Multi-View Personality Profiling Based on Longitudinal Data

Kseniya Buraya[1], Aleksandr Farseev[21], and Andrey Filchenkov[1]

[1] ITMO University, 49 Kronverksky Pr., St. Petersburg, 197101, Russia
ksburaya@corp.ifmo.ru, afilchenkov@corp.ifmo.ru
[2] SoMin Research, 221 Henderson Rd., Singapore, 159557 farseev@gmail.com

Abstract. Personality profiling is an essential application for the marketing, advertisement and sales industries. Indeed, the knowledge about one's personality may help in understanding the reasons behind one's behavior and his/her motivation in undertaking new life challenges. In this study, we take the first step towards solving the problem of automatic personality profiling. Specifically, we propose the idea of fusing multi-source multi-modal temporal data in our computational "PersonalLSTM" framework for automatic user personality inference. Experimental results show that incorporation of multi-source temporal data allows for more accurate personality profiling, as compared to non-temporal baselines and different data source combinations.

Keywords: User Profiling · Social Networks · Personality Profiling.

1 Introduction

User profiling plays an important role in various applications. One of the major components of user profiling is personality profiling, which is the identification of one's mental and emotional characteristics, such as personality type or mental status. These personal attributes allow for better understanding of the reasons behind one's behaviour [22], the selection of suitable individuals for particular tasks [27], and motivation of people in undertaking new life challenges.

There are several personality scales adopted by the research community. One of the most widely embraced typologies is called MBTI [18]. MBTI typology is designed to exhibit psychological preferences on how people perceive the world around them and distinguishes 16 personality types. It consists of four binary personality classes that form human personality type when being combined. Social scientists discovered that social media services exceedingly affect and reflect the way people communicate with the world and among themselves [11], which suggests that MBTI typology naturally fits social media research and can be used for assigning personality labels to user data when inferring users' behaviors and activities on social media.

Several studies addressed the problem of personality profiling from social science perspective [22, 24]. However, most of these works are descriptive in nature and based on manually collected data, which do not scale well to large-scale observations. At the same time, others [23, 28] utilized the advantages of social network data for automatic personality profiling. However, most of these works are based on the data collected from a single data source (i.e. Twitter) or of a single data modality (i.e. text), which may lead to sub-optimal results in the real-world scenario. Indeed, taking into account that most of the social media users participate in more than one social network in their daily life [7], it is reasonable to utilize the data from multiple sources and modalities for automatic personality profiling. Another important aspect of social media data is its temporality, which is the tight dependence on user behavior on his/her temporal and spatial environment. For example,[2] found that one's relationship status (i.e. single/not single, which is closely related to one's personality) can be predicted from the history of users' check-ins. However, the temporal aspect of multi-source social media data was not yet comprehensively addressed in the literature [2]. Considering that user personality do not change a lot over time [18], it is essential to consider data temporality at the data modeling stage.

In this work, we focus on using multi-source multimodal temporal data for automatic personality profiling. We believe that the judicious fusion of multi-source heterogeneous temporal information sources would enrichment each other and facilitate more accurate detection of user personality traits, as compared to using single-source static data. We chose Twitter, Instagram, and Foursquare as the main data sources due to they are among the largest and diverse social media networks [4]. Specifically, we harvested Twitter as the textual data source; Instagram as the image data source; and Foursquare as the location data source.

Predicting user personality profiling from multi-source temporal data is a challenging problem due to the following issues:

- **Cross-Network User Account Disambiguation.** It is hard to align the accounts of the same user from different social media resources.
- **Incomplete Multi-Source Data Fusion.** Most social media users participate only in distinct sets of social media services (e.g. Twitter + Instagram or Foursquare + Twitter) and not always active in all of them. Both these factors introduce the problem of block-wise missing data, which is a significant challenge.
- **Incorporation of temporal data aspect.** Construction of multi-source learning models that take into account the temporal data dependencies is essential but was not well studied yet.

