Egyptian Informatics Journal 22 (2021) 239–244

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Psychological Human Traits Detection based on Universal Language Modeling

Kamal El-Demerdash *, Reda A. El-Khoribi, Mahmoud A. Ismail Shoman, Sherif Abdou

Department of Information Technology, Faculty of Computers and Artificial Intelligence, Cairo University, Egypt

ARTICLE INFO

Article history:
Received 7 March 2020
Revised 5 September 2020
Accepted 8 September 2020
Available online 31 October 2020

Keywords:
Big Five Personality Model
Personality Traits
LSTM
NLP
Text Analytics
Deep Learning
ULMFiT

ABSTRACT

Personality Traits Detection is one of the important problems as a text analytics task in Natural Language Processing (NLP). Text analytics is the process of finding out insight knowledge over written text. Although most deep learning models give high performance, they often lack interpretability. Computer Vision (CV) has been affected significantly with inductive transfer learning, however training from scratch and task-specific modifications are still wanted in many NLP techniques.

This paper addresses the problem of personality traits classification. We adopted the use of the Universal Language Model Fine-Tuning (ULMFiT) in personality traits detection. The model makes use of transfer learning rather than the classical shallow methods of word embedding and proved to be the most powerful model in many NLP problems.

The basic advantage of using this model is that there is no need to do feature engineering before classification. When applied to benchmark dataset, the proposed method shows a statistical accuracy improvement of about 1% compared to the state-of-the-art results for the big five personality traits.

1. Introduction

Personality traits greatly affect our life; they impact our lifestyle options, luxury, wellness, and many other predilections [1]. As a result of the growth in technology and the dramatic spread of social media services over the last few years, new opportunities are now available to study personality from the digital traces individuals leave behind. These traces are primarily in text form, such as shared comments and observations, and other media such as images and videos. People frequently use social networks to share their activities and observations and broadcast their opinions about topics and events. This is a fertile source of information to help us understand human needs and emotional states, and is therefore considered to be a rich source for personality traits detection. In last few years the interest of the scientific society has radically grown towards personality recognition because it's applications proved to be a valuable resource in social computing and social network analysis [3].

Automatic detection of personality traits has many remarkable practical applications. Personality traits help to identify certain illnesses. It is well-known that there is a connection between irritation and heart disease [2].

This relation between the personality trait and the disease is extensively investigated and considered by health authorities. Personality detection is also helpful in forensics. It may be helpful in some cases when the investigators know the personality traits of the individuals who were at the crime scene, decreasing the pool of potential suspects [3].

1.1. Research motivation

In many scientific articles the deep learning structures and NLP tasks are treated [3]. These articles use technical vocabulary which sometimes can be difficult to understand. There is now a huge demand for the rating of online documents because of their rapid spread in many different languages [3].

In this context, transfer learning or domain adaptation has been widely used in machine learning, especially in the era of deep neural networks, could help to reuse models were developed and

* Corresponding author.
E-mail addresses: k.eldemerdash@grad.fci-cu.edu.eg (K. El-Demerdash), r.abdelwahab@fci-cu.edu.eg (R.A. El-Khoribi), m.essmael@fci-cu.edu.eg (Mahmoud A. Ismail Shoman), s.abdou@fci-cu.edu.eg (S. Abdou).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

https://doi.org/10.1016/j.eij.2020.09.001

trained in a source task to another target task [4]. The power of transfer learning is very clear when the features learned from the source or base task are general and can be repurposed to the target tasks. Nowadays, most CV models base extracting the feature to a pretrained models like AlexNet, ResNet, MS-COCO, etc. [4]. However in NLP models, the idea did not have such success until Howard and Ruder have proposed the ULMFiT [5]. Deep pretraining word representation was one of the biggest achievements of ULMFIT as moving from shallow to deep models. The concept of ULMFiT has proved the ability of transfer learning for the NLP world. Language Modeling (LM) was the password to ULMFiT to be state-of-the-art. The ULMFiT makes use of the Average stochastic gradient descent-Weighted Dropout Long Short Term Memory (AWD-LSTM) LM which was proposed by Stephen Merity [6]. LSTM is a specific type of the Recurrent Neural Networks (RNNs).

