



دانشکده مهندسی کامپیوتر

دستیار آزمایشگاه

آزمایشگاه پردازش زبان طبیعی دانشگاه علم و صنعت ایران

مهسا قادران

استاد کارآموزی

دکتر محمد طاهر پیلهور

اسفند ماه ۱۳۹۹

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

تأییدیه‌ی صحت و اصالت نتایج

بسمه تعالی

اینجانب مهسا قادران به شماره دانشجویی ۹۶۵۲۱۴۱۶ دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیه‌ی مطالب مندرج در این گزارش حاصل ۳۰۰ ساعت حضور و کار اینجانب در شرکت/کارخانه آزمایشگاه پردازش زبان طبیعی دانشگاه علم و صنعت ایران و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری شده 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 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.

نام و نام خانوادگی : مهسا قادران

امضا و تاریخ : ۱۳۹۹/۱۲/۲۰

تشکر و قدردانی:

از استاد گرامی، جناب آقای دکتر محمدطاهر پیلهور و آقای دکتر صالح اعتمادی بسیار سپاسگزارم که در این دوره مرا راهنمایی نمودند تا بتوانم تحقیقات در زمینه مورد علاقه‌ام را انجام دهم. همچنین از راهنمایی‌های جناب آقای مجید زرهرن که در به فرجام رساندن این دوره مرا یاری نمودند، تشکر می‌نمایم.

چکیده

امروزه با گسترش شبکه جهانی اینترنت و استفاده روزافزون از رسانه‌های اجتماعی، همواره حجم گسترده‌ای از متون در زمینه‌های متفاوت، میان کاربران منتشر می‌شود. همه متون قابل استناد نیستند و کاربران با توجه به سلیقه و خواسته خود می‌توانند محتوای مورد نظرشان را در شبکه‌های اجتماعی بدون هیچ‌گونه اعتبارسنجی و نظارتی منتشر کنند. وجود نهادی برای ارزیابی دائم حجم عظیم محتوای در حال انتشار و مقابله با محتوای جعلی ناصحیح می‌تواند موثر واقع شود چرا که انتشار محتوای جعلی می‌تواند پیامدهای جبران‌ناپذیری برای فرد و جامعه در برداشته باشد. کارآموزی اینجانب در آزمایشگاه زبان‌های طبیعی دانشگاه علم و صنعت گذرانده شده است. با توجه به شرایط خاص ناشی از شیوع همگانی ویروس کرونا و محدودیت‌های حاکم، درصد قابل توجهی از کارآموزی اینجانب به صورت دورکاری انجام گرفت و ارائه کارها به صورت منظم و روزانه صورت گرفته است.

واژه‌های کلیدی: هوش مصنوعی - پردازش زبان طبیعی - تشخیص موضع ادعا به متن - یادگیری عمیق -

تشخیص اخبار جعلی

فصل ۱ معرفی حوزه کارآموزی

- ۱
۱-۱ مقدمه ۲
۱-۱-۱ پژوهش‌های آزمایشگاه ۲
۲-۱-۱ نحوه همکاری اعضا ۲
۳-۱-۱ امکانات ۳

فصل ۲ مشروح فعالیت‌های انجام شده در محل استقرار

- ۱-۲ مقدمه ۵
۲-۲ معرفی واحد محل استقرار کارآموز ۵
۱-۲-۲ برنامه ریزی و زمان بندی ۵
۲-۲-۲ بررسی تحقیقات مشابه پیشین ۶
۳-۲-۲ پیاده سازی ۶
۴-۲-۲ کار گروهی ۷
۵-۲-۲ مدیریت گروه ۷
۳-۲ شرح پروژه‌ها و فعالیت‌های انجام شده توسط کارآموز ۹
۱-۳-۲ تحقیقاتی ۹
۲-۳-۲ آزمایشگاهی ۱۰
۳-۳-۲ طراحی ۱۱
۴-۳-۲ ترجمه متون علمی ۱۱
۵-۳-۲ برنامه نویسی ۱۴
۶-۳-۲ مدیریتی ۱۴
۷-۳-۲ نتیجه گیری ۱۴

فصل ۳ نتیجه‌گیری و پیشنهادها

- ۱-۳ مقدمه ۱۶
۱-۳-۱ خلاصه فعالیت‌های انجام شده ۱۶
۲-۳ اعلام پیشنهادهایی برای رفع چالش‌های حوزه/واحد کارآموزی ۱۷

فصل ۴ مراجع

۱۸

فصل ۱

معرفی حوزه کار آموزی

۱-۱ مقدمه

آزمایشگاه پردازش زبان‌های طبیعی دانشگاه علم و صنعت ایران تحت نظر دکتر محمد طاهر پیله و دکتر صالح اعتمادی می‌باشد. در این آزمایشگاه گروه‌های دانشجویی مشتاق، در حال انجام تحقیقات و پژوهش‌های علمی در این زمینه می‌باشند. هدف اصلی این آزمایشگاه پیاده‌سازی ابزار و نوآوری در روش‌هایی است که بتوانند از متون به زبان فارسی اطلاعات مورد نیاز تعریف شده را به صورت هوشمند و بدون دخالت انسان با بالاترین دقت استخراج کنند و یا با توجه به وظیفه تعریف شده متن مورد نظر را تولید کنند.

۱-۱-۱ پژوهش‌های آزمایشگاه

پروژه‌ها پس از بررسی مسائل روز و با همفکری و تحت نظر اساتید، توسط یک یا چند دانشجو در مقاطع تحصیلاتی مختلف، تکمیل و مشخص می‌شوند. نقطه قوت تعدادی از پروژه‌ها این است که به گونه‌ای تعریف شده‌اند که مکمل یکدیگر باشند و پس از اتمام، بتوانند با هم ادغام شوند و به فعالیت‌های تحقیقاتی و یا صنعتی والاتر منجر شوند.

پروژه‌های متعددی در این آزمایشگاه تعریف شده است. برای مثال می‌توان به استخراج هویت شخص با توجه به متن‌های نوشته شده توسط همان شخص، تشخیص اخبار جعلی، تولید ابزار برای گفت‌وگو در شبکه‌های اجتماعی، جداسازی زیرنویس فیلم با توجه به گوینده آن، طراحی سیستم جهت پرسش و پاسخ اشاره کرد. هر یک از پروژه‌های تعریف شده زیر شاخه‌ای از علم پردازش زبان طبیعی قرار می‌گیرند و انجام پروژه‌های تعریف شده، غالباً در ابتدا هر پروژه به تعدادی زیرپروژه شکسته می‌شود و نهایتاً پس از به نتیجه رسیدن بخش‌های کوچکتر و تجمیع آن‌ها پروژه نهایی شکل می‌گیرد.