Inspired by the challenges above, in this study we seek to address the following research questions:

1. Is it possible to perform user personality profiling more accurately by learning from multiple incomplete data sources?
2. Is it possible to improve the performance of user personality profiling by leveraging on temporal aspect of data?
3. Which data sources contribute the most to user personality profiling?

To answer these research questions, we present our **idea of temporal learning from multiple social networks for automatic personality profiling**.

Specifically, we utilize multi-modal longitudinal data from Twitter, Instagram, and Foursquare in our multi-source learning framework “PersonaLSTM”. The framework simultaneously fuses multi-source multi-modal temporal data and performs personality predictions by following MBTI personality scale. The issue of block-vise incomplete data is solved by applying non-negative matrix factorization [13], while the data temporality is efficiently incorporated by using long-short-term memory neural networks [8]. The experimental results reveal the superiority of our proposed framework over non-temporal baselines and different data source combinations.

2 Problems with Current Approaches

It is worth mentioning that multiple research groups tackled the task of personality profiling from the Computational Social Science's point of view [9, 1]. These works all observe that users' personality score is related to their behavior on social media platforms. Furthermore, Youyou et. al. [30] mentioned that personality is a major driving force behind people's interactions, behavior, and emotions. Schwartz et. al. [26] analyzed the Facebook messages of 75000 volunteers and demonstrated the correlations between words usage and personality traits. Unfortunately, these works are descriptive in nature and do not tackle the problem for automatic personality profiling.

Meanwhile, the problem of single-source personality profiling was addressed by several research groups. For example, Kosinski et. al. [12] conducted an extensive correlation analysis over 180, 000 Facebook users by using different data representations, such as the size of individual social graph, the number of uploaded photos, and the number of attended events; and reported encouraging results on user extraversion prediction [12]. Later, Verhoeven et. al. [28] revealed that such MBTI categories as “introversion — extraversion” and “thinking — feeling” can be successfully predicted from Twitter data, while the prediction of the other two categories is more challenging. Finally, Wei et al. [29] incorporated tweets, avatars, emoticons, and responsive patterns for predicting personality traits from Sina Weibo[3]. The aforementioned works made significant contributions to the field of automatic personality profiling. However, they are all limited due to the use of the single-source data, which is inadequate in the real-world scenario.

At the same time, several research works addressed the problem of user profiling from the multi-source learning perspective. One of the research groups [5] utilized multiple social networks for the task of user demographics profiling, while in [2] they demonstrated that the incorporation of multi-source data helps to increase the accuracy of relationship status prediction. Finally, Nie et. al. [20] proposed an approach for seamless integration of information from multiple social networks for career path prediction. These works are related to our study regarding the incorporation of multi-source data for individual user profiling. However, they do not incorporate the temporal aspect of multimedia data, which is one of the essential components of our study.

[3] http://weibo.com

Finally, several works were dedicated to the usage of temporal data for user profiling. For example, Liu et. al. [14] proposed a compositional recurrent neural network architecture to learn text representations at the character, word, and sentiment level for the task of personality trait inference. Another work [10] incorporated temporal aspect of the data for sentiment classification by using LSTMs. Finally, [25]) proposed "temporal continuity"-based version of non-negative matrix factorization for emerging topic detection and reported its efficiency for Twitter stream analytics.

Even though the related works significantly contributed to user personality profiling, they did not address the problem of automatic personality profiling from multi-source temporal perspective. This work is the first attempt to fill this research gap.

3 Data Description

3.1 MBTI Scale

To obtain results on the type of personality, it is necessary to take the MBTI test, which consists of the answering a series of questions (from 72 to 222). The test is scored by evaluating each answer in terms of what it reveals about the taker. Each question is relevant to one of the MBTI category [15]: Extroversion/Introversion, Sensing/Intuition, Thinking/Feeling, Judging/Perceiving.

In this article, we name the MBTI category with the first letters of the two labels. It should be emphasized, that the labels for each category are not exactly inverse of each other. This is because the human can have in his/her character the part from both labels of MBTI category. The selected MBTI tests help to identify the predominant label for every MBTI category. Therefore we didn't choose any dominant label for each category.

