Personality Detection on Persian Dataset

Phase 1

Mohammadjavad Mehditabar - 98522049

NLP Phase 1

May 27, 2023

Contents

1 Data Overview

Due to the lack of Persian datasets in the personality detection field, gathering data from an adequate source is the main part of our research. We came up with the idea of collecting data from publicly available sources, thanks to social media, which is the best and most applicable source. Because it contains a variety of people and has a wide range of data, which is useful for us to collect. As a result, we decided to choose Twitter as our data origin due to its popularity, accessibility, and also comprise a lot of texts. We have reached three different methods of collecting data

2 Data Processing

2.1 Bio Search

this dataset contains Iranian users with specified MBTI type of themselves in their bio. There were some difficulties in distinguishing the language of some bios. Hence they were solved by manual checking and some consideration of the Persian language. Labels of this data are based on mentioned types in their bio, which we recognized by searching 16 MBTI types in their bio and automatically detecting them.
This dataset consists of users' publicly available tweets and MBTI labels. Figure 1 shows a sample of our collected data based on a mentioned MBTI type on their bio.

Figure 1: A sample of bio search

2.2 Tweet Search

In this method, we tried collecting user tweets using a Python library called Twint. Through this approach, we have searched for 16 possible MBTI labels in users' tweets in which one of these labels is mentioned. Note that queried tweets are based on recent tweets in the Persian language. We noticed that The type users mentioned in their tweets, may be irrelevant to themselves, or they're saying some fact about this personality type or mentioning someone else's type. Moreover, after filtering those users, we performed some elimination in a total separate step for the users who have assigned themselves to multiple MBTI types. So these users are absolutely invalid for our dataset. Hence there

are chances that auto labeling makes a mistake in finding the correct label, thus users are labeled manually so that ambiguity will be resolved.
This dataset contains the gender of the user besides their tweets and MBTI label. Figure 2 shows a sample of our collected data based on a mentioned MBTI type on their tweet.

Figure 2: A sample of tweet search

2.3 Questionnaire

As our last method, we used a questionnaire to obtain more accurate data. We can consider this data as the golden data. The questionnaire was implemented in a web platform service with 60 standard well-known MBTI tests in which users filled their Twitter IDs and answered those questions. We received a few more data from users as well as their gender, location, degree of education, and age for analyzing and adding more features to our data to perform well on the dataset. Labeling them was based on their answer and how close they were to each of four binary class types (i.e. I/E, S/N, T/F, P/J), which were calculated after completing the questionnaire. Tweets of them are collected by the Python library Tweepy, which is a helpful tool to gain all their tweets. Users are verified through the unique token generated for each user. With this technique, we'll prevent fake data from entering our dataset. Also, we filtered users to have at least 150 tweets, and their account is public. Our questionnaire was utterly ethical since all users completed it voluntarily, and their account was public.

2.4 Data Cleaning

We follow a relatively straightforward strategy to clean our textual data. Firstly, we find and replace some special patterns with unique tokens using regular expressions. More specifically, we replace URLs with [LINK], usernames with [USERNAME], emojis with [EMOJI], and smileys with [SMILEY]. We also take a step further to filter out non-Persian characters. For this purpose, we carefully select specific ranges of the Unicode characters to be allowed in our data. We find that this step substantially improves the performance of our models. Finally, we only preserve users that have more than 100 tweets after performing the above-mentioned cleanings.
Also I should mention that, we used another approach which is cleaning more agressively. In this method we remove any non-Persian characters and all tweets are splitted with just

space, not any other punctuation marks. Currently we are using this method.
After removing not-qualified users which were about 4356, we've got 3876 users at the end.

3 Data Format

3.1 Folder Structure

Data folder contains four subfolders. explaination of each folder :

- raw : this also contains one subfolder and one json file:
  - crawled : this folder is the raw crawled data from twitter that contains all tweets of a user which mentioned his/her MBTI type in one of tweets.
  - datasets.json : this file is the json file which has every username, label and their related raw tweets(which later we can use pandas library to load json file)
- clean : this also contains one subfolder and one json file:
  - crawled : this folder is the cleaned crawled data from twitter that contains all tweets of a user which mentioned his/her MBTI type in one of tweets.
  - datasets.json : this file is the json file which has every username, label and their related cleaned tweets(which later we can use pandas library to load json file)
- sentence broken : sentence broken json file which has 4 columns : username, tweets, label, sentence tokenized tweets with *hazm* library
- word broken : word broken json file which has 4 columns : username, tweets, label, word tokenized tweets with *hazm* library

src folder contains two subfolders. explaination of each folder :

