REPHRAIN

Protecting citizens online

You Are What You Read: Inferring Personality From Consumed Textual Content

Adam Sutton - University of Bristol

Almog Simchon - University of Bristol

Matthew Edwards - University of Bristol

Stephen Lewandowsky - University of Bristol

September 2023

UK Research and Innovation · University of Bristol · UCL · The University of Edinburgh · King's College London · University of Bath

You Are What You Read: Inferring Personality From Consumed Textual Content

Adam Sutton and **Almog Simchon** and **Matthew Edwards** and **Stephan Lewandowsky**
University of Bristol
United Kingdom

Abstract

In this work we use consumed text to infer Big-5 personality inventories using data we have collected from the social media platform Reddit. We test our models on two datasets, sampled from participants who consumed either fiction content ($N = 913$) or news content ($N = 213$). We show that state-of-the-art models from a similar task using authored text do not translate well to this task, with average correlations of $r = .06$ between the model's predictions and ground-truth personality inventory dimensions. We propose an alternate method of generating average personality labels for each piece of text consumed, under which our model achieves correlations as high as $r = .34$ when predicting personality from the text being read.

1 Introduction

While authored text has previously been used for personality prediction (Eichstaedt et al., 2021), consumed text (the linguistic attributes of the text that people choose to read) has received no such attention. It is unclear if consumed text can be indicative of personality. Methods used in similar tasks may also not transfer to this domain, as a consumed piece of text is not unique to a single reader.

However, predicting reader personalities may help understand and reduce the impact of psychological micro-targeting, particularly in the domain of political advertising. Facebook has a psychological micro-targeting patent registered (Nowak and Eckles, 2014), and previous work indicates personality targeted messages increase desirable outcomes for advertisers (Matz et al., 2017).

As recent work has shown that targeted political advertising online has been more more effective than traditional methods (Zarouali et al., 2020; Goldberg et al., 2021; Tappin et al., 2022; Joyal-Desmarais et al., 2022), our work aims to reverse engineer the process of such psychological targeting, with the intention of developing countermeasures to remove or reduce the impact of this targeting. Here we first demonstrate that consumed text can be used to infer personality. This is significant, as we show that personality prediction of content consumers is possible even where there are not structural connections to known cultural touchpoints (which has been demonstrated previously by, e.g., Youyou et al. (2015)). Prediction of consumer personality from consumed text is highly tranferrable, being in principle applicable to any platform where users might read text. In the future we aim to develop tools for users that would flag articles or text that our model predicts could be congruent with their personality inventory.

In this paper we sample ∼1,100 participants from the social media website Reddit, using their public data and provided personality inventories to show that consumed text can also be indicative of the consumer's personality. Our models achieve Pearson's $r > 0.3$ between predicted personality dimension values and those provided using standard instruments. We also show that models that have achieved state-of-the-art performance when applied to produced text do not achieve suitable performance on consumed text.

2 Background

In the field of psychology, constructs such as personality are quantified using validated tools. One such tool is a personality inventory, where the outcome is usually represented as a numerical value for multiple personality dimensions. One example of this is the Big-5 model, which uses a questionnaire to capture people's personality along five dimensions (Soto and John, 2017; Goldberg, 1993). These scales enable measurement of personality, and in combination with access to large feature-rich datasets from social media they have enabled attempts at estimating people's personality from their behaviour (Bachrach et al., 2012; Schwartz et al., 2013). Machine learning has improved to

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 28–38

the point where automated personality judgements can outperform humans at the same task (Youyou et al., 2015). This has also been found with textual content: various studies have shown that text produced by a user can be used to estimate their personality (Eichstaedt et al., 2021).

Language models have played a large part in the improvement of performance in many downstream natural language tasks in recent years (Pennington et al., 2014). The most recent development to have a substantial performance impact is attention (Vaswani et al., 2017; Bahdanau et al., 2014), which enables word representations that are dynamically generated based on surrounding text (i.e. "bark" will have different representations for a "dog's bark" and "tree bark"). This has resulted in a new generation of attention-based language models that reported state-of-the-art-performance for multiple NLP tasks (Devlin et al., 2018). These architectures are still being iterated on to improve performance (Zhong et al., 2022; Patra et al., 2022).

Attention has also been useful for personality modelling in the domain of produced text. Lynn et al. (2020) defines "message-level attention", which is based on the assumption that "not all documents are equally important". Models using this form of attention take multiple produced messages from an author and weigh the importance of each message according to a learned attention mechanism, in order to predict that author's personality. Lynn et al. (2020) represent the current state-of-the-art performance for this task, while also providing some interpretability of the model via message weights.

