Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال پنجم، شماره اول، بهار و تابستان ۱۴۰۶، صفحات ۲۲ - ۱

تولید و ارزیابی مجموعه‌دادگان مهارت‌های صریح و ضمنی در پروفایل‌های فارسی با مدل‌های زبانی

علی حسین پورنادری[۱]، محمدعلی جوادزاده[۲*]، حسین حسینی[۳]

[۱] دانشجوی کارشناسی ارشد هوش مصنوعی و رباتیکز، دانشکده هوش مصنوعی و علوم شناختی، دانشگاه جامع امام حسین (ع)، تهران، ایران

[۲] استادیار، دانشکده هوش مصنوعی و علوم شناختی، دانشگاه جامع امام حسین (ع)، تهران، ایران

[۳] دانشجوی دکتری هوش مصنوعی و رباتیکز، دانشکده هوش مصنوعی و علوم شناختی، دانشگاه جامع امام حسین (ع)، تهران، ایران

مقاله پژوهشی

تاریخ دریافت:
۱۴۰۵/۰۴/۱۲

تاریخ پذیرش:
۱۴۰۵/۰۵/۲۷

کلیدواژه‌ها:
پردازش متنی، پروفایل مهارت‌های ضمنی، مجموعه‌دادگان Jhuskill، مدل‌های زبانی پیش‌آموزش‌دیده، ارزیابی پایایی

نویسنده‌مسئول:
alihpn@ihu.ac.ir

چکیده

استنتاج مهارت‌های کاربران از توصیف‌های متنی در پروفایل‌های شغلی، یکی از چالش‌های اساسی در سیستم‌های توصیه‌گر حوزه استخدام و پلتفرم‌های کاریابی است. پیچیدگی اصلی زمانی است که مهارت به‌صورت صریح ذکر نشده، بلکه در قالب جملات روایی و بافت‌های ضمنی پنهان‌شده است. با وجود پیشرفت‌های اخیر در مدل‌های زبانی، تاکنون مجموعه‌دادگان استانداردی برای پژوهش‌های زبان فارسی در این حوزه وجود نداشته که به‌طور عمیق به بازنمایی هم‌زمان مهارت‌های صریح و ضمنی بپردازد. در این پژوهش، نخستین مجموعه‌دادگان فارسی شامل ۱۰۴۳۰ پروفایل حاشیه‌نویسی‌شده است که از پلتفرم‌های کاریابی کارلنسر و پونیشا ارائه می‌شود و علاوه بر مهارت‌های صریح کاربران، چهار لایه اطلاعاتی دیگر را نیز پوشش می‌دهد: مهارت‌های ضمنی (در سه سطح دشواری)، توصیف روایی، وزن زمینه و آزمون پایایی. ساخت مجموعه‌دادگان با رویکرد ترکیبی «مدل زبانی بزرگ به‌علاوه حلقه انسانی» انجام شده است. برای اعتبارسنجی، چهار مدل زبانی mBERT، ParsBERT، RoBERTa و FaBERT بر روی پنج وظیفه طبقه‌بندی مهارت، تشخیص مهارت ضمنی، تولید توصیف روایی، طبقه‌بندی وزن زمینه و وظیفه سرتاسری ارزیابی شدند. نتایج نشان می‌دهد مدل FaBERT با کسب F1= ۰/۹۱ در طبقه‌بندی صریح و F1= ۰/۷۶ در تشخیص مهارت ضمنی، به‌ترتیب ۳/۴٪ و ۸/۶٪ بهبود نسبت به بهترین مدل پایه (ParsBERT) نشان داد و معماری فاین‌تیون‌شده آن در تمامی وظایف عملکردی برتر دارد. معرفی این مجموعه‌دادگان و مدل ارزیابی‌شده، بستری مؤثر برای بهبود سامانه‌های جستجوی مهارت‌محور، تطابق شغلی هوشمند و تحلیل پروفایل‌های متنی در حوزه مدیریت منابع انسانی می‌شود.

: 10.22034/ABMIR.2026.24873.1268

E-ISSN: 2821-2037 /The Author 2026. Published by Yazd University This is an open access article under the CC BY 4.0 License https://creativecommons.org/licenses/by/4.0/).

۱- مقدمه

امروزه با تحول در حوزه کسب‌وکارهای دیجیتال، شناسایی خودکار مهارت‌های نیروی کار به یکی از محورهای اساسی نظام‌های مدیریت منابع انسانی، پلتفرم‌های فریلنسری و سامانه‌های توصیه‌گر شغلی تبدیل‌شده است [۱]. هرروزه هزاران پروفایل شغلی در پلتفرم‌های کاریابی مانند Linkedin، Upwork و کارلنسر و جابینجا ایجاد می‌شوند که حاوی داده‌های متنی توصیفی و غیرساختاریافته از سابقه‌کاری، پروژه‌های انجام شده و دستاوردهای حرفه‌ای هستند. استخراج ساختاریافته مهارت‌ها از این متون، علاوه‌بر تسهیل جستجوی کارجویان و کارفرمایان، پایه‌گذار سامانه‌های تطابق هوشمند[1] و رتبه‌بندی نامزدهای شغلی است [۲, ۳]. در سطح جهانی، سرمایه‌گذاری روی فناوری‌های استخراج مهارت از متن رشد تصاعدی داشته است؛ پلتفرم‌های بزرگ مانند LinkedIn، گوگل و مایکروسافت سامانه‌های اختصاصی مانند Skill Graph و LinkedIn Talent Insights را برپایه مدل‌های زبانی پیش‌آموزش‌دیده و مجموعه‌دادگان حاشیه‌نویسی‌شده در مقیاس بزرگ توسعه داده‌اند [۳]. بااین‌حال، این سامانه‌ها اغلب با داده‌های اولیه زبان انگلیسی، آموزش‌دیده و برای زبان انگلیسی طراحی‌شده‌اند و عملکرد آن‌ها در زبان‌های کم‌منبع مانند فارسی به دلیل نبود مجموعه‌دادگان استاندارد به‌شدت کاهش می‌یابد. در ایران، با رشد پلتفرم‌های فریلنسری مانند کارلنسر، پونیشا و جابینجا، نیاز به سامانه‌های هوشمند استخراج مهارت و تطابق هوشمند برای شغل بیش‌ازپیش احساس می‌شود. از طرفی، حجم بالای پروفایل‌ها (صدها هزار پروفایل فعال در سایت‌های کاریابی) حاشیه‌نویسی دستی کامل را، با چالشی بزرگ روبه‌رو ساخته است. از طرف دیگر، تنوع زیاد مشاغل فریلنسری ایرانی نیازمند روشی است که ضمنی‌ترین نشانه‌های مهارت‌ها را تشخیص دهد.

یکی از اصلی‌ترین چالش‌ها در پژوهش‌های این حوزه، تمایز میان مهارت صریح و مهارت ضمنی است. مهارت صریح عبارتی است که در متن با واژگان مستقیم بیان‌شده باشد؛ برای نمونه، جمله «من در پروژه‌های پایتون و جنگو فعالیت داشته‌ام» به‌طور آشکار مهارت‌های «پایتون» و «جنگو» را بیان می‌کند [۵]. مهارت‌های ضمنی از طریق جملات روایی، بافت موقعیتی، فعل‌های کنشی، ابزارها یا دستاوردهای کمی قابل استنتاج است. برای مثال، جمله «کاهش ۳۰ درصدی زمان بارگذاری وب‌سایت با بهینه‌سازی کوئری‌های دیتابیس» به‌صراحت به مهارت «مدیریت پایگاه داده» اشاره نمی‌کند، اما ترکیب مفاهیم «بهینه‌سازی کوئری» و «کاهش درصدی زمان بارگذاری» به مهارت‌هایی مانند «بهینه‌سازی عملکرد» دلالت دارد. در پژوهش‌هایی که در سال‌های اخیر صورت گرفته نشان می‌دهد که ۳۵ تا ۵۰ درصد مهارت‌های موجود در پروفایل‌های شغلی به‌صورت ضمنی در جملات روایی ظاهر می‌شوند [۳، ۵]. تشخیص این مهارت‌ها نیاز به درک بافت گسترده جملات، دانش زمینه‌ای از حوزه شغلی و توانایی استدلال علّی است، درحالی‌که مدل‌های مبتنی بر تطابق الگو در مواجهه با جملات غیرساختارمند و عامیانه عملکرد ضعیفی دارند [۵].

از پژوهش‌های انجام‌شده قبلی، مجموعه‌دادگان استانداردی مانند SkillBERT (۲۵ هزار پروفایل، مهارت صریح [٤]، ESCoN [١٤] و SkillGraph-Narratives (۱۵۰۰ پروفایل با مهارت ضمنی و روایت) [۷] ایجادشده‌اند. هرچند مجموعه‌دادگان اخیر گامی مهم محسوب می‌شوند، اما منحصراً برای زبان‌های غیرفارسی (اغلب به زبان انگلیسی) و با هزینه بسیاربالا تهیه‌شده‌اند. در زبان فارسی که جزو زبان‌های کم‌منبع بشمار می‌رود تاکنون مجموعه‌دادگانی برای تشخیص مهارت‌های صریح و ضمنی معرفی نشده‌است. بدین‌ترتیب، شکاف ادبیات در زبان فارسی در این‌حوزه، وجود یک مجموعه‌دادگان استاندارد با حاشیه‌نویسی مهارت‌های صریح و ضمنی با درنظر گرفتن کم‌منبع بودن زبان فارسی است، شکافی که پژوهش‌حاضر به‌طورمستقیم به آن می‌پردازد.

نوآوری‌های اصلی این پژوهش عبارت‌اند از:

- ارائه نخستین مجموعه‌دادگان فارسی با ۱۰۴۳۰ پروفایل حاشیه‌نویسی‌شده از پلتفرم‌های کاریابی کارلنسر و پونیشا

[1] Skill-Based Matching

- پوشش دادن شش لایه اطلاعاتی شامل مهارت‌های صریح، مهارت‌های ضمنی در سه سطح دشواری، قطعه شاهد، توصیف روایی، وزن زمینه و آزمون پایایی.
- این مجموعه‌دادگان به روش ترکیب حلقه انسانی (داوران) و مدل زبانی (gpt-4o) ساخته شده‌است
- برای ارزیابی این مجموعه‌داده‌گان بر روی پنج وظیفه پایین‌دستی با چهار مدل زبانی (mBERT، ParsBERT، RoBERTa و FaBERT) انجام شده و برتری مدل FaBERT اثبات شد.

پس از مقدمه، در بخش دوم به‌مرور نظام‌مند کارهای مرتبط می‌پردازد. در بخش سوم روش‌شناسی ساخت مجموعه‌دادگان تشریح می‌شود. بخش چهارم مدل‌های پایه و تنظیمات ارزیابی را ارائه می‌دهد. بخش پنجم نتایج کمی و کیفی را تحلیل می‌کند. بخش ششم به بحث و محدودیت‌ها اختصاص دارد و بخش هفتم با نتیجه‌گیری و مسیرهای آینده پایان می‌یابد.

۲- کارهای مرتبط

در این بخش، ابتدا کارهای مرتبط را در پنج بخش، مدل‌های زبانی پیش‌آموزش‌دیده فارسی، پژوهش‌های استخراج مهارت (صریح و ضمنی)، مجموعه‌دادگان پروفایل شغلی، تولید توصیف روایی و کاربردهای نظام‌های تطابق شغلی بررسی می‌گردند. در آخر این بخش، جدول مقایسه‌ای شکاف‌های پژوهشی را مشخص می‌کند.

۲-۱ مدل‌های زبانی پیش‌آموزش‌دیده برای زبان فارسی

توسعه مدل‌های ترانسفورمر برای فارسی با ParsBERT آغاز شد [۱۶] که بر روی پیکره ۲/۱ میلیون سندی آموزش دید و در وظایفی مانند تشخیص موجودیت‌های اسمی و طبقه‌بندی متن عملکرد بهتری از mbert داشت. بااین‌حال، ParsBERT عمدتاً روی داده‌های خبری آموزش‌دیده و توکن‌های خارج از واژگان در متون تخصصی شغلی چالش‌برانگیز است [۱۷]. FaBERT [۱۰] با معماری Roberta، پنهان‌سازی پویا و واژگان BPE (۷۰ هزار توکن) ارائه شد و روی پیکره ۷/۸ میلیون سندی آموزش دید. در پژوهش‌های دیگری نشان‌داده‌شده که FaBERT در تحلیل متون فرینسری نسبت به ParsBERT کاهش خطای نسبی ۸ تا ۱۲ درصدی دارد [۱۹]. بااین‌حال، هیچ‌یک از این مدل‌ها برای استخراج مهارت‌های ضمنی فاین‌تیون نشده‌اند. مدل‌های چندزبانه مانند mBERT و XLM-RoBERTa [۲۶] نیز به دلیل سهم اندک فارسی در پیکره پیش‌آموزش (کمتر از ۲٪)، درک ضعیفی از جملات فارسی دارند [۲۲]. پژوهش از FaBERT، ParsBERT، mBERT و RoBERTa به‌عنوان مدل‌های مبنا استفاده می‌کند.

