Why Is MBTI Personality Detection from Texts a Difficult Task?

Sanja Štajner, Seren Yenikent
Symanto Research
Nuremberg, Germany
{sanja.stajner, seren.yenikent}@symanto.com

Abstract

Automatic detection of the four MBTI personality dimensions from texts has recently attracted noticeable attention from the natural language processing and computational linguistic communities. Despite the large collections of Twitter data for training, the best systems rarely even outperform the majority-class baseline. In this paper, we discuss the theoretical reasons for such low results and present the insights from an annotation study that further shed the light on this issue.

1 Introduction

Apart from being long and requiring to be administrated by a skilled human assessor in artificial circumstances (laboratory conditions), the traditional questionnaire-based personality tests often introduce *social desirability bias* (Krumpal, 2011) and *the reference-group effect* (Heine et al., 2002), which can be introduced either from the subject's or assessor's side. To avoid those biases, it was suggested that analysing person's writing provides more objective assessment of one's personality than the traditional questionnaires (Stachl et al., 2019).

1.1 The MBTI

The original MBTI model was based on the comprehensive theoretical work of Carl Jung (1921), and was further developed by Myers and Briggs by adding the fourth dimension (judgment/perception) and several decades of extensive practical use within the industrial and educational settings (Briggs-Myers and Myers, 1995). Today, the MBTI is one of the most widely used non-clinical psychometric assessments, regularly used in understanding team building processes in work environments (Kuipers et al., 2009), for career suggestions (Garden, 1997), and in marketing and consumer behavior (Gountas and Gountas, 2001).

The MBTI lays out a binary classification based on four distinct functions, and draws the typology of the person according to the combination of those four values (e.g. INFP, ESTJ):

- **E**xtraversion/**I**ntroversion (EI) - preference for how people direct and receive their energy, based on the outer or inner world
- **S**ensing/**IN**tuition (SN) - preference for how people take information in, by five senses or by interpretation and meanings
- **T**hinking/**F**eeling (TF) - preference for how people make decisions, by relying on logic or emotions towards people and special circumstances
- **J**udgment /**P**erception (JP) - how people deal with the world, by organizing it or staying open for new information

While many studies investigated linguistic characteristics of the Big 5 personality traits (Mairesse et al., 2007; Furnham, 1990; Pennebaker and King, 1999; Gill and Oberlander, 2002; Scherer, 2003; Pennebaker and King, 1999; Gill and Oberlander, 2003), to the best of our knowledge, there have been no studies reporting on linguistic characteristics of different MBTI types. The most probable reason for this is a different nature of the MBTI framework. Unlike the Big 5 model that originated from lexical analyses (Cattell, 1946; Tupes and Christal, 1961; Goldberg, 1982; Costa and McCrae, 1992), the MBTI fundamentally makes use of the behavioral implications in theoretical and professional contexts. Hence, the available data rarely refers to any linguistic contexts, but more to practical results of the questionnaires. However, as linguistic data has been shown as one of the best indicators of personality-related characteristics such as behaviors and motivations (Pennebaker

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3580–3589
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

and King, 1999; Tausczik and Pennebaker, 2010) and given that the MBTI makes use of behavioral implications to study personality, there is a room for linguistic representation of the concept with proper datasets and methods.

1.2 Automatic Detection of MBTI from Texts

Unlike automatic detection of the Big 5 personality traits that in the last 15 years has been attempted at from various types of texts, e.g. essays (Argamon et al., 2005), personal weblogs (Oberlander and Nowson, 2006), and Facebook posts (Kosinski et al., 2013), the automatic detection of MBTI gained popularity only recently and exclusively using Twitter data. Attempts were made for various languages: English (Plank and Hovy, 2015), six Western European languages (Verhoeven et al., 2016), and Japanese (Yamada et al., 2019). All those studies, despite using large training datasets (over 1M instances) and various features (word and character n-grams, or count-based meta-features such as number of followers, favourites, etc.), barely managed to outperform the majority-class baseline, and even that only in some of the four MBTI dimensions. The best English models were trained on over 1M Twitter instances using logistic regression classifier and combining linguistic and count-based meta-features. Nevertheless, they outperformed the majority-class baseline only on the IE and TF dimensions, achieving the accuracy of 72.5% and 61.5% on those binary tasks (Plank and Hovy, 2015).

