مقاله پژوهشی

ارائه‌ی مدل عمیق CNN-BiLSTM برای شناسایی کارآموز

Doi: 10.30508/kdip.2024.429604.1081

مهدی ابراهیمیان دهکردی[۱] | شهلا نعمتی (نویسنده مسئول)[۲] | محمد احسان بصیری[۳]

۱-کارشناسی ارشد، گروه مهندسی کامپیوتر، دانشگاه شهرکرد، شهرکرد، ایران
۲-استادیار گروه مهندسی کامپیوتر، دانشگاه شهرکرد، شهرکرد، ایران
۳-دانشیار گروه مهندسی کامپیوتر، دانشگاه شهرکرد، شهرکرد، ایران

تاریخ دریافت: ۱۴۰۲/۰۹/۱۸
تاریخ پذیرش: ۱۴۰۲/۱۲/۰۶
صفحه: ۰۰ - ۰۰

چکیده

با افزایش تمایل شرکت‌ها و سازمان‌ها، برای بکارگیری کارآموزان در موقعیت‌های مختلف، انتخاب فرد مناسب برای مشارکت در دوره‌های کارآموزی اهمیت بسیاری پیدا کرده است. کسی که برای کارآموزی انتخاب می‌شود، اگرچه باید در زمینه‌های کاری مورد نظر، دانش و مهارت نسبی داشته باشد؛ اما لازم نیست، متخصص و با تجربه باشد؛ زیرا این‌گونه افراد معمولاً دستمزد بالایی طلب می‌کنند. وب‌سایت‌های پرس‌وجوی انجمنی با کاربران فراوانی که دارند، می‌توانند به‌عنوان یکی از منابع شناخت کارآموز مورداستفاده قرار گیرند. در پژوهش‌های پیشین برای شناخت کارآموزان بالقوه ویژگی‌های آماری مانند تعداد پاسخ، تعداد حوزه‌های تخصصی، طول پاسخ‌ها و موارد مشابه پیشنهاد شده است؛ اما محتوای پاسخ‌های کاربر تاکنون برای شناخت کارآموزان استفاده نشده است. این محتوای متنی منبعی غنی برای تشخیص گستردگی یا عمق دانش کاربر است و می‌تواند کمک شایانی به شناخت کارآموزان بالقوه کند. در این پژوهش یک مدل یادگیری عمیق با نام CNN-BiLSTM برای تشخیص افراد مناسب برای کارآموزی براساس متن پاسخ‌هایی که در وب‌سایت‌های پرس‌وجوی انجمنی ارسال می‌کنند، پیشنهاد شده است. علاوه بر این، از سه مدل یادگیری ماشین و چهار مدل پرکاربرد یادگیری عمیق نیز برای مقایسه استفاده شده است. بر اساس نتایج به‌دست‌آمده مدل‌های یادگیری عمیق در مقایسه با الگوریتم‌های یادگیری ماشین عملکرد بهتری داشته‌اند. همچنین در بین مدل‌های یادگیری عمیق، مدل پیشنهادی توانسته دقت بهتری نسبت به سایر مدل‌های مورداستفاده برای شناسایی کارآموزان بالقوه نشان دهد.

کلمات کلیدی: بازیابی کارآموز، یادگیری عمیق، وب‌سایت‌های پرس‌وجو، طبقه‌بندی متن

۱- مقدمه

در سال‌های اخیر با توسعه‌ی فناوری‌های نوظهور، شرکت‌ها تمایل بیشتری برای استخدام کارآموزان و استفاده از روش‌های آموزش حین کار برای آماده‌سازی آن‌ها در موقعیت‌های حرفه‌ای پیدا کرده‌اند. این موضوع به‌قدری اهمیت دارد که برخی از شرکت‌های بزرگ، اقدام به راه‌اندازی دانشگاه‌ها و مراکز کارآموزی جدیدی کرده‌اند (رستمی و نشاطی[1]، ۲۰۲۱). کارآموزی در منابع مختلف به شیوه‌های گوناگونی تعریف شده، اما کلی‌ترین تعریف عبارت است از به‌کارگیری مشروط افراد به‌صورت نیمه وقت یا تمام وقت برای مدتی محدود با تمرکز بر یادگیری مهارت‌های مشخص (مارتز، استابل، و مارکس[2]، ۲۰۱۴). هر شرکت، برای انتخاب کارآموز ممکن است معیارهای متفاوتی را در نظر بگیرد. البته به طورکلی یک کارآموز مناسب باید حداقل دانش کلی مورد نیاز شرکت را دارا باشد و بتواند زمینه‌های کاری مربوط را پوشش دهد. همچنین دوره‌ی کارآموزی نباید برای شرکت هزینه‌ی زیادی ایجاد کند؛ زیرا ممکن است کارآموز در نهایت استخدام نشود و شرکت را ترک کند؛ بنابراین کارآموز نباید از بین افراد خبره و باتجربه انتخاب شود؛ زیرا معمولاً این افراد خبره دستمزد نسبتاً بالاتری دریافت می‌کنند (رستمی و نشاطی، ۲۰۲۱)

با گسترش فناوری‌های بازیابی اطلاعات و کاربردهای گوناگون آن و همچنین حجم عظیم داده‌های ذخیره شده در منابع قابل‌دسترسی، شناسایی کارآموزان می‌تواند از طریق بازیابی اطلاعات موجود از افراد در بستر اینترنت انجام شود. یکی از این منابع، جوامع آنلاین پرس‌وجو مانند stack overflow هستند. این وب‌سایت‌ها به دلیل گستردگی و قابلیت جذب مشارکت بالا به منبع مهمی از اطلاعات تبدیل شده‌اند (ناریانان، اولک و فوکامی[3]، ۲۰۱۰). بازیابی افراد در این وب‌سایت‌ها، بر اساس تعاملات کاربران و آثاری که از خود به‌جا می‌گذارند (مانند پرسش‌هایی که می‌پرسند، پاسخ‌هایی که به پرسش‌های دیگران ارائه می‌دهند، نظراتی که ثبت می‌کنند و...) انجام می‌شود و براین‌اساس می‌توان تشخیص داد که هر یک از کاربران در چه زمینه‌هایی بیشتر فعالیت می‌کنند و میزان دانش و مهارت آنها در هر یک از این زمینه‌ها تا چه اندازه است؛ به این معنا که دانش آنها به‌صورت کلی و متوسط است یا در آن زمینه متخصص و خبره هستند (درگاهی نوبری، ستوده و نشاطی[4]، ۲۰۱۷).

همان‌گونه که در شکل شماره (۱) نشان داده شده است، کاربران موجود در جوامع آنلاین پرس‌وجو را می‌توان براساس گستردگی دانش آنها به دسته‌های زیر تقسیم کرد (دونوفریرو، سانچز، و اسپورر[5]، ۲۰۱۰؛ رستمی و نشاطی، ۲۰۲۱؛ کومار و پدانکار[6]، ۲۰۱۶).

