Personality Detection on Persian Dataset

Phase 1

Mohammadjavad Mehditabar - 98522049

NLP Phase 1

May 21, 2023

Contents

1 Data Overview **2**

2 Data Processing **2**

2.1 Bio Search 2

2.2 Tweet Search 2

2.3 Questionnaire 3

2.4 Data Cleaning 3

3 coda explaination **4**

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.

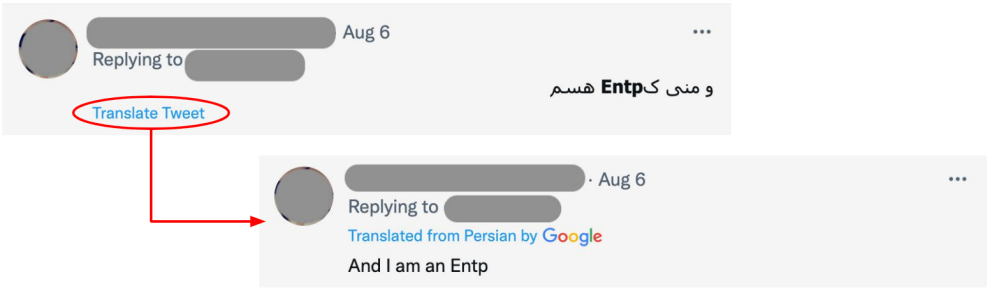

Figure 2: A sample of tweet search

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.

The following will give you an overview of what you can do with this template.

Problem 1

Type your problem here.

Personally I recommend Mathpix (https://mathpix.com/), which can easily export your ProblemBook.pdf to LaTeX code.

Solution. Write your solution here.
Example of equations. $x + 1 = 2$. Or

$$x - 1 = 0$$

Example of a list of equations.

$$x = 1$$
$$y = 2$$

Example of a matrix.

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Example of a lemma.

Lemma. *This is a lemma.*

3 coda explaination

پاراگراف فارسی ممد اومد انجیا sda منم منم منن چطوری پسر sad ستون سلام میو میو

ISTP	ISTJ	ISFP	ISFJ	INTP	INTJ	INFP	INFJ	ESTP	ESTJ	ESFP	ESFJ	ENTP	ENTJ	ENFP	ENFJ	

This is a Persian word: سیب. It means apple. so what
Example of a proof.

Proof. Write your proof here. □

Example of including a picture.

Example of referring to a piece of code. my main problem is *whatthehell* is something happening

```
1 print("Hello World!")
```

Example of a table.

	Mean	SD
Fall 2077	7.046512	1.714552
Fall 1977	9.102941	1.568919

Overall, this is a quite basic template for assignments, and above are only some basic features. I included enough packages and set a few environments. You may modify them or add features to fit your personal preference. Enjoy using it!