Personality-based Deep Learning for Hate Speech Detection

Anonymous ACL submission

Abstract

An essential factor in the fight against hate speech is the advancement of effective computational algorithms for automatically detecting it. Earlier research has put forth a range of computational methods aimed at automating hate speech detection. However, these approaches have predominantly overlooked significant insights from the psychology literature, which delves into the connection between personality traits and hate. To this end, we propose a novel framework for detecting hate speech focusing on people's personality factors reflected in their writing. Our framework has two components: (i) a knowledge distillation model for fully automating the process of personality inference from text and (ii) a personality-based deep learning model for hate speech detection. Our approach is unique in that it (i) incorporates low-level personality factors, which have been largely neglected in prior literature, into automated hate speech detection and (ii) proposes multi-head-self-attention-inspired deep learning components for fully exploiting the intricate relationship between personality and hate. In particular, the latter aids the model in untangling intermediate personality factors, the potential existence of which has been suggested by recent research in psychology. We evaluate our model with two real-world datasets. The results show that our model significantly outperforms state-of-the-art baselines. From an academic viewpoint, our study paves the way for future research by incorporating personality aspects into the design of automated hate speech detection. From a business standpoint, our model offers substantial assistance to online social platforms and governmental bodies facing challenges in effectively moderating hate speech.

1 Introduction

The motivations for businesses to control hate speech are manifold. First, the prevalence of hate speech can have a detrimental effect on trust among online community members, which may lead to their defection from the community (Nasi et al. 2015). In addition, the pervasiveness of hate speech makes social media sites unattractive for advertisers since they tend not to risk advertising on a site known for hate speech (Fortuna and Nunes 2018). Finally, online platforms face public pressure to deal with hate speech. Well-known social media services such as Facebook and Twitter have been constantly encountering criticism for being passive on the matter (Davidson et al. 2017).

Accordingly, many business platforms (e.g., social media sites, news publishers, and search engines) present strong incentives to create mechanisms for regulating hate speech. Among different approaches (e.g., human content moderation, counter speech, education, etc.), controlling hate speech using automated systems is especially important considering an extremely large amount of business and social costs incurred when solely relying on manual content moderation by humans. However, efforts to build automated methods have not been successful yet in both industry and academia (Mathew et al. 2019). One of the main reasons for this is, while the study on hate speech detection is an interdisciplinary field that encompasses business, psychology, linguistics, etc., most of its methodological focus has been put on the computational perspective. That is, current research on automated hate speech detection does not incorporate theories or empirical evidence from social science.

Specifically, in social science, there is a vast literature on hate speech from various perspectives (e.g., historical, cultural, sociological, economic, and political) (Green et al. 2001). One perspective closely related to the automated detection of hate speech but largely neglected in the previous literature is personality and hate.

In the literature of automated speech detection, there are a few studies which have used high-level personality factors (e.g., BIG 5 personality factors) as machine learning features (Elzayady et al. 2023; Lee and Ram 2020), their approaches may not lead to a desired outcome as there have been conflicting results reported in psychology regarding the relationship between these features and hate. For instance, previous studies have reported mixed results on the relationship between extraversion, one of the five most referred personality factors (i.e., BIG5 personality factors), and hate behaviors. While Galo and Smith (1998) found that extraversion was positively related to higher levels of physical and verbal aggression, anger, and hostility, ElSherief et al. (2018) observed that individuals who engage in hate speech on Twitter often exhibited lower levels of extraversion. We argue that such conflicting effects of high-level personality traits on hate make it challenging for deep learning algorithms to uncover the complex relationship between personality and hate.

To this end, in this study, we propose a deep learning approach, grounded in the recent literature of personality, which focuses on low-level personality factors (i.e., 30 personality subfactors of the BIG 5 personalities), and possible intermediate personality factors bridging the lower-level and the higher-level personality traits (Depue and Collins 1999; DeYoung et al. 2007). To achieve this, we utilize the architecture of multi-head self-attention layers and apply it within the context of hate speech detection, (Vaswani et al., 2017). We test our method with multiple real-world datasets and show that it outperforms extant cutting-edge approaches including proprietary methods developed by Google. In the following section, we review the relevant literature and discuss extant research gaps which serve as the foundation for our proposed model.

2 Related Work

2.1 Automated Hate Speech Detection

The field of research on hate speech is quite recent from the computational perspective (Fortun and Nunes 2018). In terms of methods, we identify two major streams of research: rule-based approaches and machine-learning approaches. The rule-based approach, in general, determines whether specific text contains hatred by referring to a hate lexicon (or dictionary) (Gitari et al. 2015). While this approach is straightforward to understand and can be easily applied in a practical setting, it has several drawbacks including the quality of classification highly depending on that of the lexicon; the process of building a quality lexicon being often cumbersome; and a lexicon developed for one setting not being able to generalize to other contexts (Zhang et al. 2015; Nobata et al. 2016).

