


Myers-Briggs personality classification from social media text using pre-trained language models


Vitor Garcia dos Santos

(University of São Paulo, Brazil)

 <https://orcid.org/0000-0003-4856-4043>, vitorgds95@gmail.com)

Ivandr  Paraboni

(University of S o Paulo, Brazil)

 <https://orcid.org/0000-0002-7270-1477>, ivandre@usp.br)

Abstract: In Natural Language Processing, the use of pre-trained language models has been shown to obtain state-of-the-art results in many downstream tasks such as sentiment analysis, author identification and others. In this work, we address the use of these methods for personality classification from text. Focusing on the Myers-Briggs (MBTI) personality model, we describe a series of experiments in which the well-known Bidirectional Encoder Representations from Transformers (BERT) model is fine-tuned to perform MBTI classification. Our main findings suggest that the current approach significantly outperforms well-known text classification models based on bag-of-words and static word embeddings alike across multiple evaluation scenarios, and generally outperforms previous work in the field.

Keywords: Natural language processing, text classification, Myers-Briggs, MBTI, personality, author profiling

Categories: I.2.7

DOI: 10.3897/jucs.70941

1 Introduction

Human personality - a set of relatively stable behaviour patterns of an individual [Allport and Allport, 1921] - has been the focus of studies in multiple disciplines, and it is well-known to computer science through personality models such as the Big Five [Goldberg, 1990] and, perhaps to a lesser extent, the Myers-Briggs Type Indicator (MBTI) [Myers, 1962]. Models of this kind associate word choices made by an individual (e.g., a customer, a social media user etc.) to pre-defined personality categories (e.g., extroverts versus introverts), allowing us to assess their personality traits for a wide range of practical applications in both natural language interpretation [dos Santos and Paraboni, 2019, dos Santos et al., 2020] and generation [Teixeira et al., 2014].

Personality assessment may however require the use of personality inventories (e.g., [John et al., 1991]) with the aid of specialists, which may become costly in large scale. As an alternative to this, studies in Natural Language Processing (NLP) and related fields have addressed the relation between language use and personality to develop methods for automatically detecting the personality traits of individuals based on text samples that they have written (e.g., on social media etc.) [Plank and Hovy, 2015, Liu et al., 2017, dos Santos et al., 2017, Wu et al., 2020].

Personality detection from text may be seen as an instance of author profiling, that is, the task of inferring an author's demographics based on text samples that they have

authored [Rangel et al., 2020, Polignano et al., 2020, Price and Hodge, 2020, López-Santill et. al., 2020, Silva and Paraboni, 2018a, dos Santos et al., 2019, Delmondes Neto and Paraboni, 2021]. As in other profiling tasks (e.g., author’s gender or age detection), studies of this kind are usually implemented with the aid of supervised machine learning based on text corpora labelled with personality information. The issue of personality classification from text with a specific focus on the MBTI personality model is the subject of the present study.

Existing work in MBTI classification from text generally follows much of the same methods seen elsewhere in NLP, which usually comprise the use of a bag-of-words model or, more recently, static word embeddings such as those provided by Word2vec [Mikolov et al., 2013] and similar approaches. We notice, however, that more recent text representation models - in particular, context-sensitive embeddings such as those provided by Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2019] - are still relatively uncommon in MBTI classification, even though these models have been shown to obtain state-of-the-art results in a wide range of NLP applications from sentiment analysis [Hoang et al., 2019] to author identification [Barlas and Stammatos, 2020], and many others.

Based on these observations, the present work addresses the use of pre-trained BERT language models for MBTI personality classification from text written in multiple languages. In doing so, our objective is to show that by fine-tuning BERT to the present task we may significantly outperform the use of other text representation models across these evaluation scenarios, with two main contributions to the field:

- 1 BERT-based models for MBTI personality classification from text in multiple languages.
- 2 Robust, cross-validation results shown to be consistently superior to those obtained by bag-of-words and static word embeddings alike, and to previous work in the field.

The remainder of this paper is structured as follows. Section 2 reviews recent approaches to MBTI personality classification from text. Section 3 introduces a number of computational models for the task - including the use of pre-trained language models and baseline alternatives - and the datasets to be taken as a basis for our experiments. Section 4 presents results obtained by these models, and Section 5 summarises our findings and describes opportunities for future work.

2 Background

In what follows we review existing work in MBTI personality classification, and briefly discuss opportunities for using pre-trained language models in this task.

2.1 Related work

Table 1 summarises a number of recent studies in MBTI personality classification from text by reporting the target language (Ar=Arabic, En=English, De=German, Du=Dutch, It=Italian, Fr=French, Pt=Portuguese, Sp=Spanish, In=Indonesian), domain (T=Twitter, R=Reddit, F=Facebook, O=online forums, E=essays, V=vlogs), machine learning

method (lr=logistic regression, svm=support vector machine, nb=Naive Bayes, rf=Random Forest, ens=ensemble, seq=BERT sequence learner, svd=singular value decomposition, xg=XGBoost), and learning features (w=word, c=character, pos=part-of-speech, u=user attributes, n=network attributes, p=psycholinguistic features from LIWC [Pennebaker et al., 2001] and MRC [Coltheart, 1981], t=LDA topics [Blei et al., 2003], w2v=Word2vec [Mikolov et al., 2013] and BERT [Devlin et al., 2019] word embeddings, s=text statistics). Further details are discussed individually as follows.

