دانشکده مهندسی کامپیوتر

کشف حقیقت متن

پروژه کارشناسی مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیک

ثمین فاتحی راویز

استاد راهنما

سید صالح اعتمادی

مهر ۱۳۹۸


بسم الله الرحمن الرحيم

تأییدیه‌ی صحت و اصالت نتایج

باسمه تعالی

اینجانب ثمین فاتحی راویز به شماره دانشجویی ۹۴۵۲۱۱۲۶ دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیه‌ی نتایج این پروژه 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: ثمین فاتحی راویز

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 دکتر سید صالح اعتمادی، صمیمانه تشکر و قدردانی کنم که قطعاً بدون راهنمایی‌های ارزنده ایشان، این مجموعه به انجام نمی‌رسید.

در پایان، بوسه می‌زنم بر دستان خداوندگاران مهر و مهربانی، پدر و مادر عزیزم و بعد از خدا، وجود الهام‌بخششان را ستایش می‌کنم و تشکر می‌کنم از آن‌ها و خواهر عزیزم، به خاطر پشتیبانی از من و گرما و امید ساطع از وجودشان.

همچنین تشکر می‌کنم از آقای امیرمحمد کاظمینی‌زاده که به‌عنوان همکار در کنار این پژوهش حضور داشته‌است و از ایشان بسیار آموختم.

ثمین فاتحی راویز

مهر ۱۳۹۸

چکیده

یافتن ویژگی‌های شخصیتی افراد می‌تواند در زمینه‌های مختلفی کاربرد داشته‌باشد. رفتارهای افراد، حالات چهره و نوشته‌های افراد می‌تواند نمودی از ویژگی‌های شخصیتی هر فرد باشد. در این مطالعه تلاش شده‌است با استفاده از نوشته‌های افراد، ویژگی‌های شخصیتی فرد تشخیص داده‌شود.

در این مطالعه، پس از آزمایش چندین مدل از جمله شبکه‌های عصبی عمیق مانند CNN و RNN جهت استخراج ویژگی‌های متن و استفاده از لایه‌ی MLP جهت رده‌بندی متون افراد، سرانجام به مدلی رسیدیم که موفق شد نتایج بهترین کارهای قبلی را نیز بهبود دهد. در این مدل ابتدا هر متن با زیرمتن‌های کوچکتر تقسیم نموده و با استفاده از شبکه‌ی BERT هر کلمه را توسط یک بردار بازنمایی کردیم. در نهایت با استفاده از Bagged SVM به رده‌بندی متون پرداختیم. بر این اساس با دقت ۰۳/۵۹ به‌عنوان میانگین دقت‌های ۵ ویژگی شخصیتی، این روش موفق شد بهترین نتیجه‌ی کارهای پیشین را بهبود دهد. همچنین این روش توانست با سرعت حداقل ۵/۵ برابر بهترین روش قبلی، عملکرد زمانی و محاسباتی نسبتاً بهینه‌ای از خود نشان‌دهد.

واژگان کلیدی: ویژگی‌های شخصیتی، شبکه‌های عصبی، رده‌بندی متن، شبکه‌های عصبی عمیق

فصل ۱

مقدمه

شخصیت افراد به صورت مجموعه‌ی رفتارها، شناخت‌ها و الگوهای احساسی ناشی از عوامل بیولوژیکی یا محیطی تعریف می‌شود[۲]. با توجه به این مسئله که شخصیت افراد در رفتارها، حرکات، چهره و نوشته‌های آن‌ها نمود پیدا می‌کند، می‌توان با بررسی این نمودها شخصیت افراد را کشف کرد. پیش‌بینی شخصیت افراد می‌توان در زمینه‌های مختلفی مورد استفاده قرار بگیرند. از جمله‌ای زمینه‌ها می‌توان به به موارد زیر اشاره کرد[۹]

دستیارهای صوتی: دستیاران صوتی خودکار امروزی مانند سیری ، دستیار گوگل ، الکسا و غیره می توانند با شناسایی شخصیت کاربر پاسخ های سفارشی ارائه کنند. همچنین دستیاران صوتی را می توان به گونه ای برنامه ریزی کرد که بر اساس شخصیت کاربر وبرای رضایت بیشتر وی، شخصیت‌های مختلفی از خود نشان دهند.

سیستم‌های توصیه‌گر: افرادی که دارای یک نوع شخصیت خاص هستند ممکن است علاقه و سرگرمی مشابهی داشته باشند. محصولات و خدماتی که به شخص توصیه می شود باید مواردی باشد که توسط کاربران دیگر با نوع شخصیتی مشابه ارزیابی مثبت شده است. به عنوان مثال، در پژوهشی سیستمی برای توصیه بازی‌ها به بازیکنان مبتنی بر شخصیت ایجاد شده‌است که از مکالمات متنی آن‌ها با سایر بازیکنان الگوبرداری می شود[۱۴].

مراقبت های ویژه و مشاوره درمانی: برای ارائه‌ی خدمات درمانی و مشاوره‌ای مناسب می‌توان این خدمات را با توجه به شخصیت فرد ارائه کرد و روانپزشک می تواند از این اطلاعات برای ارائه مشاوره بهتر استفاده

کند.

استخدام افراد: در مدیریت منابع انسانی، ویژگی‌های شخصیتی افراد می‌تواند در مناسب بودن برخی مشاغل برای آن‌ها موثر باشد. مسئله غربالگری کاندیدای شغل از دیدگاه میان رشته ای روانشناسان و دانشمندان یادگیری ماشین را مورد بحث قرار می دهد[۶].

در این مطالعه قصد داریم به بررسی شخصیت افراد با توجه به نوشته‌هایشان بپردازیم. big-five personality یکی از معروف‌ترین تقسیم بندی‌های برای ویژگی‌های شخصیتی است که شخصیت افراد را در پنج بعد بررسی می‌کند[۸]

- Openness to experience: کنجکاوی و ابتکار در مقابل محافظه‌کاری.
- Conscientiousness: کارآمدی و سازمان‌یافتگی در مقابل رفع‌تکلیف به صورت سریع و بی‌دقت.
- Extraversion: برون‌گرایی در مقابل درون‌گرایی.
- Agreeableness: رفتار دوستانه و دل‌سوزانه در مقابل رفتار غیردوستانه و پرچالش.
- Neuroticism: حساس و عصبی در مقابل آرام و با اعتماد به نفس.