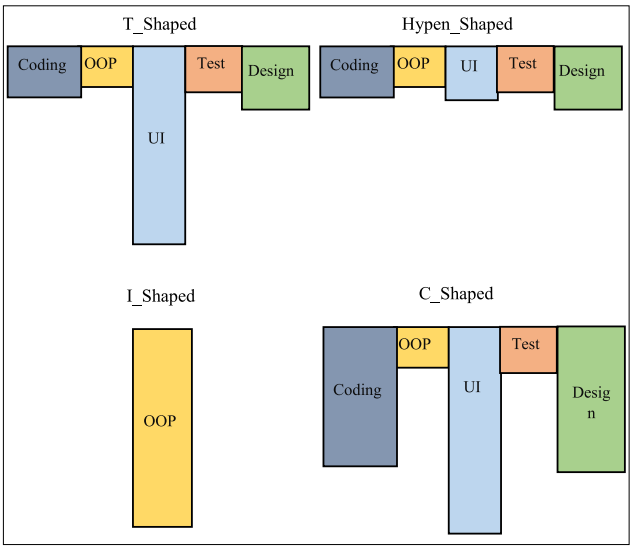
- I_shaped: فقط در یک زمینه‌ی مهارتی تخصص دارند؛
- T_shaped: در یک زمینه، دانش تخصصی و در یک یا چند زمینه، دانش متوسط دارند؛
- C_shaped: در چند زمینه، دانش تخصصی دارند؛
- Hiphen_shaped (خط پیوند): که در هیچ‌یک از زمینه‌های

1- Rostami & Neshati
2- Maertz, Stoeberl, & Marks
3- Narayanan, Olk, & Fukami
4- Dargahi Nobari, Sotudeh, & Neshati
5- Donofrio, Sanchez, & Spohrer
6- Kumar, & Pedanekar

مهارتی، متخصص و خبره نیستند اما در بعضی از زمینه‌ها دانش کلی و متوسط دارند؛

- No_shaped: در هیچ‌یک از زمینه‌های مورد نظر مهارت ندارند.

افراد نوع خط پیوند می‌توانند گزینه‌ی مناسبی برای کارآموزی باشند و با مشارکت در دوره‌های کارآموزی مهارت خود را در زمینه‌ی مورد نیاز افزایش دهند.



شکل (۱): انواع کاربران موجود در جوامع آنلاین پرس‌وجو براساس گستردگی دانش

در این تحقیق برای اولین بار یکی از مهم‌ترین بخش‌های مسأله‌ی بازیابی کارآموز یعنی شناسایی کاربران نوع خط پیوند معرفی شده است. در روش پیشنهادی، از متن تولید شده توسط کاربران در وب‌سایت‌های پرس‌وجوی انجمنی برای شناسایی کاربران نوع خط پیوند استفاده نمودیم. برای این منظور، مدلی مبتنی بر یادگیری عمیق پیشنهاد می‌شود. آموزش سرتاسری و یادگیری بازنمایی، یادگیری عمیق را از رویکردهای یادگیری ماشین سنتی متمایز می‌کند و آن را به ابزاری قدرتمند برای پردازش زبان طبیعی تبدیل می‌کند (لیکان[1]، ۲۰۱۸). در روش پیشنهادی از شبکه کانولوشنی برای استخراج ویژگی‌های مناسب از متن و از شبکه LSTM برای استخراج ارتباط معنایی بین کلمات در متن کاربر استفاده شده است. در ادامه به نوآوری‌های تحقیق حاضر اشاره شده است:

- معرفی مسأله شناسایی کاربران نوع خط پیوند به عنوان گزینه‌های مناسب برای انتخاب شدن به عنوان کارآموز.
- استفاده از متن تولید شده توسط کاربران به جای روش‌های آماری مبتنی بر تعداد اسناد تولید شده کاربر در پژوهش‌های مربوط به بازیابی کارآموز (رستمی و نشاطی، ۲۰۲۱).
- استفاده از یادگیری عمیق به جای روش‌های آماری مبتنی بر آنتروپی در پژوهش‌های پیشین.
- مقایسه‌ی روش پیشنهادی با روش‌های مرسوم مبتنی بر یادگیری عمیق و یادگیری ماشین سنتی.

در ادامه بخش‌های بعدی مقاله به‌صورت زیر سازمان‌دهی شده‌اند: در بخش دوم مرور مختصری بر پژوهش‌های پیشین انجام شده در زمینه‌های مرتبط ارائه شده است. در بخش سوم مفاهیم اولیه‌ای که در این پژوهش به کار

1- Lecun

برده شده، بیان شده است. در بخش چهارم طرح مسأله شرح داده شده و بخش پنجم شامل روش پیشنهادی و جزئیات پیاده‌سازی است و درنهایت، بخش ششم شامل نتایج حاصل از پژوهش و پیشنهادهایی برای کارهای آینده است.

۲- مبانی نظری

تحقیقات مربوط به پیشینه‌ی این مقاله، را می‌توان در سه موضوع بازیابی افراد خبره، مسیریابی پرسش‌ها در وب‌سایت‌های پرس‌وجو و بازیابی کارآموز دسته‌بندی کرد.

بازیابی افراد خبره

در زمینه‌ی بازیابی افراد خبره (کسانی که در موضوع مورد نظر متخصص باشند)، ژائو و همکاران از روش مدل‌سازی زبانی استفاده کرده‌اند (ژائو، تانگ و دوو[1]، ۲۰۱۹). در این روش، مسأله‌ی بازیابی افراد تبدیل به یک مسأله‌ی بازیابی متن می‌شود و اسناد مربوط به افراد شامل مقالات، پرس‌وجوها در اینترنت، یادداشت‌ها در شبکه‌های اجتماعی و... با موضوع مورد نظر مطابقت داده می‌شوند. آنها همچنین در مقاله‌ی خود، روش جدیدی به نام شبکه عصبی کانولوشنی محدود را ارائه کردند که ایده‌ی اصلی آن تبدیل مسأله‌ی موجود، به مسأله‌ی تشخیص تصویر است. در این مدل با توجه به کلمات اسناد مربوط به فرد و کلمات پرس‌وجو، یک ماتریس مشابهت ساخته می‌شود. این ماتریس به‌عنوان یک تصویر در نظر گرفته شده و به ورودی یک شبکه عصبی کانولوشن داده می‌شود تا میزان شباهت پرس‌وجو با اسناد مربوط به فرد را پیش‌بینی کند.

یوان و همکاران، مدل تطبیق متن عمیق با نام Match-SRNN را برای مدل‌سازی اطلاعات تعامل بین متون، برای پیش‌بینی افراد خبره معرفی کردند (یوان، ژانگ، تانگ، هال و کابوتا[2]، ۲۰۲۰). این مدل شامل یک شبکه تنسور عصبی برای ثبت برهمکنش‌های سطح کلمه و یا حرف، یک شبکه عصبی بازگشتی فضایی که روی تنسور تعاملات حرف اعمال می‌شود تا تعاملات کلی را به‌صورت بازگشتی ثبت کند و یک تابع امتیازدهی خطی برای محاسبه‌ی امتیاز نهایی تطبیق، است.

