



دانشکده مهندسی کامپیوتر

## مجموعه داده فارسی برای تشخیص شخصیت در بستر تویتر

پایان نامه برای دریافت درجه کارشناسی در رشته مهندسی کامپیوتر  
گرایش هوش مصنوعی و رباتیک

نام دانشجو

زهرا انوریان

شماره دانشجویی

۹۵۵۲۱۰۵۴

استاد راهنما

سید صالح اعتمادی

خرداد ۱۴۰۰



## تأییدیه‌ی هیأت داوران جلسه‌ی دفاع از پایان‌نامه

نام دانشکده: دانشکده مهندسی کامپیوتر

نام دانشجو: زهرا انوریان

عنوان پایان‌نامه: مجموعه‌داده فارسی برای تشخیص شخصیت در بستر توئیتر

تاریخ دفاع: خرداد ۱۴۰۰

رشته: مهندسی کامپیوتر

گرایش: هوش مصنوعی و رباتیک

ردیف	سمت	نام و نام خانوادگی	مرتبۀ دانشگاهی	دانشگاه	امضا
۱	استاد راهنما	سید صالح اعتمادی	استادیار	علم و صنعت ایران	
۲	استاد مدعو داخلی	بهروز مینایی	دانشیار	علم و صنعت ایران	

ب

## تأییدیه‌ی صحت و اصالت نتایج

باسمه تعالی

اینجانب زهرا انوریان به شماره دانشجویی ۹۵۵۲۱۰۵۴ دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیه‌ی نتایج این پایان‌نامه حاصل کار اینجانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری‌شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. در صورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی ...) با اینجانب رفتار خواهد شد و حق هرگونه اعتراض در خصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم. در ضمن، مسؤولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده‌ی اینجانب خواهد بود و دانشگاه هیچ‌گونه مسؤولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی: زهرا انوریان

تاریخ و امضا:

## مجوز بهره‌برداری از پایان‌نامه

بهره‌برداری از این پایان‌نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنما به شرح زیر تعیین می‌شود، بلامانع است:

- بهره‌برداری از این پایان‌نامه برای همگان بلامانع است.
- بهره‌برداری از این پایان‌نامه با اخذ مجوز از استاد راهنما، بلامانع است.
- بهره‌برداری از این پایان‌نامه تا تاریخ ..... ممنوع است.

استاد راهنما: سید صالح اعتمادی

تاریخ:

امضا:

## قدردانی

سپاس خداوندگار حکیم را که با لطف بی‌کران خود، آدمی را زیور عقل آراست.  
در آغاز وظیفه خود می‌دانم از زحمات بی‌دریغ استاد راهنمای خود، جناب آقای اعتمادی صمیمانه تشکر  
و قدردانی کنم که قطعاً بدون راهنمایی‌های ارزنده ایشان، این پروژه به انجام نمی‌رسید.  
و همچنین تشکر می‌کنم از آقای محمدمهدی عبدالله‌پور که به عنوان همکار در کنار این پژوهش حضور  
داشتند و از ایشان بسیار آموختم.  
در پایان، بوسه می‌زنم بر دستان خداوندگاران مهر و مهربانی، پدر و مادر عزیزم و بعد از خدا، ستایش می‌کنم  
وجود مقدس‌شان را و تشکر می‌کنم از خانواده عزیزم به پاس عاطفه سرشار و گرمای امیدبخش وجودشان،  
که بهترین پشتیبان من بودند.

زهره انوریان

خرداد ۱۴۰۰

## چکیده

در سال‌های اخیر، شناخت ویژگی‌های شخصیتی افراد از طریق شبکه‌های اجتماعی به موضوعی جالب در هر دو زمینه پردازش زبان طبیعی و علوم اجتماعی تبدیل شده است. تحقیقات روانشناختی همچنین نشان می‌دهد برخی از ویژگی‌های شخصیتی با رفتار زبانی ارتباط دارند. مدل‌های پردازش زبان طبیعی می‌توانند از این همبستگی برای مدل‌سازی و پیش‌بینی صفات شخصیتی، بر اساس حجم عظیمی از داده‌های موجود که به لطف رسانه‌های اجتماعی مدرن در دسترس است، بهره بگیرند. پیش از اینکه بخواهیم اولین مجموعه داده در زبان فارسی را از طریق شبکه اجتماعی توئیتر جمع‌آوری و تدوین کنیم، هیچ مجموعه داده‌ای در این زمینه در زبان فارسی وجود نداشته است. همانطور که در این مقاله مورد بحث قرار گرفت، ما یک مجموعه داده جدید ساخته‌ایم که دارای برجسب شاخص‌های مدل مایرز-بریگز و متشکل از ۱۵۵۲۵۳۲ توئیت است. همچنین روش‌های جمع‌آوری اطلاعات خود را ارائه داده‌ایم و در مورد چالش‌ها و نتایج آن‌ها به طور مفصل بحث کرده‌ایم. به عنوان مبنایی برای سایر محققان برای پیشرفت بیشتر، یک مدل را با تنظیم دقیق تغییرات معماری برت، پارس‌برت، که قبلاً روی متون‌های فارسی آموزش دیده است، معرفی کرده‌ایم. سرانجام، این مدل را با استفاده از روش اعتبارسنجی متقاطع طبقه‌ای تکرارشونده K بار، مجموعه داده را ارزیابی و نتایج را منتشر نمودیم.

واژگان کلیدی: ویژگی‌های شخصیتی، مجموعه داده، علوم داده، داده‌های اجتماعی، پارس‌برت

# فهرست مطالب

ح	فهرست تصاویر
خ	فهرست جداول
۱	فصل ۱: مقدمه
۱	۱-۱ شرح مسأله
۲	۲-۱ انگیزه‌های پژوهش
۳	۳-۱ دستاوردهای پایان‌نامه
۳	۴-۱ ساختار پایان‌نامه
۴	فصل ۲: مفاهیم پایه
۴	۱-۲ مجموعه داده
۵	۲-۲ شبکه‌های عصبی عمیق
۶	۳-۲ جانمایی کلمات
۷	۴-۲ ترنسفورمر
۸	۵-۲ مدل برت
۱۰	فصل ۳: مروری بر کارهای مرتبط
۱۰	۱-۳ بررسی کارهای انجام شده در این حوزه
۱۰	۱-۱-۳ توییت‌ر
۱۱	۲-۱-۳ ردیت



۱۱	۳-۱-۳ پاندورا
۱۲	۳-۱-۴ فیس بوک
۱۳	۳-۲ دلیل عدم انتخاب مدل پنج عامله
۱۴	فصل ۴: جمع آوری مجموعه داده
۱۴	۴-۱ جستجو کلیدواژه
۱۵	۴-۲ پرسشنامه
۱۶	۴-۳ مقایسه روش های پیشنهادی
۱۷	۴-۴ جمع آوری تویت ها
۱۸	۴-۵ تمیزکردن داده
۱۹	فصل ۵: تحلیل و ارزیابی مجموعه داده
۱۹	۵-۱ تحلیل مجموعه داده
۲۰	۵-۲ آماده سازی داده
۲۲	۵-۳ معرفی مدل پایه
۲۲	۵-۴ نتایج بدست آمده
۲۴	فصل ۶: نتیجه گیری و کارهای آینده
۲۴	۶-۱ نتیجه گیری
۲۴	۶-۲ کارهای آینده
۲۶	مراجع
۲۸	واژه نامه فارسی به انگلیسی
۳۰	واژه نامه انگلیسی به فارسی

## فهرست تصاویر

- ۲-۱ بخشی از یک نمونه مجموعه داده . . . . . ۵
- ۲-۲ ساختار یک نمونه شبکه عصبی . . . . . ۶
- ۲-۳ مدل انتقال توالی توسط ترنسفورمر . . . . . ۷
- ۲-۴ دو نمونه از مدل برت . . . . . ۸
- ۴-۱ نسبت اعتبار ارسالی‌های پرسشنامه اولیه . . . . . ۱۶
- ۴-۲ نسبت اعتبار ارسالی‌های پرسشنامه نهایی . . . . . ۱۶
- ۵-۱ توزیع ویژگی‌های شخصیتی بر اساس جنسیت در مجموعه داده . . . . . ۲۰
- ۵-۲ مقایسه توزیع MBTI در مجموعه داده‌ی ما با توزیع MBTI در جمعیت ایران . . . . . ۲۰
- ۵-۳ مراحل انجام روش اعتبارسنجی متقارن طبقه‌ای تکرارشونده K بار . . . . . ۲۱
- ۵-۴ توزیع تعداد کلمات نمونه‌ای از داده‌های آموزش (تعداد بین برابر ۶۴) . . . . . ۲۲

## فهرست جداول

۴-۱ آمار کلی روش‌های جمع‌آوری داده‌ها . . . . . ۱۷

۵-۱ میانگین متوسط کلان امتیاز F1 نسبت به نتایج پنج تکرار اعتبارسنجی متقابل طبقه‌ای K بار ۲۳

