Personality Trait Prediction Using Text Data from Social Media

Luís Guedes de Sousa[1] and João Silva Sequeira[1]

Instituto Superior Técnico, University of Lisbon, 1049-001 Lisbon, Portugal,
{luis.g.sousa, joao.silva.sequeira}@tecnico.ulisboa.pt

Abstract. Currently there is a plethora of personality models in use in Psychology. Arguably, the dominant model is the so called Big Five model, which identifies five traits that compose human personality. With the rise of computers in the past decades, automatic personality assessment models have been gaining popularity. Additionally, with the rise of social media, there is a large amount of data containing information on the personality of the users. Regarding text-based personality prediction, large pre-trained language models have been gaining popularity, achieving state-of-the-art results in multiple cases. In this work, two datasets containing textual data, from the Facebook and Twitter social media platforms, were used to predict the Big Five personality traits. The proposed model resorts to the large pre-trained language model, Sentence-BERT, to extract sentence embeddings and a neural network as a regression model. The results obtained with the Twitter dataset outperformed state-of-the-art by around 16-18%, though with the Facebook dataset underperformed when compared with the state-of-the-art. Additionally, a comparative analysis was performed of how data from different sources can be combined and applied to one another, in the scope of personality trait prediction.
Keywords: Personality Traits, Facebook, Twitter, Big Five, Machine Learning

1. Introduction

Personality psychology originated as a field within Psychology in the early 20th century. Multiple models were created to describe human personality, one of the most dominant among these being the Big Five model, [19].

Personality assessment can be useful in many contexts, such as self-improvement, job performance, and marketing. By modelling and describing people's personality, their behaviour and way of thinking can be explained and even predicted in some cases. Furthermore, with the rise in the use of ICT devices, e.g., smartphones, and social networks, the demand for automated personality prediction systems increased, as they can enhance these systems' interaction capabilities. Similarly, social robotics is also becoming ubiquitous and there is a clear need to endow the robots with social skills, for which being able to assess the personality of interacting people and behaving in a socially acceptable way is relevant for acceptance.

Traditional personality assessment methods, often involving psychological evaluations or formula-based tests, are time and resource-intensive, making them impractical for large populations or inaccessible individuals, such as social network users. Machine learning methods with their ability to predict personality indirectly, through the use of user data, be it textual, video or audio data, are becoming valid alternatives. Recently, large pre-trained language models have proved to achieve state-of-the-art results in various natural language processing tasks, though they have not been extensively studied for personality prediction. This study aims to evaluate these models' effectiveness in personality prediction compared to traditional machine learning approaches, through the use of social media textual data.

The paper is organized as follows. Section 2 provides background information on the Big Five model and useful machine learning concepts. Section 3 reviews relevant literature. Section 4 describes the methodology, and Section 5 Section 6 discusses the conclusions and future work.

2. Background

This section describes a few concepts necessary to understand the contents of this work, namely the Big Five model, natural language processing (NLP) and large pre-trained language models.

2.1. The Big Five Model

Trait theory is an area of Psychology that focuses on human personality through personality traits, which remain relatively consistent over time and across different situations, and that can vary in intensity. Personality assessment is a domain in this field which involves measuring these traits, usually done resorting to a professional's assessment or, in recent times, to automatic questionnaires.

Over the years, several personality trait models have been developed, with one of the most influential being the Big Five model, or Five-Factor model, [19]. This model describes human personality through

five dimensions, each related to various cognitive and behavioural patterns. The Big Five traits and the associated coarse measurements scales are:

- **Extroversion** – High - social, expressive, assertive; Low - reserved, low in energy, inexpressive
- **Neuroticism** – High - emotionally unstable, anxious, self-conscious; Low - emotionally stable, relaxed, confident
- **Agreeableness** – High - altruistic, empathetic, cooperative; Low - antagonistic, untrustworthy, insensitive
- **Conscientiousness** – High - organized, goal-oriented, thoughtful; Low - disorganised, impulsive, procrastinatory.
- **Openness** – High - curious, creative, willing to change; Low - close-minded, resist change, practical

This model describes traits on a spectrum, offering a continuous measurement of personality traits, making it well suited for automatic personality trait assessment.

2.2. Natural Language Processing Concepts

Natural Language Processing (NLP), is the field that relates computer science and linguistics, studying computational methods for analysing and extracting information from text. In the scope of this work, it is necessary to understand some NLP concepts, namely, tokenization, embeddings and transformers/self-attention.

Tokenization

Tokenization is usually necessary in order to feed text into a ML models. This is done in order to structure the text so the models can interpret them. Tokenization consists of splitting the text in parts and can be performed at a sentence level, word level or even character level, with the most popular approach to be at the word level. This discretization process allows for texts to be combined and transformed into array structures, which are well-suited for Machine Learning (ML) models.

