Predicting Responses to Psychological Questionnaires from Participants' Social Media Posts and Question Text Embeddings

Huy Vu[1], **Suhaib Abdurahman**[2], **Sudeep Bhatia**[3], **Lyle Ungar**[3]
[1]Stony Brook University, [2]Free University of Berlin, [3]University of Pennsylvania
hvu@cs.stonybrook.edu, suhaib.abdurahman@gmail.com,
bhatiasu@sas.upenn.edu, ungar@cis.upenn.edu

Abstract

Psychologists routinely assess people's emotions and traits, such as their personality, by collecting their responses to survey questionnaires. Such assessments can be costly in terms of both time and money, and often lack generalizability, as existing data cannot be used to predict responses for new survey questions or participants. In this study, we propose a method for predicting a participant's questionnaire response using their social media texts and the text of the survey question they are asked. Specifically, we use Natural Language Processing (NLP) tools such as BERT embeddings to represent both participants (via the text they write) and survey questions as embeddings vectors, allowing us to predict responses for out-of-sample participants and questions. Our novel approach can be used by researchers to integrate new participants or new questions into psychological studies without the constraint of costly data collection, facilitating novel practical applications and furthering the development of psychological theory. Finally, as a side contribution, the success of our model also suggests a new approach to study survey questions using NLP tools such as text embeddings rather than response data used in traditional methods.

1 Introduction

Psychologists conduct personality research in order to understand what aspects and factors consistently distinguish people from each other on an individual level. This is relevant because personality influences important life outcomes such as occupational and educational success and even physical and mental health (Judge et al., 1999; Roberts et al., 2007).

Traditionally, psychologists measure personality through questionnaires, by having participants read and answer questions on a rating scale, for instance from "strongly disagree" to "strongly agree". However, acquiring questionnaire data in psychological research is often a tedious and costly process, as study participants must be recruited and motivated to complete questionnaires. This problem is particularly pronounced for longer surveys, which suffer from low completion rates and careless responses due to low participant motivation (Niessen et al., 2016; Van de Mortel et al., 2008; Raghunathan and Grizzle, 1995; Champion and Sear, 1969). Therefore, the ability to predict questionnaire responses would be of great use to researchers.

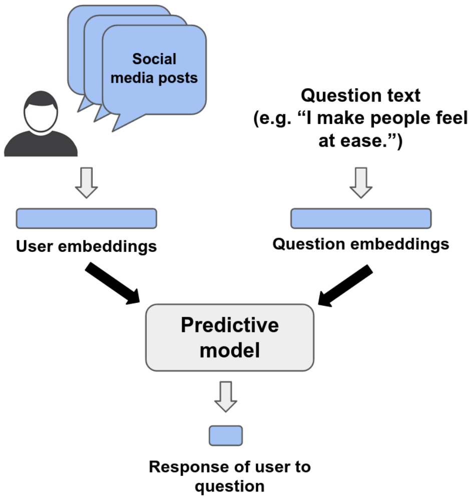

Social media posts
Question text (e.g. "I make people feel at ease.")
User embeddings
Question embeddings
Predictive model
Response of user to question

Figure 1: Overview of proposed task: analyzing users' social media text and questions text to predict responses.

The main contribution of this paper is to address this issue. We propose a system that uses the participants' social media texts and the question texts to predict the participants' responses. The system extracts BERT embeddings from the two input components and then trains a predictive model. After training, we can predict the responses for every new participant, new question or both, only requiring

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1512–1524
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

the participants' social media texts and the question texts. If our approach is successful, it will greatly reduce the costs of collecting response data for psychologists, especially when the number of new participants or questions is large.

Moreover, the success of our model also suggests a new approach to analyse psychological questionnaires by using Natural Language Processing (NLP). Traditionally, psychologists analyse questionnaires using only the participants' responses to the questionnaires, rather than the text and lexicon of the questions themselves (Cook and Beckman, 2006; Crocker and Algina, 1986). For instance, participants' responses are used to measure the similarity between two questions. However, these traditional approaches have high requirements for available response data which are costly to collect and moreover, lack the flexibility of integrating new participants or questions into their studies. In contrast, our novel approach of applying NLP into questionnaire research, as implemented in our model, offers the possibility of extending existing survey datasets and questionnaires to new subject populations and to new theoretical constructs, greatly improving the generalizability of psychological research and opening up many practical applications for personality research.

2 Related Work

2.1 Personality questionnaires

One of the most widely known and researched psychological personality models is the Five Factor or "Big Five" personality model. This comprehensive model categorizes human personality traits into five bipolar categories: *Openness to Experience*, *Conscientiousness*, *Extraversion*, *Agreeableness* and *Neuroticism* (Goldberg, 1993).

These categories are meant to describe a person's characteristic behaviors throughout different contexts of their daily life. The NEO-PI-R is one of the most established and widely accepted BIG 5 questionnaires (Costa and McCrae, 1989; Costa Jr and McCrae, 2008). As a proxy to the NEO-PI-R, this study uses the 100 question set from the publicly available International Personality Item Pool (IPIP), which is a large collection of questions for use in psychometric testing (Goldberg et al., 2006). This set of questions has been widely used in previous research such as Kulkarni et al. (2018); Park et al. (2015). Examples of questions measuring different categories are: ”I have a vivid imagination” (openness) or ”I do not mind being the center of attention” (extraversion). Each question is rated on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree). For each BIG 5 category, there are 20 questions which either increase or decrease the score of that specific category. In this paper, we call this the ”direction” of the questions. Examples of questions that share a category but have opposite directions are: ”I am easy to satisfy” (agreeableness - positive), ”I suspect hidden motives in others” (agreeableness - negative). The full list of 100 questions, along with their categories and directions can be found at: `https://ipip.ori.org/newBigFive5broadKey.htm`

When measuring personality using questionnaire responses, psychologists commonly ”reverse” the responses to negative questions to bring them in line with the positive questions. For example, a response of 1 to a negative question will be reversed to become 5 before further analysis. In this paper, we also use reverse-coding to pre-process all response data.