Comparison of performances of the Big 5 and MBTI computational models trained on Twitter data showed that type of architecture and settings practically have no influence on the MBTI detection from such data (Celli and Lepri, 2018) indicating thus that Twitter data might not contain sufficient amounts of lexical signals.

1.3 Goals and Contributions

The main goal of our study is to investigate why the automatic detection of MBTI personality traits from texts does not outperform even the simple majority-class baseline and why its performance on Twitter data is not influenced by the architecture type and settings. Furthermore, we shed some light on the natural complexity of the task and discuss the theoretical constraints of the task. We pose two hypotheses:

- H1: *Twitter data does not contain enough signals for MBTI personality detection.*
- H2: *Textual data does not resonate well with MBTI personality scores from questionnaires.*

To test those hypotheses, we collect a new dataset, write the guidelines for human annotation, conduct an extensive human annotation task, and provide both a quantitative and qualitative analysis of the results.

The main contributions of our study are the following:

1. Proposing the guidelines that translate behavioural characteristics of the MBTI dimensions into linguistic cues in texts;
2. Releasing a new dataset for MBTI analysis from textual data;
3. Testing the two above-mentioned hypotheses (H1 and H2);
4. Revealing the specificities of the MBTI constructs that make them difficult for automatic detection from texts.

2 Datasets

To test our hypotheses, we used two datasets, a subset of the MBTI-Twitter dataset (Plank and Hovy, 2015), and a dataset that we collected especially for this purpose, the MBTI-MTurk dataset.[1]

2.1 MBTI-Twitter Dataset Selection

To test our first hypothesis (H1), we randomly selected 96 user data from the MBTI-Twitter dataset (Plank and Hovy, 2015), six for each of the 16 MBTI types, using the version with 50 concatenated tweets for each user/instance. Out of those 50 concatenated tweets, we only retained the first 10 tweets for two reasons: (1) to maintain the task managable for human annotation; and (2) to have the length of posts from each user roughly comparable to those we collected (MBTI-MTurk).

2.2 MBTI-MTurk Dataset Compilation

Our goal was to compile an 'ideal' dataset of short text snippets that would maximize the potential of finding linguistic signals of the four MBTI personality dimensions.

[1] Available upon request for research purposes.

We posted a human intelligence task (HIT) using the Amazon Mechanical Turk (MTurk) crowdsourcing platform. The task consisted of three open-ended questions:

1. You might have done an MBTI personality test in the past. If you did, and you know the MBTI personality type you obtained, please write it here:
2. What is your favourite type of vacation and why?
3. Which are your favourite hobbies and why?

The first question was optional, while the other two questions were required for completing the task and getting monetary compensation. For the last two questions, we enforced the answers to be at least 300 characters long, to allow for a sufficient length to capture some linguistic signals. The last two questions were selected following the assumption that how people spends their free time would be the most natural version of their personality. Research suggests that activities engaged during leisure time are the outcomes of what the person wants to be doing by choice with high levels of intrinsic motivation (Kuykendall et al., 2015). Therefore, we considered that the types of vacation and hobbies are representative realms of the participants' true personalities.

Data Collection. We posted our HIT in 20 batches spanning a two-months period. From all collected data, we first filtered out those that did not contain the answer to the MBTI personality type question.[2]

Quality Control. All collected answers were manually checked to filter out those answers that were copied from the internet (about one third of all participants tried to trick the system by copying texts from internet about favourite types of vacations and hobbies).