Embeddings

Similarly to tokenization, the concept of embeddings is also related to the transformation of raw text into a format that can be understood by ML models (see for instance [2]). Textual embeddings consist of numerical representations of sequences of text, that represent the underlying meaning of the text through vectors of numbers. By converting raw text into an array-like structure of numbers encoding its meaning, these can be interpreted by ML models, even if not by humans.

A significant advancement in NLP was the shift from static to contextualized word embeddings. Static word embedding models assign the same vector to a word regardless of the context of the word in the text, which may result in loss of relevant information. These embeddings can be problematic in many contexts, such as for encoding the meaning of homonyms, e.g., “rock” as a stone vs. the music genre. Contextualized embeddings, however, are obtained in such a way, where the context of a word is considered, thereby better capturing its meaning based on its usage in the text. Tokenization and extraction of embeddings are commonly performed together, as the former is necessary for the discretization of text, while the latter converts each of the obtained parts into a numerical representation, yielding a structure that can be interpreted by ML models.

Transformers and Self-attention

The Transformer architecture, as it is currently used, was proposed in [30]. Transformers offer several advantages over previous models like RNNs, LSTMs, and GRUs, such as their ability to utilize parallel computing, which speeds up training, as well as the ability to model dependencies across long sequences of text. Some of the most successful and popular models in NLP are transformer-based, such as BERT, GPT-3, and XLNet. These models have achieved state-of-the-art results in various NLP tasks.

A key concept in the context of transformer models is the self-attention mechanism, which allows the model to weight the importance of different tokens in an input sequence relative to each other. It uses three vectors for each token, query (Q), key (K), and value (V), which are computed from the input embeddings, to compute attention scores. These scores indicate the significance of each token and are used to create a weighted sum of the input tokens, resulting in a contextualized representation that incorporates information from the entire sequence, for each token.

Figure 1 represents the attention mechanism in a blocks diagram. Vectors Q, K, V are combined to form an attention head. Multiple heads, capturing different dependencies, in parallel, can be concatenated and fed into a fully-connected layer a global attention score.

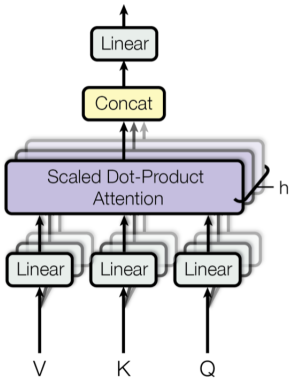

Fig. 1: Multi-head attention block, [30]

A transformer model consists of an encoder and a decoder (see the architecture in Figure 2). The encoder processes the input sequence to produce contextualized token representations, incorporating positional encodings to retain the order of tokens. This is done making use of self-attention mechanisms. The encoder uses these contextualized representations, as well as its own past outputs,to generate an output sequence. This process is performed with masked multi-head attention to prevent future information from influencing the predictions during training. The decoder's final output is produced through a feed-forward network and a softmax layer to generate probabilities for the next token in the sequence.

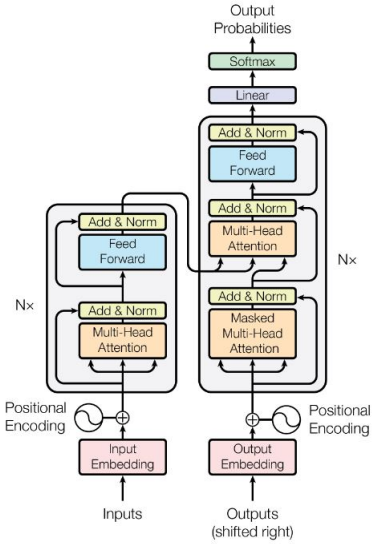

Fig. 2: Transformer architecture, [30]

2.3. Large Pre-trained Language Models

Large pre-trained Language Models (PTM) have been growing in popularity in the past few years in the NLP and ML communities, achieving state-of-the-art performance in many NLP tasks [4, 8, 22, 31]. Therefore, they are a natural choice to consider when faced with an NLP problem. PTM are language models trained on vast amounts of textual data, usually using deep learning. Their utility is based on

the concept of transfer learning, where a model is trained on one task and then adapted for another, leveraging the knowledge gained from the initial task.

PTM are thus first trained on general NLP tasks, using large amounts of unlabelled data, to gain a general understanding of natural language. The models can then leverage this knowledge, and can be fine-tuned to adapt to specific downstream tasks. There are three main approaches to implementing transfer learning in PTMs.

- **Pre-train then fine-tune:** The PTM is pre-trained on a large corpus and then fine-tuned on a labeled dataset specific to the downstream task. This method can achieve good performance without requiring the training of a model from scratch, but is computationally expensive and requires large task-specific datasets.
- **Using pre-trained contextualized embeddings:** In this approach, the PTM is used as is, to extract embeddings, which can be used as features for another model that makes predictions based on them. This approach is less computationally expensive, as it only requires a forward pass through the PTM, and doesn't require large task-specific datasets in order to obtain good performance in NLP tasks, [31].
- **Prompt Based Learning:** This method consists of adding natural language to the input text,and feeding it to the model as a prompt, thus, instructing it to perform the task. By introducing the textual prompt, the task becomes similar to those performed during pre-training, allowing for the knowledge previously acquired by the model to translate better into the new task. This method is computationally cheaper, requires no weight updates, and can perform well with limited labelled data.