2.2 Predicting questionnaire responses

First, the cold start problem, meaning that for every new participant for whom we want to predict responses, we lack the initial information necessary to determine their similarity to the other participants in the data set. While using advanced participant information such as participants' social media text embeddings can help with that problem to some extent (Sedhain et al., 2014), a second issue remains: For every new question we add to the questionnaire, we lack information on how any new participant would answer it, meaning we cannot make any predictions for novel questions. For both problems, some responses need to be elicited from each participant and for each question before predictions can be made.

Our approach avoids this bottleneck by using a predictive model that can make predictions using the new participants' social media text or new questions' text embeddings, without requiring any response data for either new participants or questions.

2.3 Characterizing users by their social media text

There is increasing interest in estimating human personality from online data, including users' social media activity. Researchers, such as Sumner et al. (2011); Roberts et al. (2007); Golbeck et al. (2011); Argamon et al. (2007); H. Andrew Schwartz (2013) have included social media features in, for instance, collaborative filtering models. One common method used to map users' social media text to a vector embedding is Latent Dirichlet Analysis (LDA), such as in Schwartz et al. (2013); Kulkarni et al. (2018) to predict users' personality. More recently, other methods have been explored for mapping users to vectors, such as by Benton and Dredze (2018); Hallac et al. (2019).

2.4 BERT embeddings

The BERT model proposed by Jacob Devlin (2019) has become increasingly popular as an out-of-the-box and powerful pre-trained language model. Based on the idea of contextualized embeddings, BERT is a multi-purpose model for many downstream tasks and able to run efficiently thanks to parallel computation advantage of using transformers (Ashish Vaswani, 2017). Because of the capacity to capture contexts in both directions, an improvement over one-sided context models such as ELMO (Matthew E. Peters, 2018), sentence embeddings from BERT prove to be very strong features for many downstream tasks (Jacob Devlin, 2019). There are two main ways to use pre-trained BERT models. The first is by adding layers at the end of BERT and then fine-tuning the whole model end-to-end for the new downstream tasks. The second is to take pre-trained BERT embeddings, such as words or sentence embeddings, as input for subsequent models. In this study, we use BERT pre-trained embeddings to capture both the participants' social media texts and the questionnaire's question texts.

3 Dataset

3.1 Data description

We collected a dataset of 1000 Facebook users, each having at least 300 Facebook posts. For each user, we randomly picked 300 posts from their entire timeline. All selected users had posted at least 1000 words in total and were less than 65 years in age. Some sample posts are: *"Someone spoiled my good mood... :("*; *"I big thanks to all my friends that wished me a happy birthday."*; *"Day one at fair was totally fun. Wish you were here"*.

All users had responded to all 100 questions in the IPIP Big5 questionnaire using a custom application (Michal Kosinski, 2015). The responses have integer values from 1 to 5. As described above, the responses of "negative" questions were reversed before further analysis.

All participants explicitly acknowledged consent for their responses and Facebook information to be used for research purposes. All research procedures were approved by the University of Pennsylvania Institutional Review Board.

3.2 User and question embedding

- Question embeddings:
  We used the pre-trained BERT embeddings to capture question text semantics. The model used is BERT Large and Uncased (24 layers, 1024 dimensions). The word embeddings in each question are averaged to get the question embeddings. Embeddings from the last four BERT layers were concatenated to create an embedding vector of size 4096. We then standardize the data and apply Principal Component Analysis (PCA) (Ian T. Jolliffe1, 2016) to reduce the dimensions down to 55 to avoid overfitting, while keeping the variation explained at 0.9.

- User embeddings:
  We used the pre-trained BERT Base Uncased (12 layers, 768 dimensions) model to extract user features as follows: For each Facebook user, we randomly selected 300 posts from their timeline and then extracted the BERT embeddings from the words in these posts. The embeddings at the last four layers were averaged to get the word embeddings, which were then averaged to get the post embeddings, which were again averaged to get the user embeddings. The user embeddings were standardized, and PCA was used to reduce their dimension from 768 to 250, again keeping an explained variation of 0.9. The main reason we chose to average the last four embedding layers instead of concatenating them, as we did with the question embeddings, is because of the Facebook data's volume (hundreds of thousands of posts vs only 100 questions).

4 Experiments

We first conducted two experiments to separately test the quality of the question and user embeddings, and then a third main experiment in which both user and question features were used together to predict the response of a user to a question. We

- Used question embeddings to build separate models for each user that predict their response to novel questions.
- Used user embeddings to build separate models for each question to predict the response of a new user to that question
- Used both user and question embeddings to predict the response of a new user to a new question.

Since text embeddings of the questions for assessment questionnaires have not been explored in previous studies, the first and third tasks are novel and play a crucial role in exploring this new prediction approach. The second task, in which the users' have been characterized by their social media posts, has been explored previously. However, we will show that BERT embeddings outperform the traditional Latent Dirichlet Analysis (LDA) used in prior work.

Our main goal is to explore the novel idea of using the text embeddings of questions and of users to predict user responses to questions. Therefore, we do not focus on designing sophisticated deep learning models. Instead, we chose to use simpler but powerful, widely used models: ridge regression and K-nearest neighbors.