On the other hand, the machine-learning approach resolves many of the issues exhibited by the rule-based approach. First, machine-learning models mostly follow the open-vocabulary approach representing text using word frequencies (e.g., TF-IDF), the distributional similarities between words (e.g., n-gram, char2vec, and word2vec), or the embeddings learned by large language models (Devlin et al. 2019; Eichstaedt et al. 2021). Thus, the tedious process of manually creating a quality lexicon can be avoided. In addition, such text representation is not peculiar to a specific context and, thus, can be used to identify hate speech in a more generalized setting.

For these reasons, many studies on automated hate speech detection have adopted the machine-learning approach. In machine learning, the types of features fed into the model are closely related to its performance. Broadly speaking, three types of features have been used in the prior literature on hate speech detection: general linguistics features; topic-specific linguistics features; and metadata.

First, general linguistics features include different variations of text quantification methods such as tf-idf, word2vec, etc., which are derived from vector space models and neural-network-based language models (Lee et al. 1997). For instance, Elzayady et al. (2023) have used the combination of tf-idf and word2vec as feature inputs for a deep learning model that includes both CNN and RNN components. Lee and Ram (2020) have proposed to use word2vec along with other text-based features to train a LSTM model for hate speech detection.

Second, topic-specific linguistic features include multiple sets of words manually labeled for identifying hate speech on specific topics (e.g., sexism, racism, and homophobia). For instance, Kwok and Wang (2013) classified racist tweets by developing a lexicon for racism with Naïve Bayes. Warner and Hirschberg (2012) suggested that hate text with different topics can be distinguished by considering high frequency stereotypical words related to each topic and tested their idea with SVM.

Lastly, metadata for hate speech detection include demographic and geographic characteristics of people (Waseem and Hovy, 2016). Also, text metadata have been used such as the number of words and the inclusion of special characters (Davidson et al. 2017). However, metadata, especially those related to people, have not been used much in literature since such data are in general hard to collect and subject to privacy concerns (Harell 2010).

2.2 Personality and Hate

In previous studies on personality-factor theories, personalities were considered as an important predictor for one's hate behavior. Specifically, Big 5 personality model (BIG5), consisting of 5 high-level personality factors – agreeableness, conscientiousness, extraversion, neuroticism, and openness – have been widely investigated. Each of these personality domains consists of 6 low-level personality factors, which capture distinct, specific characteristics of each personality domain (Paunonen and Ashton 2001).

Previous studies in psychology have extensively associated high-level personality factors with hate behavior. First, agreeableness, in literature, has been reported as one of the strongest predictors for hate behavior. This personality factor is related to one's tendency to pursue social harmony. In general, previous studies have found a negative association between agreeableness and hate (Heaven 1996; Barlett and Anderson 2012). Second, researchers have reported that conscientiousness, a personality factor closely connected to how a person controls oneself, is negatively associated with hate behavior (Jovanović et al. 2011; Jones et al. 2011). Third, extraversion is a personality factor related to being ambitious and sociable. Previous studies have reported mixed results on the relationship between extraversion and hate behavior (Galo and Smith 1998; ElSherief et al. 2018; Burton et al. 2007). Fourth, neuroticism is a personality factor related to emotional instability. In literature, a majority of studies have highlighted a positively association between neuroticism and hate behavior (Egan and Lewis 2011; Jovanović et al. 2011; Becerra-Garcia et al. 2013). Lastly, openness is related to one's willingness to experience new things. Previous studies have found either mixed results or no evidence supporting the association between openness and hate (Egan and Lewis 2011; Hosie et al. 2014; Barlett and Anderson 2012).

2.3 Research Gaps

One of the research gaps that motivated our study is that previous studies on automated hate speech detection have neglected the importance of the association between personality and hate despite the plethora of evidence in psychology literature. Even in a few exceptions, only the role of high-level personality factors has been highlighted (e.g., Elzayady et al. 2023; Lee and Ram 2020). This may not yield consistent benefits for hate speech detection, as previous studies have presented conflicting findings concerning the association between these factors and hate speech.

Some recent studies in psychology have suggested the possible existence of intermediate personality traits bridging the lower-level and the higher-level personality factors, which have not been a focus of the previous hate speech detection studies based on personality. DeYoung et al. (2007) have conducted factor analyses and found two distinct intermediate personality traits that lie between each personality domain and its related facets. Also, Depue and Collins (1999) have suggested the possibility of intermediate personality factors that connect personality facets from two or more distinct personality domains. We expect that exploring these lower-level and intermediate personality factors has the potential to improve the performance of a deep learning model.

3 Method

The task to be solved in this study is to identify whether given text belongs to one of the following categories: hate or non-hate speech. In other words, our objective is to find an optimal function, $f: t_i \rightarrow \{0,1\}$, where t_i is an element (i.e., potential hate speech) in our dataset $T = \{t_1, t_2, \ldots, t_n\}$. Each element in the range of the function f, respectively, implies the following classification categories: 0 = non-hate speech and 1 = hate speech.

