Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Hang Jiang March 28, 2018

Automatic Personality Prediction with Attention-based Neural Networks

By
Hang Jiang

Jinho D. Choi
Advisor

Program in Linguistics

Jinho D. Choi
Advisor

Roberto Franzosi
Committee Member

Marjorie Pak
Committee Member

Shun Yan Cheung
Committee Member

2018

Automatic Personality Prediction with Attention-based Neural Networks

By
Hang Jiang

Jinho D. Choi
Advisor

An Abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelors of Arts with Honors

Program in Linguistics

April 2018

Abstract

Automatic Personality Prediction with Attention-based Neural Networks

By Hang Jiang

Previous works related to automatic personality prediction focus on using traditional classification models with linguistic features, but neural networks with pre-trained word embeddings, which have achieved huge success in text classification, have never been introduced for the task. This research aims to present a novel approach to automatic personality prediction using convolutional neural networks (CNN) and long short-term memory (LSTM) networks with attention mechanism. Our models are experimented on both monologue corpus, Essays dataset [1], and new multiparty dialogue corpus, called Friends dataset [2]. We first create the corpus, Friends dataset, by annotating personalities from the popular Big Five theory [3, 4] on the multiparty dialogues from the TV show, Friends, through crowdsourcing and make a comprehensive analysis of the annotation. Our annotated corpus comprises 4 seasons with an average inter-annotator agreement below 0.1. We also propose novel attention-based CNN and LSTM models to overcome the limitations of the basic CNN and LSTM by encoding long-term contextual information and providing a global view of the document. Our analysis shows word embeddings and attention mechanism can effectively improve the performance of our model on the essays dataset by ignoring noise in the corpus. Besides, our results show the challenges for human beings to agree on the task if only text is provided from dialogues. This explains the reason why all the models cannot perform well on the Friends dataset.

Automatic Personality Prediction with Attention-based Neural Networks

By

Hang Jiang

Jinho D. Choi

Advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelors of Arts with Honors

Program in Linguistics

2018

Acknowledgements

I would like to first thank my parents in China. They have supported me financially and mentally throughout my four years at Emory University. Also, they give me the biggest freedom to pursue my interest in my college life. Without their selfless support, we could never finish my degree at Emory without concerns.

I also want to thank my advisor, Jinho D. Choi. In my sophomore year, he introduced me to the field of in Natural Language Processing, which perfectly combines my interest in computer science and linguistics. He cultivates the computational linguist inside me and makes it possible for me to pursue my dream after graduation. I am grateful to Jinho's guidance and resources that have made this thesis possible.

I want to thank my other committee members including Dr. Pak, Dr. Franzosi, and Dr. Cheung as well. They were able to spare their time and gave valuable advice to my paper during my thesis year. I think they are part of my best memory at Emory.

At last, I want to thank my Emory friends including Kaixin Ma and Huiying Zhu, who have helped me during my thesis multiple times. I extremely appreciate their help.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

How language can be used to determine the personality of a person has been an important topic in psycholinguistics. It is believed that personalities are "the organized and relatively enduring traits and mechanisms that influence one's interaction with the therapsychic, physical, and social environments" [9]. According to the Lexical Hypothesis [10], those personality traits are encoded in the use of language [9]. Traditionally, personality traits are analyzed by self-reports tests, such as the Big Five Inventory Questionaire [7]. Alternatively, researchers have designed many linguistic markers to analyze one's utterances and predict his or her personalities, which make automatic personality recognition a feasible classification problem for computational psycholinguistics [8]. Previous studies have focused on developing new datasets, creating new linguistic features, and conducting feature reduction techniques to improve the field [9, 11–16]. Instead of using human-designed linguistic features, this work uses word embeddings as linguistic features, which are pre-trained word vectors using unsupervised learning algorithms. This research aims to introduce novel attention-based models to the task of automatic personality prediction and to create a new dataset, Friends dataset, to extend the task of personality classification to multiparty dialogues.

1.1 Automatic Personality Prediction

Automatic personality prediction is a task of identifying a person's personalities based on the language input. The Big Five Hypothesis is typically used for the task[3, 4]. This hypothesis describes one's personalities from five dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. We will describe the hypothesis in more details in the Chapter 2.

There are two variations of the task in our project. One is to identify the personalities of the writer by reading a monologue essay. The other version is to identify the personalities

of a speaker given the conversation corpus in which the speaker is involved in. The second version is inherently more challenging [17] for human beings than the first one because it needs a participant to grasp the interpersonal relationships and other contextual information from a short conversation before he or she could make a reasonable judgment. Some studies [8, 17] have provided multimodal information such as audio data along with the conversation text for the participants and show that prosodic information is helpful for personality identification. In our study, we focus only on text data and explores the boundaries of using only text for the task. Therefore, our task can be seen as a variation of text classification [1, 9]: the main difference is that we are identifying the personalities of a writer or speaker behind the text, whereas typical text classification problems such as sentiment analysis focus on classifying the content of the text. This situation naturally makes our task more challenging than traditional text classification problems.

The contextual information is necessary for automatic personality prediction. Therefore, we need to extract not only local features such as n-grams but also global features from the corpus. For the Essays data, we want to pay attention to the important linguistic cues by ignoring noise. For the Friends data, we want the model to pay attention to the utterances of the target speaker. To achieve this goal in both datasets, attention mechanism is added to both CNN and LSTM models.

1.2 Related Works

One of the earliest studies [8] that focus on classifying personality traits based on text is by Pennebaker and King. They extracted linguistic features from the Essays dataset using a text analysis tool LIWC and a psycholinguistic dictionary. Their findings show automatic classification of personality based on the Big Five Hypothesis is plausible. Subsequent studies were able to present promising methods by either introducing new linguistic features [8, 15, 17] or applying feature reduction techniques like Principal Component Analysis (PCA) and Information Gain [9]. Although there have been advancement in the field of personality trait classification, human-designed linguistic features play a primary role in the success of the works. In our work, we introduce word embeddings to represent the corpus and use neural networks to extract linguistic features automatically to do automatic personality prediction.

1.3 Motivation

There are two reasons to work on the task of automatic personality prediction. First of all, automatic personality prediction has applications in a number of domains. Studies [18–20]

have shown that it can be used to identify leaders of suspected terrorists from their utterances. Dating websites also analyze users' text messages and match them efficiently [21]. Given the fact that many real-life data are dialogue-based, we decide to build a new dialogue dataset and explore the feasibility of the task on it as well as on monologue data.

Secondly, this task serves as a stepping stone to a bigger task called Character Mining, which consists of three subtasks called character identification, attribute extraction, and knowledge base construction [2]. Character mining aims to construct knowledge about certain characters through information extraction. Therefore, by exploring the task of personality prediction, we are working towards extracting attributes of characters from their language input.