Study	Language	Domain	Method	Features
Plank & Hovy	En	T	lr	w,u,n
ben Verhoeven et al.	De,Du,It,Fr,Pt,Sp	T	svm	w,c
Lukito et al.	In	T	nb	w,s,pos
Alsadhan & Skillicorn	Ar,De,Du,En,It,Fr,Pt,Sp	T,O,E,F	svd	w
Gjurković & Šnajder	En	R	lr,mlp,svm	c,w,p,t,n
Keh & Cheng	En	O	seq	BERT
Katiyar et al.	En	T,O	nb	w
Wu et al.	En	R	lr	BERT
Das & Prajapati	En	O	ens	w,w2v
Abidin et al.	En	O	rf	s
Khan et al.	En	O	xg	w
Amirhosseini & Kazemian	En	O	xg	w

Table 1: Related work

The work in [Plank and Hovy, 2015] is among the first of its kind to address the issue of MBTI personality classification in an open-vocabulary approach, that is, without resorting to personality lexicons or similar resources. The work addresses personality classification in the Twitter domain by using logistic regression over word n-grams, user (e.g., user’s gender) and network (e.g., number of social media followers etc.) features. Results are shown to outperform a majority class baseline.

The work in [ben Verhoeven et al., 2016] introduces the TwiSty corpus, a large multilingual Twitter dataset labelled with MBTI information in six languages (German, Dutch, French, Italian, Portuguese, and Spanish.) The corpus conveys 34 million tweets written by over 18 thousand users. A significant portion of the data (about 59%) concerns Spanish texts, which makes the other languages much less represented (e.g., 2.2% in German, and 2.6% in Italian.) Since some personality traits are naturally rarer than others, the corpus is also heavily imbalanced across MBTI classes. To illustrate the use of the corpus data, results from a linear SVM classifier and majority class are presented.

The study in [Lukito et al., 2016] addresses MBTI classification from Twitter data in the Indonesian language by comparing a number of models based on Naive Bayes classification, TF-IDF and part-of-speech counts. Among these, standard Naive Bayes text classification is found to be the overall best strategy.

The work in [Alsadhan and Skillicorn, 2017] presents a comprehensive investigation of both Big Five and MBTI personality classification in multiple corpora and languages. The method uses single value decomposition (SVD) to discriminate extreme personality traits (e.g., introvert versus extrovert). For most languages available from the TwiSty corpus, results are found to outperform those in [ben Verhoeven et al., 2016].

The work in [Gjurković and Šnajder, 2018] introduces the MBTI9K corpus, a large

collection of Reddit posts labelled with MBTI information. The corpus conveys 354.996 posts written by 9,872 users in the English language, and it is also heavily imbalanced across MBTI classes. The use of the data is illustrated by a number of experiments involving logistic regression, SVM and multi-layer perceptron (MLP) classifiers using a range of alternative text features (e.g., word and character n-grams, psycholinguistics-motivated features etc.) Results show that MLP classifiers using the entire feature set generally obtains best results.

The work in [Keh and Cheng, 2019] is among the first to use pre-trained language models for MBTI personality classification from text, and also for personality-dependent language generation. To this end, a pre-trained BERT [Devlin et al., 2019] model is fine-tuned to classify texts taken from a purpose-built dataset of online discussions about personality. The authors suggest that the BERT model presents accuracy above 70% in the task, and point out that this is considerably superior to the results observed in other domains such as the MBTI9k Reddit corpus [Gjurković and Šnajder, 2018]. However, the analysis does not present any baseline results obtained from the same corpus, so it remains unclear whether the model is indeed superior to existing work, or whether personality classification from personality-related texts (e.g., in which users presumably discuss their personality traits, personality test results etc.) may be simply more straightforward than performing the same task based on more general social media text.

The work in [Katiyar et al., 2020] investigates a practical application of MBTI classification by focusing on social media data (e.g., blogs, Twitter, and Stack Overflow) as a means to recruit project teams. To this end, a model based on Naive Bayes classification and TF-IDF counts is evaluated using a set of 40 Twitter and Stack Overflow users, whose results suggest that it may be possible to infer both personality traits and technical skills from text to facilitate recruitment.

The work in [Wu et al., 2020] addresses the issue of author-dependent word embeddings for author profiling classification by introducing a model called *Author2Vec*. The possible use of this formalism is illustrated by discussing two downstream applications, namely, depression detection and MBTI personality classification from text. The latter makes use of a logistic regression classifier built from a subset of the MBTI9k corpus [Gjurković and Šnajder, 2018] conveying about half of the original corpus data. The model is found to outperform a number of alternative regression models based on static Word2vec embeddings [Mikolov et al., 2013], TF-IDF counts, and LDA topic modelling [Blei et al., 2003].

Finally, a number of recent studies in MBTI personality classification have made use of a social media corpus available from Kaggle¹ whose details regarding domain and data collection methods remain scarce. These include the work in [Das and Prajapati, 2020], which compares boosting, bagging, and stacking ensemble methods using concatenated TF-IDF counts and word embeddings; the work in [Abidin et al., 2020], which uses random forest and features based on text statistics (e.g., sentence length, punctuation etc.); and the studies in [Khan et al., 2020] and [Amirhosseini and Kazemian, 2020], both of which using XGBoost ensemble learning [Chen and Guestrin, 2016] over word counts.

¹ <https://www.kaggle.com/datasnaek/mbti-type>

2.2 Summary

Existing work in MBTI personality classification from text based on pre-trained language models such as BERT remain few. The two main exceptions are the study in [Keh and Cheng, 2019], which does not provide sufficiently complete evaluation details for further analysis, and the work in [Wu et al., 2020], which is mainly focused on a novel author-oriented word embedding formalism, and which uses only a small subset of the MBTI9k corpus in [Gjurković and Šnajder, 2018] as a working example of how models of this kind may be built. This motivates a more comprehensive investigation of the present task along these lines, and taking into account languages other than English.

3 Materials and methods

We envisaged a series of experiments in supervised machine learning from text to compare standard text classification models - based on bag-of-words and static word embeddings alike - with those built by fine-tuning a pre-trained BERT language model to perform MBTI classification. In doing so, we would like to show that the BERT-based approach outperforms the alternatives by a large margin.