دستاورد‌های تحقیقات فعالین آزمایشگاه به یاری تجربه و دانش اساتید و دانشجویان در مقاطع بالاتر به صورت مقاله در آمده و برای بررسی و چاپ به کنفرانس‌ها و مجلات معتبر فرستاده می‌شوند. از مزایای دیگر این آزمایشگاه می‌توان به استفاده عملی از پژوهش‌های مستعد، در صنعت اشاره نمود. چنین پژوهش‌هایی پس از کار تحقیقاتی پایان نمی‌یابند و سعی می‌شود از نتیجه کار انجام گرفته، در صنعت کشور نیز از آن استفاده و مورد بهره برداری واقع شود.

۱-۱-۲ نحوه همکاری اعضا

هر گروه متشکل از یک یا چند دانشجو می‌باشد که طی جلسه‌های منظم با اساتید راهنما کارها را پیش می‌برند. جلساتی نیز به صورت میان گروهی، متشکل از گروه‌های فعال در آزمایشگاه به بحث در محوریت تحقیقات انجام شده می‌پردازند و نتایج تحقیقات و دست‌آوردها را به اشتراک می‌گذارند. در این جلسات

فعالیت‌های گروه‌ها توسط استاد راهنما و همچنین افراد سایر گروه‌ها بررسی و هدایت می‌شود تا هر گروه در بهترین راستا، پژوهش خود را پیش ببرد.

این گونه جلسات گروهی باعث می‌شود دانشجویان تجربیات ارزشمند خود، اعم از موفقیت‌ها و شکست‌ها را با سایرین به اشتراک بگذارند و همچنین در ادامه راه بتوانند از تجربه سایر گروه‌ها استفاده کنند. علاوه بر این، وجود جلسات منظم انگیزه گروه‌ها برای ادامه‌ی مسیر را نیز تقویت می‌کند. از نقاط قوت این آزمایشگاه رقابت سالم، همکاری و فضای دوستانه حاکم بین اعضا با یکدیگر در جهت بهبود و ارتقاء سطح کل آزمایشگاه است.

برای به سرانجام رسیدن پروژه‌های تعریف شده با توجه به نیاز، دانشجویان مقطع کارشناسی که به زمینه مربوطه علاقه‌مند باشند نیز به پروژه ملحق می‌شوند. این فرصت خوبی برای دانشجویان کارشناسی می‌باشد تا در زمینه‌ی مورد علاقه‌شان مطالعات و پژوهش عمیق‌تری انجام دهند و با نحوه‌ی انجام و مدیریت پروژه‌هایی با ابعاد بزرگتر نیز آشنا شوند و آمادگی لازم برای انجام چنین پروژه‌هایی را به دست آورند.

۱-۱-۳ امکانات

این آزمایشگاه تحت نظر دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت می‌باشد. یک آزمایشگاه با تعدادی سیستم با قدرت‌های سخت افزاری مختلف و واحدهای پردازنده گرافیکی جهت انجام تحقیقات پروژه‌های هوش مصنوعی در اختیار دانشجویان قرار داده شده است. البته با توجه به تعداد فعالین در آزمایشگاه امکان استفاده برای همه اعضا به صورت همزمان وجود ندارد و بر حسب نیاز افراد گروه از سیستم‌های موجود در آزمایشگاه برای انجام کارهای تحقیقاتی استفاده می‌کنند.

فصل ۲

مشروح فعالیت‌های انجام شده در محل استقرار

۱-۲ مقدمه

انسان‌ها با استفاده از زبان با یکدیگر ارتباط برقرار می‌کنند. زبان، سیستم پیچیده‌ای است که یادگیری آن توسط انسان سال‌ها به طول می‌انجامد. با وجود این همچنان اشخاص متفاوت برداشت‌های متفاوتی از یک متن ثابت می‌کنند و واضح است که فهم و یادگیری زبان و در لایه عمیق‌تر استخراج مفاهیم از زبان‌های طبیعی پیچیده باشد.

هوش مصنوعی در جای‌جای زندگی ما راه پیدا کرده است. پردازش و فهم اطلاعات مورد نیاز از متن یکی از گرایش‌های علم هوش مصنوعی است. با توجه به گسترش علم هوش مصنوعی به ویژه مباحثی مانند یادگیری ماشین و یادگیری عمیق استخراج مفاهیم، دسته بندی و حتی تولید متون امکان پذیر گشته است. با توجه به گسترش روز افزون استفاده از رسانه‌های اجتماعی، هر کاربر می‌تواند خبر مورد علاقه (بعضاً با هدف نامناسب پیشین) خود را در شبکه‌های اجتماعی نشر دهد. همچنین با توجه به این که معمولاً اخبار کذب، اخباری هستند که احساسات را بیشتر تحریک می‌کنند و در نتیجه با سرعت بیشتری در شبکه‌های اجتماعی پخش می‌شوند و می‌توانند آسیب‌های جبران‌ناپذیری به جامعه و افراد جامعه و یا گروه خاص مد نظر وارد کنند. محتوی منتشر شده در فضای مجازی و رسانه اجتماعی، غالباً فاقد اعتبار می‌باشند و عموم مردم توجه کمتری به این موضوع دارند. به همین دلیل تشخیص اخبار جعلی و کذب از بین محتوایاتی که همواره در شبکه‌ها در حال رد و بدل شدن است و همچنین در قدم بعدی جلوگیری از انتشار آن یا اخطار به کاربران، تاثیر بسزایی در کاهش انتشار این اخبار می‌تواند داشته باشد و در نتیجه می‌توان از آثار مخرب آن جلوگیری کرد.

پروژه تشخیص اخبار جعلی در این آزمایشگاه جهت پیاده‌سازی سیستم خودکار تشخیص میزان صحت اخبار جعلی تعریف شده است. اینجانب در دوران کارآموزی، در این پروژه فعالیت داشتم.

۲-۲ معرفی واحد محل استقرار کارآموز

پروژه به طور کلی شامل سه مرحله می‌باشد که در ادامه به شرح هر مرحله پرداخته می‌شود.

