دانشکده مهندسی کامپیوتر

مجموعه‌داده فارسی برای تشخیص شخصیت در بستر توییتر

پایان‌نامه 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 توییتر

تاریخ دفاع: خرداد ۱۴۰۰

رشته: مهندسی کامپیوتر

گرایش: هوش مصنوعی و رباتیک

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

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. درصورت 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 درخصوص 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 عبداله‌پور 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.

زهرا 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 مجموعه‌داده در زبان فارسی را از طریق شبکه اجتماعی توییتر جمع‌آوری و تدوین کنیم، هیچ مجموعه‌داده‌ای در این زمینه در زبان فارسی وجود نداشته است. همانطور که در این مقاله مورد بحث قرار گرفت، ما یک مجموعه‌داده جدید ساخته‌ایم که دارای برچسب شاخص‌های مدل مایرز_بریگز و متشکل از ۱۵۵۲۵۳۲ توییت 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، مجموعه‌داده را ارزیابی و نتایج را منتشر نمودیم.

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

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مقدمه

۱ـ۱ 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[۱] [۹]، سیستم‌های توصیه[۲] [۲۱]، تبلیغات [۱۱]، تشخیص قطبیت کلمات[۳] [۱۸] و تحلیل شبکه‌های اجتماعی [۱] مفید باشد. مدل روانشناختی مایرز-بریگز[۴] 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[۵] (E) ـ درون‌گرا[۶] (I)
- حسی[۷] (S) ـ شمی[۸] (N)

---

[۱] Job Screening
[۲] Recommendation System
[۳] Word Polarity Detection
[۴] MBTI: Myers-Briggs Type Indicators
[۵] Extroversion
[۶] Introversion
[۷] Sensing
[۸] Intuition

- منطقی[۹] (T) ـ احساسی[۱۰] (F)
- قضاوتی[۱۱] (J) ـ ادراکی[۱۲] (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[۱۳] با نوشته‌های افراد که با ویژگی‌های شخصیتی‌شان برچسب زده شده‌اند، مورد نیاز است. تاکنون چندین مجموعه‌داده برای کار پیش‌بینی ویژگی‌های شخصیتی به زبان انگلیسی و برخی از زبان‌های دیگر جمع‌آوری شده است، اما هیچ مجموعه‌داده‌ای در این زمینه برای زبان فارسی تهیه نشده است. با جستجوی عبارات تعیین‌کننده هویت و تهیه یک پرسشنامه، ما یک مجموعه‌داده با بیش از ۱.۵ میلیون توییت به زبان فارسی با برچسب نوع شخصیت MBTI آن‌ها جمع‌آوری کردیم. سپس از مدل برت[۱۴] برای ارزیابی اولین مجموعه‌داده فارسی استفاده نمودیم.

۱ـ۲ انگیزه‌های پژوهش

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

[۹] Thinking
[۱۰] Feeling
[۱۱] Judging
[۱۲] Perceiving
[۱۳] Dataset
[۱۴] BERT: Bidirectional Encoder Representations from Transformers

۱ـ۳ دستاوردهای پایان‌نامه

ما در این پژوهش توانستیم مجموعه‌داده‌ای با **۱۵۵۲۵۳۲** توییت که از **۹۳۸** کاربر توییتر بدست آمده است، جمع‌آوری کنیم. لازم به ذکر است که توانسته‌ایم در مقایسه با برخی از مجموعه‌داده‌های 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 پارس‌برت (که گونه‌ای از مدل برت است) بر روی مجموعه‌داده انجام شد، توانستیم به درصد دقت **۵۲.۸٪** برسیم.