# فصل ۱

## مقدمه

### ۱-۱ شرح مسأله

شخصیت، مجموعه مشخصه‌های رفتار، شناخت و الگوهای عاطفی است که از عوامل بیولوژیکی و محیطی نشأت می‌گیرد [۲]. در سال‌های اخیر تحقیقات زیادی در زمینه شناسایی شخصیت انجام شده است که می‌تواند در زمینه‌های مختلف از جمله غربالگری شغل<sup>۱</sup> [۹]، سیستم‌های توصیه‌ای<sup>۲</sup> [۲۱]، تبلیغات [۱۱]، تشخیص قطبیت کلمات<sup>۳</sup> [۱۸] و تحلیل شبکه‌های اجتماعی [۱] مفید باشد. مدل روانشناختی مایرز-بریگز<sup>۴</sup> به الگوهایی از جمع‌آوری اطلاعات، نحوه تصمیم‌گیری و نحوه زندگی افراد بر اساس انتخاب شیوه زندگی اشاره دارد [۱۰]. چهار ویژگی پیوسته در مدل MBTI وجود دارد [۱۴]:

• برون‌گرا<sup>۵</sup> (E) – درون‌گرا<sup>۶</sup> (I)

• حسی<sup>۷</sup> (S) – شمی<sup>۸</sup> (N)

---

<sup>۱</sup> Job Screening

<sup>۲</sup> Recommendation System

<sup>۳</sup> Word Polarity Detection

<sup>۴</sup> MBTI: Myers-Briggs Type Indicators

<sup>۵</sup> Extroversion

<sup>۶</sup> Introversion

<sup>۷</sup> Sensing

<sup>۸</sup> Intuition

● منطقی<sup>۹</sup> (T) – احساسی<sup>۱۰</sup> (F)

● قضاوتی<sup>۱۱</sup> (J) – ادراکی<sup>۱۲</sup> (P)

از آنجا که نوشته‌های افراد نشان‌دهنده هویت آن‌ها است، می‌توان از نوشته‌های آن‌ها برای تشخیص شخصیت آن‌ها استفاده کرد [۱۲]. برای این منظور، مجموعه داده‌هایی<sup>۱۳</sup> با نوشته‌های افراد که با ویژگی‌های شخصیتی‌شان برچسب زده شده‌اند، مورد نیاز است. تاکنون چندین مجموعه داده برای کار پیش‌بینی ویژگی‌های شخصیتی به زبان انگلیسی و برخی از زبان‌های دیگر جمع‌آوری شده است، اما هیچ مجموعه داده‌ای در این زمینه برای زبان فارسی تهیه نشده است. با جستجوی عبارات تعیین‌کننده هویت و تهیه یک پرسشنامه، ما یک مجموعه داده با بیش از ۱.۵ میلیون توییت به زبان فارسی با برچسب نوع شخصیت MBTI آن‌ها جمع‌آوری کردیم. سپس از مدل برت<sup>۱۴</sup> برای ارزیابی اولین مجموعه داده فارسی استفاده نمودیم.

## ۱-۲ انگیزه‌های پژوهش

یکی از اصلی‌ترین انگیزه‌های این پژوهش جمع‌آوری اولین مجموعه داده‌ی این حوزه به زبان فارسی است. از آنجا که در طی سال‌های اخیر افراد زیادی تلاش کردند تا با استفاده از مجموعه داده‌های مختلف با توجه به متون افراد، ویژگی‌های شخصیتی آن‌ها را پیش‌بینی کنند، بنابراین ما تصمیم گرفتیم که مجموعه داده‌ای در این زمینه و به زبان فارسی جمع‌آوری کنیم تا دیگر محققان بتوانند مدل‌هایی با استفاده از این مجموعه داده طراحی کنند و به هدف خود، پیش‌بینی ویژگی‌های شخصیتی افراد با درصد دقت مناسب، برسند. همچنین با توجه به چالش‌های این پژوهش که در ادامه آن‌ها را به طور کامل توضیح خواهیم داد، می‌توانند تصمیم به بهبود و گسترده‌تر کردن این مجموعه داده بگیرند.

<sup>۹</sup> Thinking

<sup>۱۰</sup> Feeling

<sup>۱۱</sup> Judging

<sup>۱۲</sup> Perceiving

<sup>۱۳</sup> Dataset

<sup>۱۴</sup> BERT: Bidirectional Encoder Representations from Transformers

### ۱-۳ دستاوردهای پایان‌نامه

ما در این پژوهش توانستیم مجموعه داده‌ای با ۱۵۵۲۵۳۲ توییت که از ۹۳۸ کاربر توییت‌ر بدست آمده است، جمع‌آوری کنیم. لازم به ذکر است که توانسته‌ایم در مقایسه با برخی از مجموعه داده‌های انگلیسی زبان داده‌ی بیشتری جمع‌آوری کنیم که در فصل‌های آینده با جزئیات به نحوه‌ی جمع‌آوری آن‌ها پرداخته خواهد شد. در نهایت پس از ارزیابی‌هایی که با استفاده از مدل پارس برت (که گونه‌ای از مدل برت است) بر روی مجموعه داده انجام شد، توانستیم به درصد دقت ۵۲.۸٪ برسیم.

### ۱-۴ ساختار پایان‌نامه

ساختار این پایان‌نامه دارای شش فصل است که در ادامه ساختار هر فصل گفته شده است. فصل دوم به معرفی تعاریف و مفاهیم پایه در ادبیات موضوع پرداخته شده است. در فصل سوم به کارهای انجام شده در حوزه جمع‌آوری مجموعه داده برای تشخیص ویژگی‌هایی شخصیتی پرداخته شده است. در فصل چهارم روش‌های پیشنهادی برای جمع‌آوری مجموعه داده و مقایسه‌ی نتایج بدست آمده از آن‌ها به طور کامل توضیح داده شده است. در فصل پنجم تحلیل و ارزیابی مجموعه داده با مدل برت مورد بررسی قرار گرفته است و در فصل ششم به جمع‌بندی و نتیجه‌گیری از پژوهش انجام شده و ارائه راه‌های پیشنهادی برای کارهای آینده پرداخته شده است.

## فصل ۲

### مفاهیم پایه

در فصل پیشین، موضوع پژوهش و هدف از انتخاب آن را مورد بررسی قرار دادیم. حال در این فصل می‌خواهیم به برخی از مفاهیم پایه‌ای مربوط به پژوهش مانند مجموعه داده، شبکه‌های عصبی عمیق<sup>۱</sup>، جانمایی کلمات<sup>۲</sup>، ترنسفورمرها<sup>۳</sup> و مدل برت که به درک بهتر مطالب در ادامه‌ی پژوهش کمک می‌کند، بپردازیم.

#### ۲-۱ مجموعه داده

مجموعه داده، به مجموعه‌ای از داده‌های آماری یا رایانه‌ای مربوط به یک پایگاه داده گفته می‌شود، که با هدف یکپارچه نمودن داده‌ها، محتویات آن را در قالب یک جدول پایگاه داده یا یک ماتریس داده‌ای، تنظیم و مرتب می‌نمایند، که در آن هر ستون از پایگاه داده، نشان‌دهنده یک متغیر خاص است و هر ردیف نیز به یکی از اعضای مجموعه داده‌ی مورد نظر مرتبط می‌باشد. در شکل ۲-۱ بخشی از یک مجموعه داده نشان داده شده است که دارای سه ستون آیدی، توییت و برچسب می‌باشد. آیدی، یک شماره‌ی منحصر به فردیست که به هر داده از مجموعه داده تعلق می‌گیرد. توییت در این مثال از مجموعه داده همان داده‌ی مورد نظر مجموعه داده است و برچسب نیز خروجی مورد انتظار شبکه عصبی مورد نظر است که در این مثال برچسب‌ها T/F یکی از شاخصه‌های مدل MBTI می‌باشد. لازم به ذکر است چندین مشخصه است که ویژگی‌ها و ساختارهای یک

---

<sup>۱</sup> Deep Neural Network

<sup>۲</sup> Word Embeddings

<sup>۳</sup> Transformers

مجموعه داده را تعریف می‌کند. این مشخصه‌ها شامل تعداد و انواع متغیرها و اقدامات آماری مختلف مانند انحراف معیار است. این مقادیر می‌توانند عددی (اعداد واقعی یا صحیح) یا مانند متن توییت کارا کتری باشند. در کل مقادیر موجود در مجموعه داده می‌توانند از هر نوع تعریف شوند و هر معیار اندازه‌گیری داشته باشند.

id	tweet	label
506	<username> <U...> بهرحال مشخصه وصل شدین شیرینییی	T
77	<retweet><username>: <hashtag> یا ...خدا بیامرزتش	T
265	<username> <user...> (والا وضعیتت زیاد پیدا نیست	T
522	...چی دیدم هورمون عشق میتونه به درمان کرونا کمک ک	F
75	...به اسم نیست بفهمه تو توییت داری چ <username>	T
63	...من از ۷ سالگی تا تقریباً ۱۶: <retweet><username>	T
338	...همین ک <username> . لولو وارد میشوید <username>	T
325	...واقعا چرا ب <username> . (=امیرجاسمی <username>	T
704	...سرمو میکوبم رو بالش یا پتو. عمل کوب <username>	T
200	...مرسی ری <username> . تو که اند مرامی <username>	F