Due to restrictions in computational power and the limited size of the available datasets, the approach chosen here uses pre-trained contextualized embeddings as features. The models chosen to extract these embeddings were the DistilBERT and Sentence-BERT model, further explained ahead.

BERT

In what concerns NLP, BERT (Bidirectional Encoder Representations from Transformers) is one of the most popular, achieving state-of-the-art results in many tasks, [8]. This model was trained using over 2500M words from Wikipedia and 800M from books. Its architecture (see Figure 3) consists of a multilayer transformer encoder stack, with 12 layers, each containing 12 self-attention heads It has a total of 110 million parameters and the feed-forward networks in the encoders have 768 neurons, resulting in an output of the same size.

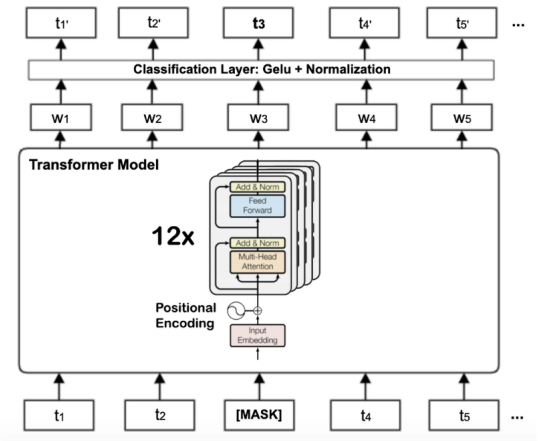

Fig. 3: BERT Base architecture, [8]

BERT is trained using two main language modelling methods: masked language modelling (MLM) and next sentence prediction (NSP). In MLM, some of the tokens that compose the input sequences are masked, that is, replaced with a generic mask token, and the model is then trained in order to predict the original masked tokens. This allows the model to learn linguistic patterns and context between words.

In the second method, NSP, the model is given pairs of sentences and must predict whether the second sentence logically follows the first. This task helps BERT understand sentence relationships.

In order to feed text into the BERT model, two tokens must be added: the [CLS] and [SEP] tokens, at the beginning and end of the sentence respectively. The embedding of the [CLS] token encodes the meaning of the whole sequence, [8].

Due to the large size of the BERT model, fine-tuning and extracting word embeddings using this model tends to be computationally expensive and of long duration. A lighter version of this model, DistilBERT, [28], which achieved 97% of its performance in popular NLP benchmarks, while being 40% smaller and 60% faster in comparison. Due to the limited computing resources available for this study, this model was selected.

Sentence-BERT

The Sentence-BERT model, or S-BERT, was introduced in [27] as a sentence embedding model based on the BERT. While it is possible to extract sentence embeddings using the original BERT model, the authors argue that this approach yields bad sentence embeddings, sometimes even performing worse than averaging GloVe word embeddings, [27], an older model. The S-BERT model was created as an improvement of the original model for sentence embeddings.

This model employs siamese networks, which consist of a pair of identical neural networks, as they share the same parameters and weights, and where parameter updating is mirrored across both sub-networks. The S-BERT model architecture consists of two siamese BERT models, which use pooling of word embeddings to obtain sentence embeddings. The model was fine-tuned to produce semantically meaningful sentence embeddings which can be compared for similarity. While the authors assert that the purpose of S-BERT sentence embeddings is not to be used for transfer learning, they found that using these embeddings as features, a new state-of-the-art in the SentEval toolkit was achieved, outperforming the other sentence embedding models.

3. Related Work

In this section, a brief overview of personality trait prediction will be performed, followed by an analysis of more traditional machine learning approaches to this task and by an overview of more recent deep learning-based approaches.

Techniques for personality trait prediction range from word counts, namely adjectives, [1, 7], to complex machine learning methods. This work focuses on machine learning approaches, which are central to the field and represent the state-of-the-art in personality prediction. Personality trait prediction can be performed based on data of various modalities, such as, text, visual data, audio, utterance behaviour, and others, either taken isolatedly or combined.

A comprehensive survey of machine learning based approaches to personality trait prediction, with an emphasis on deep learning approaches is presented in [21]. The authors claimed that the state-of-the-art in personality trait prediction is achieved using deep learning techniques with multimodal data. Additionally, they found that for unimodal data the best performance was also achieved using deep learning-based approaches.

The impact of personality in language use is addressed in [25], by analysing various texts, using LIWC, a text analysis computer program. Given a textual input, this program will calculate the score for several categories based on the percentage of words in the text that fall into that category. This program ended up becoming one of the most popular feature extraction tools for personality prediction from text, [21, 23]. The authors found correlations between many LIWC categories and the Big Five personality traits, which indicated that personality has an effect on language use. The dataset collected in this study (Essays) became one of the most popular for personality trait prediction.