۱ـ۴ ساختار پایان‌نامه

ساختار این پایان‌نامه دارای شش فصل است که در ادامه ساختار هر فصل گفته شده است. فصل دوم به معرفي تعاریف و مفاهیم پایه در ادبیات موضوع پرداخته شده است. در فصل سوم به کارهای انجام شده در حوزه جمع‌آوری مجموعه‌داده برای تشخیص ویژگی‌هایی شخصیتی پرداخته شده است. در فصل چهارم روش‌های پیشنهادی برای جمع‌آوری مجموعه‌داده و مقایسه‌ی نتایج بدست آمده از آن‌ها به طور کامل توضیح داده شده است. در فصل پنجم تحلیل و ارزیابی مجموعه‌داده 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 مجموعه‌داده، شبکه‌های عصبی عمیق[۱]، جانمایی کلمات[۲]، ترنسفورمرها[۳] و مدل برت که به درک بهتر مطالب در ادامه‌ی پژوهش کمک می‌کند، بپردازیم.

۲ ـ ۱ مجموعه‌داده

مجموعه‌داده، 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 مجموعه‌داده‌ی مورد نظر مرتبط می‌باشد. در شکل ۲ ـ ۱ بخشی از یک مجموعه‌داده نشان داده شده است که دارای سه ستون آیدی، توییت و برچسب می‌باشد. آیدی، یک شماره‌ی منحصر به فردیست که به هر داده از مجموعه‌داده تعلق می‌گیرد. توییت در این مثال از مجموعه‌داده همان داده‌ی موردنظر مجموعه‌داده است و برچسب نیز خروجی مورد انتظار شبکه عصبی مورد نظر است که در این مثال برچسب ها T/F یکی از شاخصه‌های مدل MBTI می‌باشد. لازم به ذکر است چندین مشخصه است که ویژگی‌ها و ساختارهای یک

---

[۱] Deep Neural Network
[۲] Word Embeddings
[۳] Transformers

مجموعه‌داده 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 کاراکتری باشند. درکل مقادیر موجود در مجموعه‌داده میتوانند از هر نوع تعریف شوند و هر معیار اندازه‌گیری داشته باشند.

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

شکل ۲ ـ ۱ : بخشی از یک نمونه مجموعه‌داده

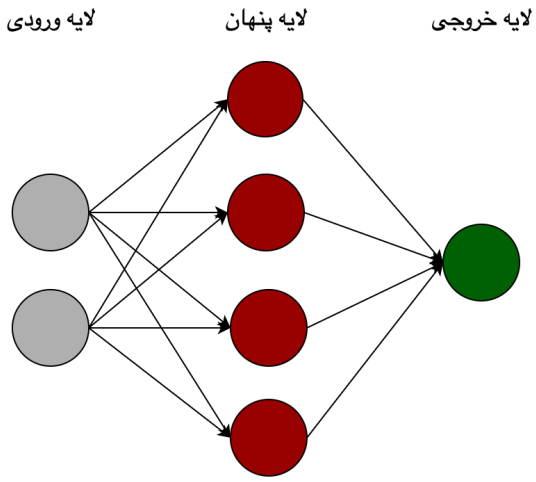
۲ ـ ۲ شبکه‌های عصبی عمیق

برای درک بهتر شبکه‌های عصبی عمیق لازم است ابتدا به شبکه‌های عصبی بپردازیم. 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 ـ 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[۴] و تابع فعال‌سازی[۵] 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،

[۴] Threshold
[۵] Activation Function

تحمل‌پذیری 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[۶] شبکه‌ی عصبی بیشتر شود، مدل پیچیده‌تر می‌شود که به این شبکه‌ها، شبکه‌های عصبی عمیق و به یادگیری آن‌ها یادگیری عمیق[۷] 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 ۲ـ۲: ساختار یک نمونه شبکه عصبی

۲ـ۳ جانمایی کلمات

برای استفاده از کلمات به عنوان ورودی جهت پردازش متن، نیاز است کلمات به شکل عددی درآیند. در پردازش زبان طبیعی[۸] جانمایی کلمات اصطلاحی است که برای نمایش کلمات برای تجزیه و تحلیل متن استفاده می‌شود، به طور معمول در قالب یک بردار[۹] 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

[۶] Hidden Layer
[۷] Deep Learning
[۸] Natural Language Processing
[۹] Vector

را می‌توان با استفاده از مجموعه‌ای از مدل‌سازی زبان[۱۰] و روش‌های یادگیری ویژگی[۱۱] به دست آورد که در آن کلمات یا عبارات به بردارهای اعداد واقعی نگاشت می‌شوند. از نظر مفهومی شامل یک جانمایی ریاضی از یک فضایی با ابعاد زیاد برای هر کلمه به یک فضایی پیوسته برداری با بعد[۱۲] 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.

