دانشکده مهندسی کامپیوتر

جمع‌آوری مجموعه‌داده نوین برای آموزش مدل‌های یادگیری ماشینی جهت دسته‌بندی نوع شخصیت افراد

آزمایشگاه پردازش زبان طبیعی دانشگاه علم و صنعت ایران

محمدمهدی عبداله‌پور

نام استاد کارآموزی:
دکتر بهروز مینایی

پاییز و زمستان ۱۳۹۹

تأییدیه‌ی صحت و اصالت نتایج

بسمه تعالی

اینجانب محمدمهدی عبداله‌پور به شماره دانشجویی ۹۶۵۲۲۲۶۷ دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیه‌ی مطالب مندرج در این گزارش حاصل ۳۰۰ ساعت حضور و کار اینجانب در آزمایشگاه پردازش زبان طبیعی دانشگاه علم و صنعت ایران 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 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.

نام و نام خانوادگی: محمدمهدی عبداله‌پور
امضا و تاریخ: خرداد ۱۴۰۰

چکیده

آزمایشگاه پردازش زبان طبیعی که در دانشکده‌ی مهندسی کامپیوتر دانشگاه علم و صنعت ایران واقع شده‌است، بر روی پروژه‌های مربوط به این حوزه از جمله پیاده‌سازی مدل‌های یادگیری عمیق، تعریف مدل‌های جدید، تولید و ارزیابی مجموعه‌داده‌های نوین و ... متمرکز است. به طور مشخص، یکی از پروژه‌های انجام شده در این آزمایشگاه تشخیص نوع شخصیتی افراد از روی متون نوشته شده‌شان بوده‌است. این پروژه با تاکید بر تعریف و پیاده‌سازی مدل‌های جدید در این حوزه، از مجموعه‌داده‌های موجود در آن زمان که همگی به زبان انگلیسی بوده‌اند برای آموزش مدل‌ها استفاده کرده‌است. در طی فرصت کارآموزی اینجانب تلاش شد که با استفاده از روش‌های گوناگون مجموعه‌داده‌ای نوین تدوین شود. این تلاش در دو بخش انجام گرفت که در این گزارش مفصل به آن‌ها پرداخته‌خواهد شد. قابل ذکر است که نتیجه‌ی نهایی که تشکیل یک مجموعه‌داده‌ی فارسی در بستر توییتر بوده از بخش دوم این تلاش حاصل گردیده‌است.

واژه‌های کلیدی: پردازش زبان طبیعی، یادگیری عمیق، داده‌کاوی، مجموعه‌داده، تشخیص نوع شخصیتی

فهرست مطالب

فصل ۱:
معرفی حوزه کارآموزی

۱.۱ مقدمه

کارآموزی اینجانب در آزمایشگاه پردازش زبان طبیعی در دانشکده‌ی کامپیوتر دانشگاه علم و صنعت ایران سپری شد. دانشجویان از مقاطع مختلف در این آزمایشگاه تحت نظر دکتر اعتمادی و دکتر مینایی از اساتید این دانشکده بر روی پروژه‌های تحقیقاتی مختلفی در زمینه داده‌کاوی و هوش مصنوعی در حیطه‌ی پردازش زبان طبیعی و بعضا در همکاری با شرکت دادماتک[۱] مشغول به کار هستند.

از جمله‌ی پروژه‌های انجام شده و یا در حال انجام در این آزمایشگاه می‌توان به تشخیص اخبار جعلی فارسی توسط آقای زرهرن، چت بات دامنه باز توسط خانم زهرا سیدی، پاسخ دادن به سوالات مبتنی بر تصویر توسط خانم هاشمی، استخراج پرسش و پاسخ بدون ناظر توسط آقای لطفی و تشخیص گوینده و مکالمه از روی متن و صدا توسط آقای پوردبیری اشاره کرد.[۲]

پروژه‌ی دیگری که به نوعی می‌توان پروژه‌های انجام شده در کارآموزی بنده را در ادامه‌ی آن قلمداد کرد تشخیص نوع شخصیتی افراد از روی متون نوشته شده‌شان است که توسط آقای کاظمینی و خانم فاتحی به عنوان پروژه‌ی پایانی کارشناسی به انجام رسیده‌است. حاصل این پروژه دو مقاله‌ی چاپ شده در کنفرانس ICDM و یک کارگاه بین‌المللی بوده است.[۱][۲]

۱.۲ ارزیابی و پیشنهادات

این آزمایشگاه بستر بسیار مناسبی برای افرادی که به تازگی وارد این حوزه می‌شوند فراهم آورده است. فرایند کار برای کارآموزان عموما به این صورت است که با نظر سرپرست آزمایشگاه، کارآموز تحت نظارت و یا در همکاری با یکی از دانشجویان مقاطع بالاتر به انجام پروژه‌ای مرتبط با پروژه‌های قبلی و یا در حال انجام دانشجوی ارشد مشغول می‌شود. این روند مزایا و معایبی دارد که در ادامه به آن پرداخته می‌شود.