1.4 Objectives

There are four main objectives of our work and they are listed as follows:

- Creating a new corpus, Friends dataset, for automatic personality prediction with thorough analysis.
- Assessing the feasibility of the task on Friends dataset with the state-of-art systems.
- Implementing existing and novel models to solve the task.
- Evaluating our approach to the task on both Essays and Friends dataset.

To the best of our knowledge, this is the first time that researchers have used neural networks, attention mechanism, or word embeddings to classify personalities. Furthermore, it is the first attempt to create a multiparty dialogue dataset for personality classification.

Chapter 2

Background

2.1 Big Five Theories

One of the most popular theories in personality variation is the Big Five [3, 4]. This theory describes one's personalities from five dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. We adopt the John and Srivastava's definitions of the five personality traits. Those definitions can be seen in the following Table 2.1.

The five personality traits have been analyzed by human-designed personality description questionnaires [3, 22, 23]. Over the past 50 years, the Big Five hypothesis has become a standard model in psychology and research adopting this model has shown these five personality traits influence various aspects of task-related individual behavior [8]. The foundation for the personality tests is the Lexical Hypothesis [10], in which "the most relevant individual differences are encoded into the language, and the more important the difference, the more likely it is to be expressed as a single word" [8]. In contrast with other individual traits such as emotion, deception, and charisma, personality traits have shown a more stable and consistent characteristics of an individual [8]. For instance, some personality traits such as extraversion and conscientiousness traits have strong positive relations, whereas other traits such as neuroticism and disagreeableness have strong negative relations [24]. Based on the Lexical Hypothesis, researchers have designed many linguistic markers, including lexical categories [1, 25–27], n-grams [28], and speech-act categories [29], to analyze one's utterances and predict his or her

Big Five Traits	**Facets**
Extraversion vs. introversion	sociable, forceful, energetic, adventurous, enthusiastic, outgoing
Agreeable vs. antagonism	forgiving, not demanding, warm, not stubborn, not show-off, sympathetic
Conscientiousness vs. lack of direction	efficient, organized, not careless, thorough, not lazy, not impulsive
Neuroticism vs. emotional stability	tense, irritable, not contented, shy, moody, not self-confident
openness vs. closeness to experience	curious, imaginative, artistic, wide interest, excitable, unconventional

Table 2.1: Definitions of Big Five Personality Traits [7]

personalities. Those features make automatic personality recognition to be a feasible classification problem for computational psycholinguistics. One of the most famous systems to extract linguistic features is called Linguistic Inquiry and Word Count (LIWC) proposed by Pennebaker and King [1], and we will introduce those linguistic features in the following section.

2.2 Linguistic Inquiry and Word Count (LIWC)

Pennebaker and King [1] find a set of linguistic markers associated with the five personality traits. They create a psycholinguistics database with those features called Linguistic Inquiry and Word Count (LIWC) dictionary. They also develop the LIWC tool to analyze the monologue essays written by students, whose personality has been evaluated by a questionnaire. They find significant correlations between the linguistic markers and personality traits. Pennebaker and King [1, 8] show that "conscientious people tend to avoid negations, negative emotion words and words reflecting discrepancies". Openness to experience is recognized by "a preference for longer words and words expressing tentativity (e.g., perhaps and maybe), as well as the avoidance of 1st person singular pronouns and present tense forms". Besides, Mehl et al. [8] have added additional markers of personality perceived by observers. They find many new correlations such as the relation between the avoidance of past tense with openness to experience.

Our paper makes use of the LIWC features for implementing the state-of-art models. Although Pennebaker and King have updated the dictionary for LIWC tool in 2007 and 2015 [30], we will continue using the linguistic features in LIWC2007 dictionary in order to easily compare our work with previous works [8, 9, 31]. The LIWC2007 dictionary includes "a total of 80 output features consisting of 4 general descriptor categories (e.g., total word count, words per sentence), 22 standard linguistic dimensions (e.g., frequency of pronouns, articles), 32 word categories tapping psychological constructs (e.g., affect, cognition), 7 personal concern categories (e.g., work, home), 3 paralinguistic dimensions (assents, fillers, nonfluencies), and 12 punctuation categories (e.g., periods, commas)" [9]. Those categories are arranged hierarchically. A word can belong to multiple categories. When calculating the LIWC features for a sentence, the LIWC tool checks each word in the sentence and see whether it is in the dictionary. If a word is in the dictionary, each of these categories that the word belongs to will increase its scores. When the tool finishes checking all the words in a sentence, it produces a sparse vector with length of 80. Each value in the vector corresponds to a category mentioned above. A part of the linguistic markers can be seen in the Table 2.2 below.

LIWC Categories and Sample Words	
CATEGORIES	**EXAMPLES**
I. STANDARD LINGUISTIC DIMENSIONS	
Pronouns	I, them, itself

Articles	a, an, the
Past tense	walked, were, had
Present tense	Is, does, hear
Future tense	will, gonna
Prepositions	with, above
Negations	no, never, not
Numbers	one, thirty, million
Swear words	*****
II. PSYCHOLOGICAL PROCESSES	
Social Processes	talk, us, friend
Friends	pal, buddy, coworker
Family	mom, brother, cousin
Humans	boy, woman, group
Affective Processes	happy, ugly, bitter
Positive Emotions	happy, pretty, good
Negative Emotions	hate, worthless, enemy
Anxiety	nervous, afraid, tense
Anger	hate, kill, pissed
Sadness	grief, cry, sad
Cognitive Processes	cause, know, ought
Insight	think, know, consider
Causation	because, effect, hence
Discrepancy	should, would, could
Tentative	maybe, perhaps, guess
Certainty	always, never
Inhibition	block, constrain
Inclusive	with, and, include
Exclusive	but, except, without
Perceptual Processes	see, touch, listen
Seeing	view, saw, look
Hearing	heard, listen, sound
Feeling	touch, hold, felt
Biological Processes	eat, blood, pain
Body	ache, heart, cough
Sexuality	horny, love, incest
Relativity	area, bend, exit, stop
Motion	walk, move, go

Space	Down, in, thin
Time	hour, day, oclock
III. PERSONAL CONCERNS	
Work	work, class, boss
Achievement	try, goal, win
Leisure	house, TV, music
Home	house, kitchen, lawn
Money	audit, cash, owe
Religion	altar, church, mosque
Death	bury, coffin, kill
IV. SPOKEN CATEGORIES	
Assent	agree, OK, yes
Nonfluencies	uh, rr*
Fillers	blah, you know, I mean

Table 2.2: A list of LIWC categories. 69 categories out of 80 categories in LIWC2007 dictionary are displayed with sample words. [31]

2.3 Essays Dataset and EAR Dataset

Two datasets are popular for the task of automatic personality prediction. Essays dataset and Electronically Activated Recorder (EAR) dataset are created by Pennebaker and King [1] and created Mehl et al. [8] respectively. A comparison of the two datasets are presented in the Table 2.3.