In this study we apply these message attention models to the domain of consumed text, alongside an alternative method that aims to predict the averaged personality profile of all known consumers of the article. We show that message attention models do not achieve desirable performance when applied to consumed text. Evaluation under averaged personality labelling shows promising performance in comparison. Our evaluation covers multiple datasets, spanning two different genres of text. We also trial the effectiveness of models predicting consumer personality using only article titles instead of the entire article. We find that the personalities of news readers are better predicted by our models than the consumers of fictional content, and predictions on the basis of news titles alone perform comparably to those informed by the content of the entire article.

3 Methods

3.1 Message Level Attention

For each personality dimension, given a set of N messages (or articles consumed) from a user u we encode each textual input ($article_i$) such that:

$$s_i = \Phi(article_i), \tag{1}$$

where Φ is the language model used to encode each consumed article. We then pass all vector representations (such that all $s_i \in S$) through another sequence model, multi-headed self-attention (MHA) (Vaswani et al., 2017):

$$S' = \text{MHA}(S). \tag{2}$$

We then apply the message attention mechanism to calculate articles that are most indicative of the personality of a given user, as proposed in Lynn et al. (2020):

$$h_i = \tanh(W_m s'_i + b_m) \tag{3}$$

$$m_i = \frac{\exp(h_i^\top h_m)}{\sum_{j=0}^{N} \exp(h_j^\top h_m)}, \tag{4}$$

where W_m and b_m are learned features for the encoders hidden state. h_m is a learned vector that judges how much attention should be paid to each article. Equation 4 is a softmax where all m_i will sum to 1.

Each value in m is a scalar that represents how important the attention mechanism considers its corresponding article vector s_i is, and scales it accordingly.

$$\hat{u} = \sum_{i=0}^{N} m_i s_i. \tag{5}$$

Equation 5 shows how the user representation is formulated using the weighted average summation of each article consumed by a user. The vector representation of the user is passed into a standard feed forward neural network such that:

$$f : \mathbb{R}^d \rightarrow \mathbb{R}, \tag{6}$$

which results in prediction of a single personality dimension for that user. Each personality dimension will have its own message attention and user representation weights calculated separately, to fine tune as accurately as possible.

3.2 Average Personality Per Article

With produced text it can be reasonably assumed that all messages produced are unique to that user. However that assumption does not hold for consumed text, as it is not intended to be unique to a single reader.

Our method assigns a single personality label for each article. The personality label for a given article is comprised of the average personality score of all participants who are known to have consumed that article. The underlying assumption of this method is that each article will target a large group of people that have an aggregate personality profile indirectly indicated in the text. We also assume that the average personality of known consumers is the likeliest approximation to the personality profile of the intended audience.

In contrast to our evaluations of message-level attention models, under this methodology only one article per training step is fed into the neural network. Labels in this model are the average personality of every user who has consumed the article.

For a given $article$ we calculate the language model representation as:

$$s = \Phi(article), \tag{7}$$

where Φ is the language model used to encode the article to a vector representation.

Unlike Section 3.1, no further processing is required to generate a vector used to predict the targeted personality, and a feed forward network is again used to estimate the personality of the average consumer of this article.

4 Dataset

Two datasets are used in our experiments, both sourced from the social media website Reddit between 2021-2022. Participants were invited to participate in a survey and gave permission for us to link their public post and comment history to their personality inventories as assessed via a Big-5 personality questionnaire (BFI-2) (Soto and John, 2017). We crawled the content of all posts our participants had commented upon, using commenting behaviour as an indication of text consumption. Our data collection and retention procedures were overseen by the relevant institutional ethics board.

Our two datasets cover different domains of content. Our news dataset contains news articles consumed by our participants from news-focused subreddits (communities dedicated to a specific topic), including *r/worldnews*, *r/politics* and *r/europe*. Our fiction dataset contains short fiction consumed by participants from subreddits devoted to sharing such content, such as *r/WritingPrompts*, *r/shortstories*, and *r/nosleep*.

Table 1: Number of users and articles that have been consumed for both datasets used in experiments. Note that these users may have consumed text from both domains.

