On the Usefulness of Personality Traits in Opinion-oriented Tasks

Marjan Hosseinia University of Houston {ma.hosseinia}*@gmail.com
Eduard Dragut Temple University edragut@temple.edu
Dainis Boumber University of Houston {dboumber}†@uh.edu
Arjun Mukherjee University of Houston arjun@cs.uh.edu

Abstract

We use a deep bidirectional transformer to extract the Myers-Briggs personality type from user-generated data in a multi-label and multi-class classification setting. Our dataset is large and made up of three available personality datasets of various social media platforms including Reddit, Twitter, and Personality Cafe forum. We induce personality embeddings from our transformer-based model and investigate if they can be used for downstream text classification tasks. Experimental evidence shows that personality embeddings are effective in three classification tasks including authorship verification, stance, and hyperpartisan detection. We also provide novel and interpretable analysis for the third task: hyperpartisan news classification.

1 Introduction

The vocabulary we use in everyday language is a rich source of information about our beliefs, thoughts, and personalities (Pennebaker et al., 2015). Many efforts in text analysis provide compelling evidence that our everyday language carries psychological cues (Gottschalk and Gleser, 1979; Stone et al., 1966; Weintraub, 1989; Pennebaker et al., 2015). With this study, *we seek to determine the personality of a given text's author as defined by the Myers-Briggs Type Indicators or MBTI* (Myers and Myers, 1995). Myers-Briggs uses four binary dimensions to classify people (Introvert–Extrovert, Intuitive–Sensing, Thinking–Feeling, Judging–Perceiving), which gives 16 different types, such as INTJ and ENTJ. This work uncovers novel insights into the personality space of authors from their online writings.

The personality signal carries the fingerprint of the individual's psyche and, even though noisy, can be useful (as shown in this work) for a variety of downstream NLP tasks, such as authorship verification, stance, and hyperpartisan detection. Personality prediction does not only benefit commercial applications and psychology but also is advantageous in health care. Recent works link personality types and social media behavior with depression and posttraumatic stress disorder (Preoţiuc-Pietro et al., 2015). This is significant because it opens new avenues for prevention care as certain personality types can anticipate mental illness and schizophrenia. (Mitchell et al., 2015).

This problem poses a non-trivial challenge because a good solution must be capable of capturing the complexity and the depth of the human psyche as expressed through text. Anything short of that will result in task-specific pattern-matching. It follows that the main technical difficulty presented by the task at hand is the discrepancy between the corpora concerning their distributions, which results in the domain shift. This is where our work becomes relevant as it aims to bridge the gap by transfer learning and universal language understanding.

The problem is also challenging because the human psyche is complex in its nature. The labels are fuzzy as the label distribution changes from population to population and the ground truth is not derived by an objective method; rather, it is a set of ideas generally agreed upon by specialists and society. Furthermore, high quality curated datasets constructed by professional psychologists are difficult to obtain due to privacy reasons.

Models have only recently reached the capacity required. Our approach uses transfer learning through language understanding for personality prediction. We create a unified dataset from the collection of user inputs of three available MBTI datasets (Gjurković and Šnajder, 2018; Mitchell, 2017; Plank and Hovy, 2015) originating from social media platforms including Reddit, Twitter, and Personality Cafe forum. We investigate how transfer learning with pretrained transformers con-

Proceedings of Recent Advances in Natural Language Processing, pages 547–556
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_062

tributes to personality prediction under a multi-label multi-class classification strategy. We analyze the relationship between personality types with the three specific tasks of stance, authorship, and hyperpartisan news classification. The results on the unified dataset show that transfer learning along with pretrained bidirectional transformer models effectively changes the Hamming loss, F1, and Jaccard similarity for multi-label personality prediction. The contributions of our paper are listed below:

- We propose to use the flow of sentiments across a document as a proxy for Myer-Briggs personality type and use a transformer-based model to predict personality type.
- We show the usefulness of personality traits on three downstream text classification tasks: authorship verification, stance, and hyperpartisan detection. The technical novelty is on transfer learning of our pretrained personality model to improve NLP downstream tasks.
- We give an in-depth statistical analysis of the effect of using personality information in the task of hyperpartisan news classification.

In the following sections we introduce and evaluate the personality model, then analyze its application in the three text classification tasks (mentioned above) using transfer learning.