Our experiments follow a 3-steps supervised machine learning pipeline that relies on training data (i.e., text documents) labelled with MBTI personality information to (1) build a text classifier model, (2) use the model to predict the class of previously unseen test data, and to produce (3) the corresponding output labels. This procedure is illustrated in Figure 1.

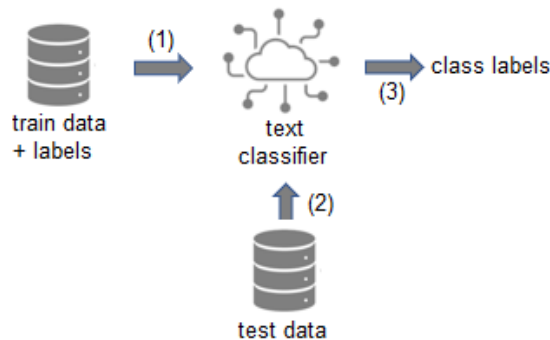


Figure 1: Experiment pipeline.

As a means to reduce the risk of overfitting, the experiments will be carried out in a 10-fold cross-validation setting, that is, each individual text classifier is built 10 times while varying the slices of data taken as train and test sets and, at the end of the evaluation, we report mean results over the 10 execution.

The models under discussion are to be evaluate using Reddit and Twitter text data in multiple languages. In all cases, MBTI personality detection will be modelled as a set of four independent binary classification tasks² corresponding to the four MBTI personality

² For instance, the EI class will be assigned the zero value when the E trait is prevalent, or the one value otherwise.

type indicators [Myers, 1962]:

1. EI: Extraversion (E) versus Introversion (I);
2. NS: Intuition (N) versus Sensing (S);
3. TF: Thinking (T) versus Feeling (F);
4. PJ: Perceiving (P) versus Judging (J).

The following sections describe the models developed for MBTI personality classification, the corpora to be taken as train/test data, and further details regarding the pre-processing and training procedures.

3.1 Models

MBTI personality classification from text will be assessed by comparing three alternatives³: BERT pre-trained language models, long short-term memory networks (LSTM) using static word embeddings, and a logistic regression bag-of-words baseline. These are discussed in turn as follows.

The main focus of the present work is the use of BERT [Devlin et al., 2019] pre-trained language models, which are presently fine-tuned for the MBTI personality classification task. To this end, we compute context-dependent DistilBert [Sanh et al., 2019] embeddings, and then feed 32-token input sequences to a network conveying a 512-neuron dense layer. This is followed by a 50% dropout layer, and by a 2-neuron dense layer that produces the binary classification result.

As an alternative to the use of pre-trained language models, we also consider the use of a sequence classifier based on a static word embeddings representation using LSTMs. Methods of this kind have been shown to obtain encouraging results in a wide range of NLP tasks, from stance and sentiment analysis [Zhang and v Wang, 2018, dos Santos and Paraboni, 2019, Pavan et al., 2020] to author profiling [Silva and Paraboni, 2018b, Ashraf et al., 2020, Escobar-Grisales et al., 2021] and others. More specifically, we compute Word2vec [Mikolov et al., 2013] skip-gram word embeddings from each corpus⁴ using a standard 300-dimension size and 8-word window. This representation is fed into a fixed LSTM architecture (i.e., with no further fine-tuning due to computational efficiency issues) comprising a 15-neuron attention layer, two LSTM layers containing 15 neurons each, and a 20% dropout layer. This is followed by a 64-neuron dense layer, and a 2-neuron softmax output layer.

Finally, we also consider a standard bag-of-words text classifier based on logistic regression over a word n-grams text representation. This approach, hereby called *Reg.word*, makes use of TF-IDF n-gram counts. Each of these input representations was subject to univariate feature selection, and the k optimal features were searched in a development dataset within the 30000-1000 range at -1000 intervals using F1 as a score function. Other logistic regression parameters were kept constant by using L2 penalty, lbfgs optimisation, balanced class weights, and 10^{-4} tolerance.

³ Code available from <https://github.com/vitorsantos95/mbti-classifier>

⁴ For this purpose, the corpus data is taken simply as a collection of word strings, that is, without access to any (MBTI) class information.

3.2 Data

The models described in the previous section are to be evaluated in multiple languages, namely, English, German, Italian, Dutch, French, Portuguese, and Spanish. In the case of English, we use the MBTI9k corpus of Reddit posts [Gjurković and Šnajder, 2018], and for the other languages we use the TwiSty corpus in the Twitter domain [ben Verhoeven et al., 2016]. Both MBTI9k and TwiSty texts are labelled with the four MBTI personality indicators, whose class distribution is summarised in Table 2. We notice however that both datasets are slightly smaller than those originally reported in [Gjurković and Šnajder, 2018] and [ben Verhoeven et al., 2016] since some of the data are no longer available online, or were removed due to noise. This issue is more prevalent in the Twitter domain in general, but it has also been raised in the context of the present Reddit dataset in [Wu et al., 2020].

Lang.	E	I	N	S	T	F	P	J
En	1,423	5,053	5,625	851	4,168	2,308	3,759	2,717
De	92,452	180,252	227,409	45,295	113,414	159,290	170,232	102,472
It	26,445	59,048	69,161	16,332	44,157	41,336	36,012	49,481
Du	70,904	33,589	76,987	27,506	35,791	68,702	66,447	38,046
Fr	249,742	481,480	566,473	164,749	297,702	433,520	451,187	280,035
Pt	27,920	32,387	44,919	15,388	27,160	33,147	32,536	27,771
Sp	243,840	277,788	334,483	187,145	199,573	322,055	302,092	219,536

Table 2: Corpus class distribution across English (En), German (De), Italian (it), Dutch (Du), French (Fr), Portuguese (Pt), and Spanish (Sp) subsets.