شکل ۲-۱: بخشی از یک نمونه مجموعه داده

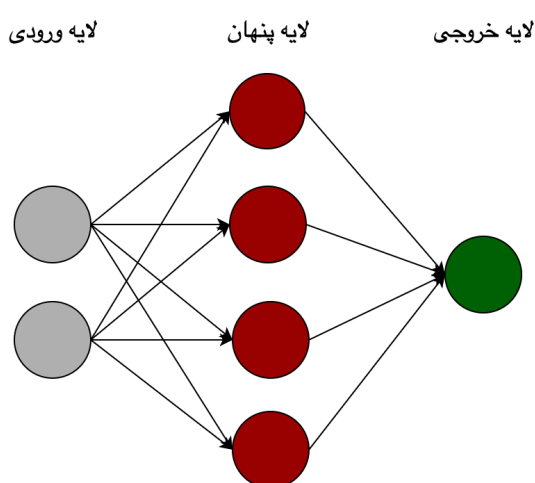
## ۲-۲ شبکه‌های عصبی عمیق

برای درک بهتر شبکه‌های عصبی عمیق لازم است ابتدا به شبکه‌های عصبی پردازیم. شبکه‌های عصبی دارای تعداد بسیار زیادی واحد کوچک به نام نورون هستند که به هم پیوسته‌اند و برای حل مسئله به طور موازی رفتار می‌کنند. شبکه‌های عصبی توسط ورودی‌ها آموزش داده می‌شوند و همانطور که در شکل ۲-۲ که یک نوع شبکه عصبی را نشان می‌دهد، مشاهده می‌کنید، شامل سه لایه ورودی، پنهان و خروجی هستند. هر کدام از عصب‌ها (یال‌ها) دارای مقدار آستانه<sup>۴</sup> و تابع فعال‌سازی<sup>۵</sup> می‌باشند که خروجی را تولید می‌کنند سپس نتیجه‌ی بدست آمده، با خروجی که انتظار داریم مقایسه می‌شود که این دو مقدار باید نزدیک به هم باشند. در نهایت مدل یاد می‌گیرد که وزن‌ها و مقدار آستانه را طوری تنظیم کند که خروجی مناسب دریافت کند. شبکه‌های عصبی دارای مزیت‌های مهمی از جمله ذخیره کردن اطلاعات در کل شبکه، توانایی کار با دانش ناکافی،

<sup>۴</sup>Threshold  
<sup>۵</sup>Activation Function



تحمل پذیری بالا در برابر داده‌های نویزی و دقت بالا در مسائل واقعی هستند اما معایبی همچون احتمال قرار گرفتن در ماکزیمم محلی، زمان آموزش زیاد و نیاز به تعیین پارامترهای تجربی را نیز دارا هستند. حال هرچه لایه‌های پنهان<sup>۶</sup> شبکه‌ی عصبی بیشتر شود، مدل پیچیده‌تر می‌شود که به این شبکه‌ها، شبکه‌های عصبی عمیق و به یادگیری آن‌ها یادگیری عمیق<sup>۷</sup> می‌گویند. به وسیله‌ی این شبکه‌های عصبی عمیق می‌توان مسائل پیچیده‌ای را حل کرد. در واقع یادگیری عمیق یک تابع است که ورودی را به خروجی تبدیل می‌کند و شبکه عصبی عمیق، ارتباط داده‌های ورودی و خروجی را پیدا می‌کند.



شکل ۲-۲: ساختار یک نمونه شبکه عصبی

## ۳-۲ جانمایی کلمات

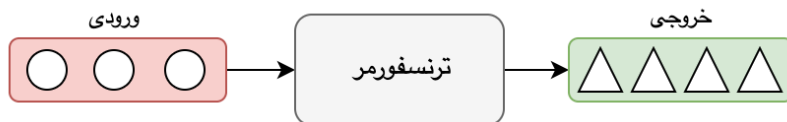
برای استفاده از کلمات به عنوان ورودی جهت پردازش متن، نیاز است کلمات به شکل عددی درآیند. در پردازش زبان طبیعی<sup>۸</sup> جانمایی کلمات اصطلاحی است که برای نمایش کلمات برای تجزیه و تحلیل متن استفاده می‌شود، به طور معمول در قالب یک بردار<sup>۹</sup> با اعداد واقعی است که معنی کلمه را رمزگذاری می‌کند به طوری که کلمات نزدیک به بردار در فضا، انتظار می‌رود از نظر معنایی مشابه باشند [۷]. جانمایی کلمات

<sup>۶</sup>Hidden Layer  
<sup>۷</sup>Deep Learning  
<sup>۸</sup>Natural Language Processing  
<sup>۹</sup>Vector

را می‌توان با استفاده از مجموعه‌ای از مدل‌سازی زبان<sup>۱۰</sup> و روش‌های یادگیری ویژگی<sup>۱۱</sup> به دست آورد که در آن کلمات یا عبارات به بردارهای اعداد واقعی نگاشت می‌شوند. از نظر مفهومی شامل یک جانمایی ریاضی از یک فضای با ابعاد زیاد برای هر کلمه به یک فضای پیوسته برداری با بعد<sup>۱۲</sup> بسیار پایین‌تر است. استفاده از روش‌های جانمایی کلمات در سال‌های اخیر بسیار رونق گرفته است. Word2Vec [۱۳] و GloVe [۱۵] دو روشی هستند که پیشتر از آن‌ها بسیار بهره می‌بردند. حال از مدل برت که در سال ۲۰۱۸ معرفی شده است [۳]، بیشتر استفاده می‌شود که در ادامه با این مدل بیشتر آشنا می‌شویم.

## ۲-۴ ترنسفورمر

ترنسفورمرها نوعی از معماری شبکه‌های عصبی هستند که محبوبیت بیشتری پیدا کرده‌اند. ترنسفورمرها اخیراً توسط OpenAI در مدل‌های زبانی مورد استفاده قرار گرفته است و همچنین توسط DeepMind نیز برای AlphaStar استفاده شده است. ترنسفورمرها برای حل مشکل انتقال توالی<sup>۱۳</sup> یا ترجمه ماشینی عصبی<sup>۱۴</sup> ساخته شده‌اند. این به معنای هر کاری است که یک توالی ورودی را به یک توالی خروجی تبدیل می‌کند که این کارها شامل تشخیص گفتار<sup>۱۵</sup>، تبدیل متن به گفتار<sup>۱۶</sup> و غیره می‌باشد.



شکل ۲-۳: مدل انتقال توالی توسط ترنسفورمر

ترنسفورمرها نیازی به پردازش داده‌های ترتیبی، به ترتیب ندارند. به عنوان مثال، اگر داده‌های ورودی جمله‌ای به زبان طبیعی باشد، ترنسفورمر نیازی به پردازش ابتدای آن قبل از انتهای آن ندارد. با توجه به این ویژگی، ترنسفورمرها نسبت به شبکه‌های عصبی بازگشتی<sup>۱۷</sup> موازی‌سازی بیشتری را فراهم می‌کنند. بنابراین

<sup>۱۰</sup> Language Modeling

<sup>۱۱</sup> Feature Learning Techniques

<sup>۱۲</sup> Dimension

<sup>۱۳</sup> Sequence Transduction

<sup>۱۴</sup> Neural Machine Translation

<sup>۱۵</sup> Speech Recognition

<sup>۱۶</sup> Text-to-Speech Transformations

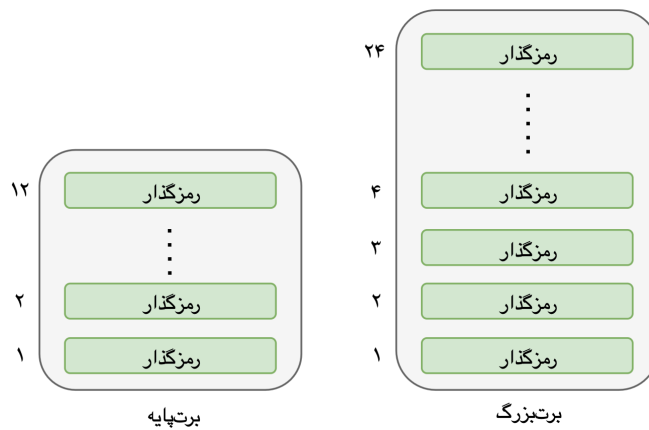
<sup>۱۷</sup> Recurrent Neural Network

زمان آموزش را کاهش می‌دهند و در نتیجه آموزش مجموعه داده‌های بزرگتر بیشتر از قبل امکان پذیر شده است. این امر منجر به توسعه سیستم‌های آموزش دیده‌ای مانند برت و GPT<sup>۱۸</sup> شده است که با مجموعه داده‌های بسیار بزرگ زبان عمومی مانند ویکی‌پدیا و کامن کرال<sup>۱۹</sup> آموزش دیده‌اند و می‌توانند برای کارهای زبانی خاص تنظیم شوند.