4.1 Testing question embeddings

Our first task sought to test the quality of question embeddings, asking how well BERT can capture the semantics of questions from a questionnaire. We did this by using question embeddings to build separate models for each user to predict their responses. Thus, For each user $u_{i_{th}}$ $(i = 1, ..., 1000)$, using 10-fold cross-validation, we trained a predictive model using 90 BERT question embeddings as input and the responses to the respective questions as labels, and then predicted responses on the 10 held-out questions. This novel task is important for this study because it shows that we can use text embeddings to capture the semantics of previously unseen questions and predict responses to those questions.

We trained a ridge regression model on the data set and optimised the regularization hyper-parameter $alpha$ for total L1-loss and correlation, using the predictions across all users. The hyper-parameter $alpha$ was tuned between $alpha = 1$ and $alpha = 1000$ (multiplied by 10 for each step). Similarly, we also trained a KNN model and optimised the number of neighbors k for total L1-loss and correlation. The hyper-parameter k was tuned between $k = 1$ and $k = 20$ (increased by 1 for each step).

The performance is measured by the correlation between the responses predictions and the groundtruth vectors as follows. For each user $u_{i_{th}}$, for $i = 1, ..., 1000$, we obtain a 10-folds (for each fold, training on 90 questions and testing on the left-out 10 questions) prediction vector $prediction\_u_{i_{th}}$, having the size of (1×100). We concatenate all prediction vectors of all users $prediction\_u_{i_{th}}$ into one single prediction vector $prediction\_u_{all}$ of size $(1 \times (1000 \times 100))$. Then the correlation between this prediction vector with the groundtruth vector $groundtruth\_u_{all}$ is calculated and reported.

We compared the models with a baseline, in which for each fold of each user, the mean of the responses on the training questions partition is used as predictions for the testing questions partition.

4.2 Testing user embeddings

Our second task used user embeddings to predict the response of a novel user to a given question. For each individual question $q_{i_{th}}$ $(i = 1, ..., 100)$, we trained a different model, predicting the response of any user with the BERT embedding of that user. I.e., for each question, we trained a separate model with 900 user embeddings as inputs and their response to the respective question as labels. The model was then tested on the 100 held-out users, using 10-fold cross-validation.

We trained the same models as in the previous task, again optimising the regularization parameter $alpha$ and the number of nearest neighbors k for total correlation and L1-loss, using the predictions across all questions. The hyper-parameter $alpha$ was tuned between $alpha = 1$ and $alpha = 100,000$ (multiplied by 10 for each step) and the number of neighbors k was tuned between $k = 5$ and $k = 450$ (increased by 5 for each step). Finally, we compared the models with a baseline, which used the mean of the responses for each individual question.

We also compared our models against the LDA method, where a user embedding is the proportion of each of a set of LDA topics in their Facebook posts. For our LDA-based personality prediction, we replicate the work of Kulkarni et al. (2018), that is, we extracted users features using 2000 publicly available LDA topics (at https://dlatk.wwbp.org/datasets.html?highlight=met_a30_2000_cp) learned from Facebook posts, which were created using the MALLET library (McCallum, 2002) with $alpha = 30$.

We seek to confirm the predictive quality of user LDA-based embeddings for predicting questionnaire responses, while also testing the relative performance of new feature extracting methods such as BERT over the older LDA (David M. Blei, 2003).

The performance is measured by the correlation between the responses predictions and the groundtruth vectors as follows. For each question $q_{i_{th}}$, for $i = 1, ..., 100$, we obtain a 10-folds (for each fold, training on 900 users and testing on the left-out 100 users) prediction vector $prediction\_q_{i_{th}}$, having the size of (1×1000). We concatenate all prediction vectors of all questions $prediction\_q_{i_{th}}$ into one sinlge prediction vector $prediction\_q_{all}$ of size $(1 \times (100 \times 1000))$. Then the correlation between this prediction vector with the groundtruth vector $groundtruth\_q_{all}$ is calculated and reported.

We also compared the models with a baseline, in which for each fold of each question, the mean of the responses on the training users partition is used as predictions on the testing users partition.

4.3 Combining user and question embeddings to predict responses

In our third task, the main predictive task of this study, we used both user and question embeddings to predict the response of a user to a question. This is a much more challenging task than the previous tasks, since the model must learn to generalize over both users and questions.

For evaluation, we divided the users and questions into 10 folds, testing on (user, question) pairs for which neither the user nor the question is in the training set. I.e., for the i^{th} loop, the i^{th} user fold and i^{th} question fold were kept as testing folds, while the model was trained on the remaining 9 user and question folds. Each training sample was created by combining one user embedding and one question embedding from the training folds. Since there were in total 1000 users and 100 questions, for each loop, we had 900 users and 90 questions for training, and 100 users and 10 questions for testing, resulting in $900 \times 90 = 81,000$ training samples, and $100 \times 10 = 1,000$ testing samples.

Again, we tested two models: ridge regression and K-nearest neighbors, as follows:

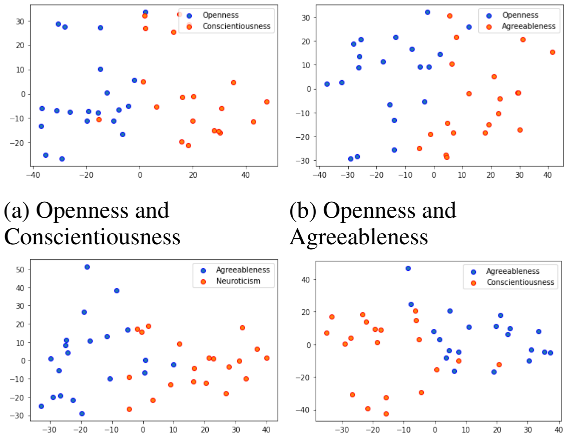
- Ridge regression:
  The embeddings of the users and questions were concatenated to one vector and used as input features for the model, with the responses of the corresponding user/question pair used as the label. Since user and question embeddings required different regularizations, we rescaled them with two separate hyperparameters $a_{question}$ and a_{user}, besides the model-wise $alpha$ weight decay for regularization. We then ran a grid search on the three hyperparameters: $a_{question}$, a_{user} and $alpha$ from 0.1 to $10,000$ (multiplied by 10 at each step) to look for the optimal set of hyper-parameters.