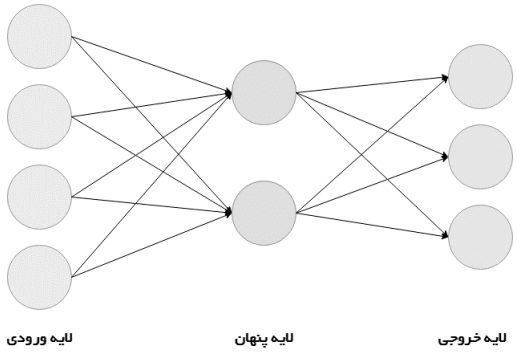
۱-۱ مفاهیم مرتبط

در سال‌های اخیر وجود شبکه‌های عصبی و روش‌های تعبیه‌ی کلمات[۱] توانسته‌است موفق چشم‌گیری در زمینه‌ی پردازش زبان طبیعی داشته باشد. در ادامه به معرفی مختصر هر یک از این مفاهیم می‌پردازیم.

۱-۱-۱ شبکه‌های عمیق

برای درک صحیح از شبکه عصبی عمیق ابتدا باید شبکه عصبی را معرفی نماییم. الگوریتم شبکه عصبی، فرایندی الهام گرفته‌شده از شبکه عصبی زیستی است که برای پردازش اطلاعات استفاده می‌شود. این الگوریتم برای پردازش داده‌ها از یک فضای شبکه‌ای شامل تعداد بسیار زیادی واحد کوچک به نام نورون استفاده می‌کند. نورون‌ها در شبکه مذکور به‌هم‌پیوسته‌اند و به‌صورت موازی برای حل یک مسئله رفتار می‌کنند. در شبکه عصبی

[۱] Word Embedding



شکل ۱-۱: یک شبکه‌ی ساده‌ی عصبی مصنوعی

سه نوع لایه‌ی ورودی، پنهان و خروجی داریم. با توجه به مسئله، تعداد ویژگی‌های ورودی و انواع خروجی در هر لایه تعدادی نورون قرار می‌گیرد. ورودی‌ها در شبکه به جریان افتاده و با تولید وزن برای یال‌های شبکه، یک‌سری خروجی تولید می‌کنند. سپس خروجی‌های تولیدشده با نورون‌های معتبر در لایه خروجی مقایسه می‌گردد. با مقایسه خروجی بدست آمده و خروجی معتبر، مقدار خطا تولیدشده بدست می‌آید. با توجه به خطا، وزن یال‌های شبکه به‌روزرسانی شده و دوباره فرایند مقایسه صورت می‌گیرد. این عملیات تکراری تا زمانی که به نتیجه مناسب برسد ادامه پیدا خواهد کرد.

شکل ۱-۱ یک شبکه عصبی را نمایش می‌دهد. در هر لایه بسته به نوع مسئله و خروجی چندین نورون قرار گرفته است. شبکه عصبی دارای مزایای زیادی است که از مهم‌ترین آن‌ها می‌توان به تحمل‌پذیری بالا در مقابل داده‌های نویز، مناسب بودن برای ورودی‌ها و خروجی‌های پیوسته، عملکرد بهنگام و دقت بالا در مسائل واقعی اشاره نمود. در مقابل فواید گفته‌شده، می‌توان به معایبی همچون زمان آموزش زیاد، نیاز به تعیین مقدار پارامترهای تجربی، احتمال قرار گرفتن در ماکزیمم محلی و قابلیت تفسیرپذیری پایین از روی مدل ساخته‌شده اشاره نمود.

شبکه عصبی عمیق نیز از نظر ساختار شبیه به الگوریتم شبکه عصبی است. در شبکه عصبی عمیق برخلاف شبکه عصبی از دو یا چند لایه پنهان تشکیل شده است. فرایند یادگیری نیز در این روش به مانند شبکه عصبی بوده، فقط با این تفاوت که جریان ورودی‌ها ابتدا به چند لایه پنهان آمده و سپس با خروجی مقایسه می‌شود. از انواع شبکه عصبی عمیق می‌توان به شبکه عصبی کانوولوشن [۲] و بازگشتی [۳] اشاره نمود. یکی از کاربردهای شبکه‌های عصبی، یافتن بازنمایی برداری برای کلمات متن می‌باشد. به این روش، تعبیه‌ی کلمات می‌گویند.

[۲] Convolutional Neural Network
[۳] Recurrent Neural Network

۲-۱-۱ Word Embedding

برای استفاده از کلمات به‌عنوان ورودی جهت پردازش متن، نیاز است کلمات به شکل عددی درآیند. برای تبدیل کلمات به صورت عددی روش‌های مختلفی مانند استفاده از بردار one hot و یا استفاده از روش TF-IDF ارائه شده‌است. افزون بر این روش‌ها، استفاده از روش‌های Word Embedding مختلف در سال‌های اخیر بسیار رونق گرفته است. Word2Vec و GloVe دو روشی هستند که سابق بر این بسیار مورد استفاده بوده‌اند و BERT نیز در سال ۲۰۱۸ معرفی‌شده‌است[۳].

پیش از این در این زمینه پژوهش‌های بسیاری صورت گرفت بوده است. این پژوهش‌ها ویژگی‌های متن افراد را به طرق مختلف استخراج، و با توجه به این ویژگی‌ها به رده‌بندی [۴] نوع شخصیت افراد می‌پرداختند. بعضی از این پژوهش‌ها با استفاده از روش‌هایی مانند ویژگی‌های LIWC [۵] یا SPLICE [۶] ویژگی‌های متون افراد را استخراج کردند و عده‌ای دیگر با استفاده از روش‌های مختلف Word Embedding مانند Word2Vec و GloVe به هر کلمه از متن یک بردار تخصیص می‌دهند. سپس با توجه به بردار کلمات متن ویژگی‌های آن را استخراج می‌کنند. اما هیچ‌یک از این روش‌ها محتوای کلمه را در قسمت خاصی از بافت متن در نظر نمی‌گیرند. با توجه به این ضعف در کارهای انجام‌شده‌ی پیشین، در این پژوهش تلاش شده با استفاده از مدل پیش‌آموزش‌یافته‌ی[۷] BERT با توجه به متن، یک بردار به هر کلمه اختصاص داده شود. در نهایت پس از رده‌بندی تیپ شخصیتی افراد با استفاده از این بردارها، توسط شبکه‌های مختلف CNN، RNN و غیره نتایج نشان داد این کار با استفاده از مدل‌های کلاسیک بهتر صورت می‌پذیرد. همچنین مدل نهایی که از رده‌بند SVM استفاده می‌کند علاوه بر دقت بهتر نسبت به کارهای پیشین، قادر است سرعت انجام این کار را چندین برابر افزایش دهد.