3.3 Procedure

The data from each corpus was subject to a 30/70 development/test split. Development sets were taken as an input to compute hyper-parameters for each model, and then discarded. Validation proper was performed using 10-fold cross validation over the previously unseen test sets.

All texts were subject to the removal of special characters (in particular, emoticons). In a pilot experiment, we also found out that stop words did not generally improve results for the task at hand and, accordingly, these were removed using NLTK [Bird, 2006] for the sake of efficiency. Other than that, all input texts were left unchanged.

From the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) obtained by each model, we computed precision, recall, F1 and accuracy scores as follows [Powers, 2011].

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1} = \frac{2*TP}{2*TP+FP+FN}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

4 Results

Table 3 summarises results obtained by the models discussed in the previous sections, and also from a majority class baseline. All results were obtained by performing 10-fold cross-validation. For brevity, in what follows we only present the mean F1 scores obtained by each model. For the full results (i.e., precision, recall, F1 and accuracy scores) we report to Table 7 at the end of this article.

Task	Model	En	De	Sp	Fr	It	Du	Pt
EI	Majority	0.43	0.40	0.35	0.40	0.41	0.40	0.35
	Reg.char	0.51	0.60	0.58	0.59	0.65	0.61	0.65
	Reg.word	0.54	0.60	0.58	0.59	0.63	0.62	0.64
	LSTM	0.83	0.73	0.71	0.72	0.80	0.82	0.80
	BERT	0.94	0.90	0.86	0.89	0.95	0.88	0.93
NS	Majority	0.46	0.45	0.39	0.44	0.45	0.42	0.43
	Reg.char	0.51	0.58	0.58	0.56	0.63	0.60	0.62
	Reg.word	0.54	0.57	0.58	0.56	0.64	0.63	0.61
	LSTM	0.82	0.75	0.68	0.74	0.79	0.82	0.79
	BERT	0.91	0.90	0.83	0.87	0.89	0.73	0.75
TF	Majority	0.39	0.37	0.38	0.37	0.34	0.40	0.35
	Reg.char	0.62	0.58	0.57	0.55	0.59	0.61	0.58
	Reg.word	0.65	0.58	0.57	0.56	0.59	0.61	0.58
	LSTM	0.82	0.73	0.69	0.72	0.78	0.81	0.81
	BERT	0.89	0.91	0.89	0.88	0.93	0.95	0.96
PJ	Majority	0.36	0.38	0.37	0.38	0.37	0.39	0.35
	Reg.char	0.57	0.57	0.58	0.56	0.62	0.59	0.57
	Reg.word	0.60	0.58	0.57	0.56	0.61	0.60	0.58
	LSTM	0.82	0.72	0.69	0.70	0.78	0.80	0.79
	BERT	0.91	0.86	0.83	0.74	0.93	0.91	0.94

Table 3: 10-fold cross-validation mean F1 results for English (En), German (De), Spanish (Sp), French (Fr), Italian (It), Dutch (Du), and Portuguese (Pt) data. Best F1 results for each class and language are highlighted.

Results from Table 3 show that BERT generally outperforms the alternatives in all but two cases - the NS task in Dutch (Du) and Portuguese (Pt) - in which case the LSTM model obtained overall best results. According to a McNemar test [McNemar, 1947], all differences between BERT and LSTM are statistically significant at $p < 0.001$ except for the NS task in Italian (It), in which case the difference is significant at $p < 0.005$ only. This outcome offers support to our main research question, that is, the use of pre-trained language models for MBTI personality classification outperforms standard text classification methods based on bag-of-words and static word embeddings alike.

As discussed in Section 3, the present experiments could not use exactly the same datasets as in previous work due to the removal of social media texts over time and, as a result, a direct comparison is not entirely possible. Bearing this limitation in mind, Table 4 presents - purely for illustration purposes - mean F1 results reported in [Gjurković and Šnajder, 2018] and [Wu et al., 2020] for the English MBTI9k corpus alongside those obtained by our present BERT model. Similarly, Table 5 presents weighted F1 results reported in both [ben Verhoeven et al., 2016] and [Alsadhan and Skillicorn, 2017] for the TwiSty corpus alongside present BERT.

Task	Gjurković & Šnajder	Wu et. al.	Current (BERT)
EI	0.83	0.69	0.94
NS	0.79	0.77	0.91
TF	0.64	0.68	0.89
PJ	0.74	0.61	0.91

Table 4: MBTI9k mean F1 results from previous work [Gjurković and Šnajder, 2018] and [Wu et al., 2020], and from the present BERT models.

Lang.	Task	Verhoeven	Alsadhan	Current (BERT)
De	EI	0.72	0.76	0.77
	NS	0.74	0.78	0.93
	TF	0.59	0.78	0.87
	PJ	0.62	0.80	0.92
Sp	EI	0.61	0.72	0.84
	NS	0.62	0.73	0.91
	TF	0.60	0.72	0.79
	PJ	0.56	0.69	0.88
Fr	EI	0.66	0.86	0.78
	NS	0.79	0.96	0.92
	TF	0.58	0.74	0.81
	PJ	0.57	0.84	0.86
It	EI	0.78	0.90	0.88
	NS	0.79	0.67	0.95
	TF	0.52	0.83	0.93
	PJ	0.47	0.79	0.91
Du	EI	0.63	0.85	0.94
	NS	0.70	0.94	0.97
	TF	0.60	0.82	0.91
	PJ	0.58	0.87	0.94
Pt	EI	0.67	0.85	0.92
	NS	0.73	0.94	0.93
	TF	0.62	0.80	0.94
	PJ	0.57	0.88	0.95

Table 5: TwiSty weighted F1 results from previous work [ben Verhoeven et al., 2016] and [Alsadhan and Skillicorn, 2017], and from the present BERT models for the German (De), Spanish (Sp), French (Fr), Italian (It), Dutch (Du), and Portuguese (Pt) languages.