به‌تازگی، پیکره‌های بزرگی مانند Matina [۲۴] با ۷۲/۹ میلیارد توکن برای آموزش مدل‌های زبانی فارسی معرفی‌شده‌اند و پیکره‌های ارزیابی جدیدی مانند MELAC [۲۰] و PK-BETS [۲۱] برای سنجش هم‌ترازی فرهنگی و دانش عمومی مدل‌های فارسی توسعه‌یافته‌اند.

۲-۲ استخراج مهارت از پروفایل‌های شغلی

پژوهش‌های اولیه دراین‌حوزه، روی تعیین موجودیت‌های مهارتی درون متون متمرکز بوده است. به‌طوری‌که از تلاش‌های اولیه در این‌حوزه مبتنی بر روش‌های قانون‌محور و تطابق با فرهنگ مهارت ESCOرا می‌توان نام‌برد [۱۳]. این روش‌ها در مواجهه با مهارت‌های پنهان درون روایت‌ها و عبارت‌های مترادف شکست می‌خورد. در پژوهش‌های بعد، روش‌های توالی‌برچسب‌زنی مانند BiLSTM-CRF بیشتر مورداستفاده قرارگرفتند. در زبان انگلیسی، مجموعه‌دادگان SkillBERT با جمع‌آوری ۲۵ هزار پروفایل Linkedin و ۱۸۰ نوع مهارت صریح، نقطه عطفی در مقیاس داده داشت اما فاقد حاشیه‌نویسی ضمنی است [۱۲]. مجموعه‌دادگان EventNarrative که در سال ۲۰۲۴ انجام شد، جزو گام‌های اولیه در حوزه تولید روایت از داده‌های ساختاریافته به‌شمار می‌رود که این مجموعه‌دادگان به‌جای پروفایل‌های شغلی، بر روایت‌های عمومی متمرکز بوده و چالش وجود یک مجموعه‌دادگان تخصصی برای استخراج مهارت از روایت‌های شغلی فارسی همچنان وجود داشت. در پژوهش‌های مبتنی بر GPT-3/5 برای تشخیص مهارت ضمنی به صحت ۶۲٪ در متون کوتاه رسیدند، اما با افزایش طول متن و تعداد مهارت‌ها، صحت به کمتر از ۴۰٪ کاهش یافت و

مشکل توهم[1] در این پژوهش‌ها شایع بود [۲۶, ۲۷]. این ناکارآمدی‌ها، ضرورت وجود داده‌های مرجع استاندارد و افزودن سنجش پایایی در این پژوهش بود.

در سال‌های اخیر، رویکردهای جدیدی برای استخراج مهارت‌های ضمنی با استفاده از مدل‌های زبانی بزرگ ارائه‌شده است. برای نمونه، SkiLLMo [۲۲] از مهندسی پرامپت برای استخراج هم‌زمان مهارت‌های صریح و ضمنی از متون استخدامی استفاده می‌کند و رویکرد LLM-Supervised [۲۳] نیز استخراج مهارت‌های ضمنی را در کنار مهارت‌های صریح در آگهی‌های شغلی پوشش می‌دهد. این پژوهش‌ها نشان می‌دهند که استفاده از مدل‌های زبانی بزرگ به‌عنوان ابزار استخراج مهارت، رویکردی رو به رشد در این حوزه است.

۲-۳ مجموعه‌دادگان پروفایل شغلی در زبان فارسی

با وجود گسترش پژوهش‌های پردازش زبان طبیعی برای زبان فارسی در حوزه‌هایی نظیر طبقه‌بندی متن، تحلیل احساسات و پرسش‌پاسخ (پیکره‌های معتبری مانند همشهری [۱۷]، HmBlogs [۱۹] و مجموعه‌داده‌های جامع ParsiNLU [۱۸] برای آن‌ها توسعه داده‌شده است.)، در حوزه تخصصی پروفایل‌های شغلی و استخراج مهارت، هنوز پژوهش‌های پردازش متن دچار فقدان منابع داده‌ای استاندارد است.

بررسی نظام‌مند مخازن علمی معتبر (شامل ACL Anthology ، IEEE Xplore و پایگاه‌های داخلی مانند مگیران و علم نت) نشان می‌دهد که تا پیش‌ازاین پژوهش، تا حد دانش و بررسی‌های انجام شده مجموعه‌داده‌گان دسترس‌پذیری برای پروفایل‌های شغلی فارسی منتشر نشده‌است که معیارهای حداقلی پژوهش‌های تطابق شغلی را دارا باشد. تلاش‌های عمده‌ای که در این‌حوزه صورت پذیرفته، بیشتر به یکی از سه دسته زیر تعلق دارند:

۱) پیکره‌های درون‌سازمانی و محرمانه[2]: برخی پلتفرم‌های داخلی جستجوی شغل، به دلیل ملاحظات حریم‌خصوصی و تجاری، از مجموعه‌دادگان داخلی خود استفاده می‌کنند، اما این منابع در دسترس عموم قرار نگرفته و فرایند حاشیه‌نویسی آن‌ها بر اساس اصول علمی مستند نشده‌است. از منظر روش‌شناسی، چنین داده‌هایی که فاقد قابلیت بازتولیدپذیری[3] هستند، نمی‌توانند مبنای ارزیابی بی‌طرفانه مدل‌ها قرار گیرند.

۲) پیکره‌های کوچک‌مقیاس با برچسب‌زنی سطحی: تعداد کمی از پژوهش‌های داخلی، به استخراج ساده عبارات کلیدی (صرفاً مهارت‌های صریح) از تعداد محدودی رزومه پرداخته‌اند، اما این تلاش‌ها ساختار استانداردی ندارند و یا لایه‌های عمیق‌تری نظیر مهارت ضمنی و وزن زمینه را پوشش نداده و پژوهش‌هایی در سطح شناسایی کلمات NER هستند.

۳) پیکره‌های عمومی که به‌اشتباه درحوزه شغلی به‌کار گرفته می‌شوند: برخی پژوهشگران از پیکره‌های خبری (مانند همشهری) برای آزمون مدل‌های خود بر روی متون شغلی استفاده می‌کنند، درحالی‌که توزیع واژگانی و ساختار نحوی این پیکره‌ها تفاوت اساسی با جملات روایی پروفایل‌های فریلنسری دارد و اعتبار نتایج چنین ارزیابی‌هایی دقیق و قابل‌اعتماد نیست.

این پژوهش با ساخت نخستین مجموعه‌دادگان ترکیبی مهارت‌های صریح، ضمنی، روایت و وزن زمینه در زبان فارسی، برای پر کردن این خلأ طراحی‌شده است.

۲-۴ تولید توصیف روایی[4]

تولید خودکار توضیحات متنی از مهارت‌های شغلی، زیرشاخه‌ای جدید در حوزه تولید زبان طبیعی (NLG[5]) است که با چالش‌های انسجام، تنوع واژگانی و وفاداری به داده مواجه است. پژوهش‌های اولیه بر مجموعه‌دادگان عمومی گراف‌به‌متن متمرکز بوده و کاربرد در حوزه رزومه محدود است. مدل‌هایی نظیر GPT-2 و Llama2 برای تولید محتوای رزومه به‌کاررفته‌اند، اما ارزیابی‌های کمی با محدودیت‌هایی در معیارهای سنتی مواجه است. در زبان فارسی، تا زمان نگارش این مقاله، هیچ پژوهشی به تولید روایت از مهارت‌ها نپرداخته است. مجموعه‌دادگان حاضر با ۱۰۴۳۰ جفت‌داده، نخستین بستر ارزیابی مدل‌های مولد فارسی را در این حوزه فراهم می‌کند.

[1] Hallucination
[2] Proprietary Corpora
[3] Reproducibility
[4] Narrative Generation
[5] Natural Language Generation

۲-۵ کاربردها در نظام‌های تطابق شغلی

مجموعه‌دادگان استخراج مهارت، یکی از زیرساخت‌های اساسی سیستم‌های تطابق شغلی[1] هستند. در سطح بین‌المللی، سیستم‌های پیشرویی نظیر LinkedIn Talent Insights از ترکیب مهارت‌های صریح و ضمنی برای تحلیل و تطابق پروفایل‌ها استفاده می‌کنند. بر اساس مستندات رسمی Linkedin، مهارت‌های صریح آن‌هایی هستند که کاربران در بخش Skills پروفایل خود درج کرده‌اند، درحالی‌که مهارت‌های ضمنی از طریق فنون پردازش زبان طبیعی از بخش‌های دیگر پروفایل نظیر عناوین شغلی، شرح موقعیت‌ها و خلاصه پروفایل استخراج می‌شوند. همچنین، پلتفرم‌هایی مانند Upwork الگوریتم‌های تطابق پیشرفته‌ای بکار می‌برند که بر اساس مهارت‌های درخواستی کارفرما و سوابق فریلنسرها عمل می‌کنند؛ بااین‌حال، جزئیات معماری این سیستم‌ها به‌صورت عمومی اعلام‌نشده است.

بررسی اکوسیستم کاریابی دیجیتال ایران نشان می‌دهد که پلتفرم‌های موجود عمدتاً بر اساس جستجوی کلیدواژه‌ای و تطابق سطحی از طریق مهارت‌های صریح عمل می‌کنند و فاقد سیستم‌های هوشمند استخراج و تطابق مهارت‌های ضمنی هستند. این نقصان، یکی از چالش‌های اصلی توسعه سامانه‌های تطابق شغلی کارآمد در زبان فارسی محسوب می‌شود. پژوهش حاضر با فراهم‌آوردن نخستین مجموعه‌دادگان ترکیبی مهارت‌های صریح، ضمنی، روایت و وزن زمینه در زبان فارسی، می‌تواند سرآغازی برای توسعه نخستین سامانه تطابق هوشمند مهارت‌محور در اکوسیستم فریلنسری ایران باشد.

۲-۶ جدول مقایسه پژوهش‌های مرتبط

برای درک بهتر جایگاه پژوهش حاضر در میان کارهای پیشین، جدول ۱ پژوهش‌های مرتبط را ازنظر زبان، نوع حاشیه‌نویسی، حجم مجموعه‌داده، نوآوری اصلی و محدودیت‌ها مقایسه می‌کند. این مقایسه نشان می‌دهد که تاکنون هیچ مجموعه‌داده‌ای برای زبان فارسی با این عمق حاشیه‌نویسی (شامل مهارت‌های صریح، ضمنی سطح‌بندی‌شده، روایت، وزن زمینه و آزمون پایایی) ارائه نشده است

همان‌طور که جدول ۱ نشان می‌دهد، این پژوهش ازنظر تنوع و عمق لایه‌های حاشیه‌نویسی بهتر از پژوهش‌های این‌حوزه در زبان انگلیسی بوده و از طرفی اولین مجموعه‌دادگان زبان فارسی در این‌حوزه است. هرچند حجم مجموعه‌دادگان از برخی نمونه‌های انگلیسی کمتر است، اما غنای اطلاعات در هر نمونه به‌گونه‌ای است که برای فاین‌تیونینگ مدل‌های زبانی کافی است.

۳- روش‌شناسی و ساخت مجموعه‌دادگان

در این بخش، مراحل طراحی، جمع‌آوری، پالایش، حاشیه‌نویسی و ارزیابی پایایی مجموعه‌دادگان به‌طور مشروح بیان می‌شود تا پژوهشگران بتوانند درصورت نیاز فرایند بازتولیدی را انجام‌دهند. رویکرد کلی مبتنی بر ترکیب «نیروی انسانی متخصص» و «کمک‌گیری از مدل زبانی بزرگ» با حلقه بازخورد انسانی است.