گذراندن دوره‌ی کارآموزی در همکاری با یکی از دانشجویان مقاطع بالاتر به خصوص در همان دانشگاه محل تحصیل مزایای منحصر به فردی دارد. اولین ویژگی مثبت این است که دانشجوی کارشناسی که به تازگی وارد این حوزه‌ی تخصصی می‌شود با فردی که به نوعی فقط یک قدم از او جلوتر است همکاری سازنده و آموزنده‌ای می‌تواند برقرار کند. به عبارت دیگر، اگر فردی با اختلاف سنی و تجربه‌ای بسیار زیاد نظارت او را بر عهده بگیرد ممکن است به دلیل عدم درک صحیح از نحوه‌ی تفکر، ارتباط کلامی دشوارتر و همچنین عدم آگاهی دقیق از دانش و پیش‌زمینه‌های کارآموز این همکاری ناکارآمد شود. مزیت دیگر این است که هماهنگی و برنامه‌ریزی برای پیش‌برد پروژه راحت‌تر و با اختلاف نظر کمتری صورت می‌گیرد.

گذراندن کارآموزی با راهنمایی یک دانشجوی مقطع بالاتر با وجود اینکه مزایایی دارد، دارای نقاط ضعفی است. کمبود تجربه و دانش در مواجهه با مشکلات فنی و غیرفنی را که در مراحل مختلف پروژه ممکن است به‌وجود بیاید می‌توان از مشکلات این نوع از همکاری قلمداد کرد. به طور دقیق‌تر، عدم رهبری

[۱] dadmatech.ir

[۲] قابل ذکر است که تمامی این پروژه‌ها به صورت تیمی انجام شده و در اینجا فقط نام یکی از افراد شاخص آورده شده است.

تیم توسط فردی با تجربه‌ی حرفه‌ای طولانی مدت (به اصطلاح سنیور) این عیب را دارد که در مواجهه با مشکلی جدید کل تیم با پیدا کردن راه‌حل به مدت بیش‌تری دست و پنجه نرم می‌کند. اما نکته‌ی مهم‌تر این است که بسیاری از کارآموزان تجربه و مهارت‌های بالایی دارند که باعث می‌شود از این نوع همکاری سود آموزشی زیادی به دست نیاورند.

برای جمع بندی باید گفت که دانشجویان باید از پیش درمورد روند اجرای پروژه‌ها برای کارآموزان در هر حوزه‌ای تحقیقات لازم را به عمل بیاورند و با توجه به مهارت‌ها و نیازهای خود در مورد سپری کردن کارآموزی در آن محیط تصمیم‌گیری کنند. اگر دانشجویی تازه‌کار است استفاده از فرصت پیش‌آمده در محیط این آزمایشگاه می‌تواند بسیار مفید و آموزنده باشد. با این حال اگر دانشجویی با مهارت و دانش کافی بخواهد در یک محیط حرفه‌ای که بر روی پروژه‌های با وسعت و اندازه‌ی بالاتری کار کند و تجربه‌هایی از نحوه‌ی مواجهه با چالش‌ها از افراد کارکشته به دست آورد بهتر است به حوزه‌های دیگر که بستر همکاری مستقیم با افراد بسیار باتجربه را فراهم می‌کنند مراجعه کند.

فصل ۲:
مشروح فعالیت های انجام شده در محل استقرار

۱.۳ مقدمه

در حال حاضر برای آموزش اکثر مدل‌های یادگیری عمیق استفاده از مجموعه‌داده‌های لیبل‌گذاری شده با حجم بالا یک ضرورت است. همانطور که پیش‌تر اشاره شد در پروژه‌ی تشخیص نوع شخصیتی افراد که توسط تیم آقای کاظمینی انجام شد برای آموزش مدل‌های عمیق از مجموعه‌داده‌های آماده که همگی آن‌ها به زبان انگلیسی بوده‌اند استفاده شده‌است. طبق برنامه‌ریزی‌های انجام‌شده ساخت یک مجموعه‌داده‌ی جدید به هدف ارتقای کیفیت مدل‌های پیشین و همچنین پیش‌برد آن در زمینه‌های دیگر از اهداف کارآموزی اینجانب تعریف شد.