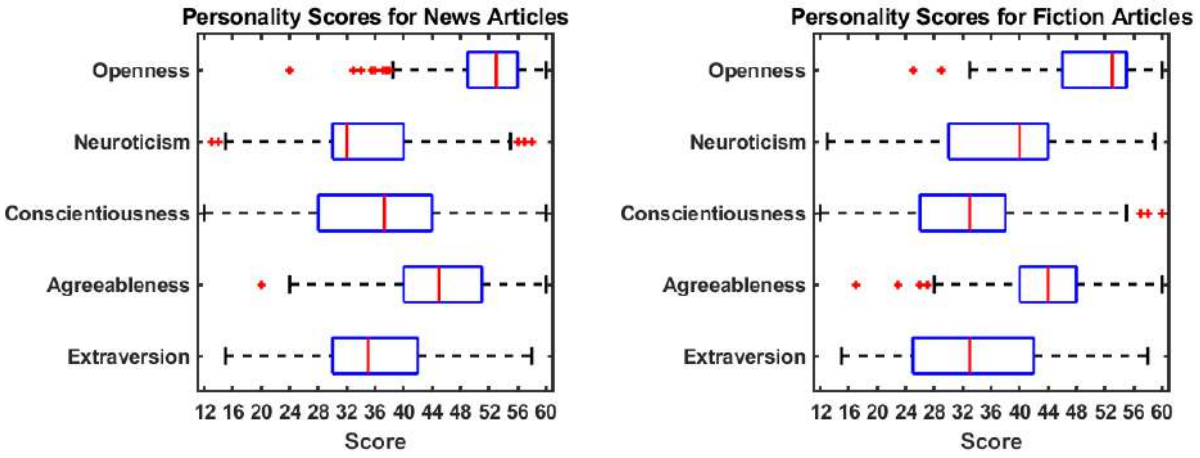
	News	Fiction
Users	213	953
Articles	19,609	4,000

News subreddits were chosen based on activity (number of users, and posts) and the majority of articles posted being URL submissions linking to news stories from external news sites. These subreddits are also moderated to remove unwanted content, such as spam or adverts. Fiction subreddits were also chosen based on activity, along with ease of crawling for the text content posted there. Text content is usually short stories which are submitted as a post, or in the case of *r/WritingPrompts* as top level comments.

Table 1 details the number of participants and articles that have been gathered through our sampling process. Active Reddit users engage with many articles, but engagement is not evenly distributed: some articles are consumed by only a single user, while other articles were consumed by hundreds of our participants. This leads to some imbalance and uncertainty in our average-personality labelling: it is possible that articles consumed by fewer of our participants give a single consumer's personality disproportionate weight.

Figure 1 shows the distribution of the Big-5 personality traits as aggregated by fiction and news articles as per our method described above. Personality labels at the article level show somewhat reduced variance compared to the per-user data (see Appendix C), but are by no means uniform. The personality distributions of news and fiction consumers are quite similar, seeming to reflect a common Reddit user personality type.

Figure 1: Box plots showing the distributions of personality scores per article. On each box, the central mark indicates the median, and the left and right edges of the box indicate the 25th and 75th percentiles, respectively. Outliers are plotted beyond the whiskers.



5 Results

In this work we primarily seek to answer the following questions:

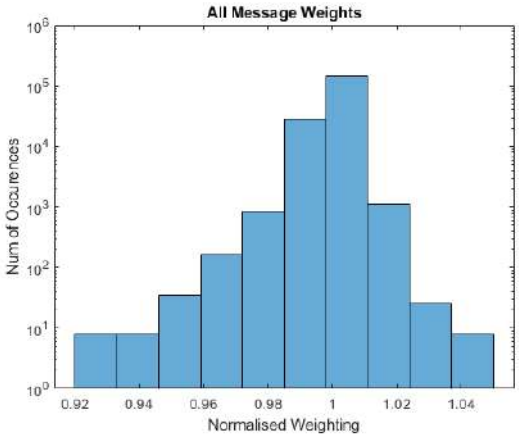
- Can a person's personality inventory be inferred from the textual content they consume?
- Do state-of-the-art models for produced textual content achieve "good" performance when applied to consumed text?
- Does the domain of consumed text affect predictive performance in these tasks?

5.1 Message-Level Attention Results

Message-level attention is considered to achieve state-of-the-art performance when predicting personality based on text that is produced by users (Lynn et al., 2020; Eichstaedt et al., 2021). It may be reasonable to presume that these models would be good for the similar task of predicting personalities based on text consumed by users. In these experiments each article is passed through a Sentence-BERT language model (Reimers and Gurevych, 2019) to generate a vector representation for each article, which is then passed to the message-level attention model as described in Section 3.1. Appendix A provides more detail on our experimental setup for all models.

Three different models are trained; one using the fiction dataset, and two using the news dataset. The textual input for the two news dataset models differed, with one model trained using the article bodies in the same manner as for the fiction dataset, and one model using the new article titles alone. Fiction titles have not been considered as a textual input due to the format of titles in the chosen fiction

Figure 2: Histogram of all normalised message weights used in weighted sums to generate user vectors. A value being close to 1 represents an article that is weighed as important as it would be to a uniformly weighted mean. Higher weights represent more informative articles according to the message attention mechanism.



subreddits; *r/WritingPrompts* titles are written as prompts for commenters to write their own fiction, while *r/shortstories* titles include authors, series tags, and other meta-text.