۲ـ۴ 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[۱۳] یا ترجمه ماشینی عصبی[۱۴] 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[۱۵]، تبدیل متن به گفتار[۱۶] و غیره می‌باشد.

شکل ۲ـ۳: 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[۱۷] موازی‌سازی بیشتری را فراهم می‌کنند. بنابراین

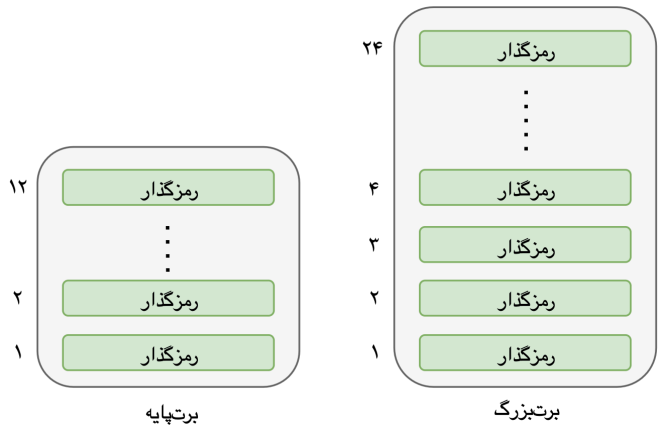
[۱۰] Language Modeling
[۱۱] Feature Learning Techniques
[۱۲] Dimension
[۱۳] Sequence Transduction
[۱۴] Neural Machine Translation
[۱۵] Speech Recognition
[۱۶] Text-to-Speech Transformations
[۱۷] Recurrent Neural Network

زمان آموزش را کاهش می‌دهند و در نتیجه آموزش مجموعه‌داده‌های بزرگتر بیشتر از قبل امکان‌پذیر شده است. این امر منجر به توسعه سیستم‌های آموزش دیده‌ای مانند برت و GPT[۱۸] شده است که با مجموعه‌داده‌های بسیار بزرگ زبان عمومی مانند ویکی‌پدیا و کامن کرال[۱۹] آموزش دیده‌اند و می‌توانند برای کارهای زبانی خاص تنظیم شوند.

۲ـ۵ مدل برت

مدل برت یک روش یادگیری ماشین[۲۰] مبنتی 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 ۲ـ۴ مشاهده می‌کنید، برت در زبان اصلی انگلیسی دارای دو مدل برت‌پایه برت‌بزرگ است. هر دو سایز مدل برت دارای تعداد زیادی لایه‌ي رمزگذاری است که که در مقاله آن‌ها را بلوک ترنسفورمر[۲۱] می‌نامد. نسخه پایه‌ی مدل دارای ۱۲ لایه ترنسفورمر و نسخه بزرگ آن دارای ۲۴ لایه ترنسفورمر می‌باشد.



شکل ۲ـ۴: دو نمونه از مدل برت

[۱۸] Generative Pre-trained Transformer
[۱۹] Common Crawl
[۲۰] Machine Learning
[۲۱] Transformer Blocks

این‌ها همچنین دارای شبکه های پیشخوان[۲۲] بزرگتر (به ترتیب ۷۶۸ و ۱۰۲۴ واحد پنهان) و سرهای توجه[۲۳] 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).

[۲۲] Feedforward-networks
[۲۳] Attention Heads

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.