۳-۱ منبع داده و معیارهای فیلترینگ

پروفایل‌های متنی از پلتفرم‌های کاریابی ایرانی جمع‌آوری شدند. داده‌ها در بازه زمانی دوهفته‌ای در بهمن ۱۴۰۴ از طریق کراول پلتفرم با کسب اجازه از سایت‌های هدف برداشت شدند. برای رعایت حریم خصوصی، تمام اطلاعات شناسه‌دار شخصی (نام، نام خانوادگی، ایمیل، شماره تماس) حذف و هر پروفایل با یک کد یکتا (UUID) جایگزین شد. همچنین امکان درخواست حذف داده‌های ناشناس‌شده کاربر نیز فراهم بود.

برای افزودن پروفایل کاربر به مجموعه‌دادگان ۴ معیار در نظر گرفته شد: (۱) داشتن حداقل ۱۵۰ کلمه متن توصیفی در بخش «درباره من» یا «سوابق پروژه»؛ (۲) وجود حداقل یک پروژه تکمیل‌شده با توضیحات متنی و نظر ثبت‌شده توسط کارفرما؛ (۳) نداشتن بیش از ۷۰٪ جملات تکراری یا بی‌معنا؛ (۴) تنوع حوزه‌های شغلی بر اساس برچسب اولیه پلتفرم: فناوری اطلاعات، طراحی گرافیک و مالتی‌مدیا، بازاریابی و سئو، تولید محتوا و ترجمه، مهندسی مخابرات، و برنامه‌نویسی و سایر. از مجموع ۱۰۵۰۰ پروفایل اولیه، پس از اعمال فیلترها و حذف نمونه‌های باکیفیت پایین، ۱۰۴۳۰ پروفایل واجد شرایط باقی ماندند.

[1] Skill Matching Systems

جدول (۱): مقایسه پژوهش‌های مرتبط درحوزه استخراج مهارت از پروفایل‌های متنی[1]

پژوهش	زبان	نوع حاشیه‌نویسی	حجم مجموعه‌دادگان	نوآوری اصلی	محدودیت اصلی
SkillBERT [۹]	انگلیسی	مهارت‌های صریح	۲۵،۰۰۰	مقیاس بزرگ، ۱۸۰ نوع مهارت	بدون مهارت ضمنی، بدون وزن
ESCO taxonomy [۱۴]	چندزبان اروپایی	سلسله‌مراتب مهارت	۱۳،۰۰۰ مهارت	استاندارد اروپایی	غیرقابل استفاده برای تشخیص خودکار از متن
SkillSpan [۶]	انگلیسی	ضمنی + صریح	۱،۴۰۰	رویکرد چندلایه در تشخیص مهارت	فاقد روایت و وزن زمینه
مجموعه‌دادگان ترکیبی انگلیسی [۲۰]	انگلیسی	ضمنی + روایت + وزن	~۱،۵۰۰	پوشش هم‌زمان لایه‌های مختلف	هزینه بالا، عدم پشتیبانی از فارسی
پژوهش حاضر	فارسی	صریح + ضمنی + روایت + وزن + آزمون پایایی	۱۰۴۳۰	نخستین مجموعه‌دادگان ترکیبی فارسی	حجم نسبتاً کوچک (قابل‌گسترش)
SkiLLMo [۲۲]	انگلیسی	صریح + ضمنی (سخت و نرم)	–	استخراج و استانداردسازی مهارت‌ها با ترکیب BERT و LLM و نگاشت به ESCO	عدم پشتیبانی از فارسی، فاقد روایت و وزن زمینه
LLM-Supervised [۲۳]	چندزبانه (انگلیسی، دانمارکی و...)	صریح + ضمنی	+k۵۰۰ جمله از +M۳ آگهی	نظارت با LLM برای آموزش رمزگذار سبک، پوشش مهارت‌های ضمنی	عدم پشتیبانی از فارسی، متمرکز بر آگهی‌های شغلی نه پروفایل

۳-۲ آمار توصیفی مجموعه‌دادگان

میانگین طول هر پروفایل ۴۲۳ واژه با انحراف معیار ۱۵۸ واژه است. میانگین تعداد مهارت‌های صریح شناسایی‌شده توسط حاشیه‌نویسان پیش از حاشیه‌نویسی ضمنی ۴/۲ با دامنه ۱ تا ۱۱ و میانگین مهارت‌های ضمنی اضافه‌شده در مرحله دوم ۳/۷ با دامنه ۲ تا ۷ است. توزیع حوزه‌های شغلی: فناوری اطلاعات ۳۲٪ (حدود ۳،۳۳۸ نمونه)، طراحی گرافیک ۲۴٪ (۲،۵۰۳ نمونه)، بازاریابی و سئو ۱۸٪ (۱،۸۷۷ نمونه)، تولید محتوا و ترجمه ۱۴٪ (۱،۴۶۰ نمونه)، مدیریت پروژه ۸٪ (۸۳۴ نمونه) و سایر حوزه‌ها ۴٪ (۴۱۷ نمونه). از ۱۰۴۳۰ پروفایل، ٪۹۰ حاوی حداقل یک مهارت ضمنی هستند. بیش از ۷۰٪ از جملات حاوی مهارت ضمنی در طبقه «کلیدی» (معادل وزن‌های بالای ۰/۷) قرار دارند، درحالی‌که این نسبت برای جملات بدون مهارت حدود ۳۵٪ است. این تفاوت نشان‌دهنده اهمیت بالاتر جملات حاوی مهارت ضمنی در تعیین وزن زمینه است.

۳-۳ فرآیند حاشیه‌نویسی چهار مرحله‌ای

همان‌طور که در شکل ۱ نشان داده‌شده‌است، در مرحله اول، حاشیه‌نویسی با پرامپت تخصصی توسط مدل زبانی برای کاهش بار انسانی، با استفاده از پرامپت‌های سیستمی و کاربری که به مدل زبانی gpt-4o-2024-11-20 داده شد، برای هر پروفایل در سه دسته: (۱) فهرستی از مهارت‌های صریح و ضمنی، (۲) قطعه شاهد برای هر مهارت ضمنی، (۳) وزن زمینه‌ای پیشنهادی برای هر جمله در پروفایل (به‌صورت سه سطح بالا، متوسط، ضعیف) تولید شد. در مرحله دوم، به‌کمک ۲ حاشیه‌نویس با سابقه در تحلیل رزومه و آشنا با داده‌کاوی به‌صورت جداگانه به‌صورت دقیق، موارد تولیدشده توسط مدل زبانی را بررسی کرده و درصورت وجود

[1] Comparison Of Related Work In Skill Extraction From Text Profiles

خطا آن‌ها را برطرف کرده و برای اطمینان از توافق بین حاشیه‌نویسان موارد تصحیح‌شده، دوباره توسط ۱ حاشیه‌نویس دیگر کنترل شد.

در مرحله سوم، به‌کمک مدل زبانی یک روایت کوتاه (۲۰–۴۰ واژه) برای هر پروفایل تولید شد که پرامپت داده‌شده شامل مهارت‌های صریح و ضمنی و قطعه شاهد بود تا یک روایت از مهارت‌های کاربر تولید شود.

در مرحله چهارم، به‌طور تصادفی ۳۰٪ از پروفایل‌های تکمیل‌شده به حاشیه‌نویسان ارجاع داده شد تا تصمیم نهایی را در سه حالت «پذیرش»، «رد» یا «ویرایش» بگیرند. این فرآیند باعث کشف ۸۰ مهارت ضمنی جدید شد که در مرحله اول بازرسی حاشیه‌نویسی یافت نشده بودند.

۳-۴ معیارهای دقیق حاشیه‌نویسی برای پنج مؤلفه اصلی

همان‌طور که در شکل ۲ دیده می‌شود مدل مفهومی داده‌های استخراج‌شده از پروفایل کاربر که شامل ۶ سطح است نشان داده‌شده‌است که شامل موارد زیر است:

۱) مهارت‌های صریح: عبارتی که عیناً در متن آمده و بدون نیاز به استنتاج معنایی قابل‌شناسایی است (مثال: تسلط به پایتون، بانکداری، هتلداری).

۲) مهارت‌های ضمنی: مهارتی که در متن ذکر نشده اما از ترکیب فعل، ابزار، دستاورد کمی یا بافت جمله قابل استنباط است.

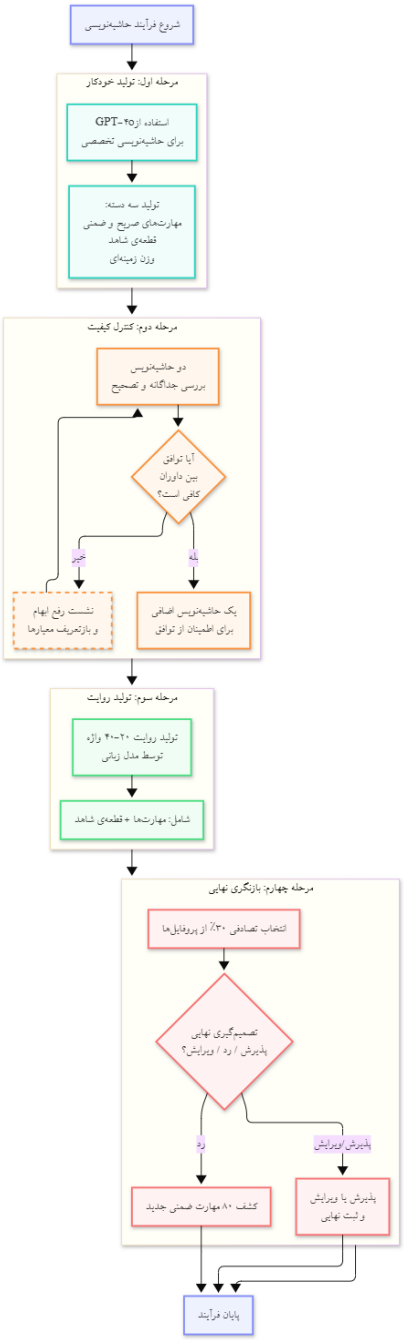
سطح ۱: یک گام استنتاج ساده (مثلاً «کاهش زمان بارگذاری» ⟶ بهینه‌سازی سایت).

سطح ۲: نیاز به ترکیب چندین جمله (مثلاً «تغییر ایندکس‌ها» + «مهاجرت به داکر» ⟶ دواپس).

سطح ۳: نیاز به دانش تخصصی خارج از متن (مثلاً A/B تست نرخ کلیک ⟶ بازاریابی مبتنی برداده).

قطعه شاهد: زیررشته‌ای از متن که حاشیه‌نویس را به آن مهارت رسانده است.

توصیف روایی[1]: جمله کاملاً طبیعی که همان مهارت را به زبانی شبیه آنچه کاربر در رزومه می‌نویسد، توضیح می‌دهد.



شکل (۱): نمای کلی فرایند ساخت و اعتبارسنجی مجموعه‌دادگان Ihuskill

وزن زمینه[2]: به‌جای عدد پیوسته، هر جمله در یکی از سه طبقه‌بندی قرار می‌دهد: کلیدی (بالا): معادل وزن‌های ۰/۸–۱، حمایت‌کننده (متوسط): معادل ۰/۷–۰/۴، و زائد (ضعیف): کمتر از ۰/۴. برای

[1] Narrative Explanation

[2] Context Weight

افزایش عینی‌بودن، در موارد اختلاف، برای حل اختلافات در تعیین وزن زمینه، از تحلیل وابستگی نحوی با ابزار Stanza استفاده شد و فاصله نحوی (Syntactic Distance) بین هسته فعل جمله و عبارت کلیدی مهارت محاسبه گردید. سپس وزن نهایی هر جمله با ترکیب خطی میانگین نظرات حاشیه‌نویسان و معکوس فاصله نحوی نرمال‌شده به‌دست آمد:

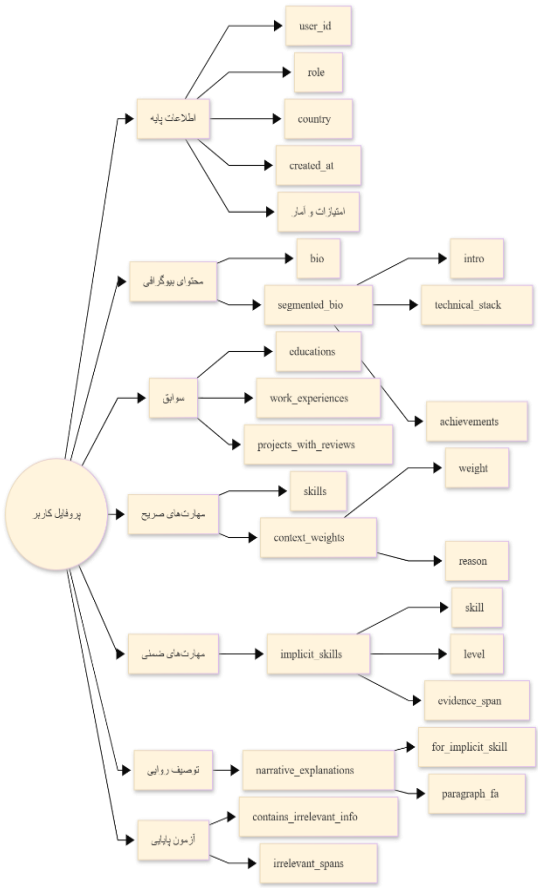
$$Weight_{final} = 0.7 \times Mean_{annotators} + 0.3 \times \left(1 - \frac{SyntacticDistance}{maxDistance}\right) \quad (۱)$$

در فرمول ۱ maxDistance بیشینه فاصله نحوی مشاهده‌شده در کل مجموعه‌دادگان است. این فرمول، توافق بین‌حاشیه‌نویسان را از k=۰/۶۹ به k=۰/۸۱ افزایش داد.