التایی و همکاران به ارائه‌ی دسته‌بندی کلی از روش‌های قابل‌استفاده برای بازیابی افراد پرداخته‌اند (التایی، کادری، اوبوسا[3]، ۲۰۱۸). بر اساس این طبقه‌بندی، روش‌های بازیابی افراد خبره به‌طورکلی به دو دسته‌ی مبتنی بر یادگیری ماشین و مبتنی بر گراف تقسیم می‌شوند که هر یک شامل رویکردهای متنوعی هستند. لیانگ و ریک پنج مدل کلی بر اساس مجموعه‌ای از اسناد ناهمگن برای شناسایی گروه‌های متخصص ارائه می‌دهند (لیانگ و ریک[4]، ۲۰۱۶). دو مورد از این مدل‌ها امتیازهای تخصص کارشناسان در یک گروه را برای کار مورد نظر جمع‌آوری می‌کنند. یکی دیگر از مدل‌ها، اسناد مرتبط با متخصصان گروه را مشخص و سپس تعیین می‌کند که اسناد تا چه حد با موضوع، مرتبط هستند. درنهایت دو مدل باقی‌مانده مستقیماً تخمین می‌زنند که آیا آن گروه، یک گروه آگاه برای یک موضوع معین است یا خیر.

۳- مسیریابی پرسش‌ها

در زمینه‌ی مسیریابی پرسش‌ها در جوامع پرس‌وجو به‌منظور یافتن کاربرانی که می‌توانند به پرسش مورد نظر، پاسخ صحیح و مناسب بدهند نیز پژوهش‌های فراوانی انجام شده است. لی و شاه در مقاله‌ی خود، پاسخ‌دهندگان بالقوه را بر اساس محتوای پرسش کاربر و پروفایل آنها پیش‌بینی می‌کنند (لی و شاه[5]، ۲۰۱۸). آنها ابتدا پروفایل کاربران را بر اساس سابقه‌ی فعالیتشان ایجاد و سپس امتیاز بین پرسش‌های جدید و پروفایل‌های کاربر را محاسبه می‌کنند. امتیاز بالاتر نشان‌دهنده‌ی شانس بالاتر پاسخ‌گویی کاربر به آن پرسش است.

اعظم و همکاران، به بررسی روش‌های مسیریابی پرسش برای یافتن پاسخ مناسب در جوامع پرس‌وجو

1- Zhao, Tang, & Du
2- Yuan, Zhang, Tang, Hall, & Cabotà
3- Al-Taie, Kadry, & Obasa
4- Liang, & de Rijke
5- Le, & Shah

پرداخته‌اند (اعظم، تازی و هوسنی[1]، ۲۰۱۷). این روش‌ها بر اساس رویکردی که برای بازیابی اطلاعات استفاده می‌کنند به سه دسته‌ی اصلی تقسیم شده است: مدل فضای برداری، مدل‌سازی موضوع و مدل‌های مبتنی بر یادگیری عمیق. آنها در پژوهش خود روش مسیریابی پرسش به نام QR-DSSM را نیز پیشنهاد کرده‌اند که از ویژگی‌های متنی برای پیش‌بینی شباهت معنایی بین پرسش‌ها و نمایه‌های پاسخ‌دهندگان بالقوه در سایت‌های پرس‌وجو، با کمک شبکه‌های عصبی عمیق استفاده کرده‌اند.

یان و ژو، مسیریابی پرسش در جوامع پرس‌وجوی انجمنی[2] برای یافتن پاسخ مناسب را به‌عنوان یک مسأله‌ی رتبه‌بندی در نظر گرفته‌اند که پاسخ‌دهندگان بالقوه را بر اساس میزان احتمالی توانایی آنها در حل مسأله‌ی جدید، رتبه‌بندی کرده‌اند (ژائو، جاوید، جاکوب، و مک نایر[3]، ۲۰۱۵). در این روش پیشنهادی، از مدل تنسور و مدل موضوعی به‌طور هم‌زمان برای استخراج روابط معنایی نهفته بین پرسش‌کننده، پرسش و پاسخ‌دهنده استفاده شده است و یک روش یادگیری بر اساس مدل‌های فوق ساخته شده تا با بهینه‌سازی چندکلاسه، رتبه‌بندی بهینه‌ی پاسخ‌دهندگان، برای پرسش‌های جدید به دست آید.

ژیو و هانگ، استفاده از شبکه عصبی تنسور کانولوشنی را برای بازیابی پرسش‌های مشابه در سایت‌های پرس‌وجوی اجتماعی پیشنهاد کرده‌اند (ژیو و هانگ[4]، ۲۰۱۵). از این شبکه عصبی برای رمزگذاری جملات در فضای معنایی و مدل‌سازی روابط آنها در یک‌لایه‌ی تنسور استفاده شده است. این مدل، تطبیق معنایی و مدل‌سازی جملات را در یک مدل واحد ادغام کرده که نه‌تنها می‌تواند اطلاعات مفید را با لایه‌های کانولوشن و پولینگ[5] جمع‌آوری کند، بلکه معیارهای تطبیق بین پرسش و پاسخ آن را نیز یاد می‌گیرد.

بازیابی کارآموز

در تعدادی از تحقیقات نیز، به موضوع یافتن کارآموز پرداخته شده است. از جمله رستمی و نشاطی در مقاله‌ی خود برای بازیابی کارآموزان، ابتدا از مدل‌های دانش تخصصی[6] و دانش متوسط[7] استفاده کرده‌اند که پیش از آن برای بازیابی افراد خبره استفاده می‌شود؛ سپس دو مدل جدید با نام مدل مبتنی بر آنتروپی[8] و مدل دانش حداکثری[9] را بر اساس احتمال شرطی پیشنهاد داده‌اند که نتایج بسیار بهتری را نسبت به مدل‌های موجود برای بازیابی کارآموز به دست آورده‌اند (رستمی و نشاطی، ۲۰۲۱).

ام بایا و همکاران، سیستم توصیه‌گر مبتنی بر موجودیت شناختی را ارائه داده‌اند که می‌تواند در فرایند اختصاص موقعیت کارآموزی به دانشجویان دانشگاه کمک کند (ام بایا، لاوال، موعلا، اوزاروت، و بوراس[10]، ۲۰۱۶). هدف این روش، ایجاد یک هماهنگی بین نیازهای شرکت‌ها و توانایی‌های دانشجویان بوده است. این امر با ایجاد یک مدل موجودیت شناختی از مشخصات فراگیران آموزشی و موقعیت‌های کارآموزی که در یک متن آزاد و بدون استفاده از واژگان کنترل شده نوشته شده بود، انجام شد که درنهایت منجر به ایجاد یک سیستم توصیه‌گر معنایی برای دانشگاه‌ها شد تا در تصمیم‌گیری برای موقعیت‌های کارآموزی به آنها کمک کند.

ژائو و همکاران، به ارائه‌ی روشی خودکار برای شناسایی مهارت‌های افراد بر اساس سابقه‌ی کاری و بهینه‌سازی آن پرداخته‌اند (ژائو و همکاران، ۲۰۱۵). سیستم پیشنهادی آنها شامل دو رویکرد ایجاد طبقه‌بندی مهارت‌ها بر اساس طبقه‌بندی ویکی‌پدیا و برچسب‌گذاری مهارت‌ها است. در جدول شماره (۱)، مروری بر پژوهش‌های پیشین آورده شده است.

