Hierarchical Modeling for User Personality Prediction: The Role of Message-Level Attention

Veronica E. Lynn, Niranjan Balasubramanian, and **H. Andrew Schwartz**
Stony Brook University
{velynn, niranjan, has}@cs.stonybrook.edu

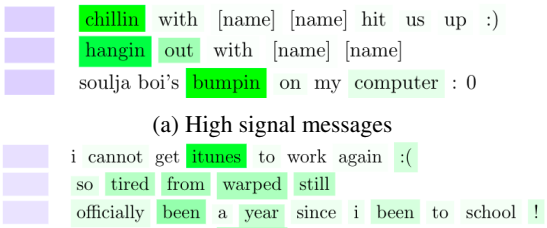
Abstract

Not all documents are equally important. Language processing is increasingly finding use as a supplement for questionnaires to assess psychological attributes of consenting individuals, but most approaches neglect to consider whether all documents of an individual are equally informative. In this paper, we present a novel model that uses message-level attention to learn the relative weight of users' social media posts for assessing their five factor personality traits. We demonstrate that models with message-level attention outperform those with word-level attention, and ultimately yield state-of-the-art accuracies for all five traits by using both word and message attention in combination with past approaches (an average increase in Pearson r of 2.5%). In addition, examination of the high-signal posts identified by our model provides insight into the relationship between language and personality, helping to inform future work.

1 Introduction

Most language-based methods for human attribute prediction assume all documents generated by a person are equally informative. However, this is not necessarily true. Figure 1 gives examples of high and low signal messages for predicting extraversion — one's tendency to be energized by social interaction. The high signal messages contain words relating to social interaction (*hangin out*, *chillin*), whereas the low signal messages, while still containing social-related words, have little clear relevance to extraversion. The former examples would ideally be weighted higher by a personality prediction model than the latter.

This paper applies the idea of modeling *document relevance* to the task of personality prediction. Inferring an individual's personality traits is a fundamental task in psychology (McCrae and Costa Jr, 1997; Mischel et al., 2007), with social scientific applications ranging from public health (Friedman and Kern, 2014) and marketing (Matz et al., 2017) to personalized medicine (Chapman et al., 2011), mental health care (Bagby et al., 1995), and even providing useful information for downstream NLP tasks (Preoţiuc-Pietro et al., 2015; Lynn et al., 2017). Recently, researchers from both NLP and psychology have turned toward more accurately assessing personality and other human attributes via language (Mairesse et al., 2007; Schwartz et al., 2013; Park et al., 2015; Kulkarni et al., 2018). The idea behind "language-based assessments" (Park et al., 2015) is that language use patterns can supplement and, in part, replace traditional and expensive questionnaire-based human assessments.



(a) High signal messages

(b) Low signal messages

Figure 1: Examples of high and low signal messages identified by our proposed model for predicting extraversion. All examples are from the same highly-extroverted user. Shading indicates strength of message-level (blue) and word-level (green) attention.

Here, we present a hierarchical neural sequence model over both the words and messages of the user and correspondingly applies attention to each level. The document-level attention learns the relative importance of each social media post for predicting personality.

Contributions. Our main contributions include:

1. A neural model for personality prediction that uses message-level attention to recover high-signal messages from noisy data.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5306–5316
July 5 - 10, 2020.

2. An empirical demonstration that shows models with message-level attention outperform those without.
3. State-of-the-art performance for language-based assessment of personality.
4. Insight into the relationship between message-level language use and personality.

2 Model Architecture

Our goal is to encode user messages into a representation that can be used to predict the personality of the user. We can use a two-step process to produce such a representation: First encode the sequences of words in each message to form message-level representations and then encode the message-level representations to form a user-level representation. Social media users write hundreds or even thousands of messages; while the messages, and the words within them, contain valuable clues to their personality, not all of it is equally valuable. An ideal representation of user text, therefore, should pay particular attention to personality-revealing portions of a user's text. Hierarchical attention is a natural fit for this problem. At the message level, a word-attention model can learn to emphasize personality related words in the message representation, while at the user-level, a message attention model can learn to emphasize personality-related messages in the overall user representation. We instantiate this idea using a hierarchical sequence architecture shown in Figure 2.

