



Natural Language Processing
Project Phase 1

Data Generation for Personality Detection

Babak Behkamkia (98521099)

Dr. Sauleh Eetemadi

-

Contents

1	Introduction	3
2	Related Datasets	3
3	Collection	3
4	Cleaning	3
5	Generation	4
6	Statistics	5
6.1	Basic statistics	5
6.1.1	raw data	5
6.1.2	clean data	6
6.2	Number of unique common words	6
6.3	Number of uncommon unique words	7
6.4	Most frequent uncommon words	7
6.5	Relative Normalized Frequency	7
6.6	TF_IDF	8
6.7	Most appeared unique words	9
7	Scripts	9

1 Introduction

Nowadays, a huge number of different products are produced worldwide. Thus, people have more options, leading producers to consider increasing their sales. Thus, they tried to determine each individual's personality and advertise the products customized to their personality type. Collecting such data is challenging for researchers because many people don't know their personality type. Thus, researchers tried to extract the personality of people from their tweets. There are other problems because they cannot determine people's personalities just by reading a tweet. Therefore, they are dependent on people again to determine their own personality. With data generation, we hope to overcome these problems. I put my code for this project on my GitHub¹.

2 Related Datasets

There is a ready dataset for this task on Kaggle (Myers-Briggs Personality Type Dataset²). This dataset consists of more than 8600 samples. Each row of this dataset presents a person's personality type and his/hers most recent 50 posts. I used this dataset in the generation phase.

3 Collection

I used Tweepy and Twitter API to crawl for tweets that contain a personality keyword like "INFJ". I repeated this procedure for all 16 personality types in MBTI and labeled them with the corresponding keyword.

I created a dataframe for each label and saved it in CSV format. These dataframes consist of 5 columns:

- author name
- author screen_name
- author id
- post
- author personality type

4 Cleaning

In this section, I describe the cleaning methods I have used to make the raw texts of tweets more processable data.

¹<https://github.com/Babakbehkamkia/Natural-Language-Processing>

²<https://www.kaggle.com/datasets/datasnaek/mbti-type>

- Necessary cleaning: Removing special characters, stop words, and Lemmatizing the words are done in this part of cleaning.
- Removing Hashtags
- Removing Links: Links do not have a semantic meaning. Thus, their presents will not improve the model.
- Removing Usernames: A name by itself doesn't contain any information. Therefore, I remove usernames.
- Removing Retweet Sign: All retweets have the string "RT" in their beginning, which is unnecessary.
- Replacing Emojis: In many related works, researchers tend to delete the emojis. However, they contain a lot of information, and even in some cases, they can change the sentence's meaning. In this work, I replaced all emojis with their corresponding text encode by using the demoji³ library in Python.

In the end, I saved the cleaned text in a new column named "cleaned_text" in the same dataframe.

5 Generation

As told earlier, collecting and annotating data for personality detection is a challenging task. Thus, I tried to generate new samples for each personality type by using GPT-3 API. I created a prompt that contains the personality, a predefined topic (e.g. depression), and an example post from Myers-Briggs Personality Type Dataset with the given personality. With this prompt, I generated some posts toward the given topic and labeled them with the corresponding personality type. Moreover, in some generated samples, the personality type is mentioned directly. Therefore, I used the [PT] token instead of the personality type.

The generated dataset is saved as a CSV file in the "data/raw" directory. Each file has three columns:

- generated post
- topic of the post
- personality type

³<https://pypi.org/project/demoji/>

6 Statistics

I used nltk library for the following:

- **nltk.tokenize.TweetTokenizer**: tokenizing the words
- **nltk.corpus.stopwords**: removing stopwords
- **nltk.stem.WordNetLemmatizer**: Lemmatizing the words
- **nltk.word_tokenize**: breaking samples to words
- **nltk.sent_tokenize**: breaking samples to sentences