In spite of this there are an increasing number of works in personality detection. It's still tricky to say what is the state-of-the-art because almost all the scientists working in the field have used different datasets and evaluation measures.

In our study, we have used the most agreeable big five model [7] to define and discover our traits from text. It is considered the most agreeable model because it calculates and epitomizes our personality through five dimensions summarized in Table 1 in next binary (yes/no) values:

In this paper, we show a technique to take out personality traits from flow of consciousness articles through a ULMFiT structure. We have trained five independent networks using one structure for the five personality traits.

Every network is a binary classifier to predict the positive or negative of corresponding trait. Our results show a significant accuracy improvement compared to the state-of-the-art results for the big five personality traits.

We organized the rest of this paper as follows:-

We give a general review of the related work in Section 2 and we show the proposed model architecture in Section 3. We explain and discuss the experimental setup, datasets used, evaluation measures and basic results in Section 4. Conclusions and future work are presented in Section 5.

2. Related work

Automate of Personality disclosure is considered as prospective and novel domain. Yash Mehta et al. [3] have made inclusive literature scanning of the recent directions and evolution in personality trait detection domain.

They have revised many topics, they showed the historic machine learning models to detect personality with an affirmation on models driven by deep learning.

They presented a general view of the most widespread techniques for personality detection in automated way using different annotated datasets. In addition to that, they discussed the personality detection applications in different industries and the latest models in machine learning to extract personality with particular concentrate on multi-modal approaches.

Table 1
Big Five Personality Traits Adapted from [1].

Five Traits	Description
Extroversion (EXT)	Is the person outgoing, talkative,and energetic versus reserved and solitary?
Neuroticism (NEU)	Is the person sensitive and nervous versus secure and confident?
Agreeableness (AGR)	Is the person trustworthy,straightforward, generous, and modest versus unreliable, complicated,meager, and boastful?
Conscientiousness (CON)	Is the person efficient and organized versus sloppy and careless?
Openness (OPN)	Is the person inventive and curious versus dogmatic and cautious?

Regarding the textual modality, they emphasise that data preparation, capturing useful features from texts are a very important stages and adopting the right technique can produce great results. Although some researchers have made thoroughly research and found a number of useful features from text such as Linguistic Inquiry and Word Count (LIWC) [8], Mairesse [9]. However, figuring out useful features from texts with close connection to user's personality is still a challenging task to explore [10]. Generally, features are extracted from text, to insert it in typical ML classifiers like support vector machine [11]. James Pennebaker et al. [12] did some previous work on detection of personality traits from plain text, they collected the essays dataset that we used in our experiments.

They defined the correlations between the essay and personality by using LIWC features [8]. To improve the result using the essays dataset, Francois Mairesse [9] added other features such as imageability. In the same context, combinations of feature sets were applied by Saif Mohammad et al. [11] to outperform results in [9], which they called the Mairesse baseline. Sun et al. [10] used the same essays, and have proposed model using bidirectional LSTMs concatenated with Convolution Neural Network (CNN) which called the 2CLSIM Model. They captured user's personality using the structural features from texts. They have introduced abstract feature combination based on closely connected sentences.

The state-of-the-art results using the essays dataset was introduced by Majumder et al. [1], they encoded the essays from word level to sentence level. They have used 3-dimensional convolution to learn the structure of an article. They gave future direction of research incorporate more features and preprocessing i.e if written text contain repeated words, sentences contains spam content or unrecognized character set etc. preprocessing filter those. Our study proposes an efficient and explainable model based on language modeling. LM aims to predict the previous word or next one given a list of words [6].

It introduces important basics in most of NLP applications through understanding the long-term dependencies and hierarchical structure of the text [6]. Although language modeling is the basic component of NLP tasks such as dialogue modeling and machine translation, LMs suffer from tragic forgetting and are too large to fit into small datasets when fine-tuned with a classifier.