Given a set of n messages from a user u, the first step of the model is to produce an encoding for each message m_i. Each word w_j^i in message m_i is fed through a Gated Recurrent Unit (GRU) (Cho et al., 2014) to produce a hidden state:

$$h_j^i = \text{GRU}(w_j^i) \tag{1}$$

We then apply an attention mechanism over the sequence of hidden states $[h_1^i, h_2^i, ..., h_l^i]$:

$$d_j^i = \tanh(W_{word} h_j^i + b_{word}) \tag{2}$$

$$\alpha_j^i = \frac{\exp({d_j^i}^\top d_{word})}{\Sigma_{k=0}^{l} \exp({d_k^i}^\top d_{word})} \tag{3}$$

$$s_i = \sum_{k=0}^{l} \alpha_k^i h_k^i \tag{4}$$

where d_{word} is a learned context vector for word-level attention, b_{word} is a bias term, and α_j^i is a normalized attention weight for h_j^i. s_i is thus a weighted combination of the hidden states representing $\{w_1^i, w_2^i, ..., w_l^i\}$.

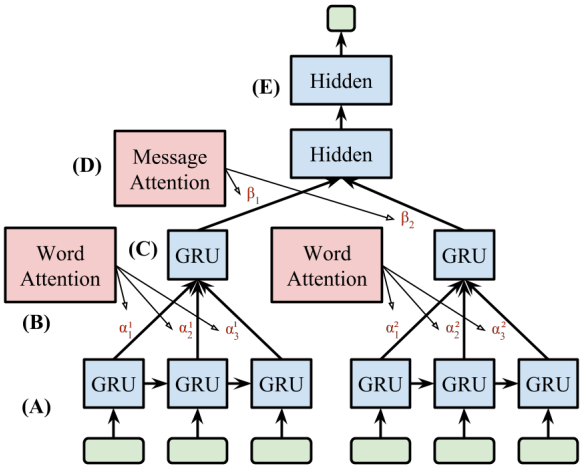


Figure 2: Diagram of our proposed model for personality prediction. (A) Each post is passed through a GRU to produce a message-level encoding. (B) A word-level attention mechanism learns weights for each of the words in the message. (C) All message representations are passed to a second GRU to produce a user-level encoding. (D) A message-level attention mechanism learns weights for each of that user's posts. (E) The user representation passes through two hidden layers and a final prediction layer.

Once we have these message representations, the next step is to encode each sequence of messages into a user representation. Each message representation s_i is passed through another encoder, also using Gated Recurrent Units:

$$h_i = \text{GRU}(s_i) \tag{5}$$

As before, the hidden states are then passed through another message-level attention mechanism:

$$e_i = \tanh(W_{message} h_i + b_{message}) \tag{6}$$

$$\beta_i = \frac{\exp(e_i^\top e_{message})}{\Sigma_{k=0}^{n} \exp(e_k^\top e_{message})} \tag{7}$$

$$u = \sum_{k=0}^{n} \beta_k h_k \tag{8}$$

As before, $e_{message}$ is a learned context vector for message-level attention. The representation for a user u is thus a weighted combination of the hidden states representing that person's messages. Once the user representation has been produced, u is further passed through some fully-connected layers before being used for prediction at the final layer.

In this way, important words and messages don't get lost to noise and are instead carried through to later portions of the model, where they can have a greater impact on the final prediction. Our model is similar in structure and motivation to the Hierarchical Attention Network proposed by Yang et al. (2016). However, our work focuses on a different level of analysis: whereas Yang et al. (2016) encode `words` → `sentences` → `documents`, our work seeks to encode `words` → `documents` → `users`. This idea of applying attention at a document level when modeling user-level attributes is, to the best of our knowledge, entirely novel. We hypothesize that *where* attention is applied is crucial and that message-level attention is of particular importance for modeling personality.

3 Dataset

We draw our data from consenting users of a Facebook application (Kosinski et al., 2013), which allowed users to take various psychological assessments and voluntarily share their data with researchers. Following the work of Schwartz et al. (2013) and Park et al. (2015), the current state of the art on this dataset, we filtered the users to those who shared their Facebook status posts, wrote at least 1,000 words across those statuses, provided their age and gender, and were less than 65 years old.

All users completed psychological measures, ranging from 20 to 100 items, that assessed their Big Five personality traits (Costa and McCrae, 1992): conscientiousness, agreeableness, neuroticism, openness to experience, and extraversion. Each of the five dimensions is represented by a normalized, continuous score representing the degree to which that trait is exhibited. We refer to these as *personality scores*. The Big Five personality traits are described more fully in Section 4.

Overall, our dataset contains Facebook statuses and personality scores for 68,687 users. To allow for direct comparisons, we use the same test set (n=1,943) as Park et al. (2015). Each of these test users completed a longer 100-item questionnaire, ensuring higher-quality scores. We sample an additional 4,998 for use as a development set, and leave the remaining 61,746 for training.