آزمون پایایی[1]: آیا اگر تنها همان قطعه شاهد به فردی نشان داده شود، او می‌تواند همان مهارت را استنباط کند؟ (بلی/خیر/تا حدی). نمونه‌هایی با پاسخ «خیر» (کمتر از ۳٪) حذف شدند.

۳-۵ توافق بین حاشیه‌نویسان[2]

برای محاسبه پایایی، ۱۵۰ پروفایل (حدود ۲۰٪) توسط هر سه حاشیه‌نویس مستقل کدگذاری شدند. مقادیر ضریب کاپا فلیس[3]: مهارت صریح ۰/۸۹ (توافق تقریباً کامل)، مهارت ضمنی ۰/۷۶ (قابل‌قبول)، سطح مهارت ضمنی ۰/۶۸ (متوسط). وزن زمینه: کاپا برابر ۰/۷۴ (قابل‌قبول) که نسبت به حالت پیوسته (ICC=۰/۶۹) بهبودیافته است. روایت: BLEU بین جفت حاشیه‌نویسان میانگین ۵۲/۴ بوده و علاوه برآن به‌منظور پوشش شباهت‌های معنایی (و نه صرفاً تطابق سطحیِ n-gram)، معیار BERTScore نیز محاسبه گردید که میانگین آن ۸۶/۳ به‌دست آمد. این مقدار نشان‌دهنده هم‌پوشانی معنایی مطلوب بین روایت‌های تولیدشده توسط حاشیه‌نویسان است و کاستی‌های معیار BLEU در جملات کوتاه روایی را جبران می‌کند. موارد اختلاف در نشست‌های هفتگی به‌اتفاق نظر رسیدند.



شکل (۲): مدل مفهومی داده‌های پروفایل کاربر شامل شش سطح اصلی

۳-۶ ساختار JSON نهایی

هر نمونه در قالب JSON ذخیره می‌شود. شکل ۳ یک نمونه واقعی از مجموعه‌دادگان را نمایش می‌دهد. فیلدهای کامل شامل segmented_bio جملات تفکیک‌شده با وزن‌های زمینه مستقل و faithfulness_test در سطح هر حاشیه‌نویسی است.

۳-۷ چالش‌ها و راهکارها

در فرایند طراحی و حاشیه‌نویسی مجموعه‌دادگان، چهار چالش اصلی شناسایی و با راهکارهای سیستماتیک مهار شدند:

۱) ابهام در سطح‌بندی مهارت‌های ضمنی (به‌ویژه سطح ۳): مهارت‌های سطح ۳ که نیازمند دانش تخصصی خارج از متن هستند، بیشترین نرخ اختلاف بین حاشیه‌نویسان را به خود

[1] Faithfulness Test
[2] Inter-Annotator Agreement
[3] Fleiss' Kappa

اختصاص دادند (κ=۰/۶۸). برای مهار این چالش، از یک متخصص حوزه بازاریابی دیجیتال و تحلیل کسب‌وکار به‌عنوان مشاور کمک گرفته‌شد و ۵۰ نمونه پرچالش به‌صورت دوسویه‌کور[1] بازبینی گردید. نتیجه این بازبینی، بازتعریف معیارهای سطح ۳ و ارائه ۱۰ مثال نقضی[2] در دستورالعمل بود.

۲) تعارض در تعیین وزن زمینه: در مواردی که یک جمله به چند مهارت هم‌زمان اشاره داشت، تخصیص وزن واحد به چالش کشیده می‌شود. راهکار: قاعده‌گذاری اولویت وزن با مهارتی که جمله ازنظر نحوی و معنایی به آن وابسته‌تر بوده و ثبت وزن‌های چندگانه در یک ساختار مجزا (آرایه‌ای از وزن‌ها به‌ازای هر مهارت) به‌جای وزن واحد برای کل جمله انجام‌می‌شود. این تغییر، توافق بین حاشیه‌نویسان را از ۰/۶۹ به ۰/۸۱ افزایش داد.

۳) خطر پذیرش کورکورانه[3] پیشنهادهای LLM: برای جلوگیری از اتکای بیش‌ازحد به خروجی مدل، از دو روش استفاده شد: (الف) نمایش نمره اطمینان[4] به‌همراه هر پیشنهاد و الزام حاشیه‌نویس به ثبت دلیل رد یا ویرایش؛ (ب) اعمال قانون حداقل یک تغییر معنادار ویرایش روایت یا اصلاح قطعه شاهد در ۱۰۰٪ نمونه‌های بازبینی‌شده در مرحله چهارم، که موجب کاهش ۲۲٪ از خطاهای پذیرش پیشنهادهای نادرست گردید.

٤) کمبود پروفایل‌هایی با مهارت در حوزه‌های شغلی نادر (مهندسی مخابرات و سایر با ۱۲٪ داده): این عدم توازن می‌توانست باعث سوگیری مدل در مرحله آموزش شود که با نمونه‌گیری موزون[5]، به‌گونه‌ای که درمجموعه آموزش، وزن هر حوزه متناسب با وارونِ فراوانی آن در جمعیت اصلی تنظیم‌شده و علاوه‌براین، ۲۰ پروفایل مکمل از حوزه‌های نادر به‌صورت هدفمند به مجموعه‌دادگان افزوده شد تا تعادل نسبی برقرار گردد.

مجموعه‌دادگان نهایی به همراه کدهای بازتولیدپذیر در مخزن گیت‌هاب و هاگینگ فیس پژوهش در دسترس خواهد بود.


```
"user_id": 473776,
"text": "شن‌های تحت وب پیشرفته پرداخته ام.",
"explicit_skills": [
  "سی شارپ",
  "دبلیو پی اف",
  "سی شارپ",
  "برنامه‌نویسی دسکتاپ ویندوز",
  "جاوا اسکریپت",
  "ای اس پی دات نت کور",
  "ای اس پی دات نت کور ام وی سی",
  "ربات تلگرام",
  "رستفول ای پی آی",
  "سی شارپ"
],
"implicit_skills": [
  {
    "skill": "کار تیمی",
    "level": 2,
    "evidence_span": "حرفه‌ای معرفی کرده ام
  },
  {
    "skill": "ارتباط مؤثر با مشتری",
    "level": 2,
    "evidence_span": "حرفه‌ای معرفی کرده ام
  },
  {
    "skill": "یادگیری مستمر",
    "level": 3,
    "evidence_span": "Agile توسعه مبتنی بر
  },
```


شکل(۳): نمونه ساختار JSON یک پروفایل حاشیه‌نویسی‌شده درمجموعه‌دادگان پیشنهادی

برای شفاف‌سازی کمی نقش «حلقه انسانی» در فرایند حاشیه‌نویسی ترکیبی، نرخ توافق اولیه بین خروجی مدل زبانی (GPT-4o) و حاشیه‌نویسان انسانی، و نیز نرخ توافق نهایی پس از اعمال بازبینی‌ها در جدول زیر گزارش‌شده است. این ارقام بر اساس بررسی ۳۰٪ از پروفایل‌های تکمیل‌شده در مرحله چهارم (معادل ۳۱۲۹ نمونه) محاسبه‌شده‌اند.

همان‌طور که جدول ۲ نشان می‌دهد، اگرچه خروجی اولیه مدل زبانی تنها در ٦٨٪ موارد با نظر حاشیه‌نویسان توافق داشته است، اما استفاده از رویکرد ترکیبی «LLM + حلقه انسانی» توانست زمان حاشیه‌نویسی را به‌طور میانگین ٦٧٪ کاهش دهد (از حدود ۱۲ دقیقه به ٤ دقیقه به‌ازای هر پروفایل). این صرفه‌جویی با پژوهش‌های مشابه درزمینه حاشیه‌نویسی کم‌هزینه با کمک LLM

[1] Double-Blind
[2] Counterexamples
[3] Automation Bias
[4] Confidence Score
[5] Stratified Sampling

همخوانی دارد؛ به‌گونه‌ای که استفاده از مدل‌های زبانی بزرگ به‌عنوان دستیار حاشیه‌نویسی می‌تواند هزینه‌ها را به‌طور چشمگیری کاهش دهد. همچنین اعمال قانون «حداقل یک تغییر معنادار» در مرحله بازبینی نهایی، میزان پذیرش کورکورانه پیشنهادهای نادرست را ۲۲٪ کاهش داد که خود گواهی بر اثربخشی حلقه بازخورد انسانی در افزایش کیفیت نهایی مجموعه‌دادگان است.

جدول (۲): نرخ توافق و هزینه تصحیح خروجی‌های مدل زبانی

شاخص	مقدار
توافق اولیه (خروجی LLM با نظر اولیه حاشیه‌نویسان)	٪۶۸
توافق نهایی (پس از بازبینی‌های انسانی و اعمال اصلاحات)	٪۸۶
درصد نمونه‌های نیازمند ویرایش جزئی	٪۲۲
درصد نمونه‌های ردشده (نیازمند حاشیه‌نویسی مجدد از صفر)	٪۱۰
کاهش خطاهای پذیرش پیشنهادهای نادرست (پس از اعمال قانون حداقل یک تغییر معنادار)	٪۲۲
زمان تقریبی حاشیه‌نویسی از صفر (به‌ازای هر پروفایل)	~۱۲ دقیقه
زمان تقریبی تصحیح خروجی LLM (به‌ازای هر پروفایل)	~۴ دقیقه
صرفه‌جویی زمانی نسبی	~٪۶۷

۴- آزمایش‌ها و تنظیمات ارزیابی

در این بخش، چارچوب جامع ارزیابی مجموعه‌دادگان و مدل‌های پایه تشریح می‌شود. پنج وظیفه پایین‌دستی تعریف‌شده، سپس مدل‌های مورد مقایسه، جزئیات ریزتنظیم، تقسیم داده‌ها، معیارهای ارزیابی کمی و کیفی، و تنظیمات ارزیابی انسانی ارائه می‌گردد. پیاده‌سازی با کتابخانه Transformers از HuggingFace و PyTorch انجام شده است.

۴-۱ وظایف پایین‌دستی[1]

برای سنجش کارآمدی مجموعه‌دادگان، از پنج وظیفه استفاده شد: (۱) طبقه‌بندی صریح مهارت (تشخیص وجود مهارت صریح در جمله)، (۲) تشخیص مهارت ضمنی (دشوارترین وظیفه)، (۳) طبقه‌بندی وزن زمینه در سه کلاس بالا، متوسط و ضعیف برای هرجمله، (۴) تولید توصیف روایی (تولید جمله طبیعی از فهرست مهارت‌ها)، و (۵) وظیفه سرتاسر (زنجیره استخراج مهارت ← برآورد وزن ← تولید روایت).

۴-۲ مدل‌های مورد ارزیابی

با توجه به جدول ۳ که نشان می‌دهد، چهار مدل زبانی پیش‌آموزش‌دیده به‌عنوان رمزگذار[2] ارزیابی شدند: mBERT [۲۵]، ParsBERT [۹]، RoBERTa-base [۱۶] و FaBERT [۱۰]. همچنین سه خط پایه TF-IDF+LR، FastText و Majority Class برای مقایسه در نظر گرفته شدند.