## ۲-۵ مدل برت

مدل برت یک روش یادگیری ماشین<sup>۲۰</sup> مبتنی بر ترنسفورمر برای پردازش زبان طبیعی است که همانطور که در بخش‌های قبل گفته شد، این مدل در سال ۲۰۱۸ توسط جکوب دولین و همکارانش در گوگل ایجاد و منتشر شد [۳].

همانطور که در شکل ۲-۴ مشاهده می‌کنید، برت در زبان اصلی انگلیسی دارای دو مدل برت پایه برت بزرگ است. هر دو سایز مدل برت دارای تعداد زیادی لایه‌ی رمزگذاری است که در مقاله آن‌ها را بلوک ترنسفورمر<sup>۲۱</sup> می‌نامند. نسخه پایه‌ی مدل دارای ۱۲ لایه ترنسفورمر و نسخه بزرگ آن دارای ۲۴ لایه ترنسفورمر می‌باشد.



شکل ۲-۴: دو نمونه از مدل برت

<sup>۱۸</sup>Generative Pre-trained Transformer

<sup>۱۹</sup>Common Crawl

<sup>۲۰</sup>Machine Learning

<sup>۲۱</sup>Transformer Blocks

این‌ها همچنین دارای شبکه‌های پیشخوان<sup>۲۲</sup> بزرگتر (به ترتیب ۷۶۸ و ۱۰۲۴ واحد پنهان) و سرهای توجه<sup>۲۳</sup> بیشتر (به ترتیب ۱۲ و ۱۶) نسبت به تنظیمات پیش فرض در پیاده‌سازی مرجع ترنسفورمر در مقاله‌ی اولیه هستند (۶ لایه رمزگذار، ۵۱۲ واحد پنهان و ۸ سر توجه).

## فصل ۳

### مروری بر کارهای مرتبط

در فصل قبل به مفاهیم پایه و تعاریف لازم مرتبط با موضوع پژوهش پرداخته شد. در این فصل می‌خواهیم کارهای مربوطه انجام شده در این حوزه را مورد بررسی قرار دهیم.

#### ۳-۱ بررسی کارهای انجام شده در این حوزه

مقالات مفید زیادی در مورد جمع‌آوری مجموعه داده با استفاده از شبکه‌های اجتماعی وجود دارد، اما همانطور که پیشتر ذکر شد، هنوز هیچ مقاله‌ای به زبان فارسی وجود ندارد. در زبان انگلیسی، مجموعه داده‌های مختلفی برای مدل‌های روانشناسی وجود دارد که در ادامه به هر یک به طور مجزا می‌پردازیم.

#### ۳-۱-۱ توییت

آقایان باربارا پلنک و دیرک هووی پژوهشی در این حوزه در بستر توییت انجام دادند که توانستند در نهایت بیش از ۱۰۲ میلیون توییت انگلیسی از ۱۵۰۰ کاربر توییت با ویژگی شخصیتی و جنسیتشان را جمع‌آوری کنند [۱۶]. آن‌ها برای جمع‌آوری این مجموعه داده ابتدا در توییت، کاربرانی که نوع شخصیت MBTI خود را اعلام کرده بودند را پیدا کردند و سپس توییت‌هایی که در آن‌ها یکی از ۱۶ نوع شخصیتی MBTI داخلشان بود را حذف کردند و به جای تمام حساب‌های کاربری ذکر شده در توییت‌ها کلمه‌ی یکتایی مانند USERNAME و همچنین برای تمام لینک‌ها کلمه‌ی یکتایی مانند LINK و همینطور برای هشتگ‌ها نیز کلمه‌ی HASHTAG را

جایگزین کردند.

### ۳-۱-۲ ردیت

آقایان ماتر جورکوویچ و یان اشنایدر در این حوزه در بستر ردیت<sup>۱</sup> مجموعه داده‌ای با ۳۵۴۹۹۶ پست که توسط ۹۸۷۲ کاربر منحصر به فرد گذاشته شده است، به همراه ویژگی شخصیتی و جنسیتشان به کمک فلیرز را جمع‌آوری کردند [۵] و آن را MBTI9K نامگذاری کردند. آن‌ها به این صورت عمل کردند که ابتدا به کمک فلیرزها که معمولاً کاربران یک توضیحاتی درباره‌ی خود مانند جنسیت و نوع شخصیت و غیره می‌نویسند، آن‌هایی که دارای کلمه‌ای از کلمات کلیدی MBTI را دارا بودند، جمع‌آوری کرد اما بعضی از آن‌ها به طور واضح کلمات را استفاده نکرده بودند به طور مثال لابه‌لای کلمه‌ی دیگری ذکر کرده بودند و به راحتی قابل شناسایی نبودند بنابراین دوباره حساب‌های کاربری جمع‌آوری شده را بررسی می‌کنند و آن‌هایی که واضحاً ویژگی شخصیتی خود را ذکر نکرده‌اند را حذف کردند زیرا الویتشان جمع‌آوری مجموعه داده‌ای با دقت بالاست البته تلاش‌هایی برای تشخیص دقیق فلیرزهای مبهم نیز کرده‌اند اما آن‌هایی شک داشتند و یا به طور مثال کاربرانی که در فلیرزهای مختلفشان ویژگی‌های شخصیتی متفاوتی را ذکر کرده‌اند نیز حذف شدند. پس از جمع‌آوری داده‌ها متوجه آن شدند که از بعضی از ویژگی‌های شخصیتی به تعداد کمی دارند در نتیجه جمله‌ی "I am an <TYPE>" را در نظرهای پست‌های مربوط به ویژگی‌های شخصیتی جستجو کردند و در نهایت یک لیستی از حساب‌های کاربری با نوع شخصیتی و جنسیتشان بدست آوردند.

### ۳-۱-۳ پاندورا

این مجموعه داده از اولین مجموعه داده‌هایی است در مقیاس بزرگ که در این زمینه توسط آقای ماتر جورکوویچ و همکارانش جمع‌آوری شده است. پاندورا دارای حدود ۱۷ میلیون نظر ردیت از بیش از ۱۰ هزار کاربر منحصر به فرد است که با ۳ مدل ویژگی شخصیتی MBTI، پنج عامله [۱۹] و Enneagram به همراه اطلاعاتی مانند سن و جنسیت و موقعیت مکانی، برچسب زده شده است. پاندورا برای قسمت نوع شخصیتی MBTI اش از مجموعه داده‌ی MBTI9K استفاده کرده و برای جمع‌آوری قسمت نوع شخصیتی Enneagram به این صورت عمل کردند که به صورت دستی آن کاربرانی که در فلیرز خود نوع شخصیتی Enneagram خود را اعلام کرده

<sup>۱</sup>Reddit

بودند را جمع‌آوری کردند و در مجموع برای نوع شخصیتی MBTI، ۹۰۸۴ کاربر و برای نوع شخصیتی Enneagram، ۷۹۳ کاربر به دست آوردند اما برای نوع شخصیتی پنج‌عامله به چالش‌های زیادی خوردند از جمله اینکه این مدل شخصیتی دارای کلید واژه‌ای مانند دیگر مدل‌های شخصیتی ذکر شده نیست پس جستجوی آن و پیدا کردن نوع شخصیتی افراد در ردیت برای این مدل شخصیتی چالش برانگیز است. بنابراین آن‌ها دیگر برای پیدا کردن ویژگی شخصیتی افراد در این مدل، فلیرزها را بررسی نکردند بلکه در نظرهایی<sup>۲</sup> که در زیر پست‌های مربوط به آزمون شخصیتی پنج‌عامله بود درصد ویژگی‌های شخصیتی این مدل را جستجو می‌کردند. از چالش‌های دیگر این مدل می‌توان یکسان نبودن نام پنج ویژگی شخصیتی آزمون‌های آنلاین متفاوت نام برد و همچنین امتیازهای این ویژگی‌ها در آزمون‌های مختلف به طرز متفاوتی مانند درصد و یا عدد خام ممکن است داده شده باشد و همچنین مبتنی بر چه توزیعی این امتیازها حساب شده است، نیز از چالش‌های دیگر این مدل می‌باشد. در نهایت با رفع تمام این چالش‌ها توانستند ۳۹۳ کاربر مجزا برای مدل پنج‌عامله جمع‌آوری کنند.

### ۳-۱-۴ فیس‌بوک

این مجموعه‌داده در سال ۲۰۰۷ توسط دیوید استیلول که اکنون مدرس دانشگاه کمبریج است، ایجاد شد و آن را myPersonality نامگذاری کرد و در سال ۲۰۰۹ دیوید به میشل کوسینسکی که اکنون مدرس دانشگاه استنفورد است، پیوست [۸]. مجموعه‌داده myPersonality یک اپلیکیشن فیس‌بوک بود که به کاربران خود اجازه می‌داد که با تکمیل کردن پرسشنامه‌ای در رابطه با ویژگی شخصیتی، در تحقیقات روانشناسی شرکت کنند و همچنین به آن‌ها بازخوردی از امتیازات آن‌ها ارائه می‌داد. این مجموعه‌داده در سال ۲۰۱۲ به دلیل کمبود وقت در نگهداری آن متوقف شد. کاربران آن‌ها می‌توانستند با به اشتراک گذاشتن داده‌های فیس‌بوک خود به تحقیقاتشان کمک کنند اما مجبور نبودند. حدود ۴۰٪ آن‌ها داوطلبانه این کار را انجام دادند. در نهایت مجموعه‌داده myPersonality در سال ۲۰۱۸ با بیش از ۶ میلیون داوطلب انتشارش متوقف شد و دلیل آن را سنگین شدن مسئولیت‌هایی نظیر حفظ مجموعه‌داده، بررسی پروژه‌ها، پاسخگویی به سوالات و مطابقت داشتن با مقررات مختلف برای این دو نفر، ذکر کردند.