On average, users in our dataset are 23 years old and 63% are female. Users had an average of 3,619 words and 165 messages, all posted to Facebook between 2009 and 2011.

Ethical Research Statement. All participants consented to sharing their status updates and personality questionnaire results for research purposes, and the study has been approved by an academic institutional review board.

4 Big Five Personality Traits

Discovery of the "Big Five" personality traits began nearly a century ago with some of the first data-driven, statistical latent variable modeling techniques (Thurstone, 1934). The goal in this decades-long pursuit was not very different from that of producing latent vector embeddings of words:[1] to use latent factor analysis to reveal underlying, stable dimensional vectors that distinguish people. However, rather than finding latent semantic dimensions of words, the models (run by hand at first) focused on how individuals answered questions about themselves. For example, modern questions include: "How much do you agree with these statements? (1) I am the life of the party; (2) I have difficulty understanding abstract ideas; (3) I like order; (4) I worry about things" (Goldberg et al., 2006).

The idea behind this data-driven approach was that if such latent dimensions could be found to be stable across time and differing populations, that suggests they are fundamental to what makes each of us different. Such work continued for decades, documented across thousands of studies to eventually arrive at the acceptance of five such factors being fundamental and consistent across time and populations (Costa and McCrae, 1992). Those fundamental human factors, the target of our human language predictive task, are described below.

The big five often goes by the acronym "OCEAN", standing for ***o****penness to experience*, ***c****onscientiousness,* ***e****xtraversion,* ***a****greeableness,* and ***n****euroticism*. High scores for *openness to experience* are correlated with philosophical and free thought, as well as an interest in the arts, music, and cinema (Schwartz et al., 2013; Kern et al., 2014). Those who score low here may be more practical, realistic, or close-minded (Costa and McCrae, 1992).

Individuals with high *conscientiousness* tend to be well organized and have a lot of self-discipline, which may be expressed through discussions of work or school-related responsibilities (Yarkoni, 2010; Kern et al., 2014). Those who score low

[1] In fact Thurstone referred to the latent variables as "vectors of the mind".

on this dimension may appear impulsive, disorganized, or unreliable. Those with high *extraversion* are likely to talk about friends, social situations, and interpersonal interaction. On the other hand, those with low extraversion may be more independent and may focus more on solo activities (e.g. watching television) (Costa and McCrae, 1992; Park et al., 2015).

Agreeableness is associated with being friendly and good-natured, while those who score low may be selfish or rude. Swearing is highly correlated with low agreeableness (Yarkoni, 2010; Schwartz et al., 2013). High *neuroticism* is strongly linked to anxiety and depression, while low neuroticism is linked to emotional stability.[2] This dimension may be expressed through feelings such as fear, sadness, or frustration (Costa and McCrae, 1992; Kern et al., 2014).

5 Evaluation

In this section, we describe the method for training and evaluating our proposed model, along with the various baseline models we compared against.

5.1 Features

Each user was represented as a sequence of their messages, from most to least recent, which were themselves represented as a sequence of word embeddings. To do so, we pre-trained 200-dimensional word2vec embeddings (Mikolov et al., 2013) over all messages belonging to the training set users. The vocabulary was limited to words that appear in at least 50 messages. Words that occurred fewer times were replaced by an out-of-vocabulary token. The Language Detection Library (Shuyo, 2010) was used to filter out non-English texts.[3]

5.2 Baseline Models

Ridge Regression (N-Grams/Topics). We compare against Park et al. (2015), which is the current state of the art on this dataset and, to the best of our knowledge, demonstrated the best published regression predictions over a Big Five personality factors from language alone. Their model uses a combination of n-gram features and LDA-based topics extracted from the training data. These features then undergo dimensionality reduction in the form of univariate feature selection and randomized principal component analysis, resulting in a total of 5106 features. These features are then used to train ridge regression models, one per personality dimension, for prediction. Because we use the same test set users as Park et al. (2015), we compare directly against their reported results.

Ridge Regression (Embeddings). In addition to the n-gram and topic-based ridge models of Park et al. (2015), we train ridge regression models using the word embeddings described in Section 5.1. These embeddings are averaged first per-message and then per-user, creating a 200-dimensional embedding per user to input to the model.

DAN. We modify the model proposed in Section 2 to use a Deep Averaging Network (Iyyer et al., 2015), rather than a GRU, at the word and/or message level. This takes the average across all word (or message) embeddings to produce a message- (or user-) level representation.

DAN + Attn. Identical to the DAN variant except takes the weighted (rather than unweighted) average using learned attention weights.