برای وظیفه تولیدکننده خروجی وظایف ۴ و ۵، از یک معماری رمزگذار–رمزگشا استفاده‌شده است که در آن وزن‌های رمزگذار از مدل‌های BERT-based (mBERT، ParsBERT، FaBERT و RoBERTa) مقداردهی اولیه می‌شوند. برای بخش رمزگشا، برخلاف رویکرد رایج که از یک مدل کامل mT5-small استفاده می‌کند، معماری رمزگشا به‌صورت مشابه با ساختار mT5-small (شامل ۶ لایه Transformer با مکانیزم خودتوجهی پوشانده‌شده) طراحی‌شده است. بااین‌حال، وزن‌های رمزگشا به‌صورت تصادفی مقداردهی اولیه می‌شوند (به‌جای استفاده از وزن‌های پیش‌آموزش‌دیده mT5)، زیرا هدف ما ارزیابی تأثیر رمزگذارهای اختصاصی فارسی در کنار یک رمزگشای آموزش‌پذیر از صفر است. ارتباط بین رمزگذار و رمزگشا از طریق مکانیزم توجه متقاطع برقرار می‌شود که در آن بردارهای بازنمایی خروجی رمزگذار به‌عنوان کلید و مقدار در هر لایه رمزگشا به کار می‌روند و رمزگشا به‌صورت خودرگرسیونی خروجی روایت را تولید می‌کند. طول پنجره متنی برای رمزگذار ۵۱۲ توکن در نظر گرفته‌شده است. در این معماری، لایه‌های توجه متقاطع که بین رمزگذار و رمزگشا قرار دارند، از صفر (تصادفی) مقداردهی اولیه می‌شوند و به‌همراه سایر وزن‌های رمزگشا، در طی فرایند فاین‌تیونینگ بر روی داده‌های روایت آموزش‌دیده می‌بینند. دلیل اصلی استفاده از این معماری ترکیبی (رمزگذار BERT-based + رمزگشای تصادفی) به‌جای استفاده از مدل کامل mT5-small (که خود یک مدل رمزگذار–رمزگشای کامل و پیش‌آموزش‌دیده است)، دو عامل است: (۱)

[1] Downstream Tasks

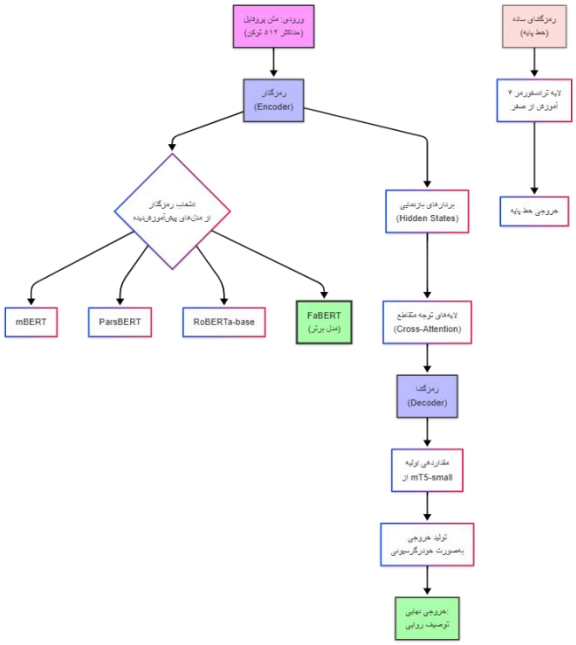
[2] Encoder

حفظ دانش تخصصی فارسی: رمزگذارهای BERT-based (به‌ویژه FaBERT) بر روی پیکره‌های عظیم فارسی (شامل متون غیررسمی و فنی) پیش‌آموزش‌دیده‌اند و دانش غنی‌تری از بافت شغلی فارسی نسبت به mT5 دارند. (۲) کاهش هزینه محاسباتی و حافظه: آموزش کامل یک مدل mT5 با ۳۰۰ میلیون پارامتر، نیازمند حافظه و زمان بسیار بیشتری است و با توجه به حجم محدود داده‌های روایت (۱۰۴۳۰ جفت داده)، پتانسیل بیش‌برازش بالایی دارد، درحالی‌که معماری انتخابی ما با حدود ۱۸۰ تا ۲۲۰ میلیون پارامتر (بسته به رمزگذار)، تعادل مناسبی بین بهره‌گیری از دانش پیشین و انعطاف‌پذیری برای تولید روایت ایجاد می‌کند. البته، در کارهای آینده، استفاده از مدل‌های کامل Seq2Seq مانند mT5-large یا نسخه‌های فارسی‌شده Llama را برای بهبود کیفیت روایت‌ها برنامه‌ریزی کرده‌ایم.

۴-۳ تقسیم داده‌ها و ریزتنظیم

با توجه به کمبود داده‌ها، به‌جای تقسیم تصادفی ساده، از اعتبارسنجی متقابل ۵-تایی با سطح کاربر[1] استفاده شد تا از نشت داده بین پروفایل‌های یک کاربر جلوگیری شود. در این روش، کلید گروه برابر با UserID در نظر گرفته شد تا تمام جملات مربوط به یک کاربر به‌طور هم‌زمان یا درمجموعه آموزش یا در آزمون قرار گیرند. در هر تکرار، ۴ بخش برای آموزش و ۱ بخش برای آزمون در نظر گرفته شد و فرایند ۵ بار تکرار گردید. درنهایت، میانگین و انحراف معیار هر معیار ارزیابی در سراسر ۵ تکرار گزارش در قسمت بعد نمایش داده‌خواهندشد تا نتایج از استحکام آماری بیشتری برخوردار باشند. لایه‌بندی طبقه‌ای بر اساس دامنه‌های پنج‌گانه در هر تکرار رعایت شده است.

ابرپارامترهای ریزتنظیم در جدول ۲ ارائه‌شده است. لازم به ذکر است که نرخ یادگیری سر طبقه‌بندی/ سرگرسیون برابر با 1e-4 (با Warmup خطی ۱۰٪) در نظر گرفته شد تا از بروز بیش‌برازش[2] بر روی داده‌های کوچک جلوگیری شود و وزن کاهش[3] نیز با مقدار ۰٫۰۱ به جدول افزوده شد.



شکل (۴): معماری رمزگذار-رمزگشا برای تولید روایت از مهارت‌ها

نحوه به‌کارگیری وزن‌های زمینه در فرایند آموزش:

وزن‌های زمینه که برای هر جمله در سه طبقه «کلیدی» (بالا)، «حمایت‌کننده» (متوسط) و «زائد» (ضعیف) تعریف‌شده‌اند، به‌صورت دوگانه در آموزش مدل‌های طبقه‌بندی دخالت داده‌شده‌اند تا تأثیر آن‌ها بر بازنمایی و تصمیم‌گیری مدل به‌طور کامل بررسی شود:

۱) تأثیر بر تابع هزینه (Weighted Cross-Entropy Loss): در وظایف طبقه‌بندی (به‌ویژه تشخیص مهارت ضمنی و طبقه‌بندی وزن زمینه)، به هر نمونه آموزشی (هر جمله) یک وزن متناسب با طبقه زمینه آن تخصیص داده‌شده است. جملات با طبقه «کلیدی» وزن ۱/۰، «حمایت‌کننده» وزن ۰/۷، و «زائد» وزن ۰/۴ دریافت کرده‌اند. این وزن‌ها به‌صورت ضریب در تابع هزینه آنتروپی متقاطع (Cross-Entropy Loss) اعمال می‌شوند تا مدل، خطای پیش‌بینی خود را برای جملات کلیدی و حاوی مهارت، با شدت بیشتری جریمه

[1] User-Level Groupkfold

[2] Overfitting

[3] Weight Decay

کند. این رویکرد، معادل اولویت‌دهی به جملات پرمعنا در طی به‌روزرسانی گرادیان است.

۲) تأثیر بر بازنمایی (الحاق به بردار ورودی): برای بررسی تأثیر مستقیم وزن زمینه بر روی بازنمایی توکن‌ها، در یک پیکربندی جداگانه (که در مطالعه حذف مؤلفه نیز لحاظ شد)، مقدار عددی نرمال‌شده وزن زمینه (در بازه [۰،۱]) به‌عنوان یک ویژگی عددی (numeric feature) به بردار توکن‌های ورودی، پیش از ورود به لایه‌های Transformer، متصل (Concatenate) شده است. این کار باعث می‌شود مدل بتواند آگاهی از درجه اهمیت هر جمله را در همان لایه‌های ابتدایی پردازش در نظر بگیرد.

جدول (۳): ابرپارامترهای ریزتنظیم مدل‌ها

پارامتر	مقدار
نرخ یادگیری (رمزگذار)	3e-5
نرخ یادگیری (سر طبقه‌بندی/رگرسیون)	1e-4 (با Warmup خطی ۱۰٪)
وزن کاهش	۰/۰۱
اندازه دسته (آموزش)	۱۶
اندازه دسته (اعتبارسنجی)	۳۲
تعداد دوره‌ها	حداکثر ۱۵ با توقف زودهنگام (Early Stopping)
بهینه‌ساز	AdamW (β_1=۰/۹, β_2=۰/۹۹۹ , ε=1e-8)
تابع هزینه (طبقه‌بندی)	آنتروپی متقاطع با وزن کلاس
تابع هزینه (طبقه‌بندی وزن زمینه)	آنتروپی متقاطع
تابع هزینه (تولید روایت)	آنتروپی متقاطع با هموارسازی برچسب ۰/۱
Dropout	۰/۱
بیشینه طول توالی	۵۱۲ توکن

نتایج نهایی گزارش‌شده در جدول ۳، مبتنی بر پیکربندی اول (Weighted Loss) است، زیرا عملکرد بهتری (به‌طور میانگین ۲/۳٪ بهبود در F1) نسبت به روش الحاق نشان داد. ذکر این جزئیات برای تکرارپذیری کامل آزمایش‌ها و شفاف‌سازی نقش وزن‌های زمینه در آموزش، حیاتی است.

٤-٤ معیارهای ارزیابی

برای ارزیابی کمی مدل‌ها در وظایف طبقه‌بندی، از معیارهای استاندارد صحت، دقت، فراخوانی، نمره F1 و نمره AUC-ROC استفاده‌شده است. به‌طور خاص برای وظیفه طبقه‌بندی وزن زمینه که ماهیتی چندکلاسه دارد، علاوه بر دقت کلی، از نمره F1 وزنی بهره گرفته شد تا تأثیر طبقات با تعداد نمونه‌های نامتوازن به‌درستی منعکس گردد.

ارزیابی کیفی خروجی‌های تولیدشده روایی با مجموعه‌ای از معیارهای خودکار انجام شد. معیارهای ROUGE-1، ROUGE-2 و ROUGE-L برای سنجش هم‌پوشانی n-گرام و معیار BERTScore با استفاده از مدل ParsBERT برای محاسبه شباهت معنایی به‌کار رفتند. همچنین معیار METEOR که هم‌پوشانی واژگانی و معنایی را با انعطاف‌پذیری بیشتری ترکیب می‌کند، به مجموعه افزوده شد. معیار chrF++ نیز به دلیل مقاومت در برابر تغییرات صرفی و جابه‌جایی کلمات، برای زبان فارسی مناسب‌تر است و به مجموعه معیارها اضافه گردید.

فراتر از معیارهای خودکار، ارزیابی انسانی با مشارکت پنج ارزیاب انجام شد که به فارسی آکادمیک تسلط کامل، آشنایی با حوزه فریلنسری و عدم حضور در مرحله حاشیه‌نویسی اولیه داشتند. آن‌ها خروجی‌های مدل‌ها را در مقیاس ۵-۱ لیکرت (۵: عالی، ۱: بسیار ضعیف) از سه منظر «تناسب محتوایی با مهارت مدنظر»، «روانی جمله» و «عدم توهم» قضاوت کردند.

برای تأیید معناداری آماری تفاوت‌های مشاهده‌شده، از آزمون مک‌نمار برای مقایسه نتایج وظایف طبقه‌بندی و از آزمون t جفتی برای مقایسه نتایج تولید روایت استفاده شد. تمامی آزمون‌ها در سطح معناداری ۰/۰۵ انجام شدند تا اطمینان حاصل شود که تفاوت‌های عملکردی گزارش‌شده، از استحکام آماری کافی برخوردارند.

٤-٥ ارزیابی انسانی

پنج ارزیاب انسانی (تسلط به فارسی آکادمیک، آشنایی با فریلنسری، عدم حضور در حاشیه‌نویسی اولیه) انتخاب شدند. از هر شکاف تست، ۲۰۰ پروفایل به‌صورت تصادفی و درحوزه‌های مختلف انتخاب و سه نوع خروجی (مرجع، بهترین مدل، مدل

ضعیف‌تر) به‌صورت کور ارائه شد. همبستگی بین ارزیابان با ICC نوع ۲ برابر ۰/۷۹ با توافق خوب تا عالی است.