Table 2 shows the 5-fold cross-validation performance of all three models that are trained using message attention. The performance of all these models is degraded in comparison to the results obtained by Lynn et al. when using produced text (Lynn et al., 2020). The model trained on fiction articles produced the best performance of the three, and the lowest variance in performance.

Message attention models learn a weighting function that weighs the relative importance of each

Table 2: 5-fold cross-validation performance of message attention models attempting to predict personality inventories from consumed text. We show results for two different domains of text: fictional stories and news articles. We also show the performance of models trained just on news titles instead of the entire news article. We report the average Pearson's *r* across all 5 folds and intervals of one standard deviation.

Domain	Fiction	News	
Content	Articles	Articles	Titles
Extraversion	**0.08 [0, 0.16]**	0.05 [-0.06, 0.16]	0.06 [-0.06, 0.18]
Agreeableness	**0.1 [0.06, 0.14]**	0 [-0.2, 0.19]	0.01 [-0.11, 0.14]
Conscientiousness	0.07 [0.04, 0.11]	0.11 [-0.04, 0.26]	**0.13 [-0.05, 0.31]**
Neuroticism	**0.07 [-0.01, 0.16]**	0.04 [-0.05, 0.14]	0.01 [-0.14, 0.17]
Openness to Experience	0.04 [0, 0.09]	-0.02 [-0.2, 0.16]	**0.1 [-0.14, 0.33]**

article when generating the user vector. We can look at the distribution of these weights for each article to see if they are finding certain articles to be more informative than others.

We extract all message-level attention weights in order to examine the distribution. In the model these weights are used to create the user vector, with the weights contributing to the weighted sum of each article. If a user's personality is predicted given N arguments then a uniform weighting would be $1/N$ for each article. Each user may have consumed a different number of news articles, so to normalise this we divide each weight we obtain by $1/N$. If a normalised weight is < 1 then the model estimated that the article is less informative than average in predicting a user's personality. If a normalised weight is > 1 then the model has estimated that the article is more important to predicting that user's personality.

Figure 2 shows a histogram of all attention weights that are used in the validation steps of all three message attention models. All attention weights in these models are close to equally weighted for every article. This indicates that the model is predicting that none of these consumed articles are more or less informative than any other in predicting a user's personality.

5.2 Average Personality Per Article

For this experiment, all articles have a corresponding label that is the average personality score of all participants who have consumed the article. Our model of consumption is that an article has been consumed if the user has commented on a Reddit post that links to that article. In these experiments the language model used is the same as the previous models, with Sentence-BERT providing a vector representation for each article. The same input language model is used for a fair comparison between message attention and average personality models. The model is described in detail in Section 3.2.

Three models are trained again using the same fiction, news article and news title inputs as described in Section 5.1, this time with an average personality label for each article.

Table 3 shows the 5-fold cross-validation performance of all three models that are trained with averaged labels for each article. Model performance is much improved when compared to the results for the message attention approach. The variance within k-fold performance is also decreased, showing a more consistent performance between models. Models trained using news article text have generally better performance than models trained using fiction, with the exception of personality dimension of Openness. Our news titles model achieves similar performance to the model trained using the entire news article.

5.3 Visualisations

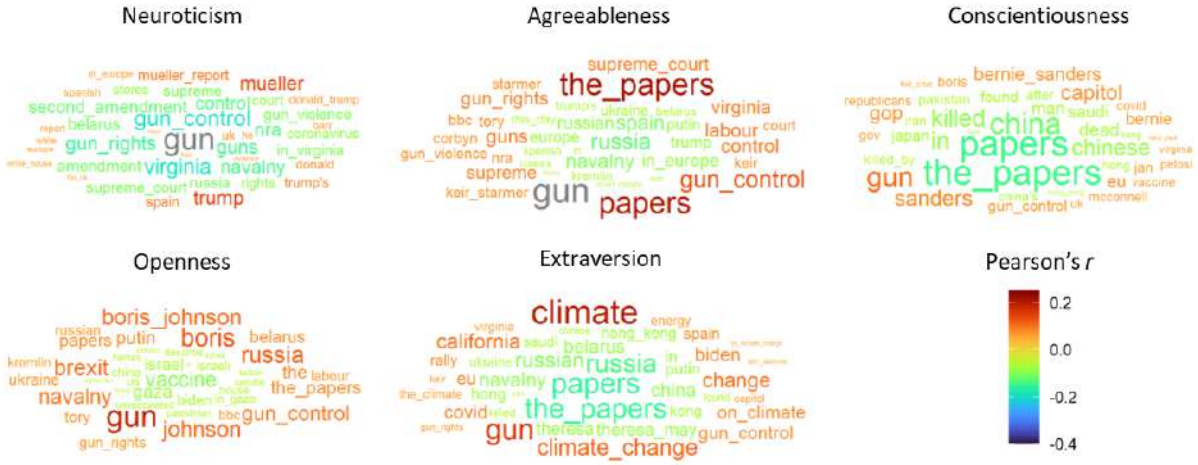
We generated word clouds to understand which words and phrases were most strongly correlated with each personality dimension. This was achieved by taking the validation set predictions from each fold and examining which n-gram phrases (1,2,3-gram) were most correlated with each personality dimension.