- data processing : this folder is 3 folder and one python file.
  - data collection : this folder is main part of crawling. It works in this way that first collect MBTI keywords in *collect_keyword_tweets.py* then after crawling that user it will save them on a csv file. after that all tweets of each user are collected in *specific_user_tweets.py* file. the *script.py* file use these two file to create our raw dataset.
  - data cleaning : this has scrpit subfolder in it which will handle the data cleaning file with *ptpd.exe* file with running *script.py* file. There is another subfolder which is main part of cleaning that has been implemented in *F#*.
  - tokenizing : this has script *tokenizer.py* which will generate sentence tokenizer and word tokenizer with *hazm* library.
  - main.py : this script is the main part for processing the above folder mentioned and it will run with this command *python main.py all*
- data analyzing : this will generate some statistic about our dataset. which is in stats folder

stats folder contains files which are generated for analyzing our dataset with different criterion which is mainly comprised by *csv, png and json* file.

3.2 Label Independency

As I mentioned we have 16 different MBTI label which are formed on top of 4 separate traits(i.e. I/E, S/N, T/F, P/J). So each user has unique label which is among these 16 label. For example a user can be *INTJ* or *ESTP*.

4 Labeling Explaination

Since there is ambiguity in each tweets with specific MBTI keyword and we don't know what they're describing in the tweet we had to label each user manually. if user was mentioning their type we would've collected that user with all of his/her tweets. So If a user said و منی که ENTP هستم. our group labeled him/her as an ENTP type and collect all of his/her tweets. And on the other hand if a user said تایپ شخصیتی ENTP واقعا عجیبه. this is not mentioning his/her type so we wouldn't pick this user.

4.1 Dataset Unit Label

our unit of labeling is for every single user which were filtered previously, we've collected all of his/her tweets and their related Label. So we train our model on users with all of their tweets to predict their personality type.

5 Data Analyzing

In this section we're going to some details about our datasets and its relevant labels.

5.1 Main Part Counts

The table below shows exact number of sentence, words, unique words and number of all rows in our dataset.

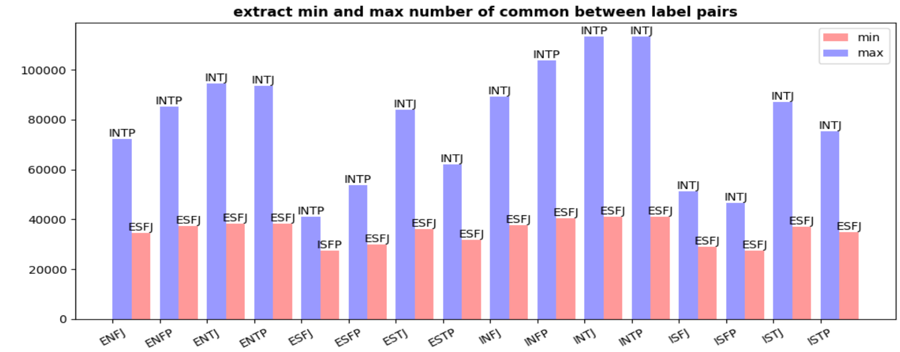
	stats
data-len	3876
sentence-len	6139872
word-len	62840810
unique-words-len	733488

5.2 Common and Not Common Words Count

In this part there are two tables showing each pair of labels how many words are common and not.