The first essays corpus is self-report monologue data and contains 2468 essays [1, 9] collected from psychology students (1.9 million words), who were asked to write freely for 20 minutes. After finishing writing, they were asked to fill one Big Five Inventory questionnaire [7], which "asks participants to evaluate on a 5 point scale how well their personality matches a series of descriptions" [8].

The second Electronically Activated Recorder (EAR) corpus is dialogue data and contains conversation extracts transcribed from records [8]. EAR dataset contains both self-report and observer reports. Both reports were done by asking them to rate descriptions of the Big Five Inventory. It is interesting to note that the correlations on EAR dataset turn out to be higher for observer reports than for self-reports. In order to protect the privacy of the participants, Mehl et al. [8] only recorded random pieces of conversations. They also anonymized the speakers and only transcribed utterances of the participants, making it impossible to reconstruct the

Dataset	Essays	EAR
Source	Written	Spoken
Report Type	self-report	self-report & observation
Number of words	1.9 million	97,468
Instances	2,468	96
Words per subject	654	1015

Table 2.3: A comparison of Essays and EAR corpus [8]. Please note that the version of essays dataset we received from Pennebaker is different from the one Mehl et al. received. Our version is the same as what Tighe et al. received.

original conversations. Besides, this dataset contains both text and audio data. Hence, the EAR dataset is suitable for building models with both speech and text modes. However, the EAR dataset has only 96 participants with 96,468 words and 15,269 utterances, so we cannot apply our proposed neural networks mode, which typically needs thousands of annotations to train, on the EAR data. Instead, we are encouraged to create a new dialogue dataset, Friends dataset, for automatic personality prediction, which contains enough instances to train models such as neural networks.

2.4 Neural Networks

In this section, we will introduce the neural networks knowledge such as word embeddings, convolutional neural networks, long short-term memory networks, and attention mechanism. We will use these techniques in the later chapter to develop our novel model.

2.4.1 Word Embeddings

Word embedding is a popular feature learning technique in natural language processing (NLP). It is inspired by the distributional semantics hypothesis [32], which states that linguistic items with similar distributions have similar meanings. By training neural networks on large language corpus, the model learns the word embedding information as vectors, which can be used to represent words. These dense vectors, different from the sparse vector that the previous bag-of-word model produces, have low dimensions and contain rich syntactic and semantic information. There are many training methods implemented by softwares such as Word2vec [33], GloVe [34], Gensim [35], and fastText [36] to learn those representations. Word embeddings such as Word2vec and fastText allow researchers to feed large corpus data directly into neural networks, which are good at automatically extracting linguistic features from those vectors. This technique has beat many state-of-art models in NLP tasks that use human-crafted linguistic features from text and significantly improved the performance of these NLP tasks such as syntactic parsing [37] and sentiment analysis [38]. This paper compares the performance of

word embeddings and the LIWC features on the essays dataset. Furthermore, we aim to boost the performance on automatic personality prediction by using word embeddings with advanced neural networks.

2.4.2 Convolutional Neural Networks

Neural network was initially proposed to model neuron's behavior in human brain. Today, this model has achieved many breakthrough in learning tasks, like classification, regression, and prediction. Because of its ability to learn non-linear and complex features, neural network is able to outperform many traditional statistical models. Neural networks with different architectures have different advantages and achieved improvements in different tasks.

CNN is a variation of feed-forward neural network, which train all the input together at once. CNN instead selects the most salient information from the input through convolution and pooling operations [39]. It was first introduced for computer vision because its convolution and pooling layers are able to transform a image from a big representation to a smaller one and extract important local shapes. Now it is adapted for text data because word embeddings allow us to represent text as a matrix similar to the representation of a image. Because text data is sequential word by word, the convolution operations on text matrix can be adapted to move vertically while the convolution operations on image data move both vertically and horizontally. By doing so, CNN therefore learns linguistic features such as n-grams easily by using filters with different lengths. So far, CNN has improved performance of many NLP tasks such as text classification [5, 40, 41]. Compared with Long Short-Term Memory (LSTM) network that is popular for NLP tasks, CNN model is computationally inexpensive and fast to train. Since personality prediction can be seen as a variant of document classification, we select the Kim's text CNN [5] as baseline and develop a more advanced CNN model to improve the benchmark on the task.

2.4.3 Long Short-Term Memory Networks

Long Short-Term Memory (LSTM) Networks are also widely used for text classification [42, 43]. In particular, we choose Bidirectional LSTM (BLSTM) [42, 44] for automatic personality prediction because this BLSTM has access to both past and future contextual information while unidirectional LSTM only has access to the past information. The BLSTM introduces a second layer to unidirectional LSTM networks such that the network contains two sub-networks flowing forward and backward respectively. Compared with CNN models, LSTM models take longer to train but can encode some contextual information. Because our input sequences are long, LSTM models are not able to encode the long-term information efficiently neither and we therefore need attention mechanism.

2.4.4 Attention Mechanism

As mentioned above, CNN and LSTM models are good at extracting local features, lacking a global view of the whole document [45], which can be important to the personality prediction task. Our datasets both contain long text input with a few sentences. If we treat each input as a document, not all words or phrases contribute equally to composing the representation of the document. Hence, we will introduce attention mechanism to extract those words that are important to the meaning of the document and merge those words to generate a paragraph vector, called document embedding [46]. This document vector is a high level representation of the document and contains useful features for document classification, which in our case is automatic personality classification.

Chapter 3

Corpus

3.1 Corpus Creation

As mentioned before, we will develop our model to fit two types of datasets, which are distinguished by monologue and multiparty dialogue data. Essays dataset is big enough to develop neural networks models on monologue data. As for EAR dataset, it is a corpus that contain only one target speaker's utterances instead of multiparty utterances for privacy reasons, and it does not have sufficient annotations to our task. Therefore, it is both novel and necessary for us to create a new corpus for the task. Our new Friends corpus is published and publicly available online. This work also introduces a systematic framework for annotating personality traits in order to get a large scale dataset for personality prediction.

3.1.1 Data Source

We choose Friends TV show as our source of multiparty dialogue data. As discussed by Chen and Choi [2], Friends TV show is a good representation of everyday life. The content of Friends is not domain-specific and the vocabulary of the show is fairly simple, making it easy to comprehend by annotators without prior knowledge about characters or the content.