۳ـ۱ بررسی کارهای انجام شده در این حوزه

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

۳ـ۱ـ۱ توییتر

آقایان باربارا پلنک و دیرک هووی پژوهشی در این حوزه در بستر توییتر انجام دادند که توانستند در نهایت بیش از ۱.۲ میلیون توییت انگلیسی از ۱۵۰۰ کاربر توییتر با ویژگی شخصیتی و جنسیتشان را جمع‌آوری کنند [۱۶]. آن‌ها برای جمع‌آوری این مجموعه‌داده ابتدا در توییتر، 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.

۳ـ۱ـ۲ ردیت

آقایان ماتژ جورکوویچ و یان اشنایدر در این حوزه در بستر ردیت[1] مجموعه‌داده‌ای 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 معمولا 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 مجموعه‌داده‌ای با دقت بالاست البته تلاش‌هایی برای تشخیص دقیق فلیرزهای مبهم نیز کرده‌اند اما آن‌هایی شک داشتند ویا 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 درنتیجه 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.

۳ـ۱ـ۳ پاندورا

این مجموعه‌داده از اولین مجموعه‌داده‌هایی است در مقیاس بزرگ که در این زمینه توسط آقای ماتژ 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، پنج‌عامله [۱۹] و Enneagram به همراه اطلاعاتی مانند سن و جنسیت و موقعیت مکانی، برچسب زده شده است. پاندورا برای قسمت نوع شخصیتی MBTIاش از مجموعه‌داده‌ی 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

[1] 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 چالش‌برانگیز است. بنابراین آن‌ها دیگر برای پیدا کردن ویژگی شخصیتی افراد در این مدل، فلیرزها را بررسی نکردند بلکه در نظرهایی[۲] 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 ویا عدد خام ممکن است داده شده باشد و همچنین مبتنی بر جه 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.

۳ـ۱ـ۴ 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 میشال 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 مسؤلیت‌هایی نظیر حفظ مجموعه‌داده، بررسی پروژه‌ها، پاسخگویی به سوالات و مطابقت داشتن با مقررات مختلف برای این دو نفر، ذکر کردند.

[۲] Comments

۳ـ۲ 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 مجموعه‌داده‌ای با برچسب 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 توییتر، 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 توییتر دشوارتر می‌نمود. راه‌حلی که برای این چالش درنظر 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 ننمودیم 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 مجموعه‌داده

در این فصل نحوه جمع‌آوری مجموعه‌داده را بررسی می‌کنیم. مجموعه‌داده از دو قسمت اصلی یعنی مجموعه توییت‌های 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 محموعه‌داده 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 توییت‌هایشان در جهت کمک به این پژوهش بود.

۴ـ۱ جستجو کلیدواژه

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

[۱] 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[۲] هستند را به عنوان گره‌های[۳] اصلی جمع‌آوری کردیم و با بررسی اینکه کدامیک از دنبال‌کنندگان یا دنبال‌شوندگان[۴] 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 جمع آوری شده از طریق ویژگی جستجوی پیشرفته، که توسط API توییتر ارائه شده، استفاده می‌کنیم.