2.1. Word embeddings

Word embeddings are a category of mechanism where separate words are symbolized as vectors with real-values in a predefined vector space [13]. Key features of this technique is the thought of representing every word in a densely distributed representation [14]. Often the representation of these words appear in tens or hundreds of dimensions. This is different from one-hot encoding technique which requires thousands or millions of dimensions for word representations. Fig. 1 illustrates the word embeddings concept.

The best feature in the word embedding technique appears when words have the same meaning, where they have a similar representation [13]. GloVe and Word2Vec [13,14] are very common path for learning word embeddings and presenting them as vectors. They are a distributed representation for text that have a great execution for deep learning methods on challenging NLP problems. The distributed representation is learned based on the usage of words. This result in *similar representations for the words*

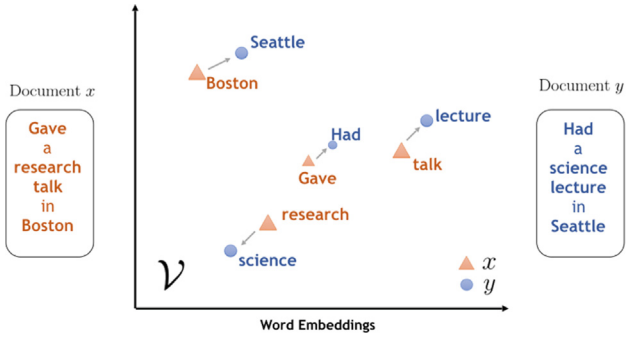

Fig. 1. Example of Word Embbedings.

that are used in similar ways, grasp their meaning in a natural manner [13].

2.2. Transfer learning

Transfer learning is an exciting concept where we try to leverage prior knowledge from one domain task and insert it into a different task [15]. This work is inspired by the way humans learn. We have an inherent ability to learn in progressively rather than leaning from scratch. Many state-of-the-art models in NLP have to be trained from scratch and require large datasets to achieve reasonable results [16]. They not only take up huge quantities of memory but are also quite time consuming. Specifically, in text classification, there might not even be enough labeled examples to begin with. Inductive transfer learning tackles exactly these challenges [15], and it is also the central concept ULMFiT is based on. Fig. 2 shows the conceptual difference to traditional machine learning.

Transfer learning aims to mimic the human ability to acquire knowledge while learning one task and to then utilize this knowledge to solve some related tasks [16].

In the traditional approach, two models are trained separately without either retaining or transferring knowledge from one to the other. On the other hand transfer learning would be to retain knowledge (e.g. weights or features) from training a model 1 and to then utilize this knowledge to train a model 2 [16]. In this case, model 1 would be called the source task and model 2 the target task [17]. In this context, there are many areas in NLP that can be utilized using transfer learning to optimize the general process of deep learning as in LM which aims to forecast the words in context. Consider we have this sentence: "I thought I would finish on time, but unfortunately ended up 5 min. . .".

It's clear in a sensible way to the reader that the next word will be a word or phrase that means "late". Effectively finding an answer to this task demands not only an understanding of linguistic construction (nouns follow adjectives, verbs have subjects and

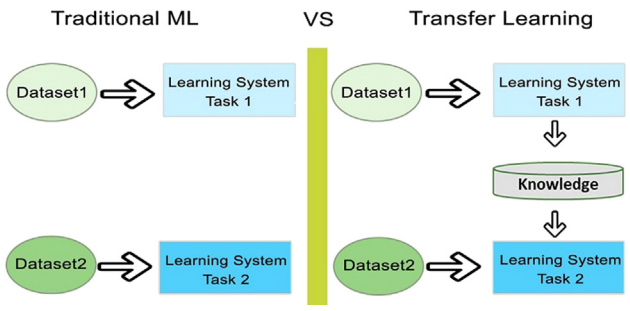

Fig. 2. Traditional ML vs. Transfer Learning [16].

objects, etc.) but also the capability to make decisions based on broad contextual evidences ("late" is a feasible choice we can use in the blank in this case because the previous text gives a clue that the sentence is talking about time). In addition, LM has a catchy property where no labeled training data is required [6]. Raw text is amply available for every potential domain. These two properties make language modeling a perfect fit for learning generalized base models.