- K-nearest neighbors: We applied KNN separately for the user and question features. For each test sample, consisting of one testing user and one testing question, we searched for the k_{user} nearest users in the training set based on their user embeddings and the $k_{question}$ nearest questions in the training set based on their question embeddings. We then took the average of the responses of each of k_{user} nearest users to each of $k_{question}$ nearest questions as the prediction value. For regularization, we ran a grid search on k_{user} from 1 to 500 (increased by 25 at each step), and $k_{question}$ from 1 to 20 (increased by 1 at each step), and report the best performing set of hyper-parameters.

The reported correlation is calculated as follows. For each k_{th} fold with $k = 1, ..., 10$, a model is trained on the training partition of questions and users $(q_{k_{th}\_training\_fold} \times u_{k_{th}\_training\_fold})$ of size (90×900) and tested on the left-out testing partition of size (10×100), which can be flatten out to a vector $prediction\_q_{k_{th}}, u_{k_{th}}$ of size (1×1000). The predictions across 10 folds are then concatenated into one vector $prediction\_q_{all}, u_{all}$ of size $(1 \times (10 \times 1000))$. We then calculate the correlation between this concatenated vector and the groundtruth vector $groundtruth\_q_{all}, u_{all}$ and report the results.

Testing question embeddings on user level		
Model	**Corr**	**L1 Loss**
KNN (k=2)	0.275	1.14
Ridge (a=10)	**0.324**	**1.033**
Baseline	0.11	1.05
Testing user embeddings on question level		
Model	**Corr**	**L1 Loss**
KNN (k=200)	0.39	0.977
Ridge (a=1000)	**0.421**	**0.906**
LDA Ridge (a=10000)	0.403	0.917
Baseline	0.39	0.924
Test users and question embeddings		
Model	**Corr**	**L1 Loss**
KNN $(k_{user} = 100, k_{question} = 11)$	**0.220**	**1.087**
Ridge $(a_{user} = 0.01, a_{question} = 0.1,$ $a = 10)$	0.197	1.013
Baseline	0	1.095

Table 1: Main experiments predictions results



(a) Openness and Conscientiousness

(b) Openness and Agreeableness

(c) Agreeableness and Neuroticism

(d) Agreeableness and Conscientiousness

Figure 2: Visualizations of question embeddings for pairs of categories. Each point is the embedding for a question, projected on the first two principal components of the question embeddings. Questions about different factors of the BIG 5-factor model separate relatively cleanly

We compared the two models ridge regression and k-NN with a baseline, which simply takes the mean of all the responses within each training folds of questions and users as predictions.

5 Results and Discussion

5.1 Main results

5.1.1 Testing question embeddings

For the first task, we compare the best performance model for ridge regression, KNN and the baseline. Table 1 shows the highest correlation as $r = 0.324$ $(p < 0.05)$ for ridge regression with a regularization parameter of $alpha = 10$, compared to the baseline correlation of $r = 0.114$ $(p < 0.05)$. Questionnaire embeddings significantly improve predictions over the baseline.

Thus question embeddings in fact do have predictive power for individual user responses. To further support this view, we visualised the question embeddings on a 2D plane for each pair of categories on a one versus one scheme in figure 2. The figures show that BERT embeddings are able to capture the differences of personality categories fairly well, and suggest their potential for use in future applications that use personality information. The figure was created by applying PCA to the questions embeddings reducing the dimensions to 2 and then plot them on a 2D plane. The full plots for all pairs of categories can be found in Appendices.

Figure 2 illustrates the utility of using text embeddings to represent questionnaire questions. Psychologists commonly measure similarity between two questions by calculating the correlation of the *responses* to those questions. This works well–if one has collected user responses to the questions. Using BERT embeddings, in contrast, requires only the question's text; we can measure the semantic similarity between sentence pairs based on their distance in the embedding space, and thus to reduce the cost of data collection.

Caveat: To put these results into context, typically psychological variables have a 'correlational upper-bound' around 0.3 to 0.4 correlation (G J Meyer, 2001). Although our tasks are slightly different in the way that we measure correlations of users' responses to the personality questionnaire rather than the personality score as in (G J Meyer, 2001), but the value range of the correlations should be similar.

5.1.2 Testing user embeddings

We now examine the second task, in which we build a separate model for each question, in order to predict the response of a given user to that question. Again, we compare the best performing models for ridge regression and KNN against the baseline. Table 1 shows the highest correlation to be $r = 0.421$ $(p < 0.05)$ for the ridge regression with a regularization parameter of $alpha = 1000$, compared to the baseline correlation of $r = 0.390$ ($p < 0.05$).

We again see a significant improvement of prediction over the baseline. This experiment thus reconfirms the utility of user embeddings in predicting personality. Moreover, the results also show improvement compared to the older LDA model, which by itself is a strong model, proving that BERT embeddings are superior in capturing personality.

It should be noted that our user embeddings require much higher regularization than the question embeddings in task one ($k = 200$ vs. $k = 2$ for KNN and $alpha = 1000$ vs $alpha = 10$ for ridge regression), which suggest a much higher level of noise in the user embeddings. This is not surprising, since the question texts are specifically designed to only measure one among five categories. User embeddings, on the other hand, are created from an aggregation of social media posts, of which each can be about any topic. It should therefore be expected that user embeddings contain more noise than the question embeddings and thus require stronger regularization to avoid overfitting.