2 Related Work

Personality prediction from text is a challenging task (Štajner and Yenikent, 2021; Yang et al., 2020) and many personality prediction approaches rely on crafted features which can range from simple ones, such as TF-IDF of word or character n-grams to the ones produced by tools such as Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015), which extracts anything from low-level information such as Part Of Speech tags and topical preferences to psychological categories. These features are often supported by various psycholinguistic word lists that aim to detect emotions and sensory experiences (Preoţiuc-Pietro et al., 2017).

Our work uses a bidirectional transformer to predict MBTI personality types using a large collection of data obtained from three existing personality datasets. Utilization of the pretrained word embeddings (Mikolov et al., 2013; Pennington et al., 2014) in many deep learning models indicates that leveraging knowledge obtained from unsupervised learning boosts the performance. Recently, language models pretrained on a large amount of raw text were shown to provide representations applicable to a wide variety of tasks with minimal fine-tuning (Radford et al., 2018; Howard and Ruder, 2018; Peters et al., 2018a). These models can be effectively generalized to many downstream tasks and adapted to different domains. Below are the three representative studies on utilizing online user-generated text for personality prediction. They are annotated with self-reported MBTI personality types of users.

Reddit9K dataset is a large-scale dataset constructed from the posts and comments of 9K Reddit users. It is labeled with MBTI indicators and covers a wide variety of topics (Gjurković and Šnajder, 2018). The authors extract user activity and linguistic features including word and character n-grams, LIWC word categories (Pennebaker et al., 2015), and two Psycholinguistic dictionaries (Preoţiuc-Pietro et al., 2017; Coltheart, 1981). Support Vector Machine (SVM), Logistic Regression (LR), and multi-layer perceptron are used to identify personality types and prove to be discriminative for personality prediction. **Twitter** dataset is a large corpus of 1.2M tweets of 1.5K users (Plank and Hovy, 2015). Experiments performed by the dataset creators show that linguistic features are reliable representatives for two out of four personality dimensions. We hypothesize that the cause of the discrepancy is the difference between the distribution of personality types in social media users and the general U.S. population. Finally, **Kaggle** dataset collects the user posts of the Personality Cafe[1] forum and covers 8.6K different people with 16 MBTI personality types (Mitchell, 2017).

3 Dataset

We use Reddit9k, Twitter, and Kaggle Myers-Briggs personality type datasets to train and evaluate our proposed model for automatic personality type prediction. In all datasets, the annotation process relies on self-reported personality types, and no questionnaire is given to the users. Previously, *MyPersonality* created from Facebook user data was a questionnaire-based dataset. However, it is not available to the public anymore. We make a unified dataset from the collection of the three available MBTI personality datasets and remove

[1] https://www.personalitycafe.com

Set	Size	Size/p.type	Size/p.dim
Train	558,352	34,897	279,176
Dev	79,776	4,986	39,888
Test	159,520	9,970	79,760

Table 1: Unified personality dataset statistics; p.:personality; # of dimensions: 4; # of types: 16

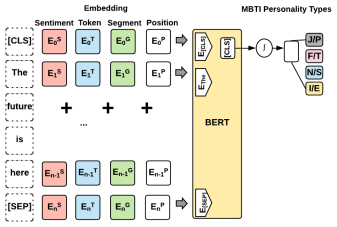

Figure 1: PersBERT model architecture

the non-English contents. We find that the new dataset is highly skewed towards two out of four personality dimensions. There are a few reasons for that. i) According to Plank and Hovy (2015) the distribution of personality types among the United States population is not balanced. ii) Users from some specific personality types tend to participate in social media platforms and express their personality types more than others. Our experiments also show that the class imbalance highly affects training, generating poor results for small classes, among evaluation methods. To alleviate the skewness of the data in training we take two (standard) steps: add class weights concerning their size in loss computation and make a balanced subset of the original dataset. We notice that the former does not improve the performance significantly, but the latter does. Hence, we create a balanced version of the dataset by over-sampling the small and under-sampling the large classes such that their final sizes become equal to the original average size of the 16 MBTI personality types before sampling. Table 1 reports the unified personality dataset statistics after balancing.