Our new Friends dataset is developed upon the work of Chen and Choi [2], who built the conversational dataset from TV show. Transcripts of the 10-season Friends TV show are formulated into clean JSON format by Chen and Choi [2]. Each season has multiple episodes, and each episode has multiple scenes. Each scene is divided into utterances, and each utterance belongs to one speaker in the scene. One utterance has at least one sentence spoken by the speaker at once.

3.1.2 Dialogue Extraction

To build our own dataset, we develop a statistical algorithm to extract sub-scenes from each scene. Sub-scenes are defined as extracts from a whole scene. The motivation to extract sub-scenes is to have annotators work on a comparatively short conversation which contains enough information to tell the personality of one speaker at once. Otherwise, annotators will spend long time reading the whole scene and have difficulties paying attention to important parts of the scene. Moreover, sub-scene extraction provides more annotations than using each scene once, which is beneficial to building a large-scale dataset.

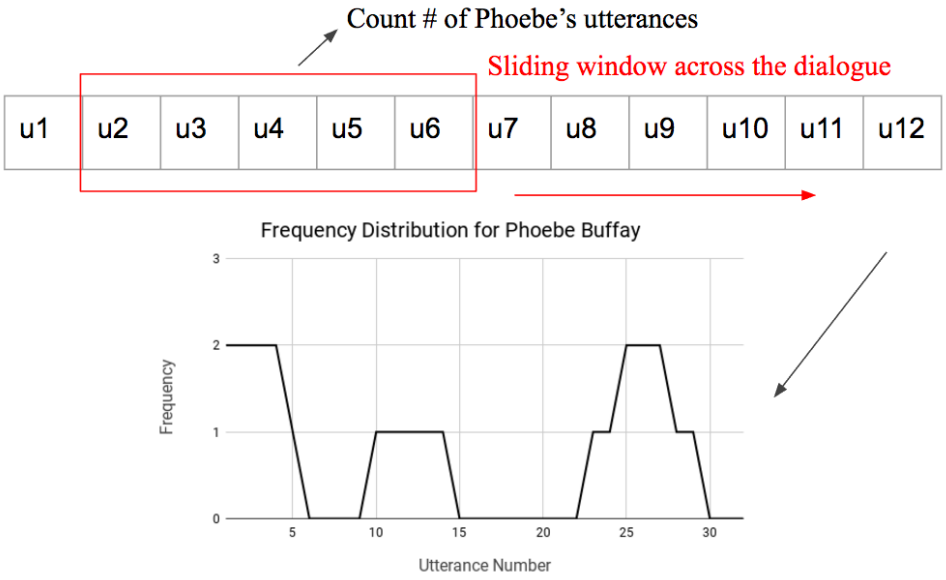

Figure 3.1: An overview of the extraction algorithm. **U** indicates utterances.

For this algorithm (Fig 3.1), we first find a set of main speakers in each scene. For each main speaker, we use a sliding-window technique to construct the frequency distribution graph in the scene. We pick the peaks in each frequency distribution graph if the peak frequency is bigger than a threshold like 2. Each peak is then used to identify the index range of consecutive sentences in which the speaker dominates temporarily. Each set of consecutive sentences extracted from the scene is called a sub-scene. In the example of Fig 3.1, the two index ranges identified should be 6 to 15 and 22 to 30. In those two sub-scenes, Phoebe Buffay is a main character. After optimizing the algorithm to get the maximum number of reliable sub-scenes, we generate 8738 sub-scenes from 10-season Friends transcripts with the minimum utterance number of a sub-scene to be 4.

3.1.3 Corpus Annotation

The first four seasons of Friends dataset are annotated through Amazon Mechanical Turk. The 4 seasons count for a total of 3545 sub-scenes out of 8738 sub-scenes. Each sub-scene is annotated by three annotators. The whole annotation has 10635 annotations and costed about 500 dollars.

To annotate each sub-scene, an annotator needs to read the sub-scene first. At the end of the page, an annotator is asked to evaluate each of five personality traits on a -1 to 1 scale for one main speaker. "1" means the annotator agrees on that personality trait and "-1' on the opposite of that personality trait. "0" means "unclear" or "unknown".

Our annotation scheme is designed to be easy-to-understand and focuses on one speaker at a time. We expect an annotator to read the whole sub-scene to pick up the context and pay attention of the utterances from the target speaker in order to make valid judgments. We found out that our design indeed facilitates the process of the annotation.

However, it is hard to make annotators agree with each other. We realize that the task is by itself more difficult than previous works [1, 17] have done. Given only excerpted dialogue text, the annotators do not have multimodal aiding data like audios or videos to help them visualize the scenes and fully understand characters. Therefore, it is expected to see annotators disagree because they might visualize different versions of the sub-scenes based on their own personalities or experience. Instead of forcing annotators to agree by rejecting numerous annotations, we decide to accept an annotation as long as we determine that the annotator pays attention to the annotation. We first manually check a couple of annotations to make some annotators have historical performances, the percentage of his or her annotations accepted. Second, we rely on the historical score, the time spent on a task, and the annotations of an annotator across tasks to decide whether one annotator works hard for the task. Third, we accept the annotation is annotator works hard; otherwise, we believe the annotator is cheating and rejects the annotation. If one's annotation is rejected frequently, we will recheck his or her historical score and decide whether to ban this annotator from our task forever.

The final agreement among the three annotators are reported as below in Table 3.1. As we can see, the inter-rater agreement, represented as Fleiss' Kappa score, is below 0.1 for all personalities, which imply that this task is probably too difficult for the online annotators. However, it is also possible that the line between 1 and 0 or -1 and 0 is very blurred and this blurred line makes the agreement rate low.

Trait	Fleiss' Kappa	Observed Agreement	Expected Agreement
Agreeable	0.053	0.414	0.381
Conscientous	0.017	0.387	0.376
Extraversion	0.016	0.455	0.446
Neuroticism	0.031	0.379	0.359
Openness	0.041	0.409	0.383

Table 3.1: Inter-rater agreement among three annotators.

Dataset	Essays	EAR	Friends
Source	Written	Spoken	Written
Report Type	self-report	self-report & observation	observation
Number of words	1.9 million	97,468	556,273
Instances	2,468	96	3,448
Words per instance	651	1015	161

Table 3.2: Comparison among the three datasets. Friends dataset we create is based on the first 4 seasons of the TV show. EAR dataset is not selected for our experiments due to its small size.

3.1.4 Annotation Adjustment

By examining the annotations we collected, we noticed two things. First, some speakers are identified as a target main speaker because this speaker has many short utterances in conversations such as "Oh", "Yeah". According to the Lexical Hypothesis [10], we need enough language input to analyze one's personality. However, those utterances include very little linguistic cues and can lead to very low agreement. Therefore, we delete sub-scenes whose target speaker has too little language input. By doing so, we have 3448 useful sub-scenes out of 3545 annotations. The statistics about our Friends dataset against the other two datasets can be seen in Table 3.2.