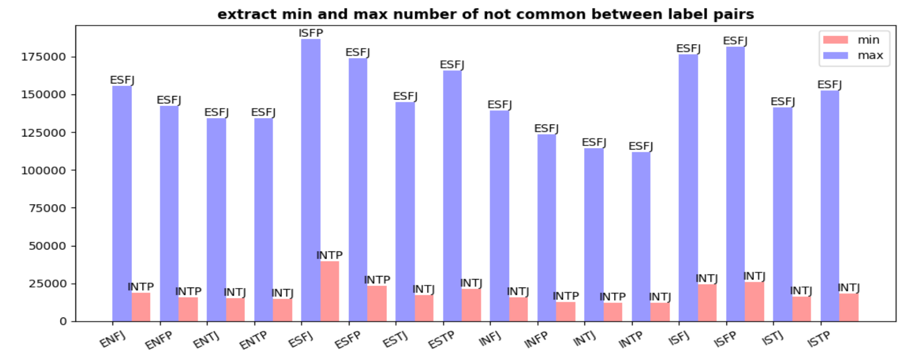
5.2.1 Common Count

	ENFJ	ENFP	ENTJ	ENTP	ESFJ	ESFP	ESTJ	ESTP	INFJ	INFP	INTJ	INTP	ISFJ	ISFP	ISTJ	ISTP
ENFJ		60709	64069	64274	34660	43057	58956	47607	61889	68767	72158	72159	41234	37879	60337	55020
ENFP	60709		73463	73897	37436	47509	66610	52903	71315	81095	84752	85171	45048	41389	69076	62026
ENTJ	64069	73463		80107	38181	48883	73312	56026	76464	86825	94514	93275	47208	42681	75419	66544
ENTP	64274	73897	80107		38357	49221	72340	55840	76166	86883	93549	93263	46800	42896	74574	66213
ESFJ	34660	37436	38181	38357		29913	36173	31864	37697	40439	41083	41097	28916	27393	37187	35001
ESFP	43057	47509	48883	49221	29913		45778	38865	48023	52146	53503	53643	34737	32378	47237	43724
ESTJ	58956	66610	73312	72340	36173	45778		52266	69600	77103	84070	82842	44413	40081	68277	61087
ESTP	47607	52903	56026	55840	31864	38865	52266		54376	59008	62125	61684	37607	34689	53343	49060
INFJ	61889	71315	76464	76166	37697	48023	69600	54376		83622	89193	88328	46091	42088	72271	63980
INFP	68767	81095	86825	86883	40439	52146	77103	59008	83622		103622	103846	49559	45002	80919	70916
INTJ	72158	84752	94514	93549	41083	53503	84070	62125	89193	103622		113254	51421	46548	87139	75369
INTP	72159	85171	93275	93263	41097	53643	82842	61684	88328	103846	113254		51197	46198	86042	74992
ISFJ	41234	45048	47208	46800	28916	34737	44413	37607	46091	49559	51421	51197		31722	45455	42060
ISFP	37879	41389	42681	42896	27393	32378	40081	34689	42088	45002	46548	46198	31722		41327	38510
ISTJ	60337	69076	75419	74574	37187	47237	68277	53343	72271	80919	87139	86042	45455	41327		62916
ISTP	55020	62026	66544	66213	35001	43724	61087	49060	63980	70916	75369	74992	42060	38510	62916	



5.2.2 Not Common Count

	ENFJ	ENFP	ENTJ	ENTP	ESFJ	ESFP	ESTJ	ESTP	INFJ	INFP	INTJ	INTP	ISFJ	ISFP	ISTJ	ISTP
ENFJ		48939	45579	45374	74988	66591	50692	62041	47759	40881	37490	37489	68414	71769	49311	54628
ENFP	85492		72738	72304	108765	98692	79591	93298	74886	65106	61449	61030	101153	104812	77125	84175
ENTJ	94788	85394		78750	120676	109974	85545	102831	82393	72032	64343	65582	111649	116176	83438	92313
ENTP	99939	90316	84106		125856	114992	91873	108373	88047	77330	70664	70950	117413	121317	89639	98000
ESFJ	18619	15843	15098	14922		23366	17106	21415	15582	12840	12196	12182	24363	25886	16092	18278
ESFP	29860	25408	24034	23696	43004		27139	34052	24894	20771	19414	19274	38180	40539	25680	29193
ESTJ	74368	66714	60012	60984	97151	87546		81058	63724	56221	49254	50482	88911	93243	65047	72237
ESTP	41352	36056	32933	33119	57095	50094	36693		34583	29951	26834	27275	51352	54270	35616	39899
INFJ	87235	77809	72660	72958	111427	101101	79524	94748		65502	59931	60796	103033	107036	76853	85144
INFP	131100	118772	113042	112984	159428	147721	122764	140859	116245		96245	96021	150308	154865	118948	128951
INTJ	152949	140355	130593	131558	184024	171604	141037	162982	135914	121485		111853	173686	178559	137968	149738
INTP	155349	142337	134233	134245	186411	173865	144666	165824	139180	123662	114254		176311	181310	141466	152516
ISFJ	27481	23667	21507	21915	39799	33978	24302	31108	22624	19156	17294	17518		36993	23260	26655
ISFP	31901	28391	27099	26884	42387	37402	29699	35091	27692	24778	23232	23582	38058		28453	31270
ISTJ	78252	69513	63170	64015	101402	91352	70312	85246	66318	57670	51450	52547	93134	97262		75673
ISTP	59100	52094	47576	47907	79119	70396	53033	65060	50140	43204	38751	39128	72060	75610	51204	



5.3 Most Uncommon Between Label Pairs

Since there are 16 MBTI labels and for computing most Uncommon words between every pair 16 X 16 = 256 so we can show tables but I showed them in a sample image that is like below

for ENFJ non common with ENFP are :