Consistency Check. Out of all HITs that contained the answer to the MBTI question, we only retained those for which we had two HITs from the same user completed with at least one month in between and where the asnwer to the MBTI question contained the same personality type both times. The assumption was that if a user provided two different MBTI results in those two instances it was for one of the following reasons: (1) the user gives random MBTI types to intentionally harm the study (Ipeirotis et al., 2010), or (2) in at least one of the MBTI dimensions, the user is somewhere in the middle range, and might have genuinely obtained two different scores on an MBTI questionnaire.

Final Dataset. The final dataset for this study consisted of 96 HITs, all completed by different users (MTurk IDs)[3]. Similar as in the case of the MBTI-Twitter dataset, the selected dataset comprises of equal number of HITs (six) per each of the 16 MBTI personality types. For the two pilot rounds, the additional 30 HITs were used, ensuring that each of the eight polarities (extravert, introvert, sensing, intuitive, etc.) is present at least twice.

3 Annotation

As there are no studies laying out linguistic characteristics of the MBTI model, we created the annotation guidelines starting from the behavioral characteristics. The MBTI's theoretical and practical framework provides detailed profiles for each type ranging from general and typical characteristics to real-time depictions of those characteristics, in the contexts of relationships, marriage, work, and learning settings (Briggs-Myers and Myers, 1995). We translated this behavioral information into linguistic and textual signals in general, as well as some specific signals relevant for the contexts we asked the participants to write about.

To define the characteristics that would linguistically distinguish the two polarities of the EI dimension, we focused on the processes of socialization depicted in the text. As the dimension of JP is mainly related to organizational preferences, we chose the linguistic signals referring to arrangement of schedules and plans. The fact that people high on J tend to be more exact and pay more attention to details, whereas people high on P show more flexible and accidental characteristics (Briggs-Myers and Myers, 1995) we translate into guidelines considering sentence structures and grammar. The SN dimension identifies preferences for the characteristics of tasks and information people would like to process, either using facts and five senses (S), or using imagination and abstraction

[2] We opted for making the MBTI personality type question optional to avoid people writing random MBTI types, as those are widely known and popular on the internet. By paying equally those who provided the answer to the MBTI question and those who did not, we tried to ensure that provided MBTI types are not just randomly chosen but rather represent the real results of the user's MBTI testing.

[3] From each user ID, that passed all above-mentioned checks, we randomly chose one answer about the hobbies and one answer about the vacations.

Extravert	Introvert
Mention of new people (e.g. *crowd*, *strangers*)	Mention of closer people rather than any group of people (e.g. *husband*, *family*)
Mention of social activities and events that contain interaction with other groups of people (e.g. *party*, *dancing*, *couchsurfing*)	Mention of individual activities or activities that can be done without interaction with other people (e.g. *by myself*, *spending time at home*)
Mention of outside world and vibrant places (e.g. *bars*, *restaurants*)	Mention of inner world, and calm and quiet places (e.g. *home*, *museum*)
We references	*I* references
Use of intensifiers and exclamation marks	Hedging
More assertive, positive, enthusiastic arguments	Less assertive arguments

Table 1: Linguistic signals of extraversion and introversion.

Sensing	Intuitive
Technical, object-based and hands-on hobbies	Inspirational and imaginative hobbies (e.g. *creating*, *exploring*)
Facts and real cases (e.g. *documentary*, *diary*)	Abstraction rather than facts (e.g. *sci-fi*, *cartoons*)
Details and examples (more adjectives and adverbs to provide details, use of the words *example*, *for instance*)	Main ideas rather than details
Needs to use the 5 senses	Needs to focus on the bigger picture
Puzzles, model planes, crafts, carving, rowing, sailing, diving, rock climbing, etc.	Painting, music, dancing, poetry, chess, literature, arts, martial arts, yoga, meditation, etc.
Simplified and straightforward writing style (short sentences)	Complex writing style (long sentences)
Clear and concise writing style	Artistic, longer, more words

Table 2: Linguistic signals of sensing and intuition.