۲-۱ کار های مربوطه

در طی سال‌ها افراد تلاش کرده‌اند مسئله‌ی پیش‌بینی شخصیت با توجه به متون افراد را به طرق مختلف با استفاده از مجموعه دادههای متنوع حل کنند. به عنوان مثال، Lambiotte و همکارش Kosinski در پژوهشی

[۴] classification
[۵] Linguistic Inquiry and Word Count
[۶] Structured Programming for Linguistic Cue Extraction
[۷] Pretrained Model

بر روی مجموعه داده‌ی شناخته‌شده‌ی MyPersonlity، که یک مجموعه داده‌ی تهیه شده از متون منتشر شده‌ی افراد در شبکه‌ی اجتماعی Facebook است، با استفاده از رگرسیون به پیش‌بینی شخصیت افراد پرداختند و سپس با ضریب همبستگی پیرسون دقت رگرسیون را محاسبه کردند [۵]. همچنین Tadesse و همکارانش نیز با استفاده از روش‌های LIWC و SPLICE به استخراج ویژگی‌های متن پرداختند و سپس با استفاده از چهار ردهبند Gradient Boosting، LR، SVM و XGB سعی به تشخیص شخصیت افراد در همین مجموعه داده داشتند[۱۲].

Howlader و همکارانش نیز با استفاده از داده‌های بیش از ۸۰۰ هزار کاربر فیسبوک از مجموعه داده‌ی MyPersonlity، ویژگی‌های متن افراد را با به کارگیری روش‌های LIWC و LDA به دست آورده و توسط چهار الگوریتم رگرسیون، شخصیت افراد را پیش‌بینی کردند که نتایج آن‌ها نشان داد که P-SVR و RBF-SVR کمترین خطا را دارند[۴]. هم‌چنین در به کارگیری روش‌های یادگیری عمیق برروی این مجموعه داده، Tandera و همکارانش با استفاده از استخراج فیچرهای LIWC و SPLICE به مقایسه‌ی روش‌های یادگیری ماشین کلاسیک مانند Naïve Bayes، SVM، LR، Gradient Boosting، LDA و مدل‌های عمیق مانند شبکه‌های RNN ، شبکه‌های CNN و MLP پرداخته است و برتری مدل های عمیق را به اثبات رسانده است[۱۳].

Pennebaker و همکارش King در سال ۱۹۹۸ در پژوهشی به جمع‌آوری مجموعه داده‌ای از انشاهای دانشجویی به‌همراه ویژگیهای شخصیتی تشخیص داده‌شده برای آن‌ها پرداختند[۱۰]. این مجموعه داده، بعدها در بسیاری از پژوهش‌ها مورد استفاده قرار گرفت. به عنوان مثال Poria و همکارانش برای استخراج ویژگی‌های این مجموعه‌داده از روش‌های LIWC و MRC استفاده کرده‌اند[۱۱]. آن‌ها در یک اقدام جدید برای تشخیص احساسات از ترکیب دو مجموعه ConcepNet و EmoSenticNet استفاده کرده‌اند و پس از درنظر گرفتن یک بردار برای هر مفهوم با استفاده از SVM به رده‌بندی انشاها پرداخته‌اند. Majumder و همکارانش نیز با استفاده از Google News Word2Vec به استخراج ویژگی‌های کلمات این مجموعه داده پرداخت. وی با استفاده از شبکه‌های عمیق و ترکیب ویژگی‌های متن با ویژگی‌های Mairesse به تشخیص شخصیت دانشجویان پرداخته است[۷]. Zuo و همکارانش نیز در این مجموعه داده ارتباطات میان تیپ‌های شخصیتی را با استفاده از الگوریتم ML-KNN وزن‌دار مدل می‌کنند. این الگوریتم درواقع یک K نزدیک‌ترین همسایه است که به‌صورت چند برچسبی آموزش می‌بیند[۱۵].

در روش‌های استخراج ویژگی از متن، علاوه بر Word2Vec از GloVe نیز استفاده شده است. به عناون مثال در پژوهشی Arnoux و همکارانش با استفاده از GloVe به استخراج ویژگی‌های متن و سپس به مقایسه

آن با ویژگی‌های LIWC پرداختند[۱]. با وجود کارها و پژوهش‌های بسیاری که در این حوزه انجام شده‌است، هیچ‌یک از آن‌ها برای استخراج ویژگی کلمات بافت[۸] متن را در نظر نمی‌گیرند. همچنین مدل‌های عمیق ارائه شده، به علت طولانی بودن متون هر فرد و ماهیت شبکه‌های عمیق، نیازمند زمان و منابع اجرایی بسیار قدرتمند هستند. Hernandez و Knight در پژوهشی با ترکیب نوشته‌های کوتاه توییتری افراد توانستند شخصیت افراد را پیش‌بینی کنند و دقت و سرعت پیش‌بینی را با این‌کار افزایش دهد.

در این پژوهش مدلی ارائه شده است که با استفاده از الگوریتم‌های عمیق و با شکستن متون بلند به زیرمتن‌های کوتاه‌تر و ترکیب پیش‌بینی شخصیت نویسنده برای هر زیرمتن، با در نظر گرفتن جایگاه کلمات در متن با دقت و سرعت بالایی به تشخیص شخصیت افراد می‌پردازد.

۳-۱ جمع‌بندی

با توجه به آنکه تشخیص شخصیت افراد در زمینه‌های مختلفی کاربردهای فراوان دارد این مسئله امروزه به مسئله‌ی مهمی در اجتماع تبدیل شده است. در زمینه‌ی تشخیص شخصیت افراد با استفاده از متن‌های نوشته‌شده توسط آن‌ها، اخیرا کارهای بسیاری انجام شده است. اما با توجه به آنکه بهترین کارهای انجام‌شده در این زمینه، بافت متن را برای تولید بردار هر کلمه در نظر نمی‌گرفتند و همچنین دارای سرعت بسیار کم و نیازمند منابع قوی محاسباتی بودند، در این پژوهش ما با ارائه‌ی روشی با سرعت چند برابر بیشتر و همچنین مبتنی بر بافت متن موفق شدیم دقت بهترین کارهای انجام شده‌ی پیشین در این حوزه را بهبود دهیم.

[۸] context

فصل ۲

روش حل مسئله

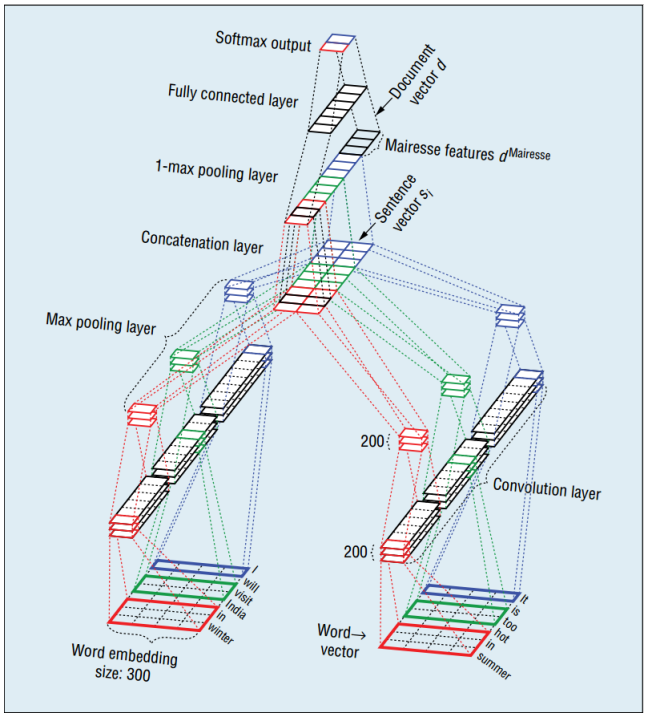
در این بخش به شرح روش انجام شده جهت پیش‌بینی شخصیت افراد با استفاده از نوشته‌هایشان می‌پردازیم.