Since the inter-rater agreement is low for the task despite the annotators work hard, we decide to add three annotations of each task and obtain a final score for each task between -3 to 3. This way, we can make use of all the annotations. After we draw the distribution of 7 classes for each personality trait, we notice that the percentage of -3 and 3 classes are both very small, around 1 to 2 %. This means it is rare to have strong agreement in the annotations. In a statistical model, classes like -3 and 3 are too small to be ever predicted. In order to make the classes more normally distributed and make use of the two small classes, we decide to combine classes 3 and 2 together, and classes -3 and -2 together. As a result, we reduce the number of classes from 7 to 5. The resulting distributions are more normally distributed.

3.2 Corpus Analysis

3.2.1 Annotation Results

After annotation adjustment, five classes on all five personality traits are more normally distributed. If we look at the dominant class for each personality trait, some personality traits still have a dominant class way above 20%. This shows that some classes such as 1 and 0 are extremely large.This again increase the difficulty of classification because the majority baseline is built too high.

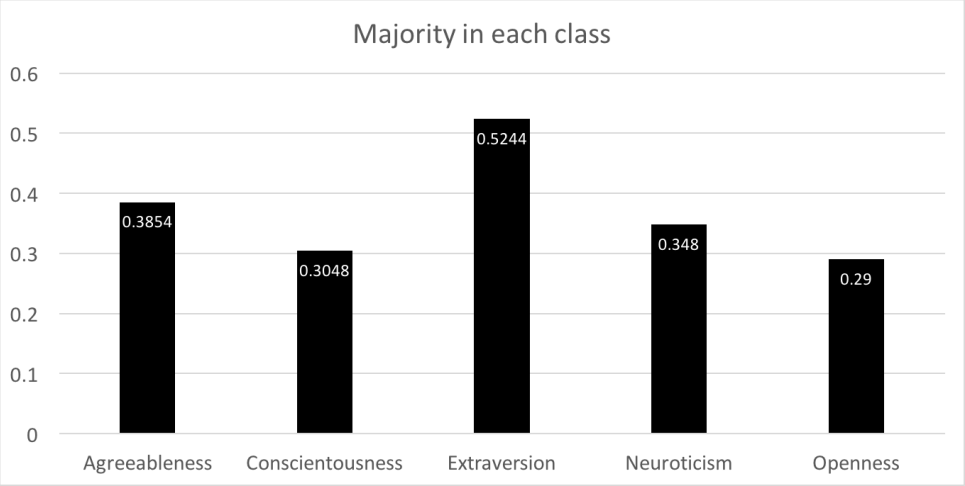

Figure 3.2: The majority distribution for each personality trait. The distribution of the dominant class for each personality trait is displayed in the graph. It shows that the data for each trait is not evenly distributed.

3.2.2 Challenges of Personality Prediction with Dialogue Text

Given the difficulties we have encountered in the annotation task, we have felt that the task is way more difficult than previous works. The difficulty is from both the dialogue structure and the lack of multimodal data. For the Essays dataset, the annotations are self-reports, which are relatively reliable because people know themselves and can answer Big Five questionnaires easily. As for the EAR dataset, observers only need to pay attention to the utterances of the participants and they have access to the prosodic cues from speech. As a study [8] has shown, prosodic cues are very important for human to infer personalities of a speaker. Therefore, the lack of audios and videos indeed poses serious challenges to the task given the low agreement of human annotators. If statistical models fail on the task, we should conclude that automatic personality prediction on text dialogue is difficult and we need to provide more information to the annotators.

Chapter 4

Approaches

4.1 Data Formulation

4.1.1 Data Split

In order to train the statistical models, we use 10-fold cross-validation [47] to split our corpus. In the 10-fold cross-validation, the original sample is randomly partitioned into 10 equal size subsamples. Out of the 10 subsamples, a single subsample is used as the validation data to test the statistical model, whereas the remaining 9 subsamples are used as training data to train the model. The cross-validation process is repeated 10 times and each time a different subsample is taken out as test data and the remaining is used for training. The 10 results from the 10 folds are averaged to generate a single score, which is the final result for the 10-fold cross-validation. The advantage of the method is to make use of all annotations for both training and validation, with each annotation used exactly once. To make results comparable for different models, we set a constant seed value for randomization so that every time we train and test a model, the same splits are generated by the random number generator.

4.1.2 Data Format

The format of Essays dataset is kept unchanged since it is monologue essays. However, the dialogue-based Friends dataset needs some formatting to distinguish the utterances of the target and non-target speakers. There are three ways proposed to format the dialogue data 4.1. The first way is to extract only the utterances of the target speaker and concatenate them together. This format will ignore the contextual information, but it distinguish sub-scenes the best. We call this format "single". The second way of formatting is to extract utterances of the target and non-target speakers, concatenate utterances respectively, and add a new line between

the two long, concatenated strings. The new line tells the machine to distinguish the target speaker's utterances and the context. We call the second format "single+context". The third way of formatting is to add a target or non-target embeddings to each utterance and the label is to help machine pay attention to utterances of the target speaker. The last format is called "target". We will run our models on the three versions of datasets and compare the results.

Original Conversation	**Single**	**Single+Context**	**Target**
Ross: Hi, Rachel.	Ross: Hi, Rachel.	Ross: Hi, Rachel.	#Targ# Ross: Hi, Rachel.
Rachel: Hi Ross.	Ross: I have a bad day.	Ross: I have a bad day.	#NonTarg# Rachel: Hi Ross.
Ross: I have a bad day.	Ross: How is your day?	Ross: How is your day?	#Targ# Ross: I have a bad day.
Rachel: Oh.		Rachel: Hi Ross.	#NonTarg# Rachel: Oh.
Ross: How is your day?		Rachel: Oh.	#Targ# Ross: How is your day?

Figure 4.1: Three ways of formatting the Friends data. These formats are proposed to help classifiers distinguish instances by recognizing the target speaker's utterances.

4.2 Models

4.2.1 LIWC-based Models

LIWC-based models refer to the state-of-art classification models used by previous studies [1, 9] together with the LIWC features. To replicate those models, we first need to use LIWC tool developed by Pennebaker and King [1] and extract a set of linguistic features for each text input. This tool analyzes text files sequentially by searching target words in its dictionary. If a target word is found in the dictionary, the corresponding word category scale will increment accordingly. To compare our models with previous studies, we will use LIWC2007 dictionary rather than the newest LIWC2015 dictionary, which provides more linguistic categories than before.

After obtaining the LIWC features, we will apply the state-of-art algorithms such as Support Vector Machine (SMO), and Linear Logistic regression (SimpleLogistic). To be consistent with previous studies, default parameters settings are used for the models. We use the best performance of those models on Friends dataset to compare with our models.