1- Azzam, Tazi, & Hossny
2- Community question answering (CQA)
3- Zhao, Javed, Jacob, & McNair
4- Qiu, & Huang
5- pooling
6- Expertise knowledge model (EKM)
7- Intermediate knowledge model (IKM)
8- Entropy based model (EBM)
9- Maximum knowledge model (MKM)
10- M'Baya, Laval, Moalla, Ouzrout, & Bouras

جدول (۱): خلاصه تحقیق‌های پیشین

ردیف	محققین	حوزه	روش پیشنهادی
۱	التایی و همکاران (۲۰۱۸)	بازیابی افراد خبره	دسته‌بندی و معرفی روش‌های مربوط به بازیابی افراد خبره
۲	ژائو و همکاران (۲۰۱۹)	بازیابی افراد خبره	استفاده از شبکه‌های عصبی کانولوشنی برای بازیابی افراد خبره
۳	یوان و همکاران (۲۰۲۰)	بازیابی افراد خبره	استفاده از مدل match-SRNN برای پیش‌بینی افراد خبره
۴	ژائو و همکاران (۲۰۱۵)	مسیریابی سؤال	مسیریابی پرسش به‌عنوان مسأله‌ی رتبه‌بندی و یافتن پاسخ‌دهندگان بالقوه
۵	ژیو و هانگ (۲۰۱۵)	مسیریابی سؤال	استفاده از شبکه عصبی تنسور کانولوشنی برای بازیابی پرسش‌های مشابه
۶	اعظم و همکاران (۲۰۱۷)	مسیریابی سؤال	ارائه روش QR-DSSM برای پیش‌بینی شباهت معنایی بین پرسش‌ها و نمایه‌های پاسخ‌دهندگان بالقوه
۷	لی و شاه (۲۰۱۸)	مسیریابی سؤال	شناسایی پاسخ‌دهندگان بالقوه بر اساس محتوای پرسش و پروفایل کاربر
۸	ژائو و همکاران (۲۰۱۵)	بازیابی کارآموز	ارائه‌ی یک روش خودکار برای شناسایی مهارت‌های افراد بر اساس سابقه کاری و بهینه‌سازی آن
۹	ام بابا و همکاران (۲۰۱۶)	بازیابی کارآموز	ارائه‌ی سیستم توصیه‌گر مبتنی بر موجودیت شناختی برای اختصاص کارآموزی به دانشجویان دانشگاه
۱۰	رستمی و نشاطی (۲۰۲۱)	بازیابی کارآموز	بررسی مدل‌های مورداستفاده برای بازیابی کارآموز و ارائه‌ی دو مدل جدید

شکاف تحقیقی

همان‌گونه که بررسی شد، اکثر پژوهش‌های پیشین فقط بر یافتن متخصصان متمرکز است، اما در این مقاله، هدف ما یافتن کاربران نوع خط پیوند به عنوان متخصصین عمومی مناسب برای یک موقعیت کارآموزی است. تنها به تأثیر شکل دانش کاربر بر مسأله بازآموزی کارآموز پرداخته شده است (رستمی و نشاطی، ۲۰۲۱). تفاوت‌های این تحقیق و تحقیق حاضر به شرح زیر است:

- مدلی آماری مبتنی بر آنتروپی و دانش حداکثری را بر اساس احتمال شرطی پیشنهاد داده‌اند که تنها از تعداد اسناد مرتبط کاربر با حوزه موضوعی مورد نیاز کارآموزی استفاده می‌کند. درحالی‌که در پژوهش حاضر از محتوای پاسخ‌ها استفاده می‌شود.
- مستقیماً به یافتن کاربران با شکل دانش خط پیوند پرداخته نشده درحالی‌که در پژوهش حاضر این مسأله به‌صورت یک مسأله دسته‌بندی تعریف و حل می‌شود؛
- صرفاً از دو روش آماری برای بازیابی کارآموز استفاده شده درحالی‌که در این پژوهش با توجه به در اختیار داشتن حجم مناسب داده‌های آموزشی، استفاده از یادگیری عمیق پیشنهاد شده است.

طرح مسأله

برای هر مأموریت کارآموزی، باید افرادی انتخاب شوند که در زمینه‌های مهارتی مورد نظر، دانش و مهارت سطح متوسط داشته باشند. همان‌گونه که اشاره شد، این افراد از نظر خبرگی به شکل خط پیوند هستند. یکی از منابع مهم برای دسترسی به چنین افرادی، وب‌سایت‌های پرس‌وجو هستند. در این وب‌سایت‌ها می‌توان با تحلیل محتوای پرس‌وجوهای انجام شده در موضوع‌های مورد نظر، کاربران به شکل خط پیوند که برای موقعیت کارآموزی مناسب هستند را یافت. در پژوهش حاضر این مسأله را یک مسأله‌ی طبقه‌بندی متن دو کلاسه می‌بینیم که در آن برچسب‌گذاری داده‌ها براساس شکل خبرگی کاربران به‌صورت رابطه‌ی شماره (۱) انجام می‌شود:

$$class(u_i) = \begin{cases} 1 & shape(u_i) \in \{I, T, C\} \\ 0 & otherwise \end{cases}$$

رابطه‌ی (۱): تبدیل شکل خبرگی به برچسب کلاس

که در آن، ui کاربر اُم، shape(ui) شکل خبرگی بر اساس دسته‌بندی نشان داده شده در شکل ۱ و class نشان‌دهنده برچسب کلاس کاربر است. لازم به ذکر است که برچسب‌گذاری اولیه مجموعه داده به‌صورت دستی توسط قره‌باغی، رستمی ونشاطی[1] (۲۰۱۸) انجام شده و در (رستمی و نشاطی، ۲۰۲۱) استفاده شده است.

در پژوهش حاضر مجموعه‌ی تمام متون مربوط به پاسخ‌های کاربران به شکل رابطه‌ی ۲ است:

$$D = \bigcup_{i \in U} D_i$$

رابطه‌ی (۲): مجموعه اسناد مربوط به کاربران

که در این رابطه $D_i = \{d_{i1}, d_{i2}, \dots, d_{in}\}$ نشان‌دهنده مجموعه پاسخ‌های کاربر اُم می‌باشد.