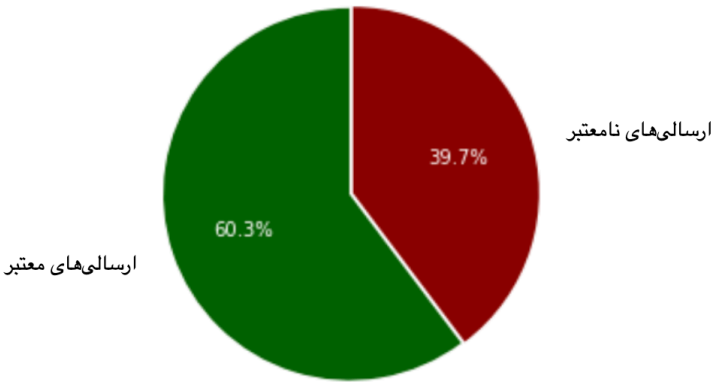
۴ـ۲ پرسشنامه

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 مجموعه‌داده قرار نمی‌گیرد، رضایت آن‌ها را جلب کردیم و از آن‌ها درخواست کردیم تا با تکمیل کردن پرسشنامه، حساب کاربری عمومی توییتر خود، جنسیت و ویژگی شخصیتی MBTI خود را در اختیارمان قرار دهند. از آنجا که بسیاری از افراد با آزمون MBTI[۵] 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 چالش‌برانگیزترین قسمت این پژوهش، همانطور که در شکل ۴ـ۱ مشاهده می‌کنید، درصد قابل توجه ارسالی‌های نامعتبر در روزهای اول انتشار پرسشنامه بود که دلیل آن ورودی نادرست در بخش حساب کاربری توییتر، با وجود جملات آگاه‌کننده، توضیحات کامل و تایید اعتبار خودکار ساده بود.

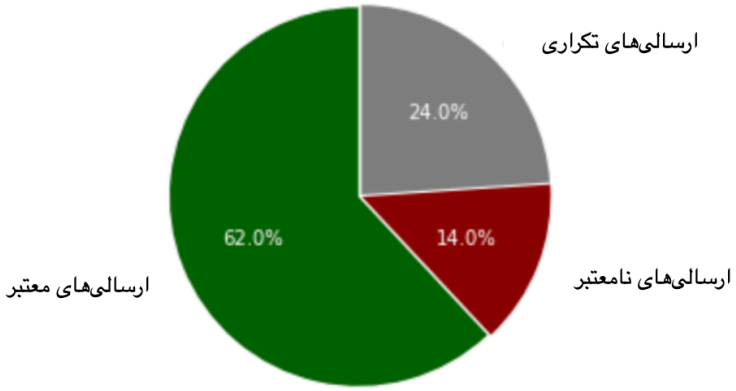
برای غلبه بر این چالش، عدم مطالعه‌ی دستورالعمل‌ها به طور کامل، زمان زیادی صرف طراحی یک

[۲] Follower
[۳] Nodes
[۴] Followings
[۵] برای مشاهده آزمون MBTI می‌توانید به سایت https://www.16personalities.com/fa مراجعه کنید.



شکل ۴ـ۱: نسبت اعتبار ارسالی‌های پرسشنامه اولیه

پرسشنامه جدید با منطق انشعاب با استفاده از فرم‌های مایکروسافت[۶] نمودیم و همچنین محیط انتشار پرسشنامه را به توییتر تغییر دادیم و همانطور که در شکل ۴ـ۲ مشاهده می‌کنید، موفق شدیم شمار ارسالی‌های نامعتبر را تا حد بسیار زیادی کاهش دهیم.



شکل ۴ـ۲: نسبت اعتبار ارسالی‌های پرسشنامه نهایی

۴ـ۳ مقایسه روش‌های پیشنهادی

روش نخست به ما کمک کرد تا بتوانیم ۱۴۴۳۶۵۸ توییت که ۹۲٪ از مجموعه‌داده را تشکیل می‌دهد، را جمع‌آوری کنیم. همانطور که در جدول ۴ـ۱ 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

[۶] 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 گاهاً اجباری 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 ۴ـ۱: 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

۴ـ۴ جمع‌آوری توییت‌ها

پیش از آغاز روند جمع‌آوری توییت‌ها باید ابتدا کاربران نامعتبر را فیلتر کنیم. در این بخش نامعتبر به معنی عمومی نبودن حساب کاریری ویا 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 بر خلاف 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.

۴ـ۵ 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 کلمه منحصر به فرد "<LINK>" و همچنین تمام هشتگ‌ها را با "<HASHTAG>" جایگذاری کردیم.

لازم به ذکر است که از ویژگی شناسایی زبان API توییتر، 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 مجموعه‌داده حفظ کنیم تا هر شخصی بر اساس استفاده و پژوهش خود از این مجموعه‌داده استفاده کند.