۱.۴ تلاش اول: مجموعه‌داده‌ای از گفت‌وگوی شخصیت‌های تخیلی

در اولین تلاش برای ساختن یک مجموعه‌داده با توجه به این نکته که مکالمات موجود در فیلم‌ها و انیمیشن‌ها یک منبع بزرگ و غنی از گفته‌های شخصیت‌هاست، جمع‌آوری و لیبل‌گذاری این نوع داده از منابع مختلف اینترنتی در دستور کار قرار گرفت.

درنتیجه‌ی جست‌وجوهای به عمل آمده، یافتیم که در حال حاضر یک مجموعه‌داده در وب‌سایت کگل[۱] موجود است که حاوی اطلاعات مفیدی است که دو قسمت نام شخصیت و صحبت گفته شده توسط او برای هدف ما مفید به نظر می‌رسد. اما برای کامل کردن مجموعه‌داده مورد نیاز باید لیبل نوع شخصیتی این شخصیت‌ها را نیز به گونه‌ای با این مجموعه‌داده تلفیق کنیم. برای این کار داده‌های موجود در یک وب‌سایت[۲] که توسط رای‌دهی افراد مشتاق جمع‌آوری شده‌است برای نیاز ما موردقبول به نظر می‌رسد. داده‌های مهم از این وب‌سایت حاوی نام شخصیت‌ها و نوع شخصیتی‌شان به براساس‌های روش‌های نام‌گذاری مختلف بوده است.

بعد از مشخص کردن منابع اصلی داده‌های خام وارد مرحله‌ی جمع‌آوری آن داده‌ها شدیم. در این مرحله با استفاده از ابزارها و مهارت‌های گوناگون که بخشی از آن‌ها در زیر آورده شده‌است داده‌های این دو منبع جمع‌آوری شد:

- زبان اف‌شارپ[۳]
- کتاب‌خانه‌های اف‌اس لب[۴]
- پایگاه‌داده مونگو[۵]
- پلتفرم داکر[۶]

---

[۱] kaggle.com

[۲] personality-database.com

[۳] fsharp.org

[۴] fslab.org

[۵] mongodb.com

[۶] docker.com

در مرحله‌ی بعد باید بر اساس نام شخصیت‌ها این دو مجموعه‌داده را با هم تلفیق (شبیه به عمل جوین در پایگاه‌داده) کرد. این کار در یک تلاش محدود توسط خانم فاتحی انجام شد اما نتیجه‌ی حاصل از لحاظ تعداد تطبیق‌ها برای هدف ما مطلوب محسوب نمی‌شد. با وجود این که چالش‌های این مرحله می‌توانست با روش‌های مختلف مورد حمله قرار بگیرد به دلایل مختلفی که مهم‌ترین آن تجمیع نیروهای تیم برای فرستادن مقاله‌ای برای یک کارگاه بین‌المللی بوده‌است تلاش‌ها برای به ثمر رساندن آن متوقف شد.

به‌طور خلاصه حاصل نهایی این مرحله کسب تجربه‌ی کار با ابزارهای نام‌برده‌شده بوده است که برای موقعیت‌های آینده بسیار مفید واقع خواهد شد. همچنین رفع چالش‌ها و ادامه‌ی مراحل این تلاش به فرصت‌های بعدی از جلمه پروژه‌ی پایانی کارشناسی موکول شد.

۱.۵ تلاش دوم: مجموعه‌داده‌ای فارسی از بستر توییتر[۱]

طبق هدف‌گذاری‌ای که از اواسط دوره‌ی کارآموزی انجام گرفت تصمیم بر آن شد که تمرکز تیم بر روی ساخت یک مجموعه‌داده‌ی فارسی و آموزش مدل‌هایی نه‌چندان پیچیده و در نهایت نوشتن یک مقاله با ساختارهای استاندارد برای شرکت در یک کارگاه بین‌المللی قرار گیرد.

با بررسی‌های انجام شده این نتیجه حاصل شد که مجموعه‌داده‌ی فارسی‌ای در این زمینه موجود نمی‌باشد و حتی منبعی از داده‌های خام مربوط در دسترس نیست. با در نظر این نکته تصمیم گرفته شد که از دو روش مبتنی بر بستر توییتر اما مجزا به حل این مساله پرداخته شود. روش اول تعریف و ساخت یک پرسشنامه‌ی مجازی و پخش آن در محیط‌های مختلف مجازی و روش دوم جمع‌آوری مستقیم قسمت‌های مختلف مجموعه‌داده از توییتر بوده‌است.