Thinking	Feeling
Logical reasoning for their actions and choices (e.g. *reading books for learning*)	Emotional reasoning for their actions and choices (e.g. *reading books for gateway feeling*)
Mention of opinions, ideas, comparisons	Mention of people, values, feelings
Direct (e.g. *reading is nice*)	Tactful, indirect (e.g. *reading feels nice*)

Table 3: Linguistic signals of thinking and feeling.

Judging	Perceiving
Holidays that include planning such as ski holidays, city tours etc. (e.g. *tour*, *pass*, *ticket*, *reservation*)	Spontaneous holidays such as going to the beach, a new city etc. (e.g. *flexible*, *spontaneous*)
Decisive, planful, organized (e.g. *plan*, *schedule*, *followed by*)	Curiosity, anticipation of change, and spontaneity
Organizers of the plans (e.g. *invite*, *organize*)	Followers of the plans (e.g. *join*, *tag along*)
Warranty (e.g. *insurance*, *make sure*)	Autonomy and impulsiveness (e.g. *suddenly*, *out of the blue*, *last minute*)
Past tense or present perfect tense	Present simple tense
Formal and structured writing style with grammatical rules followed as much as possible (e.g. *I like ski holidays and sometimes prefer city tours.*)	Informal writing style with grammar mistakes (e.g. *I like going to the beach. Also, do art sometimes.*)

Table 4: Linguistic signals of judging and perceiving.

(N). The TF dimension is related to preferences for decision-making processes. Hence, we focus on the reasoning aspect of the linguistic characteristics, whether it is logical or emotional.

Tables 1– 4 provide linguistic signals for each label. They were provided to the hired annotators as the main annotation guidelines.

3.1 Annotators

Two annotators, one with a PhD degree in psychology and the other in computational linguistics, were hired to annotate all instances in both datasets. Both annotators underwent a six-months (paid) training which covered reading the extensive literature on personality assessment (e.g. traditional questionnaire approaches and dictionary-based methods), the MBTI framework and its use cases.

3.2 Annotation Procedure

We asked the annotators to assign to each instance (either coming from the MBTI-Twitter dataset or from Hobbies and Vacation questions in the MBTI-MTurk dataset)[4], for each of the MBTI dimensions separately, one of the following four labels: either of the two polarities (e.g. E or I for the EI dimension) etc.), *unsure* (in cases where they saw signals from both polarities and are not confident to make a binary decision), or *not enough signal* (in cases where they did not find any signals for any of the two polarities).

Both annotators were first asked to complete two pilot rounds of annotation so that they can be fully familiarised with the annotation guidelines and the procedure. In each pilot round, the dataset they were ask to annotate consisted of 15 MBTI-Twitter and 15 MBTI-MTurk user-instances which were not used for the final round. After each pilot round, a question and answering session was organized to address all potential issues with the guidelines or the procedure. Furthermore, after each pilot round, the annotators showed their annotations and commented their decisions to the other annotator to calibrate their annotations. During the pilot rounds, the annotators were asked to mark the parts of the instances which guided their decisions for assigning certain polarity.

The annotations and text mark-ups obtained during the pilot rounds are used in Section 4 to show how the proposed guidelines were used in practice, as well as to point out the most challenging aspects of the annotation process.

After finishing both pilot rounds, the annotators were given the final dataset which consisted of 96 user-instances from the MBTI-Twitter dataset and 96 user-instances from the MBTI-MTurk dataset. In total, each annotator annotated 192 user-instances for each of the four MBTI personality dimensions. The annotators were instructed to have enough breaks to avoid the fatigue effect.

4 Findings from the Pilot Rounds

The annotation and mark-up obtained during the pilot rounds revealed several important characteristics of the task.

4.1 Middle Cases

One of the recurring issues found during the pilot rounds was the case of people whose answers truly belong to the middle of the spectrum, as in the following case (the signals for introversion are shown in italics, those for extraversion are shown in bold, and those for the JP dimension are shown underlined):

(1) “I like travelling to some new places. *Mostly* I like travelling to a city I have never been before, but from time to time I *can* also enjoy just going to a nice seaside place and relax for a week without any fixed plans and sightseeing schedules. The only type of holidays **I really don’t like** is **just staying at home**. I find that OK if just for a day or two maximum, but longer than that **I get bored**.”