مجموعه داده

برای اجرا و ارزیابی مدل‌ها در این پژوهش، از مجموعه داده‌ای استفاده شده است که توسط قره‌باغی و همکاران (۲۰۱۸) معرفی و رستمی و نشاطی (۲۰۲۱) نیز از آن برای بازیابی کارآموز استفاده کرده‌اند. این مجموعه داده بر اساس پرس‌وجوهایی ساخته شده است که از اوت ۲۰۰۸ تا مارس ۲۰۱۵ در وب‌سایت stack overflow انجام و دارای سه زیرمجموعه‌ی Java، C# و Android است. هر یک از این زیرمجموعه‌ها شامل پرسش‌های دارای برچسب[2] مربوط به آن موضوع و پاسخ‌های ارائه شده برای هر پرسش هستند. برای هر مجموعه، تعدادی از زمینه‌های مهارتی اصلی از ۲۰۰ برچسب پرتکرار آن مجموعه، استخراج و سطح دانش کاربر در هر حوزه‌ی مهارتی با توجه به تعداد پاسخ‌های پذیرفته شده‌ی کاربر در آن حوزه تعیین شده است. سپس در هر مجموعه، کاربران با توجه به سطح دانش آنها در هر حوزه‌ی مهارتی دسته‌بندی می‌شوند. با توجه به این‌که در میان دسته‌بندی موجود، تنها افرادی که در گروه خط پیوند قرار می‌گیرند برای کارآموزی مناسب هستند، می‌توان آنها را بر اساس عضو بودن یا نبودن در این دسته، برچسب‌گذاری کرد (رابطه ۱). با توجه به اینکه، متن پرسش‌ها و پاسخ‌های کاربران در این مجموعه داده موجود نیست، این متن‌ها نیز از پایگاه داده‌ی وب‌سایت، بر اساس شناسه‌ی کاربر و شناسه‌ی پرسش و پاسخ، استخراج و به مجموعه داده افزوده شدند. مشخصات مجموعه داده در جدول شماره (۲) آمده است.

جدول (۲): مشخصات مجموعه داده‌ی استفاده شده

نام مجموعه	تعداد داده‌های آزمایش	تعداد داده‌های ارزیابی	تعداد داده‌های آموزشی
Android	۱۲۰۰۰	۱۰۰۰۰	۳۵۰۰۰
Java	۱۵۰۰۰	۱۱۰۰۰	۵۵۰۰۰
#C	۱۵۰۰۰	۱۲۰۰۰	۵۵۰۰۰

1- Gharebagh, Rostami, & Neshati
2- Tag

همچنین نمونه‌ای از متون کاربران در جدول شماره (۳) آورده شده است.

جدول (۳): مثال‌هایی از نوشته‌های کاربران. کلاس خبرگی ۱ کاربر نوع خط پیوند و ۰ کاربر غیر خط پیوند را نشان می‌دهد

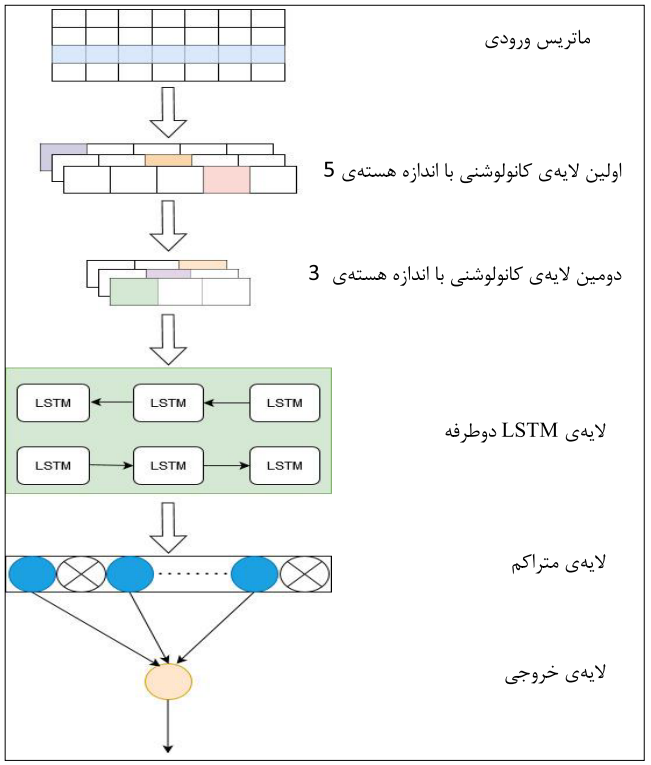
نام مجموعه	کلاس خبرگی	متن
Java	۱	Ask your ISP why this is, and if you don't get an answer, switch ISPs again
		I have no experience with electrical connections, but you can download the iPhone developer kit
		Maybe you can use easymock to isolate the classes you want to test, and have the mock objects receive the events and check that they're fired
Java	۰	you need to create a compound join statement to produce the correct cursor. What you have there will always throw an error because you haven't specified children correctly
		So, after much digging, on PyCharm vs Eclipse, eclipse automatically designate sources in a project, while PyCharm requires you to select the appropriate folders as source folders vs template folders
		using the given structures above, it is possible, and was confirmed that you can solve this with a set of tasklets. It is a SIGNIFICANT speed up over the iterative method

روش پیشنهادی

در این تحقیق برای شناسایی کاربران نوع خط پیوند به عنوان کارآموزان مناسب، مدل پیشنهادی با نام CNN_BiLSTM معرفی می‌شود. در مدل پیشنهادی (شکل شماره ۲)، از لایه‌های زیر استفاده شده است:

- لایه‌ی تعبیه‌ساز: در این لایه ابعاد ورودی برابر با تعداد واژه‌های منحصربه‌فرد در مجموعه‌داده (۲۰۰۰ واژه) در نظر گرفته می‌شود و ابعاد خروجی آن مقدار ۱۰۰ است. همچنین باتوجه‌به این که در مرحله قبل طول همه دنباله‌ها یکسان و برابر با ۲۰۰ شد، طول دنباله‌های ورودی لایه تعبیه ساز نیز مقدار ۲۰۰ می‌گیرد.
- لایه‌ی حذف تصادفی[1]: این لایه برای افزایش قدرت تعمیم مدل در نظر گرفته شده و نرخ ۰/۱ برای آن تنظیم شده است.
- لایه‌های کانولوشن[2](CNN): دو لایه‌ی کانولوشنی یک‌بعدی هر یک با ۱۰ فیلتر به اندازه‌های ۳ و ۵ و تابع فعال‌سازی relu برای استخراج ویژگی و کاهش ابعاد داده‌ها به صورت پی‌در پی قرار می‌گیرند.
- لایه‌ی LSTM دو طرفه[3]: برای استخراج ارتباط معنایی بین ویژگی‌ها و زمینه[4] از این لایه با تعداد سلول ۶۴ استفاده شده است. این لایه با داشتن حافظه‌ی موقت و پردازش دوطرفه‌ی داده‌ها، می‌تواند عملکرد مناسبی در وظایف پردازش متن داشته باشد.
- لایه‌ی متراکم[5]: برای کاهش ابعاد فضای ویژگی نهایی از این لایه با اندازه‌ی ۱۰ و تابع فعال‌سازی relu استفاده شده است.
- لایه‌ی حذف تصادفی: مشابه لایه‌ی نرخ تصادفی اولیه در خروجی تعبیه‌ساز.
- لایه‌ی متراکم نهایی: برای انجام دسته‌بندی این لایه با یک سلول و تابع فعال‌سازی سیگموئید در نظر گرفته شده است.