Figure 3 shows the word clouds for news titles when using the average-label method, and the n-grams that most correlate with each personality dimension. The word clouds show that words related to article content, rather than stylistic features, are most correlated with personality features of the text's consumer. The particular phrases visualised also represent major news stories that occurred during the period of data collection.

Table 3: 5-fold cross-validation performance of average-label models attempting to predict personality inventories. We show results for two different domains of text: fictional stories and news articles. We also show the performance of models trained just on news titles instead of the entire news article. We report the average Pearson's *r* across all 5 folds and and intervals of one standard deviation.

Domain	Fiction	News	
Content	Articles	Articles	Titles
Extraversion	0.29 [0.27, 0.30]	**0.3 [0.28, 0.31]**	0.27 [0.26, 0.28]
Agreeableness	0.12 [0.09, 0.14]	0.21 [0.2, 0.22]	**0.23 [0.22, 0.24]**
Conscientiousness	0.13 [0.09, 0.16]	**0.25 [0.23, 0.27]**	0.23 [0.21, 0.25]
Neuroticism	0.13 [0.11, 0.16]	0.33 [0.31, 0.35]	**0.34 [0.32, 0.35]**
Openness to Experience	**0.23 [0.2, 0.27]**	0.17 [0.15, 0.20]	0.15 [0.13, 0.17]

Figure 3: 1,2,3-gram word cloud of news titles. This shows top 20 positive and top 20 negative features that best predict each personality dimension. The colour of the n-grams indicate correlation to the dimension. The size of the n-gram is the absolute value of that correlation.



6 Discussion

Our experiments have tackled a novel personality prediction task using consumed rather than produced text to infer personality. We demonstrate that the structure of this task is important, showing a significant difference in performance between the state-of-the-art message attention model and average-labelling models. The message-level attention mechanism used by Lynn et al. (2020) does not perform well on this problem, and our examination of attention weights suggests this is because the model rarely finds any consumed text to be especially informative with regard to a participant's personality. We also find that the source of the corpus also has an impact on performance, with models trained upon news text better suited to predicting personality.

6.1 A failure of message attention

There may be multiple factors that explain the poor performance of message-level attention models on this task. The fundamental assumption of these models that is some pieces of text are more important than others when predicting personality. This assumption seems to hold true when using produced texts, given the performance of those models. However, this does not appear to have held for consumed text.

One factor that may be key to explaining this result is that produced text can usually be assumed to be somewhat unique. There may be some comments that are commonly produced such as common questions or short phrases in reply to another user, but the majority of comments that are produced by a typical user would generally be unique text sequences. In contrast, news articles and pieces of fiction submitted to Reddit are intended to reach

a wide audience, and so would take the role of common factors linked to the many unique personality profiles of all their consumers. Each article that is consumed by multiple people that have differing personality scores could confound the message attention mechanism, essentially providing the same input and expecting multiple different outputs.

The average-label approach to predicting the personality profile that consumes an article demonstrates encouraging predictive performance, accompanied by a reduction in variance between folds. These results instill confidence in the method's capability to infer the overall personality that a consumed article may elicit. Although our results may not match the outcomes achieved in other personality prediction tasks such as generated text, they serve as a solid foundation for further advancement.

Some consideration should be given to the difference between the tasks. Message attention models are modelling a user's personality given all of the text they have consumed. The average-label frame models the average personality of a single article. Are these similar enough tasks for a fair comparison of performance? Is it viable to use average label models as part of a model that would predict users?

6.2 Do we need more samples?

The different nature of the proposed models also leads to a large difference in the number of samples. We gathered 213 participants who consumed news articles, whereas we gathered 953 participants who read fiction. Contemporary work involving produced text generally has samples in the tens of thousands (Lynn et al., 2020; Eichstaedt et al., 2021).