۱-۲-۲ برنامه ریزی و زمان بندی

هنگام انجام پروژه‌های بلند مدت و در ابعاد بزرگ، از اصول اولیه برای رسیدن به نتیجه‌ی مطلوب، تقسیم کردن پروژه به تعدادی فاز کلی و همچنین مشخص کردن زمانی برای تمام کردن مراحل است.

مشخص کردن هدف‌های کوتاه مدت نه تنها نظم گروه را افزایش می‌دهد، بلکه باعث می‌شود روند انجام کار برای همه‌ی اعضای گروه واضح‌تر بوده و از طرفی انگیزه‌ی بیشتری برای اعضای گروه جهت ادامه اجرای پروژه و اتمام به موقع آن را ایجاد می‌کند. این پروژه به سه فاز کلی: ۱. تحقیق و بررسی پژوهش‌ها و فعالیت‌های انجام شده. ۲. جمع‌آوری مخزن اطلاعات مورد نیاز. ۳. پیاده‌سازی و تقسیم شده بود و زمان تقریبی اتمام هر فاز نیز برای همه اعضای تیم مشخص بود.

۲-۲-۲ بررسی تحقیقات مشابه پیشین

اولین قدم برای انجام هر پروژه‌ای بررسی و مطالعه‌ی تحقیقات روز انجام شده در آن زمینه می‌باشد. در گرایش پردازش زبان‌های طبیعی با توجه به این که هر زبان قواعد و ویژگی‌های منحصر به فرد خودش را دارد و مطالعات و پژوهش‌های اندکی مستقیماً روی زبان فارسی انجام گرفته است، بررسی کارهای انجام شده در زبان فارسی از این نظر که زبان فارسی مورد بررسی قرار گرفته است از اهمیت بالایی برخوردار است. چالش‌هایی که در انجام پژوهش‌های به زبان فارسی با آن‌ها مواجه شده‌اند و نحوه‌ی رفع آنها در انجام پروژه نقش بسزایی ایفاء می‌کند. همچنین ایده‌ها و روش‌هایی که در زبان‌های غیر فارسی استفاده شده است و نتایج مطلوبی که به دست آمده نیز می‌توانند در زبان فارسی شبیه‌سازی شده، و نتیجه آنها در این زبان نیز مورد بررسی قرار گیرد و از آنها برای ادامه‌ی کار ایده گرفت. چرا که با وجود کلمات و قواعد گوناگون بین زبان‌های متفاوت از نظر ارتباط مفهومی بین کلمات زبان‌ها تا اندازه‌ای می‌توانند شبیه به هم باشند و مانند همدیگر عمل کنند.

در این پروژه نیز پژوهش‌هایی که در گروه آزمایشگاه از قبل انجام شده بود، توسط افراد جدید به طور دقیق مورد مطالعه قرار گرفت و توضیحات لازم درباره‌ی جزئیات کار داده شد. علاوه بر آن مقالات روز و همچنین گزارش‌های شرکت‌های پیش‌تاز که در این زمینه فعالیت می‌کنند مورد بررسی قرار گرفت.

۳-۲-۲ پیاده‌سازی

پس از انجام تحقیقات و مطالعات لازم و اطمینان از به‌روز بودن اطلاعات، با الهام از پژوهش‌های انجام شده و با توجه به نقاط ضعف و قوت و تفاوت‌های بین زبان انگلیسی و فارسی لازم بود تا ایده‌های موجود برای انجام پروژه از ابعاد مختلف مانند در دسترس بودن اطلاعات و قابل جمع‌آوری بودن آن، وجود امکانات سخت افزاری و نرم افزاری لازم برای پیاده‌سازی ایده، تخمین زمان پیاده‌سازی و به سرانجام رساندن ایده، تخمین از میزان عملکرد ایده در صورت تجربیات پیشین و یا مطالعات انجام گرفته بررسی و امکان‌سنجی شوند. پس از انجام بررسی‌های لازم از میان ایده‌ها و روش‌ها موجود معمولاً به انجام یک روش اکتفا نمی‌کنند و با توجه به زمان و هزینه، از میان ایده‌های عملی به ترتیب اولویت ایده‌هایی که

انتظار می‌رود خروجی بهتری داشته باشند، پیاده‌سازی می‌شوند. نهایتاً نتایج به دست آمده از چند روش با هم مقایسه شده و بهترین آنان به عنوان خروجی نهایی سیستم در نظر گرفته می‌شود. برای پیاده‌سازی این پروژه، پس از چند جلسه صحبت و ایده پردازی توسط همه اعضای گروه، ایده‌ها مورد بررسی قرار گرفتند و چند ایده به عنوان کاندید جهت پیاده‌سازی انتخاب شدند. ایده‌هایی که از مقالات الهام گرفته شده بودند، ابتدا برنامه در زبان اصلی مقاله اجرا شد و نتایج آن ثبت گردید و پس از اعمال تغییرات لازم و اجرای آن روی زبان فارسی، نتایج به دست آمده با زبان فارسی مقایسه شد. با استفاده از نتایج حاصله از مقایسه خروجی یک روش بر دو زبان متفاوت می‌توان تصمیمات بهتری برای بهبود نتایج در زبان فارسی اعمال کرد. نهایتاً از بین روش‌های مختلف آزمایش شده از بهترین روش به عنوان خروجی برای تشخیص اخبار جعلی استفاده شد.

۲-۲-۴ کار گروهی

امروزه اهمیت کار گروهی بر همگان ثابت شده است. انجام کار گروهی نیازمند داشتن مهارت است. معیارهایی مانند پایبندی به گروه و توانایی نرمش با اعضا گروه و همکاری در راستای بهبود پروژه از مهارت‌های لازم برای کار گروهی است.

در بسیاری از مواقع کار گروهی با گروه کاری اشتباه گرفته می‌شود. هنگامی که صحبت از گروه می‌شود بدین معناست که خواسته همه‌ی افراد گروه با هدف گروه باید هم‌راستا باشد. در کار گروهی افراد گروه به یکدیگر انگیزه و روحیه لازم را برای ادامه کار می‌دهند. اهمیت فعالیت گروهی امروزه بر کسی پوشیده نیست و شرکت‌های پیشرفته دنیا حتی چیدمان شرکت‌ها و ساختمان‌ها را بر هدف ارتباط بیشتر کارمندان با همدیگر و پیشرفت کارها به صورت گروهی طراحی می‌کنند.