The words *mostly* and *can* represent hedging and are thus signals for introversion (Table 1), whereas the word *really* is an intensifier and thus a signal for extraversion (Table 1). The negation of the fixed plans and schedules signalizes a truly mid-range personality along the JP dimension, as it shows that the person does not like fixed plans and schedules, but is still aware of their existence and thus mentions them.

The ‘middle cases’ were frequent also for the other two dimensions. The following example was annotated as *unsure* for the SN dimension by both annotators (the signals for sensing are shown in bold, and the signals for intuition are shown in italics):

[4] The answers to the Hobbies question were used for annotating only SN and TF dimensions, and the answers to the Vacations question were used for annotating only EI and JP dimensions.

(2) “I enjoy doing sports, although its been a while since I have been active. Such like **volleyball**, **gym fitness**, and I am trying to get into *yoga*. On the complete opposite side, I also really love *baking and exploring new recipes to cook*. In my spare time I am trying to learn to appreciate being outdoors more, as I usually spend too much time binging netflix.”

The noticeable amount of such ‘middle cases’ is, however, not caused by the flaw of the guidelines, but rather reflects the drawbacks of the binary nature of the MBTI framework. Those people who have characteristics of both polarities, which is a common case (Pittenger, 1993), are by the traditional questionnaire-based assessment placed in one of the two groups (e.g. extravert or introvert). Analysis of the short posts can reveal the presence of signals for both polarities but their ratio in the short posts might be different from that obtained by the questionnaire-based assessment.

4.2 Content vs. Style

As can be seen in Tables 1– 4, the annotation guidelines contain pointers regarding both content (i.e. lexical choices and argumentation) and style (i.e. grammatical and stylistic preferences). The pilot phases discovered that in approximately 10% of the cases, a given instance shows the content-based signals of one polarity and the stylistic-based signals of the opposite polarity. One such example is the following, where the content cues are clearly of the judging polarity (shown in italics), while the stylistic cues (sentence structure, grammatical errors, and typos) are of the perceiving polarity (shown in bold):

(3) “I *would like my vacation to be well organised* so **i wont** have to deal with **anything just** enjoy the flow. I like visiting new places, that is **me** perfect vacation. **Its** not interesting for me going several times in one place just because **its quite**. **Don’t** like big buildings, crowded places and cemented environments**, i** like nature and historical places”

In all such cases, the polarity that was signalized by the stylistic preferences was in line with the reported official MBTI label. It is known that people can consciously change the content of their answers, but not the style (Chung and Pennebaker, 2007). Therefore, the content of the answers to the open-end questions can also suffer from social desirability bias, similar as the answers to the questionnaires.

4.3 Writing Style in Twitter

Another issue that was revealed during the pilot rounds was that twitter posts, by their nature, often have incorrect grammar, punctuation, and ill-formed sentences. That makes it difficult for the annotators to assign J or P labels unless there are lexical cues (which are extremely rare in Twitter), because otherwise, focusing on writing style (Table 4) the great majority of posts would be annotated as P. Similarly, the nature of the Twitter posts to overuse intensifiers and exclamation marks makes many posts stylistically extraverted, and thus, the last two points in the annotation guidelines (Table 1) may lead to false (extravert) positives.

Two examples of instances from the MBTI-Twitter dataset are given later, in Table 6.

5 Final Annotation Results

The annotation of the final dataset that consists of 96 instances from the MBTI-Twitter dataset and 96 instances from the MBTI-MTurk dataset revealed further particularities of each dataset, and shed some light on how feasible the task is, even for trained human annotators.

5.1 The Presence of the MBTI Signals

As expected, in the final datasets, the instances from the MBTI-MTurk dataset were reported to have enough signals across all four MBTI dimensions, as opposed to the instances in the MBTI-Twitter dataset (Table 5). In the MBTI-Twitter dataset, many instances were reported to have insufficient signals across the JP and SN dimensions.