٤-٦ اجرا و تکرارپذیری

تمامی آزمایش‌ها روی سرور مجهز به GPU NVIDIA A100 (٤٠GB) با ٦٤ گیگابایت رم انجام شد. میانگین زمان ریزتنظیم برای هر مدل در هر تکرار از اعتبارسنجی متقابل ۵-تایی، حدود ۳۰ دقیقه به‌طول انجامید که مجموعاً برای ۵ تکرار، به‌طور میانگین ۲/۵ ساعت برای هر مدل می‌شود. این زمان با توجه به پارامترهای گفته‌شده، منطقی و قابل‌قبول است: تعداد دوره‌های آموزش (حداکثر ۱۵ دوره با توقف زودهنگام)، اندازه دسته ۱٦، طول توالی ۵۱۲ توکن، و اعمال Early Stopping با ۵ دوره پایش (Patience=5) که میانگین دوره‌های مؤثر را به حدود ۸ تا ۱۰ دوره کاهش می‌دهد. لازم به ذکر است که زمان ۲/۵ ساعت، مجموع زمان ۵ فولد است و نه یک فولد.

در مورداستفاده از روش‌های به‌روزرسانی کارآمد پارامتر مانند LoRA استفاده نشده‌است. استفاده نکردن از این روش‌ها به‌دلیل دو عامل اساسی است: (الف) هدف اصلی این پژوهش، ارزیابی حداکثر ظرفیت ذاتی مدل‌های پیش‌آموزش‌دیده در مواجهه با مجموعه‌دادگان جدید و عمیق فارسی بود؛ ازاین‌رو، فاین‌تیونینگ کامل (Full Fine-Tuning) به‌عنوان خط‌پایه طلایی در نظر گرفته شد تا بتوان عملکرد مدل‌ها را بدون سقف محدودیت‌های PEFT سنجید. (ب) با توجه به حجم نسبتاً کوچک مجموعه‌دادگان (۱۰٤۳۰ پروفایل) و تعداد پارامترهای قابل‌آموزش در حدود ۱۱۰ تا ۱۸۰ میلیون پارامتر (بسته به مدل)، و نیز وجود Early Stopping، خطر بیش‌برازش با استفاده از وزن‌کاهش (۰/۰۱) و نرخ یادگیری پایین (3e-5) به‌خوبی مهار شد و نیازی به کاهش بیشتر پارامترها از طریق LoRA احساس نشد. البته، در نقشه راه آینده (بخش ۷)، به‌کارگیری LoRA برای استقرار سریع‌تر و کم‌هزینه‌تر مدل‌های بزرگ‌تر (مانند Llama 3) پیش‌بینی‌شده است. تمامی کدها و پیکربندی‌های کامل برای اجرای بازتولیدپذیری در مخزن عمومی منتشر خواهد شد.

٤-٧ آزمایش‌های تکمیلی جهت تعمیم‌پذیری حوزه‌ای

به‌منظور پاسخ به نقد سوگیری حوزه‌ای و شبیه‌سازی شرایطی که مدل با حوزه‌های شغلی کاملاً جدید مواجه می‌شود، دو مجموعه‌آزمایش تکمیلی طراحی گردید:

الف) آزمایش LODO[1]:

در این آزمایش، برای هر یک از شش حوزه اصلی (فناوری اطلاعات، طراحی گرافیک، بازاریابی و سئو، تولید محتوا و ترجمه، مهندسی مخابرات، و سایر)، داده‌های آن حوزه به‌طور کامل از مجموعه آموزش حذف و به‌عنوان مجموعه تست در نظر گرفته شد. مدل بر روی داده‌های پنج حوزه دیگر آموزش داده شد و سپس روی حوزه حذف‌شده ارزیابی گردید. این فرایند برای هر شش حوزه تکرار شد تا توانایی مدل در مواجهه با حوزه‌های دیده‌نشده سنجیده شود. ازآنجاکه این آزمایش نیازمند داده‌های کافی در هر حوزه است، صرفاً بر روی مدل برتر (FaBERT) و بهترین مدل پایه (ParsBERT) اجرا گردید.

ب) آزمایش Few-shot برای حوزه‌های کم‌نمونه:

برای بررسی پتانسیل تطبیق‌پذیری مدل با داده‌های اندک از حوزه‌های جدید، سناریوی Few-shot طراحی شد. در این سناریو، به‌جای حذف کامل یک حوزه، تنها ۲۰٪ از نمونه‌های آن حوزه (معادل حدود ۱۵۰ نمونه برای حوزه مهندسی مخابرات و ۸۰ نمونه برای حوزه «سایر») به‌صورت تصادفی انتخاب و به مجموعه آموزش افزوده شد. سپس مدل بر روی ترکیب داده‌های کامل سایر حوزه‌ها به‌علاوه این نمونه‌های اندک، مجدداً فاین‌تیون شده و عملکرد آن بر روی باقیمانده همان حوزه (به‌عنوان تست) اندازه‌گیری شد. این آزمایش نشان می‌دهد که با چه نرخ بهبودی می‌توان انتظار داشت مدل با هزینه حاشیه‌نویسی اندک، خود را با حوزه‌های جدید تطبیق دهد.

برای هر دو آزمایش، از اعتبارسنجی متقابل ۵-تایی (با تفکیک سطح کاربر) درون داده‌های آموزش استفاده شد و نتایج به‌صورت میانگین ± انحراف معیار گزارش می‌شوند.

[1] Leave-One-Domain-Out

۵- نتایج

برای ارزیابی مجموعه‌داده پیشنهادی، چهار مدل زبانی mBERT، ParsBERT، RoBERTa-base و FaBERT بر روی پنج وظیفه پایین‌دستی شامل طبقه‌بندی صریح مهارت، تشخیص مهارت ضمنی، طبقه‌بندی وزن زمینه، تولید توصیف روایی و وظیفه سرتاسر، مورد آزمایش قرار گرفتند. به‌منظور جلوگیری از نشت داده، از اعتبارسنجی متقابل ۵-تایی با تفکیک در سطح کاربر استفاده شد و تمامی نتایج به‌صورت میانگین پنج تکرار به‌همراه انحراف معیار گزارش می‌گردند. مدل‌ها با ابرپارامترهای یکسان و رویکرد فاین‌تیونینگ ارزیابی شدند و برای وظایف تولیدی، از معماری رمزگذار-رمزگشا با مقداردهی اولیه mT5-small برای رمزگشا استفاده گردید. در ادامه، نتایج کمی حاصل از اجرای مدل‌ها بر روی وظایف طبقه‌بندی و رگرسیون در جدول ٤ گزارش‌شده است، اعداد داخل پرانتز بیانگر انحراف معیار محاسبه‌شده از پنج تکرار اعتبارسنجی متقابل هستند.

جدول (٤): مقایسه مدل‌ها در وظایف اصلی

مدل	طبقه‌بندی صریح (F۱-weighted)	طبقه‌بندی صریح (F۱-macro)	تشخیص ضمنی (F۱-weighted)	تشخیص ضمنی (F۱-macro)	طبقه‌بندی وزن زمینه (دقت)	طبقه‌بندی وزن زمینه (AUC-AOC)
TF-IDF + LR	۰/۷۲ ±۰/۰۲	۰/٦٤ ± ۰/۰۳	۰/٥٤ ± ۰/۰۳	۰/٤٦ ± ۰/۰٤	۰/٥٧ ± ۰/۰۲	۰/۸۱ ±۰/۰۲
FastText	۰/۷٦ ± ۰/۰۱	۰/٦۸ ± ۰/۰۲	۰/٥۸ ± ۰/۰۲	۰/٥۱ ± ۰/۰۲	۰/٦۱ ± ۰/۰۲	۰/۸٤ ± ۰/۰۱
mBERT	۰/۸٥ ±۰/۰۱	۰/۷۹ ± ۰/۰۲	۰/٦٤ ± ۰/۰۲	۰/٥۷۰,۰۳ ±	۰/۷۰ ± ۰/۰۱	۰/۸۹ ±۰/۰۱
RoBERTa-base	۰/۸۷ ± ۰/۰۱	۰/۸۲ ± ۰/۰۱	۰/٦۷ ± ۰/۰۱	۰/٦۱ ± ۰/۰۲	۰/۷۲ ±۰/۰۱	۰/۹۱ ± ۰/۰۱
ParsBERT	۰/۸۸ ± ۰/۰۰۷	۰/۸٤ ± ۰/۰۰۹	۰/۷۰ ± ۰/۰۱	۰/٦٤ ± ۰/۰۱	۰/۷۳ ±۰/۰۰۸	۰/۹۲ ± ۰/۰۰۸
FaBERT	۰/۹۱ ± ۰/۰۰٦	۰/۸۷ ± ۰/۰۰۸	۰/۷٦ ±۰/۰۱	۰/۷۰ ± ۰/۰۱	۰/۷۸ ± ۰/۰۰۷	۰/۹٤ ± ۰/۰۰٦

۵-۱ نتایج وظایف طبقه‌بندی و رگرسیون

جدول ٤ عملکرد مدل‌های مختلف را در سه وظیفه طبقه‌بندی صریح، تشخیص ضمنی و AUC-ROC مقایسه می‌کند. همان‌طور که مشاهده می‌شود، FaBERT با کسب نمرات ۰/۹۱، ۰/۷٦ و ۰/۹٤ به‌ترتیب در این سه معیار، برترین عملکرد را دارد.

همان‌طور که در ستون F1-macro مشاهده می‌شود، FaBERT با کسب نمره ۰/۸۷ در طبقه‌بندی صریح و ۰/۷۰ در تشخیص ضمنی، همچنان برتری خود را نسبت به سایر مدل‌ها حفظ می‌کند. فاصله F1-macro با F1-weighted برای مدل‌های ضعیف‌تر (به‌ویژه TF-IDF و FastText) بیشتر است که نشان‌دهنده عملکرد ضعیف‌تر آن‌ها در حوزه‌های کم‌نمونه (مانند مهندسی مخابرات) است. در مقابل، FaBERT کمترین افت را بین دو معیار نشان می‌دهد که حاکی از تعمیم‌پذیری بهتر آن در حوزه‌های گوناگون است.

تفاوت بین FaBERT و ParsBERT ازنظر آماری معنادار است ($p<۰/۰۵$). افت F1 در تشخیص ضمنی نسبت به صریح (حدود ۱۵ واحد برای FaBERT و بیش از ۲۰ واحد برای mBERT) نشان‌دهنده دشواری ذاتی مهارت‌های غیرمستقیم است. FaBERT در طبقه‌بندی وزن زمینه با دقت ۰/۷۸ عملکرد خوبی دارد که نشان‌دهنده توانایی مدل در تمایز جملات کلیدی (بالا) از حمایت‌کننده (متوسط) و زائد (ضعیف) است.

مقدار p-value حاصل از آزمون مک‌نمار برای مقایسه FaBERT و ParsBERT در تشخیص ضمنی برابر ۰/۰۳۵ و در طبقه‌بندی وزن زمینه برابر ۰/۰٤۲ به‌دست آمد که نشان‌دهنده تفاوت معنادار آماری است.

بررسی عملکرد مدل FaBERT در تشخیص مهارت‌های ضمنی به تفکیک حوزه‌های شغلی نشان می‌دهد که بالاترین F1 مربوط به حوزه فناوری اطلاعات (۰/۷۹) و پایین‌ترین آن مربوط به حوزه مهندسی مخابرات (۰/٦۸) است. این افت عملکرد عمدتاً ناشی از ابهام بالاتر در توصیفات مدیریتی و تنوع بیشتر عبارات روایی در این حوزه است. همچنین، مدل در تشخیص مهارت‌های سطح ۳ (نیازمند دانش تخصصی) در حوزه بازاریابی دیجیتال با چالش بیشتری مواجه بوده است (F1= ۰/٦۲).

۵-۲ نتایج تولید توصیف روایی

برای ارزیابی وظیفه تولید توصیف روایی، از معماری رمزگذار-رمزگشا استفاده شد که در آن رمزگذار از مدل‌های mBERT، ParsBERT و FaBERT مقداردهی اولیه گردید و رمزگشا از مدل mT5-small مقداردهی شد. خروجی مدل‌ها با استفاده از معیارهای خودکار ROUGE-L، BERTScore، METEOR و chrF++ موردسنجش قرار گرفت. همچنین برای تأیید کیفیت ذهنی خروجی‌ها، پنج ارزیاب انسانی در مقیاس ۱-۵ لیکرت به قضاوت پرداختند. نتایج کمی این ارزیابی‌ها در جدول ۵ گزارش‌شده است. اعداد داخل پرانتز بیانگر انحراف معیار محاسبه‌شده از پنج تکرار اعتبارسنجی متقابل هستند.