4.2.2 Convolutional Neural Networks

We will use Kim's CNN model as the baseline CNN. In this CNN model [5], every word is represented using a word vector and a document can therefore be represented by a matrix. We then use filters of lengths 2, 3, 4 to extract features related to 2,3,4 grams from the text. We concatenate the features together to form a new matrix, which represents the original text in more local details. In the following step, we will perform the global max pooling operation along the vertical axis to extract the most salient features from each dimension space, forming a document embedding with the same dimensions as the word embeddings. This model is merely for baseline establishment, so we do not expect the model to completely outperform the traditional state-of-art models. A conceptual understanding of Kim's CNN can be seen in Fig 4.2.

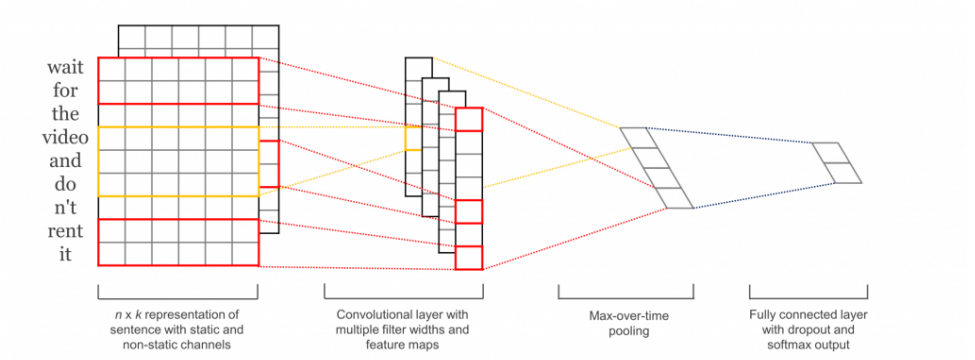

Figure 4.2: An overview of CNN [5]

4.2.3 Long Short-Term Memory Networks

Sochuster and Paliwal's Bidirectional LSTM (BLSTM) model is used for baseline establishment in our paper (Fig.4.3). In this model, an adaptive gating mechanism is used to decide how to keep the previous state and memorize extracted features of the current input. A classic LSTM model proposed by Hochreiter and Schmidhuber [48] processes a sequence word by word. At each time step, the model has access to both the past context and the current input (the current word). Bidirectional LSTM, proposed by Sochuster and Paliwal [44], has access to both the past and the future context through two sub-networks for the forward and backward sequence. The outputs of two sub-networks are combined to represent the document so that both past and future information is contained.

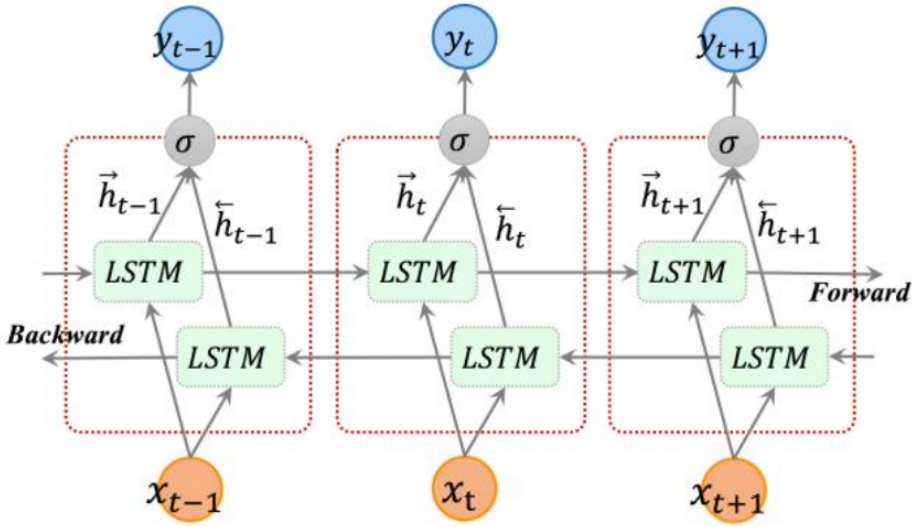

Figure 4.3: An overview of BLSTM [6]

4.2.4 Attention-based CNN and BLSTM

We will adopt the attention mechanism proposed by Yang et al. [46] and apply this attention mechanism to both CNN and BLSTM networks respectively. The introduced attention layer will produce a document representation of the output vectors by CNN or BLSTM and feed the document representation to a fully connected layer. To adapt the attention layer to CNN, we replace the global max pooling layer with a 2-dimensional max pooling layer to return a 2-dimensional matrix similar to the output of BLSTM.

4.3 Evaluation Metrics

Mehl et al.[8] use both ranking and classification models for automatic personality prediction; therefore, their paper uses both ranking loss and classification accuracies for evaluation. Edward et al. [9] use accuracy, precision and F1 score for his classification models. Both papers use 10-fold cross-validation. To be consistent with the evaluation metrics adopted by previous studies [8, 9, 17], we use classification accuracy to measure the performance of the models. Results are averaged over a 10-fold cross-validation.

Chapter 5

Experiment

Based on the objectives of the work, there are two major goals we want to achieve. First, we aim to improve the benchmarks for the task of automatic personality prediction on essays dataset with novel approaches. Second, we want to introduce the new Friends dataset and test the feasibility of the task on our multiparty dialogue dataset with existing and novel models.

First, we introduce word embeddings as linguistic features to the task and evaluate the performance of the two embeddings with MLP classifier.

Second, we implement state-of-art text classification model based on CNN and LSTM. On top of these models, we create new models with attention mechanism and evaluate the new models on both datasets.

5.1 Task Feasibility

The task of automatic personality prediction on monologue text dataset has shown to be a feasible task by previous studies. However, nobody has shown that this task is also feasible on dialogue text corpus. In order to evaluate the feasibility of the task, we will develop the state-of-art models and see how the models perform on our new corpus. It will not be surprising if the models perform worse on the dialogue corpus because our corpus is more challenging than previous dialogue datasets and even human beings show little consensus. If the state-of-art models are not successful on the new corpus, we should start designing new models to adapt to the different structure of dialogue data.

After extracting LIWC features using Pennebaker's LIWC tool, we feed the dataset into Weka to build the two best-performing classification models (SMO, and simpleLogistic) mentioned in the paper [1] without any feature reduction.

Trait	Majority	LIWC features	Word Embeddings
Agreeable	0.5308	**0.5790**	0.5551
Conscientous	0.5081	0.5662	**0.5859**
Extraversion	0.5174	0.5596	**0.5669**
Neuroticism	0.5004	0.5672	0.5693
Openness	0.5154	0.5762	**0.5944**

Table 5.1: Comparison between LIWC features and word embeddings. Multilayer Perceptron, the baseline neural network, is in both cases. FastText is used to train our word embeddings on Friends and other large-scale datasts.