4 Personality Embedding

We build a general model to predict four MBTI personality dimensions and to infer personality embedding. The MBTI dimensions are expressed as Booleans (0/1). The personality dimensions are IE, or Introversion (I)/Extroversion (E); NS, or iNtuition (N)/Sensing (S); FT, or Feeling (F)/Thinking (T); and JP, or Judging (J)/Perceiving (P). Under this scheme, each instance can have multiple labels with four classes. The combination of these four classes gives $2^4 = 16$ MBTI personality types. We consider multi-labeling classification to learn the personality dimensions together. Our experiments show that sub-sampling creates a small training set with poor final results while over-sampling creates a huge dataset with hundreds of redundant examples. So, the models cannot differentiate the 16 classes (personality types) with the under-sampled small training data or they fail to predict the unseen examples of minority classes correctly in the over-sampled dataset as they get overfitted by the redundant examples. Finally, we find that by converting the 16 classes into four in the multi-labeling scheme and applying over-sampling and sub-sampling simultaneously we can better overcome the class imbalance. The following section describes the proposed personality prediction model.

4.1 PersBERT Model

The use of pretrained language models and transformers shows significant improvements in various NLP problems. The bidirectional based transformer models such as BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019) overcome previously published language models trained on one direction (e.g. ULMFiT) (Howard and Ruder, 2018) or the shallow concatenation of left and right direction of text input (e.g. ELMo) (Peters et al., 2018b) in various text classification tasks. We use BERT architecture as the basis of our personality prediction model. BERT takes position, segment, and token embedding as input to compute the importance of a token in a sequence. For personality classification purposes, we take into account the sentiment of sentences in an input sequence aside from the standard BERT input. According to (Tausczik and Pennebaker, 2010), the level of emotion and sentiment expression by people in their opinions and the way they express their emotions define how people feel about the world. People's everyday language is a rich source of their beliefs, thinking patterns, and personality. Because personality speaks of stable differences in characteristic patterns of thinking, feeling, and behaving, it is connected with emotion and sentiment (Corr and Matthews, 2009). In this regard, some tools such as LIWC are designed to organize the words in psychologically meaningful categories and to identify emotion in language. They are also widely used in Psycholinguistic studies (Tausczik and Pennebaker, 2010). The connec-

tion between language, emotion, and personality elevates opinionated user generated content into a valuable resource for mining people's personalities.

In our approach we split the input sequence into linguistic sentences.[2] The sentiment of each sentence is one of positive, negative, or neutral; it can be inferred using any sentiment analysis tool. We give the utilized tools in Section 4.3. The sentiments of input tokens are embedded using a $3 \times k$ matrix that is randomly initialized, where k is the size of the hidden states of the model. Then, these sentence-wise sentiment embeddings are accumulated with the three standard embeddings of the BERT model ($E_t^{token}, E_t^{position}, E_t^{segment}$) to form the input embedding (E_t for token t). So, $E_t = E_t^{token} + E_t^{position} + E_t^{segment} + E_t^{sentiment}$.

Figure 1 shows the model architecture. The input embeddings are given to the BERT sentence classification model that takes a sequence of linguistic sentences as one single input compared to the sentence-pair model that takes two inputs (e.g. a question and its answer). A fully connected layer forms a classifier that squeezes the pooled output (x) of the BERT model to four personality dimensions (I/E, N/S, F/T, and J/P). The hidden state of [CLS] token ($h_{[\text{cls}]}$) is used as the input of the pooling layer. So, $x = \tanh(W^p h_{[\text{cls}]} + b^p)$, $\text{logit} = W^c x + b^c$ where W^c, W^p, b^p, and b^c are the layers' parameters. Similar to other multi-label multi-class problems, the loss is the overall binary cross entropy among all classes, $L = \frac{1}{CN} \sum_{i \in N, c \in C} y_{c,i} \log \sigma(y'_{c,i}) + (1 - y_{c,i}) \log(1 - \sigma(y'_{c,i}))$.

where N is the number of examples, C number of classes, σ sigmoid function and y, y' are true labels and logits (input of probability function) respectively. We refer to the proposed model as PersBERT for the rest of the paper.