As a means to illustrate the most relevant word features for each task, we performed eli5 prediction explanation⁵ to compute the terms more strongly correlated with each class, using as an example the word-based *Reg. word* classifier and the English MBTI9k dataset. This, despite being outperformed by our main BERT models, is more suitable to interpretation.

Selected features are illustrated in Table 6, in which word weights represent the change (decrease/increase) of the evaluation score when the specific feature is shuffled, keeping in mind that MBTI classes are not independent and, due to class imbalance, words that would intuitively be more associated with a particular MBTI type may have been selected by association with another, concomitant type (e.g., if users labelled as extraverts also happen to be mostly of the thinking type etc.)

⁵ <https://eli5.readthedocs.io/en/latest/>

Weight	EI	Weight	NS	Weight	TF	Weight	PJ
+0.113	job	+0.051	yourself	+0.035	thank	+0.051	comcast
+0.097	may	+0.047	after	+0.033	baby	+0.036	25b2
+0.090	never	+0.047	job	+0.030	amazing	+0.034	now
+0.086	them	+0.046	trans	+0.027	still	+0.033	nice
+0.084	free	+0.046	up	+0.025	two	+0.033	story
+0.081	im	+0.046	etc	+0.025	pregnancy	+0.032	awesome
+0.080	aren	+0.045	than	+0.025	actually	+0.032	always
+0.076	10	+0.044	all	+0.024	team	+0.029	others
+0.073	couple	+0.043	end	+0.024	from	+0.029	isn
+0.073	wiki	+0.041	know	+0.023	these	+0.029	game
...
-0.058	comcast	-0.030	honestly	-0.018	how	-0.022	let
-0.058	temple	-0.030	without	-0.018	fucking	-0.022	well
-0.059	death	-0.030	run	-0.018	ve	-0.022	nothing
-0.061	est	-0.030	help	-0.018	lol	-0.023	soylent
-0.062	city	-0.031	amazing	-0.018	point	-0.023	edit
-0.062	vs	-0.031	running	-0.019	something	-0.023	still
-0.062	week	-0.031	totally	-0.019	mind	-0.023	sex
-0.062	usually	-0.032	thing	-0.019	op	-0.023	completely
-0.062	own	-0.032	won	-0.020	lt	-0.023	lmao
-0.062	album	-0.032	which	-0.020	female	-0.024	clinton

Table 6: Top-10 positive and negative word weights for each classification task using the Reg.word logistic regression classifier.

Finally, the following text examples illustrate correctly classified instances for each MBTI type, in which the most relevant (word) features are highlighted.

In Figure 2 Extraversion correlates positively with 'team', and negatively with 'technology', whereas Introversion correlates positively with 'game' and negatively with 'dancing'.

that's because it's done as a **team**, **and** there **are** others higher up than **him**. do **we** **really** think that a **technology** that threatens the world's largest and **most** **blood-thirsty** banking cartels can't fund these activities? yes.

i personally want to see **santa** **hat** **baron** if **you** have 2 or more **christmas** skins in a **game**. **same** in my **game** except it's the **dancing** plague. i keep getting **those** global events every 6 months.

Figure 2: Extraversion (top) and introversion (bottom) features.

In Figure 3, Intuition correlates positively with 'people' and negatively with 'because' (which suggests reasoning), whereas Sensing strongly correlates with 'because' and more concrete concepts (e.g., cars, school, kids etc.)

it has a big biohazard symbol on the door to scare people away and all that jazz. there's even a badge scanner next to the door; it's just never on; it's just security theater. ironic because the break room is locked.

/rant i avoid driving near dunkin donuts on my way to work, because cars will *block the road* waiting in the drive-thru. sigh have kids, they say. they're a blessing, they say. then they just can't wait for the little snowflakes to go back to school...lol.

Figure 3: Intuition (top) and Sensing (bottom) features.

In Figure 4, Thinking correlates positively with concrete concepts (e.g., model, engineer) and negatively with sentiment-charged words (e.g., 'like', 'love'.) By contrast, Feeling correlates positively with sentiment, and negatively with 'would' (which might suggest reasoning.)

i love my spacemouse. it's like flying through the model once you get used to it. working full time as a design engineer on cad;
so excited, you're making smite more professional by the minute, would love to see more stuff like this.

Figure 4: Thinking (top) and Feeling (bottom) features.

Finally, in Figure 5, Perceiving correlates with a certain preference to take in information (e.g., 'inform'), and Judging correlates positively with decision-making terms (e.g., 'work').

you could also inform the brotherhood once you get the quest mass fusion and do spoils of war, which will get you banished.
hahaha that entire show is gold will forever be my favorite series. maybe tied with got.
e honestly... no one. i love the company i work for, even if i am just a lowly associate.

Figure 5: Perceiving (top) and Judging (bottom) features.

These examples, although suboptimal for the reasons discussed above, in our view suggest a reasonable consistency with the MBTI guidelines. Examples in which the logistic regression classifier does not make the right decision, by contrast, include, the selection of 'party' as a prominent feature for Extraversion even when the term is used in its political (and not leisure) sense. In the case of our BERT model, however, errors of this kind are arguably less likely to occur given the model's context sensitivity.

5 Final remarks

This work has addressed the issue of MBTI personality classification from text with the aid of pre-trained BERT language models. The present approach has been compared

against alternatives based on bag-of-words and static word embeddings representations, and its results were found to be consistently superior in a number of evaluation scenarios involving multiple target languages in the Reddit and Twitter domains, and by a significant margin.