6.1 Basic statistics

6.1.1 raw data

	sample num	sentence num	word num	unique word num
ENFP	3142	5758	81077	8107
ENFJ	1205	1700	32802	2842
ESTJ	764	1277	27318	952
ESFJ	494	1486	15653	1345
ESTP	580	995	16953	1523
ESFP	361	856	10176	1554
ENTJ	426	634	10552	2142
ENTP	1419	2442	31964	4725
INFP	5423	10094	145891	11956
INTP	3649	5914	92065	9511
INFJ	1000	1590	25460	3642
INTJ	2048	3387	49572	6986
ISFP	1000	1895	28085	2907
ISTP	1000	1955	28596	2358
ISFJ	1000	1930	29031	2390
ISTJ	1000	1769	28896	2375

6.1.2 clean data

	sample num	sentence num	word num	unique word num
ENFP	3142	3139	34786	4461
ENFJ	1205	1183	14100	1635
ESTJ	764	764	11513	579
ESFJ	494	494	6503	824
ESTP	580	579	7861	923
ESFP	361	361	4266	963
ENTJ	426	426	4543	1309
ENTP	1419	1406	12943	2725
INFP	5423	5419	59540	5890
INTP	3649	3646	38042	4857
INFJ	1000	1000	9901	2165
INTJ	2048	2041	20236	3795
ISFP	1000	1000	11235	1599
ISTP	1000	1000	12310	1341
ISFJ	1000	999	12090	1404
ISTJ	1000	1000	12386	1412

6.2 Number of unique common words

	ENFP	ENFJ	ESTJ	ESFJ	ESTP	ESFP	ENTJ	ENTP	INFP	INTP	INFJ	INTJ	ISFP	ISTP	ISFJ	ISTJ
ENFP		907	365	532	571	616	774	1279	2061	1842	1084	1536	883	723	793	812
ENFJ	907		264	389	397	457	549	752	1030	933	663	843	572	478	537	539
ESTJ	365	264		191	192	209	271	295	396	365	279	341	239	209	234	251
ESFJ	532	389	191		255	273	365	415	565	526	388	489	365	302	350	340
ESTP	571	397	192	255		309	356	485	616	598	418	523	388	347	341	360
ESFP	616	457	209	273	309		403	550	667	617	456	568	423	380	388	387
ENTJ	774	549	271	365	356	403		691	888	840	571	780	478	419	444	472
ENTP	1279	752	295	415	485	550	691		1459	1400	865	1221	717	603	642	644
INFP	2061	1030	396	565	616	667	888	1459		2262	1259	1837	996	846	868	912
INTP	1842	933	365	526	598	617	840	1400	2262		1146	1750	931	778	812	835
INFJ	1084	663	279	388	418	456	571	865	1259	1146		1063	640	555	609	597
INTJ	1536	843	341	489	523	568	780	1221	1837	1750	1063		797	699	739	775
ISFP	883	572	239	365	388	423	478	717	996	931	640	797		513	505	525
ISTP	723	478	209	302	347	380	419	603	846	778	555	699	513		464	473
ISFJ	793	537	234	350	341	388	444	642	868	812	609	750	505	464		505
ISTJ	812	539	251	340	360	387	472	644	912	835	597	775	525	473	505	