۲-۱ پژوهش پایه

در این پژوهش آخرین کار معتبر انجام شده در زمینه‌ی تشخیص شخصیت را که دارای بهترین دقت ارایه شده در هر تیپ و به صورت میانگین داشت، پایه‌ی کار قرار دادیم.[۷] در مقاله ی مذکور Majumder و همکارانش تلاش کردند با استفاده از Word2Vec هر کلمه ی هر انشا را به صورت یک بردار بازنمایی کنند. سپس با استفاده از چند لایه‌ی شبکه‌ی عصبی به تشخیص ویژگی‌های شخصیتی افراد بپردازند.

تصویر ۲-۱ مراحل کار پوریا و همکارانش را نشان می دهد. طبق این مراحل، در بهترین الگوریتم ارائه‌شده در این مقاله، ابتدا با استفاده از مدل پیش‌ساخته‌ی Google News Word2Vec، هر کلمه‌ی هر متن به‌صورت یک بردار بازنمایی شد. برای تمامی N-gramهای ۱، ۲ و ۳ تایی با استفاده از یک لایه‌ی کانولوشن با ۲۰۰ فیلتر، ویژگی‌های هر جمله استخراج و با بیشینه‌گیری بر روی ویژگی‌های جملات یک متن، بردار ویژگی متن ساخته می‌شد. سپس در این الگوریتم سه بردار حاصل از ویژگی‌های ۱، ۲ و ۳ گرام‌ها پشت سر یک‌دیگر قرار گرفته و ویژگی‌های هر متن به صورت یک بردار ۶۰۰ تایی بازنمایی می‌شود. در مرحله‌ی بعد این بردار را به یک لایه‌ی تماماً متصل [۱] انتقال داده و سپس با استفاده از یک لایه‌ی softtmax به رده‌بندی ویژگی‌های

[۱] fully connected



شکل ۱-۲: مراحل روش تشخیص ویژگی‌های شخصیتی با استفاده از متن افراد در پژوهش پایه

شخصیتی افراد می‌پردازد.

در پژوهش حاضر ابتدا به پیاده‌سازی و اجرای دوباره‌ی روش اصلی مقاله‌ی قبلی پرداختیم. به منظور پیاده‌سازی این مدل نیز از ابزار keras استفاده کردیم. در نتیجه‌ی این پیاده‌سازی، توانستیم نتایج را تا حد خوبی باز تولید کنیم. اما آموزش این مدل بسیار کند و نیازمند منابع قوی محاسباتی بود. پس از پیاده‌سازی مدل پایه، تلاش کردیم با استفاده از آزمایش مدل های مختلف، دقت به دست آمده در کار قبلی را بهبود دهیم. در این راستا از چند جهت مدل را بررسی کردیم.

- وردامبدینگ

در قسمت وردامبدینگ برای آن‌که به هر کلمه یک بردار اختصاص دهیم می‌توان از روش‌های مختلفی استفاده کرد که ما ابتدا در ادامه‌ی کارهای پیشین با استفاده از شبکه‌ها و رده‌بندهای مختلف را آزمایش کردیم. تا به امروز کارهای پیشین تنها وردامبدینگ‌های و را مورد استفاده قرار داده بودند در حالی که هیچ‌یک از این دو روش بافت جمله در نظر نمی‌گیرند. به همین منظور در قسمت وردامبدینگ مدل BERT را جایگزین Word2Vec کردیم. شبکه‌ی BERT که از آن در این پژوهش استفاده شده است،

دارای ۱۲ لایه است که از خروجی هر یک یا ترکیبی از لایه‌ها می‌توان بازنمایی برداری از متن‌های دانشجویی به‌دست آورد.

- استخراج ویژگی

در این قسمت ابتدا سعی در استخراج ویژگی‌های جملات کردیم. پوریا و همکارانش با توسط یک لایه‌ی کانولوشن، ویژگی‌های جملات را با استفاده از بردارهای حاصل از Word2Vec برای کلمات، استخراج کرده بودند. در این پژوهش ما از این روش استخراج ویژگی جمله با بردارهای حاصل BERT استفاده کردیم. علاوه برآن براساس کارهای آینده‌ی همان مقاله، از RNN ها برای استخراج ویژگی‌های جملات استفاده کردیم. هر دوی این روش‌ها بسیار زمان‌گیر و نیازمند منابع قدرتمند محاسباتی بودند. بنابراین در ادامه از روش‌های سریع‌تر مانند میانگین‌گیری و یا ماکزیمم‌گیری استفاده کردیم. در حالتی دیگر نیز بدون به دست آوردن ویژگی‌های جملات، با استفاده از بردارهای توکن‌ها به صورت مستقیم ویژگی‌های متن را به دست آوردیم.

- رده‌بندی

در این قسمت، ویژگی های استخراج شده برای جملات را برای پیش‌بینی برچسب به رده‌بندهایی مانند لایه‌ی Fully Connected و لایه‌ی softmax دادیم.

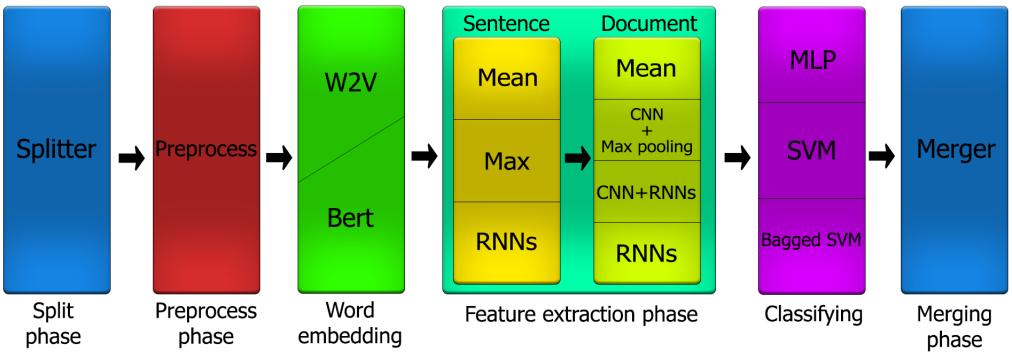
همچنین تلاش کردیم با استفاده از DocBERT، RegLSTM و RegLSTM مسئله را انجام دهیم. در ادامه نیز رده بندهای کلاسیک را جایگزین شبکه‌های عمیق کردیم.