3.1. Traditional Machine Learning approaches to text-based Personality Trait Prediction

One of the first approaches to automatic personality trait prediction was done in [17] using the Essays dataset and EAR data. Their approach resorted to LIWC (Linguistic Inquiry and Word Count) and MRC (Psycholinguistic database) for feature extraction in both datasets, as well as using other speech-related features for the EAR data, and used SVM and Naive Bayes to perform classification. Testing different combinations of the feature sets, they achieved average accuracies of 59.09% and 66.58% for the Essays and EAR datasets respectively.

The two studies in [12, 13], presented Big Five personality trait prediction, one with Twitter data and one with Facebook data. In the Twitter study, [12], the authors used LIWC, MRC and sentiment

features, as well as some structural features, with a ZeroR model (see, for instance, [9]). They found that they could predict personality scores within 11-18% of their ground truth values. In the Facebook study, resorting to LIWC and other statistical features and using the M5 Rules regression model (see, for instance, [26]), they found that they could, similarly, predict scores within 11% of their true values.

In [15] over 3000 bloggers were classified in the lower, middle or upper third of the the Big Five personality trait spectrums, using LIWC and bi-gram features with an SVM classifier, achieving an accuracy of 76.81%.

Binary classification of Facebook users using the MyPersonality dataset in [11]. They extracted LIWC features, network features and time-related features, and used different configurations of these feature sets to perform classification, msing the SVM, KNN and Naive Bayes models, achieving an average F1 score of 0.54, with the best performing approaches in each trait. Additionally, they found that combining this dataset with the Essays dataset for training the model, improved results. They also trained the model on both datasets separately and applied them to one another, finding comparable results than using only one dataset. The authors therefore claimed that personality prediction models benefit from combining data from different domains and generalise well across them.

Three separate datasets, the MyPersonality Facebook dataset, [5], the Youtube vlogs dataset and the PAN15 Twitter dataset, were used in [10] to predict Big Five traits. Extracting features from text (using LIWC, MRC psycholinguistic database, NRC emotion lexicon and SPLICE), as well as sentiment analysis features, they predicted the scores using multiple univariate and multivariate regression algorithms, with the best performing one being the ERCC algorithm. They achieved average RMSE results for the Facebook, Youtube and Twitter datasets were, respectively, 0.718, 0.731 and 0.179. Additionally, they performed cross-domain experiments, by comparing the performance of the model trained with combined and non-combined data and reached opposite conclusions in relation to their 2013 study, finding that the ability of a model to benefit for combining data of different domains and to generalise well across them, depends much on the similarity and quality of the datasets.

The Big Five personality trait scores of 171 Twitter users were predicted in [3] using a dataset containing multilingual data from the 2015 PAN Author Profiling competition (though only the English data was used). Using solely LIWC features with a ERCC regression model, with 10-fold CV, they achieved an average MAE of 0.1414.

3.2. Deep Learning approaches to text-based PTP

A comparison between traditional machine learning and deep learning approaches is presented in [29]. They extracted LIWC and SPLICE features, as well as network statistic features, and, while in the former approach, they used traditional classification models, such as SVM and Naive Bayes, in the deep learning approach, they used many types of neural networks. The deep learning approach outperformed the traditional ML approach consistently, with the best model configuration being a LSTM + CNN-1D network, which achieved an average accuracy of 74.17%.

Using a deep learning-based approach, [18] performed Big Five personality trait prediction on the Essays dataset [25], as a binary classification task. The features used in this study consisted of Mairesse document level features (as in [17]) and document-level embeddings, obtained by passing Word2Vec embeddings through a neural network with pooling layers. Using SVM and MLP as classification models, they achieved an average accuracy of 58.83% across the five traits.

In [14], the four Meyer-Briggs model traits were identified from the MBTI Kaggle dataset, performing binary classification of each of the four MBTI traits. The GloVe embeddings of the 2500 most common words in the Kaggle posts were used as features. Testing multiple neural network architecture for classification, they achieved an average accuracy of 67.78% using a LSTM.

The study in [20] compared the performance of traditional closed-vocabulary psycholinguistic features and open-vocabulary features in personality trait prediction, using two popular datasets: Essays, [25], and MBTI Kaggle. In the former approach, they used Mairesse document-level features, [17], and similar dictionary-based features. In the latter, they used pre-trained word embeddings, obtained using the BERT, RoBERTa and ALBERT pre-trained language models. Using MLP and SVM classifiers, they found that the approach resorting to pre-trained word embeddings, specifically with the BERT model, yielded the best results, with average accuracies of 60.6% and 77.1%, for the Essays and MBTI Kaggle datasets respectively.