2.3. Universal language model fine-tuning overview

UMLFiT depends on the most general inductive transfer learning setting for NLP: Given a static source task TS and any target task TT with TS not equal TT, it would like to improve performance on TT and the ideal source task will be the LM [5].

ULMFiT is an efficient and functional transfer learning model that can be used in many tasks in NLP. The actual model structure showed in Fig. 3. The model introduces techniques that are key for fine-tuning a language model by making use of the state-of-the-art AWD-LSTM LM, the same 3-layer LSTM architecture with the same hyper-parameters and no additions other than tuned dropout hyper-parameters are used [5]. The classifier layers above the base LM encoder is simply a pooling layer (maximum and average pool) followed by a fully connected linear layer block.

Ruder Said that, *"language modeling is particularly suited for capturing facets of language that are important for target tasks*." In the same context Ruder expects that, *"It only seems to be a question of time until pretrained word embeddings will be dethroned and replaced by pretrained language models in the toolbox of every NLP practitioner*." UMLFiT has already captured facets of language and produced impressive empirical results. It incorporates several fine-tuning techniques which are broadly applicable and could boost performance for other methods. The overall models significantly outperforms the state-of-the-art on six text classification tasks including three tasks for sentiment analysis, reducing the error by 18–24% on the majority of datasets [5]. The code and pre-trained models are open source on (http://nlp.fast.ai/ulmfit).

3. Proposed architecture

After having introduced the basic ideas underlying ULMFiT, we can now focus on the high-level architecture of the proposed model for personality task using ULM model showed in Fig. 4. The complete model consists of the combination of a pre-trained ULM model and additional task-specific layers for the desired tasks. Once a ULM model is developed, the learning process becomes limited to learning the parameters of the additional layers. This transfer learning process is referred to as fine-tuning with

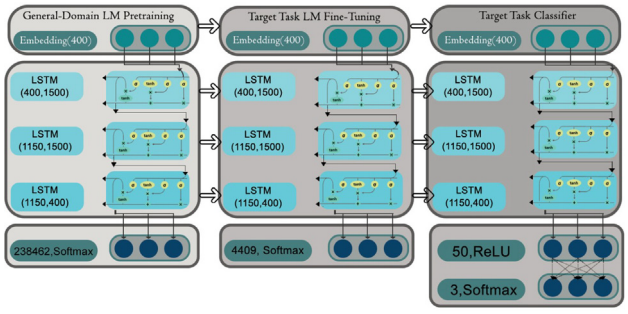

Fig. 3. ULMFiT consists of three main stages: 1- The LM is trained on a general-domain corpus to capture general features of the language in different layers, 2- The full LM is fine-tuned on target task data using fine-tuning to learn task-specific features, 3- The classifier is fine-tuned on the target task [16].

ULM and this is the main benefit of using ULMs. We have trained five independent networks using one structure for the five personality traits on a well-known dataset typically used to compare personality detection techniques.

For text preprocessing contrary to images in CV, text can't immediately be transformed to numbers to insert into a model. The first step we need to do is to preprocess our text so that we alter the raw texts to lists of words or tokens (that is called tokenization step). Then convert these tokens into numbers (that is called numericalization step).

These numbers are then passed to embedding layers that will transform them in arrays of floats before inserting them into a model. For the purpose of providing an overview to the proposed model, it can be split into three steps:

1 Getting our data preprocessed and ready for modeling
2 Fine-tuning a language model to our dataset,
3 Building a classifier on top of the encoder of the LM

3.1. General-domain LM pretraining

In the first step, a LM is pretrained on a large general-domain corpus, on Wikitext-103 consisting of 28,595 preprocessed Wikipedia articles and 103 million words [5]. Now, the model is able to predict the next word in a sequence (with a certain degree of certainty).