6.3 Number of uncommon unique words

	ENFP	ENFJ	ESTJ	ESFJ	ESTP	ESFP	ENTJ	ENTP	INFP	INTP	INFJ	INTJ	ISFP	ISTP	ISFJ	ISTJ
ENFP		4282	4310	4221	4242	4192	4222	4628	6229	5634	4458	5184	4294	4356	4279	4249
ENFJ	4282		1686	1681	1764	1684	1846	2856	5465	4626	2474	3744	2090	2020	1965	1969
ESTJ	4310	1686		1021	1118	1124	1346	2714	5677	4706	2186	3692	1700	1502	1515	1489
ESFJ	4221	1681	1021		1237	1241	1523	2719	5584	4629	2213	3641	1693	1561	1528	1556
ESTP	4242	1764	1118	1237		1268	1530	2678	5581	4584	2252	3672	1746	1570	1645	1615
ESFP	4192	1684	1124	1241	1268		1466	2588	5519	4586	2216	3622	1716	1544	1591	1601
ENTJ	4222	1846	1346	1523	1520	1466		2652	5423	4486	2332	3544	1952	1812	1825	1777
ENTP	4628	2856	2714	2719	2678	2588	2652		5697	4782	3160	4078	2890	2860	2845	2849
INFP	6229	5465	5677	5584	5581	5519	5423	5697		6223	5537	6011	5497	5539	5558	5478
INTP	5634	4626	4706	4629	4584	4586	4486	4782	6223		4730	5152	4594	4642	4637	4599
INFJ	4458	2474	2186	2213	2252	2216	2332	3160	5537	4730		3834	2484	2396	2351	2383
INTJ	5184	3744	3692	3641	3672	3622	3544	4078	6011	5152	3834		3800	3738	3681	3657
ISFP	4294	2090	1700	1693	1746	1716	1952	2890	5497	4594	2484	3800		1914	1993	1961
ISTP	4356	2020	1502	1561	1570	1544	1812	2860	5539	4642	2396	3738	1914		1817	1807
ISFJ	4279	1965	1515	1528	1645	1591	1825	2845	5558	4637	2351	3681	1993	1817		1806
ISTJ	4249	1969	1489	1556	1615	1601	1777	2849	5478	4599	2383	3657	1961	1807	1806	

6.4 Most frequent uncommon words

According to the number of labels (16), I can only show part of the table in this section. There is a total of 120 (choosing 2 out of 16) possible combinations for the rows of this table.

	word 0	word 1	word 2	word 3	word 4	word 5	word 6	word 7	word 8	word 9
ENFP/ENFJ	sinb	vote	virgo	yangyang	literal	wayv	facewe	sauropod		
ENFJ/ENFP	gold	unwavering	desire	empathetic	headbandage	skazvt	gel	eya	ford	teddy
ENFP/ESTJ	enfp	enfj	hi	junhoe	donghyuk	intp	ikon	passport	junhwan	yunhyeong
ESTJ/ENFP	exol	estj	lucky	bgm	glrf	mama	exo	backhyuns	050623	kyoong
ENFP/ESFJ	junhoe	donghyuk	ikon	passport	junhwan	yunhyeong	bobby	chanwoo	w	1
ESFJ/ENFP	demiboy	gopard	izzie	gold	serving	nurturing	jia		jianna	kye
ENFP/ESTP	junhoe	donghyuk	ikon	passport	junhwan	yunhyeong	bobby	chanwoo	idol	game
ESTP/ENFP	jung	ahyeon	april	2007	position	rapper	kaden	bg	lyon	army
ENFP/ESFP	junhoe	donghyuk	ikon	passport	junhwan	yunhyeong	bobby	chanwoo	idol	sinb
ESFP/ENFP	sean	herb	04	lep	1er	lbfm	nmixx	artistic	riam	hergenshin
ENFP/ENTJ	junhoe	donghyuk	ikon	passport	junhwan	yunhyeong	bobby	chanwoo	sinb	helloo
ENTJ/ENFP	entjartist	brain	brawn	table	savvy	jihoons	doyoung	entjs	style	opportunism
ENFP/ENTP	junhoe	donghyuk	ikon	passport	junhwan	yunhyeong	bobby	chanwoo	idol	sinb
ENTP/ENFP	youngk	xh	uid	position	charge	heittthey	rara	8540672	93	endless
ENFP/INFP	junhoe	donghyuk	ikon	passport	junhwan	yunhyeong	chanwoo	sinb	preseason	baekho
INFP/ENFP	weather	ssearaim		reii	emotion	butterfly		230510	743pm	tag
ENFP/INTP	sinb	sam	yangyang	wayv	facewe	sauropod	preseason	option	baekho	yeonjun
INTP/ENFP	lizie	stanlist	lina	int	moncaratfournot	ssearaim	galy	baemon	blackpink	twic
ENFP/INFJ	junhoe	isfp	donghyuk	ikon	passport	junhwan	yunhyeong	bobby	chanwoo	sinb
INFJ/ENFP	jungwoo	doyoung	jaehyun	reset	kofi	bffs	mae	loona	taeyang	bust