In [16], Big Five personality trait prediction was performed as a regression, using the MyPersonality dataset, [5]. They used the BERT model, as well as a multilingual version of the model, to extract sentence embeddings, by using the word embedding relative to the [CLS] token. As classification model,

they utilised a neural network with a single fully connected layer of size 300, obtaining an average MSE of 0.3026, which outperformed previous state-of-the-art results by around 30%.

An innovative approach to Big Five personality trait prediction, combining multiple pre-trained language models (specifically the BERT, RoBERTa and XLNet models) into a single architecture was presented in [6]. They performed this task using two popular datasets, namely the MyPersonality dataset [5] and Bahasa Twitter data. They fed each of these model embeddings into a separate self-attention attention head, followed by a fully-connected layer and performed model averaging of the output probabilities. Average accuracies of 76.75% and 76.98%, were obtained for the Facebook and Twitter datasets, respectively, marking new state-of-the-art results.

4. Methodology

This section describes the architecture and datasets used (see Figure 4).

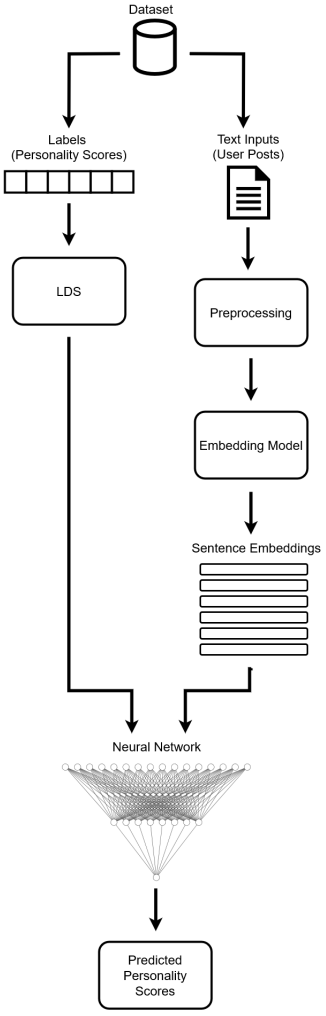

Fig. 4: Proposed model architecture for Personality Trait Prediction.

4.1. Datasets

Two datasets were selected to be used in this study, namely (i) the MyPersonality Facebook dataset and (ii) the PAN15 Twitter dataset. The Facebook dataset, available from [5], consists of Facebook data collected through MyPersonality, which was a popular Facebook application that allowed users to take real psychometric tests and get feedback on them. This dataset contains 9917 textual Facebook statuses collected from 250 users, as well as personality scores scores from the users (on a 1-5 scale) and other parameters containing more information about the post, such as its date.

The Twitter dataset was released as part of the 2015 Author Profiling competition, [24], organized by the PAN organization, where participants used the available data to predict the users' demographics from their writing. It contained Twitter data from English, Spanish, Italian and Dutch Twitter users, each split into two sets of training and test data, which consisted of textual posts (or tweets), Big Five personality scores (ranging from -0.5 to 0.5) and information regarding their age and gender. In this study, only the data relative to the English speaking users was used, which contained over 27344 tweets in total from 294 users.

A statistical analysis of the score distributions of both datasets was performed, and it was found that both datasets were imbalanced, containing less samples of low scoring users in most traits. Additionally, This imbalance was especially significant in the Openness trait of both datasets. The Twitter dataset also proved to have a more pronounced imbalance in most traits.

4.2. Prediction Scenarios

Personality trait prediction can be performed using data from a single domain or multiple domains, e.g., from different social media platforms. It is relevant to compare these two scenarios, as this choice can affect the model's generalization ability and its overall performance. Two studies analysing these two scenarios reached conflicting conclusions: one study, [11], found that using data from multiple domains improved performance and generalization, while another, [10], found little to no improvement due to contextual differences.

In this study, two types of experiments were conducted to compare these scenarios. In the first scenario, models were trained and tested on the same data, from a single domain (Facebook or Twitter). In the second scenario, data from both domains was used, and two types of experiments were performed, (i) using data from a single dataset to train the model and testing in the other one and (ii) using data from both datasets to train the model and testing it in each of the two datasets separately. These experiments aimed to analyse the model's ability to generalize across domains and whether combining data from different domains results in improved performance compared to using data from a single domain.

4.3. Preprocessing

Two preprocessing scenarios were tested and compared, one more preprocessing heavy approach, and a lighter one. The lighter preprocessing approach consisted of: lowercasing, URL removal, emoji removal, and the replacement of mentions and hashtags with the custom tokens [USER] and [HASHTAG], respectively. The second scenario, performed these operations, as well as punctuation removal, digit removal, contraction expansion, e.g., can't -¿ can not), and lemmatization. The lighter preprocessing approach outperformed the heavier one, so it was used in all subsequent experiments.

Additionally, two other preprocessing operations were performed: the conversion of both datasets' labels to a standardized scale of [0, 1] and the removal of users with fewer than 10 posts or fewer than 100 tokens.