At this stage the model learns the general features of the language, e.g. that the typical sentence structure of the English language is subject-verb-object.

3.2. Target task LM fine-tuning

Following the transfer learning approach, the knowledge gained in the first step should be utilized for our personality classification task. However, the personality task dataset (i.e. the essays) is likely to come from a different distribution than the source task dataset.

To address this issue, the LM is consequently fine-tuned on the data of our task. We train the model on the forward and backward LMs for both the general-domain and our task specific dataset. This stage converges faster as it only needs to adjust to the characteristics of the target data, and it allows to train a solid LM even for small datasets [5]. Just as after the first step, the model is at this point able to predict the next word in a sequence. Now however, it has also learned the essays dataset-specific features of the language, which we used in our experiments.

3.3. Target task classifier fine-tuning

Since ultimately, in our case, we do not want our model to predict the next word in a sequence but to provide a personality traits classification, in this step the pretrained LM is expanded by two linear blocks so that the final output is a probability distribution over the trait labels (i.e. EXT, NEU, AGR, CON and OPN).

For training the overall classification model, the model is trained in an end to-end way in a supervised learning framework. The aim of the training is to optimize all the parameters to minimize the objective function as much as possible. We trained the architecture by using slanted triangular learning rates (STLR). Fig. 5 shows the STLR which change the learning rate for each iteration in triangular fashion. It is used for ULMFiT as a function of the number of training iterations, we used only once cycle as recommend by the authors [5]. The model was trained by discriminative fine-tuning which uses different learning rate for each layer group and it is trained gradually by freezing and unfreezing layers for different groups. As followed in [5] the parameters in our task-specific classifier layers are the only ones that are learned from scratch. The first linear layer takes as the input the pooled last hidden layer states.

4. Experiments and results

In this section, we present the experiments to evaluate the effectiveness of our model for personality detection. To evaluate the proposed method, we tested it on a well-known dataset typically used to compare personality detection techniques.

4.1. Experimental setup

Lately, we noticed public availability of systematically annotated datasets. Table 2 provides an overview of common datasets with text modality for personality-detection. Accordingly, scientists can now focus on creating and using superior models and architectures rather than on data collecting and preprocessing [3].

It is worth mentioning that the MBTI traits are more complex to predict than the big five traits and are lacking in resources [20]. However, the MBTI personality measure is the most popular measurement used across the world right now [3].

4.2. Datasets

To evaluate the proposed UMLFiT we have used the stream-of-consciousness essay dataset compiled by James Pennebaker [12], which used to compare personality detection in text modality.

The size of the used dataset is 2,467 valid unidentified essays per class (1.9 million words). The essays labeled with the authors' personality traits: EXT, NEU, AGR, CON, and OPN in binary (yes/no) values. we experimented with all the essays. The labels of stream-

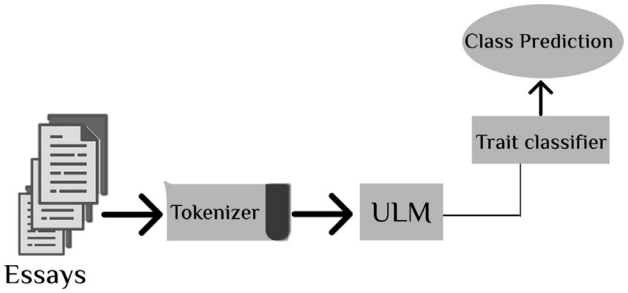

Fig. 4. High Level Architecture of the Proposed Model.