1-Dropout
2- Convolutional neural network (CNN)
3- Bidirectional Long Short-Term Memory (Bi-LSTM)
4- Context
5- Dense



شکل (۲): ساختار مدل پیشنهادی

همچنین برای مقایسه و ارزیابی کارایی مدل پیشنهادی از چهار مدل پرکاربرد یادگیری عمیق شامل شبکه‌ی تماماً متصل[1]، LSTM دوطرفه، شبکه عصبی کانولوشنی[2] و مدل پیش‌آموخته‌ی برت[3] استفاده شده است. علاوه بر این مدل‌ها، از سه الگوریتم پرکاربرد یادگیری ماشین سنتی یعنی KNN، Naïve-Bayes و SVM نیز استفاده شد تا مقایسه‌ای نیز بین مدل‌های یادگیری عمیق و یادگیری ماشین سنتی صورت گیرد. ساختار این مدل‌ها با زبان برنامه‌نویسی پایتون و با استفاده از کتابخانه‌ی کراس[4] که از کتابخانه‌های پرکاربرد در زمینه‌ی یادگیری عمیق محسوب می‌شود، پیاده‌سازی شده‌اند. برای اجرای مدل‌ها، فایلی که هر ردیف آن شامل شناسه‌ی کاربر، برچسب و متن پاسخ‌ها است باید به‌عنوان ورودی به مدل داده شود اما قبل از آن باید پیش‌پردازش انجام گیرد. مرحله‌ی پیش‌پردازش، شامل گام‌های زیر است:

- حذف علامت‌های غیرالفبایی؛
- حذف واژگان تکراری؛
- تبدیل حروف بزرگ انگلیسی به حروف کوچک؛
- ریشه‌یابی لغات.

تمام مراحل پیش‌پردازش به‌وسیله‌ی کتابخانه‌ی nltk در پایتون انجام شدند. انجام مرحله‌ی پیش‌پردازش علاوه بر کاهش حجم فایل ورودی می‌تواند باعث افزایش دقت مدل‌ها نیز شود. در گام بعدی، واژگان متن‌ها از یکدیگر تفکیک و یک شاخص واژگان برای ۲۰۰۰ کلمه‌ی پرتکرار ساخته می‌شود. در این شاخص، واژه‌ها بر اساس تعداد تکرار به‌صورت نزولی مرتب شده و به هر کلمه یک عدد صحیح مثبت، نسبت داده می‌شود. با جایگزینی این اعداد به‌جای کلمات، متن‌ها تبدیل به دنباله‌هایی از اعداد صحیح خواهند شد. این دنباله‌ها برای ورود به مدل باید طول یکسانی داشته باشند. در این پژوهش، طول دنباله‌ها برابر با ۲۰۰ در نظر گرفته شده است. دنباله‌هایی که طول بیشتری داشته باشند، کوتاه خواهند شد و

1- Fully-connected
2- Convolutional neural network (CNN)
3- BERT
4- Keras

دنباله‌هایی که طول کمتری داشته باشند به ابتدای آن‌ها صفر افزوده می‌شود.

پیش از اجرای مدل‌ها، باید مشخص شود که چه تعدادی از داده‌ها برای آموزش و چه تعدادی برای آزمایش مورداستفاده قرار می‌گیرند. در این پژوهش، حدوداً ۱۵درصد از کل داده‌ها برای آزمایش و ۱۵ درصد از باقی‌مانده برای ارزیابی در هر مرحله از اجرای مدل، به‌صورت تصادفی انتخاب می‌شوند.

۴- یافته‌های تحقیق

برای ارزیابی نتایج حاصل از اجرای مدل‌ها از معیارهای صحت[1]، اتلاف[2] و معیار F1 استفاده شده است که در رابطه‌ی شماره (۳) آورده شده است.

$$precision = \frac{tp}{(tp + fp)}$$

$$recall = \frac{tp}{(tp + fn)}$$

$$F_1 = \frac{2 \times precision \times recall}{(precision + recall)}$$

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \times \log\left(p(y_i)\right) + (1 - y_i) \times \log\left(1 - p(y_i)\right)$$

$$Accuracy = \frac{tp+tn}{(tp+fp+tn+fn)}$$

رابطه‌ی (۳): روابط استفاده شد برای ارزیابی مد

در رابطه‌ی شماره (۳)، منظور از tp، tn، fp و fn به ترتیب مثبت درست، منفی درست، مثبت کاذب و منفی کاذب است. Loss نشان‌دهنده‌ی اتلاف دودویی است و در آن، p(yi) احتمال کلاس ۱ و ۱-p(yi) احتمال کلاس ۰ است. هنگامی‌که پیش‌بینی مدل متعلق به کلاس ۱ است، بخش اول فرمول فعال می شود و قسمت دوم ناپدید می شود و بالعکس.

تنظیم پارامترها

برای همه مدل‌های یادگیری عمیق مورداستفاده، مقدار max_words برابر با ۲۰۰۰ و مقدار max_len برابر با ۲۰۰ قرار داده شده است. در الگوریتم‌های یادگیری ماشین سنتی نیز، مقدار max_features برای بردارهای TF-IDF برابر ۲۰۰۰ در نظر گرفته شده است.

اولین مدل یادگیری عمیق مدل متراکم است که برای کامپایل کردن این مدل، از تابع بهینه‌ساز آدام[3] با نرخ یادگیری ۰/۰۰۰۱، تابع هزینه binary_crossentropy و معیار سنجش صحت استفاده می‌شود. در تابع fit نیز پس از مشخص‌کردن داده‌های آموزشی و داده‌های ارزیابی، تعداد دوره‌های اجرای مدل برابر با ۲۰ و اندازه دسته (batch_size) برابر با ۱۲۸ تعریف می‌شود. همچنین با استفاده از دستور callback، درصورتی‌که پس از ۵ دوره، میزان صحت برای داده‌های ارزیابی صعودی نباشد، اجرای سایر دوره‌ها متوقف می‌شود.

دومین مدل LSTM دوجهته است. این مدل نیز مانند مدل قبلی در ابتدا از لایه تعبیه‌سازی و حذف تصادفی بهره می‌برد با این تفاوت که ابعاد تعبیه سازی آن برابر با ۱۲۸ در نظر گرفته شده است. در ادامه دو لایه LSTM دوجهته با ۶۴ واحد و تابع فعال‌ساز پیش‌فرض tanh قرار می‌گیرند که دنباله‌های ورودی را در دو جهت (روبه‌جلو و رو به عقب) پردازش می‌کنند. در ادامه نیز یک لایه حذف تصادفی با نرخ ۰/۱ و لایه متراکم با یک نورون قرار می‌گیرند تا خروجی موردنظر را تولید کنند. برای کامپایل کردن این مدل از تابع بهینه‌ساز آدام با نرخ یادگیری ۰/۰۰۰۱، تابع هزینه binary_crossentropy و معیار

1- Accuracy
2- loss
3- Adam

سنجش صحت استفاده شده است. تعداد دوره‌ها برابر با ۲۰ و اندازه دسته ۲۵۶ در نظر گرفته شده است.