4.4. Feature Extraction

As referred in Section 3, the state-of-the-art in personality trait prediction from text is dominated by approaches incorporating large pre-trained language models, such as BERT or XLNet, either as feature extraction tools or as actual classification/regression models. Due to their large size and extensive training data, these models have a solid knowledge of language, achieving good performance in many NLP tasks without any fine tuning.

A determining factor in the choice of the embedding model to use in this work, was the fact that computing resources were limited. Since word embedding models yield embeddings of large size for each token (768 in the case of the BERT models), this can result in significantly large feature sets which cause computationally expensive and long training steps. For this reason, the choice to use sentence embeddings was made, using the DistilBERT and Sentence-BERT models.

The extraction of the DistilBERT sentence embeddings was implemented using the HuggingFace Python library *transformers*, with the *distilbert-base-uncased* model version. This process started with

a text preparation step, where the [CLS] and [SEP] tokens were added, tokenization of the text was performed, together with padding/truncation to the length of 50. Following this, the processed text was fed to the DistilBERT model, where a forward pass was performed, and the average of the output of the last four layers was extracted, thus yielding word embeddings of the text. Finally, the word embedding of the [CLS] token was extracted and used as a sentence embedding, as explained in Section 3.

In what concerns the SentenceBERT embeddings, the process was simpler, as the process of text preparation and embedding extraction was performed automatically by the model, with a single function (*Sentence-Transformer.encode* function of the *sentence_transformers* Python library). This model yielded sentence embeddings of size 384, while the DistilBERT model yielded sentence embeddings of size 768.

The embeddings of each users posts were extracted yielding an array of shape *(number of posts, embedding size)*, and since the number of posts per user varied, these arrays were padded with dummy embeddings (arrays of zeros), to the maximum length of all users. These arrays were combined into a single Tensor of shape *(number of users, maximum number of posts, embedding size)*.

4.5. Label Distribution Smoothing

One aspect that had to be adressed was the fact that the datasets were imbalanced in certain personality traits, as referred in Section 4.1. A technique named Label Distribution Smoothing, or LDS, was performed to deal with imbalanced data. With this technique the imbalance in the label distribution is perceived by the model differently than the objective label distribution. This happens as, in regression, labels are continuous and their distances are meaningful, which means that different imbalances have different impacts. Unlike classification, where equally underrepresented samples are equally affected, in regression, a value in a sparsely represented area is more affected by imbalance than a similar value in a densely represented area, despite having the same number of observations.

LDS addresses this problem by yielding an adjusted label distribution, that is, that more closely reflects this perceived imbalance, which can be used to offset the imbalance training process. By performing the convolution between a kernel function and the label distribution, the adjusted label distribution is obtained. In this work, the inverse of the LDS estimated label density the samples were used weights which were applied to the loss function.

4.6. Prediction Model

As seen above, the state-of-the-art is dominated not only by approaches resorting to large pre-trained language models, but also by the use of deep neural networks as classifiers. In this work, neural networks were chosen as the regression model, as this model is suitable for tasks with large datasets and can learn complex patterns in data.

Multiple architectures of neural networks were tested, with a different number of hidden layers, with different types of hidden layers, different regularization techniques and with different hyperparameters. Recurrent architectures, with GRU and LSTM layers, were compared with MLP architectures, with the MLP architectures found to perform better. In what concerns to regularization, the techniques which showed to improve results, reducing overfitting, were L2 regularization, dropout and early stopping.

Mean Average Error (MAE), Root mean Squared Error (RMSE) and Mean Squared Error (MSE) were used as loss functions, as well as the versions of these functions adjusted with the LDS weights. The model was trained and evaluated, using 5-fold cross-validation. The final version of the model, which yielded the highest results, was characterized by the parameters displayed in Table 1.

5. Results

This section presents the results obtained for both scenarios described in Section 4.2. These results will also be compared to studies using similar datasets.

5.1. Single Domain Learning

As previously referred in Subsection 4.2, in these experiments, the proposed model was trained and evaluated using data from a single dataset exclusively, for both the Twitter and Facebook datasets. The metrics used to evaluate the model were MAE, RMSE and MAE. Before calculating the MSE metric, the objective and predicted personality scores had to be converted to a 1-5 scale, as to be able to compare it with [16]. Multiple model configurations were evaluated for both datasets, and, three parameter choices were found to impact results significantly.

The first one was the choice of the embedding model, with the choice of using S-BERT embeddings improving average MAE results by up to 8% and 10%, for the Facebook and Twitter datasets respectively, when compared to the DistilBERT embeddings. Secondly, the use of an MLP architecture proved to

Table 1: Parameters of the final neural network.