To solve the above problem, we propose a novel framework illustrated in Appendix A. Largely, it consists of two components: an automated personality inference method based on knowledge distillation and a deep learning model, which extends the multi-head self-attention mechanism, for detecting hate speech based on personality traits (Hinton et al. 2015; Vaswani et al. 2017). In the following, we elaborate on the details of these two components.

3.1 Personality Inference Method

As the first component of our hate speech detection framework, we develop a computational method to infer personality from text. Our approach for personality inference is inspired by the knowledge distillation and its details are describe below (Hinton et al., 2015).

Previously, several studies have proposed automated methods for personality inference using textual cues such as syntactic and lexical features (Mairesse et al., 2007; Pennebaker and King, 1999). Among them, a model developed by IBM, aka Personality Insights (IBMPI) (IBM Cloud, 2020) is cutting-edge in terms of its performance (i.e., accuracy) and comprehensiveness in the types of personality traits covered (i.e., both high- and low-level personality factors).

However, IBMPI is a proprietary product which does not reveal the details of its methodology and has been discontinued recently. Thus, we develop our own personality inference method leveraging knowledge distillation. Knowledge distillation is an approach for transferring knowledge learned by a complex deep learning (aka a distilling model) model to a simple model (aka a distilled model) (Hinton et al., 2015). Typically, knowledge distillation is implemented to minimize the loss between the softmax outputs of knowledge-distilled and knowledge-distilling models (Gou et al., 2021). Specifically, its objective is to minimize the cross entropy, $-\sum \hat{y}_{x_i \setminus t} \cdot \log y_{x_i \setminus t}$, where $\hat{y}_{x_i \setminus t}$ and $\log y_{x_i \setminus t}$ are outcomes of distilled and distilling models softened by a parameter t (aka temperature) and x_i is input of a data point i (Hinton et al., 2015). Note that by increasing the temperature value, a distilled model can present greater generalizability. On the other hand, by lowering the temperature value, a distilled model will more closely mimic the behavior of a distilling model.

As depicted in Appendix A, we use the output of IBMPI (i.e., distilling model) scores on 35 high-level and low-level personality factors as the source of knowledge distillation. These scores are used to optimize the parameters of our personality inference model using a pre-trained language model (i.e., distilled model) (Vaswani et al., 2017). Specifically, input text is processed with a tokenizer and fed into a language model. Then it produces, in the last hidden layer, a set of embedding vectors that capture different aspects of text. Among the embedding vectors, we use one for classification (i.e., CLS embedding) as input for the fine-tuning layer. The fine-tuning layer is a single dense layer that consists of 100 neurons with the tanh activation. The outputs of the fine-tuning layer are connected to the prediction layer comprising 35 neurons with the sigmoid activation. Each of the 35 neurons represents one of the 35 personality traits. Following the suggestion provided by Hinton et al. (2015), we set the temperature (i.e., t) as 2.5. Note that, since we do not know the detailed architecture of IBMPI, we softened, with the temperature, the outcomes of IBMPI assuming they were activated by the sigmoid function. For instance, if a personality score for input x is 0.2, based on the assumption that it is activated by $1/(1+e^{-x})$, we derive its softened score, 0.36, by converting the activation function into $1/(1+e^{-x/2.5})$ (temperature = 2.5).

3.2 Hate Detection Method

Based on the personality scores inferred using our approach detailed in Section 3.2., we develop an automated method for detecting hate speech. As illustrated in Appendix A, Our approach consists of multiple subunits: those for semantic encoding, individual personality factors, local intermediate personality traits, and global intermediate personality traits. Our methodological contribution lies on (i) incorporating both high-level and low-level personality factors in automated hate speech detection and (ii) proposing multi-head-self-attention-inspired deep learning components that capture intermediate personality factors. We elaborate the details below.

Subunit 1: Semantic encoding

While the focus of this study is to examine the value of personality factors in hate speech detection, text semantics also play a vital role in achieving high performance. Thus, following recent developments in the domain of automate hate speech detection, we apply a pre-trained language model to infer latent semantics of text at the document level and feed the information into a fine-tuning layer to capture important semantics of text regarding hate speech (Alatawi et al., 2021). We used the CLS embedding of the pre-trained language model as the summary of text semantics, and it was further processed by a single fine-tuning layer with 768 neurons and the tanh activation. The output of the fine-tuning layer is then concatenate with those of the other subunits described below (Devlin et al., 2019).

Subunit 2: Raw personality scores

As discussed above, previous literature in psychology has paid close attention to the connection between personality and hate. To capture this connection and enhance the performance of hate speech detection, we feed 5 high-level and 30 low-level personality scores derived from our personality inference method into the concatenated layer of the personality detection method (refer to Appendix A).

Subunit 3 and 4: Local and global intermediate personality traits

One of the research gaps that we identified from the psychology literature was a lack of focus on the connection between intermediate personality traits and hate behavior (DeYoung et al., 2007; Goldberg, 1999). Nevertheless, when it comes to addressing hate speech from a personality perspective, it is worth noting that there has not been a single study that has specifically focused on the intermediate personality factors. To fill this research gap, we developed deep learning components, inspired by multi-head self-attention layers of the transformer, for inferring the intermediate personality traits (Vaswani et al., 2017). Specifically, we introduce two subunits that help identify the intermediate personality factors from local and global perspective.