5.1.1 LIWC vs word embeddings

The large number of misspellings in both datasets pose a serious challenge to the application of pre-trained word vectors because they are unlikely to appear in the pre-trained word embeddings [2]. However, this problem can be solved by utilizing a character-level word embeddings because it is able to compose similar word embeddings for those misspelled or irregular words as their corresponding standard spelling. Specifically, we use fastText [36] n-character embeddings trained on a dataset which combines New York Times corpus, the Wikipedia text dump, the Amazon Book Reviews, and the transcripts from several TV shows including Friends in our paper.

To see whether word embeddings also contain the linguistic cues necessary to the task of automatic personality prediction, we design a small experiment on the essays dataset. Specifically, we feed the essays dataset into the same MLP model using LIWC features and word embeddings respectively. The results (Table 5.1) below demonstrate that the same model has a better accuracy in 3 out of 5 personality traits on the same dataset. As a result, we are able to confirm that word embeddings are effective linguistic features for automatic personality prediction. In the following experiments, we can keep using pre-trained word vectors for our task.

5.1.2 Attention-based models vs vanilla models on Essays dataset

We first implement the basic CNN and BLSTM models [5, 42]. After fine tuning the models with a grid search algorithm, we are able to get state-of-art results with both basic CNN and BLSTM models. However, the vanilla models still cannot beat all the best performance of previous studies [9]. This is because the features extracted by CNN or BLSTM are mostly local information for the CNN model, lacking a global view of the whole document which can be important to the task. Our attention mechanism can serve to encode long-term and global contextual information. By adding attention mechanism, both models effectively pick the most salient words from the document to compose a document embedding for classification. With

Trait	Majority	Classic Models (Tighe et al.)	MLP	CNN	Att-CNN	BLSTM	Att-BLSTM
Agreeableness	53.08	57.5	55.51	57.38	**57.82**	56.64	**58.85**
Conscientousness	50.81	56	58.59	57.74	**60.13**	57.83	**59.55**
Extraversion	51.74	55.7	56.69	56.28	**58.75**	59.17	**59.32**
Neuroticism	50.04	58.3	56.93	57.09	**58.51**	57.69	**59.56**
Openness	51.54	61.95	59.44	63.49	**63.65**	63.02	**63.99**

Table 5.2: Performance of basic vs. attention-based models on the Essays dataset in accuracy. Classic models refer to the state-of-art models implemented by Tighe et al. [9]; MLP refers to Multilayer Perceptron; CNN and BLSTM are implementations of Kim [5] and Zhou et al. [42]; Att-CNN and Att-BLSTM stand for attention-based CNN and BLSTM respectively. The attention mechanism is similar to the one proposed by Zhou et al. [43].

little fine tuning on new models, we outperform all the state-of-art scores achieved by Tighe et al [9]. Besides, we observe that attention-based CNN and BLSTM outperform their vanilla models respectively, revealing that the attention mechanism is improving the baseline models consistently.

Trait	Formats	Majority	Classic Models	CNN	Att-CNN	BLSTM	Att-BLSTM
	single	38.54	38.72	38.81	38.57	**39.39**	39.07
Agreeableness	single+context	38.54	-	38.69	38.6	**38.98**	38.83
	target	38.54	-	38.69	38.57	**39.36**	38.8
	single	30.48	30.97	32.86	30.77	**32.19**	31.49
Conscientousness	single+context	30.48	-	32.05	30.66	**32.05**	31.41
	target	30.48	-	31.81	30.66	**32.66**	31.58
	single	52.44	52.44	52.55	52.55	**52.99**	52.87
Extraversion	single+context	52.44	-	52.44	52.44	**52.67**	52.58
	target	52.44	-	52.47	52.44	**52.55**	52.73
	single	34.8	35.21	36.11	35.24	**35.61**	35.21
Neuroticism	single+context	34.8	-	35.21	35.06	**35.64**	35.01
	target	34.8	-	35.64	35.44	**36.08**	35.5
	single	29	28.65	30.42	29.79	**30.68**	30.86
Openness	single+context	29	-	30.19	29.18	**30.8**	30.13
	target	29	-	30.34	29.79	**30.63**	30.05

Table 5.3: Performance of basic vs. attention-based models on the Friends dataset in accuracy. Three formats of Friends dataset are evaluated. Classic models refer to the two state-of-art models (SMO and simpleLogistics) implemented by Tighe et al. [9]; MLP refers to Multilayer Perceptron; CNN and BLSTM are implementations of Kim [5] and Zhou et al. [42]; Att-CNN and Att-BLSTM stand for attention-based CNN and BLSTM respectively. The attention mechanism is similar to the one proposed by Zhou et al. [43].

5.2 Performance on Friends dataset

We follow the three formatting method proposed in Chapter 3 to create "single", "single+context", and "target" datasets, three versions of the Friends dataset. We run all the models for each dataset, but none of the models work for any of the dataset. All the scores are merely the same as baseline.

We first construct two state-of-art models (SMO and simpleLogistics) and test the models on the Friends dataset. We pick the higher score of the two models as the score for classic models. None of the two models do significantly better than the majority baseline. This behavior of the two state-of-art models suggest that the task of automatic personality prediction is probably not feasible on our dataset. To confirm this idea, we test all of our models on the dataset too. As expected, none of our models perform significantly better than the majority baseline neither. However, there are some interesting phenomenon we notice.

First of all, it is interesting to note that the basic CNN and BLSTM do not converge on any training set while the attention-based CNN and BLSTM converge on all training datasets for 10-fold cross-validation. Nevertheless, the convergence of attention-based models only leads to lower scores on the test set. This behavior reveals again that the annotation task is challenging for annotators and machine cannot find consistency in the annotations. That attention-based model converges is probably only because they memorize all the cases on the training set while vanilla CNN cannot. This explains why attention-based models do worse than basic models respectively on the Friends dataset because attention-based models cannot generalize memorized cases on training sample into test sample. Secondly, we also see that the best results from our models still slightly outperform those from the classic models [9]. Out of our four models, BLSTM does the best in all five traits.

As a matter of fact, we have tried different ways of merging the annotations. Class numbers of 2, 3, 4, 6 are all attempted. However, we do not have any annotation that ever works on a statistical model. Therefore, we decide to just merge the two tiny classes very their neighboring class. By doing so, our data are kept less misrepresented and have a relatively healthy distribution of their classes. All the scores are in Table 5.3.