## ۲-۳ دلیل عدم انتخاب مدل پنج‌عامله

ما پس از بررسی‌هایی که انجام دادیم متوجه آن شدیم که برخی از دانشمندان سوالاتی در مورد میزان اعتبار ویژگی‌های شخصیتی MBTI وارد کردند اما در آزمون استخدامی بسیاری از شرکت‌های بزرگ از آن استفاده می‌شود و همچنین شهرت زیادی در میان مردم دارد اما ویژگی شخصیتی پنج‌عامله در جامعه روان‌شناسی با مقبولیت بیشتری مواجه شده است و نسبت به ویژگی‌های شخصیتی MBTI قابل استنادتر است اما جمع‌آوری مجموعه داده‌ای با برجسب ویژگی شخصیتی پنج‌عامله چالش‌هایی به همراه داشت که ادامه دادن پژوهش را برای ما دوچندان دشوار می‌کرد. در ادامه این چالش‌ها را بررسی کردیم و راه‌حل‌هایی برای آن‌ها مطرح نمودیم. نوع شخصیتی پنج‌عامله دارای شهرت زیادی در میان مردم ایران نیست و به این معنی است که افراد بسیار کمی با آن آشنایی دارند و در آزمون آن شرکت کرده‌اند. همچنین در زمان انجام پژوهش، در بستر توییت، توییت‌های کمی درباره‌ی این مدل شخصیتی موجود بود که تأییدکننده‌ی شهرت کم این مدل شخصیتی در میان مردم بود و همچنین به دلیل نداشتن کلید واژه‌ای برای این مدل شخصیتی امکان جستجوی ویژگی‌های شخصیتی را در بستر توییت دشوارتر می‌نمود. راه‌حلی که برای این چالش در نظر داشتیم به این صورت بود که توضیح کاملی از این مدل شخصیتی و برتری آن نسبت به مدل شخصیتی MBTI در پرسشنامه بدهیم و لینک آزمون آن را نیز قرار دهیم تا افرادی که تمایل به کمک در تحقیقات دارند، به راحتی در آزمون شرکت کنند اما با توجه به نتایجی که از انتشار اولیه پرسشنامه بدست آوردیم و داده‌های نامعتبری که با وجود توضیحات کامل دریافت نمودیم، متوجه آن شدیم که این راه‌حل، بدون حمایت مالی کمکی به دریافت داده‌های معتبر و زیاد نمی‌کند. چالش دیگری که این مدل شخصیتی برای ما داشت، انگلیسی بودن آزمون اصلی این مدل بود که منجر به این می‌شد که افرادی که حتی توضیحات را خوانده و تمایل به کمک در تحقیقاتمان دارند، به دلیل ضعف در فهم زبان انگلیسی از تکمیل کردن پرسشنامه صرف نظر کنند. راه‌حلی که برای این چالش در نظر گرفتیم، ترجمه‌ی آزمون اصلی بود به این صورت که پاسخ‌های وارد شده توسط فرد مورد نظر را در آزمون اصلی به طور خودکار وارد کنیم و سپس پاسخ دریافت شده از آن را ترجمه کرده و در قالب ایمیل برای فرد مورد نظر ارسال کنیم. لازم به ذکر است که این راه‌حل را آزمایش نمودیم اما چالش اصلی ما شهرت کم این مدل شخصیتی در میان مردم بود که با شرایط ذکر شده، رسیدن به داده‌ی زیاد را برای ما غیرممکن نموده بود. در نتیجه ما تصمیم به ادامه دادن پژوهش با ویژگی شخصیتی MBTI گرفتیم.



## فصل ۴

### جمع‌آوری مجموعه داده

در این فصل نحوه جمع‌آوری مجموعه داده را بررسی می‌کنیم. مجموعه داده از دو قسمت اصلی یعنی مجموعه تویت‌های کاربرانی که دارای حساب کاربری عمومی در توییتر هستند و ویژگی‌های شخصیتی هر فرد، تشکیل شده است. برای دستیابی به این شکل از داده، لازم است ویژگی‌های شخصیتی هر فرد را به دست آوریم. همانطور که نتیجه گرفتیم، این کار را می‌توان در دو روش اصلی انجام داد. اولین و آسانترین روشی که توسط آقایان باربارا پلنک و دیرک هووی هم استفاده شده است [۱۷]، جستجوی کلید واژه و یا جملاتی که نشانگر ویژگی شخصیتی باشد که در مورد این پژوهش، کلید واژه‌های MBTI مانند "INTP" یا جملاتی مانند "من ESTJ هستم"، می‌باشد. این روش کمک کرد تا ۹۲٪ از اطلاعات مجموعه داده را جمع‌آوری کنیم. دومین روشی که برای جمع‌آوری داده استفاده نمودیم، توزیع پرسشنامه برای پرسیدن مستقیم ویژگی‌های شخصیتی MBTI از افراد است به شرط داشتن حساب کاربری عمومی در توییتر و مهم‌تر از همه تمایل به اشتراک‌گذاری تویت‌هایشان در جهت کمک به این پژوهش بود.

#### ۴-۱ جستجو کلیدواژه

در بستر توییتر، دو منبع اصلی بالقوه وجود دارد که افراد ویژگی شخصیتی MBTI خود را ذکر می‌کنند: بیو و تویت. با این حال، جستجو در این دو بخش باید از نظر روش‌شناسی<sup>۱</sup> با هم متفاوت باشند زیرا توییتر ابزار

<sup>۱</sup>Methodology

مناسبی برای یافتن کلمات کلیدی در بیو ارائه نمی‌دهد. اگرچه برخی از بسترهای شخص ثالث ادعا می‌کنند که این ویژگی را در اختیار قرار می‌دهند، اما کمیت و کیفیت داده‌های آن‌ها، به ویژه در زبان فارسی، ناشناخته است و همچنین منبع مالی کافی برای تأمین هزینه‌ها نداشتیم. از این رو، ما چندین کاربر شناخته شده که دارای تعداد زیادی دنبال‌کننده<sup>۲</sup> هستند را به عنوان گره‌های<sup>۳</sup> اصلی جمع‌آوری کردیم و با بررسی اینکه کدامیک از دنبال‌کنندگان یا دنبال‌شوندگان<sup>۴</sup> آن‌ها یکی از کلمات کلیدی MBTI را در بیو خود قرار داده است، جستجو را ادامه و گسترش دادیم. علاوه‌براین از رهنمودهای ذکر شده توسط باربارا پلنک و دیرک هووی همچنین در توییت‌های جمع‌آوری شده از طریق ویژگی جستجوی پیشرفته، که توسط API توییت ارائه شده، استفاده می‌کنیم.

## ۲-۴ پرسشنامه

برای جمع‌آوری ویژگی‌های شخصیتی افراد، پرسشنامه‌ای را طراحی و در کانال‌های دانشگاهی مختلف و گروه‌های متفرقه تلگرامی توزیع کردیم. پس از اینکه آگاهی کافی درباره‌ی پژوهش خود به افراد دادیم و اطمینان دادیم که هیچ‌گونه داده‌ی شخصی مانند حساب کاربری آن‌ها در مجموعه داده قرار نمی‌گیرد، رضایت آن‌ها را جلب کردیم و از آن‌ها درخواست کردیم تا با تکمیل کردن پرسشنامه، حساب کاربری عمومی توییت خود، جنسیت و ویژگی شخصیتی MBTI خود را در اختیارمان قرار دهند. از آنجا که بسیاری از افراد با آزمون MBTI<sup>۵</sup> آشنایی نداشتند اما مایل به کمک در این پژوهش بودند، ما از دستورالعمل‌ها و لینک‌هایی برای کمک به آن‌ها در انجام آزمون استفاده کردیم.

شایان ذکر است که برای ما چالش برانگیزترین قسمت این پژوهش، همانطور که در شکل ۴-۱ مشاهده می‌کنید، درصد قابل توجه ارسالی‌های نامعتبر در روزهای اول انتشار پرسشنامه بود که دلیل آن ورودی نادرست در بخش حساب کاربری توییت، با وجود جملات آگاه‌کننده، توضیحات کامل و تایید اعتبار خودکار ساده بود.