First, as depicted in Appendix B-(a), the local intermediate personality traits (LIPs) capture the intricate interactions among low-level personality factors within each high-level personality factors they belong to (e.g., the interactions among altruism, cooperation, modesty, morality, sympathy, and trust, which are the low-level factors of the same high-level factor, agreeableness). In Appendix B-(b), we provide a detailed illustration of how the multi-head self-attention mechanism is applied to infer the LIPs (Vaswani et al., 2017). Specifically, the set of all 35 personality traits, P, consists of the following 6 disjoints subsets: (i) $P_H = \{p_i | 1 \leq i \leq 5\}$, a set of the 5 high-level personality factors (i.e., agreeableness, conscientiousness, extraversion, neuroticism, and openness); (ii) $P_A = \{p_i | 6 \leq i \leq 11\}$, the 6 low-level personality factors which belong to agreeableness; (iii) $P_C = \{p_i | 12 \leq i \leq 17\}$, the 6 low-level personality factors which belong to conscientiousness; (iv) $P_E = \{p_i | 18 \leq i \leq 23\}$, the 6 low-level personality factors which belong to extraversion; (v) $P_N = \{p_i | 24 \leq i \leq 29\}$, the 6 low-level personality factors which belong to neuroticism; and (vi) $P_O = \{p_i | 30 \leq i \leq 35\}$, the 6 low-level personality factors which belong to openness. For P_A, P_C, P_E, P_N, and P_O, we applied two-head self-attention layers to capture two LIPs, in total, getting 10 LIPs. (i.e., LIP 1 -10 in Appendix B-(b)). We set the number of attention heads based on the findings of DeYoung et al. (2007) who showed that there are two distinct aspects within each of the five high-level personality factors. However, the number of attention heads can be hyper-parameterized, since determination of the optimal number of intermediated personality factors needs further investigation (Jang et al., 2002).

In addition to LIPs, we also develop a deep learning component for globally identifying intermediate personality traits (i.e., GIPs). As opposed to LIPs, GIPs capture the complex relationships among low-level personality factors that belong to different high-level personality factors (e.g., the relationship between a low-level personality factor of agreeableness and that of neuroticism). The decision to incorporate the GIP component was driven by insights from the personality literature, which suggested the potential for cross-domain interactions. In other words, low-level personality factors that are part of different high-level personality factors can combine to form compound intermediate personality traits (Depue and Collins, 1999). The architecture of the GIP component is similar to that of the LIP component except that all low-level personality factors are jointly fed into a multi-head self-attention layer (refer to Appendix B-[c]). Specifically, as represented in Appendix B-(d), the low-level personality factors, from p_6 to p_{35}, are processed by a ten-head self-attention layer to produce 10 distinct global intermediate personality traits (i.e., GIP 1 – 10).

As a result, we produce 20 scores (i.e., LIP 1 – 10 and GIP 1 – 10) from the subunit 3 and 4. We concatenate these outputs with those from the subunit 1 and 2 and feed them into the final layer of our personality detection method. The final layer generates the probability of given text using all the input sources and makes classification whether it is hate-speech or not.

4 Experiment

4.1 Data and Evaluation Framework

Prior to the development of our hate speech detection method, we first trained and tested the personality inference model. To train the model, we used data collected from Wikipedia (henceforth, we call this data *WikiHate*) (Conversation AI, 2018). WikiHate consists of comments collected from the Wikipedia's talk pages. Our personality inference model was trained and tested on a subset of WikiHate whose personality scores were calculated by IBMPI (i.e., the data within the dashed box). Specifically, among 64,888 comments (the number of hate comments: 16,222; the number of non-hate comments: 48,666) in WikiHate, we randomly

(a) WikiHate		(b) PersEssay**					
	MAE		Agreeable-ness	Conscientious-ness	Extraversion	Neuroticism	Openness
BERT	0.0060	Our Model	60.35	60.19	56.78	57.14	64.33
RoBERTa	0.0056	Mairesse et al. (2007)	55.35	55.28	55.13	58.09	59.57
ELECTRA	0.0133	Majumder et al. (2017)	56.71	56.71	58.09	57.33	61.13

*The shaded cells indicate the best results.
** Accuracy for each personality domain is reported in percentage.

Table 1. The summary of our persoanlity inference method

sampled 20,000 comments and calculated their personality scores using IBMPI. Then, using these comments and personality scores as a dataset for the knowledge distillation process described above, we developed and evaluated our personality inference model. We applied five-fold cross validation for training and testing processes. Then, our personality inference method was applied to the rest of the data to infer personality scores.