Chapter 6

Conclusion

This work introduces a novel approach to automatic personality prediction. We integrate attention mechanism into Kim's CNN model [5] and Sochuster and Paliwal's BLSM [44] to extract global features from text corpus. The attention-based models perform significantly better than Tighe et al.'s state-of-art models [9] and establish a new benchmark on the Essays dataset. Besides, this work creates a new, challenging dataset for the task of automatic personality prediction. This dialogue-based corpus, Friends dataset, is more challenging than previous corpora such as EAR and essays datasets not only because it has a distinct dialogue structure but also because it only provides a short conversation to human beings to annotate. The low agreement in our new dataset accounts for the bad performance of all the models. This low inter-rater agreement also shows two implications for future researchers. First, text only provides limited information when we analyze dialogues. To tell the personality of someone in a conversation, humans need more information such as audios and videos. Second, it is difficult for human to make direct judgment about one's personality, i.e. extroverted or introverted. Even if we provide definitions of the five personality traits for annotators to consult during work, annotators constantly complain that it is hard for them to make decisions. Instead of asking them to explicit judge personalities of a character, we should use the Big Five Inventory (BFI) [7] and design indirect questions to ask participants to fill as many studies [1, 8, 9, 17] have suggested. This alternative approach is more expensive, but it will provide reliable annotations. In summary, our work shows the potential for automatic personality prediction by improving the benchmarks on the essays dataset. By constructing the new, difficult Friends dataset, we at least shows that building dialogue datasets, different from collecting sentiment annotations, is a challenging task and more research should be invested into this area. We have shared our dataset online publicly and welcome any researcher to further analyze and annotate the data.

6.1 Future work

There are several future directions that are promising to improve our current work. First, we can develop a system that provides annotators text, audio, and video of a conversion extract to improve the quality of annotations. This new system will overcome the limitations that humans face in annotating personalities on textual dialogues. Second, we can integrate linguistic features such as LIWC features and word embeddings to improve the benchmark on essays dataset. Shin et al. Shin et al. [41] have shown that integrating lexicons and word embeddings is helpful to improving the benchmarks on sentiment analysis, which is also a text classification task.

Bibliography

[1] James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

[2] Yu-Hsin Chen and Jinho D Choi. Character identification on multiparty conversation: Identifying mentions of characters in tv shows. In *SIGDIAL Conference*, pages 90–100, 2016.

[3] Warren T Norman. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, 66(6):574, 1963.

[4] Lewis R Goldberg. Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology*, 2(1):141–165, 1981.

[5] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[6] Zhiyong Cui, Ruimin Ke, and Yinhai Wang. Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction. *arXiv preprint arXiv:1801.02143*, 2018.

[7] Oliver P John, Eileen M Donahue, and Robert L Kentle. The big five inventory—versions 4a and 54, 1991.

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

[9] Edward P Tighe, Jennifer C Ureta, Bernard Andrei L Pollo, Charibeth K Cheng, and Remedios de Dios Bulos. Personality trait classification of essays with the application of feature reduction. In *SAAIP@ IJCAI*, pages 22–28, 2016.

[10] Gordon W Allport and Henry S Odbert. Trait-names: A psycho-lexical study. *Psychological monographs*, 47(1):i, 1936.

[11] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 149–156. IEEE, 2011.

[12] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[13] Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934, 2015.

[14] Kuei-Hsiang Peng, Li-Heng Liou, Cheng-Shang Chang, and Duan-Shin Lee. Predicting personality traits of chinese users based on facebook wall posts. In *Wireless and Optical Communication Conference (WOCC), 2015 24th*, pages 9–14. IEEE, 2015.

[15] Saif M Mohammad and Svetlana Kiritchenko. Using nuances of emotion to identify personality. *Proceedings of ICWSM*, 2013.

[16] Soujanya Poria, Alexandar Gelbukh, Basant Agarwal, Erik Cambria, and Newton Howard. Common sense knowledge based personality recognition from text. In *Mexican International Conference on Artificial Intelligence*, pages 484–496. Springer, 2013.

[17] François Mairesse and Marilyn Walker. Automatic recognition of personality in conversation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 85–88. Association for Computational Linguistics, 2006.

[18] Robert Hogan, Gordon J Curphy, and Joyce Hogan. What we know about leadership: Effectiveness and personality. *American psychologist*, 49(6):493, 1994.

[19] Simon Tucker and Steve Whittaker. Accessing multimodal meeting data: Systems, problems and possibilities. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 1–11. Springer, 2004.

[20] Sam Nunn. Preventing the next terrorist attack: The theory and practice of homeland security information systems. *Journal of Homeland Security and Emergency Management*, 2(3), 2005.

[21] M Brent Donnellan, Rand D Conger, and Chalandra M Bryant. The big five and enduring marriages. *Journal of Research in Personality*, 38(5):481–504, 2004.

[22] Dean Peabody and Lewis R Goldberg. Some determinants of factor structures from personality-trait descriptors. *Journal of personality and social psychology*, 57(3):552, 1989.

[23] Lewis R Goldberg. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216, 1990.

[24] David Watson and Lee Anna Clark. On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model. *Journal of personality*, 60(2):441–476, 1992.

[25] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[26] Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862, 2006.

[27] Lisa A Fast and David C Funder. Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of personality and social psychology*, 94(2):334, 2008.

[28] Jon Oberlander and Alastair J Gill. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, 42(3):239–270, 2006.

[29] Klaus Vogel and Sigrid Vogel. L'interlangue et la personnalite de l'apprenant. *IRAL: International Review of Applied Linguistics in Language Teaching*, 24(1):48, 1986.

[30] James W Pennebaker, Roger J Booth, Ryan L Boyd, and Martha E Francis. Linguistic inquiry and word count: Liwc 2015 [computer software]. pennebaker conglomerates, 2015.

[31] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[32] Magnus Sahlgren. The distributional hypothesis. *Italian Journal of Disability Studies*, 20: 33–53, 2008.

[33] Yoav Goldberg and Omer Levy. word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

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

[35] Keyvan Khosrovian, Dietmar Pfahl, and Vahid Garousi. Gensim 2.0: a customizable process simulation model for software process evaluation. In *International Conference on Software Process*, pages 294–306. Springer, 2008.

[36] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[37] Richard Socher, John Bauer, Christopher D Manning, et al. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 455–465, 2013.

[38] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[40] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 373–374. ACM, 2014.

[41] Bonggun Shin, Timothy Lee, and Jinho D Choi. Lexicon integrated cnn models with attention for sentiment analysis. *arXiv preprint arXiv:1610.06272*, 2016.

[42] Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. *arXiv preprint arXiv:1611.06639*, 2016.

[43] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 207–212, 2016.

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

[45] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1201–1211. Association for Computational Linguistics, 2012.

[46] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

[47] Tadayoshi Fushiki. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21(2):137–146, 2011.

[48] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.