این پروژه کار به صورت گروهی انجام پذیرفت و اعضای گروه یاد گرفتند ایده‌های همدیگر را به دقت گوش کرده و به آن بیاندیشند و با هم به پرورش ایده‌ها و برطرف کردن مشکلات بپردازند. در انجام کار گروهی چالش‌هایی نظیر اختلاف نظر افراد گروه، همکاری و فعالیت نامساوی بین اعضای گروه و ایجاد تکروری در انجام کارهای گروه انکار ناپذیر است و آنچه اهمیت دارد مهارت افراد برای رفع چالش‌ها و مشکلات پیش آمده است.

۲-۲-۵ مدیریت گروه

روش‌های مدیریت گروه تاثیر بسزایی در پیشرفت گروه داشته و ایجاد دلبستگی و فعالیت بیشتر را بین اعضای گروه را فراهم می‌کند. با مشاهده گروه‌های متفاوت متوجه می‌شویم، بعضی از گروه‌ها موفقیت‌های چشمگیری به دست می‌آورند و این در حالی است که بعضی نیز علی‌رغم تلاش‌های گسترده و جدی صورت گرفته بعد از مدتی سرد شده و از هم می‌پاشند. یکی از عوامل موثر در موفقیت و ادامه‌ی فعالیت گروه اعمال

روش‌های مدیریت مناسب در گروه است. روش‌های نوین و موفق مدیریت گروه مانند اسکرام [۱] که مبتنی بر تشویق اعضای گروه به فعالیت و همکاری بیشتر است، نتایج خوبی به دنبال داشته است و امروزه طرفداران بسیاری دارد. در این روش از مدیریت چهار نوع جلسه وجود دارد. که در ادامه به توضیح مختصر آنها می‌پردازیم.

۱-۲-۵-۲-جلسات برنامه‌ریزی:

این جلسات به جهت برنامه‌ریزی برای بازه زمانی مشخص پیش رو (حدوداً دو هفته) برگزار می‌شود. در این جلسات همه افراد گروه‌ها و مدیریت کلی سیستم حاضر می‌شوند و هدف‌گذاری‌های کوتاه مدت را انجام می‌دهند. در این جلسات مشخص می‌شود وظیفه هر فرد در گروه در این مدت چیست و در پایان این مدت جمعا چه انتظاری از گروه هست که انجام داده باشد.

۲-۲-۵-۲-جلسات ارائه:

در پایان زمان مشخص شده، جلسه‌ای برگزار می‌شود و گروه حاصل تلاشی که در این مدت انجام داده است را ارائه می‌دهند، محصول خروجی به دقت توسط افراد بررسی شده و در صورت وجود هرگونه ایراد و یا ناهماهنگی برای تصحیح به گروه گفته می‌شود.

۳-۲-۵-۲-جلسات ارزیابی عملکرد:

پس از تمام شدن ارائه، افراد هر گروه به صورت جداگانه درباره نحوه عملکرد کل گروه و تک تک اعضا صحبت می‌کنند و سعی می‌شود مشکلات را شناسایی کرده و درباره آنها صحبت کنند و راه‌حلهایی اتخاذ کنند تا در دوره‌ی بعدی مشکلات از بین رفته یا کم‌رنگ‌تر شوند و همچنین درباره‌ی نقاط قوت همدیگر صحبت می‌کنند تا در اعضای گروه حس مثبتی ایجاد و تا جای ممکن این نقاط قوت در دوره‌های بعدی نیز تکرار شوند.

۴-۲-۵-۲-جلسات روزانه:

برگزاری این جلسات به صورت منظم و روزانه می‌باشد. این جلسات به گونه‌ای طراحی شده‌اند که روزانه افراد گروه خلاصه‌وار گزارش کار خود را ارائه می‌دهند و آنچه که می‌خواهند انجام دهند را برای گروه توضیح داده و در صورتی که با مشکلی مواجه شدند آن را در گروه مطرح می‌کنند تا با استفاد از نظرات اعضای گروه مشکل را مرتفع نمایند. جلسات روزانه باعث می‌شوند تا اعضای گروه انگیزه بیشتری برای فعالیت مداوم داشته باشند و بیشتر در جریان کار هم گروهی‌های خود قرار بگیرند و از تجربیات همدیگر استفاده کنند و گروه با سرعت بیشتری در رسیدن به هدف پیش برود. لازم به ذکر است که این نوع از روش‌های مدیریتی در هر گروه تعداد

اعضا ۳ تا ۷ نفر تعریف می‌شود و در جلسات روزانه هر نفر حدود ۵ دقیقه گزارش کلی وضعیت کار خود را می‌دهد.

۵-۲-۲- مدیریت پروژه در کارآموزی

در این پروژه مدیریت گروه تقریباً شبیه به اسکرام برگزار شد. به این صورت که جلسات برنامه‌ریزی و ارایه در بازه‌های یک ماهه انجام می‌گرفت و گزارش‌ها به صورت کتبی روزانه و جلسات حضوری دو روز درمیان انجام گرفت. وجود جلسات حضوری باعث پیشرفت بهتر کار می‌شد و در تیم با مشاهده‌ی فعالیت اعضای گروه انگیزه‌ی بیشتری برای ادامه‌ی فعالیت در گروه ایجاد می‌شد.

۲-۳ شرح پروژه‌ها و فعالیت‌های انجام شده توسط کارآموز

فعالیت‌های صورت گرفته در این دوره به چندین بخش تقسیم می‌شود که در ادامه به شرح مختصری از هر کدام از موارد پرداخته می‌شود.