4.2 Multi-class Multi-label Baselines

We mentioned earlier that personality is connected to emotion and sentiment (Tausczik and Pennebaker, 2010; Corr and Matthews, 2009). Also, automatic prediction of MBTI personality is being considered under a multi-label setting. Thus, we choose baselines with various architectures that are widely used in sentiment analysis or multi-label classification. They are listed as follows: **Kim-CNN** & **XML-CNN** are two CNN-based neural network models. The former is one of the initial and successful applications of Convolutional Neural Network (CNN) for text classification (Kim, 2014). And the latter is designed for extreme multi-label text classification where the number of labels can exceed even a few thousand (Liu et al., 2017). Its architecture inherits Kim-CNN's model specification with an additional dynamic max-pooling layer that highlights important information across different parts of a document. XML-CNN was able to beat most of the deep learning baselines in six benchmark datasets. **DocBERT** is the BERT model with a fully connected layer that converts the hidden state of the BERT pooling layer to C activations for C-class classification (Devlin et al., 2019). The pooling layer pools the model by taking the hidden state corresponding to the classification token ([CLS]) of the input sequence through non-linearity ($\tanh$). We fine-tune DocBERT for classification and initialize it with pretrained BERT-base-uncased weights. Lastly, **Hierarchical Attention Network (HAN)** is a recurrent neural network model that mirrors the hierarchical structure of the English language (Yang et al., 2016). Applying attention mechanisms in word and sentence-level enables this model to find crucial parts of the document for the downstream classification task. The model outperforms its competitive baselines in sentiment analysis of user reviews dataset including Yelp, Amazon, and IMDB.

4.3 Evaluation

We train the models on 30 epochs with the batch size of 16 or 32. Training is controlled by early stopping with patience = 5, which will be stopped after 5 consequent epochs of no improvement of the highest F1 score gained. The test set is evaluated using the model with the best F1 of the dev set. We use Google News (GNews) (He et al., 2020; Liu et al., 2015) and FastText token embedding in our experiments for the two CNN-based (Kim-CNN & XML-CNN) and LSTM-based baselines (HAN) (Mikolov et al., 2018). However, DocBERT and PersBERT models' parameters are initialized with their corresponding BERT-base-uncased model weights. The BERTAdam optimizes these two models with the learning rate of $2e-5$ recommended in (Devlin et al., 2019). We set the sequence length = 256 for all models. Similar to DocBERT, all parameters of PersBERT are updated during backpropagation. Training PersBERT with more than $5K$ examples (Table1) takes

[2] we use NLTK sentence tokenizer (Loper and Bird, 2002).

Method	Jacc.	Hamm.	Ma.-F1	Mi.-F1
Kim-CNN, GNews	46.82	41.76	62.82	63.78
Kim-CNN, FastText	45.83	39.31	62.23	62.86
XML-CNN, GNews	44.72	45.76	56.69	61.80
XML-CNN, FastText	47.97	40.96	64.0	64.83
HAN, GNews	46.62	41.18	63.03	63.59
HAN, FastText	46.29	38.48	62.83	63.29
DocBERT	86.03	7.46	92.47	92.49
PersBERT	**86.97**	**6.94**	**93.03**	**93.03**

Table 2: Personality prediction on the unified dataset of Table 1; Jacc.:Jaccard, Hamm.:Hamming

5 days on a TITAN RTX GPU with batch size=16.

For evaluating multi-label personality prediction, we use Jaccard Similarity, Hamming loss, Macro-F1, and Micro-F1 scores. For more information about the measures the reader is directed to (Wu and Zhou, 2017). We use scikit-learn library (Pedregosa et al., 2011) for evaluation measures and other statistical methods. We utilize VADER, a rule-based model for the general sentiment analysis task, to infer the sentiment of sentences (Hutto and Gilbert, 2014). VADER gives us a compound sentiment score between -1 and +1. The scores between -1 and -0.05 indicate negative, the ones grater than 0.05 show positive sentiment, and the scores between -0.05 and + 0.05 have a neutral sentiment. Each token inherits the sentiment of the sentence in which the token appears. For the two classification ([CLS]) and separator [SEP] tokens, we use neutral embedding. Although VADER is a token-based sentiment tool, we use sentence-wise sentiment instead of token-wise for two reasons: i) our intuition is to let the model learn the transition of sentiment across sentences and not tokens; this follows from the assumption that the change of sentiment from sentence to sentence may indicate one's personality. ii) BERT uses sub-words units known as Word-pieces and each VADER lexicon may be composed of multiple Word-pieces. Thus, we must assign the sentiment of an entry in VADER lexicon to all its Word-pieces. For example, the sentiment of 'huggable' must be assigned to its three sub-words in our model: ['hug', '##ga', '##ble']. Also, our experiments on the dev set show that token-wise sentiment avoids learning the transition of sentiments and does not improve the model performance as much as sentence-wise sentiment.