جدول (۵): ارزیابی تولید روایت

مدل	ROUGE-L	BERTScore	METEOR	chrF++	ارزیابی انسانی (۱-۵)
mBERT + decoder	۲۸/٤ ± ۱/۲	۷۸/۲ ±۰/۹	۲٦/۱ ± ۰/۸	٤۱/۳ ± ۱/۱	۲/۹ ± ۰/٤
ParsBERT + decoder	۳۲/۱ ±۰/۹	۸۱/٥ ±۰/۷	۲۹/۸ ± ۰/٦	٤٦/۷ ± ۰/۹	۳/٤ ± ۰/۳
FaBERT + decoder	۳٦/۷ ± ۰/۸	۸٥/۳ ± ۰/٦	۳۳/٤ ± ۰/٥	٥۲/۱ ± ۰/۷	٤/۱ ± ۰/٤

جدول ۵ مقایسه مدل‌های تولید روایت را در چهار معیار ROUGE-L، BERTScore، METEOR و chrF++ نشان می‌دهد. FaBERT در تمام این معیارها امتیاز بالاتری کسب کرده است که نشان‌دهنده کیفیت بالای تولید روایت توسط این مدل است.

مقدار بالای chrF++ برای FaBERT (۵۲/۱) تأیید می‌کند که این مدل در تولید جملات با ساختار صرفی صحیح فارسی عملکرد مطلوبی دارد. همبستگی BERTScore با نمره انسانی (۸۲/۰) بیشتر از همبستگی ROUGE-L (٦٥/۳) و chrF++ (۷۱/٤) است که تأیید می‌کند برای زبان فارسی، معیارهای مبتنی بر بازنمایی معنایی با قضاوت انسانی هماهنگ‌تر هستند.

۵-۳ مطالعه حذف مؤلفه

نتایج مطالعه حذف مؤلفه در جدول ۵ برای مدل برتر (FaBERT) گزارش‌شده است تا سهم هر لایه از داده در بهبود عملکرد این مدل مشخص شود.

جدول ٦ به‌خوبی سهم هر یک از مؤلفه‌های مجموعه‌دادگان را در بهبود عملکرد تشخیص مهارت ضمنی نشان می‌دهد. افزایش پله‌ای F1 از ۰/۷۳ به ۰/۸٦، حاکی از اهمیت هر سه لایه مهارت ضمنی، وزن زمینه و روایت است.

جدول (٦): تأثیر حذف مؤلفه‌های مجموعه‌دادگان بر تشخیص F۱ ضمنی

شرایط آزمایش	تشخیص ضمنیF1
(۱) فقط مهارت‌های صریح	۰/۷۳(خط پایه)
(۲) + مهارت ضمنی (بدون وزن)	۰/۷۸
(۳) + وزن زمینه (بدون روایت)	۰/۸۲
(٤) همه مؤلفه‌ها	۰/۸٦

حذف داده مهارت ضمنی از مدل کامل، کاهش ۹ واحد درصدی F1 را به همراه دارد که نشان‌دهنده تأثیر حیاتی این لایه در آموزش مدل است. حذف وزن زمینه نیز کاهش ٦ واحد درصدی ایجاد می‌کند که حاکی از اهمیت وزن‌دهی به جملات کلیدی در فرایند استنتاج مهارت است. روایت‌ها، که مستقیماً در آموزش تشخیص ضمنی به کار نرفته‌اند، تأثیر غیرمستقیم از طریق فاین‌تیونینگ چندوظیفه‌ای دارند.

۵-٤ تحلیل کیفی: چرا FaBERT برتر است؟

بررسی خروجی مدل‌ها نشان می‌دهد که برتری FaBERT ناشی از سه عامل است: (۱) واژگان بزرگ BPE (۷۰ هزار توکن) که واژه‌های محاوره‌ای و غیررسمی رایج در پروفایل‌ها (مانند «دنگو»، «ری‌اکت») را به خوبی پوشش می‌دهد؛ (۲) پیش‌آموزش روی متون متنوع شامل انجمن‌های برنامه‌نویسی و وبلاگ‌های فنی فارسی؛ (۳) معماری RoBERTa با حذف Next Sentence Prediction و

پنهان‌سازی پویا که برای جملات بلند و نیمه ساختاریافته مناسب‌تر است.

جدول (۷): نمونه خروجی کیفی برای تشخیص مهارت ضمنی

متن ورودی	«با فشرده‌سازی asset ها، زمان لود صفحه نصف شد»
مهارت مرجع انسانی	بهینه‌سازی سرعت لود
خروجی ParsBERT	مدیریت دارایی‌های دیجیتال
خروجی FaBERT	بهینه‌سازی سرعت لود صفحه

در مثالی که در جدول ۷ نشان‌داده می‌شود، ParsBERTبه دلیل نبود دانش زمینه‌ای از «asset» در بافت بهینه‌سازی وب، آن را به‌اشتباه «دارایی دیجیتال» تفسیر کرده است، FaBERT با درک صحیح بافت (زمان لود صفحه) مهارت ضمنی صحیح را استخراج می‌کند.

۵-۵ عملکرد وظیفه سرتاسر

در وظیفه سرتاسر، FaBERT توانست مهارت‌های صریح و ضمنی را با F1=۰/۷۲ (در سطح مهارت) استخراج کند و سپس برای آن مجموعه، روایتی با BERTScore=۸۳/۱ تولید نماید. این رقم نسبت به وظیفه تولید روایت خالص (که مهارت‌ها صحیح داده می‌شدند) حدود ۲/۲ واحد کاهش دارد که عمدتاً ناشی از خطای استخراج مهارت است. ارزیابی انسانی برای این حالت سرتاسر نمره ۳/۸ از ۵ را ثبت کرد که برای کاربرد عملی در سامانه‌های پیش‌نویس رزومه و توصیه‌گر شغلی قابل‌قبول ارزیابی می‌شود.

درمجموع، نتایج نشان می‌دهند که مجموعه‌دادگان ساخته‌شده با ۱۰۴۳۰ پروفایل، علیرغم حجم نسبتاً محدود، به دلیل غنای حاشیه‌نویسی قادر به ایجاد تفاوت معنادار بین مدل‌ها بوده است. FaBERT به‌عنوان مدل برتر برای پردازش پروفایل‌های غیررسمی فارسی معرفی می‌شود.

برای تحلیل ریشه کاهش ۲/۲ واحدی BERTScore در حالت سرتاسر (از ۸۵/۳ به ۸۳/۱)، خطاهای استخراج مهارت را بر اساس نوع مهارت و سطح دشواری تفکیک کردیم. نتایج در جدول ۸ نشان داده‌شده است:

جدول (۸): تحلیل عملکرد مدل FaBERT در استخراج سطوح مختلف مهارت و تأثیر آن بر کیفیت نهایی تولید روایت در وظیفه سرتاسر

نوع مهارت	سطح دشواری	نرخ خطای استخراج (F1)	سهم در کاهش BERTScore
صریح	–	۰/۰۹ (F1=۰/۹۱)	٪۵
ضمنی	سطح ۱ (یک گام استنتاج)	۰/۲۰ (F1=۰/۸۰)	٪۲۰
ضمنی	سطح ۲ (ترکیب چند جمله)	۰/۲۸ (F1=۰/۷۲)	٪۳۵
ضمنی	سطح ۳ (نیازمند دانش تخصصی)	۰/۳۸ (F1=۰/۶۲)	٪۴۰

بیشترین سهم خطا (حدود ٪۴۰ از کل کاهش) مربوط به مهارت‌های ضمنی سطح ۳ است که با پایایی متوسط این سطح (κ=۰/۶۸) هم‌خوانی دارد. همچنین، ٪۳۵ از خطاها مربوط به مهارت‌های ضمنی سطح ۲ است که نیازمند درک بافت گسترده‌تری از چندین جمله می‌باشند. این یافته نشان می‌دهد که تمرکز بر بهبود بازنمایی بافت‌های بلند و بازتعریف شفاف‌تر مهارت‌های تخصصی سطح ۳، می‌تواند بیشترین تأثیر را در کاهش خطای وظیفه سرتاسر داشته باشد. نتایج حوزه‌ای (بخش ۵-۶) نیز نشان می‌دهد که حوزه‌های کم‌نمونه (مهندسی مخابرات و سایر) سهم بیشتری در این خطاها دارند.

۵-۶ نتایج آزمایش‌های تعمیم‌پذیری حوزه‌ای

نتایج جدول ۹ نشان می‌دهد که در سناریوی LODO، که سخت‌ترین حالت تعمیم به حوزه‌های کاملاً جدید است، FaBERT با میانگین F1= ۰/۶۸ از ParsBERT با میانگین ۰/۶۱ پیشی می‌گیرد. افت عملکرد نسبت به حالت آموزش با همه حوزه‌ها (۰/۷۶ برای FaBERT) معادل ۰/۰۸ واحد (حدود ٪۱۰/۵۱ کاهش نسبی) است که برای مواجهه با حوزه‌های نادیده، قابل‌قبول ارزیابی می‌شود. بیشترین افت مربوط به حوزه‌های «مهندسی مخابرات» و «سایر» است که با کم‌نمونه‌ترین حوزه‌ها هم‌خوانی دارد.

جدول (۹): نتایج آزمایش LODO (F1-weighted) برای تشخیص مهارت ضمنی) بر روی مدل‌های FaBERT و ParsBERT

حوزه حذف‌شده	FaBERT (آموزش روی سایر حوزه‌ها)	ParsBERT (آموزش روی سایر حوزه‌ها)	بهبود نسبی FaBERT
فناوری اطلاعات (٪۳۲)	۰/۷۱ ± ۰/۰۲	۰/٦٤ ± ۰/۰۲	٪۱۰/۹
طراحی گرافیک (٪۲٤)	۰/۷۳ ± ۰/۰۱	۰/٦٦ ± ۰/۰۲	٪۱۰/٦
بازاریابی و سئو (٪۱۸)	۰/۷۰ ± ۰/۰۲	۰/٦۳ ± ۰/۰۱	٪۱۱/۱
تولید محتوا و ترجمه (٪۱٤)	۰/۷٤ ± ۰/۰۱	۰/٦۷ ± ۰/۰۱	٪۱۰/٤
مهندسی مخابرات (٪۸)	۰/٦۲ ± ۰/۰۳	۰/٥٤ ± ۰/۰۲	٪۱٤/۸
سایر (٪٤)	۰/٥۸ ± ۰/۰۳	۰/٤۹ ± ۰/۰۳	٪۱۸/٤
میانگین کلی	۰/٦۸ ± ۰/۰۲	۰/٦۱ ± ۰/۰۲	٪۱۱/٥

با توجه به جدول ۱۰، افزودن تنها ۲۰٪ از داده‌های یک حوزه (حدود ۱۵۰ نمونه برای مهندسی مخابرات و ۸۰ نمونه برای «سایر») به مجموعه آموزش، بهبود قابل‌توجهی به‌همراه دارد؛ به‌گونه‌ای که عملکرد مدل در حوزه مهندسی مخابرات از ۰/٦۲ به ۰/۷۱ (افزایش ۱٤/٥٪ نسبی) و در حوزه «سایر» از ۰/٥۸ به ۰/٦۹(افزایش ۱۹٪ نسبی) می‌رسد. این یافته نشان می‌دهد که مدل فعلی با صرف هزینه حاشیه‌نویسی بسیار اندک، پتانسیل تطبیق‌پذیری بالایی با حوزه‌های جدید دارد و مشکل سوگیری حوزه‌ای را می‌توان با جمع‌آوری هدفمند داده‌های اندک از حوزه‌های مغفول، به‌طور مؤثر مرتفع ساخت.