۲-۳-۱ تحقیقاتی

- شبکه‌های عصبی: از مباحث مرتبط با شبکه‌های عمیق و یادگیری ماشین، ابتدا مطالعات ابتدایی در این زمینه صورت گرفت. پس از آشنایی با مفاهیم روش‌های دسته‌بندی مانند Bayes و Random Forest و SVM و نحوه عملکرد آن‌ها مورد مطالعه قرار گرفتند. پس از آن در زمینه انواع شبکه‌های عصبی که جهت پردازش زبان طبیعی کاربرد دارند مطالعه و یادگیری انجام شد.
- متعادل سازی مجموعه دادگان: با توجه به اینکه در بسیاری از پایگاه‌های داده، از بین دسته‌بندی انجام شده، از هر دسته به میزان متقارن وجود ندارد، یکی از مباحث چالش برانگیز در هوش مصنوعی، استفاده از روش‌هایی است که به رغم عدم تقارن در دسته‌ها و وجود دسته‌های با تعداد نمونه کمتر، شبکه‌های عصبی بتوانند به خوبی اطلاعات مربوط به این دسته را یاد بگیرند. یکی از مباحثی که در این دوره مورد مطالعه قرار گرفته است، بررسی روش‌های متعادل‌سازی پایگاه‌های داده نامتقارن و مقایسه آنها با یکدیگر است. هر کدام از روش‌های موجود، به طور کلی از دو دیدگاه به این مسئله نگاه می‌کنند. ۱. با استفاده از الگوریتمی به تعداد داده‌هایی که در دسته کوچک‌تر هستند اضافه کند و به کل تعداد داده‌ها بیافزاید و یا حتی در صورتی که تعداد نمونه‌ها در یک دسته خیلی بیشتر سایر

دسته‌ها باشد، با استفاده از الگوریتم‌هایی تعدادی از این نمونه‌ها را کمتر کنند، به صورتی که کمترین میزان ممکن اطلاعات از دست رود و یا از ترکیبی از دو روش گفته شده استفاده می‌کنند. ۲. دیدگاه دوم تغییر شبکه‌ی عصبی است به طوری که برای داده‌هایی که در دسته‌ی کوچک‌تر هستند، اهمیت بیشتری قائل باشد.

- ارزیابی صحیح شبکه عصبی: با توجه به این که ارزیابی نحوه عملکرد شبکه امر مهمی است، انتخاب نمونه داده برای ارزیابی شبکه از اهمیت بالایی برخوردار است و در صورتی که اشتباه انجام شود در عمل شبکه آن طور که انتظار می‌رود کار نمی‌کند. هنگامی که مجموعه داده متقارن نباشد، برای ارزیابی شبکه عصبی پس از انجام آموزش شبکه، نمی‌توان از روش‌های معمول برای ارزیابی استفاده کرد. در نتیجه بخش دیگری از تحقیقاتی که انجام داده‌ام درباره روش‌های تقسیم صحیح داده‌ها به دسته‌ی آموزش و ارزیابی در هنگامی است که تعداد نمونه‌های موجود از یک یا چند دسته کمتر از سایر دسته‌ها باشد.
- پردازش زبان طبیعی: به صورت تخصصی در زمینه پردازش زبان طبیعی، روش‌های تشخیص احساسات در متن، تشخیص موجودیت زبانی از متن و همچنین روش‌های تشخیص نقش کلمات در جمله را مورد بررسی قرار داده شد و چندین مقاله در زبان انگلیسی و همچنین تحقیقات انجام شده در زبان فارسی مورد مطالعه قرار گرفت.

۲-۳-۲ آزمایشگاهی

در کنار تحقیقات انجام شده درباره روش‌های متعادل‌سازی دادگان، روش‌های متفاوت مورد آزمایش قرار گرفتند. روش‌های موجود جهت متعادل‌سازی دادگان یا بر پایه داده‌افزایی در گروهی است که داده کمتری دارد و یا داده‌کاهی در گروهی که داده بیشتری دارد. با توجه به اندازه دادگان در این پروژه روش‌های مبتنی بر داده‌افزایی انتخاب شدند و بهترین روش در مدل‌هایی به کار گرفته شد. روش‌هایی که بیشترین تاثیر مثبت در عملکرد شبکه را داشتند، عبارت‌اند از [۲] Smote و [۳] Adasyn و همچنین هایپرپارامترهای شبکه مانند نحوه محاسبه میزان خطا محاسبه شده برای هر دسته، متناسب با معکوس تعداد داده‌های موجود در آن دسته قرار گرفت و این موضوع به شبکه هنگام آموزش کمک می‌کند که به داده‌هایی که دسته کوچک‌تر هستند توجه بیشتری شود. البته این مقادیر طوری باید تنظیم شوند که مورد اول و دوم باعث نشوند یک دسته بیش از اندازه برای شبکه اغراق شود.

پس از انجام تحقیقات درباره NER، به بررسی ابزارهای موجود در زبان فارسی پرداخته شد و در غالب یک سیستم، پیاده‌سازی شد. عملکرد سیستم به این صورت است که هنگام دادن یک متن به عنوان ورودی به سیستم، موجودیت‌های آن متن را به عنوان خروجی نمایش می‌دهد.

با توجه به این که معمولاً اخبار جعلی حاوی کلماتی با بار معنایی منفی هستند و بیشتر حس منفی را به خواننده القا می‌کنند، استفاده از ویژگی بار احساسی جملات به عنوان ورودی مدل، در بهبود عملکرد مدل تاثیرگذار است. برای تشخیص احساسات با استفاده از یکی از دادگان جمع‌آوری شده موجود در زبان فارسی

برای هر کلمه بار احساسی معادل آن قرار داده شد. با توجه به نیاز پروژه، ارزیابی مثبت یا منفی بودن بار معنی کلمات موجود در جمله به عنوان خروجی تشخیص احساسات در نظر گرفته شد.

اصل مقاله‌ی [۴] FEVER که روی زبان انگلیسی پیاده سازی شده است، بار دیگر اجرا و ارزیابی گردید. با توجه به محدودیت منابع سخت افزاری، از بخشی از کل مجموعه صفحه‌های ویکیپدیایی که در مقاله استفاده شده بود، مدل شبکه عصبی را آموزش و ارزیابی کردیم. روند کلی برنامه‌های نوشته شده به این صورت است که ابتدا صفحه‌های وب سایت ویکیپدیا خوانده می‌شوند و پیش پردازش‌های لازم روی متن هر صفحه انجام می‌گیرد. برای هر صفحه با توجه به الگوریتم‌های منتخب، مشابه ترین صفحات موجود در ویکیپدیا در دادگان موجود، پیدا می‌شود. یکی از الگوریتم‌هایی که برای این کار استفاده می‌کند، استفاده از شباهت بردارها است. در قدم بعدی این صفحات و یک جمله موجود در دیتاست را به شبکه عصبی می‌دهد و شبکه رابطه‌ی یک جمله با این صفحات را یاد می‌گیرد.