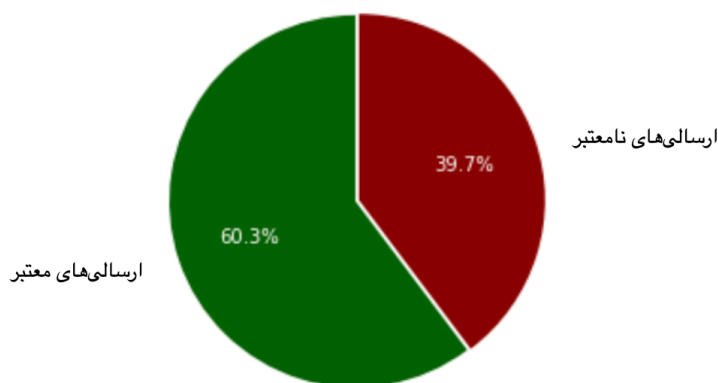
برای غلبه بر این چالش، عدم مطالعه‌ی دستورالعمل‌ها به طور کامل، زمان زیادی صرف طراحی یک

<sup>۲</sup> Followers

<sup>۳</sup> Nodes

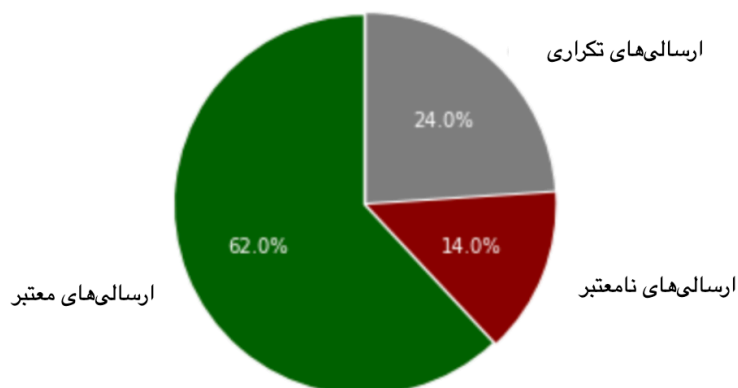
<sup>۴</sup> Followings

<sup>۵</sup> برای مشاهده آزمون MBTI می‌توانید به سایت <https://www.16personalities.com/fa> مراجعه کنید.



شکل ۴-۱: نسبت اعتبار ارسالی‌های پرسشنامه اولیه

پرسشنامه جدید با منطق انشعاب با استفاده از فرم‌های میکروسافت<sup>۶</sup> نمودیم و همچنین محیط انتشار پرسشنامه را به تویتر تغییر دادیم و همانطور که در شکل ۴-۲ مشاهده می‌کنید، موفق شدیم شمار ارسالی‌های نامعتبر را تا حد بسیار زیادی کاهش دهیم.



شکل ۴-۲: نسبت اعتبار ارسالی‌های پرسشنامه نهایی

### ۴-۳ مقایسه روش‌های پیشنهادی

روش نخست به ما کمک کرد تا بتوانیم ۱۴۴۳۶۵۸ توییت که ۹۲٪ از مجموعه داده را تشکیل می‌دهد، را جمع‌آوری کنیم. همانطور که در جدول ۴-۱ مشاهده می‌کنید، اختلاف زیادی بین نتایج بدست آمده از دو روش استفاده شده، وجود دارد که از نظر ما این اختلاف دلایل مختلفی دارد. ابتدا اینکه متقاعد کردن افراد

<sup>۶</sup> برای مشاهده پرسشنامه طراحی شده می‌توانید به لینک <https://cutt.ly/hzU73Jo> مراجعه کنید.

برای تکمیل کردن پرسشنامه کاری بسیار سخت و طاقت‌فرسا می‌باشد. برای غلبه بر این چالش، بسیاری از محققان از روش‌های انگیزشی و گاه‌آجباری استفاده می‌کنند که برای ما غیرممکن و غیراخلاقی تلقی می‌شد. دوم، یافتن جامعه هدفی که به این موضوع علاقه‌مند هستند و یا در این باره دانش و آگاهی دارند، چالش‌برانگیز بود. برای این منظور، یک ایده می‌تواند این باشد که از افرادی که دارای دنبال‌کنندگان زیادی هستند بخواهیم که پرسشنامه‌ی طراحی شده را مجدد توزیع کنند. این ایده را با درخواست مستقیم از ۲۱ کاربر برای کمک در انجام این پژوهش آزمایش کردیم و فقط دو نفر از آن‌ها اعلام همکاری (توزیع مجدد پرسشنامه‌ی ما) کردند. عدم حمایت مالی، کمبود ابزار مناسب و همچنین تجربه‌ی ناکافی را می‌توان از دلایل دیگر عدم توانایی ما در دستیابی به نتایج مورد انتظار از روش دوم دانست. با این حال، این روش در مقایسه با روش اول، این مزیت را دارد که نیازی به مداخله مداوم ندارد و می‌تواند به طور موازی پیش برود.

جدول ۴-۱: آمار کلی روش‌های جمع‌آوری داده‌ها

	روش ۱		روش ۲	
	بیو	توییت	پرسشنامه	کل
شمار توییت‌ها	۳۰۹۳۶۴	۱۱۳۴۲۹۴	۱۰۸۸۷۴	۱۵۵۲۵۳۲
شمار کاربرها	۲۱۰	۶۵۳	۷۵	۹۳۸

#### ۴-۴ جمع‌آوری توییت‌ها

پیش از آغاز روند جمع‌آوری توییت‌ها باید ابتدا کاربران نامعتبر را فیلتر کنیم. در این بخش نامعتبر به معنی عمومی نبودن حساب کاربری و یا عدم دارا بودن بیشتر از ۱۰۰ توییت است. ما برای جمع‌آوری توییت‌ها از توییت‌ر دو روش را آزمایش کردیم. ابتدا با استفاده از سلنیوم [۶] به جمع‌آوری توییت‌ها پرداختیم اما مدتی بعد متوجه آن شدیم که این روش برخلاف قوانین توییت‌ر است و با استفاده از روش‌هایی مانند reCAPTCHA [۲۰] هر چند دقیقه یکبار مانع جمع‌آوری توییت‌ها می‌شد. اگرچه ممکن است راه‌های دیگری برای رفع کردن این مشکل وجود داشته باشد اما اصرار بر ادامه دادن با این رویکرد برای ما به دلیل زمان‌بر بودن آن، مناسب نبود و می‌توانست مسائل حقوقی نیز پیش بیاورد. با وجود محدودیت‌های استفاده زیاد از API رسمی توییت‌ر، آن را به عنوان ابزار اصلی خود برای جمع‌آوری توییت‌ها انتخاب کردیم و توانستیم ۱۵۵۲۵۳۲ توییت از ۹۳۸ کاربر جمع‌آوری کنیم.

## ۴-۵ تمیزکردن داده

پس از جمع‌آوری داده‌ها، نیاز است تا آن‌ها را تمیز و تا حدودی ناشناس و آماده برای استفاده کرد. ما نیز به این منظور تمام حساب‌های کاربری را با یک کلمه منحصر به فرد و ناشناخته "`<USERNAME>`"، تمام لینک‌ها را با یک کلمه منحصر به فرد "`<LINK>`" و همچنین تمام هشتگ‌ها را با "`<HASHTAG>`" جایگذاری کردیم.

لازم به ذکر است که از ویژگی شناسایی زبان API تویتر، برای حذف توییت‌های غیر فارسی بهره بردیم و همچنین در نظر داریم که شکلک‌های موجود در متن توییت‌ها را با یک کلمه منحصر به فرد دیگر مانند "`<EMOJI>`" جایگذاری کنیم، اما به دلیل ارزش بالای آن در برخی مدل‌های پردازش زبان طبیعی، تصمیم گرفتیم آن‌ها را در مجموعه داده حفظ کنیم تا هر شخصی بر اساس استفاده و پژوهش خود از این مجموعه داده استفاده کند.

## فصل ۵

### تحلیل و ارزیابی مجموعه داده

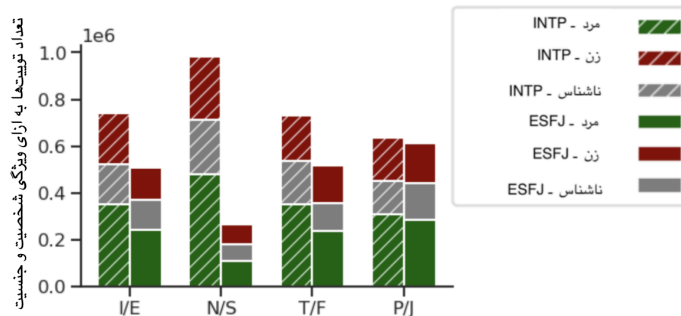
در این فصل می‌خواهیم به تجزیه و تحلیل مجموعه داده‌ی بدست آمده و همچنین ارزیابی آن بپردازیم. برای ارزیابی مجموعه داده و به نمایش گذاشتن مدلی که روی مجموعه داده‌ی ما آموزش دیده است، ما ابتدا داده‌ها را مجدداً قالب‌بندی کردیم سپس با استفاده از روش اعتبارسنجی متقابل طبقه‌ای تکرارشونده K بار<sup>۱</sup> یک مدل مبتنی بر برت را آموزش دادیم و اعتبارسنجی کردیم.