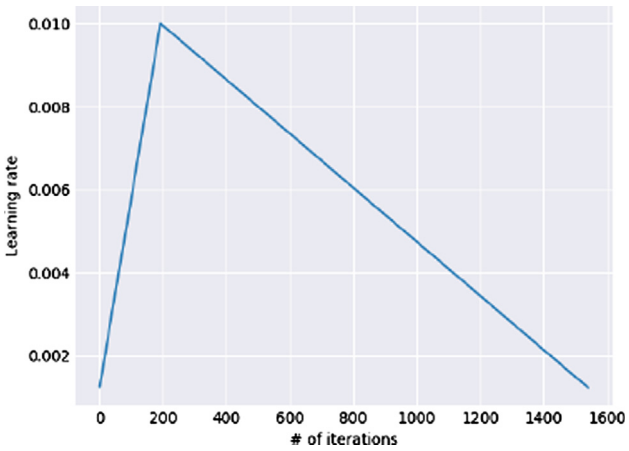

Fig. 5. The slanted triangular learning rate [5].

Table 2
Popular Personality-Detection Datasets. Adapted from [3].

Dataset	Personality Measure
(Pennebaker and King 1999) [12]	Big Five
(Tausczik and Pennebaker 2010) [18]	Big Five
MBTI Kaggle	MBTI
Italian FriendsFeed (Celli et al. 2010) [19]	Big Five

of-consciousness essays data come from the author's own questionnaire, which can be explained as autognosis [10]. They collected volunteers to write stream-of-consciousness essays in a controlled environment and then they requested the authors of the essays to locate their own big five personality traits [12]. They defined the correlations between the essay and personality by using LIWC features. For example, people who got high grades on extraversion usually use more positive emotion words (e.g., great, amazing, happy) while those higher in neuroticism use first-person singulars (e.g., I, mine, me) much more [3].

4.3. Evaluation measures

We use the classes accuracy for head-to-head comparison with the state-of-the-art accuracy results.

Regarding the hyperparameters, we use the AWD-LSTM language model with an embedding size of 400, 3 layers, 1150 hidden activations per layer [6]. We use backpropagation through time (BPTT) of 80 to enable gradient propagation for large input sequences for LM. We use mixed precision training which enables us to use a higher batch size of 256 by training part of the model in FP16 precision, and also speeds up training by a factor 2 to 3 on modern GPUs.

We train the model using gradual unfreezing (partially training the model from everything but the classification head frozen to the whole model training by unfreezing one layer at a time) and differential learning rate (deeper layer gets a lower learning rate). We apply dropout of 0.3 to layers, 0.2 to RNN layers, 0.3 to input embedding layers, 0.05 to embedding layers, and weight dropout of 0.5 to the RNN hidden-to-hidden matrix. We fine-tuning the forward and backward LM for 10 epochs for each one. Fine-tuning took 200 to 400 s which depends on each class. The classifier is a model that is a bit heavier, so we have lower the batch size of 32.

The classifier needs a little less dropout, so we pass dropout of 0.5 to multiply all the dropouts by this amount. For validation operations, we split our dataset as test and train data. We set aside a random 20% of all the essays to build our validation set. We fine-tune the classifier for 5 epochs. It is worth mentioning that many configurations used in our experiments are not shown here to avoid clutter, we showed the configurations which provide us the best traits accuracy.

4.4. Results and discussion

We compared our model with state-of-the-art competitive accuracy results reported by Majumder et al. [1] in the original paper, which has been confirmed by the recent published research in the domain [3].

Table 3
Proposed Model Traits Accuracy Compared to the state-of-the-art Results.

Trait	EXT	NEU	AGR	CON	OPN
State of the art Results	58.09	59.38	56.71	57.30	62.68
Proposed Results	58.85	59.88	59.25	57.97	63.30

Compared to the state-of-the-art accuracy results shown in Table 3, the proposed UMLFiT is competitive and has the state-of-the-art accuracy results for personality classification task using stream-of-consciousness essays dataset without do feature engineering before classification.

The proposed model outperformed the AGR trait with a significant statistical margin about 2.54% and with competitive statistical margins at least 0.5% in rest of traits. We trained the model on the forward and backward LMs for both the general-domain and our task specific dataset, both LMs backward and forward are used to build two versions of the same proposed architecture. For our best accuracy results, the final decision is the ensemble of both.