Sequence Network (SN). Similar to our proposed model but using the final state of each GRU, rather than word or message attention.

Transformer (TN). This variant of our proposed model uses a two-layer transformer (Vaswani et al., 2017) with double-headed attention, rather than a GRU, at the message or word level.

BERT. Whereas our proposed model *learns* message-level representations, we instead experiment with using pre-trained BERT embeddings (Devlin et al., 2019) as our message representations. These 768-dimension message embeddings are produced by averaging across all BERT token embeddings for each message (Matero et al., 2019).

5.3 Training

All models were implemented using PyTorch (Paszke et al., 2017), with the exception of Ridge Regression which used scikit-learn (Pedregosa et al., 2011). One model was trained for each of the five personality dimensions. All deep learning models use two feed-forward layers with 512 hidden units each, followed by a final prediction layer. The GRU layers have a hidden size of 200 to match the number of embedding dimensions. Similarly, we learn a projection down to 200 dimensions for our BERT embeddings.

All hyperparameters (dropout and learning rate

[2] Some versions of the Big Five flip this dimension and call it "emotional stability".

[3] Even without this step, the models tended to artificially exclude non-English texts by assigning them very low attention weights.

word-to-message	message-to-user	*OPE*	*CON*	*EXT*	*AGR*	*NEU*
DAN	DAN	.579	.516	.509	.474†	.516
SN	SN	.601	.506	.512	.431	.523
DAN + Attn	DAN + Attn	.615†	.506	.530†	.499†	.528†
DAN + Attn	SN + Attn	.605	.510	.535†	.501†	**.560**†
SN + Attn	DAN + Attn	.625	.497	.539†	**.519**†	.532†
SN + Attn	SN + Attn	**.626**	**.521**	**.552**†	.509†	.541
TN (Attn)	SN + Attn	.544	.474	.513†	.483†	.526

Table 1: Comparison of Disattenuated Pearson R of different models for personality prediction on the test set users (n=1943), using different architectures to aggregate from word to message level and message to user level. † indicates statistically significant improves over the `SN (No Attention)` baseline, based on a paired t-test on the errors of each model.

word-to-message	message-to-user	*OPE*	*CON*	*EXT*	*AGR*	*NEU*
SN + Attn	SN + Attn	**.626**	**.521**	.552	.509	.541
BERT	DAN	.602	.512	.537	.505	.520
BERT	SN	.597	.511	.520	.522	.507
BERT	DAN + Attn	.613	.511	**.570**†	.533†	.536
BERT	SN + Attn	.610	.519	.544	**.538**†	**.547**†
BERT	TN (Attn)	.590	.501	.526	.523	.516

Table 2: Performance as Disattenuated Pearson R measures when using pre-trained BERT embeddings (Devlin et al., 2019) at the message level, compared to our proposed model which learns message-level representations. † indicates statistically significant improvement over the `SN + Attn` model based on a paired t-test on the errors of each approach.

for deep models; alpha for ridge) were tuned over the development set for a single personality dimension (*OPE*), with the best parameters being used to train models for the remaining dimensions. The deep models were trained using a batch size of 64. Training lasted for a maximum of 20 epochs, with most models stopping after around 10 epochs due to early stopping with a patience of two epochs. To reduce memory requirements during training, each user's post history was "chunked" into sequences of at most 500 messages each. For example, a user with 1250 messages total would be divided into three instances with 500, 500, and 250 messages. This was only done for the training set; the testing and tuning sets used all messages at once.

6 Results

Our evaluation aims to answer the following:

1. How successful are attention-based models at predicting personality?
2. What is the distribution of high signal versus low signal messages?
3. What is the relative importance of message-level attention over word-level attention?

6.1 Attention for Personality Prediction

Table 1 compares the performance of our proposed model, `SN+Attn`, against variations using different architectures to aggregate from the word to message level and message to user level. Model performance is given as the disattenuated Pearson correlation coefficient[4] between the predicted and questionnaire-based personality scores.

Overall the models with attention outperform those without. Perhaps surprisingly, the `SN+Attn` at the message level typically outperformed the `DAN+Attn`, which may be due to the messages forming a sort of personal narrative, containing repeated themes and follow-ups to previous messages. The `SN+Attn` also tended to outperform the `DAN+Attn` at the word level. Our proposed model, using `SN+Attn` at both word and message level, is best for three out of five dimensions.

Table 2 shows the performance when using pre-trained BERT embeddings (Devlin et al., 2019) as our message representations, rather than learning them as part of the model. As before, we see that message-level attention is generally beneficial, and additionally we find that the BERT-based models outperform our proposed model in 3 out of 5 cases.