To see if number of samples was the cause for the large gap in performance, we created a model that would predict personality from the produced text our participants posted on Reddit. We use the same message attention model as is used in previous work (Lynn et al., 2020). Our results (given in Appendix B) show performance much improved relative to that of our consumed text models, and with confidence intervals within range of state-of-the-art performance. This demonstrates that the number of samples alone does not explain the large decrease in performance between the produced and consumed text tasks.

6.3 Textual features

Average-label models also may be over-fitting to textual features that are repeated multiple times across each corpus. This may be particularly true with both article corpora. Efforts have been made to clean the text for repeating signals of this form (e.g., the author bylines for news articles) but we cannot be certain of removing all such indicators from our crawled article content. To mitigage this effect, we have trained models using L2 regularization. L2 regularization imposes a larger cost on the loss function for larger weights, thus decreasing the impact of over-fitting. Appendix D shows that L2 regularization on average-label models reduces performance, but these regularised models still outperform message-level attention models.

Our visualisations presented in Section 5.3 show how words and phrases correlate with personality dimensions. Content is picked up rather than writing style when looking at the word clouds, suggesting that consumption of particular topics may be more indicative of personality than the style in which the content is presented. These results may be seen as consistent with similar works involving user generated content and personality. Facebook likes of topics and media content have also been found to be congruent with personality (Youyou et al., 2015). Our visualisations of the news dataset also show that the model is correlating predictions with certain news topics dominant at the time of data collection. This may be an artefact of the small time period of data collection from users; while all articles that participants have consumed have been crawled, their activity is more likely to contain recent content.

N-grams that appear to be predictive of a high neuroticism score (such as 'gun control') have an inverse correlation with the other four dimensions. This is consistent with theoretical and other quantitative research into the general factors of personality, and the broader interrelation between those four dimensions when contrasted with neuroticism (Van der Linden et al., 2010; Musek, 2007).

6.4 Further pointers

News content in general appears to out-perform fictional content when used as a predictor of personality. Three personality dimensions appear to be less predictable from fictional content than from news, while extraversion remains predictable with good performance across all three datasets. Openness is

however easier to predict when using fictional content as an input. Behavioural research may reveal if these patterns exist outside of these models.

To model consumption from observable posting behaviour, we assumed that if a participant had commented on a Reddit thread, that participant had read the article which began the thread. We cannot say with certainty that this is true, and especially cannot be confident that a user has read any specific part of an article, as commenting without reading is an unfortunately common behaviour on many social media platforms. To explore this, we compared predictions of personalities using news article text and just the title of the article (which is the first thing a user will see on entering a thread), finding that predictions using the titles alone were often as good as (and for some personality dimensions, better than) using the full article text. We tentatively conclude that when making predictions on the basis of text consumption, some scepticism may be warranted as to whether a user has fully consumed a given text.

7 Conclusions

In this paper we have shown that personality can be inferred based on the text that a user has consumed. To our knowledge, this is the first work using consumed textual content to model personality that reaches comparable performances to produced content. The performance achieved by average-label modelling can be seen as a baseline for personality modelling using consumed text.

Message attention models do not achieve acceptable levels of performance when applied to the domain of consumed text. We show that this may be due to the weighting function giving no especial weight to any consumed text, in combination with the lack of unique textual content for each user, which gives confounding feedback to the model during training.

We used three different datasets to train and evaluate our models: pieces of fiction, news articles, and news titles. Personality is shown to be more reliably inferred from news content than fiction content. Models trained upon news article titles, with less textual content, achieved similar performance to models trained upon whole news articles, which may reveal that a condensed set of features are most important for modelling personality.

Future work in this field should involve further investigation as to how message attention models may be adapted to this context, as well as establishing resources to enable new approaches to this problem in the form of a shared task. Due to participant privacy concerns, our datasets cannot be released, which forms a hurdle to reproduction and development. A publicly available dataset would be beneficial, so new work can be evaluated on a standardised dataset. An ideal dataset would also provide access to more training samples, along with greater assurance that the textual content has been consumed by the users.

Acknowledgements

This research was supported by funding from REPHRAIN: National Research Centre on Privacy, Harm Reduction and Adversarial Influence online (UKRI grant: `EP/V011189/1`). The authors were also supported by a large grant from the Volkswagen Foundation ("Reclaiming individual autonomy and democratic discourse online"). S.L. was also supported by funding from the Humboldt Foundation in Germany and by an ERC Advanced Grant (PRODEMINFO).

