5/57.7/57.8</td>
<td>78.6/87.1/79.6/68.6/78.5</td>
</tr>
<tr>
<td>average_word_embeddings_komninos</td>
<td>62.5/57.9/55.3/56.6/58.5/58.1</td>
<td>77.0/86.2/74.3/63.0/75.1</td>
</tr>
<tr>
<td>average_word_embeddings_levy_dependency</td>
<td>61.4/55.9/54.0/53.3/56.7/56.3</td>
<td>77.0/86.2/70.2/60.5/73.5</td>
</tr>
<tr>
<td>nli-bert-base</td>
<td>64.0/60/58.7/58.2/60.4/60.2</td>
<td>77.6/86.4/70.8/62.5/74.3</td>
</tr>
<tr>
<td>nli-bert-base-cls-pooling</td>
<td>63.8/59.7/57.7/59.1/60.1/60.1</td>
<td>77.6/86.3/71.1/62.2/74.3</td>
</tr>
<tr>
<td>nli-bert-base-max-pooling</td>
<td>63.0/58.0/57.7/57.4/58.4/58.7</td>
<td>77.5/86.2/69.7/61.8/73.8</td>
</tr>
<tr>
<td>nli-bert-large</td>
<td>63.5/59.8/57.1/58.7/60.8/60.0</td>
<td>77.6/86.3/71.2/62.2/74.3</td>
</tr>
<tr>
<td>nli-bert-large-cls-pooling</td>
<td>63.6/59.2/57.9/58.7/60.1/59.9</td>
<td>77.5/86.3/71.3/62.7/74.4</td>
</tr>
<tr>
<td>nli-bert-large-max-pooling</td>
<td>63.0/58.1/58.3/58.5/59.1/59.4</td>
<td>77.5/86.2/70.8/61.9/74.1</td>
</tr>
<tr>
<td>nli-distilbert-base-max-pooling</td>
<td>62.5/58.8/58.5/57.8/59.4/59.4</td>
<td>77.6/86.2/71.4/62.3/74.4</td>
</tr>
<tr>
<td>nli-distilbert-base-v2</td>
<td>62.4/57.0/57.5/57.5/60.2/58.9</td>
<td>77.5/86.2/68.8/61.7/73.6</td>
</tr>
<tr>
<td>nli-mptnet-base-v2</td>
<td>63.2/58.5/59.5/58.7/61.5/60.3</td>
<td>81.0/87.3/77.9/71.5/79.4</td>
</tr>
<tr>
<td>nli-roberta-base</td>
<td>64.2/58.8/59.7/59.1/60.6/60.5</td>
<td>81.0/87.2/78.1/69.3/78.9</td>
</tr>
<tr>
<td>nli-roberta-large</td>
<td>62.0/59.1/58.9/59.2/59.0/59.6</td>
<td>77.7/86.3/72.0/62.4/74.6</td>
</tr>
<tr>
<td>nli-roberta-base-v2</td>
<td>62.8/59.7/58.9/59.3/60.8/60.3</td>
<td>77.9/86.5/72.0/63.1/74.9</td>
</tr>
<tr>
<td>paraphrase-distilroberta-base-v1</td>
<td>65.0/57.8/59.3/59.0/59.7/60.2</td>
<td>80.1/87.1/76.2/70.7/78.5</td>
</tr>
<tr>
<td>paraphrase-xlnr-multilingual-v1</td>
<td>63.6/58.1/58.8/57.3/59.8/59.5</td>
<td>79.1/86.6/74.2/67.8/77.0</td>
</tr>
<tr>
<td>stsb-bert-base</td>
<td>64.0/59.1/57.7/58.1/60.6/59.9</td>
<td>78.1/86.5/72.4/63.4/75.1</td>
</tr>
<tr>
<td>stsb-bert-large</td>
<td>62.4/56.9/58.0/58.1/61.4/59.4</td>
<td>77.5/86.5/71.3/62.4/74.4</td>
</tr>
<tr>
<td>stsb-distilbert-base</td>
<td>62.8/58.0/58.0/57.1/59.3/59.1</td>
<td>78.5/86.5/73.1/64.6/75.7</td>
</tr>
<tr>
<td>stsb-distilroberta-base-v2</td>
<td>63.8/58.9/58.5/58.9/59.8/60.0</td>
<td>81.1/87.2/77.3/71.0/79.2</td>
</tr>
<tr>
<td>stsb-mptnet-base-v2</td>
<td>64.2/58.6/58.7/59.0/61.1/60.3</td>
<td>81.1/87.5/78.0/69.1/78.9</td>
</tr>
<tr>
<td>stsb-roberta-base</td>
<td>63.4/58.2/57.4/57.8/59.5/59.3</td>
<td>80.3/86.8/76.1/65.8/77.2</td>
</tr>
<tr>
<td>stsb-roberta-base-v2</td>
<td>63.4/58.7/59.7/58.9/60.6/60.3</td>
<td>81.0/87.3/77.5/70.3/79.0</td>
</tr>
<tr>
<td>stsb-roberta-large</td>
<td>62.7/58.4/57.6/58.0/59.7/59.3</td>
<td>80.1/86.6/74.2/65.4/76.6</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>O</th>
<th>C</th>
<th>E</th>
<th>A</th>
<th>N</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.086</td>
<td>0.488*</td>
<td>0.208</td>
<td>0.662**</td>
<td>0.533**</td>
<td>0.700**</td>
</tr>
</tbody>
</table>