نهایتاً به روشی رسیدیم که درآن با تکه‌تکه کردن متن انشای هر فرد و دادن هر تکه به SVM و سپس رأی گیری بین برچسب‌های پیش‌بینی شده برای تکه‌های هر متن، با دقت و سرعت خوبی تیپ شخصیتی افراد را با توجه به دست‌نوشته‌هایشان تشخیص دادیم.

برخی روش‌های استفاده شده برای هر قسمت کار در شکل ۱-۲ نشان داده شده‌است.

۲-۲ روش‌های آزمایش‌شده

در ادامه به بررسی جزئی‌تر آزمایش‌های انجام‌شده براساس رده‌بندهای مورد استفاده می‌پردازیم.



شکل ۲-۲: روش‌های استفاده‌شده برای هر مرحله از تشخیص ویژگی‌های شخصیتی با استفاده از متن افراد

۲-۲-۱ رده‌بندهای مبتنی شبکه‌های عمیق

- W2V + LSTM + LSTM : M1

ابتدا در ادامه‌ی پژوهش قبلی و انجام کارهای آینده‌ی آن پژوهش، از دولایه‌ی RNN پشت سر هم استفاده کردیم. در این روش ابتدا با استفاده از یک لایه‌ی LSTM ویژگی‌های هر جمله را استخراج نموده و با انتقال ویژگی‌های جملات پشت سر هم به یک لایه‌ی دیگر LSTM ویژگی‌های متن هر فرد را به‌دست می‌آوریم. این روش بسیار کند و نیازمند منابع است. با توجه به محدودیت منابع و مشاهده‌ی نتایج ضعیف اولیه، استفاده از این روش متوقف شد.

- BERT + Conv2D + Dense : M2

در ادامه در مدل آزمایش‌شده‌ی بعدی وردامبدینگ را از Word2Vec به تغییر دادیم. BERT یک مدل آموزش‌دیده از قبل است که در سال ۲۰۱۸ معرفی شده‌است. در این مدل ما بردارهای تولیدشده برای هر جمله توسط وردامبدینگ BERT را به یک لایه‌ی کانولوشن انتقال داده و ویژگی‌های هر جمله را استخراج کردیم. سپس با استفاده از یک لایه‌ی MaxPooling و ماکزیمم‌گیری بین ویژگی‌های جملات یک متن، بردار ویژگی‌های متن را به‌دست آوردیم. در مرحله‌ی آخر نیز متن‌های افراد را با استفاده از دو لایه‌ی Dense و softmax رده‌بندی کردیم.

- BERT + Mean + Dense : M3

در ادامه با استفاده از بردارهای برگرفته از شبکه‌ی BERT برای مرحله‌ی استخراج ویژگی به جای استفاده از شبکه‌های کانولوشن از روش میانگین‌گیری استفاده کردیم. به این صورت که ویژگی‌های هر جمله میانگین بردارهای کلمات آن جمله بود. سپس ویژگی‌های هر متن توسط میانگین‌گیری بین ویزگی‌های جملات به‌دست می‌آید. برای رده‌بندی متون نیز از دو لایه‌ی Dense و softmax پشت سر هم استفاده کردیم.

- BERT + RNN + Dense : M4

در این حالت برای استخراج ویژگی‌های متن به ترتیب بردارهای جملات متن را پشت سر هم به یک لایه‌ی RNN دادیم. در این مرحله از سه روش SimpleRNN، LSTM و GRU به صورت جداگانه استفاده کردیم. و برای رده‌بندی از لایه‌ی dense و softmax بهره بردیم.

- BERT + Conv2D + RNN + Dense : M5

در این حالت برای استخراج ویژگی‌های جملات مانند مدل M2 از لایه‌ی کانولوشن استفاده کردیم. سپس مانند مدل M4 برای استخراج ویژگی‌های متن با استفاده از ویژگی‌های جملات از لایه‌ی RNN بهره گرفتیم. در نهایت نیز با استفاده از لایه‌ی Fully Connected و softmax به رده‌بندی برای تشخیص تیپ شخصیتی افراد پرداختیم.

۲-۲-۲ رده‌بندهای کلاسیک

استفاده از مدل‌های عمیق علاوه بر نیاز به منابع بسیار قوی محاسباتی، بسیار زمان‌بر بود. با توجه به این محدودیت‌ها تصمیم به استفاده و آزمایش مدل‌های کلاسیک مانند SVM گرفتیم.

- W2V + Mean + SVM : M6

در این حالت بردارهای استخراج شده برای توکن‌ها حاصل از Word2Vec را با میانگین‌گیری تبدیل به ویژگی‌های متن افراد کردیم. سپس با استفاده از روش SVM به رده‌بندی متون پرداختیم.

- BERT + Mean + SVM : M7

در این حالت با استفاده از مدل BERT برای هر جمله یک بردار استخراج کردیم که میانگین بردارهای کلمات آن جمله بود. سپس با میانگین‌گیری بین بردار جملات به صورت معمولی و وزن‌دار ویژگی‌های متن را استخراج نمودیم. حال بردار ویژگی‌های تولیدشده برای هر متن را به‌عنوان ورودی کلاسیفایر SVM دادیم و تیپ‌های شخصیتی را از یک‌دیگر تشخیص دادیم.

- BERT + Bagging SVM : BB-SVM

در این روش که روش نهایی این پژوهش بود و منتج به بهترین نتیجه شد،ابتدا مانند آن‌چه در بخش ۲-۳ آمده‌است هر متن را به چندین زیرمتن با حداکثر ۲۵۰ توکن تبدیل کردیم. سپس با استفاده از مدل BERT بردار ویژگی هر زیرمتن را استخراج کردیم. تبدیل هر متن به زیرمتونش به این دلیل انجام می‌شود که حداکثر اندازه‌ی مجاز برای ورودی شبکه‌ی BERT پانصد و دوازده توکن است. با استفاده از بردار ویژگی استخراج‌شده برای هر زیرمتن و انتقال آن به کلاسیفایرهای موازی SVM به تشخیص تیپ شخصیتی افراد پرداختیم. مراحل مدل BB-SVM در بخش ۲-۳ با جزئیات بیشتر شرح‌داده‌شده است.