These results support our first hypothesis (H1) that Twitter posts (even when grouped per user) do not always contain linguistic signals to allow for personality detection, even for the trained human annotators. Furthermore, these results give a potential explanation to the question why previously proposed binary classification models for the JP and SN dimensions did not manage to outperform even the majority-class baselines.

In Table 6, we present examples of two Twitter users, one for whom both annotators reported insufficient signals for the JP and SN dimensions, and another one whose personality was correctly annotated across all four dimensions by both annotators.

Annotator	Statistic	MBTI-Twitter				MBTI-MTurk			
		EI	SN	TF	JP	EI	SN	TF	JP
A (CL)	Not enough signal	2	31	8	59	0	0	0	0
	Unsure	14	14	22	7	31	23	38	33
	Confident (class assigned)	84	55	71	33	69	77	62	67
B (Psychologist)	Not enough signal	10	17	10	33	0	0	0	0
	Unsure	29	17	19	25	23	39	54	7
	Confident (class assigned)	61	67	71	42	77	61	46	93

Table 5: Human annotation statistics (in percentage of cases, out of the total number of 96 instances) on the MBTI-MTurk dataset (Holidays and Hobbies) and the subset of the MBTI-Twitter dataset.

Insufficient signals for any dimension	Correctly annotated (ESFP) by both annotators
“what happened in the fandom why are the makorra shippers so angry ? @URL / 5mfxado0lp photoset : calmorrison : aer-dna : sweetlikepoison 528 : aer-dna : ” just the four of us . ” remember when team ... @URL / hdcrf 7ljsc omg remind me to not go into the tags i just saw a post that essentially said makorra was baited like how ... @URL / jxdtfcmlld dylanftsw : shut the fuck up about “ shitty writing ” or the end of the show not making sense or being rushed .... @URL / r62qnjt7di photo : hellkatespangled : sleepy girlfriends and lazy art @URL / 5oq9n8P2Kd wall-maria-around-ba-sing-se : so i saw this gif : and all i could think was how it made her look like an ... @URL / gom 5newfum photo : insomniadiesdown : some kuvira sketches @URL / ahrgfwwqsb photo : joeldosreisviegas : dariucdraws :...”	“omg i just met doug adams from season 12 of top chef , here in portland , the dude was super mellow and friendly ! :D @USER craaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaling in my skiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii @USER he was super chill . i shook his hand and told him i loved his attitude and was rooting for him , and he was just like , ” cool ! ” @USER hey cat , you guys get home okay ? @USER i sure hope so ! i’ve loved doug since the beginning of this season his attitude is great , he’s talented , he never whines @USER conversion ” therapy ” is indeed the greatest sham it’d be an amazing step to see it banned , and lgbt’s treated normally @USER it’s essentially the last bastion of clinical homophobia the day it dies is the day attitudes towards us improve drunk tonight and just can’t get over how awesome bjork is <3 ...”

Table 6: An MBTI-Twitter instance which did not contain sufficient signals and could not be annotated (left) as opposed to another MBTI-Twitter instance which was correctly annotated by both annotators across all four dimensions (right). In both cases, we do not present the full instance for the space constraints.

5.2 Assigned Classes

We found that in the MBTI-MTurk dataset, a relatively high number of instances (ranging between 7% and 54% depending on the dimension and annotator, with most of them being in the range between 23% and 39%) was reported to have mixed signals (labelled as *unsure*).

As expected, the number of instances for which the annotators felt confident enough to assign one of the polarities was higher in the MBTI-MTurk than in the MBTI-Twitter dataset, in most of the cases. The difference was most noticeable in the JP dimension, where the number of instances with an assigned polarity doubled.