As an additional data source for evaluation, we employed one collected from Stormfront.org, a white supremacist web forum (henceforth, we call this data *SupremacistHate*; De Gilbert et al., 2018). A majority of users on Stormfront are white nationalists who can be characterized by their pseudo rationalism (Meddaugh and Kay, 2009). SupremacistHate contains 10,944 sentences classified as either hate or non-hate comments. Among the total, 10.9% (i.e., 1,196) and 89.1% (i.e., 9,748) were hate and non-hate comments, respectively.

4.2 The Evaluation of Our Personality Inference Method

For the development of our personality inference method, we use the following models and compare their performance: BERT, RoBERTa, and ELECTRA (Devlin et al., 2019; Liu et al., 2019; Clark et al., 2020). We used mean absolute error (MAE) to evaluate their performance, which measures the absolute difference between the personality scores inferred using IBMPI and our method. As an additional validation process, we applied our method to another dataset called PersEssay. This dataset includes essays written by students and ground-truth binary labels for high-level personality factors. Since our model produces continuous scores, we convert them into a binary format using the threshold of 0.5. That is, if a score exceeds 0.5, it is considered as 1, and 0 otherwise. As baselines, we selected automated personality detection models from the previous literature, which are built upon the same dataset (i.e., Mairesse et al., 2007; Majumder et al., 2017). As metrics for evaluation, we report classification accuracy of each of the high-level personality factors.

The results of our personality inference method are summarized in Table 1. First, in Table 1-(a), we report the performance of our personality inference method developed using WikiHate. Overall, all three models that we tested (i.e., BERT-based, RoBERTa-based, and ELECTRA-based) produced strong results. The MAE of the BERT-based model was only 0.006, which implies that the average difference between the sum of personality-domain and -trait scores predicted by our model and IBMPI is 0.006. We observe similar results for RoBERTa-based and ELECTRA-based models, whose MAEs are 0.0056 and 0.0133, respectively.

We further examined the performance of our method using PersEssay. As mentioned above, the problem here is to classify given documents into BIG5 personality domains (i.e., a multi-class multi-label classification problem). In Table 1-(b), we report the results of our method for each personality domain compared with those of the baselines mentioned above. Note that, among the three models that we build upon WikiHate, we used the best performing one (i.e., RoBERTa) as our model for the PersEssay classification task. The results show that our model outperforms the baselines in agreeableness, conscientiousness, and openness (by 3.64%[1], 3.48%, and 3.20% in accuracy when compared to the best results of the baselines) and produces comparable results in extraversion and neuroticism.

4.3 The Evaluation of Our Hate Detection Method

Ss Based on the same set of models used to develop our personality inference method (i.e., BERT, RoBERTa, and ELECTRA), we applied

[1] Note that we use absolute percentage points for reporting performance comparison.

the wrapper method to measure the impact of each component of our methodological framework (i.e., semantic encoding of text, raw personality scores, LIPs and GIPs) (Dash and Liu, 1997). For instance, using BERT, we developed and compared the following models: BERT (a BERT model with semantic encoding), BERT + RAW (a BERT model with semantic encoding and raw personality scores), BERT + LIP (a BERT model with semantic encoding and LIPs), BERT + GIP (a BERT model with semantic encoding and GIPs), BERT + RAW + LIP, BERT + RAW + GIP, BERT + LIP + GIP, and BERT + RAW + LIP + GIP. Therefore, for each type of the transformer-based methods (i.e., BERT, RoBERTa, and ELECTRA), we develop eight variations to evaluate our design. In addition to these models, we added an additional, cutting-edge model for performance comparison, Google Perspective, which is a commercial tool developed by Google for identifying the level of hate in text (Google, 2019). For evaluation, we used the following metrics: precision, recall, F-measure, accuracy, and area under the precision-recall curve.

The performance of our hate detection method is summarized in Table 2. First, Table 2-(a) reports the results on WikiHate. The best performing model in terms of F1-score was RoBERTa trained with all personality features (i.e., RAW, LIP, and GIP). Compared to the vanilla RoBERTa model, it improved F1-score by 7.44%, respectively. On the other hand, ELECTRA trained with all personality features outperformed the rest of the models in recall and AUCPR. Comparing to vanilla ELECTRA, the two metrics were improved by 1.55% and 3.51%, respectively. Additionally, when compared to Google Perspective, our method produced better outcomes. Specifically, RoBERTa+RAW+LIP+GIP, one of our best models, improved the F1-score of Google Perspective by 2.10%.

A detailed exploration of the results provides us with some interesting findings. First, the LIP element in PERSONA resulted in the largest degree of improvement in detecting hate speech. For all the three baselines (i.e., BERT, RoBERTa, and ELECTRA), when a single personality component (i.e., RAW, LIP, or GIP) was added to their vanilla models, the LIP component contributed to the strongest performance boost (improvement in F1-score by 3.60% - 7.67%). Particularly, RoBERTa+LIP and ELETRA+LIP produced F1-scores that are comparable to RoBERTa+RAW+LIP+GIP and ELECTRA+RAW+LIP+GIP, respectively. Among the other two personality components (i.e., RAW and GIP), the GIP component was more effective in identifying hate speech than the RAW component. GIP increased the F1-scores of vanilla models by 3.17% - 4.49% while RAW improved them by 1.78% - 4.23%. We observed the similar trend in model performance when two