۲-۳-۳ طراحی

سایر اعضای تیم دیتاست مورد نیاز تشخیص اخبار جعلی و همچنین تشخیص موضوع متن را جمع آوری کردند و همچنین ویژگی‌های دیگری نیز مانند استفاده از درونگرایی کلمات به دست آمده حاصل از مدل BERT [۵] استخراج شدند. به عنوان قدم آخر نیاز بود تا مدلی طراحی شود که این ویژگی‌ها و در صورت لزوم ویژگی دیگری را به عنوان ورودی بگیرد و خروجی مدل میزان صحت خبر را نمایش دهد. هنگام جمع آوری دیتاست ویژگی‌هایی مانند وب سایت منتشر کننده‌ی خبر نیز ذخیره شده است. با مشورت در گروه تصمیم گرفته شد که با استفاده از دیتاست جمع آوری شده اطلاعاتی درباره میزان قابل اعتماد بودن هر منبع خبری محاسبه شود که نهایتاً با توجه به سابقه و بسایت خبری در دادگان در انتشار اخبار جعلی و صحیح، این ویژگی نیز استخراج شد. در قدم آخر با مشورت اعضای گروه شبکه‌ای عمیقی طراحی شد و پس از اجرا، سعی بر بهبود طراحی مدل شد که بتواند یک ادعا را با دقت بهتری اعتبار سنجی کند.

۲-۳-۴ ترجمه متون علمی

۱-۴-۳-۲- ترجمه مقاله Detection: Stance Persian [۶]

این مقاله یکی از مقالات به چاپ رسیده، توسط آزمایشگاه پردازش زبان‌های طبیعی دانشگاه علم و صنعت در سال گذشته بوده است. با توجه به این که این پروژه در ادامه این مقاله تعریف شده بود، برای آشنایی با کلیت کار پس از انجام تحقیقات اولیه درباره مفاهیم پردازش زبان‌های طبیعی این مقاله را با دقت و جز به جز ترجمه کردم و قسمت‌های آن را با جزئیات مطالعه کردم. هدف اصلی این مقاله جمع آوری

دیتاست فارسی با هدف تشخیص موضوع یک جمله در برابر یک متن است. از بین ۲۱۲۴ خبر جمع آوری شده، ۵۳۴ ادعا تولید شده است. در این مقاله ۴ رابطه (برچسب) برای ارتباط متن و ادعا در نظر گرفته شده است. که عبارتند از: موافق، مخالف، مرتبط و نامربوط. در این پروژه برای ساختن ادعا از شایعات و اخبار استفاده شده است. در جمع آوری این دیتاست دقت بالایی به کار گرفته شده است، به طوری که برای برچسب‌هایی که برای هر ادعا زده شده است دو شخص به عنوان سرگروه، دوباره برچسب‌های زده شده را بازبینی کرده اند. پس از جمع آوری دیتا برای استخراج بردار کلمات از دو روش [۷] BoW و [۸] TF-IDF استفاده کرده است. علاوه بر آن ویژگی‌هایی مانند دو بخشی بودن متن ادعا و سوالی بودن آن نیز برای دسته بندی استفاده کرده است. دسته بندی رابطه‌ی بین ادعا و متن در این مقاله از دو دیدگاه کلی ادعا و متن خبر و همچنین ادعا و تیترا خبر استفاده شده است. در این مقاله برای طبقه‌بندی هر کدام از دیدگاه‌های ذکر شده روش‌های دسته بندی Forest Random [۹]، SVM [۱۰]، Logistic Regression [۱۱] و Naive Bayes [۹] بررسی شده اند و نهایتاً با توجه به این که دیتاست گردآوری شده متوازن نیست، خروجی مهمی این حالات با حالتی مقایسه شده‌است که مدل آموزش دیده همواره خروجی را برابر با دسته‌های که از آن بیشترین نمونه موجود است، بدهد. بهترین دقتی که در مقاله به آن رسیده است ۶۸٪ در پیدا کردن رابطه‌ی تیترا و ادعا با استفاده از الگوریتم Forest Random [۱۲] می‌باشد.

۲-۴-۳-۲- ترجمه مقاله FEVER [۴]

نام این مقاله مخفف شده کلمه‌ی Fact Extraction and Verification به معنی 'استخراج صحت متن و اثبات آن می‌باشد. این مقاله تعدادی ادعا تولید کرده است و تلاش بر این دارد که با توجه به اسناد موجود در ویکی‌پدیا میزان صحت این ادعا را مشخص کند و همچنین برای اظهار خود ادله‌ی لازم از متون مرتبط را بیان کند. با توجه به کاربرد نزدیکی که این مقاله به پروژه گروه دارد این مقاله را با دقت کامل مطالعه کردم و آن را ترجمه نمودم.

در این مقاله تولید دیتاست قدم مهم و حیاتی به شمار می‌آید. برای جمع‌آوری اطلاعات مورد نیاز این گروه از دو ابزار جداگانه برای افزایش دقت افرادی که اطلاعات را جمع آوری می‌کنند و جلوگیری از استفاده‌ی آنان از اطلاعات پیشین در نظر گرفته شده است. در ابزار اول تنها هدف، تولید ادعا از صفحات ویکی‌پدیا است. به این صورت که فرد پاراگراف اول از یک صفحه ویکی‌پدیا را مشاهده می‌کند و باید یک ادعا از آن تولید کند به صورتی که از عبارات و کلمات جدید استفاده کرده باشد. ادعاهای تولید شده می‌توانند منفی باشند و یا نام یا عبارتی از آن‌ها تغییر کرده باشد. این قوانین به طور کامل در ابزار برای کاربران معرفی شده است.

در قدم بعدی افرادی که می‌خواهند صحت ادعا را از سایت ویکی‌پدیا بررسی کنند بدون در نظر گرفتن هیچ دانش پیشین این کار را انجام می‌دهند. با توجه به این که صحت ادعای مورد نظر را با اسناد موجود در ویکی‌پدیا بتوان اثبات کرد یا نه، یکی از برچسب‌های تایید یا رد و یا اطلاعات ناکافی است، انتخاب می‌شود. در صورتی که متنی از ویکی‌پدیا به عنوان متن مرتبط در نظر بگیرند توضیح مختصر ویکی‌پدیا موجود در پاراگراف اول نیز برای آنان نمایش داده می‌شود و همچنین در دیتاست ذخیره می‌شود. با این روش سعی شده است که تا جای ممکن افرادی که ادعا را برچسب می‌زنند و شواهد و ادله خود را ویکی‌پدیا پیدا می‌کنند بدون در نظر گرفتن دانش قبلی این عمل را انجام دهند.