6.5 Reletive Normalized Frequency

Due to the same problems as the above section, I only show the first 20 rows of this table.

	word 0	word 1	word 2	word 3	word 4	word 5	word 6	word 7	word 8	word 9
ENFP/ENFJ	sam	another	ten	idol	v	yet	3rd	game	ask	telling
ENFP/ESTJ	sheher	interactive	cry	game	amp	e	according	virgo	another	v
ENFP/ESFJ	enfp	according	game	there	cat	hello	intp	option	yet	havent
ENFP/ESTP	enfp	isfp	according	w	people	sinb	there	many	enfj	first
ENFP/ESFP	according	enfp	enfj	another	many	time	whats	tear	intp	qrt
ENFP/ENTJ	enfp	according	idol	game	minor	vote	isfp	many	active	v
ENFP/ENFP	helloo	vote	according	enfp	game	telling	isfp	cat	virgo	others
ENFP/INFP	game	ceo	lt	sam	rabbit	standard			enfp	headphone
ENFP/INTP	ten			3rd	ceo	lt	helloo	acc	nbtis	haii
ENFP/INFJ	enfp	according	virgo	idol	pronoun	enfj	intp	havent	3rd	game
ENFP/INTJ	bobby	enfp	chamwoo	active	option	3rd	ht	voting	member	1st
ENFP/ISFP	helloo	05	please	idol	pronoun	enfp	option	yet	yeonjun	dream
ENFP/ISTP	enfp	enfj	vote	please	according	name	w	many	today	said
ENFP/ISFJ	enfp	w	enfj	according	intj	ten	05	yet	qrt	full
ENFP/ISTJ	wayv	please	idol	option	yet	flag	l	mainly	there	acc
ENFJ/ESTJ	sheher	guy	according	interactive	born	leader	boy	let	help	cry
ENFJ/ESFJ	according	guy	enfj	scorpio	intp	enfp	lee	hello	isfp	born
ENFJ/ENFP	enfj	isfp	according	leader	guy	enfp	help	heart	born	scorpio
ENFJ/ESFP	enfj	according	intp	leader	boy	guy		15	gel	enfp
ENFJ/ENTJ	according	minor	enfj	isfp	sheher	name	others	enfp	help	3

6.6 TF_IDF

I used **TfidfVectorizer** from **sklearn** in order to compute TF_IDF.