References

Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. 2012. Personality and patterns of Facebook usage. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 24–32.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Johannes C Eichstaedt, Margaret L Kern, David B Yaden, HA Schwartz, Salvatore Giorgi, Gregory Park, Courtney A Hagan, Victoria A Tobolsky, Laura K Smith, Anneke Buffone, et al. 2021. Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, 26(4):398.

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

Matthew H Goldberg, Abel Gustafson, Seth A Rosenthal, and Anthony Leiserowitz. 2021. Shifting republican views on climate change through targeted advertising. *Nature Climate Change*, 11(7):573–577.

Keven Joyal-Desmarais, Alexandra K Scharmer, Molly K Madzelan, Jolene V See, Alexander J Rothman, and Mark Snyder. 2022. Appealing to motivation to change attitudes, intentions, and behavior: A systematic review and meta-analysis of 702 experimental tests of the effects of motivational message matching on persuasion. *Psychological Bulletin*, 148(7-8):465.

Veronica Lynn, Niranjan Balasubramanian, and H Andrew Schwartz. 2020. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5306–5316.

S C Matz, M Kosinski, G Nave, and D J Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48):12714–12719.

Janek Musek. 2007. A general factor of personality: Evidence for the Big One in the five-factor model. *Journal of Research in Personality*, 41(6):1213–1233.

Michael Nowak and Dean Eckles. 2014. Determining user personality characteristics from social networking system communications and characteristics. US Patent 8,825,764.

Barun Patra, Saksham Singhal, Shaohan Huang, Zewen Chi, Li Dong, Furu Wei, Vishrav Chaudhary, and Xia Song. 2022. Beyond English-centric bitexts for better multilingual language representation learning. *arXiv preprint arXiv:2210.14867*.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

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.

Christopher J Soto and Oliver P John. 2017. The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1):117–143.

Ben M Tappin, Chloe Wittenberg, Luke Hewitt, David Rand, et al. 2022. Quantifying the persuasive returns to political microtargeting. *(Working Paper)*.

Dimitri Van der Linden, Jan te Nijenhuis, and Arnold B Bakker. 2010. The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, 44(3):315–327.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

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

Brahim Zarouali, Tom Dobber, Guy De Pauw, and Claes de Vreese. 2020. Using a personality-profiling algorithm to investigate political microtargeting: Assessing the persuasion effects of personality-tailored ads on social media. *Communication Research*, 49(8):1066–1091.

Qihuang Zhong, Liang Ding, Yibing Zhan, Yu Qiao, Yonggang Wen, Li Shen, Juhua Liu, Baosheng Yu, Bo Du, Yixin Chen, et al. 2022. Toward efficient language model pretraining and downstream adaptation via self-evolution: A case study on SuperGLUE. *arXiv preprint arXiv:2212.01853*.

Appendix

A Experimental Setup

All models were implemented using PyTorch and PyTorch Lightning. One model was trained for each personality dimension. Language modelling was performed using the 'sentence-transformers/all-distilroberta-v1' Sentence-BERT model, which provides a 768-dimensional representation for each piece of text. The learning rate for each model was selected using PyTorch Lightning. Training lasted for 8 epochs, although most stopped after 4 epochs, due to early stopping. Samples were uniformly sampled and split into 5 folds.

A.1 Users

The batch size was set to 1, due to hardware constraints. The maximum number of articles per user was also limited to 100. 512 input tokens were taken from each article. If an article was shorter than 512 tokens, it would be padded. If the article was longer, it would be truncated to the first 512 tokens. All articles were passed to the language model, and messages were split between 5 Titan-X GPUs for parallel computation.

After generating the embeddings, mean pooling was applied to the token embeddings to obtain sentence vectors. These sentence vectors were then processed through a multi-headed attention mechanism, followed by dot product attention on the outputs. This produced scalar values corresponding to each message. These scalars were used as weights for the weighted sum of the token embeddings. Finally, two feed forward layers used the user vector to generate the personality prediction.

A.2 Articles

The batch size for articles was set to 16. Since the model is relatively small, only a single GPU was required for processing. Each article was represented with 512 tokens and undergoes the same truncation or padding process as the users' articles. Once the token embeddings were computed, sentence vectors were generated using mean pooling. These sentence vectors were then passed through two feed-forward layers to generate a personality prediction. When training models with L2 regularization (as specified in Appendix D) the weight decay parameter (λ) was set to 0.001.

A.3 Titles

The batch size for titles was also set to 16 to ensure comparable training with models that use articles as input text. However, each title was limited to 128 tokens in length. The titles are padded or truncated as necessary during pre-processing. As with articles, when doing L2 regularization the weight decay was set to 0.001.