در پیاده‌سازی مدل شبکه عصبی کانولوشنی، مانند مدل‌های قبل ابتدا یک لایه تعبیه سازی و یک لایه حذف تصادفی قرار می‌گیرد. پس از آن یک لایه کانولوشنی یک‌بعدی، یک هسته کانولوشنی یک‌بعدی را با ۱۲۸ فیلتر، اندازه هسته ۵ و تابع فعال‌سازی relu روی داده‌ها اعمال می‌شود. باتوجه‌به طول دنباله ورودی (۲۰۰)، طول تنسور خروجی از این لایه برابر با ۱۹۶ خواهد بود. همچنین با وجود ۱۲۸ فیلتر در این لایه، قالب خروجی آن به شکل (,۱۲۸,۱۹۶None) است. این تنسور خروجی در مرحله بعد وارد یک لایه ادغام حداکثری یک‌بعدی (GlobalMaxPooling۱D) می‌شود. این لایه با درنظرگرفتن حداکثر مقدار در بعد زمان، نمایش ورودی را پایین می‌آورد. در ادامه‌ی این مدل، یک لایه‌ی متراکم با ۱۰ نورون و تابع فعال‌ساز relu، یک لایه حذف تصادفی با نرخ ۰/۱ و یک لایه متراکم با یک نورون و تابع فعال‌ساز سیگموئید داده‌ها را برای طبقه‌بندی در یکی از کلاس‌های دوگانه، پردازش می‌کنند. پارامترهای کامپایل کردن این مدل مشابه با مدل LSTM دوجهته است. تعداد دوره‌ها برای اجرای مدل ۱۵ و اندازه‌ی دسته برابر با ۶۴ تعریف شده است.

دیگر مدل مورداستفاده، مدل پیش آموخته‌ی برت است. انواع مختلفی از مدل‌های برت وجود دارند که بر اساس نوع کاربرد می‌توان از آنها استفاده کرد. در اینجا از مدل small BERT با ۴ لایه و اندازه‌ی لایه‌ی پنهان ۵۱۲ استفاده شده است. برای پیاده‌سازی، در اولین لایه که مربوط به داده‌های ورودی است، به آنها نام «text» اختصاص‌داده‌شده و نوع آنها برابر با «string» در نظر گرفته می‌شود. پس از آن داده‌ها وارد لایه پیش‌پردازش می‌شوند تا متن‌ها تبدیل به تنسورهایی از شناسه‌های عددی شوند. تنسور خروجی از این لایه، به لایه BERT فرستاده می‌شود. در گام بعدی یک لایه حذف تصادفی با نرخ ۰/۱ و یک لایه متراکم با یک نورون و تابع فعال‌ساز سیگموئید قرار می‌گیرند تا طبقه‌بندی نهایی را تولید کنند. پارامترهای مربوط به کامپایل کردن این مدل، مانند مدل قبل مقداردهی می‌شوند. تعداد دوره‌های اجرای این مدل ۵ و اندازه دسته آن ۵۱۲ در نظر گرفته شده است.

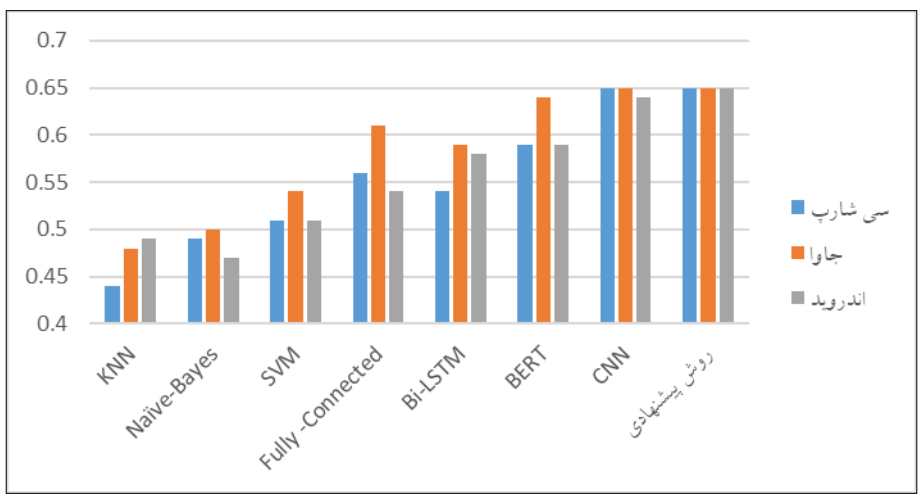
نتایج پیاده‌سازی براساس معیارهای صحت و اتلاف در جدول‌های شماره (۴) و (۶) و براساس معیار F1 در شکل شماره (۳) آورده شده است.

جدول (۴): مقایسه‌ی مدل‌ها برای مجموعه‌ی C#

نوع مدل	نام مدل	صحت	اتلاف
یادگیری ماشین	KNN	۰/۷۳	-
	Naïve-Bayes	۰/۷۴	-
	SVM	۰/۷۸	-
یادگیری عمیق	Fully -Connected	۰/۸۰	۰/۳۷
	Bi-LSTM	۰/۸۲	۰/۳۷
	BERT	۰/۸۲	۰/۴۴
	CNN	۰/۸۴	۰/۳۵
	روش پیشنهادی	۰/۸۳	۰/۳۶

جدول (۵): مقایسه‌ی مدل‌ها برای مجموعه‌ی Java.			
نوع مدل	نام مدل	صحت	اتلاف
یادگیری ماشین	KNN	۰٫۷۲	-
	Naïve-Bayes	۰٫۷۶	-
	SVM	۰٫۷۸	-
یادگیری عمیق	Fully -Connected	۰٫۸۱	۰٫۳۹
	Bi-LSTM	۰٫۸۱	۰٫۴۰
	BERT	۰٫۸۲	۰٫۴۲
	CNN	۰٫۸۲	۰٫۳۸
	روش پیشنهادی	۰٫۸۳	۰٫۳۶

جدول (۶): مقایسه‌ی مدل‌ها برای مجموعه‌ی Android			
نوع مدل	نام مدل	صحت	اتلاف
یادگیری ماشین	KNN	۰٫۷۲	-
	Naïve-Bayes	۰٫۷۴	-
	SVM	۰٫۷۷	-
یادگیری عمیق	Fully -Connected	۰٫۸۱	۰٫۴۹
	Bi-LSTM	۰٫۸۲	۰٫۴۳
	BERT	۰٫۸۳	۰٫۴۴
	CNN	۰٫۸۳	۰٫۴۱
	روش پیشنهادی	۰٫۸۴	۰٫۳۸



شکل (۳): نتایج پیاده‌سازی براساس معیار F1

بر اساس نتایج به‌دست‌آمده از اجرای مدل‌ها و مقایسه آن‌ها با هم به نظر می‌رسد میزان صحت و معیار F1، مدل‌های یادگیری عمیق مورداستفاده، به یکدیگر نزدیک است باوجوداین، مدل پیشنهادی و مدل کانولوشنی عملکرد نسبتاً بهتری داشته‌اند. براساس مقدار تابع هزینه نیز همان‌گونه که در هر سه جدول قابل‌مشاهده است، مدل پیشنهادی و مدل کانولوشنی مقدار کمتری را نشان می‌دهند. مقایسه‌ی کلی بین روش‌های یادگیری عمیق و الگوریتم‌های یادگیری ماشین سنتی نیز نشان‌دهنده‌ی برتری مدل‌های یادگیری عمیق در این زمینه است.