#### ۵-۱ تحلیل مجموعه داده

پس از جمع‌آوری توییت‌های ۹۳۸ کاربر منحصر به فرد و آماده‌سازی آن‌ها برای استفاده، به تجزیه و تحلیل آن‌ها پرداختیم. شمار کل توییت‌های جمع‌آوری شده ۱۵۵۲۵۳۲ توییت است و از این رو به طور متوسط، هر کاربر دارای ۱۶۵۵ توییت است. ما تصمیم گرفتیم برچسب‌ها را به چهار دسته I/E, N/S, T/F, P/J تقسیم کنیم. با این فرض که این چهار ویژگی در واقع مستقل باشند. شکل ۵-۱ عدم تعادل در داده‌های مربوط به ویژگی‌های I/E و N/S را نشان می‌دهد.

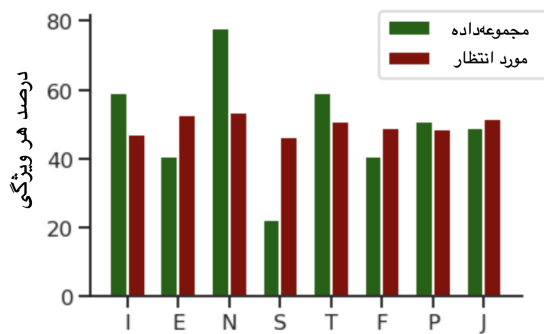
در ایران با بیش از ۸۲ میلیون نفر جمعیت، ۶۲۵۱۹ نفر آزمون MBTI را در وب‌سایت رسمی اش شرکت کرده‌اند. درصد افرادی که تشکیل‌دهنده هر یک از این ویژگی‌های شخصیتی هستند را بدست آوردیم و مقایسه‌ای با درصدهای بدست آمده از مجموعه داده‌ی خود انجام دادیم. همانطور که در شکل ۵-۲ مشاهده

<sup>۱</sup> Repeated Stratified K-fold Cross-Validation



شکل ۵-۱: توزیع ویژگی‌های شخصیتی بر اساس جنسیت در مجموعه داده

می‌کنید، افراد درون‌گرا کمتر از شبکه‌های مجازی در جامعه حضور دارند زیرا این افراد بهتر قادر به فعالیت و معاشرت در فضای مجازی هستند و در آنجا به راحتی ابراز احساسات می‌کنند.



شکل ۵-۲: مقایسه توزیع MBTI در مجموعه داده با توزیع MBTI در جمعیت ایران

## ۲-۵ آماده سازی داده

برای تزریق داده‌ها به مدل برت لازم است مجموعه توییت‌های مربوط به هر کاربر را به چند بخش تقسیم کنید، زیرا ورودی این مدل در شمار توکن‌هایی<sup>۲</sup> که می‌تواند پردازش کند، دارای محدودیت است. همچنین بسیار مهم است که به خاطر داشته باشید که برای تقسیم داده‌ها به دسته آموزش، آزمایش و ارزیابی<sup>۳</sup>، همه‌ی توییت‌ها در یک مجموعه قرار داشته باشند.

<sup>۲</sup> Tokens  
<sup>۳</sup> Validation

از آنجا که ما قصد تنظیم بیش از حد پارامترها را نداریم، از  $K$  بار<sup>۴</sup> تودرتو<sup>۵</sup> استفاده نکردیم از  $K$  بار با  $K$  برابر با ۵ استفاده کردیم و همانطور که در شکل ۳-۵ نشان داده شده است، روند تقسیم داده‌ها را پنج بار با تقسیم‌بندی تصادفی بارزیابی مدل جلوگیری کنیم. همانطور که پیشتر اشاره شد، برای کاهش تأثیر عدم تعادل برخی از دسته‌ها بر تقسیمات، از  $K$  بار استفاده کردیم. علاوه‌براین، قبل از تزریق داده‌ها به مدل، با نمونه‌برداری از گروهی که شمار داده‌ها در آن بیشتر است، داده‌ها را در هر دسته متعادل می‌کنیم.

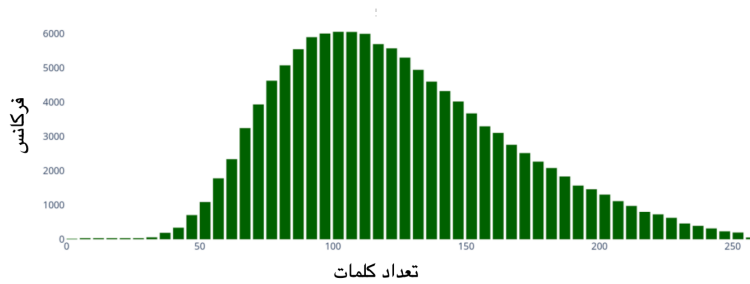


شکل ۳-۵: مراحل انجام روش اعتبارسنجی متقارن طبقه‌ای تکرارشونده  $K$  بار

در نهایت، تصمیم گرفتیم داده‌ها را در هر مجموعه به چند توییت به اندازه ۱۰ تقسیم کنیم. توزیع شمار توکن‌ها در یک نمونه آزمایش در شکل ۴-۵ نشان داده شده است. با توجه به منابع و توزیع ذکر شده، تصمیم گرفتیم حداکثر طول ۲۵۶ توکن برای هر ورودی مدل برت در نظر بگیریم. توکن‌های اضافی حذف می‌شوند و آن ورودی‌هایی که توکن‌های کمتری دارند نیز تکمیل نشده هستند.

<sup>۴</sup>K-fold  
<sup>۵</sup>Nested





شکل ۴-۵: توزیع تعداد کلمات نمونه‌ای از داده‌های آموزش (تعداد بین برابر ۶۴)

### ۳-۵ معرفی مدل پایه

ما از مدل پارس‌برت، یک مدل مبتنی بر برت، که بر روی متون فارسی از قبل آموزش دیده شده و از یک رده‌بند<sup>۶</sup> لاجستیک رگرسیون<sup>۷</sup> در بالای رمزگذاری [CLS] استفاده کردیم [۴]. ما با استفاده از داده‌های خود با ثابت چهار تکرار برای هر دوره، این مدل را دقیق تنظیم کردیم<sup>۸</sup>. همانطور که پیشتر ذکر شد، ۵ بخش با ۵ بار تکرار و چهار ویژگی به عنوان دسته داریم. از این رو ۱۰۰ تکرار باید انجام دهیم تا به نتایج ارزیابی دقیق برسیم اما به دلیل نداشتن ابزار مناسب و قوی برای اجرای این تعداد روی کل حجم مجموعه داده، ما قبل از هر اجرا با استفاده از نمونه‌برداری فرعی<sup>۹</sup> به صورت تصادفی حجم ورودی مدل را به قدری کاهش دادیم تا مدت اجرای هر آموزش به ۱۵ الی ۲۰ دقیقه کاهش یابد.

### ۴-۵ نتایج بدست آمده

ما عمدتاً از معیار متوسط کلان امتیاز F1<sup>۱۰</sup> برای ارزیابی عملکرد مدل بر روی مجموعه داده خود استفاده کردیم. نتایج در جدول ۲ با جمع کردن میانگین هر معیار برای هر تکرار ارائه شده است. شایان ذکر است که قبل از آموزش مدل، داده‌های خود را متعادل<sup>۱۱</sup> می‌کنیم از این رو پایه اصلی اکثریت برای همه ۵۰٪ است. در آزمایشات ما، میزان فراخوانی بسیار بالاتر از دقت بود و میانگین کلی ۵۲.۸٪ بود.

Classifier<sup>۶</sup>  
 Logistic Regression<sup>۷</sup>  
 Fine-tune<sup>۸</sup>  
 Subsampling<sup>۹</sup>  
 F1-score macro average<sup>۱۰</sup>  
 Balance<sup>۱۱</sup>

جدول ۵-۱: میانگین متوسط کلان امتیاز F1 نسبت به نتایج پنج تکرار اعتبارسنجی متقابل طبقه‌ای K بار

شمار تکرار	I/E	N/S	T/F	P/J
۱	.	.	.	.
۲	.	.	.	.
۳	.	.	.	.
۴	.	.	.	.
۵	.	.	.	.
میانگین	.	.	.	.

## فصل ۶

### نتیجه‌گیری و کارهای آینده

#### ۱-۶ نتیجه‌گیری

باتوجه به جالب بودن و مورد توجه قرار گرفتن این موضوع یعنی تشخیص شخصیت افراد، در حوزه پردازش زبان طبیعی، ما نیز در این پژوهش سعی کردیم تا اولین مجموعه‌داده فارسی این زمینه را جمع‌آوری کنیم تا دیگر محققان بتوانند از آن استفاده و مدل‌هایی با مجموعه‌داده فارسی طراحی کنند و به نتایج مناسب برسند و یا افراد علاقه‌مند به این زمینه با استفاده از روش‌ها و تحلیل‌های انجام شده و همچنین چالش‌های مطرح شده در این پژوهش، راه آسان‌تری برای جمع‌آوری مجموعه‌داده فارسی در پیش داشته باشند.

درنهایت ما توانستیم مجموعه‌داده‌ای با استفاده از نوشته‌های افراد در بستر توییتر به زبان فارسی جمع‌آوری کنیم که همانطور که پیشتر ذکر شد، در مقایسه با برخی از مجموعه‌داده‌های انگلیسی زبان با توجه به شرایط ذکر شده، داده بیشتری تهیه کنیم.