Table 3 compares our proposed model against the state-of-the-art. Unsurprisingly, `Ridge (Embeddings)` is the worst-performing model overall. Although Park et al. (2015) also used ridge

[4]Disattenuated Pearson correlation helps account for the error of the measurement instrument (Murphy and Davidshofer, 1988; Kosinski et al., 2013). Following Lynn et al. (2018), we use reliabilities: $r_{xx} = 0.70$ and $r_{yy} = 0.77$.

	d	*OPE*	*CON*	*EXT*	*AGR*	*NEU*
Ridge (Embeddings)	200	.538	.500	.505	.444	.505
Our Proposed Model	200	.626	.521	.552	.509	.541†
Ridge with PCA (N-Grams/Topics) (Park et al., 2015)	5106	.627	.518	.558	.545	.531
Ridge with PCA (N-Grams/Topics) + Our Proposed Model	5306	**.657**†	**.538**†	**.583**†	**.557**†	**.564**†

Table 3: Combining our best model with that of Park et al. (2015) obtains new state-of-the-art performance in terms of Disattenuated Pearson R. Number of input dimensions (d) is shown for each model. † indicates a statistically significant improvement over Park et al. (2015) based on a paired t-test on the errors of each approach.

regression, their models used significantly more features (d=5106 (dimensionally reduced, supervised, from an original of over $d > 50,000$) compared to our d=200). Finally, we find that by averaging the z-scored predictions of our proposed model and `Ridge (N-Grams/Topics)`, we obtain the overall best performance, outperforming current state-of-the-art. This suggests that the models are able to learn complementary information.

These results show that neural models with attention are better able to predict personality than those without. Because some messages are of more relevance than others, attention allows the model to better separate the signal from noise. In addition, combining the predictions of the best attention-based model, `SN+Attn`, with those from Park et al. (2015), the previous best, advances the state-of-the-art results over all 5 factors by a signficant margin ($p < .05$ from a paired t-test on error) and an average increase of .025, demonstrating the complementary value in these methods.

6.2 Message Attention Distribution

Results suggest not all text is equally informative when it comes to personality prediction, which is why attention helps. Figure 3 shows the distribution of standardized message-level attention weights, obtained from our proposed model, for 100 randomly-sampled test set users. Sampled users had 742 messages on average. The figure shows that any single user's messages encompass a range of relative importance. *OPE* skews negative, indicating that most messages of a user are of little relevance with a few being very relevant, while *NEU* was slightly more likely to mark messages as relevant but with less variance. By incorporating that concept of message (and word) importance via attention, we can produce better user-level representations from which to predict personality.

6.3 Effects of Word and Message Attention

Thus far we have demonstrated the importance of attention for personality prediction. However, our proposed model incorporates attention at two different levels of analysis: word and message level. We examine each attention mechanism's impact on the overall performance of the model.

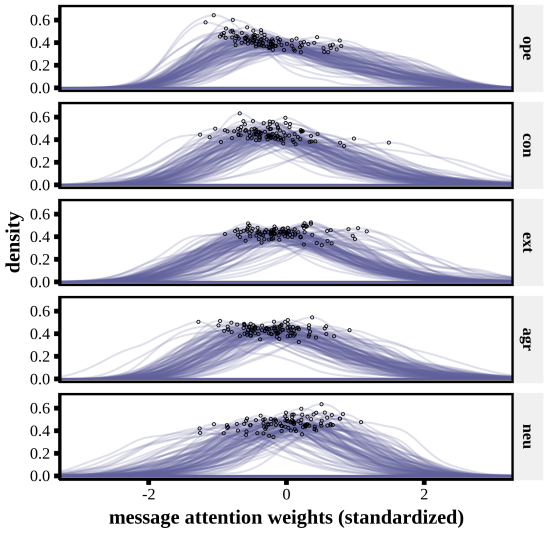


Figure 3: Standardized distribution of message-level attention weights for 100 randomly-sampled test set users with at least 20 messages. The black dot indicates the max density per user (i.e. the most frequent attention weight for that person).

	OPE	*CON*	*EXT*	*AGR*	*NEU*
No Attn	.601	.506	.512	.431	.523
Word Only	.612†	.510	.516†	.456†	.541†
Msg Only	.621†	.511	.535†	**.521**†	**.544**†
Word + Msg	**.626**	**.521**	**.552**†	.509†	.541

Table 4: Ablation demonstrating the importance of using word- and message-level attention. All models are sequence networks (SNs) with or without attention at the word and message levels. † indicates statistically significant improvements ($p < 0.05$) over the No Attention baseline based on a paired t-test on the errors of each approach.