In terms of accuracy of the PredLabels and SimScores, 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 SimScores, 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, MP-Net 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 \cite{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. \cite{67}. Studies have shown that there is a strong association between sleep deficiency and mortality \cite{111}, resulting in 5 of the top 15 causes of death in the U.S. \cite{57}. Insufficient sleep also impacts negatively on decision-making and reaction time, making judgments, and workplace productivity \cite{81, 99}. The most common comorbidities of sleep deprivation are anxiety, depression, bipolar disorder, and ADHD \cite{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.” \cite{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 \cite{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. \cite{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
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.
### Table 4.1: Summary of PM journal topics

<table>
<thead>
<tr>
<th>Topic</th>
<th>Theme</th>
<th>Highest LIWC Scores</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Detailed positive productive day, Hopes, Interaction</td>
<td>‘family’, ‘AllPunc’, ‘food’, ‘motion’, ‘friend’, ‘affiliation’, ‘visual’, ‘ethnicity’, ‘Social’, ‘reward’, ‘article’, ‘want’, ‘male’, ‘prosocial’, ‘det’, ‘we’, ‘socrefs’, ‘Comma’, ‘female’, ‘Tone’, ‘Dic’, ‘polite’</td>
<td>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%)</td>
<td></td>
</tr>
</tbody>
</table>

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

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

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-
<table>
<thead>
<tr>
<th>Topic</th>
<th>Theme</th>
<th>Highest LIWC Scores</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Detailed positive last night sleep</td>
<td>'Apostrophe', 'certitude', 'time', 'Drives', 'affiliation', 'Cognition', 'cogproc', 'insight', 'tentat', 'risk', 'focusfuture', 'reward', 'want', 'money', 'Social', 'socothev', 'polite', 'ethnicity', 'focuspresent', 'achieve', 'you', 'ipron', 'Period', 'Dec', 'AllPunc', 'we'</td>
<td>Fell asleep right away and slept fine. Woke up nice during the night (don’t know why) but immediately went back to bed. It was 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%)</td>
</tr>
<tr>
<td>2</td>
<td>Brief report about Satisfactory Sleep</td>
<td>'moral', 'Conversation', 'emo_pos', 'emotion', 'tone_pos', 'Affect', 'fulfill', 'allure', 'adj', 'verb', 'advverb', 'Physical', 'focuspast', 'polite', 'family', 'netspeak', 'filler', 'Cloud', 'Exclaim', 'Tone', 'sexual', 'assent', 'nonflu'</td>
<td>Slept ok(96.5%) / Slept fine (95.8%) / I slept great(95.1%)</td>
</tr>
<tr>
<td>3</td>
<td>Longest reports, several woke ups and some because of the CPAP</td>
<td>'WC', 'cause', 'auxverb', 'attention', 'Perception', 'acquire', 'need', 'wellness', 'home', 'discrep', 'number', 'tech', 'differ', 'memory', 'Culture', 'male', 'female', 'socrets', 'prep', 'conj', 'article', 'WPS', 'det', 'Linguistic', 'ppron', 'i', 'function', 'space', 'motion', 'Authentic', 'shehe', 'Comma', 'they', 'pronoun'</td>
<td>Slept good overall except didn’t get to bed until midnight due to cleaning the house and getting home late. Made myself wear the cpap mask and i woke up at 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 little clearer thinking from using cpap machine. (99.1%)</td>
</tr>
<tr>
<td>4</td>
<td>exhausted day with sleepiness or illness signs</td>
<td>'curiosity', 'quantity', 'power', 'allalone', 'substances', 'mental', 'illness', 'feeling', 'tone_neg', 'auditory', 'emo_neg', 'emo_anx', 'emoanger', 'emo_sad', 'sweat', 'visual', 'prosocial', 'health', 'conflict', 'Analytic', 'comm', 'friend', 'Lifestyle', 'leisure', 'work', 'food', 'relig', 'negate', 'QMark', 'BigWords', 'fatigue', 'lack', 'death'</td>
<td>Headache at 3:30am(97.5%) / I passed out from my pain medication(97.4%) / Super sleepy and passed out(97.3%) / (Sleepy(97.3%) / Exhausted(97.2%))</td>
</tr>
</tbody>
</table>

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”, “emoanger”, “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
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 4.8: Polarity box plot of Positive topics of AM journals

Figure 4.9: Word cloud of Positive 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 be 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
Chapter 5. Conclusions and Future Work

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 5.2 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


[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


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