#### ۲-۶ کارهای آینده

ما قصد داریم برای بهبود داده‌های جمع‌آوری شده، عکس‌های افراد را به ParsTSet اضافه کنیم تا از ویژگی‌های صورت نیز برای بهبود مدل‌سازی و درصد دقت بدست آمده، استفاده کنیم. علاوه بر این غلبه بر چالش‌های موجود در جمع‌آوری داده‌ها با استفاده از آزمون‌های شناخته شده شخصیتی پنج‌عامله، می‌تواند دروازه‌ای برای

پیشرفت بیشتر در عملکرد مدل‌سازی ویژگی‌های روانشناختی یک متن به زبان فارسی در نظر گرفته شود و همچنین جمع‌آوری داده‌های بیشتر به بهبود عملکرد مجموعه داده کمک می‌کند.

## مراجع

- [1] Balmaceda, J. M., Schiaffino, S., and Godoy, D. How do personality traits affect communication among users in online social networks? *Online Information Review* (2014).
- [2] Corr, P. J., and Matthews, G. *The Cambridge handbook of personality psychology*. Cambridge University Press, 2020.
- [3] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 4171–4186.
- [4] Farahani, M., Gharachorloo, M., Farahani, M., and Manthouri, M. Parsbert: Transformer-based model for persian language understanding. *ArXiv abs/2005.12515* (2020).
- [5] Gjurkovi, M., and Šnajder, J. Reddit: A gold mine for personality prediction. in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media* (New Orleans, Louisiana, USA, June 2018), Association for Computational Linguistics, pp. 87–97.
- [6] Jason Huggins, Paul Gross, J. T. W. Selenium automates browsers. <https://www.selenium.dev/>, 2004.
- [7] Jurafsky, D., and Martin, J. H. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 2009.
- [8] Kosinski, M., Stillwell, D., and Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences* 110, 15 (2013), 5802–5805.

- [9] Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Wicaksana, A. S., Born, M. P., and König, C. J. Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. in *Explainable and interpretable models in computer vision and machine learning*. Springer, 2018, pp. 197–253.
- [10] Martin, C. R. *Looking at type: The fundamentals*. Center for Applications of Psychological Type, 1997.
- [11] Matz, S. C., Kosinski, M., Nave, G., and Stillwell, D. J. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences* 114, 48 (2017), 12714–12719.
- [12] Mehta, Y., Majumder, N., Gelbukh, A., and Cambria, E. Recent trends in deep learning based personality detection. *Artificial Intelligence Review* (2019), 1–27.
- [13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26 (10 2013).
- [14] Myers, I. B. *The myers-briggs type indicator: Manual* (1962).
- [15] Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. volume 14, pp. 1532–1543.
- [16] Plank, B., and Hovy, D. 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* (2015), pp. 92–98.
- [17] Plank, B., and Hovy, D. Personality traits on twitter—or—how to get 1,500 personality tests in a week. pp. 92–98.
- [18] Poria, S., Cambria, E., Hazarika, D., and Vij, P. A deeper look into sarcastic tweets using deep convolutional neural networks. *Proceedings of COLING* (10 2016).
- [19] Rothmann, S., and Coetzer, E. P. The big five personality dimensions and job performance. *SA Journal of Industrial Psychology* 29, 1 (2003).
- [20] team, G. What is recaptcha? <https://www.google.com/recaptcha/about/>.
- [21] Yang, H.-C., and Huang, Z.-R. Mining personality traits from social messages for game recommender systems. *Knowledge-Based Systems* 165 (2019), 157–168.

# واژه‌نامه فارسی به انگلیسی

Activation Function	تابع فعالسازی
Attention Heads	سرهای توجه
Balance	متعادل
Classifier	رده‌بند
Comments	نظرات
Dataset	مجموعه داده
Deep Learning	یادگیری عمیق
Deep Neural Networks	شبکه‌های عصبی عمیق
Dimension	بعد
Extroversion	برونگرایی
F1-score macro average	متوسط کلان امتیاز F1
Feature Learning Techniques	روش‌های یادگیری ویژگی
Feedforward-networks	شبکه‌های پیشخوان
Feeling	احساس
Fine-tune	با دقت تنظیم کردن
Fold	بار
Follower	دنبال‌کننده
Following	دنبال‌شونده
Hidden Layer	لایه پنهان
Introversion	درونگرایی
Intuition	شهود
Job Screening	غربالگری شغل
Judging	قضاوت
Language Models	مدل‌های زبانی
Machine Learning	یادگیری ماشین
MBTI: Myers-Briggs Type Indicators	شاخص‌های نوع مایرز-برگز

Methodology . . . . .	روش‌شناسی . . . . .
Natural Language Processing . . . . .	پردازش زبان طبیعی . . . . .
Nested . . . . .	تودرتو . . . . .
Neural Machine Translation . . . . .	ترجمه ماشین عصبی . . . . .
Node . . . . .	گره . . . . .
Perceiving . . . . .	ادراک . . . . .
Recommendation System . . . . .	سیستم‌های توصیه . . . . .
Recurrent Neural Network . . . . .	شبکه‌های عصبی بازگشتی . . . . .
Repeated Stratified K-fold Cross-Validation . . . . .	اعتبارسنجی متقابل طبقه‌بندی شده k بار . . . . .
Sensing . . . . .	حس کردن . . . . .
Sequence Transduction . . . . .	انتقال توالی . . . . .
Speech Recognition . . . . .	تشخیص گفتار . . . . .
Text-to-Speech Transformations . . . . .	تبدیل متن به گفتار . . . . .
Thinking . . . . .	تفکر . . . . .
Threshold . . . . .	آستانه . . . . .
Vector . . . . .	بردار . . . . .
Word Embeddings . . . . .	جانمایی کلمات . . . . .
Word Polarity Detection . . . . .	تشخیص قطبیت کلمات . . . . .



# واژه‌نامه انگلیسی به فارسی

Threshold	آستانه
Feeling	احساس
Perceiving	ادراک
Repeated Stratified K-fold Cross-Validation	اعتبارسنجی متقابل طبقه‌بندی شده k بار
Sequence Transduction	انتقال توالی
Fine-tune	با دقت تنظیم کردن
Fold	بار
Vector	بردار
Dimension	بعد
Extroversion	برونگرایی
Natural Language Processing	پردازش زبان طبیعی
Activation Function	تابع فعالسازی
Text-to-Speech Transformations	تبدیل متن به گفتار
Neural Machine Translation	ترجمه ماشینی عصبی
Word Polarity Detection	تشخیص قطبیت کلمات
Speech Recognition	تشخیص گفتار
Thinking	تفکر
Word Embeddings	جانمایی کلمات
Nested	تودرتو
Sensing	حس کردن
Introversion	درونگرایی
Follower	دنبال‌کننده
Following	دنبال‌شونده
Classifier	رده‌بند
Methodology	روش‌شناسی
Feature Learning Techniques	روش‌های یادگیری ویژگی

Attention Heads	سرهای توجه
Recommendation System	سیستم‌های توصیه
MBTI: Myers-Briggs Type Indicators	شاخص‌های نوع مایرز-برگز
Feedforward-networks	شبکه‌های پیشخوان
Recurrent Neural Network	شبکه‌های عصبی بازگشتی
Deep Neural Networks	شبکه‌های عصبی عمیق
Intuition	شهود
Job Screening	غربالگری شغل
Judging	قضاوت
Node	گره
Hidden Layer	لایه پنهان
Balance	متعادل
Dataset	مجموعه داده
Language Models	مدل‌های زبانی
F1-score macro average	متوسط کلان امتیاز F1
Comments	نظرات
Deep Learning	یادگیری عمیق
Machine Learning	یادگیری ماشین

**Abstract:**

In recent years, recognizing individuals' personality traits through social media has become an interesting topic in both fields of natural language processing and social sciences. Psychological research also shows that some personality traits are associated with language behavior. NLP models can take advantage of this correlation to model and predict personality traits based on the huge amount of data made available thanks to modern social media. However, in Persian, there had been no dataset relating to this task before we attempted to collect and compile the first dataset via Twitter. As discussed in this article, we have constructed a novel dataset labeled with Myers-Briggs Type Indicators (MBTI) consisting of 1,552,532 tweets. We have presented our data collection methodologies and discussed their challenges and results in detail. As a baseline for other researchers to further improve upon, we have introduced a model by fine-tuning a variation of BERT architecture (ParsBERT), pre-trained on Persian corpora. Finally, we validated this model using the repeated stratified K-fold cross-validation method and published the results.

**Keywords:** Personality Traits, Dataset, Data Science, Social Data, ParsBERT



**Iran University of Science and Technology  
Computer Engineering Department**

# **A Persian Dataset for Personality Detection on Twitter**

**A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree  
of Master of Science in Computer Engineering**

**By:**

**Zahra Anvarian**

**Student ID:**

**95521054**

**Supervisor:**

**Dr. Sauleh Etemadi**

**April 2021**