Table 4 shows ablation results for word and message attentions. As expected, adding any attention results in improvements over the `No Attn` model. In addition, using only message-level attention generally outperforms using only word-level attention. This may be because message-level attention oc-

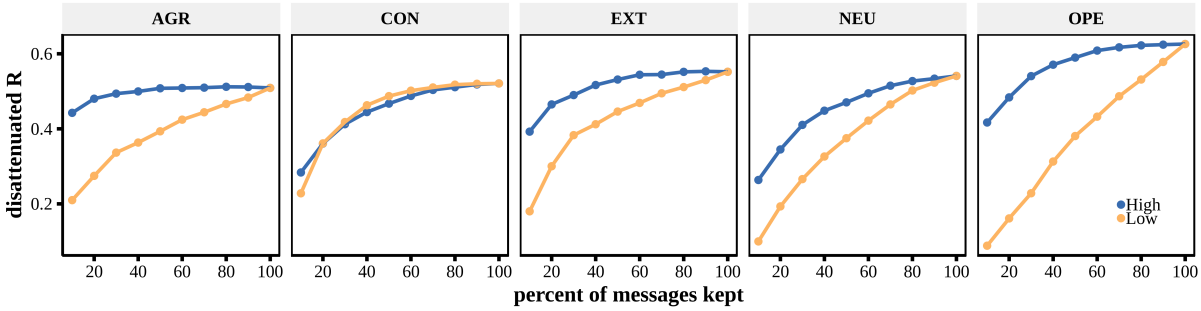


Figure 4: Performance of our model when keeping only the top n percent highest or lowest weighted messages.

curs later in the model, where its impacts are less likely to get washed out by downstream layers.

While adding message attention provides the single largest boost, in 3 out of 5 cases combining it with word attention results in additional gains. This may be because the word-level attention helped the model to better encode longer messages: the average message length for the top 5% highest-weighted messages were, on average, 4.4 tokens longer for `Word+Msg` than for `Msg Only`.

The inclusion of message-level attention appears to have little direct impact on the word-level attention. On examination, `Word+Msg` and `Word Only` typically assigned roughly the same word-level attention weights to the same sentences. This suggests the strength of adding message-level attention is in learning how best to weight messages, rather than how to better represent each individual message.

We further explore the impact of the learned message-level attention weights. Figure 4 shows our proposed model's performance when evaluated over the top n percent highest or lowest weighted messages, as learned by our model. We see that performance is much better when using high-attention messages than low-attention ones in all cases but *CON*, which we saw in Table 4 did not benefit much from message-level attention. Another note of interest is that *AGR* plateaus very quickly for high attention messages, which suggests that high-signal messages are rare but extremely predictive.

In conclusion, while adding any attention is helpful, message-level attention provides overall larger gains than word-level attention.

7 Qualitative Value of Identifying Informative Text

The high-signal text identified by our attention-based models potentially provides additional, qualitative value for researchers interested in the relationship between language and personality. Bag-of-words approaches to language modeling can identify attribute-relevant words (e.g. word clouds), but this can be limiting as it lacks the context in which the words appear. By contrast, a personality researcher interested in how high extraversion, for example, manifests itself in one's language use can use our learned attention weights to identify whole messages that may warrant further study.

Table 5 shows examples of messages that received high and low attention weights from the `SN+Attn` model for users at the extreme ends of each personality dimension. Overall, the high-attention messages are thematically relevant to the target personality dimension. For example, the messages for *conscientiousness* focus on work and school responsibilities, while those for *extraversion* discuss social interactions. The high-attention words, highlighted in green, are also consistent with each personality dimension. For example, *openness to experience* highlights philosophical words (*weird*, *nothingness*, *trippy*) while *agreeableness* favors swear words (*shit*). In contrast, the low-attention messages have little relevance.

To test whether our high-signal text might be of qualitative value to researchers, we asked two experts on personality (psychologists with past research in the area) to view 100 paired messages sets (20 per dimension) and select which set was more informative of the individual's personality. Each paired set consisted of 5 messages within the top third of message weights and 5 in the bottom third for a given user. To reduce the frequency of long messages, we only selected messages whose length was at most 20 characters above or below that user's average message length. The users themselves were randomly sampled from those in the top or bottom 10th percentile of each dimension and who had at least 20 messages total. Note that personality psychologists, though experts in how