Despite the positive initial results, however, the current set of experiments is only a first step towards more general, domain-independent MBTI personality classification from text. One obvious limitation of the present approach is, for instance, the focus on only two MBTI language resources (namely, the MBTI9k [Gjurković and Šnajder, 2018] and TwiSty [ben Verhoeven et al., 2016] corpora.) Although covering seven languages in two linguistic domains, more work needs to be done to investigate the present task in other text genres.

Furthermore, we notice that the present discussion has been limited to the issue of fine-tuning BERT models for the MBTI classification task for which training data is readily available. Outside the present Twitter and Reddit domains, however, text corpora labelled with MBTI information may be scarce, and it may be necessary to resort to domain adaptation methods. These may include, for instance, the use of BERT-based adversarial adaptation with distillation (AAD) method proposed in [Ryu and Lee, 2020] for cross-domain sentiment analysis, among others. A study along these lines in the context of the present personality classification task is also left as future work.

Finally, we notice that in recent years there has been a surge in transformer-based language models, including ELMo [Peters et al., 2017], XLNet [Yang et al., 2019], RoBERTa [Liu et al., 2019], GPT-3 [Brown et al., 2020], and many others. In most cases, these models are yet to be applied to the present MBTI classification task.

Acknowledgements

This work has received support from the University of São Paulo.

Task Model	English			German			Spanish			French			Italian			Dutch			Portuguese					
	Acc	P	F1	Acc	P	F1	Acc	P	F1	Acc	P	F1	Acc	P	F1	Acc	P	F1	Acc	P	F1			
MajORITY	0.77	0.77	0.77	0.43	0.66	0.66	0.66	0.40	0.53	0.53	0.53	0.35	0.65	0.65	0.65	0.40	0.69	0.69	0.69	0.41	0.67	0.67	0.67	0.40
	0.56	0.55	0.26	0.51	0.63	0.49	0.46	0.60	0.59	0.57	0.56	0.58	0.63	0.46	0.46	0.59	0.70	0.42	0.52	0.65	0.68	0.80	0.75	0.61
	0.54	0.58	0.26	0.54	0.63	0.52	0.46	0.60	0.58	0.61	0.54	0.58	0.64	0.37	0.46	0.59	0.71	0.47	0.55	0.63	0.68	0.82	0.73	0.62
EI	0.99	0.98	0.98	0.83	0.80	0.56	0.79	0.73	0.72	0.66	0.72	0.71	0.96	0.48	0.80	0.72	0.88	0.72	0.87	0.80	0.88	0.96	0.88	0.82
	0.87	0.67	0.69	0.94	0.86	0.69	0.86	0.90	0.85	0.82	0.85	0.86	0.86	0.72	0.84	0.89	0.93	0.84	0.93	0.95	0.92	0.96	0.93	0.88
	0.86	0.86	0.86	0.46	0.83	0.83	0.83	0.45	0.64	0.64	0.64	0.39	0.77	0.77	0.77	0.44	0.80	0.80	0.80	0.45	0.73	0.73	0.73	0.42
NS	0.63	0.64	0.90	0.51	0.75	0.84	0.86	0.58	0.61	0.66	0.71	0.58	0.74	0.89	0.79	0.56	0.81	0.93	0.85	0.63	0.71	0.80	0.80	0.60
	0.66	0.70	0.89	0.54	0.78	0.89	0.85	0.57	0.61	0.69	0.69	0.58	0.73	0.87	0.80	0.56	0.76	0.84	0.86	0.64	0.70	0.81	0.79	0.63
	0.99	0.99	0.99	0.82	0.85	0.90	0.85	0.75	0.76	0.93	0.75	0.68	0.85	0.97	0.85	0.74	0.92	0.96	0.93	0.79	0.90	0.95	0.91	0.82
TF	0.95	0.97	0.95	0.91	0.92	0.94	0.93	0.90	0.88	0.93	0.89	0.83	0.90	0.93	0.90	0.87	0.93	0.98	0.92	0.89	0.96	0.96	0.98	0.73
	0.64	0.64	0.64	0.39	0.58	0.66	0.66	0.37	0.61	0.53	0.53	0.38	0.59	0.59	0.59	0.37	0.51	0.51	0.51	0.34	0.65	0.65	0.65	0.40
	0.64	0.66	0.75	0.62	0.59	0.61	0.50	0.58	0.59	0.49	0.47	0.57	0.59	0.41	0.50	0.55	0.60	0.73	0.59	0.59	0.65	0.49	0.49	0.61
PJ	0.65	0.65	0.76	0.65	0.58	0.58	0.50	0.58	0.60	0.45	0.47	0.57	0.58	0.38	0.48	0.56	0.61	0.71	0.60	0.59	0.66	0.48	0.50	0.61
	0.99	0.99	0.99	0.82	0.77	0.63	0.77	0.73	0.75	0.47	0.80	0.69	0.79	0.51	0.97	0.72	0.84	0.85	0.84	0.78	0.87	0.76	0.85	0.81
	0.93	0.95	0.94	0.89	0.89	0.86	0.88	0.91	0.85	0.75	0.83	0.89	0.85	0.78	0.85	0.88	0.93	0.94	0.93	0.93	0.94	0.89	0.92	0.95
BERT	0.58	0.58	0.58	0.36	0.62	0.62	0.62	0.38	0.57	0.57	0.57	0.37	0.61	0.61	0.61	0.38	0.57	0.57	0.57	0.37	0.63	0.63	0.63	0.39
	0.57	0.54	0.65	0.57	0.60	0.65	0.69	0.57	0.59	0.66	0.64	0.58	0.61	0.76	0.66	0.56	0.62	0.57	0.55	0.62	0.65	0.80	0.69	0.59
	0.59	0.61	0.66	0.60	0.61	0.73	0.67	0.58	0.59	0.65	0.64	0.57	0.60	0.74	0.65	0.56	0.63	0.49	0.58	0.61	0.64	0.80	0.69	0.60
BERT	0.99	0.99	0.99	0.82	0.77	0.92	0.76	0.73	0.73	0.90	0.71	0.69	0.76	0.93	0.74	0.72	0.84	0.78	0.84	0.78	0.86	0.93	0.86	0.81
	0.91	0.92	0.93	0.91	0.90	0.93	0.90	0.86	0.86	0.90	0.86	0.83	0.82	0.91	0.82	0.74	0.93	0.91	0.92	0.93	0.93	0.95	0.94	0.91
	0.91	0.92	0.93	0.91	0.90	0.93	0.90	0.86	0.86	0.90	0.86	0.83	0.82	0.91	0.82	0.74	0.93	0.91	0.92	0.93	0.93	0.95	0.94	0.94

Table 7: 10-fold cross-validation mean accuracy (Acc), precision (P), recall (R) and F1 results. Best F1 results for each class and language are highlighted.