Hyperparameter	Values
Number of hidden layers	3
Type of hidden layers	Fully Connected
Nº of neurons per H. layer	256, 128 and 64
Activation function	Leaky ReLU
Loss function	Weighted MSE
Regularization	L2, Dropout and Early Stopping
Learning rate	5×10^{-4}
Batch size	32
Number of epochs	80

outperform the recurrent architectures, yielding average MAE results, up to 16% and 15% lower. Due to computational resources limitations, the recurrent architectures tested were less complex, containing less layers and neurons per layer, than the MLP architecture. Finally, batch size choice also resulted in significant variations in results for the Twitter dataset, with an improvement in average MAE of around 5% while for the Facebook dataset, results did not vary significantly. The remaining parameter changes did not yield significant variations in performance.

The results of these experiments were compared to the studies reviewed in Section 3, which used similar data from the same sources (MyPersonality and PAN15). These studies, used for comparison, were [3], which used a similar PAN15 Twitter dataset, [10], in which the MyPersonality dataset was used as well as the similar PAN15 Twitter dataset, and [16], in which the MyPersonality dataset was used. A ZeroR regression model with 5-fold CV was used as baseline model. All of these results, as well as the results of best performing model configuration of this study, with the parameters depicted in Table 1, can be seen in Tables 2 and 3.

Table 2: Results comparison for the Facebook dataset

Model	RMSE	Avg. MSE
Farnadi et al. 2016	0.1810	
Leonardi et al. 2020		**0.3026**
Baseline Model	0.1806	0.5371
Proposed Model	**0.1762**	0.5167

Table 3: Results comparison for the Twitter dataset

Model	Avg MAE	Avg. RMSE
Arroju et al. 2015	0.1414	
Farnadi et al. 2016		0.1770
Baseline Model	0.1395	0.1688
Proposed Model	**0.1158**	**0.1476**

From the results obtained, not only was the performance in the Twitter dataset significantly better than in the Facebook dataset, the impact that the model configuration had on the results was also larger, as in the Facebook dataset, results did not vary significantly in most configurations. This might be attributed to the higher quality of the Twitter dataset, when compared to the Facebook dataset. This was due to the fact that the former was larger in size as well as more homogeneous. Additionally, the performance of the model was better for the Openness and Agreeableness traits, compared to the remaining ones, which indicates that these can be more easily predicted than the other traits.

Comparing the obtained results to the baseline model, the model outperformed in all all traits in both datasets, except for the Extroversion dimension in the Facebook dataset. For the Twitter dataset the proposed model outperformed [3] by around 18%, in terms of decrease in average MAE, and outperformed [10] by around 16% in terms of decrease in average RMSE. For the Facebook dataset, while the proposed model outperformed the results in [10] by approximately 11%, in terms of decrease in average RMSE, it significantly underperformed when compared to [16], yielding an average MSE approximately 70% higher.

The fact that the proposed model yielded state-of-the-art results in the Twitter dataset and was outperformed by [16], which employed a similar approach, based in PTM embeddings and neural networks, supports the argument that using pre-trained embeddings with neural networks as a regression model yields poential for state-of-the-art improvements.

5.2. Cross-domain Learning

An additional objective of this study was to analyse how personality trait prediction models behaved when using data from different domains or contexts. To analyse this, the best performing model of the experiments in the previous section was tested in two settings. Firstly, in order to to understand if combining data from different domain resulted in improvements in performance, the model was trained using a combination of both datasets and was subsequently evaluated on each of the datasets separately, and compared with the previous section's results. In the second setting, an experiment was performed by training the same model in one of the datasets and evaluating it on the other one, in order to understand how well personality trait prediction models could generalise from one context to another.

The results of these experiments for the Facebook and Twitter datasets can be seen in Tables 4 and 4.

Table 4: Cross-domain results for the model evaluated on the Facebook dataset.

Training Dataset	Avg. MAE
Facebook	0,1402
Twitter	0,1610
Twitter + Facebook	0,1626

Table 5: Cross-domain results for the model evaluated on the Twitter dataset.

Training Dataset	Avg. MAE
Twitter	0,1158
Facebook	0,1610
Twitter + Facebook	0,1204

In the first experiment, the model trained with the combined datasets performed worse than the one trained and evaluated with a single dataset, yielding an average MAE 16% and 4% higher, for the Facebook and Twitter datasets respectively, supporting the findings in [10], that combining data from different domains to perform personality trait prediction does not result in better performance. Similarly, in the second experiment, with training the model in one dataset and evaluating it on the other, resulted in significantly worse performance than by using the same dataset for both operations, with the former approach yielding an average MAE 40% and 15% higher than in the latter approach.

In conclusion, there are two aspects affecting the ability of a model to benefit from data from different sources, as well as generalising from data of one source and another: the quality of the datasets and their similarity in context. In the case of this study, the quality of the Twitter dataset was higher, as is evident by the results, and, it is apparent that the two datasets differ too much in context to be effectively combined or applied to one another.

6. Conclusions

In this study, the aim was to create a model capable of predicting the Big Five personality traits of social media users, resorting to large pre-trained language model embeddings. The proposed model yielded satisfactory results, outperforming the previous state-of-the-art in the Twitter dataset by around 16-

18%, but also performing considerably worse than the state-of-the-art for the Facebook dataset, which employed a similar approach to the proposed model. These results suggest that an architecture using large pre-trained model embeddings as features combined with neural network regression models, yields state-of-the-art results in personality trait prediction.