High OPE		trippy day ahead ....
		nothingness at last ....
		shutter island was good .. .
		they are over ... yah
		my phone is not working ...
High CON		stoked on the exam schedule !
		40 % math midterm ? thank god 3/4 count .
		got a co-op job interview ! woo !
		just had some damn good pears . note to self : buy more ? damnit .
		found free bag of skittles in the vending machine , jackpot .
High EXT		at the beach with keira ! ! !
		getting ready for brittany's dance recital tonight ! !
		had fun at nathans barmitzvah last night ! ! !
		i have made 72 cupcakes in the last 3 days ! ! ! ! lol
		just finished my science project :)
Low AGR		sooo excited for new school year :) going top make it awesome
		grudges are so ridiculous and pointless ¿ - ¿
		ahh shit almost 1 ! ? i need to finish this paper ! ! !
		that sure was a fun ride home O.o
		wants to just skip to the next weekend .
High NEU		can't believe i got that done in time .....
		packing to go back to school makes me sad .
		losing things and is getting extremely frustrated . :(
		is amazed at how similar cameras are to your eyes .
		whhhaaa ? it's only wednesday ...

Table 5: Random selection of messages that received high (top) and low (bottom) attention weights from the SN+Attn model. Blue shades indicate strength of message-level attention and green indicates word-level attention. Each set of messages is from a single user, with that user having a personality score in the top or bottom 10th percentile. For brevity, only messages with 70 or fewer characters were included.

personality manifests in behaviors like language, are not trained necessarily to identify it from microblog posts. The goal here is not to simply validate the attention, but to shed some light on where message attention helps and whether it is consistent with expectations from personality theory.

Table 6 shows the percentage of instances where each expert identified the high-attention set as most informative, and their inter-rater agreement. Judges showed a preference towards the high-attention messages for *OPE* and *AGR*, while *CON* and *NEU* were no better than chance. These findings are somewhat consistent with Table 4, which showed that *OPE* and *AGR* benefited from message-level attention more than *CON*. Not only were *EXT* judgements no better than chance, but there was virtually no agreement among experts. This suggests that for some personality dimensions, individual messages have more or less relevance for personality, while for other dimensions there is little difference between messages (or at least it is difficult for both experts and our approach to capture differences).

In general, our proposed model seems to identify text that is informative of one's personality, both in terms of individual words and the overarching themes of the message as a whole, though this is easier for some dimensions than others. Modeling document relevance is useful, then, not just as a means to boost performance but as a tool to aid those seeking to better understand language.

8 Related Work

Personality modeling from language is becoming increasingly important for many social scientific

	Percent Preferred High		
	Expert 1	*Expert 2*	*Cohen's* κ
OPE	75%	75%	.60
CON	55%	55%	.60
EXT	55%	45%	.08
AGR	75%	75%	.76
NEU	40%	55%	.79

Table 6: Personality experts picked which of a pair of message sets were most informative for prediction. Each pair contained five of the highest and five of the lowest-weighted messages for a user. Table shows the percentage of instances where the expert selected the high-attention message set as most informative, as well as Cohen's κ inter-rater agreement.

applications. For example, Preoţiuc-Pietro et al. (2015) found personality features to be highly predictive of depression and PTSD. Lynn et al. (2017) demonstrated that the performance of document classification models can be improved by adapting to a variety of human factors, including personality. Personality has also been shown to be useful for deception detection (Fornaciari et al., 2013) and recommendation systems (Roshchina et al., 2011).

Most research on personality modeling focuses on the Big Five, or Five-Factor Model (Costa and McCrae, 1992). Personality is traditionally measured using questionnaires, but cost and scalability issues make computational methods preferable.

Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) features are popular for personality modeling (Yarkoni, 2010; Schwartz et al., 2013; Gjurković and Šnajder, 2018), as they readily provide insight into the type of language that correlates with certain personality dimensions. However, using predefined lexica is limiting; Schwartz et al. (2013) and Park et al. (2015) showed significantly improved prediction when using topics and n-grams extracted from their training set. When working with a very limited amount of data, Arnoux et al. (2017) found pre-trained word embeddings to be effective.

Deep learning approaches to personality prediction are limited. Majumder et al. (2017) used a convolutional neural network (CNN) with max pooling, alongside traditional document features (e.g. word count). Their best results were obtained when they filtered out sentences that did not contain strong emotion words (as determined via lexica) during preprocessing. This supports our intuition that some messages contain stronger signal than others, though our approach allows the model to identify such cases.

Yu and Markov (2017) also used CNNs with max- and average-pooling to predict personality over Facebook statuses. They experimented with fully-connected neural networks and bidirectional recurrent neural networks, but ultimately CNNs performed best. Both Majumder et al. (2017) and Yu and Markov (2017) used datasets that were significantly smaller than ours (n=2467 and n=9917, respectively) and their problems were framed as binary classification rather than regression[5].