References

- [Abidin et al., 2020] Abidin, N. H. Z., Remli, M. A., Ali, N. M., Phon, D. N. E., Yusoff, N., Adli, H. K., and , A. H. B. (2020). “Improving intelligent personality prediction using myers-briggs type indicator and random forest classifier,” *International Journal of Advanced Computer Science and Applications*, 11(11).
- [Allport and Allport, 1921] Allport, F. H. and Allport, G. W. (1921). “Personality traits: Their classification and measurement,” *Journal of Abnormal And Social Psychology*, 16:6–40.
- [Alsadhan and Skillicorn, 2017] Alsadhan, N. and Skillicorn, D. (2017). “Estimating personality from social media posts,” in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 350–356.
- [Amirhosseini and Kazemian, 2020] Amirhosseini, M. H. and Kazemian, H. (2020). “Machine learning approach to personality type prediction based on the Myers-Briggs type indicator,” *Multimodal Technologies and Interaction*, 4(1).
- [Ashraf et al., 2020] Ashraf, M. A., Nawab, R. M. A., and Nie, F. (2020). “A study of deep learning methods for same-genre and cross-genre author profiling,” *Journal of Intelligent & Fuzzy Systems*, 39:2353–2363.
- [Barlas and Stamatatos, 2020] Barlas, G. and Stamatatos, E. (2020). “Cross-domain authorship attribution using pre-trained language models,” in Maglogiannis, I., Iliadis, L., and Pimenidis, E., editors, *Artificial Intelligence Applications and Innovations*, pages 255–266, Cham. Springer International Publishing.
- [ben Verhoeven et al., 2016] ben Verhoeven, Daelemans, W., and Plank, B. (2016). “TwiSty: A multilingual twitter stylometry corpus for gender and personality profiling,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1632–1637, Portorož, Slovenia. European Language Resources Association (ELRA).
- [Bird, 2006] Bird, S. (2006). “NLTK: the natural language toolkit,” in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.
- [Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, 3(4-5):993–1022.
- [Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). “Language models are few-shot learners,” *CoRR*, abs/2005.14165.
- [Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). “XGBoost: A Scalable Tree Boosting System,” in *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16*, pages 785–794, New York, NY, USA. Association for Computing Machinery.
- [Coltheart, 1981] Coltheart, M. (1981). “The MRC psycholinguistic database,” *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 33(4):497–505.
- [Das and Prajapati, 2020] Das, K. and Prajapati, H. (2020). “Personality identification based on MBTI dimensions using natural language processing,” *International Journal of Creative research Thoughts*, 8(6):1653–1657.
- [Delmondes Neto and Paraboni, 2021] Delmondes Neto, J. P. and Paraboni, I. (2021). “Multi-source BERT stack ensemble for cross-domain author profiling,” *Expert Systems*, e12869.

- [Devlin et al., 2019] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). “BERT: pre-training of deep bidirectional transformers for language understanding,” in Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.
- [dos Santos et al., 2017] dos Santos, V. G., Paraboni, I., and Silva, B. B. C. (2017). “Big five personality recognition from multiple text genres,” in *Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence* vol. 10415, pages 29–37, Prague, Czech Republic. Springer-Verlag.
- [dos Santos et al., 2020] dos Santos, W. R., Funabashi, A. M. M., and Paraboni, I. (2020). “Searching Brazilian Twitter for signs of mental health issues,” in *12th International Conference on Language Resources and Evaluation (LREC-2020)*, pages 6113–6119, Marseille, France. ELRA.
- [dos Santos and Paraboni, 2019] dos Santos, W. R. and Paraboni, I. (2019). “Moral Stance Recognition and Polarity Classification from Twitter and Elicited Text,” in *Recent Advances in Natural Language Processing (RANLP-2019)*, pages 1069–1075, Varna, Bulgaria. INCOMA Ltd.
- [dos Santos et al., 2019] dos Santos, W. R., Ramos, R. M. S., and Paraboni, I. (2019). “Computational personality recognition from facebook text: psycholinguistic features, words and facets,” *New Review of Hypermedia and Multimedia*, 25(4):268–287.
- [Escobar-Grisales et al., 2021] Escobar-Grisales, D., Vásquez-Correa, J. C., and Orozco-Arroyave, J. R. (2021). “Gender recognition in informal and formal language scenarios via transfer learning,” *CoRR*, abs/2107.02759.
- [Gjurković and Šnajder, 2018] Gjurković, M. and Šnajder, J. (2018). “Reddit: A gold mine for personality prediction,” in *Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*, pages 87–97, New Orleans, USA. Association for Computational Linguistics.
- [Goldberg, 1990] Goldberg, L. R. (1990). “An alternative description of personality: The Big-Five factor structure,” *Journal of Personality and Social Psychology*, 59:1216–1229.
- [Hoang et al., 2019] Hoang, M., Bihorac, O. A., and Rouces, J. (2019). “Aspect-based sentiment analysis using BERT,” in *22nd Nordic Conference on Computational Linguistics*, pages 187–196, Turku, Finland. Linköping University Electronic Press.
- [John et al., 1991] John, O. P., Donahue, E., and Kentle, R. (1991). “The Big Five inventory - versions 4a and 54,” Technical report, Inst. Personality Social Research, University of California, Berkeley, CA, USA.
- [Katiyar et al., 2020] Katiyar, S., Kumar, S., and Walia, H. (2020). “Personality prediction from stack overflow by using naive bayes theorem in data mining,” *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9.
- [Keh and Cheng, 2019] Keh, S. S. and Cheng, I. (2019). “Myers-Briggs personality classification and personality-specific language generation using pre-trained language models,” *CoRR*, abs/1907.06333.
- [Khan et al., 2020] Khan, A. S., Ahmad, H., Asghar, M. Z., Saddozai, F. K., Arif, A., and Khalid, H. A. (2020). “Personality classification from online text using machine learning approach,” *International Journal of Advanced Computer Science and Applications*, 11(3):460–476.
- [Liu et al., 2017] Liu, F., Perez, J., and Nowson, S. (2017). “A language-independent and compositional model for personality trait recognition from short texts,” in *15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 754–764, Valencia, Spain. Association for Computational Linguistics.