در روش اول در ابتدا با استفاده از فرم‌های گوگل[۲] یک پرسشنامه‌ی ساده تدوین و در محیط‌های مختلفی از جمله کانال‌های تلگرامی[۳] و بستر توییتر منتشر شد که از پاسخ‌دهندگان شناسه‌ی توییترشان و نوع شخصیتی‌شان (بر اساس روش ام‌بی‌تی‌آی[۴]) را می‌پرسید. اما به دلالیل گوناگونی که مهم‌ترین آن‌ها عدم توجه خوانندگان به توضیحات و درنتیجه وارد کردن اطلاعات نادرست در پرسشنامه بود پرسشنامه از حالت ساده خارج و با استفاده از منطق‌های شرطی و کنترل‌های خودکار یک پرسشنامه‌ی دیگر این بار فقط در بستر توییتر منشتر شد که کیفیت داده‌های ورودی را بسیار بهبود بخشید. قابل ذکر است که در نهایت پرسش‌نامه با استفاده از فرم‌های مایکروسافت[۵] به انجام رسید که علاوه‌بر ظاهر زیباتر، عملکرد و ویژگی‌های بیشتری را در اختیار قرار می‌دهد.

---

[۱] twitter.com

[۲] docs.google.com/forms

[۳] telegram.org

[۴] en.wikipedia.org/wiki/Myers–Briggs_Type_Indicator

[۵] forms.office.com

در روش دوم با استفاده از دو ابزار سعی شد که اطلاعات کاربرانی که نوع شخصیتی خود را در توییتر اعلام کرده‌اند جمع‌آوری شود. در این راستا ابتدا از ابزار سلنیوم[۱] برای جمع‌آوری داده‌ها استفاده شد که به دلیل مقابله‌ی توییتر با استفاده از کپچا[۲] عملا بسیار کند و ناکارآمد محسوب و در نهایت کنار گذاشته شد. با وجود تمام محدودیت‌ها در ادامه با استفاده از رابط برنامه‌نویسی رسمی توییتر[۳] اطلاعات افراد جمع‌آوری و با استفاده از زبان پایتون[۴] پیش‌پردازش‌های لازم جهت تمیز کردن و مخفی کردن هویت کاربران انجام شد.

در نهایت بیش از یک میلیون توییت از بیش از ۹۰۰ کاربر جمع‌آوری شد که بیش‌تر از ۹۰ درصد آن‌ها از روش دوم به دست آمد. در ادامه با توجه به وقت کم باقی مانده فقط یک مدل ابتدایی از برت[۳] که روی زبان فارسی از پیش آموزش دیده بود استفاده شد که دقت حاصل از آن بسیار پایین محسوب می‌شود. با بررسی‌های انجام شده این دقت پایین می‌تواند حاصل چندین مورد باشد که به‌طور خلاصه در پایین آورده شده است.

- عملکرد ضعیف روش ارزیابی شخصیت امبی‌تی‌آی که در بسیاری از جوامع علمی نیز مورد تاکید قرار گرفته است.
- عملکرد ضعیف مدل پارس‌برت[۴] در مقایسه با هم‌نوع انگلیسی آن
- استفاده از دسته‌بند بسیار ساده در بالای بردارهای حاصل از مدل برت
- شک در صحت برخی از لیبل داده‌ها که با استفاده از روش خوداظهاری به دست آمده است.

با این وجود از فرصت باقی مانده برای تکمیل نوشتن مقاله و تهیه‌ی گزارش‌های لازم برای آن استفاده شد.

در جمع‌بندی، حاصل این تلاش برای اینجانب به عنوان کارآموز، تجربه‌ی بسیار ارزشمند نوشتن مقاله و آشنایی با چالش‌ها و روش‌های مختلف جمع‌آوری داده در یک زبان با منابع محدود و همچنین آشنایی بیشتر با جنبه‌های مختلف علم روان‌شناسی بوده‌است.

[۱] selenium.dev

[۲] en.wikipedia.org/wiki/CAPTCHA

[۳] developer.twitter.com/en/docs/twitter-api

[۴] python.org

فصل ۳:
مراجع

فهرست مراجع

1: Amirmohammad Kazameini, Samin Fatehi, Yash Mehta. Sauleh Eetemadi, Erik Cambria, Personality Trait Detection Using Bagged SVM over BERT Word Embedding Ensembles, 2020
2: Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, Sauleh Eetemadi, Bottom-Up and Top-Down: Predicting Personality with Psycholinguistic and Language Model Features, 2020
3: Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018
4: Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, Mohammad Manthouri, ParsBERT: Transformer-based Model for Persian Language Understanding, 2020