['درلیست', 'فالوازشما', 'بکاکان', 'برقوباد', 'فالوکنید', 'مثیسیمی', 'چهقدر', 'میخواسم', 'باوا', 'هانیل']

for ENFJ non common with ENTJ are :

['اددشدن', 'فالوازشما', 'بکاکان', 'برقوباد', 'آج', 'مثیسیمی', 'یونجون', 'ویشی', 'هانیل', 'طنازم']

for ENFJ non common with ENTP are :

['درلیست', 'فالوازشما', 'بکاکان', 'برقوباد', 'مثیسیمی', 'چهقدر', 'ویشی', 'هانیل', 'عشی', 'میخوایی']

for ENFJ non common with ESFJ are :

['آذربایجان', 'درلیست', 'اددشدن', 'فالوازشما', 'باافتخار', 'بکاکان', 'فالوی', 'برقوباد', 'فالوکنید', 'فالوبک']

for ENFJ non common with ESFP are :

Figure 3: most uncommon words between every 2 label

5.4 Relative Normalized Frequency

This criterion is also same as previous one that can't be fit in a page so I show some part of it

for ENFJ non common with ENFP are :

[['و', 0.96], ['که', 0.99], ['به', 1.04], ['از', 1.03], ['من', 0.92], ['این', 1.05], ['رو', 1.11], ['تو', 0.9500000000000001], ['یه', 1.06], ['با', 1.03]]

for ENFJ non common with ENTJ are :

[['و', 0.91], ['که', 0.9500000000000001], ['به', 0.96], ['از', 0.96], ['من', 1.07], ['این', 0.97], ['رو', 0.9400000000000001], ['تو', 1.01], ['یه', 1.01], ['با', 0.97]]

for ENFJ non common with ENTP are :

[['و', 0.9400000000000001], ['که', 0.99], ['به', 1.02], ['از', 0.99], ['من', 1.04], ['این', 0.98], ['رو', 0.9400000000000001], ['تو', 0.98], ['یه', 1.04], ['با', 1.01]]

for ENFJ non common with ESFJ are :

[['و', 1.04], ['که', 1.08], ['به', 1.06], ['از', 1.04], ['من', 0.99], ['این', 1.0], ['رو', 1.21], ['تو', 1.0], ['یه', 1.16], ['با', 1.1]]

for ENFJ non common with ESFP are :

[['و', 0.9400000000000001], ['که', 0.96], ['به', 1.06], ['از', 1.05], ['من', 0.9500000000000001], ['این', 1.05], ['رو', 1.06], ['تو', 0.9400000000000001], ['یه', 0.98], ['با', 1.03]]

for ENFJ non common with ESTJ are :

[['و', 0.87], ['که', 1.0], ['به', 0.96], ['از', 0.9400000000000001], ['من', 1.15], ['این', 1.04], ['رو', 1.04], ['تو', 1.06], ['یه', 1.09], ['با', 0.99]]

Figure 4: RNF between every two pair

5.5 TF-IDF

this is a criterion that computes text frequency in a document and inverse document frequency. The exact formula is in like this:

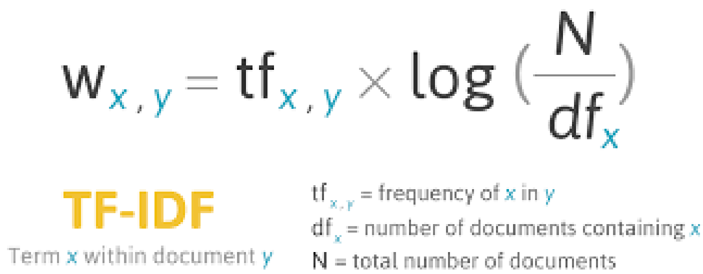


Figure 5: tf-idf formula

And according to the computation of above formula, results in the following table:

19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	
که	به	از	من	این	رو	تو	یه	با	می	هم	ولی	خیلی	بود	دیگه	در	تا	نه	اون	کنم	ENFJ
که	به	از	من	تو	این	رو	یه	با	هم	ولی	خیلی	بود	می	دیگه	نه	تا	در	کنم	اون	ENFP
که	به	از	من	رو	این	تو	یه	با	هم	می	در	خیلی	بود	ولی	دیگه	تا	نه	اون	ها	ENTJ
که	به	از	من	رو	این	تو	یه	با	هم	ولی	خیلی	می	بود	در	دیگه	نه	تا	اون	کنم	ENTP
که	به	از	من	این	تو	رو	یه	با	می	هم	خیلی	ولی	بود	دیگه	نه	تا	در	کنم	اون	ESFJ
که	به	از	من	تو	این	رو	یه	با	هم	می	خیلی	ولی	بود	دیگه	در	نه	تا	اون	دارم	ESFP
که	به	از	من	این	رو	تو	با	یه	هم	می	در	بود	خیلی	ولی	دیگه	تا	نه	اون	برای	ESTJ
که	به	از	من	رو	این	تو	یه	با	هم	ولی	خیلی	بود	در	می	دیگه	نه	تا	اون	داره	ESTP
که	به	از	من	این	رو	تو	یه	با	هم	می	ولی	خیلی	بود	در	دیگه	تا	نه	اون	کنم	INFJ
که	به	از	من	تو	این	رو	یه	می	با	هم	ولی	خیلی	بود	دیگه	در	تا	نه	کنم	اون	INFP
که	به	از	رو	این	من	تو	یه	هم	با	می	ولی	در	خیلی	بود	دیگه	تا	نه	اون	ها	INTJ
که	به	از	من	این	رو	تو	یه	با	هم	می	ولی	در	خیلی	بود	دیگه	تا	نه	اون	کنم	INTP
که	به	از	رو	من	این	تو	یه	هم	با	می	ولی	خیلی	بود	دیگه	در	تا	نه	اون	کنم	ISFJ
که	از	من	به	این	تو	رو	یه	با	خیلی	ولی	هم	بود	دیگه	می	تا	نه	کنم	در	چرا	ISFP
که	به	از	رو	من	این	تو	یه	هم	با	می	ولی	خیلی	در	دیگه	بود	نه	تا	اون	میشه	ISTJ
که	به	از	من	این	تو	رو	یه	با	هم	ولی	خیلی	بود	می	در	دیگه	تا	نه	اون	میشه	ISTP

The above table shows tf-idf between every 2 pairs of label. and not suprisingly perposition words are in top occuring based on this criterion. we see that the word "که" happened more than any other perposition. And it seems in twitter Iranian use the word ”very” so often that it'll be translated to "خیلی". And the verb with most factor of tf-idf is "میشه" that it's informal version of "می‌شود". and finally there is result that make sense which I believe twitter users are more Intorvert as a result they should use more the word ”me” in their tweets that its translation is "من". so these are the facts that can be understood from the tf-idf table.

5.6 Unique Words Most And Least Frequent

In this part we calculate which words happened the most and the least. For sake of fitting into the page we just select first 30 of each part.

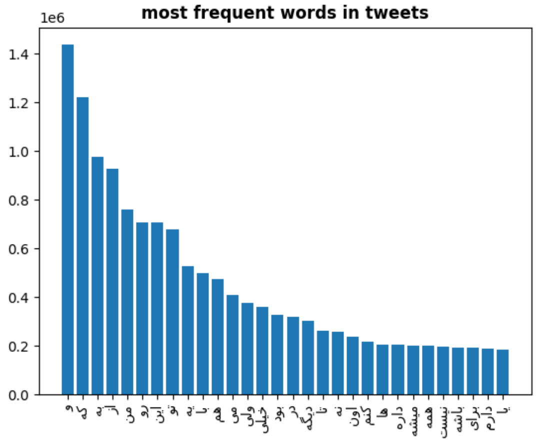


Figure 6: Most Frequent Words

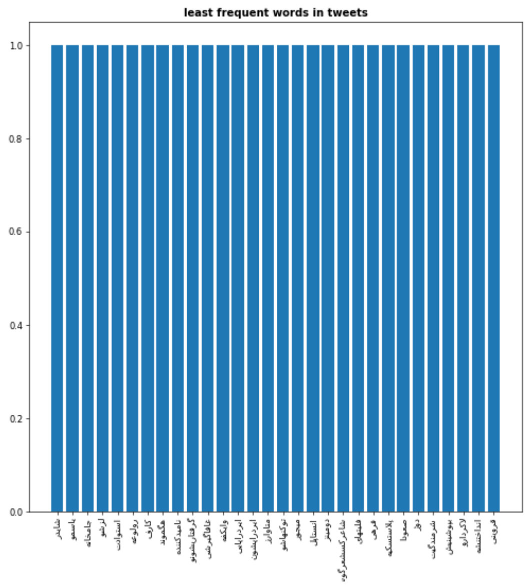


Figure 7: Least Frequent Words

Most occuring words are almost perposition and common verbs which happens to be around 1,4 milion. but for least frequent words all of them are one.