3.2 Dataset Collection

The related works in the field of multi-source social media modeling proposed various approaches to solving the problem of cross-network user account disambiguation [6, 4]. In this work, we adopt the so-called "cross-linking user account mapping" strategy, where Twitter is used as a "sink" that accumulates Instagram and Foursquare re-posts as well as Twitter tweets in one information channel. To obtain personality-related ground truth, we utilized Twitter search API[4] to perform a search for the results of trusted online MBTI tests, such as 16 Personalities[5], Jung Typology Test[6], and MBTI Online[7]. After collecting tweets with MBTI test results, we extracted MBTI ground truth labels from them.

[4] http://dev.twitter.com/rest/public

[5] http://16personalities.com

[6] http://humanmetrics.com/

[7] http://mbtionline.com

As a result, we obtained MBTI labels for $15,788$ Twitter users. After collecting ground truth results and users' Twitter profiles, we downloaded all possible tweets, photos, and check-ins for each user.

Table 1. Dataset Statistics

	Twitter	**Instagram**	**Foursquare**
#users	15788	10254	3090
	tweets	**images**	**check-ins**
#posts	122,584,534	4,789,519	420,603

4 Data Representation

In this section, we overview the features that we extracted from our collected dataset.

Heuristically-Inferred And Lexicon Features. First, we counted the number of URLs, the number of hashtags and the number of user mentions. Second, we calculated the number slang words, the number of emotion words, the number of emoticons, and the average sentiment score. Third, we computed the linguistic style features, such as the number of repeated characters in words, number of misspellings, and number of unknown to spell checker words. Lastly, we utilized several crowd-sourced lexicons, which are associated with controversial subjects from the US press and healthiness categories. We also calculated the average level of user Twitter activity during eight daytime durations (3-hour intervals) that could indirectly be related to users' activities. In total, we've extracted 53 heuristically-inferred and lexicon features.

Linguistic Features. We extracted LIWC features [21] that were found to be a powerful mechanism for personality, age, and gender prediction purposes [23]. For each user, we extracted 64 LIWC features.

LDA Features. For each user, we merged all his/her tweets into "documents" (one user - one document) and then projected these documents into a latent topic space by applying Latent Dirichlet Allocation (LDA). As a result, for each user, we extracted 50 LDA features[8].

Visual Features. We automatically mapped each Instagram photo to 1000 ImageNet [3] image concepts by using pre-trained GoogleNet model. We then summed up the predicted concept occurrence likelihoods for each user and divided the obtained vector to the total number of images posted by this user. In total, we extracted $1,000$ image features for every user.

Location Features. We utilized 886 Foursquare venue categories to compute location features. For each user, we counted the total number of his/her Foursquare check-ins in venues of each venue category. Then, we divided the number of check-ins in each category by the total number of check-ins of the user.

5 Personality Profiling

This section presents our multi-view temporal data learning approach for the task of automatic personality profiling.

[8] We empirically set $\alpha = 0.5$, $\beta = 0.1$, $T = 50$ topics for $1,000$ LDA iterations.

McCrae et al. [17] stated that personality traits in adulthood continuously evolve, but such changes are happening rare and over long periods of time. The above observation inspired us to incorporate temporal aspect in the form of not large time intervals. We thus divided user timelines into $k = 10$ time intervals of 142 days each, so that within **such summary interval user's personality is not expected to change significantly**. Such data division approach is expected to be helpful in identifying temporal user behavior patterns. According to [17] during all these time intervals user' personality will not significantly change, which means that we can use the same personality type for all time periods.

The original MBTI scheme assumes that each MBTI category represents distinct sides of a human character. For example, Mattare et al. [16] demonstrated that "sensing – intuition" category is related to human entrepreneur skills. In this work, we thus focus on the prediction of each MBTI category separately, so that they can later be aggregated into one of 16 MBTI personality types.