		(a) WikiHate				(b) SupremacistHate			
		Precision	Recall	F1	AUCPR	Precision	Recall	F1	AUCPR
BERT	Vanilla	63.34	91.00	74.69	88.56	61.89	47.78	53.93	59.34
	RAW	67.03	89.00	76.47	88.81	57.35	50.63	53.78	56.14
	LIP	71.47	86.56	78.29	88.65	68.75	45.25	54.58	59.54
	GIP	69.88	87.89	77.86	89.10	63.27	49.05	55.26	61.70
	RAW+LIP	70.00	89.44	78.53	89.32	65.95	48.42	55.84	61.00
	RAW+GIP	71.51	87.00	78.50	89.73	65.22	47.47	54.95	60.55
	LIP+GIP	72.91	86.44	79.10	89.18	69.91	50.00	58.30	63.30
	RAW+LIP+GIP	76.48	84.56	80.32	89.41	65.85	51.27	57.65	63.25
RoBERTa	Vanilla	68.55	90.11	75.16	90.31	53.35	52.85	53.10	55.90
	RAW	70.82	90.33	79.39	91.23	60.54	50.00	54.77	59.20
	LIP	79.56	85.22	82.29	91.39	62.02	56.33	59.04	64.44
	GIP	71.54	89.67	79.59	90.82	66.27	53.48	59.19	66.53
	RAW+LIP	73.28	87.78	79.88	90.58	66.80	52.85	59.01	63.95
	RAW+GIP	71.17	90.78	79.79	90.95	63.67	56.01	59.59	63.81
	LIP+GIP	74.15	89.56	81.13	91.05	64.95	59.81	62.27	67.63
	RAW+LIP+GIP	78.81	86.78	82.60	91.40	60.94	61.71	61.32	63.59
ELECTRA	Vanilla	63.22	89.56	74.12	88.26	68.05	36.39	47.42	53.22
	RAW	67.83	90.89	77.68	90.49	64.80	40.19	49.61	54.33
	LIP	77.09	87.11	81.79	91.18	64.68	51.58	57.39	62.87
	GIP	72.40	90.67	78.61	91.29	61.87	54.43	57.91	60.21
	RAW+LIP	75.17	84.78	79.69	88.14	61.82	43.04	50.75	58.96
	RAW+GIP	74.20	85.00	79.23	88.79	62.83	44.94	52.40	59.64
	LIP+GIP	69.00	87.56	77.18	87.94	63.00	54.43	58.40	59.22
	RAW+LIP+GIP	73.21	91.11	81.19	91.77	64.34	49.68	56.07	61.29
Google Perspective		73.64	88.78	80.50	90.99	51.77	36.19	42.60	48.32

* The shaded cells indicate the best results; ** The scores of these metrics are not reported; *** All metrics are reported in percentage.

Table 2. The summary of our proposed method

of the three personality components were included. That is, a model with LIP+GIPs in general outperformed that with RAW+LIPs or RAW+GIPs. These results indicate the importance of intermediate personality traits in effective hate speech detection.

In Table 2-(b), we summarized model performance on SupremacistHate. Aligning with the results on WikiHate, models with LIP and GIP produces good results in general. BERT+LIP+GIP recorded the highest precision rate of 69.61% while RoBERTa+RAW+LIP+GIP reported the highest recall rate of 61.71%. RoBERTa+LIP+GIP produced the best results in terms of F1-score and AUCPR (62.27% and 67.63%, respectively). We argue that these results further validate the aeffectiveness of our design that utilizes personality features in a unique way. It is also important to note that Google Perspective did not perform well on SupermacistHate. This is partly because of Google Perspective being a proprietary software and not being able to fine-tune it on SupremacistHate.

5 Discussion and Implications

The implications of this study are manifold. First, from the methodological perspective, we introduced an automated hate speech detection framework based on personality traits inferred from text. Our method is the first to focus on the intricate relationship of personality and hate. That is, based on the recent discovery of psychology literature, we designed our method, using the multi-head self-attention mechanism, to capture not only low-level but also intermediate personality factors (i.e., LIP and GIP), which have been largely neglected in prior literature. This significantly improved the performance of our personality-based approach in detecting hate comments, outperforming state-of-the-art baselines including Google Perspective across multiple contexts.

Second, from the theoretical perspective, we extended theories of personality factors formed in psychology literature in a hate-speech context. Specifically, several recent studies in psychology have suggested the possible existence of intermediate personality traits bridging the lower-level and the higher-level personality factors (DeYoung et al., 2007). We incorporated this new perspective into our design process and the results strongly suggest that there is indeed a need for more detailed exploration of these intermediate personality factors.