5.1.3 Combining user and question embeddings to predict responses

For this task, we reported the best performance model for ridge regression and KNN along with the baseline in Table 1.

We find the best correlation to be $r = 0.22$ ($p < 0.05$) for the KNN model ($k_{question} = 11, k_{user} = 100$), significantly higher than the baseline. It is thus possible to predict a user's response to a question using their social media text embeddings and the question text itself, even when neither user nor the question have ever been seen before. This is in stark contrast to collaborative filtering methods, which, for any new user or new question, always require some initial responses, as described in section 2.2.

The best performing model in this task has a correlation of $r = 0.22$ ($p < 0.05$), is better than baseline, but not as accurate as it would have been had one seen either the user (as in 4.2) or the question (as in 4.1) before. Generalizing over both users and questions is, not surprisingly, harder than generalizing over just one of them. The model is required to learn about two types of information, user and question embeddings, at the same time and across all users and all questions rather than on the individual user level or question level.

Testing question embeddings for each user		
Model	**Corr**	**L1 Loss**
KNN (k=5)	0.234	1.237
Ridge (a=1000)	**0.325**	**1.107**
Baseline	0.046	1.153

Table 2: Testing questions embeddings for each user with non-reversed responses.

We also find that, the ridge regression doesn't perform as well as the KNN model, in contrast to the first two experiments. A simple linear model concatenating users and questions is not able to compute how similar a question and a user are. (Beyond being a nonlinear relationship, remember that these embeddings are of different sizes.) KNN is a simple non-linear approach, and thus outperforms ridge regression. We expect that a reasonably designed neural network or deep learning model could improve these results substantially.

5.2 Additional Analysis

5.2.1 Testing questions embeddings without reverse-coding responses

As mentioned in section 2, it is common to reverse-code questionnaire responses; i.e., to transform the responses of negative questions (e.g. from a to $(5 - a + 1)$) to bring them in line with the positive questions. This transformation makes the prediction tasks easier because the model does not have to learn the direction (positive or negative) of the questions. However, we want to test whether our model can still perform well without reverse-coding information.

Since this task relies heavily on how well the BERT embeddings capture the direction of the questions, we reproduce the experiment in section 4.1 but with non-reverse-coded responses. The best performing ridge regression and KNN models are reported along with the baseline.

Table 2 shows our models' performance on non-reverse-coded responses. The ridge regression model, although confronted with a more challenging task, still has a correlation of up to 0.325 ($p < 0.05$) as high as when using reverse-coded responses. The KNN model has a correlation of 0.234 ($p < 0.05$), still significantly better than the baseline. This proves that even without direction information of the questions, our model can still perform well. We also find that in this sce-

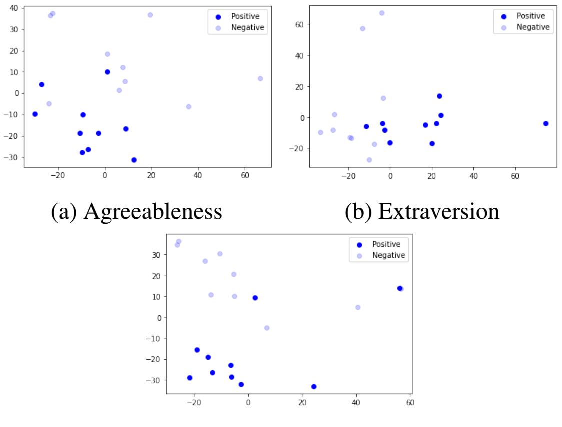


Figure 3: Visualizations of opposite direction questions within one category

Testing user embeddings on each question	
Model	**In-category L1-loss (O,A,E,C,N)**
KNN (k=200)	[0.84, 0.93, 1.07, 0.93, 1.12]
Ridge regression (a=1000)	[0.79, 0.86, 0.98, 0.87, 1.02]
Testing question embeddings on each user	
Model	**In-category L1-loss (O,A,E,C,N)**
KNN (k=5)	[0.91, 1.16, 1.02, 1.1, 1.22]
Ridge regression (a=100)	[0.91, 0.96, 0.93, 0.86, 1.16]

Table 3: Predictions results for each category: Openness (O), Agreeableness (A), Consciousness (C), Extraversion (E) and Neuroticism (N).

nario, the baseline has much more difficulty in giving good predictions, with a very low correlation ($r = 0.046, p < 0.05$) and high L1 loss. This is probably caused by the value range being distributed more uniformly between 1 and 5 without reverse coding.

In order for the model to perform much better than the baseline without reverse-coding, the questions embeddings must be able to capture not only the categories (O,C,E,A or N) but also the direction (negative and positive) of the questions. Indeed, this can be seen in the 2D plots in Figure 3, which show that using text embeddings, we can visually separate positive and negative questions within one category.

5.2.2 Prediction results for each category

For task 4.1 and 4.2, we additionally looked into the models' performance on each BIG 5 category. Table 3 shows the results of the best performing model for the first two tasks. The complete results for all regularization configurations can be found in Appendices.

- Regarding the predictions using user embeddings, table 3 shows the best performance in category *Openness*, followed by *Agreeableness*. The worst performance can be found in category *Neuroticism*. This might be partially explained by user activity on social media. Posts usually center around activities, experiences and feelings (Lai and To, 2015). These terms are usually associated with the first two categories.

- For the predictions using question embeddings, the results in table 3 show relatively inconsistent results for the two models. This might be caused by the relatively small sample of question embeddings (100) compared to the user embeddings (1000). However, what is consistent over both models is the lower performance of *Neuroticism* and *Agreeableness*. While *Neuroticism* is consistent with the results for the user embeddings, *Agreeableness* is surprising and opposite to the explanations stated previously. As such, future research regarding category-specific performance should be conducted to gain further insight into these differences.