A closer look at the instances for which both annotators were confident enough to assign either of the polarities revealed even more prominent differences between the two datasets (Table 7). In the MBTI-Twitter dataset, only in 46% of the instances the annotators agreed on the EI dimension, and only in 50% cases they agreed on the JP dimension. In the MBTI-MTurk dataset, in contrast, the annotators agreed across those two dimensions (EI and JP) in 100% and 78% of the cases, respec-

Statistic	MBTI-Twitter				MBTI-MTurk			
	EI	SN	TF	JP	EI	SN	TF	JP
Both annotators confident	53	**30**	43	**15**	62	**54**	38	**69**
Annotators agree	**46**	*81*	61	**50**	**100**	*69*	62	**78**
Annotator A agrees with the gold label	77	64	64	53	78	54	77	44
Annotator B agrees with the gold label	47	44	54	47	60	54	85	42
Both annotators agree with the gold label	77	54	57	50	75	62	54	43

Table 7: Inter-annotator agreement statistics on the MBTI-MTurk dataset (Holidays and Hobbies) and the subset of the MBTI-Twitter dataset. The percentage of instances, out of those for which both annotators were confident to assign a polarity, for which both annotators assigned the same polarity is presented in the row "Annotators agree". Similarly, the next two rows present the percentage of cases in which each annotators label was equal to the gold label (the provided MBTI label) out of all cases for which the annotator was confident to assign a polarity (as opposed to the other two labels: 'unsure' and 'not enough signal'). The percentage of cases in which both annotators agree with the gold label is calculated taking into account only those instances on which both annotators agreed.

tively. Nevertheless, if we calculate the percentage of cases in which the 'gold' label was the same as the shared label of the two annotators (the row 'both annotators agree with the gold label' in Table 7), we find that, surprisingly, it only happens in up to 50% of the cases for the JP dimension. This indicates that the linguistic signals for the JP dimension captured from text, even with high confidences of the annotators and their inter-annotator agreement, are not correctly associated with the results of the traditional questionnaire-based MBTI personality assessment. The only MBTI dimension for which the percentage of cases in which both annotators agreed with the gold label is noticeably above the majority-class baseline (50%) in both datasets is the EI dimension. These results indicate that the EI dimension is the only dimension for which a noticeable association between the results of the traditional questionnaire-based MBTI personality assessment and the textual-analysis-based MBTI personality assessment was found. This evidence supports our second hypothesis (H2) for three out of four MBTI dimensions.

6 Discussion and Conclusions

The results of presented analyses shed some light on possible causes of the poor performances of the automatic MBTI personality detection systems proposed so far. Furthermore, they indicate that the linguistic cues found in short texts do not seem to directly correspond to the results of the questionnaire-based results, which are commonly used as the 'gold labels' in classification experiments. This is in line with academic studies that showed the psychometric inadequacy of the questionnaire (Pittenger, 1993; Boyle, 1995).[5]

The high number of instances without enough signal and with mixed signals in the MBTI-Twitter dataset across all MBTI dimensions, and especially the JP dimension, leads to high amounts of noise in the training datasets. Therefore, it is not a surprise that the best systems trained on those datasets rarely outperform even the majority-class baseline.

Even in the cases where user-instances contain sufficient signals and the agreement between the human annotators is high, the agreement between the annotators and the gold label is still very low for three out of four dimensions (Table 7). These results indicate that the constructs set up by the traditional questionnaire-based personality assessment might not have the exact equivalent translation into the linguistic cues. The only exception for this is the EI dimension where the agreement between the annotators and the gold label reaches 75%-77%. This is somewhat expected as the EI is the highest correlated dimension between the MBTI and Big 5 models amongst all dimensions (Furnham, 1996), and in the Big 5 model, the EI has a good linguistic correspondence.

Finally, the results of the in-depth analyses performed during the two pilot rounds (Section 4) revealed three phenomena that need to be taken into

[5] Due to the methodology followed to develop and improve the questionnaire (i.e. qualitative methods such as observations and introspection), the MBTI has received considerable criticism for not relying on a scientifically proven background (i.e. data-driven approaches).

account when using the linguistically-based MBTI personality analysis from texts:

(1) The content and the style of the text sometimes exhibit signals of the opposite polarities;

(2) Many people naturally express signals of the opposite polarities, as they probably belong to the middle ranges of those personality dimensions;

(3) The language used in Twitter shows specific stylistic characteristics in terms of tonality, use of exclamation marks, sentence structure and grammar, thus making everyone seem more extraverted than they are.