Lastly, our study has practical implications for businesses and society. First, hate speech burdens businesses with additional costs for hiring content moderators and our method can assist in reducing these costs. Social media businesses are increasingly facing regulations from governmental authorities to restrict hate speech on their platforms. As a response, they have employed tens of thousands of workers solely for moderating inappropriate content. For example, Facebook have been spending more than 500 million dollars a year to its outsourcing vendors for regulating toxic content on the platform (Santariano and Isaac, 2021). Second, hate speech entails tremendous social costs as well. Facebook's leaked internal report revealed that cyberbullying on people's bodies on Instagram made teenage girls extremely obsessed with their appearance causing anxiety and depression (Callahan, 2021; Wells et al., 2021). Another research has identified the association between online hate and suicide-related behaviors (Sumner et al., 2021). On top of that, there is a growing body of evidence that hate speech causes psychological problems for those who are hired to monitor it, content moderators. Content moderators of social media companies, continuously being exposed to toxic content, tend to suffer from post-traumatic stress disorder (Arsht and Etcovitch, 2018). Consequently, both businesses and society are putting more and more interests in building automated systems for effectively detecting hate speech and we claim that our framework can play a significant role in such tasks.

This study is not without limitations, and we plan to extend our study in the future. For example, personality traits were deduced based on the writing level rather than the individual level, potentially impacting the accuracy of the results. In addition, a more comprehensive examination should be carried out to assess the applicability of our hate detection method to various subcategories of hate (e.g., sexism, racism, ageism).

References

Arsht, A. & Etcovitch, D. (2018). The Human Cost of Online Content Moderation. Retrieved from https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation.

Alatawi, H. S., Alhothali, A. M., & Moria, K. M. (2021). Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT. IEEE Access, 9, 106363-106374.

Barlett, C. P., & Anderson, C. A. (2012). Direct and indirect relations between the Big 5 personality traits and aggressive and violent behavior. Personality and Individual Differences (52:8), pp. 870-875.

Becerra-García, J. A., García-León, A., Muela-Martínez, J. A., & Egan, V. (2013). A controlled study of the Big Five personality dimensions in sex offenders, non-sex offenders and non-offenders: relationship with offending behaviour and childhood abuse. The Journal of Forensic Psychiatry & Psychology, (24:2) pp. 233-246.

Burton, L. A., Hafetz, J., & Henninger, D. (2007). Gender differences in relational and physical aggression. Social Behavior and Personality: an international journal (35:1), pp. 41-50.

Callahan, M. (2021). Why Does Instagram Have a Negative Effect on Teenagers' Mental Health? Retrieved from https://news.northeastern.edu/2021/09/20/negative-effects-of-instagram/.

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Conversation AI. (2018). Toxic comment classification challenge: identify and classify toxic online comments, Retrieved from https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

Dash, M., & Liu, H. (1997). Feature selection for classification. Intelligent data analysis, 1(1-4), 131-156.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Eleventh international aaai conference on web and social media.

Depue, R. A., & Collins, P. F. (1999). Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion. Behavioral and brain sciences, 22(3), 491-517.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, January). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. Journal of personality and social psychology (93:5), pp. 880.

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C. A., Tobolsky, V. A., Smith, L. K., Buffone, A., Iwry, J., Seligman, M., & Ungar, L. H. (2021). Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. Psychological Methods, 26(4), 398.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. In Twelfth International AAAI Conference on Web and Social Media.

Elzayady, H., Mohamed, M. S., Badran, K. M., & Salama, G. I. (2023). A hybrid approach based on personality traits for hate speech detection in Arabic social media. International Journal of Electrical and Computer Engineering, 13(2), 1979.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (51:4), pp. 85.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. Personality psychology in Europe, 7(1), 7-28.

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. International Journal of Computer Vision, 129(6), 1789-1819.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering (10:4), pp. 215-230.

Google (2019). Perspective API. Retrieved from https://bit.ly/2QsmbA0.

Green, D. P., McFalls, L. H., & Smith, J. K. (2001). Hate crime: An emergent research agenda. Annual review of sociology (27:1), pp. 479-504.

Harell, A. (2010). Political tolerance, racist speech, and the influence of social networks. Social Science Quarterly (91:3), pp. 724-740.

Heaven, P. C. (1996). Personality and self-reported delinquency: Analysis of the Big Five personality dimensions. Personality and individual differences (20:1), pp. 47-54.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

IBM Cloud. (2020). The science behind the service. Retrieved from https://ibm.co/2vFybHl.

Jang, K. L., Livesley, W. J., Angleitner, A., Riemann, R., & Vernon, P. A. (2002). Genetic and environmental influences on the covariance of facets defining the domains of the five-factor model of personality. Personality and individual Differences, 33(1), 83-101.

Jones, S. E., Miller, J. D., & Lynam, D. R. (2011). Personality, antisocial behavior, and aggression: A meta-analytic review. Journal of Criminal Justice (39:4), pp. 329-337.

Jourova, V. (2019). How the Code of Conduct Helped Countering Illegal Hate Speech Online. Retrieved from https://ec.europa.eu/info/sites/info/files/hatespeech_infographic3_web.pdf.