B Produced Text Models

Models trained from produced text have the same model as titles, where the length of the text is limited to 128 tokens due to comments being shorter. Approximately 10,000 comments were used in training these models.

Table 4: 5-fold cross-validated prediction performance when using message attention to predict users' personality scores from the text they have produced. The dataset used here is sampled from the same 1,116 participants used in our consumed text models, but with predictions made using text they produced via their comments. We report the average performance across each fold as well as 95% confidence intervals.

Personality Dimension	Pearson's *r* [95% CI]
Extraversion	0.32 [0.22, 0.43]
Agreeableness	0.31 [0.20, 0.42]
Conscientiousness	0.33 [0.27, 0.38]
Neuroticism	0.33 [0.21, 0.45]
Openness to Experience	0.32 [0.20, 0.44]

Table 4 shows the 5-fold cross-validation performance of a message-level attention model, using our participants' produced text to predict their personality inventories rather than the text they have consumed. The model used in this experiment is the same as the model described in Section 3.1, which achieved underwhelming performance when using consumed text.

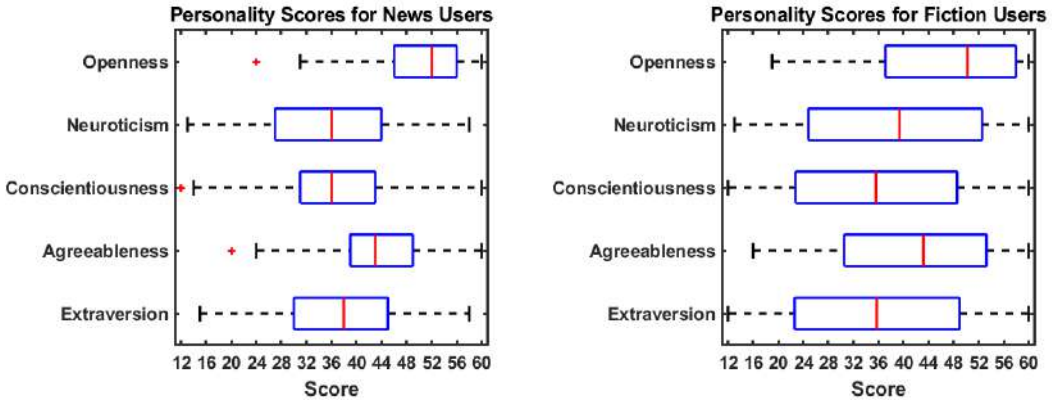
This level of performance more closely resembles state-of-the-art metrics that have been documented when using produced text for personality prediction, with state of the art performance within the confidence intervals for each dimension (Eichstaedt et al., 2021).

This shows that message-level attention models can perform well with a lower number of samples when using produced text, however consumed text may not be an ideal medium for this model architecture.

Table 5: 5-fold cross-validation performance of average-label models attempting to predict personality inventories when using L2 regularization. These experiments were intended to probe whether over-fitting is evident in our average-label models. Bold indicates which models performed the best for each dimension.

Domain	Fiction	News	
Content	Articles	Articles	Titles
Extraversion	0.17 [0.15, 0.2]	0.23 [0.21, 0.25]	**0.28 [0.27, 0.29]**
Agreeableness	0.1 [0.05, 0.14]	0.17 [0.15, 0.18]	**0.21 [0.19, 0.22]**
Conscientiousness	0.11 [0.08, 0.14]	**0.2 [0.19, 0.21]**	**0.2 [0.19, 0.22]**
Neuroticism	0.08 [0.06, 0.1]	0.29 [0.28, 0.31]	**0.31 [0.3, 0.32]**
Openness to Experience	0.09 [0.08, 0.1]	**0.12 [0.12, 0.13]**	**0.12 [0.11, 0.13]**

Figure 4: Box plots showing the distributions of personality scores per user. On each box, the central mark indicates the median, and the left and right edges of the box indicate the 25th and 75th percentiles, respectively. Outliers are plotted beyond the whiskers.



C User Personality Distributions

Figure 4 shows the distribution of the Big-5 personality traits for fiction and news consumers amongst our participants. Figure 1 from the main body displays scores as aggregated on a per-article basis.

D Regularized Models

Table 5 presents the 5-fold cross-validation performance of average-label models when using L2 regularization. The decreases in performance may be explained by over-fitting in the original models without L2 regularization.

The models using the news title dataset are generally now the better-performing models and also see the lowest performance impact from regularisation. This may indicate that our news article representations contain noisy features as a byproduct of crawling, and models without regularization over-fit to those features.