References

Shlomo Argamon, Sushant Dhawle, Moshe. Koppel, and James W. Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*.

Gregory J. Boyle. 1995. Myers-briggs type indicator (mbti): Some psychometric limitations. *Australian Psychologist*, 30(1).

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

Raymond B. Cattell. 1946. *The description and measurement of personality*. Yonkers-on-Hudson.

Fabio Celli and Bruno Lepri. 2018. Is Big Five Better than MBTI? A Personality Computing Challenge Using Twitter Data. In *CLiC-it*.

Cindy Chung and James W. Pennebaker. 2007. The psychological functions of function words. In *Social communication*, pages 343–359. New York: Psychology Press.

Paul T. Costa, Jr and Robert R. McCrae. 1992. *Revised NEO Personality Inventory (Neo-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual*. Psychological Assessment Resources.

Adrian Furnham. 1990. *Handbook of Language and Social Psychology*, chapter Language and personality. Winley.

Adrian Furnham. 1996. The big five versus the big four: the relationship between the myers-briggs type indicator (mbti) and neo-pi five factor model of personality. *Personality and Individual Differences*, 21(2).

Anna Garden. 1997. Relationships between mbti profiles, motivation profiles, and career paths. *Journal of Psychological Type*, 41.

Alastair J. Gill and Jon Oberlander. 2002. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363—-368.

Alastair J. Gill and Jon Oberlander. 2003. Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; neuroticism is a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 456—-461.

Lewis R. Goldberg. 1982. From Ace to Zombie: Some explorations in the language of personality. *Advances in personality assessment*, 1:203–234.

John Gountas and Sandra Gountas. 2001. A new psychographic segmentation method using jungian mbti variables in the tourism industry. *Consumer psychology of tourism, hospitality and leisure*, 2.

Steven J. Heine, Darrin. R. Lehman, Kaiping Peng, and Joe Greenholtz. 2002. What's wrong with cross-cultural comparisons of subjective likert scales?: The reference-group effect. *Journal of Personality and Social Psychology*, 82(6):903—-918.

Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*.

Carl G. Jung. 1921. *Psychological Types: Volume 6*. Routledge.

Michal Kosinski, David Stillwell, and Thore Graepell. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 15:5802–5805.

Ivar Krumpal. 2011. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality Quantity*, 47(4).

Ben S. Kuipers, Malcolm J. Higgs, Natalia V. Tolkacheva, and Marco C. de Witte. 2009. The influence of myers-briggs type indicator profiles on team development processes: An empirical study in the manufacturing industry. *Small Group Research*, 40(4).

Lauren Kuykendall, Louis Tay, and Vincent Ng. 2015. Leisure engagement and subjective well-being: A meta-analysis. *Psychological Bulletin*, 141(2).

François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Int. Res.*, 30(1):457–500.

Jon Oberlander and Scott Nowson. 2006. Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*.

James W. Pennebaker and Laura A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296–1312.

David J. Pittenger. 1993. Measuring the MBTI... and coming up short. *Journal of Career Planning and Employment*, 54:48–52.

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

Klaus R. Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40 (1-2):227—-256.

Clemens Stachl, Florian Pargent, Sven Hilbert, Gabriella M. Harari, Ramona Schoedel, Sumer Vaid, Sam Gosling, and Bühner Markus. 2019. Personality Research and Assessment in the Era of Machine Learning.

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

Ernest C. Tupes and Raymond E. Christal. 1961. Recurrent personality factors based on trait ratings. *USAF ASD Technical Report*, pages 61–97.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1632–1637, Portoroz, Slovenia. European Language Resources Association (ELRA).

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2019. Incorporating textual information on user behavior for personality prediction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 177–182, Florence, Italy. Association for Computational Linguistics.