۲-۳ بهترین مدل: BB-SVM

در بهترین روش انجام شده در این پژوهش برای تشخیص ویژگی‌های شخصیتی افراد با استفاده ز متن‌هایشان، به ترکیب یادگیری عمیق و رده‌بندهای کلاسیک پرداختیم. به این منظور با استفاده از مدل BERT یک بازنمایی برداری از متن به دست آورده و رده‌بندی کردن متون را با استفاده از SVM انجام دادیم. مراحل دقیق‌تر این روش در شماره های ۱ تا ۶ بیان شده‌اند:

۱. تکه‌تکه کردن و تقسیم به زیرمتون [۲]

در اولین مرحله از این روش هر متن را به چند زیرمتن تقسیم کردیم. در این روش ابتدا کل مجموعه داده را به ۱۰ قسمت به صورت رندوم با توزیع یکنواخت تقسیم کردیم. (با توجه به آنکه ارزیابی با 10-Fold Cross Validation انجام می شود، هر بار یک قسمت از ده قسمت به عنوان داده‌ی test برداشته می‌شوند.) سپس تمامی جملات با بیشتر از ۲۰۰ کلمه را به چند قسمت شکستیم به طوری که طول هر جمله بیشتر از ۲۰۰ نباشد. و هر تکه حاصل را به عنوان یک جمله در متن در نظر گرفتیم.

[۲] Splitter

در مرحله‌ی بعد به تقسیم متن به تکه‌های کوچک‌تر پرداختیم. به این صورت که اگر متن D شامل مجموعه جملات $\{s_۱, ..., s_n\}$ باشد و l_i (به‌طوری‌که $i \in \{۱, ..., n\}$) نشان دهنده‌ی طول جملات s_i باشد، آن گاه کلیه‌ی متون از ابتدا به چند قسمت تقسیم می‌شوند به طوری که در زیرمتن S با جملات $\{s_k, ..., s_{k+m}\}$ که: $\sum_{j=k}^{k+m} l_j \leq ۲۰۰$ و $\sum_{j=k}^{k+m+} l_j > ۲۰۰$

بنابراین هر متن به چند زیرمتن تقسیم می‌شود به طوری که هیچ جمله‌ای شکسته نشود (مگر آن‌که طول آن جمله بیشتر از ۲۰۰ کلمه باشد). برچسب‌های هر متن (برای هر ۵ تیپ شخصیتی) برای تمامی زیرمتون آن تکرار شده‌اند. لازم به ذکر است تقسیم‌بندی داده‌ها به ۱۰ قسمت پیش از شکستن متون به زیرمتن‌ها باعث می‌شود تمامی زیرمتن‌های مربوط به یک متن تنها در یکی از ده قسمت قرار بگیرند. این بدان معناست که تمامی زیرمتون یک متن یا تنها برای train استفاده می‌شوند یا تنها برای validation و یا test. همچنین شکستن متن افراد به چند قسمت علاوه بر رعایت محدودیت‌های حداکثر تعداد کلمات، موجب افزایش داده‌های آموزشی ما نیز شد.

۲. پیش‌پردازش [۳]

در این پژوهش preprocessing مشابه مقاله‌ی پایه انجام شده‌است. در این حالت تعداد توکن‌های هر تکه متن پس از پردازش متن، حداکثر به ۲۵۰ رسید.

۳. وردامبدینگ [۴]

در ادامه‌ی این روش به ازای هر توکن در یک زیرمتن، خروجی لایه‌ی ۱۱ مدل BERT را به عنوان بردار آن توکن در نظر گرفتیم. بنابراین اندازه‌ی خروجی در این مرحله به صورت W'×768 خواهد بود. که در آن W' ماکزیمم تعداد کلمات همه‌ی زیرمتون است.

۴. استخراج ویژگی [۵]

در مرحله ی بعد بردار های تمامی tokenهای یک زیر داکیومنت را میانگین گرفتیم و با ویژگی های mairesse کنار هم قرار داده و به عنوان ویژگی‌های متن در نظر گرفتیم. بنابراین خروجی این مرحله

[۳] Preprocessor
[۴] Word Embedding
[۵] Feature Extraction

بردار ویژگی‌های متن به صورت R^{852} خواهد بود که در آن R نشان‌دهنده‌ی مجموعه‌ی اعداد حقیقی است.

۵. رده‌بندی [۶]

حال ویژگی‌های هر زیرمتن را برای پیش‌بینی برچسب به رده‌بند SVM دادیم. پژوهش‌های پیشین نشان داده‌است اعمال بگینگ روی رده‌بند SVM [۷] می‌تواند موجب افزایش دقت پیش‌بینی در کارهای مختلف شود.(رفرنس)به همین جهت ویژگی‌های مورد نظر را به ۱۰ کلاسیفایر دادیم تا به صورت موازی پیش‌بینی را انجام دهند.

۶. پساردهبندی [۸]

اگر فرض کنیم $\{d_1, ..., d_n\}$ تمامی زیرمتن‌های متن D باشند و مجموعه $\{p_1, ..., p_n\}$ مجموعه‌ی برچسب‌های پیش‌بینی‌شده برای تمامی زیرمتون باشند، در این مرحله بین تمامی برچسب‌های پیش‌بینی‌شده برای زیرمتن‌های یک متن رأی اکثریت گرفتیم. برچسبی که برای اکثریت زیرمتن‌ها پیش‌بینی‌شده باشد به‌عنوان برچسب پیش‌بینی‌شده برای D در نظر گرفته می‌شود. در صورتی که تعداد برچسب‌های درست و نادرست پیش‌بینی‌شده برای زیرمتن‌های یک متن برابر باشد، ویژگی‌های در نظر گرفته شده برای تمامی زیرمتن‌های $\{d_1, ..., d_n\}$ را پس از میانگین‌گیری به مدل آموزش داده‌شده دادیم. بنابراین برچسب پیش‌بینی شده برای ترکیب ویژگی‌های زیرمتن‌ها به عنوان برچسب متن معرفی خواهد شد.

با توجه پژوهش‌های پیشین نشان داده‌است که پشت‌سر هم قراردادن چهار لایه‌ی آخر مدل BERT بهترین بازنمایی برداری از هر کلمه را در پی دارد[۳]. بنابراین الگوریتم ارائه‌شده در مراحل ۱ تا ۶ را دوباره اجرا کردیم، یک بار با این تفاوت که به جای لایه‌ی یازدهم، چهار لایه‌ی آخر را پشت‌سرهم گذاشته و به‌عنوان بردار هر توکن در نظر گرفتیم. مشاهدات نشان داد این مسئله می‌تواند دقت مدل را افزایش دهد.

[۶] Classification
[۷] Bagging
[۸] Post-Classification

فصل ۳

یافته‌ها و نتایج

در این بخش به بررسی نتایج روش‌های بیان‌شده در فصل گذشته بر روی مجموعه داده‌ی انشاهای دانشجویی دانشگاه تگزاس می‌پردازیم.