Experimental results of multi-label MBTI personality prediction on the unified dataset (Table 1)

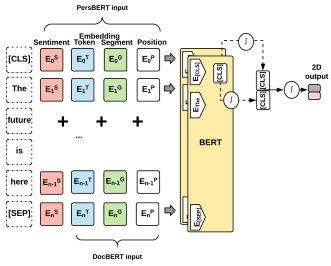

Figure 2: DocBERT+PersBERT model architecture

are provided in Table 2. They indicate that PersBERT trained on 256 tokens of input sequence achieves the best results among baselines in all multi-class multi-label evaluation measures. An F1 improvement of $+0.5\%$ and -0.52% Hamming loss reduction on $\geq$ 159K test instances compared to DocBERT shows that adding sentiment embedding of sentences to the input distinguishes the personality types more accurately. Apart from that, both transformer-based models, DocBERT and PersBERT, show significant improvement over the two CNN models with two different pretrained embeddings as well as the HAN (about $+40\%$ in Jacc. and -30% in Hamming loss). We believe that the two masking and next sentence prediction techniques used in BERT's pretraining enable the model to better understand the relationship between the words of a sentence in both directions, as well as the relationship between the sentences. This leads to an enriched language model and a remarkable improvement in identifying personality signals from individuals' language compared to other baselines.

5 Transfer Learning

We aim to study if the knowledge gained from our personality prediction model, PersBERT, helps opinion-oriented problems. Personality is closely connected with *opinions* and how people form opinions; hence, we choose three tasks, i.e., hyperpartisan news detection, authorship verification, and stance detection (Hosseinia et al., 2020), that are designed around opinion mining. We create our transfer learner named **DocBERT + PersBERT** by connecting the pretrained personality model, PersBERT, to DocBERT (Figure 2). Empirical results show that these transformer models achieve the best results among the baselines introduced in Section 4.2. We share a fully connected layer, classifier layer, between the DocBERT and PersBERT models. The layer takes the concatenation of the

News dataset	train	dev	test
by-article	386	129	130
by-publisher	164,944	118,510	149,794

Table 3: Hyperpartisan news dataset

output vector of pooling layers and converts them to C classes. Recall that in DocBERT and PersBERT, the pooling layer pools the model by taking the hidden state corresponding to [CLS] token of input sequence using a non-linear activation ($\tanh$). Hence,

$$x_{[\text{cls}]}^{\text{PersBERT/DocBERT}} = \tanh(W^{p/d} h_{[\text{cls}]} + b^{p/d}),$$
$$z = W^{c}[x_{[\text{cls}]}^{\text{DocBERT}}; x_{[\text{cls}]}^{\text{PersBERT}}] + b^{c}$$

where $W^{p/d}, b^{p/d}$ are the parameters of PersBERT or DocBERT pooling layers. $[;]$ denotes concatenation, W^c is the $2 * k \times C$ classifier weight matrix and k is the size of the pooled vectors (hidden state). Finally, the classifier output, z, is normalized with a Softmax function for downstream $C-$class classification tasks. Next, we introduce the three text classification problems for our evaluations.

5.1 Hyperpartisan News Detection

The term "hyperpartisan news" is used to define the extremely biased news in favor of the right or the left political spectrum. SemEval2019 task 4 proposes hyperpartisan news detection and has released only the training and dev sets of two versions of the hyperpartisan news dataset. In the first version, "news by-publisher", all articles are labeled by the overall bias of the publisher as provided by BuzzFeed[3] journalists or MediaBiasFactCheck.com while in "news by-article" dataset documents are labeled manually by the agreement of the journalists (Kiesel et al., 2019). Because the test set is not released yet, we use SemEval dev set as test and split its training set into new training and dev sets with no publishers in common. Likewise, we create new sets for the "news by-article" dataset. Our topic modeling analysis on the "news by-publisher" training set reveals that it is highly imbalanced in terms of news classes. We use Non-negative Matrix Factorization (NMF) to estimate topic distribution in news. For some topics, top documents belong to only one class. To avoid the models to learn topics but hyperpartisanship we sample from the training set so that the resulting set includes an equal number of unique examples per topic. We only apply sub-sampling on the training set and keep the dev and test set intact. This process increases the F1 score of DocBERT model by 5% on the dev set. Table 3 provides the dataset statistics.