مدل پیشنهادی که علاوه‌بر لایه‌ی LSTM از لایه کانولوشن نیز استفاده می‌کند، مانند مدل کانولوشنی پاسخ بهتری نسبت به مدل‌های متراکم و LSTM دارد. دلیل این امر می‌تواند در قدرت بالای لایه‌ی کانولوشن برای استخراج ویژگی و کاهش ابعاد فضای ویژگی‌ها باشد. همچنین می‌توان نتیجه گرفت که لایه‌ی LSTM کمک چندانی به بهبود نتایج نداشته است. این نتیجه نشان‌دهنده‌ی این است که توالی بین کلمات کمتر از خود کلمات مهم بوده‌اند. به‌طورکلی می‌توان نتیجه گرفت، با استفاده از مدل پیشنهادی یا مدل کانولوشنی می‌توان با دقت قابل قبولی، بازیابی کارآموزان را بر اساس متن پاسخ‌هایی که به پرسش‌های دیگران داده‌اند، انجام داد.

۵- نتیجه‌گیری

بازیابی افراد با ویژگی‌ها و مهارت‌های مشخص، از جمله موضوعاتی است که در بسیاری از تحقیقات به آن پرداخته شده است؛ اما به بازیابی افرادی که برای موقعیت‌های کارآموزی، مناسب باشند کمتر توجه شده است. شناسایی چنین افرادی از منابع مختلفی امکان‌پذیر است که وب‌سایت‌های پرس‌وجوی انجمنی مانند stack overflow ازجمله این منابع هستند. در این پژوهش برای اولین بار یکی از مهم‌ترین بخش‌های مسأله‌ی بازیابی کارآموز یعنی شناسایی کاربران نوع خط پیوند معرفی شده است. در روش پیشنهادی، از متن تولید شده توسط کاربران در وب‌سایت‌های پرس‌وجوی انجمنی برای شناسایی کاربران نوع خط پیوند استفاده نمودیم. برای این منظور، مدلی مبتنی بر یادگیری عمیق پیشنهاد شد که در آن هم از لایه LSTM و هم از لایه کانولوشن برای استخراج و کاهش ابعاد ویژگی‌ها استفاده شده است.

برای اجرا و ارزیابی مدل‌های این پژوهش از سه مجموعه داده با موضوعاتی در زمینه‌ی مهندسی نرم‌افزار استفاده شده است، که هر یک از این موضوع‌ها شامل زمینه‌های مهارتی متنوعی هستند. نتایج حاصل از پیاده‌سازی مدل پیشنهادی و سایر مدل‌های یادگیری عمیق و یادگیری ماشین سنتی نشان داد که مدل پیشنهادی به همراه مدل مبتنی بر کانولوشن می‌تواند با دقت خوبی کاربران مناسب برای کارآموزی را شناسایی کند. برای پژوهش‌های آینده می‌توان به دنبال پیاده‌سازی مدلی بود که علاوه بر تشخیص سطح مهارت و توانایی افراد، زمینه‌های مهارتی آن‌ها را نیز بر اساس محتوای متن‌ها یا برچسب‌ها تشخیص دهد. همچنین می‌توان موضوع بازیابی کارآموز را برای متن‌های فارسی، بر اساس پرس‌وجوهای موجود در وب‌سایت‌ها و شبکه‌های اجتماعی، پیاده‌سازی کرد.

منابع:

1-Al-Taie, M. Z., Kadry, S., & Obasa, A. I. (2018). Understanding expert finding systems: domains and techniques. *Social Network Analysis and Mining*, *8*, 1-9.
2-Azzam, A., Tazi, N., & Hossny, A. (2017, April). Text-based question routing for question answering communities via deep learning. In *Proceedings of the Symposium on Applied Computing* (pp. 1674-1678).
3-Dargahi Nobari, A., Sotudeh Gharebagh, S., & Neshati, M. (2017, August). Skill translation models in expert finding. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (pp. 1057-1060).

4-Donofrio, N., Sanchez, C., & Spohrer, J. (2010). Collaborative innovation and service systems: Implications for institutions and disciplines. *Holistic engineering education: Beyond technology*, 243-269.
5-Gharebagh, S. S., Rostami, P., & Neshati, M. (2018). T-shaped mining: A novel approach to talent finding for agile software teams. *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, 411–423.
6-Kumar, V., & Pedanekar, N. (2016, February). Mining shapes of expertise in online social Q&A communities. In *Proceedings of the 19th ACM conference on computer supported cooperative work and social computing companion* (pp. 317-320).
7-Le, L. T., & Shah, C. (2018). Retrieving people: Identifying potential answerers in community question answering. *Journal of the association for information science and technology*, *69*(10), 1246-1258.
8-Lecun, Y. (2018). PERSPECTIVES Special Topic: Machine Learning Deep learning for natural language processing: advantages and challenges. *Natl. Sci. Rev*, *5*(1), 22-24.
9-Liang, S., & de Rijke, M. (2016). Formal language models for finding groups of experts. *Information Processing & Management*, *52*(4), 529-549.
10-M'Baya, A., Laval, J., Moalla, N., Ouzrout, Y., & Bouras, A. (2016, November). Ontology based system to guide internship assignment process. In *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)* (pp. 589-596). IEEE.
11-Narayanan, V. K., Olk, P. M., & Fukami, C. V. (2010). Determinants of internship effectiveness: An exploratory model. *Academy of Management Learning & Education*, *9*(1), 61-80.
12-P. Maertz Jr, C., A. Stoeberl, P., & Marks, J. (2014). Building successful internships: lessons from the research for interns, schools, and employers. *Career Development International*, *19*(1), 123-142.
13-Qiu, X., & Huang, X. (2015, June). Convolutional neural tensor network architecture for community-based question answering. In *Twenty-Fourth international joint conference on artificial intelligence*.
14-Rostami, P., & Neshati, M. (2021). Intern retrieval from community question answering websites: A new variation of expert finding problem. *Expert Systems with Applications*, *181*, 115044.
15-Yan, Z., & Zhou, J. (2015). Optimal answerer ranking for new questions in community question answering. *Information Processing & Management*, *51*(1), 163–178.
16-Yuan, S., Zhang, Y., Tang, J., Hall, W., & Cabotà, J. B. (2020). Expert finding in community question answering: a review. *Artificial Intelligence Review*, *53*, 843-874.
17-Zhao, M., Javed, F., Jacob, F., & McNair, M. (2015, January). SKILL: A system for skill identification and normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 29, No. 2, pp. 4012-4017).
18-Zhao, Y., Tang, J., & Du, Z. (2019). EFCNN: A Restricted Convolutional Neural Network for Expert Finding. In *Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, PAKDD 2019, Macau, China, April 14-17, 2019, Proceedings, Part II 23* (pp. 96-107). Springer International Publishing.