- [Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” CoRR, abs/1907.11692.
- [López-Santill et al., 2020] López-Santill et al., R. (2020). “Richer document embeddings for author profiling tasks based on a heuristic search,” *Information Processing & Management*, 57(4).
- [Lukito et al., 2016] Lukito, L. C., Erwin, A., Purnama, J., and Danoekeoesoemo, W. (2016). “Social media user personality classification using computational linguistic,” in 8th International Conference on Information Technology and Electrical Engineering (ICITEE), pages 1–6.
- [McNemar, 1947] McNemar, Q. (1947). “Note on the sampling error of the difference between correlated proportions or percentages,” *Psychometrika*, 12(2):153–157.
- [Mikolov et al., 2013] Mikolov, T., Wen-tau, S., and Zweig, G. (2013). “Linguistic regularities in continuous space word representations,” in Proc. of NAACL-HLT-2013, pages 746–751, Atlanta, USA. Association for Computational Linguistics.
- [Myers, 1962] Myers, I. B. (1962). “The Myers-Briggs type indicator,” Consulting Psychologists Press.
- [Pavan et al., 2020] Pavan, M. C., dos Santos, W. R., and Paraboni, I. (2020). “Twitter Moral Stance Classification using Long Short-Term Memory Networks,” in 9th Brazilian Conference on Intelligent Systems (BRACIS). LNAI 12319, pages 636–647. Springer.
- [Pennebaker et al., 2001] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). “Inquiry and Word Count: LIWC,” Lawrence Erlbaum, Mahwah, NJ.
- [Peters et al., 2017] Peters, M. E., Ammar, W., Bhagavatula, C., and Power, R. (2017). “Semi-supervised sequence tagging with bidirectional language models,” in Proc. of ACL-2017, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.
- [Plank and Hovy, 2015] Plank, B. and Hovy, D. (2015). “Personality traits on Twitter—or—How to get 1,500 personality tests in a week,” in 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 92–98, Lisbon. Association for Computational Linguistics.
- [Polignano et al., 2020] Polignano, M., de Gemmis, M., and Semeraro, G. (2020). “Contextualized BERT sentence embeddings for author profiling: The cost of performances,” in Computational Science and Its Applications (ICCSA)-2020, LNCS 12252, pages 135–149, Cham. Springer.
- [Powers, 2011] Powers, D. M. W. (2011). “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation,” *Journal of Machine Learning Technologies*, 2(1):37–63.
- [Price and Hodge, 2020] Price, S. and Hodge, A. (2020). “Celebrity profiling using twitter follower feeds,” in Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece. CLEF and CEUR-WS.org.
- [Rangel et al., 2020] Rangel, F., Rosso, P., Zaghouni, W., and Charfi, A. (2020). “Fine-grained analysis of language varieties and demographics,” *Natural Language Engineering*, page 1-21.
- [Ryu and Lee, 2020] Ryu, M. and Lee, K. (2020). “Knowledge distillation for BERT unsupervised domain adaptation,” CoRR, abs/2010.11478.
- [Sanh et al., 2019] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” arXiv preprint arXiv:1910.01108.
- [Silva and Paraboni, 2018a] Silva, B. B. C. and Paraboni, I. (2018a). “Learning personality traits from Facebook text,” *IEEE Latin America Transactions*, 16(4):1256–1262.
- [Silva and Paraboni, 2018b] Silva, B. B. C. and Paraboni, I. (2018b). “Personality recognition from Facebook text,” in 13th International Conference on the Computational Processing of Portuguese (PROPOR-2018) LNCS vol. 11122, pages 107–114, Canela. Springer-Verlag.

[Teixeira et al., 2014] Teixeira, C. V. M., Paraboni, I., da Silva, A. S. R., and Yamasaki, A. K. (2014). “Generating relational descriptions involving mutual disambiguation,” in *Computational Linguistics and Intelligent Text Processing (CICLing-2014)*, Lecture Notes in Computer Science 8403, pages 492–502, Kathmandu, Nepal. Springer.

[Wu et al., 2020] Wu, X., Lin, W., Wang, Z., and Rastorgueva, E. (2020). “Author2vec: A framework for generating user embedding,” arXiv preprint arXiv:2003.11627.

[Yang et al., 2019] Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. (2019). “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, volume 32, pages 5753–5763, Vancouver, Canada. Curran Associates, Inc.

[Zhang and v Wang, 2018] Zhang, L. and v Wang, a. B. L. (2018). “Deep learning for sentiment analysis: A survey,” *WIREs Data Mining and Knowledge Discovery*, 8(4):e1253.