5.2 Stance Detection

Stance detection identifies if an opinion supports an idea or contradicts it. We use the new version of Procon dataset (Hosseinia et al., 2019) in our evaluation. The dataset covers the argumentative opinions of different controversial issues, ranging from education and immigration to birth control. The dataset has $4,264$ instances and we split it into $(70\%, 10\%, 20\%)$ for train, dev and test, respectively. As each instance in Procon dataset is a pair of a question about an issue and an opinion about it, we use the BERT sentence-pair model for both DocBERT and DocBERT + PersBERT models. Thus, the input of the two BERT-based models is formed as *[CLS] question [SEP] opinion [SEP]* where *[CLS], [SEP]* are reserved tokens used by BERT for classification and separation of the two input parts respectively (Devlin et al., 2019).

5.3 Authorship Verification

Authorship Verification (AV) identifies whether a pair of documents are written by the same author. It has applications in plagiarism detection, forensic analysis, and sockpuppet detection, to name a few. We examine our model on three standard PAN AV datasets[4]. Each dataset contains one training and one test set. We split the original training set into $(70\%, 30\%)$ for training and dev, and evaluate on the original test set. Similarly, each instance in the AV dataset is a pair of documents, so, we use the BERT sentence-pair model for both DocBERT and DocBERT + PersBERT models. The input is formed as *[CLS] first document [SEP] second document [SEP]* where documents are written by one or two unknown author(s) and may contain several (linguistic) sentences.

5.4 Results and Analysis

The settings, training strategy, and baselines are the same as PersBERT's (Table2, Section4.3) for the three aforementioned opinion-oriented clas-

[3] https://www.buzzfeed.com

[4] https://pan.webis.de/data.html

	DocBERT			DocBERT+PersBERT		
Dataset	P	R	F1	P	R	F1
Procon	72.89	77.65	75.20	77.09	82.87	**79.87**
PAN2014E	65.09	69.70	67.32	62.41	83.84	**71.55**
PAN2014N	60.33	73.74	66.36	65.15	86.87	**74.46**
PAN2015	44.26	75.60	55.83	59.11	68.80	**63.59**
News by-art.	77.27	72.34	74.73	84.62	70.21	**76.74**
News by-pub.	61.48	38.33	**47.22**	57.04	33.29	42.05

Table 4: Effect of personality in three opinion-oriented tasks; P: Precision; R: Recall

sification tasks. All parameters of the transformer models, including DocBERT, PersBERT, and DocBERT + PersBERT are updated during backpropagation. For AV datasets with pairs of inputs, the overall length of the two input documents does not exceed 512 tokens while the maximum length of input for other datasets is 256.

Transfer learning results are provided in Table 4 for six different datasets. We only report DocBERT results because it achieves much higher performance compared to other baselines for the benchmark datasets. For the stance detection task (Procon dataset) DocBERT + PersBERT beats DocBERT by more than 4.5% of F1 overall. We plot the distribution of 16 personality types among the pro and con classes for two different issues (Figure 3). From the plots we find that i) personality distribution among pro and con classes depends on the underlying topics and varies in different topics and ii) pro and con-arguments have different personality types based on the underlying topic. We thus note that personality distribution provides distinctive signals for stance in each topic. This personality difference between stance classes shows why adding personality information to the BERT model results in more accurate differentiation of proponent and opponent arguments, which in turn improves the final F1. The results on all three authorship verification datasets exhibit a similar trend. DocBERT + PersBERT outperforms the BERT classification model by more than $4\%, 8\%$, and 7.7% in PAN2014 Essay, PAN2014 Novel, and PAN2015 respectively.

Despite the improvements of F1 in "news by-article" results using personality information, we do not see the same effect on hyperpartisan "news by-publisher" results. There is a reduction by -5.17% of F1 in "news by-publisher" when personality information is added. We hypothesize that there are two main reasons for this behavior: first,

Training set	$\overline{\text{Entropy}}$	Model	B.Acc.
topic "econ"	0.2795	DocB.	50.02
		DocB.+PersB.	**55.51**
topic "life"	0.6837	DocB.	**52.82**
		DocB.+PersB.	50.15

Table 5: Effect of topic and personality based sub-sampling of training set on "news by-publisher"; B.Acc.: balanced accuracy; test and dev sets are the original sets

		DocB.			DocB.+PersB.		
Train data	$\overline{\text{Entropy}}$	**P**	**R**	**F1**	**P**	**R**	**F1**
all	0.99	61.48	38.33	47.22	57.04	33.29	42.05
sub-sampled	0.09	61.69	44.09	51.43	58.03	47.16	**52.04**$*$

Table 6: "News by-publisher" results with entropy-based sampling; $*$:p-value of McNemar's test $\leq 10^{-5}$

the PersBERT model is trained on social media data while news data is formal and usually follows its publisher's strict writing regulations. It may lead to hiding the author's informal writing and personality features. This difference between the language of news and social media data challenges the effect of transfer learning between the two domains. Secondly, it is expected that personality distribution differs between mainstream and hyperpartisan classes for different news topics, similar to what we observed in stance detection. The following section provides additional experiments for hyperpartisanship detection.