همچنین در مقاله به چالش‌های موجود در هنگام جمع‌آوری اطلاعات اشاره کرده است. در گروه برای جلوگیری از برخوردن به چالش‌های ذکر شده ایده‌هایی مطرح و بعضاً پیاده‌سازی شد. برای مثال در مقاله بیان شده است که تولید ادعاهای منفی پیچیده برای افراد کار سخت‌تری بوده است. با توجه به این مشکل، تصمیم گرفته شد، برای هر لغت در فرهنگ لغت کلمات متضاد و موافق به فردی که می‌خواهد ادعا را تولید کند نمایش داده شود. در این صورت فرد علاوه بر این که می‌تواند از فرهنگ لغت قوی‌تری استفاده کند برای تولید ادعای منفی نیز با چالش کمتری رو به رو خواهد بود.

پس از این که برچسب ادعاها و شواهد موجود جمع‌آوری شد، برای سنجش صحت اطلاعات جمع‌آوری و تولید شده از چند روش مختلف بهره گرفته شده است. ۱. هر ادعا توسط ۵ فرد برچسب زده شده است و نتیجه‌ی نهایی حاصل توافق این افراد می‌تواند باشد. ۲. استفاده از فرد برچسب زن با مهارت بیشتر برای بررسی درصدی از نمونه‌ها که به صورت کاملاً تصادفی انتخاب شده اند. ۳. در قدم آخر، جهت بررسی صحت اطلاعات (به صورتی تصادفی)، تعدادی از نمونه‌ها را انتخاب کرده و دوباره برچسب زده‌اند. ادعاهایی که به درستی تولید نشده بودند و یا برچسب‌زنان به توافق نهایی نرسیده بودند، حذف گردیدند. نهایتاً پس از این ارزیابی دوباره، ۹۱ درصد از نمونه‌ها باقی ماندند.

در قدم بعدی با استفاده از روش TF-IDF [۸] و یا DrQA [۱۳] به تعداد k تا از مشابه‌ترین اسناد ویکی‌پدیا به ادعا را محاسبه می‌کند. براساس کل متن قسمت چکیده‌ی موجود در اسناد انتخاب شده از ویکی‌پدیا آموزش شبکه را انجام می‌دهد. نهایتاً با انتخاب تعداد L جمله از اسنادی که بیشتر از سایرین به ادعا شبیه هستند شبکه را آموزش می‌دهد. روش دوم، هدف نهایی مقاله است. در این روش سیستم می‌تواند از اسناد ویکی‌پدیا و با ذکر دقیق جملات به عنوان سند، میزان صحت ادعا را مشخص سازد.

جهت دسته بندی شبکه‌ی عصبی از دو مدل شبکه‌ی عصبی MLP^1 و Decomposable Attention

[۱۴] استفاده شده است. پس از آموزش هر دو مدل و مقایسه روش‌های آزمایش شده، بیشترین دقت هنگام استخراج جملات مربوط به مدل DA^2 با میزان ۳۲.۵ درصد دقت بوده است.

^۱ Multilayer perceptron

^۲ Decomposable Attention

۲-۳-۵ برنامه نویسی

علاوه بر موارد ذکر شده در قسمت آزمایش‌ها، برنامه‌هایی جهت راه اندازی ابزار FEVER [۴] در زبان فارسی توسعه داده‌ام و همچنین وب سایت جهت نمایش خروجی و عملکرد مدل‌های آموزش دیده طراحی و پیاده‌سازی شده است.

۲-۳-۶ مدیریتی

در ابتدای کار تمام نقشه‌ی راه کاملاً برای همه‌ی اعضای گروه واضح نبود و پس از گذشت چند جلسه و مشورت اعضا تصمیم بر این شد که نقشه‌ی مشخص‌تری در گروه و در دسترس همه‌ی اعضا باشد. و این تصمیم نظم و انگیزه‌ی بیشتری در افراد گروه ایجاد کرد. نهایتاً جلسه‌های دو روز در میان، به صورت منظم از سر گرفته شد و اعضای هر گروه گزارش فعالیت‌ها و نتیجه‌ی به دست آمده را اعلام می‌کردند و همچنین به صورت مکتوب نیز گزارش فعالیت‌های انجام شده و در دست اجرا در اختیار اعضای گروه قرار می‌گرفت. برای انجام کارهای در مقیاس بزرگتر زمان دو تا سه هفته‌ای در نظر گرفته شد و این کارها با همفکری اعضای گروه به وظایف کوچکتر شکسته و در پایان زمان مشخص شده کار محول شده تحویل گرفته می‌شد.

۲-۳-۷ نتیجه گیری

حاصل فعالیت‌ها و کارهای انجام شده نهایتاً با استفاده از مجموعه اطلاعات جمع آوری شده توسط گروه، ابزار تشخیص راستی سنجی متن با توجه به اسناد موجود در سایت ویکی‌پدیا به زبان فارسی راه اندازی شد و همچنین در کنار آن توسعه‌هایی جهت بهبود عملکرد ابزار تشخیص اخبار جعلی صورت گرفت.

فصل ۳

نتیجه‌گیری و پیشنهادها

۳-۱ مقدمه

محیط‌های آکادمیک مانند دانشگاه، به پرورش علمی دانشجویان و آشناسازی آنان با مباحث روز می‌پردازند. اما نداشتن یک رزومه خوب و تجربه عملی کافی، مشکلی است که بسیاری از دانش‌آموختگان پس از فارغ‌التحصیلی با آن مواجه می‌شوند، چراکه برای شروع موفقیت‌آمیز یک موقعیت شغلی، علاوه بر دانش به روز و کافی، فرد نیازمند مهارت‌های ایجاد ارتباط موثر با افراد و داشتن روحیه مسئولیت‌پذیری می‌باشد. کارآموزی این فرصت ویژه را برای دانشجویان فراهم می‌کند تا بتوانند دانشی که آموخته‌اند را با محیط کار تطبیق دهند و مهارت‌های لازم برای شروع کار را به دست بآورند. کارآموزی با توجه به علاقه و بازار کاری که دانشجو خواهان آن است، می‌تواند در زمینه تحقیقاتی و یا میدانی و صنعتی باشد. مهمترین مسئله آن است که زمینه و محیطی که دانشجو برای انجام کارآموزی انتخاب می‌کند، در زمینه‌ی کاری و مورد علاقه‌ی دانشجو باشد تا بتواند تجربیات و دانش مفیدی در زمینه‌ی مطلوب خود را به دست بیاورد. همچنین توانایی ایده‌پردازی و حل مسایل را بهتر بیاموزد و در صورت تمایل شرکت و دانشجو بعد از اتمام کارآموزی در آن شرکت به صورت حرفه‌ای شروع به کار کند.