۳-۱ نتایج و تفسیر آن‌ها

روش‌های اولیه‌ی آزمایش‌شده در ادامه‌ی پژوهش پایه در این پژوهش ، مبتنی بر یادگیری عمیق بوده و با توجه به نیازمندی بارز این روش‌ها به منابع قوی محاسباتی و زمان‌بر بودن این روش‌ها، در ادامه به بررسی روش‌های کلاسیک برای رده‌بندی متون افراد جهت پیش‌بینی تیپ شخصیتی آن‌ها پرداختیم.

- روش‌های مبتنی به یادگیری عمیق

مدل‌های آزمایشی M1 تا M5 به‌عنوان نمایندگان مدل‌های آزمایش‌شده در این گزارش ذکرشده‌اند. هر یک از این مدل‌ها مشتقاتی داشته‌است. همچنین تعدادی از این روش‌ها در لایه‌های مختلف BERT آزمایش شده‌اند. برخی از روش‌های آزمایش‌شده به علت محدودیت منابع و زمان، با آغاز آزمایش و مشاهده‌ی نتایج اولیه کنار گذاشته‌شده و برخی نیز ادامه یافته اما با روش 10-fold cross validation آزمایش نشده‌اند.

در ادامه بخشی از نتایج به دست آمده در این پژوهش، در جدول ۳-۱ مشاهده می‌شود. تمامی این مدل‌ها با استفاده از 10-fold cross validation صورت گرفته‌اند.

نام مدل	معماری					تیپ‌های شخصیتی					
	تبدیل به تکه‌ها با اندازه‌ی ۲۰۰ حداکثر توکن	وردامبدینگ	استخراج ویژگی جمله	استخراج ویژگی متن	رده‌بند	EXT	NEU	AGR	CON	OPN	میانگین
S_1	**خیر**	**W2V**	**CNN**	**Max**	**MLP**	**58.09**	**59.38**	**56.71**	**57.3**	**62.68**	**58.83**
S_2	**خیر**	**W2V**	**CNN**	**Max**	**MLP**	**58.09**	**57.33**	**56.71**	**56.71**	**61.13**	**57.99**
M_2	خیر	BERT(لایه‌ی 11)	میانگین‌گیری	CNN+Max	MLP	56.36	57.52	56.39	55.85	60.98	57.42
M_3	خیر	BERT(لایه‌ی 10)	میانگین‌گیری	میانگین‌گیری	MLP	**58.82**	58.62	56.19	55.30	59.57	57.70
M_4	خیر	BERT(لایه‌ی 11)	میانگین‌گیری	GRU	MLP	54.37	55.48	52.08	52.54	58.18	53.62
M_5	خیر	BERT(لایه‌ی 11)	میانگین‌گیری	CNN+GRU	MLP	55.29	58.44	56.03	56.40	60.92	57.42
M_6	خیر	W2V	-	میانگین‌گیری	SVM	56.03	58.87	**57.62**	55.91	60.25	57.74
M_7	بله	BERT(چهار لایه‌ی آخر)	-	میانگین‌گیری	SVM	58.59	**59.78**	56.45	57.71	61.26	58.76
BB-SVM	**بله**	**BERT(چهار لایه‌ی آخر)**	**-**	**میانگین‌گیری**	**Bagging-SVM**	**59.30**	**59.39**	**56.52**	**57.84**	**62.09**	**59.03**

شکل ۱-۳: نتایج مدل‌های آزمایش‌شده

سطر اول جدول نشان‌دهنده‌ی بهترین دقت میانگین به‌دست‌آمده در بهترین روش استفاده‌شده در پژوهش پایه است. سطر دوم نیز، بهترین دقت به‌دست‌آمده در پژوهش پایه برای هر ویژگی شخصیتی را نشان می‌دهد.

در ادامه به بررسی مدل‌های ارائه‌شده در روش حل مسئله می‌پردازیم.

- W2V + LSTM + LSTM : M1

همان‌طور که اشاره شد این مدل بسیار زمان‌بر و نیازمند منابع قوی محاسباتی بود. با مشاهده‌ی چند عدد گزارش‌شده به‌عنوان دقت این مدل و امیدبخش نبودن روند آموزش این مدل، از ادامه‌ی آن صرف نظر کردیم.

- BERT + Conv2D + Dense : M2

نتیجه‌ی این مدل روی لایه‌ی یازدهم شبکه‌ی BERT در سطر سوم جدول نشان‌داده‌شده‌است. این مدل

نیز علی‌رغم زمان‌بر بودن و احتیاج به منبع محاسباتی قوی، نتایج چندان خوبی در مقایسه با مدل پایه درپی نداشت.

— BERT + Mean + Dense : M3

این مدل بهترین نتیجه را در لایه‌ی ۱۰ شبکه‌ی BERT ارائه می‌داد که این نتایج در سطر چهارم جدول قابل مشاهده است. این مدل موفق شده‌است دقت تشخیص ویژگی اول شخصیتی را نسبت به مطالعه‌ی پایه بهبود دهد.

— BERT + RNN + Dense : M4

این مدل نیز چنان‌که نتایج آن در در سطر پنجم جدول مشاهده می‌شود نتوانسته نتیجه‌ی مدل پژوهش پایه را بهبود بخشد.

— BERT + Conv2D + RNN + Dense : M5

این شبکه با استفاده از LSTM، SimpleRNN و GRU به عنوان لایه‌ی استخراج ویژگی متن مورد آزمایش قرار گرفته است. به نظر می‌رسید بهترین نتیجه حاصل از لایه‌ی GRU است که نتایج این مدل نیز در سطر ششم جدول نشان داده‌شده‌است. بر این اساس، بهبودی در نتایج مدل پژوهش پایه با استفاده از لایه‌ی کانولوشن برای استخراج ویژگی جملات و استفاده از لایه‌های RNN ایجاد نشده است.

- روش‌های منتهی به رده‌بندهای کلاسیک

با توجه به نیازهای شبکه‌های عمیق به منابع قدرتمند محاسباتی و همچنین زمان‌بر بودن فرایند آموزش در این شبکه‌ها، در ادامه‌ی پژوهش به آموزش مدل‌های کلاسیک پرداختیم. با آن‌که سرعت رده‌بندی به‌طور قابل توجهی افزایش یافت، دقت رده‌بندی هم موفق شد به خوبی بهبود یابد.

— W2V + Mean + SVM : M6

در این مدل که از رده‌بند کلاسیک SVM در آن برای رده‌بندی هر ویژگی شخصیتی استفاده شده‌بود، دقت تشخیص یکی از ویژگی‌های شخصیتی(ویژگی سوم) نسبت به بهترین دقت گزارش‌شده توسط پژوهش قبلی برای این ویژگی بهبود یافت.