5.4.1 A Deeper Look into Hyperpartisanship

The experiments reported above show that the proposed approach is not useful in identifying hyperpartisanship in the "news by-publisher" dataset. However, we anticipate there are some *connections* between personality and hyperpartisanship as opinion forms a bridge between these two concepts. We design the following experiments to investigate the hidden connection. Articles of the large "news by-publisher" dataset (Table 3) cover a wide variety of topics. So, we investigate whether personality types vary in mainstream and hyperpartisan classes for separate news topics. We first model topics of the news training set using the Non-negative Matrix Factorization (NMF) algorithm for 20 topics. Then, we choose distinct articles for each topic and induce MBTI personality dimensions using PersBERT (Section 4.1). We select these two topics to minimize the influence of per topic-personality distribution. Later in this section, we measure the relationship between personality distribution and

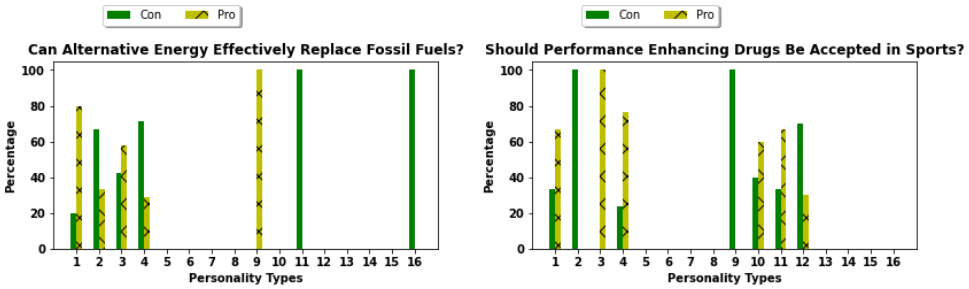

Figure 3: MBTI personality types distribution among proponents and opponents for Can Alternative Energy Effectively Replace Fossil Fuels? (left), Should Performance-Enhancing Drugs Be Accepted in Sports? (right)

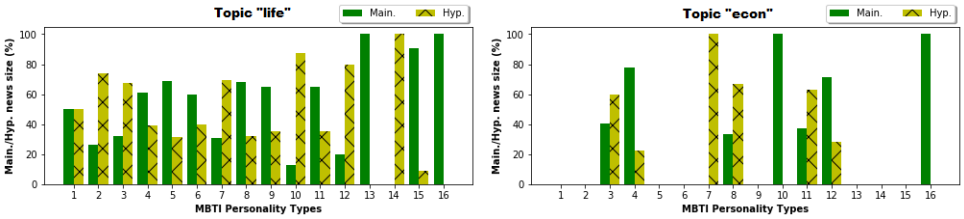

Figure 4: Personality distribution among main. and hyper. news for topics "life" (left) and "econ" (right)

news topics using *entropy*. As noted before, a tuple of four dimensions gives us an MBTI personality type. We plot 16 MBTI personality types versus two news classes for the two selected topics (Figure 4). According to the figure, there is a remarkable difference between several personality types of the two news classes for the topic "econ." 100% of personality types 7 and 10 are from hyperpartisan and mainstream news classes, respectively; about 80% of personality type 4 belongs to the mainstream, while 70% of type 8 forms hyperpartisan news. However, the plot of topic "life" shows less difference in news distribution among personality classes. We report average *entropy* of the two news classes across all personality types to measure the difference of personality distributions between the two news classes (Table 5).

$$\overline{\text{entropy}} = \frac{1}{|T|} \sum_{t \in T} \sum_{i \in I} -p_{i,t} \log_2(p_{i,t})$$