6 Conclusion

Our study proposes a novel task: predicting responses of participants to a personality questionnaire, using their social media texts and the texts of the questions they are asked. Unlike prior work, we are able to successfully make out of sample predictions for both new survey questions and new participants. Our approach could potentially reduce the cost of data collection for psychologists, but more importantly our findings showcase a novel method for improving the generalizability of personality research. They also open up many novel applications that rely on existing social media and survey data to make predictions for out-of-sample participants and survey questions. Finally, our results offer the promise of improving psychological research by representing survey questions with informative text embeddings, which can be used by researchers and theorists to better understand the core dimensions of personality. We look forward to future work that integrates psychological theory with novel advances in natural language processing, to better measure, predict, and understand what distinguishes humans from each other.

References

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. 2017. Attention is all you need. *Advances in Neural Information Pro-cessing Systems.*

Adrian Benton and Mark Dredze. 2018. Using author embeddings to improve tweet stance classification. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 184–194.

Dean J Champion and Alan M Sear. 1969. Questionnaire response rate: A methodological analysis. *Social Forces*, pages 335–339.

David A Cook and Thomas J Beckman. 2006. Current concepts in validity and reliability for psychometric instruments: theory and application. *The American journal of medicine*, 119(2):166–e7.

PT Costa and RR McCrae. 1989. Neo five-factor inventory (neo-ffi). *Odessa, FL: Psychological Assessment Resources*, 3.

Paul T Costa Jr and Robert R McCrae. 2008. *The Revised NEO Personality Inventory (NEO-PI-R).* Sage Publications, Inc.

Linda Crocker and James Algina. 1986. *Introduction to classical and modern test theory.* ERIC.

Michael I. Jordan David M. Blei, Andrew Y. Ng. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research.*

L D Eyde G G Kay K L Moreland R R Dies E J Eisman T W Kubiszyn G M Reed G J Meyer, S E Finn. 2001. Psychological testing and psychological assessment. a review of evidence and issues. *Am Psychol*, 56(2).

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 149–156. IEEE.

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

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.

Margaret L. Kern Lukasz Dziurzynski Stephanie M. Ramones Megha Agrawal Achal Shah Michal Kosinski David Stillwell Martin E. P. Seligman Lyle H. Ungar H. Andrew Schwartz, Johannes C. Eichstaedt. 2013. *Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach.*

Ibrahim R Hallac, Semiha Makinist, Betul Ay, and Galip Aydin. 2019. user2vec: Social media user representation based on distributed document embeddings. In *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–5. IEEE.

Jorge Cadima Ian T. Jolliffe1. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of The Royal Society A.*

Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *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).*

Timothy A Judge, Chad A Higgins, Carl J Thoresen, and Murray R Barrick. 1999. The big five personality traits, general mental ability, and career success across the life span. *Personnel psychology*, 52(3):621–652.

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).

Linda SL Lai and Wai Ming To. 2015. Content analysis of social media: A grounded theory approach. *Journal of Electronic Commerce Research*, 16(2):138.

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

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

Samuel D Gosling Vesselin Popov David Stillwell Michal Kosinski, Sandra C Matz. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist.*

Thea F Van de Mortel et al. 2008. Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, The*, 25(4):40.

A Susan M Niessen, Rob R Meijer, and Jorge N Tendeiro. 2016. Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, 63:1–11.

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(6):934.

Trivellore E Raghunathan and James E Grizzle. 1995. A split questionnaire survey design. *Journal of the American Statistical Association*, 90(429):54–63.

Brent W Roberts, Nathan R Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R Goldberg. 2007. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological science*, 2(4):313–345.

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.