5.1 The PLSTM Framework

Long Short Term Memory neural networks (LSTMs) is a particular category of Recurrent Neural Networks (RNNs) that are capable of learning long-term temporal data dependencies. Such property fits well to our problem of temporal multi-source learning for personality profiling.The architecture of our PLSTM framework is illustrated on Figure **??**.For the prediction of the full MBTI profile, four separate models will need to be trained for four MBTI categories, each of which consists of two labels. Each model predicts one final label for each MBTI category. For each model, the features extracted from the data that corresponds to the m time periods are used as inputs to the corresponding m LSTM layers. Each layer except the first one uses the learned information from the previous layers. In this way, the last layer represents the information for m periods of time. The output of the last layer is connected to the softmax layer, which consists of two neurons. Each neuron in the softmax layer corresponds to the probabilities of each label for a MBTI category. The final prediction is made by selecting the label with greater probability.

5.2 Missing Data Problem

As mentioned before, one of the major challenges in multi-source temporal data analysis is the modeling of block-wise missing data. Indeed, the number of users in our dataset who participate in all three social networks simultaneously is relatively small: 702 users. Moreover, the number of users who, at the same time, has performed activities in all $k = 10$ time periods is only 128, which is not sufficient for effective LSTM training. To tackle this problem, we utilize non-negative matrix factorization (NMF) [13] separately for the data from each period of time. For every time interval we try to recover the missing data modalities of those users who have contributed to at least one social network during that time interval. After applying NMF we obtained in total 5001 users with the automatically filled matrix of activities in all social networks and in all time periods (both with the real data and after filling missing data with NMF). For our final experiments, we selected only this group of users.

6 Evaluation

To answer our proposed research questions, we carried out two experiments. The first experiment aims to evaluate the importance of temporal data utilization; while the second one compares the results obtained by models trained on different data source combinations.

For evaluation, we divided the dataset into training, test and evaluation sets. First, we selected all users with the activities in any social network in all $k = 10$ time intervals. Second, we divided users into training (70%) evaluation (10%), and test (20%) sets, preserving the original distribution of data among MBTI types and the level of user activity in three social networks.

Although we perform binary classification on each label in each category, we cannot prioritize the label in each category since they are not exactly the inverse of each other. Because of this, we use “Macro-F1” metrics for evaluation in our experiments. This metrics represents the averaged “Recall”, “Precision” and “F1” measures across two labels in each.

6.1 Model Training

In order to feed in the data into our LSTM neural network, for every user timeline in our dataset, we divided the data into $k = 10$ equal time periods of 142 days each. We then extracted the multi-source features for each time period as described in section above. We further defined a “window” as m consecutive time periods of user activity. We vary “window” size to find the best number of consecutive time periods for determining the dependencies between users' social activities and their personality and to test the sensitivity of time durations in inferring different personality concepts. We built an independent LSTM neural network for each of four personality categories. The trained LSTM models for all MBTI categories are then aggregated into the PersonalLSTM framework. It is noted, that the number of windows for train, test and validation sets for each *window* size is different. Let S be the number of windows for *window* os size m, then $S = k - m + 1$, where $k = 10$ is the total number of time periods in the dataset. From this formula, it follows, that the largest number of windows is for *window* size $m = 2$, while the smallest is for $m = 10$.

6.2 Baselines that does not consider temporal aspect

We selected Gradient Boosting, Logistic Regression, and Naive Byes classifiers as our non-temporal baselines. These approaches achieved high performance when solving the problem of relationship status prediction [2] based on the early-fusion multisource data (when feature vectors from data modalities are combined into one feature vector). The task of relationship status profiling was reported to be semantically similar to user personality profiling [2], which implies that the above algorithms can be used as strong baselines for our task. The features for non-temporal baselines were computed as described in Section above and based on the whole history of user activities in our dataset. The key idea is to train the baseline models based on the whole period of data (from 1 January 2013 to 31 December 2016) and then compare them to our proposed temporal approach, which also utilizes the whole data, but divided the data into 10 periods of time.