When applied an ANOVA test, our proposed method shows a significant statistical accuracy improvement of about 1% compared to the state-of-the-art results for the big five personality traits. We used Pytorch to build the whole model and make use of Fastai libraries for applying the training strategies and fine-tuning the language models. We ran our experiments on Google's Colaboratory environment with GPU acceleration.

5. Conclusions and future work

In this paper, we introduced a solution to personality traits classification task. We used the ULMFiT model to classify five personality traits. Our results show an accuracy improvement over the state-of-the-art accuracy results for all the five traits. This emphasises the perfection of transfer learning in personality traits detection from text as a NLP task. We didn't do feature engineering before classification as an advantage of using UMLFiT.

In the future, we plan to further improve the accuracy of the classifier through reinforcement learning. We also plan to train multi-label classifier instead of using one architecture for the five personality traits.

References

[1] Majumder N, Poria S, Gelbukh A, Cambria E. Deep learning-based document modeling for personality detection from text. IEEE Intell Syst 2017;32(2):74–9.
[2] Yilmaz T, Ergil A, Ilgen B. Deep learning-based document modeling for personality detection from turkish texts. In: Proceedings of the future technologies conference (FTC).
[3] Mehta Y, Majumder N, Gelbukh A, Cambria E. Recent trends in deep learning based personality detection. Artif Intell Rev 2019. https://doi.org/10.1007/s10462-019-09770-z..
[4] Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep learning for computer vision: A brief review. Comput Intell Neurosci 2018;02:1–13..
[5] Howard J, Ruder S. Universal language model fine-tuning for text classification. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers). p. 328–39.
[6] Merity S, Keskar NS, Socher R. Regularizing and optimizing lstm language models. In: International conference on learning representations; 2018..
[7] Digman J. Personality structure: Emergence of the five-factor model. Ann Rev Psychol 1990;41:417–40.
[8] Pennebaker J, Martha F, Roger, B. Linguistic inquiry and word count (liwc). In: vol 71. Lawrence Erlbaum Associates, Mahway; 2001..
[9] Mairesse F, Walker MA, Mehl MR, Moore RK. Using linguistic cues for the automatic recognition of personality in conversation and text. J Artif Intell Res 2007;30:457–500.
[10] Sun X, Liu B, Cao J, Luo J, Shen X. Who am i? Personality detection based on deep learning for texts. In: IEEE international conference on communications (ICC). IEEE; 2018. p. 1–6.
[11] Mohammad S, Kiritchenko S. Using hashtags to capture fine emotion categories from tweets. Comput Intell 2015;31(2):301–26.
[12] Pennebaker J, King LA. Linguistic styles: Language use as an individual difference. J. Person Soc Psychol 1999;77(6):1296–312.
[13] Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: EMNLP; 2014..
[14] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, December 05–10, 2013, Lake Tahoe, Nevada, p. 3111–3119..
[15] Pan SJ, Yang Q. A survey on transfer learning'. In: J. personality and social psychology, IEEE transactions on knowledge and data engineering, vol. 22, issue 10; 2010..

[16] Faltl S, Schimpke M, Hackober C. Ulmfit: State-of-the-art in text analysis by seminar information systems (WS18/19), February 7, 2019. https://humboldt-wi.github.io/blog/..
[17] Sarkar D. A comprehensive hands-on guide to transfer learning with real-world applications in deep learning. Towards Data Science, Nov. 14, 2018 [Blog]..
[18] Tausczik YR, Pennebaker JW. The psychological meaning of words: Liwc and computerized text analysis methods. J Lang Soc Psychol 2010;29(1):24–54.
[19] Celli F, Lascio FMLD, Magnani M, Pacelli B, Rossi L. Social network data and practices: the case of friendfeed. In: International conference on social computing, behavioral modeling, and prediction. Springer; 2010. p. 346–53.
[20] Furnham A. The big five versus the big four: the relationship between the myers-briggs type indicator (mbti) and neo-pi five factor model of personality. Personal Individ Differ 1996;21(2):303–7.