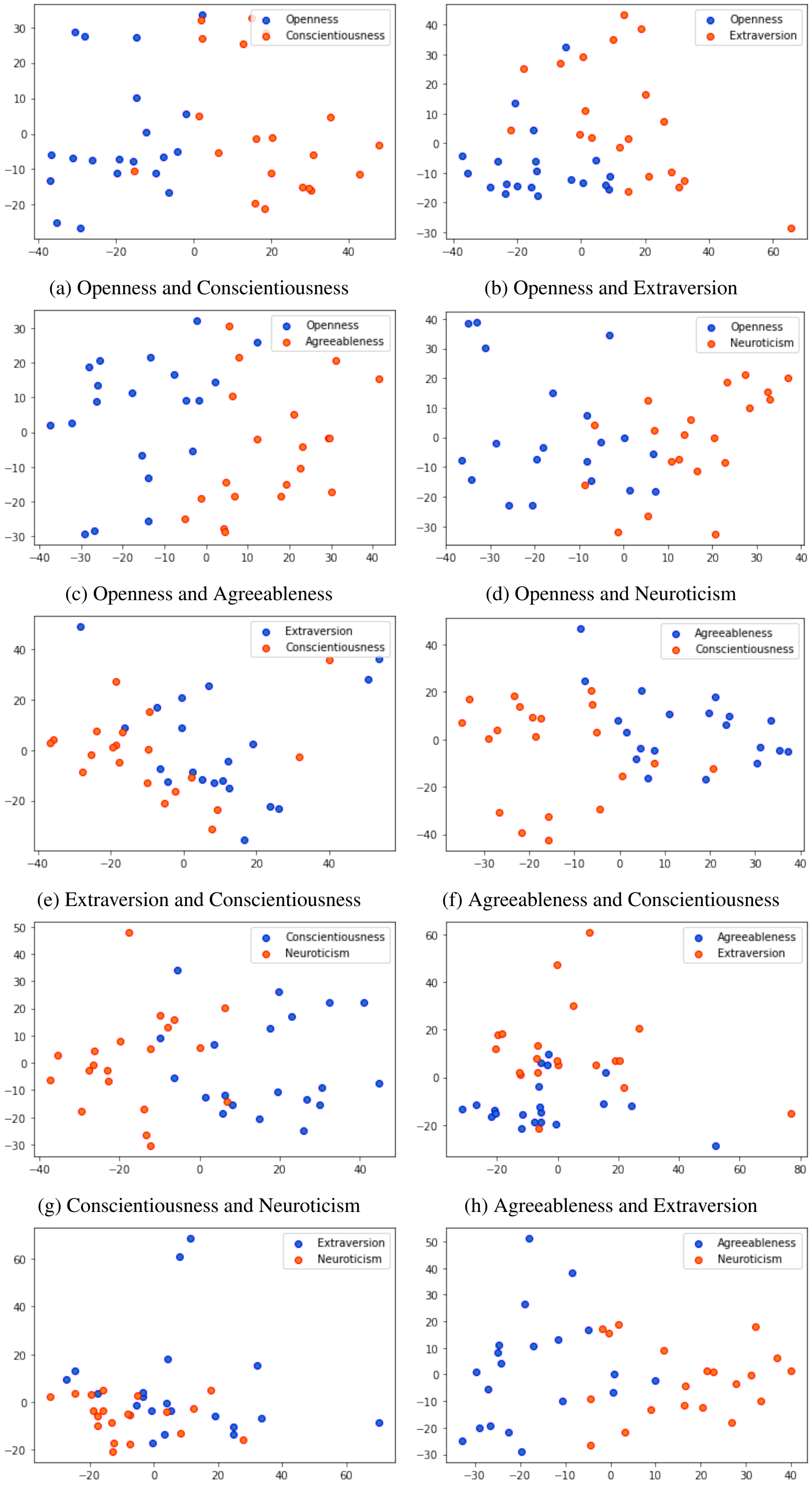
Suvash Sedhain, Scott Sanner, Darius Braziunas, Lexing Xie, and Jordan Christensen. 2014. Social collaborative filtering for cold-start recommendations. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 345–348.

Chris Sumner, Alison Byers, and Matthew Shearing. 2011. Determining personality traits & privacy concerns from facebook activity. *Black Hat Briefings*, 11(7):197–221.

A Appendices

Appendices include:

- The full visualizations of questions embeddings , for each pair of categories and opposite directions within each category.
- Full results of task 1 and 2 described in 4.1 and 4.2 with choices of regularizations, correlations and L1 loss for each category separately.



(a) Openness and Conscientiousness

(b) Openness and Extraversion

(c) Openness and Agreeableness

(d) Openness and Neuroticism

(e) Extraversion and Conscientiousness

(f) Agreeableness and Conscientiousness

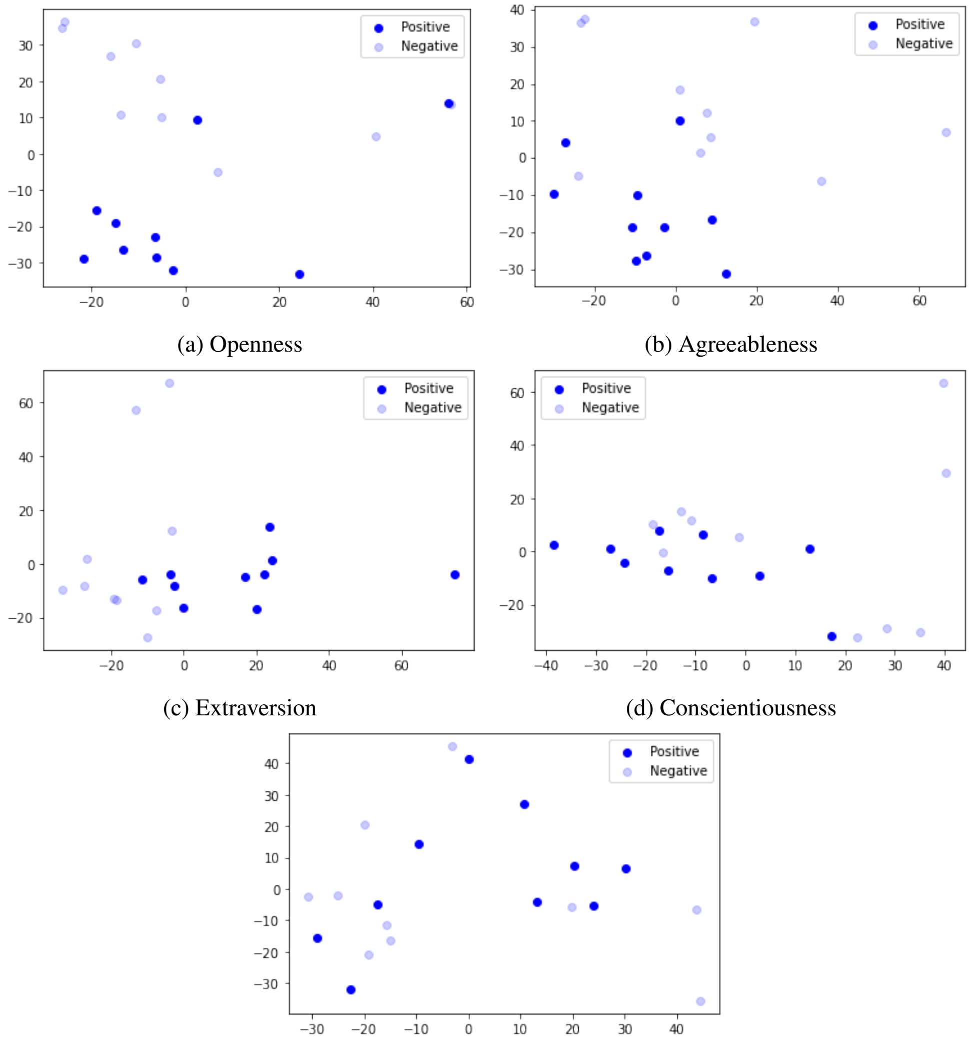
(g) Conscientiousness and Neuroticism

(h) Agreeableness and Extraversion

(i) Extraversion and Neuroticism

(j) Agreeableness and Neuroticism

Figure 4: Visualization of embeddings for each pair of categories. Each dot represents a question from the respective BIG 5 category. The Visualizations show that sentence embeddings are able to separate questionnaire questions by category.