فصل ۵

تحلیل و ارزیابی مجموعه‌داده

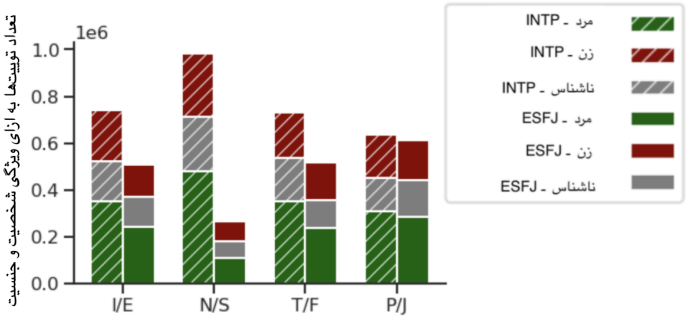
در این فصل می‌خواهیم به تجزیه و تحلیل مجموعه‌داده‌ی بدست آمده و همچنین ارزیابی آن بپردازیم. برای ارزیابی مجموعه‌داده و به نمایش گذاشتن مدلی که روی مجموعه‌داده‌ی ما آموزش دیده است، ما ابتدا داده‌ها را مجدداً قالب‌بندی کردیم سپس با استفاده از روش اعتبارسنجی متقابل طبقه‌ای تکرارشونده K بار[1] یک مدل مبتنی بر برت را آموزش دادیم و اعتبارسنجی کردیم.

۵ـ۱ تحلیل مجموعه‌داده

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 ۵ـ۱ عدم تعادل در داده‌های مربوط به ویژگی‌های I/E و N/S را نشان می‌دهد.

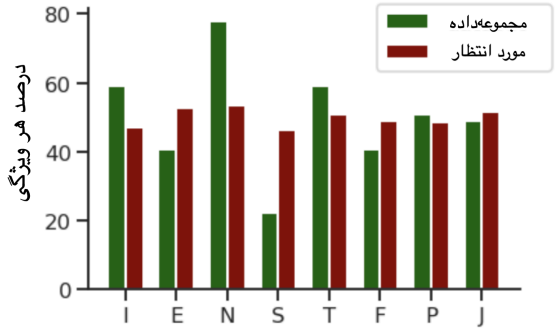
در ایران با بیش از **۸۲** میلیون نفر جمعیت، **۶۲۵۱۹** نفر آزمون MBTI را در وب‌سایت رسمی‌اش شرکت کرده‌اند. درصد افرادی که تشکیل‌دهنده هر یک از این ویژگی‌های شخصیتی هستند را بدست آوردیم و مقایسه‌ای با درصدهای بدست آمده از مجموعه داده‌ی خود انجام دادیم. همانطور که در شکل ۵ـ۲ مشاهده

[1] Repeated Stratified K-fold Cross-Validation



شکل ۵ـ۱: توزیع ویژگی‌های شخصیتی بر اساس جنسیت در مجموعه‌داده

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.



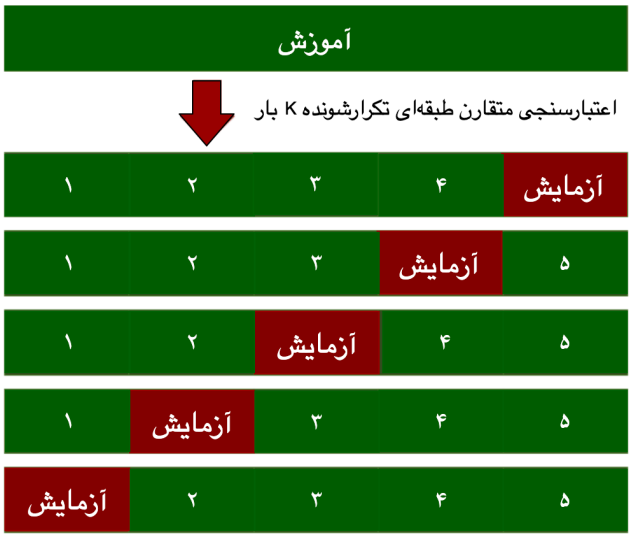
شکل ۵ـ۲: مقایسه توزیع MBTI در مجموعه داده‌ی ما با توزیع MBTI در جمعیت ایران

۵ـ۲ آماده‌سازی 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[۲] 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[۳]، همه‌ی توییت‌ها در یک مجموعه قرار داشته باشند.