- BERT + Mean + SVM : M7

تفاوت این مدل با مدل M6 در مرحله‌ی Word Embedding است که در این مدل، BERT جایگزین Word2Vec شده‌است. به دلیل آنکه پژوهش‌ها نشان داده‌است بهترین نتیجه‌ی حاصل از شبکه‌ی BERT، حاصل از پشت سر هم گذاشتن بردارهای چهار لایه‌ی آخر این شبکه‌است، این مدل با همین ورودی مورد آزمایش قرار گرفت و نتایج نشان داده این روش می‌تواند نسبت به پژوهش پایه دقت تشخیص ویژگی دوم شخصیتی مجموعه‌ی داده‌ی انشاهای داشجویی را بهبود بخشد.

- BERT + Bagging SVM : BB-SVM

مدل نهایی ارائه‌شده در این پژوهش در سطر آخر جدول مشخص شده‌است. این مدل که جزئیات آن در بخش ۳-۲ شرح داده‌شده‌است، توانست به‌صورت میانگین دقت تشخیص را انجام دهد. این مدل علاوه‌بر آنکه با استفاده از BERT به‌عنوان Word Embedding بهره برده‌است، و با استفاده از میانگین‌گیری بین بردارهای tokenهای هر متن، ویژگی‌های هر متن را استخراج می‌کند. در ادامه برای رده‌بندی متون افراد آن‌ها را به Bagged SVM می‌دهد. نتایج به‌دست آمده نشان می‌دهد این روش بهترین دقت میانگین را بین روش‌های آزمایش‌شده داشته‌است.

۳-۲ نتیجه‌گیری

در این پژوهش تلاش شده‌است با توجه به اهمیت تشخیص ویژگی‌های شخصیتی افراد در کاربردهای مختلف، با توجه به نوشته‌هایشان ویژگی‌های شخصیتی آن‌ها را تشخیص دهیم. بدین منظور پس از آزمایش مدل‌های مختلف شبکه‌های عصبی عمیق و همچنین آزمایش رده‌بندهای کلاسیک، به روشی رسیدیم که نسبت به بهترین روش معرفی شده در کارهای پیشین، علاوه بر بهبود قابل‌توجه دقت، پیش‌بینی را با سرعت بسیار بیشتر انجام می‌دهد.

۳-۳ کارهای آینده

در ادامه‌ی این پژوهش و جهت بهبود عملکرد آن می‌توان شبکه‌ی BERT را با مجموعه داده‌ی آموزشی، آموزش داد و وزن‌های لایه‌های انتهایی آن که بیشتر وابسته به یک موضوع خاص هستند را با توجه به داده به‌روزرسانی نمود. همچنین می‌توان با آزمایش مدل نهایی با مجموعه داده‌های مختلف، مدل را به طریقی که در تمامی مجموعه‌داده‌ها به خوبی ویژگی‌های شخصیتی افراد را تشخیص دهد.

مراجع

[1] Arnoux, P.-H., Xu, A., Boyette, N., Mahmud, J., Akkiraju, R., and Sinha, V. 25 tweets to know you: A new model to predict personality with social media. in *Eleventh International AAAI Conference on Web and Social Media* (2017).

[2] Corr, P. J., and Matthews, G. *The Cambridge handbook of personality psychology*. Cambridge University Press Cambridge, .2009

[3] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805.* (2018).

[4] Howlader, P., Pal, K. K., Cuzzocrea, A., and Kumar, S. Predicting facebook-users' personality based on status and linguistic features via flexible regression analysis techniques. in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (2018), ACM, pp. .345–339

[5] Lambiotte, R., and Kosinski, M. Tracking the digital footprints of personality. *Proceedings of the IEEE 102*, 12 (2014), .1939–1934

[6] 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. .253–197

[7] Majumder, N., Poria, S., Gelbukh, A., and Cambria, E. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems 32*, 2 (2017), .79–74

[8] Matthews, G., Deary, I., and Whiteman, M. Personality traits cambridge university press. *Cambridge, UK* (1998).

[9] Mehta, Y., Majumder, N., Gelbukh, A., and Cambria, E. Recent trends in deep learning based personality detection. *arXiv preprint arXiv:190803628.* (2019).

[10] Pennebaker, J. W., and King, L. A. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology 77*, 6 (1999), .1296

[11] Poria, S., Gelbukh, A., Agarwal, B., Cambria, E., and Howard, N. Common sense knowledge based personality recognition from text. in *Mexican International Conference on Artificial Intelligence* (2013), Springer, pp. .496–484

[12] Tadesse, M. M., Lin, H., Xu, B., and Yang, L. Personality predictions based on user behavior on the facebook social media platform. *IEEE Access 6* (2018), .61969–61959

[13] Tandera, T., Suhartono, D., Wongso, R., Prasetio, Y. L., et al. Personality prediction system from facebook users. *Procedia computer science 116* (2017), .611–604

[14] Yin, H., Wang, Y., Li, Q., Xu, W., Yu, Y., and Zhang, T. A network-enhanced prediction method for automobile purchase classification using deep learning. in *PACIS* (2018), p. .111

[15] Zuo, X., Feng, B., Yao, Y., Zhang, T., Zhang, Q., Wang, M., and Zuo, W. A weighted ml-knn model for predicting users' personality traits. in *2013 International Conference on Information Science and Computer Applications (ISCA 2013)* (2013), Atlantis Press.

Abstract:

Personality trait detection has a wide range of applications, including marketing, customer relationship management, and online safety. Machine learning approaches for automatic personality trait detection based on various user input has attracted much attention in recent years. In this paper, we study the application of deep learning to personality trait detection based on user essays in text format. Our work outperforms state-of-the-art by 1.03 percent in average personality trait accuracy, where the previous state-of-the-art was only able to improve its predecessors accuracy by 0.55 percent in the same metric. Also, our approach is more computationally efficient with a training time speed up of 450 percent using a combination of classical and deep learning methods. We provide experimental results for a wide range of deep learning architectures, including various word-embeddings and different approaches to sentence and document vector representation. Our best performing architecture combines BERT word embeddings to represent a document and then uses an SVM classifier for classification. Finally, we get higher accuracy and more efficient memory usage by splitting each document into several chunks as a preprocessing step. We use majority vote from each chunk to determine the final class for each document and outperform all prior work on this task.

Keywords: Personality Traits, Neural Networks, Text Classification, Deep Neural Networks

Iran University of Science and Technology
Computer Engineering Department

Discover the reality of the text

Bachelor of Science Thesis in Computer Engineering

By:

Samin Fatehi

Supervisor:

Sayyed Sauleh Eetemadi

September 2019