6.3 Evaluation against non-temporal baselines

First, we evaluated the performance of PersonaLSTM framework for different window sizes. We varied the window size m from 2 to 10 and trained LSTM network for each window size.

For each personality attribute the best performing window size is different. For example, the window size $m = 3$ (3 time periods of 142 days each, approximately 14 months) demonstrated the best performance for predicting MBTI category “extraversion – introversion”. This category is related to the so-called “energy and motivation resources” for people of different types [18]. The small window size of the best-performing model could indicate that the “energy” acquisition resources do not depend on the temporal data aspect.

At the same time, the category “sensing – intuition” can be accurately predicted with the window size $m = 9$. We hypothesize that it is because the category depends on behavior patterns over long periods of time. The above observation is consistent with the description of “sensing – intuition” category in literature. Specifically, Mayers et al. (myers1985manual) reported that “sensing” people pay attention to physical reality, while intuition people pay attention to impressions and the meaning of the information.

The best performing window size for the category “thinking – feeling” is $m = 7$ (approximately 2.7 years). Mayers et. al. (myers1985manual) mentioned that people of “feeling” type tend to be more aware of other humans’ feelings and can “...relate more consistently well to people” [18]. At the same time, as noted in Nasca et al. (nasca1994impact), “thinking” people are considered to be less emotional. From the above definitions, it follows that the prediction results of “thinking – feeling” category can be explained by their dependence on users’ reactions to surrounding life events.

Finally, for the category “judging – perception”, the optimal window size is $m = 5$ (approximately 2 year interval). Based on Myers et al. (myers1985manual) definition, “judging – perception” category separates judging people and perceiving individuals. At the same time, from Nasca et al. (nasca1994impact) it follows that people of “judging – perception” category “...do not show any particular relationship with communication apprehension” [19]. The above suggests that there are no strict communications patterns in social media data for predicting “judging – perception” category. At the same time, the results reveal that the temporal data helps to increase the quality of the predictions, which suggests that the temporal-enabled models can identify more complex behavior patterns that are not directly related to user conversations.

Based on the above results, we selected window sizes for temporal models that demonstrated the best results for each MBTI category and included them in the final configuration of PersonalLSTM framework. We then compared them with non-temporal baselines. The obtained results are presented in Table 2. From the Table, it can be seen that PersonalLSTM outperformed non-temporal baselines for "sensing – intuition" and "judging – perception" categories. At the same time, "introversion – extroversion" and "thinking – feeling" prediction performance is lower than those obtained by Gradient Boosting. Our obtained results are also consistent with Myers's definitions [18]. Precisely, "sensing – intuition" and "judging – perception" categories describe the "...reactions to different life changes" [18], which goes well with the temporality of social media data. At the same time, "introversion – extroversion" and "thinking – feeling" categories are more about life perception, which requires longer periods of data to be feed into personality profiling models. Based on the experimental outputs, we can **positively answer to our second research question**[9]. Specifically, we claim that it is possible to improve the performance of automatic personality profiling by leveraging on temporal data aspect for two MBTI categories: "sensing – intuition" and "judging – perception".

Table 2. Results obtained by non-temporal baselines and PLSTM framework. Evaluation metrics: macro F_1

	Gradient Boosting	Logistic Regression	Naive Bayes	LSTM
E/I	**0,715**	0,600	0,440	0,541
S/N	0,495	0,425	0,33	**0,724**
T/F	**0,670**	0,435	0,440	0,534
J/P	0,510	0,395	0,470	**0,750**

6.4 Evaluation Against Data Source Combinations

To understand the importance of different data sources for personality profiling as well as to answer our first and third research questions, in this section we compare the results obtained by PersonalLSTM trained based on different data source combinations.