[۲] Tokens
[۳] Validation

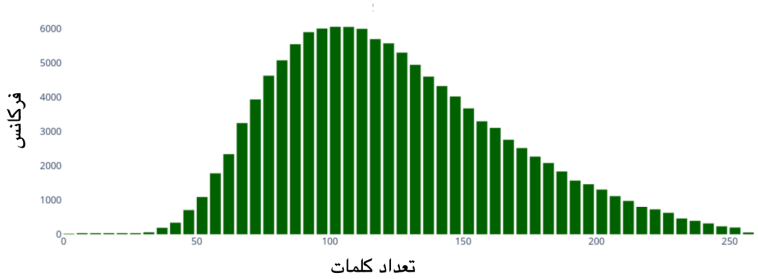
از آنجا که ما قصد تنظیم بیش از حد پارامترها را نداریم، از K بار[۴] تودرتو[۵] استفاده نکردیم از K بار با K برابر با ۵ استفاده کردیم و همانطور که در شکل ۵ـ۳ نشان داده شده است، روند تقسیم داده‌ها را پنج بار با تقسیم‌بندی تصادفی بارزیابی مدل جلوگیری کنیم. همانطور که پیشتر اشاره شد، برای کاهش تأثیر عدم تعادل 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 ۵ـ۳: 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 ۵ـ۴ نشان داده شده است. باتوجه 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.

[۴] K-fold
[۵] Nested



شکل ۵ـ۴: توزیع تعداد کلمات نمونه‌ای از داده‌های آموزش (تعداد بین برابر ۶۴)

۵ـ۳ 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[۶] لاجستیک رگرسیون[۷] 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[۸]. 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 مجموعه‌داده، ما قبل از هر اجرا با استفاده از نمونه‌برداری فرعی[۹] به صورت تصادفی حجم ورودی مدل را به قدری کاهش دادیم تا مدت اجرای هر آموزش به ۱۵ الی ۲۰ دقیقه کاهش یابد.

۵ـ۴ نتایج بدست آمده

ما عمدتاً از معیار متوسط کلان امتیاز F1[۱۰] برای ارزیابی عملکرد مدل بر روی مجموعه‌داده 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[۱۱] 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.

[۶] Classifier
[۷] Logistic Regression
[۸] Fine-tune
[۹] Subsampling
[۱۰] F1-score macro average
[۱۱] Balance

جدول ۵ـ۱: 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	۵۶.۶۹	۵۷.۷۵	۵۶.۹۳	۵۸.۱۵
۲	۵۶.۲۷	۵۸.۱	۵۷.۵۱	۵۶.۳۱
۳	۵۵.۳۲	۵۵.۹۲	۵۵.۹۳	۵۵.۱۴
۴	۵۷.۲۱	۵۷.۷۸	۵۷.۲۴	۵۷.۹۳
۵	۵۸.۴۱	۵۶.۹۷	۵۵.۱۲	۵۸.۴۸
میانگین	۵۶.۷۶	۵۷.۳	۵۶.۵۵	۵۷.۲

فصل ۶

نتیجه‌گیری و کارهای آینده

۶ـ۱ 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 ویا 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.

۶ـ۲ 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. علاوه‌براین 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 مجموعه‌داده 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] Gjurković, 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

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

واژه‌نامه فارسی به انگلیسی

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

واژه‌نامه انگلیسی به فارسی

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

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

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