Additionally, it was found that sentence embeddings extracted with the SentenceBERT model are more suitable for personality trait prediction than with the DistilBERT model. In regards to the Big Five traits, the results indicate that the Openness and Agreeableness dimensions are easier to predict than the other three traits. On the cross-media experiments, it was found that the ability these models to utilize data from different sources and to generalise from one context to another is dependent on the similarity and quality of the data.

In order to improve on these results in the future, an approach using a larger dataset as well as a more complex model is advised, or even fine-tuning a large-pre-trained language model. Additionally, the use of word embeddings instead of sentence embeddings might also result in an improvement in performance.

Acknowledgements
This work was supported by LARSyS FCT funding (DOI: 10.54499/LA/P/0083/2020, 10.54499/UIDP/50009/2020, and 10.54499/UIDB/50009/2020).

References

1. Gordon W. Allport and Henry S. Odbert. Trait-Names. A Psycho-lexical Study. *Psychological Monographs*, 47(1), 1936.
2. Felipe Almeida and Geraldo Xexéo. Word Embeddings: A Survey. *arXiv e-prints*, page arXiv:1901.09069, January 2019.
3. Mounica Arroju, Aftab Hassan, and Golnoosh Farnadi. Age, gender and personality recognition using tweets in a multilingual setting. In *CLEF 2015 working notes*, 2015.
4. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
5. Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. Workshop on Computational Personality Recognition: Shared Task – Technical Report. In *Procs. AAAI Workshop*, 07 2013.
6. Hans Christian, Derwin Suhartono, Andry Chowanda, and Kamal Zamli. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal of Big Data*, 8, 05 2021.
7. D.M. Condon, J. Coughlin, and S.J. Weston. Personality Trait Descriptors: 2,818 Trait Descriptive Adjectives Characterized by Familiarity, Frequency of Use, and Prior Use in Psycholexical Research. *Journal of Open Psychology Data*, 10(1):1–9, 2022.
8. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Procs. 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, June 2019. Association for Computational Linguistics.
9. Hana Esmaeel. Analysis of classification learning algorithms. *Indonesian Journal of Electrical Engineering and Computer Science*, 17(2):1029–1039, February 2020.
10. Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine De Cock. Computational personality recognition in social media. *User Modeling and User-Adapted Interaction*, 26, 06 2016.
11. Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine Cock. Recognising personality traits using facebook status updates. *AAAI Workshop - Technical Report*, 7:14–18, 01 2013.
12. Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 149–156, 2011.
13. Jennifer Golbeck, Cristina Robles, and Karen Turner. Predicting personality with social media. In *Procs. Conference on Human Factors in Computing Systems*, pages 253–262, 05 2011.
14. R.K. Hernandez and I. Scott. Predicting myers-briggs type indicator with text. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.
15. Francisco Iacobelli, Alastair J. Gill, Scott Nowson, and Jon Oberlander. Large scale personality classification of bloggers. In Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin, editors, *Affective Computing and Intelligent Interaction*, pages 568–577, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

16. Simone Leonardi, Diego Monti, Giuseppe Rizzo, and Maurizio Morisio. Multilingual transformer-based personality traits estimation. *Information*, 11(4), 2020.
17. Francois Mairesse, Marilyn Walker, Matthias Mehl, and Roger Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res. (JAIR)*, 30:457–500, 09 2007.
18. Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79, 2017.
19. Robert R. McCrae and Paul T. Costa Jr. A Five-Factor Theory of Personality. In L.A. Pervin and O.O. John, editors, *Handbook of Personality: Theory and research*. New York: Guilford, 2nd edition, 1999.
20. Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189, 2020.
21. Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53, 04 2020.
22. Bonan Min, Hayley Ross, Elior Sulem, Amir Veyseh, Thien Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 06 2023.
23. Veronica Ong, A.D.S. Rahmanto, Williem Williem, and Derwin Suhartono. Exploring personality prediction from text on social media: A literature review. *Internetworking Indonesia Journal*, 9:65–70, 01 2017.
24. PAN Organization. Author profiling 2015. https://pan.webis.de/clef15/pan15-web/author-profiling.html, 2015. Accessed: 07/02/2024.
25. James Pennebaker and Laura King. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77:1296–312, 01 2000.
26. J.R. Quinlan. Learning with continuous classes. In *Procs. 5th Australian Joint Conference on Artificial Intelligence*, volume 92, page 343–348. Singapore: World Scientific, 1992.
27. Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*, 2019.
28. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.
29. Tommy Tandera, Hendro, Derwin Suhartono, Rini Wongso, and Yen Lina Prasetio. Personality prediction system from facebook users. *Procedia Computer Science*, 116:604–611, 2017. Discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017).
30. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
31. Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. Pre-Trained Language Models and Their Applications. *Engineering*, 2022.