Where T is the set of all personality types, $I = [Hyp., Main.]$, $|\cdot|$ denotes the size, and $p_{i,t} = \frac{n_{i,t}}{\sum_{i \in I} n_{i,t}}$ is the proportion of news class if in personality class t. The smaller the entropy, the more the two news classes have different distributions among the 16 personality types. We train both DocBERT and DocBERT+PersBert on about 500 articles from topic "life" and "econ," separately and evaluate it on the original test set. According to Table 5, training on topic "econ" with lower entropy results in higher balanced accuracy of DocBERT+PesrBERT with an improvement of $> 5\%$ for the sequence length$= 256$. It shows that topic-based sub-sampling gives us a more distinctive representation of personality types that contributes to better hyperpartisan news detection. On the other hand, training on the data with higher entropy (topic "life") results in lower accuracy of DocBERT+PersBERT compared to DocBERT indicating that adding unbiased personality features makes the differentiation between the two news classes harder. Moreover, we sub-sampled the whole training data such that the average entropy does not exceed the low amount of 0.1. The sampling gives us $111,614$ training examples for $\overline{\text{entropy}} \leq 0.1$. Results in Table 6 reveal that training on data with unbalanced personality distribution of news classes results in remarkable improvements for both models with higher F1 of DocBERT+PersBERT.

6 Conclusion

We believe that this work lays the foundation for leveraging personality signals in a variety of opinion-oriented tasks. We first proposed a novel model, PersBERT, that jointly models the sentence-specific sentiment and personality information building upon the BERT architecture to predict the MBTI personality dimensions. Our pretrained personality transformer improves BERT results and other baselines in benchmark datasets on the personality task. Further, our proposed model was

used on different downstream NLP tasks providing major improvements showing that the subtle signals of user sentiment and their connection with personalities captured by our model are useful in real-world NLP tasks. It is worthwhile to note that the performance comes from training using only short sequences of online user posts (i.e. noisy data for personality). We believe the improvements of our model can be more pronounced if trained upon large-scale gold standard personality datasets (e.g. curated using controlled experiments which is a potential future work). We find that personality signals are more distinctive in authorship verification and stance detection than hyperpartisan news detection where the data is sourced from formal and more supervised writings. However, our personality embeddings can still be used for an effective sub-sampling even in hyperpartisan news detection. Our architecture allows for novel analysis and insights that were previously unknown and have the potential to improve various other NLP tasks which we defer for future exploration.

Acknowledgments

Research was supported in part by grants NSF 1838147, NSF 1838145, ARO W911NF-20-1-0254. The views and conclusions contained in this document are those of the authors and not of the sponsors. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Philip J Corr and Gerald Matthews. 2009. *The Cambridge handbook of personality psychology*. Cambridge University Press Cambridge.

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

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

Louis A Gottschalk and Goldine C Gleser. 1979. *The measurement of psychological states through the content analysis of verbal behavior*. Univ of California Press.

Lihong He, Chao Han, Arjun Mukherjee, Zoran Obradovic, and Eduard C. Dragut. 2020. On the dynamics of user engagement in news comment media. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 10(1).

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2019. Pro/con: Neural detection of stance in argumentative opinions. In *Social, Cultural, and Behavioral Modeling*, pages 21–30, Cham. Springer International Publishing.

Marjan Hosseinia, Eduard C. Dragut, and Arjun Mukherjee. 2020. Stance prediction for contemporary issues: Data and experiments. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, SocialNLP@ACL 2020, Online, July 10, 2020*, pages 32–40. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124. ACM.

Qingyuan Liu, Eduard C. Dragut, Arjun Mukherjee, and Weiyi Meng. 2015. Florin: A system to support (near) real-time applications on user generated content on daily news. 8(12).

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mitchell. 2017. Myers-briggs personality type dataset.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20.

Isabel Briggs Myers and Peter B Myers. 1995. *Gifts differing: Understanding personality type*. Davies-Black Publishing.

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

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Jeffrey Pennington, Richard Socher, and Christopher 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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proc. of NAACL*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018b. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98.

Daniel Preoţiuc-Pietro, Jordan Carpenter, and Lyle Ungar. 2017. Personality driven differences in paraphrase preference. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 17–26.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Sanja Štajner and Seren Yenikent. 2021. Why is mbti personality detection from texts a difficult task? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3580–3589.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Walter Weintraub. 1989. *Verbal behavior in everyday life*. Springer Publishing Co.

Xi-Zhu Wu and Zhi-Hua Zhou. 2017. A unified view of multi-label performance measures. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3780–3788. JMLR. org.

Fan Yang, Eduard Dragut, and Arjun Mukherjee. 2020. Predicting personal opinion on future events with fingerprints. In *COLING*, pages 1802–1807.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.