اینجانب با توجه به علاقه‌ای که به انجام فعالیت‌های تحقیقاتی به ویژه در زمینه‌ی هوش مصنوعی دارم، برای واحد کارآموزی آزمایشگاه پردازش زبان‌های طبیعی دانشگاه علم و صنعت را انتخاب کردم. دست‌آورد کارآموزی برای اینجانب، کسب دانش در زمینه پردازش زبان طبیعی با جزئیات بیشتر و انجام کار به صورت جدی بوده است.

۳-۱-۱ خلاصه فعالیت‌های انجام شده

ابتدای کارآموزی به یادگیری و آشنایی با مفاهیم دانش مربوطه و همچنین آشنایی با فعالیت‌های پیشین انجام شده در زمینه‌ی پروژه در سطح آزمایشگاه و در سطح بین‌الملل پرداخته شد. همزمان با یادگیری انجام وظیفه‌های کوچک جهت یادگیری بهتر مطالب صورت گرفت.

در قدم بعدی با توجه به تحقیقات و مطالعات صورت گرفته جلساتی تشکیل شد تا ایده‌های موجود بررسی شوند و راهکارهای ممکن برای بهبود نتایج تحقیقات پیشین ارائه گردید. با پیاده‌سازی ایده‌های نتیجه‌گیری شده در جلسات و همچنین انجام مطالعات در حین این راه، دقت بهتری از مدل‌های شبکه عصبی موجود گرفته شد.

با توجه به مضمون پروژه یکی از مقالات موفق پیاده‌سازی شده که به حوزه پروژه نزدیک بود مورد مطالعه‌ی دقیق و بررسی قرار گرفت و در قدم بعدی اقدامات لازم برای تبدیل آن به زبان فارسی انجام شد. همچنین نیز وب‌سایتی جهت نمایش خروجی و نحوه‌ی کارکرد سیستم طراحی و پیاده‌سازی گردید. در آخرین مرحله پژوهش‌های انجام گرفته شده توسط گروه به صورت مقاله نوشته شد.

۲-۳ اعلام پیشنهادهایی برای رفع چالش‌های حوزه/واحد کارآموزی

زمان و تلاش قابل توجه برای اتمام کارآموزی نیاز است، در صورتی که این امکان باشد کارآموزی در زمان تعطیلات تابستان انجام شود، بازدهی بهتری خواهد داشت. به دلیل محدودیت‌های پیش آمده به دلیل شیوع گسترده ویروس کرونا، گذراندن دوره کارآموزی به صورت کامل در تابستان امکان پذیر نبوده و در نتیجه تداخل واحدهای درسی و همچنین انجام مسئولیت‌های محول شده توسط سرپرست کارآموزی بار کاری زیادی را متحمل دانشجو می‌کند.

در صورتی که کارآموزی به صورت حضوری برگزار شود، دانشجو می‌تواند از تجربیات افراد بهره بیشتری ببرد و همچنین بیشتر با محیط کار آشنا خواهد شد. درمیان گذاشتن سوال‌ها و مشکلاتی که کارآموز با آن‌ها مواجه شده است چاره‌اندیشی جهت رفع آن‌ها در لحظه از مزایا کارآموزی حضوری می‌باشد. با توجه به شرایط و دستورالعمل‌های قرنطینه دوران کرونا برقراری ارتباط با سرپرست کارآموزی کار سخت‌تری است. انجام پروژه‌های در زمینه‌ی هوش مصنوعی نیازمند سخت افزارهای مناسب این کار است و نیاز به سخت افزار قدرتمندتری به نسبت کامپیوتر شخصی دانشجو می‌باشد. در صورت نبود سخت‌افزار مناسب، مدت زمان اجرای برنامه‌ها خیلی طولانی شده و یا حتی در واقعی امکان پذیر نمی‌باشد و نمی‌توان شبکه مورد نظر را آموزش داد. در صورتی که واحد کارآموزی بتواند سخت افزار مورد نیاز انجام پروژه را در اختیار دانشجو قرار دهد عملکرد و خروجی بهتری در زمان کوتاه‌تری می‌توان گرفت.

فصل ٤

مراجع

- [١] S. Ken and B. Mike, *Agile Software Development with Scrum* Prentice Hall PTR, .٢٠٠١
- [٢] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," in ٢٠١٧ *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, .٢٠١٧
- [٣] H. Haibo, B. Yang, E. A. Garcia, and L. Shuao, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in ٢٠٠٨ *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, .٢٠٠٨
- [٤] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a Large-scale Dataset for Fact Extraction and VERification," New Orleans, Louisiana, .٢٠١٨
- [٥] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova", "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Association for Computational Linguistics*, ٢٠١٩: arxiv .
- [٦] Z. Majid *et al.*, "Persian Stance Classification Data Set," presented at the Conference for Truth and Trust Online ٢٠١٩, .٢٠١٩
- [٧] W. Pu, N. Liu, S. Yan, J. Yan, K. Xie, and Z. Chen, "Local Word Bag Model for Text Categorization," in *Seventh IEEE International Conference on Data Mining (ICDM ٢٠٠٧)*, .٢٠٠٧
- [٨] C. Sammut and G. I. Webb, "TF-IDF," in *Encyclopedia of Machine Learning*: Springer US, .٢٠١٠
- [٩] G. I. Webb, "Naïve Bayes," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb Eds.: Springer US, .٢٠١٠
- [١٠] A. H. Marti, "Support Vector Machines," *IEEE Intelligent Systems*, .١٩٩٨
- [١١] T. Hastie, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. springer, .٢٠٠٩
- [١٢] B. Leo, "Random Forests," *Mach. Learn.*, .٢٠٠١
- [١٣] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions," ٢٠١٧: Association for Computational Linguistics .
- [١٤] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A Decomposable Attention Model for Natural Language Inference," *ACL*, .٢٠١٦