	ENFP	ENFJ	ENTJ	ESFJ	ESTP	ESFP	ENTJ	ENTP	ISFP	INTP	INFJ	INTJ	ISFP	INTP	INFJ	INTJ
0	day	isfp	paper	aka	chance	it	new	92	philippines	jishuan	new	ten	inf	enfp	enfp	admirer
1	sun	nj	spide	surprising	position	if	put	rara	stare	jushoe	most	new	cause	li	fact	according
2	stib	file	see	hokim	aditi	creativity	became	rice	art	yundeyong	looking	infj	goodlooking	blackpink	fun	isfp
3	interactive	ribbon	number	finding	south	newjeans	like	hottthey	weather	contacting	aditi	personalitylovely	day	luc	know	enfj
4	bobby	hi	em	sheher	rapper	looking	add	found	hello	enfp	bfa	literal	jdlf	sheher	day	most
5	sheher	looking	paper	main	most	stay	face	add	put	lma	enfp	erv	active	findom	most	isfp
6	face	sheher	gt h	future	11th	artistic	character	character	character	enfp	gg	most	face	jeongin	hot	diva
7	chamwo	aditi	always	new	april	aditi	enfp	friend	findom	hi	lodi	li	15	stonyjay	new	chant
8	yundeyong	bobby	bucklyuns	looking	looking	lee	jishuan	hi	day	interactive	enfp	wayv	trixx	welcomed	probably	bobby
9	jishuan	most	kyong	interactive	lg	enfp	bringing	add	add	stan	fun	put	lth	casuals	sheher	in
10	passport	chamwo	050823	14	place	stage	enfp	new	character	lmae	fact	searipool	sky	lmael	looking	chamwo
11	lma	lma	aditi	see	in	lep	brain	looking	li	new	know	facewen	li	aditi	extrovert	lma
12	donghyuk	jishuan	bobby	in	train	bobby	galileo	ak	sheher	day	entirement	new	enfp	lma	yundeyong	enfp
13	jishuan	donghyuk	chamwo	lma	lma	enfp	aditi	yong	interactive	stardust	infj	add	aditi	lmaenfp	enfp	passport
14	most	donghyuk	lma	enfp	most	lma	most	new	sheher	in	most	sheher	interactive	stonyjay	donghyuk	enfp
15	looking	passport	gtf	denahay	lma	lma	enfp	aditi	aditi	aditi	enfp	character	interactive	new	deoyong	jishuan
16	new	yundeyong	lma	gypard	lma	in	entjartist	gon	most	most	deoyong	looking	most	most	jishuan	jishuan
17	aditi	enfp	enfp	most	2007	enfp	aditi	yong	looking	looking	jishuan	aditi	looking	in	enfp	enfp
18	in	in	lody	most	skyeon	herb	in	in	in	in	infj	in	in	in	jeongwo	aditi
19	enfp	enfj	enfp	enfj	enfp	enfp	enfp	enfp	enfp	enfp	jeongwo	enfp	enfp	enfp	enfp	enfp

6.7 Most appeared unique words

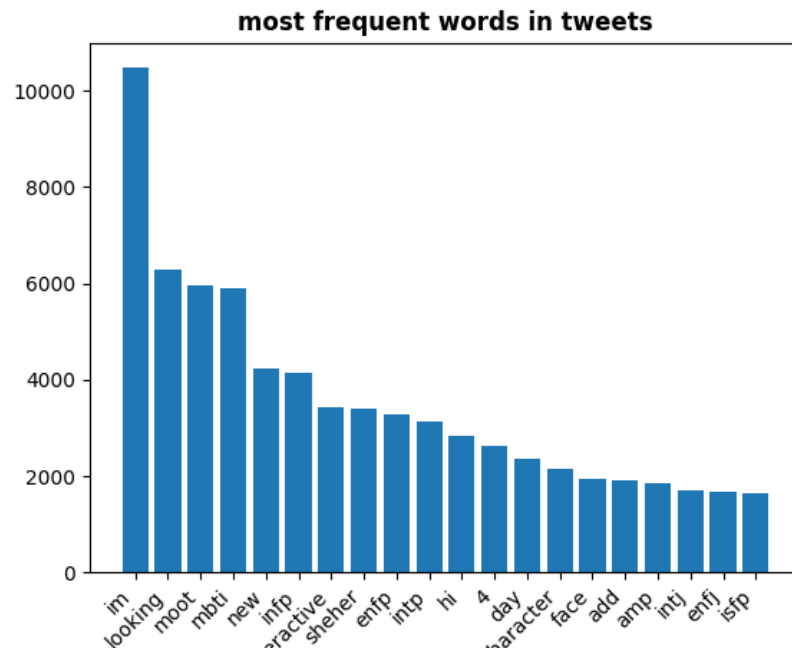


Figure 1: most Frequent Words

7 Scripts

For this work, bash and batch scripts are provided to run each part or the entire project. The necessary instructions are inside those files.