(a) Openness

(b) Agreeableness

(c) Extraversion

(d) Conscientiousness

(e) Neuroticism

Figure 5: Visualization of embeddings for both question directions (positive vs. negative) in each category. The visualizations show that sentence embeddings can distinguish the direction of a questionnaire question reasonably well (for categories Openness, Extraversion and Agreeableness).

Fixed users, testing questions embeddings				
Model	**Corr**	**L1 Loss**	**In-construct Corr** **(O, A, E, C, N)**	**In-construct L1** **(O, A, E, C, N)**
KNN (k=1)	0.257	1.104	[0.33, 0.15, 0.2, 0.27, 0.22]	[0.85, 1.13, 1.17, 1.06, 1.3]
KNN (k=2)	**0.275**	**1.14**	**[0.24, 0.12, 0.28, 0.25, 0.36]**	**[1.25, 1.2, 1.15, 1.07, 1.03]**
KNN (k=5)	0.262	1.083	[0.25, 0.1, 0.32, 0.19, 0.26]	[0.91, 1.16, 1.02, 1.1, 1.22]
KNN (k=10)	0.201	1.118	[0.23, 0.08, 0.32, 0.19, 0.12]	[0.86, 1.16, 1.02, 1.09, 1.47]
KNN (k=20)	0.141	1.153	[0.18, 0.11, 0.27, 0.16, -0.01]	[0.86, 1.11, 1.05, 1.11, 1.63]
Ridge regression (a=1)	0.311	1.116	[0.32, 0.21, 0.29, 0.26, 0.29]	[0.93, 1.2, 1.14, 1.12, 1.19]
Ridge regression (a=10)	**0.324**	**1.033**	**[0.34, 0.23, 0.32, 0.28, 0.28]**	**[0.88, 1.09, 1.04, 1.01, 1.13]**
Ridge regression (a=100)	0.302	0.966	[0.32, 0.25, 0.35, 0.3, 0.17]	[0.91, 0.96, 0.93, 0.86, 1.16]
Ridge regression (a=1000)	0.167	1.031	[0.19, 0.19, 0.32, 0.24, -0.06]	[1.03, 1.0, 0.95, 0.89, 1.28]
Fixed questions, testing users embeddings				
Model	**Corr**	**L1 Loss**	**In-construct Corr** **(O, A, E, C, N)**	**In-construct L1** **(O, A, E, C, N)**
KNN (k=5)	0.3	1.126	[0.2, 0.31, 0.14, 0.18, 0.15]	[0.91, 1.07, 1.22, 1.14, 1.29]
KNN (k=10)	0.338	1.07	[0.22, 0.35, 0.16, 0.21, 0.19]	[0.87, 1.0, 1.18, 1.07, 1.24]
KNN (k=15)	0.348	1.049	[0.23, 0.36, 0.16, 0.24, 0.2]	[0.85, 0.99, 1.16, 1.02, 1.22]
KNN (k=30)	0.371	1.014	[0.24, 0.37, 0.18, 0.27, 0.22]	[0.84, 0.96, 1.13, 0.96, 1.18]
KNN (k=50)	0.378	1.002	[0.25, 0.38, 0.19, 0.27, 0.23]	[0.83, 0.95, 1.11, 0.95, 1.16]
KNN (k=200)	**0.39**	**0.977**	**[0.24, 0.39, 0.22, 0.27, 0.25]**	**[0.84, 0.93, 1.07, 0.93, 1.12]**
KNN (k=450)	0.387	0.982	[0.24, 0.39, 0.22, 0.26, 0.25]	[0.83, 0.93, 1.08, 0.94, 1.13]
Ridge regression (a=1)	0.341	0.999	[0.23, 0.33, 0.21, 0.26, 0.22]	[0.88, 0.95, 1.08, 0.96, 1.12]
Ridge regression (a=10)	0.346	0.992	[0.24, 0.33, 0.22, 0.26, 0.22]	[0.87, 0.95, 1.07, 0.96, 1.11]
Ridge regression (a=100)	0.372	0.955	[0.25, 0.36, 0.24, 0.28, 0.24]	[0.84, 0.91, 1.03, 0.91, 1.08]
Ridge regression (a=1000)	**0.421**	**0.906**	**[0.28, 0.41, 0.27, 0.32, 0.28]**	**[0.79, 0.86, 0.98, 0.87, 1.02]**
Ridge regression (a=10000)	0.412	0.911	[0.26, 0.4, 0.25, 0.29, 0.27]	[0.8, 0.87, 0.98, 0.88, 1.02]
Ridge regression (a=100000)	0.399	0.921	[0.24, 0.38, 0.23, 0.27, 0.25]	[0.81, 0.89, 0.98, 0.89, 1.03]

Table 4: Full results for testing questions embeddings on the individual user level and testing users embeddings on the individual question level. BIG 5 Categories as: Openness (O), Agreeableness (A), Extraversion (E), Consciousness (C) and Neuroticism (N).