The evaluation results are reported in Table 3. From the table, it can be seen that the incorporation of single data source always demonstrates lover prediction performance, as compared to multi-source utilization. Among the single data sources, the location data demonstrates the best performance for all categories except "sensing – intuition", which can be better predicted by leveraging textual data. These results are consistent with the definition of "sensing – intuition" type. Specifically, the category characterizes the way users receive new information from the outside world [18], which conforms well with the richness of the textual data modality.

From the results of models that were trained on bi-source combinations, it can be seen that the best performance (ranging from 49.5% to 59.6%) for all MBTI

[9] Is it possible to improve the performance of user personality profiling by leveraging on temporal data aspect?

Table 3. The results for combination of data of different modalities, the best window size is shown in brackets. Evaluation metrics: Macro F_1.

	E/I	S/I	T/F	J/P
Text (T)	0,362 (2)	0,456 (8)	0,380 (6)	0,469 (9)
Media (M)	0,349 (2)	0,409 (5)	0,493 (9)	0,472 (7)
Location (L)	0,511 (4)	0,43 (4)	0,493 (9)	0,494 (5)
T, M	0,349 (3)	0,594 (3)	0,465 (9)	0,543 (2)
M, L	0,511 (4)	0,605 (5)	0,496 (4)	0,595 (3)
T, L	0,521 (5)	0,618 (3)	0,495 (3)	0,596 (7)
T, M, L	**0,541 (6)**	**0,724 (9)**	**0,534 (7)**	**0,750 (5)**

categories was achieved by text and location combination. This is consistent with the results of models training on single modality data, where text and location were found to be the best among all single-source baselines. Thus, the results of experiments with a single source and bi-source data make it possible to **answer our third research question**[10] **that text and location modalities contribute the most to the task of automatic user personality profiling**.

Finally, it is noted that PersonalLSTM trained based on all three modalities, outperformed all other multi-source baselines, which **positively answers to our first**[11] **research question**. Specifically, we would like to highlight that that **the incorporation of multiple sources results from 28% boosting of personality profiling performance, as compared to single-source utilization**. It is also worth mentioning that the best-performing window size varies from 5 to 9, which is the period from 1.5 years to 2.5 years. Such differences in the window size may be caused by differences in the essence of the each of the MBTI categories. Since each of MBTI categories characterizes the disjoint personality traits, the best time interval for determining each of MBTI category is also different.

6.5 Full MBTI Type Prediction

In this experiment, we aim to predicted the full MBTI type of users. First, we predicted each of the four MBTI categories separately and then combined the predicted categories into the full profile. The achieved results showed an accuracy of 20.3%. Even though these results are better than random prediction, their quality is insufficient for use in real-life settings. For real life applications, such as marketing and recommendations it is more reasonable to try to improve the quality of predictions of each MBTI category separately.

7 Conclusions and Future Work

In this work, we presented the first study of learning from multimodal temporal data of the task of automatic user personality profiling. Our proposed PersonalLSTM framework consists of multiple LSTM models trained based on temporal

[10] Which data sources contribute the most to user personality profiling?

[11] Is it possible to perform user personality profiling more accurately by learning from multiple incomplete data sources?

data from three social networks (Twitter, Instagram, Foursquare). The evaluation results demonstrate that for two MBTI categories the incorporation of temporal model increases the quality of automatic personality profiling, while the two MBTI categories can be better predicted by non-temporal models. At the same time, in all cases when the models are trained based on multiple data sources, the personality profiling performance is significantly better as compared to the single-source baselines. To facilitate further research, we released our multi-source multimodal temporal dataset for public use.

Our further research will include: (1) the extension of our current dataset; (2) the incorporation of new image features that will include visual sentiment estimations; (3) the incorporation of more detailed temporal data; and (4) the corresponding temporal regularization of PersonalLSTM framework.