9 Conclusion

Language-based personality prediction is an important task with many applications in social science and natural language processing. We presented a hierarchical sequence model with message- and word-level attention that learns to differentiate high- and low-signal messages. Our approach, which novelly models the idea that all messages are not equally valuable for psychological regression tasks, achieves new state-of-the-art results for personality prediction and provides insight into the relationship between language and personality. Our analysis demonstrates that the level of abstraction at which attention is applied can have a significant impact on a model's overall performance. Finally, this work highlights the critical role of *document relevance* as we progress with further human-centered natural language processing.

Acknowledgments

This work is supported in part by the National Science Foundation under Grant IIS-1815358. Data set used in grateful collaboration with Michal Kosinski and David Stillwell. We thank Google for supporting this research through the Google Cloud Platform credits. Thanks also to social and personality psychologists Sandra Matz and David Yaden for their help with the expert evaluation task.

References

Pierre-Hadrien Arnoux, Anbang Xu, Neil Boyette, Jalal Mahmud, Rama Akkiraju, and Vibha Sinha. 2017. 25 tweets to know you: A new model to predict personality with social media. *ICWSM*.

R Michael Bagby, Russell T Joffe, James DA Parker, Valery Kalemba, and Kate L Harkness. 1995. Major

[5]Personality theory suggests factors are better represented as continuous dimensions than discrete types (McCrae and Costa Jr, 1989).

depression and the five-factor model of personality. *Journal of Personality Disorders*, 9(3):224–234.

Benjamin P Chapman, Brent Roberts, and Paul Duberstein. 2011. Personality and longevity: knowns, unknowns, and implications for public health and personalized medicine. *Journal of aging research*, 2011.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

P.T. Costa and R.R. McCrae. 1992. *Revised NEO Personality Inventory (Neo-PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Psychological Assessment Resources.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tommaso Fornaciari, Fabio Celli, and Massimo Poesio. 2013. The effect of personality type on deceptive communication style. In *Intelligence and Security Informatics Conference (EISIC), 2013 European*, pages 1–6. IEEE.

Howard S Friedman and Margaret L Kern. 2014. Personality, well-being, and health. *Annual review of psychology*, 65:719–742.

Matej Gjurković and Jan Šnajder. 2018. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of Peoples Opinions, Personality, and Emotions in Social Media*, pages 87–97.

Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. 2006. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84–96.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Association for Computational Linguistics*.

Margaret L Kern, Johannes C Eichstaedt, H Andrew Schwartz, Lukasz Dziurzynski, Lyle H Ungar, David J Stillwell, Michal Kosinski, Stephanie M Ramones, and Martin EP Seligman. 2014. The online social self: An open vocabulary approach to personality. *Assessment*, 21(2):158–169.

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

Vivek Kulkarni, Margaret L Kern, David Stillwell, Michal Kosinski, Sandra Matz, Lyle Ungar, Steven Skiena, and H Andrew Schwartz. 2018. Latent human traits in the language of social media: An open-vocabulary approach. *PloS one*, 13(11):e0201703.

Veronica E. Lynn, Alissa Goodman, Kate Niederhoffer, Kate Loveys, Philip Resnik, and H. Andrew Schwartz. 2018. CLPsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46, New Orleans, LA. Association for Computational Linguistics.

Veronica E. Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In *Empirical Methods in Natural Language Processing*, pages 1146–1155.

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

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

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Robert R McCrae and Paul T Costa Jr. 1989. Reinterpreting the myers-briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality*, 57(1):17–40.

Robert R McCrae and Paul T Costa Jr. 1997. Personality trait structure as a human universal. *American psychologist*, 52(5):509.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

Walter Mischel, Yuichi Shoda, and Ozlem Ayduk. 2007. *Introduction to personality: Toward an integrative science of the person*. John Wiley & Sons.

Kevin R Murphy and Charles O Davidshofer. 1988. *Psychological Testing: Principles and Applications*. Pearson.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 21–30.

Alexandra Roshchina, John Cardiff, and Paolo Rosso. 2011. A comparative evaluation of personality estimation algorithms for the twin recommender system. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 11–18. ACM.

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. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Nakatani Shuyo. 2010. Language detection library for java.

Louis Leon Thurstone. 1934. The vectors of mind. *Psychological review*, 41(1):1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. 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.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373.

Jianguo Yu and Konstantin Markov. 2017. Deep learning based personality recognition from Facebook status updates. In *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, pages 383–387. IEEE.