Jovanović, D., Lipovac, K., Stanojević, P., & Stanojević, D. (2011). The effects of personality traits on driving-related anger and aggressive behaviour in traffic among Serbian drivers. Transportation research part F: traffic psychology and behaviour (14:1), pp. 43-53.

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In Twenty-seventh AAAI conference on artificial intelligence.

Lee, D. L., Chuang, H., & Seamons, K. (1997). Document ranking and the vector-space

Lee, K., & Ram, S. (2020). PERSONA: personality-based deep learning for detecting hate speech.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of artificial intelligence research (30), pp. 457-500.

Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. IEEE Intelligent Systems (32:2), pp. 74-79.

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P. & Mukherjee, A. (2019). Thou shalt not hate: Countering online hate speech. In Proceedings of the International AAAI Conference on Web and Social Media, (13:1), pp. 369-380.

Meddaugh, P. M., & Kay, J. (2009). Hate speech or “reasonable racism?” The other in Stormfront. Journal of Mass Media Ethics, 24(4), 251-268.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In Proceedings of the 25th international conference on world wide web, pp. 145-153.

Paunonen, S. V. & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. Journal of personality and social psychology (81:3), pp. 524.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. Journal of personality and social psychology (77:6), pp. 1296.

Santariano and Isaac (2021). The silent partner cleaning up Facebook for $500 million a year, Retrieved from https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html.

Sumner, S. A., Ferguson, B., Bason, B., Dink, J., Yard, E., Hertz, M., Hilkert, B., Holland, K., Mercado-Crespo, M., Tang, S., & Jones, C. M. (2021). Association of online risk factors with subsequent youth suicide-related behaviors in the US. JAMA network open, 4(9), e2125860-e2125860.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems, pp. 5998-6008).

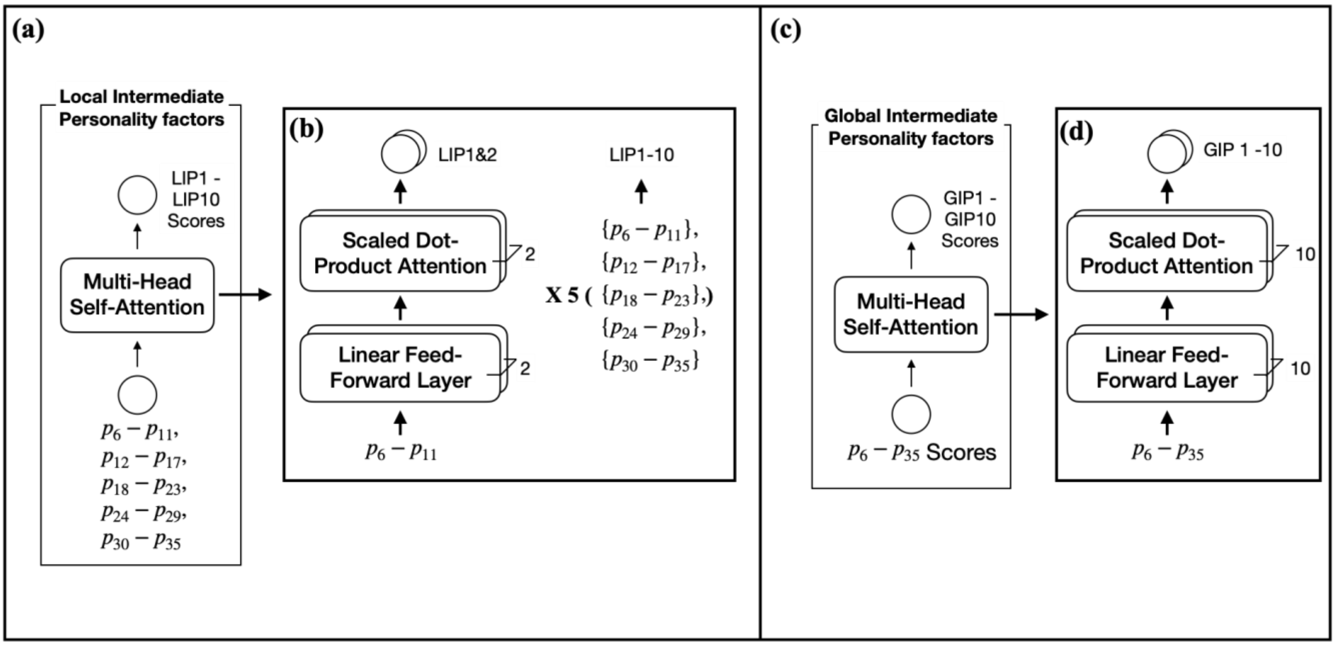
Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In Proceedings of the second workshop on language in social media, pp. 19-26.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop, pp. 88-93.

Wells, G., Horwitz, J., & Seetharama, D. (2021). Facebook Knows Instagram Is Toxic for Teen Girls. Retrieved from https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739.

Zhang, D., Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and SVMperf. Expert Systems with Applications (42:4), pp. 1857-1863.

A Appendix: Methodological Framework



B Appendix: Processes of Inferring Local and Global Intermediate Personality Factors