References

1. Adali, S., Golbeck, J.: Predicting personality with social behavior. In: Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on. pp. 302–309. IEEE (2012)
2. Buraya, K., Farseev, A., Filchenkov, A., Chua, T.S.: Towards user personality profiling from multiple social networks. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (2017)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 248–255. IEEE (2009)
4. Farseev, A., Akbari, M., Samborskii, I., Chua, T.S.: 360 user profiling: past, future, and applications. ACM SIGWEB Newsletter (Summer), 4 (2016)
5. Farseev, A., Nie, L., Akbari, M., Chua, T.S.: Harvesting multiple sources for user profile learning: a big data study. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. pp. 235–242. ACM (2015)
6. Farseev, A., Samborskii, I., Chua, T.S.: bbridge: A big data platform for social multimedia analytics. In: Proceedings of the 2016 ACM on Multimedia Conference. pp. 759–761. ACM (2016)
7. Farseev, A., Samborskii, I., Filchenkov, A., Chua, T.S.: Cross-domain recommendation via clustering on multi-layer graphs. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 195–204. ACM (2017)
8. Gers, F.A., Schraudolph, N.N., Schmidhuber, J.: Learning precise timing with lstm recurrent networks. Journal of machine learning research **3**(Aug), 115–143 (2002)
9. Golbeck, J., Robles, C., Turner, K.: Predicting personality with social media. In: CHI'11 extended abstracts on human factors in computing systems. pp. 253–262. ACM (2011)
10. Huang, M., Cao, Y., Dong, C.: Modeling rich contexts for sentiment classification with lstm. arXiv preprint arXiv:1605.01478 (2016)
11. Kaplan, A.M., Haenlein, M.: Users of the world, unite! the challenges and opportunities of social media. Business horizons **53**(1), 59–68 (2010)
12. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences **110**(15), 5802–5805 (2013)

13. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
14. Liu, F., Perez, J., Nowson, S.: A language-independent and compositional model for personality trait recognition from short texts. arXiv preprint arXiv:1610.04345 (2016)
15. Martin, C.R.: Looking at type: The fundamentals. Center for Applications of Psychological Type (1997)
16. Mattare, M.: Revisiting understanding entrepreneurs using the myers-briggs type indicator®. Journal of Marketing Development and Competitiveness **9**(2), 114 (2015)
17. McCrae, R.R., Costa, P.T.: Personality in adulthood: A five-factor theory perspective. Guilford Press (2003)
18. Myers, I.B., McCaulley, M.H., Most, R.: Manual, a guide to the development and use of the Myers-Briggs type indicator. Consulting Psychologists Press (1985)
19. Nasca, D.: The impact of cognitive style on communication. NASSP Bulletin **78**(559), 99–107 (1994)
20. Nie, L., Zhang, L., Wang, M., Hong, R., Farseev, A., Chua, T.S.: Learning user attributes via mobile social multimedia analytics. ACM Transactions on Intelligent Systems and Technology (TIST) **8**(3), 36 (2017)
21. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates **71**(2001), 2001 (2001)
22. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: Our words, our selves. Annual review of psychology **54**(1), 547–577 (2003)
23. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: CLEF. sn (2015)
24. Rushton, S., Morgan, J., Richard, M.: Teacher's myers-briggs personality profiles: Identifying effective teacher personality traits. Teaching and Teacher Education **23**(4), 432–441 (2007)
25. Saha, A., Sindhwani, V.: Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In: Proceedings of the fifth ACM international conference on Web search and data mining. pp. 693–702. ACM (2012)
26. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one **8**(9), e73791 (2013)
27. Song, X., Nie, L., Zhang, L., Akbari, M., Chua, T.S.: Multiple social network learning and its application in volunteerism tendency prediction. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 213–222. ACM (2015)
28. Verhoeven, B., Daelemans, W., Plank, B.: Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In: 10th International Conference on Language Resources and Evaluation (LREC 2016) (2016)
29. Wei, H., Zhang, F., Yuan, N.J., Cao, C., Fu, H., Xie, X., Rui, Y., Ma, W.Y.: Beyond the words: Predicting user personality from heterogeneous information. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. pp. 305–314. ACM (2017)
30. Youyou, W., Kosinski, M., Stillwell, D.: Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences **112**(4), 1036–1040 (2015)