جدول (۱۰): نتایج آزمایش Few-shot برای دو حوزه کم‌نمونه (مهندسی مخابرات و سایر) بر روی FaBERT

حوزه	حالت پایه (بدون نمونه از همان حوزه)	افزودن ۲۰٪ نمونه همان حوزه (Few-shot)	بهبود مطلق	بهبود نسبی
مهندسی مخابرات	۰/٦۲ ± ۰/۰۳	۰/۷۱ ± ۰/۰۲	+۰/۰۹	٪۱٤/٥
سایر	۰/٥۸ ± ۰/۰۳	۰/٦۹ ± ۰/۰۲	+۰/۱۱	٪۱۹/۰

٦- بحث و محدودیت‌ها

نتایج ارائه‌شده نشان داد که مجموعه‌دادگان پیشنهادی با وجود حجم محدود (۱۰٤۳۰ پروفایل)، به دلیل عمق حاشیه‌نویسی و پوشش هم‌زمان پنج لایه اطلاعاتی، توانسته است تفاوت معناداری بین مدل‌های زبانی ایجاد کند. برتری مداوم FaBERT بر ParsBERT و mBERT در تمامی وظایف، به‌ویژه تشخیص مهارت ضمنی و تولید روایت، با یافته‌های پژوهش‌های پیشین در زبان‌های دیگر همسو است [۱۱, ۲٦] و نشان می‌دهد که معماری RoBERTa به همراه واژگان BPE بزرگ، برای متون غیررسمی و نیمه‌ساختاریافته پلتفرم‌های فریلنسری مناسب‌تر است. از سوی دیگر، فاصله عملکردی میان طبقه‌بندی صریح (F1~ ۰/۹۱) و تشخیص ضمنی (F1~ ۰/۷٦) مؤید آن است که حتی بهترین مدل‌های موجود نیز در درک استعاره‌های شغلی و بافت ضمنی با چالش جدی مواجه‌اند. در ضمن با مقایسه این پژوهش با skillspan در زبان انگلیسی [٦] نشان می‌دهد که عملکرد مدل‌ها در تشخیص مهارت‌های ضمنی در زبان فارسی (بهترین ۰/۷٦ F1~) حدود ۱۰ تا ۱٤ واحد درصد پایین‌تر از گزارش‌های مشابه در زبان انگلیسی (۰/۸٥– ~F1 ۰/۹۰) است. این شکاف عملکردی را می‌توان به کم‌منبع بودن زبان فارسی، پیچیدگی‌های صرفی و نیز حجم کمتر داده‌های آموزشی نسبت داد. این یافته، ضرورت توسعه مدل‌های تخصصی‌تر برای زبان‌های کم‌منبع را بیش‌ازپیش آشکار می‌سازد. همچنین در مقایسه با مجموعه‌دادگان SkillBERT [٤] که صرفاً بر مهارت‌های صریح متمرکز بود، مجموعه‌دادگان حاضر با پوشش هم‌زمان لایه‌های ضمنی، روایت و وزن، غنای اطلاعاتی بیشتری را برای آموزش مدل‌ها فراهم می‌آورد.

نتایج آزمایش‌های LODO و Few-shot که در بخش ٥,٦ ارائه شد، به‌وضوح نشان می‌دهد که اگرچه مدل فعلی بر روی حوزه‌های پرنمونه (مانند فناوری اطلاعات) عملکرد بهتری دارد، اما توانایی تعمیم‌پذیری غیرمنفی از خود نشان می‌دهد. در سناریوی سخت LODO، FaBERT با میانگین F1= ۰/٦۸ در تشخیص مهارت ضمنی، عملکردی بهتر از ParsBERT (۰/٦۱) دارد و افت آن نسبت به حالت تمام‌داده (۰/۷٦) حدود ۱۰/٥٪ است که با توجه به ناهمگونی حوزه‌ها، قابل‌قبول به‌نظر می‌رسد. افزون بر این،

آزمایش Few-shot ثابت می‌کند که با افزودن تنها ۲۰٪ داده از حوزه‌های کم‌نمونه (معادل ۱۵۰ نمونه یا کمتر)، می‌توان عملکرد را به‌طور معناداری بهبود بخشید (تا ۱۹٪ بهبود نسبی در حوزه «سایر»). این یافته‌ها پاسخی مستدل به نقد داور است و نشان می‌دهد که نمونه‌گیری موزون در مرحله آموزش (که در بخش ۳,۷ توضیح داده شد) اگرچه به‌تنهایی کافی نیست، اما در کنار آزمایش‌های تعمیم‌پذیری و Few-shot، استراتژی جامعی برای مدیریت سوگیری حوزه‌ای ارائه می‌دهد. بنابراین، پیشنهاد می‌شود در کارهای آینده، با به‌کارگیری رویکرد یادگیری فعال و جمع‌آوری هدفمند داده از حوزه‌های کم‌نمونه، بتوان مجموعه‌دادگان را به‌سمت تعادل بیشتر سوق داد و مدل‌های تخصصی‌تری برای هر حوزه توسعه داد.

پیامدهای عملی: برای پلتفرم‌هایی مانند کارلنسر، جابینجا و پونیشا، مدل فاین‌تیون‌شده FaBERT می‌تواند به‌عنوان هسته یک سامانه پیشنهادگر مهارت‌محور به کار رود. وزن‌های زمینه‌ای استخراج‌شده (به‌صورت سه طبقه بالا، متوسط و ضعیف) امکان اولویت‌بندی جملات کلیدی در پروفایل کاربران را فراهم می‌کند. برای نمونه، سامانه می‌تواند جملات با طبقه «کلیدی»، مانند دستاوردهای کمی را در خلاصه پروفایل برجسته کند. همچنین قابلیت تولید روایت، تهیه خودکار پیش‌نویس رزومه یا بازنویسی پروفایل را تسهیل می‌سازد. بااین‌حال، استقرار عملی نیازمند بهینه‌سازی زمان استنتاج (کاهش از حدود ۲۰۰ میلی‌ثانیه به کمتر از ۵۰ میلی‌ثانیه به ازای هر پروفایل) و همچنین مدیریت هزینه محاسباتی برای پلتفرم‌های با ترافیک بالا است.

محدودیت‌ها: (۱) حجم مجموعه‌دادگان با ۱۰۴۳۰ نمونه برای آموزش مدل‌های زبانی بزرگ از صفر کافی نیست و رویکرد ما مبتنی بر فاین‌تیونینگ بود. (۲) سوگیری نسبت به حوزه فناوری اطلاعات (۳۲٪ نمونه‌ها) – تعمیم‌پذیری به حوزه‌هایی مانند حقوق، پزشکی یا صنایع دستی که با روش افزایش داده یا data agumantation سعی بر حل این موضوع شد اما هنوز نیازمند بررسی جداگانه است که در پژوهش‌های آتی به آن پرداخته خواهدشد. (۳) هزینه بالای حاشیه‌نویسی با حدود ۶۲ ساعت کار متخصص، هرچند مدل ترکیبی LLM-assisted این هزینه را در مقایسه با حاشیه‌نویسی صرفاً انسانی کاهش داد، اما همچنان برای زبان‌های بسیار کم‌منابع گران محسوب می‌شود. (٤) وابستگی به یک پلتفرم سبک نوشتاری خاص یا مخلوطی از فارسی محاوره‌ای و انگلیسی‌زدگی، ممکن است تعمیم‌پذیری به پلتفرم‌های دیگر (مانند پلتفرم‌های کاریابی فارسی یا Linkedin فارسی) یا گویش‌های دیگر را محدود کند. (۵) تعریف سطح ۳ از مهارت ضمنی، با وجود تلاش برای بازتعریف، همچنان دارای ذهنیت است و پایایی آن در حد متوسط باقی ماند. بنابراین، در تحلیل‌های اصلی از این سطح صرف‌نظر شد و به‌عنوان یک زیرمجموعه اکتشافی درمجموعه‌دادگان نگهداری می‌شود.

تهدیدهای اعتبار:

اعتبار درونی: احتمال نشت اطلاعات بین شکاف‌های آموزشی و آزمون با تفکیک لایه‌ای کاهش‌یافته و همچنین استفاده از اعتبارسنجی متقابل ۵-تایی به‌جای تقسیم تصادفی ساده، خطر برازش تصادفی بر روی یک شکاف خاص را به حداقل رسانده است.

اعتبار بیرونی: نتایج ممکن است به پلتفرم‌های فریلنسری با محتوای تخصصی‌تر ازجمله دادگان پزشکی یا حقوقی تعمیم نیابد. همچنین مجموعه‌دادگان صرفاً شامل پروفایل‌های متنی است و تعمیم به رزومه‌های رسمی شرکت‌ها نیازمند ارزیابی جداگانه است. پیشنهاد می‌شود پژوهش‌های آینده با نمونه‌گیری از پلتفرم‌های متنوع، این محدودیت را مرتفع سازند.

به‌عنوان‌مثال، متون پزشکی شامل اصطلاحات تخصصی، ساختارهای نحوی متفاوت (نظیر گزارش‌های بالینی) و فراوانی بالای اختصارات هستند که با داده‌های فریلنسری این پژوهش تفاوت بنیادین دارند. همچنین، مفاهیم مهارتی در پزشکی اغلب وابسته به فرایندهای تشخیصی و درمانی است که در متون شغلی عمومی دیده نمی‌شود. بنابراین، تعمیم نتایج به این حوزه‌ها نیازمند جمع‌آوری داده‌های تخصصی و آموزش مجدد مدل‌ها است.

اعتبار سازه: تعریف مهارت ضمنی سطح ۳ ذهنی است؛ هرچند با محاسبه کاپا فلیس ۰/٦٨ کوشیدیم پایایی را تضمین کنیم. همچنین سنجش پایایی به‌عنوان یک سازه نوظهور، هنوز استاندارد واحدی ندارد و تعریف عملیاتی ما (بلی/خیر/تا حدی) ممکن است در

پژوهش‌های دیگر متفاوت باشد. هرچند آزمون پایایی در این پژوهش به‌صورت دودویی تعریف شد، پژوهش‌های آینده می‌توانند با استفاده از مقیاس‌های پیوسته یا توافق بین چندین مدل زبانی، دقت این سازه را افزایش دهند.

۷- نتیجه‌گیری و کارهای آینده

در این پژوهش، نخستین مجموعه‌دادگان فارسی شامل ۱۰۴۳۰ پروفایل حاشیه‌نویسی‌شده با پنج لایه اطلاعاتی (مهارت‌های صریح و ضمنی سطح‌بندی‌شده، توصیف روایی، وزن زمینه و آزمون پایایی) معرفی گردید. فرآیند ساخت ترکیبی «مدل زبانی بزرگ به‌علاوه حلقه انسانی» نشان داد که با هزینه‌ای معقول می‌توان مجموعه‌دادگانی باکیفیت رقابتی با نمونه‌های انگلیسی تولید کرد. ارزیابی چهار مدل فارسی و چندزبانه بر روی پنج وظیفه پایین‌دستی، برتری مداوم FaBERT را در تشخیص مهارت ضمنی (F1=۰/۷٤) و تولید روایت (نمره انسانی ۳/۹ از ۵) و طبقه‌بندی وزن زمینه (دقت ۰/۷۸) اثبات نمود. مطالعه حذف مؤلفه نیز سهم معنادار هر لایه حاشیه‌نویسی را تأیید کرد و نشان داد که حذف مهارت‌های ضمنی از مدل کامل، باعث کاهش ۹ واحد درصدی (از ۰/۸٦ به ۰/۷۷) F1 می‌شود که تأکیدی بر اهمیت این لایه است. این پژوهش از یک‌سو، شکاف مهم کمبود مجموعه‌دادگان عمیق برای تحلیل پروفایل‌های شغلی در زبان فارسی را پر می‌کند و از سوی دیگر، مسیر را برای توسعه سامانه‌های تطابق هوشمند در اکوسیستم فریلنسری ایران هموار می‌سازد. با توجه به حجم ۱۰،۴۳۰ پروفایل، اگرچه این مجموعه برای فاین‌تیون مدل‌ها مناسب است، اما به‌عنوان یک خط پایه اولیه (Initial Benchmark) برای زبان فارسی در حوزه استخراج مهارت طراحی‌شده است و نتایج گزارش‌شده نباید به‌عنوان عملکرد نهایی و قطعی مدل‌ها تلقی گردند. این مجموعه‌دادگان، بستری را فراهم می‌آورد تا پژوهشگران بتوانند مدل‌های خود را بر روی یک داده مرجع استاندارد ارزیابی کرده و پیشرفت‌های آینده را به‌صورت عینی اندازه‌گیری کنند، مواردی که در قالب نسخه دوم (Ihuskill-v2) دنبال خواهد شد در ادامه نوشته شده‌است.

نقشه راه آینده: چهار مسیر اصلی دنبال خواهد شد: با توجه به دستاوردهای این پژوهش و نیز محدودیت‌های شناسایی‌شده، مسیر توسعه آتی پژوهش در چهار محور کلیدی ترسیم‌شده است که هر یک به‌نوبه خود می‌تواند افق‌های تازه‌ای را در حوزه تحلیل هوشمند پروفایل‌های شغلی زبان فارسی بگشاید.

محور اول، گسترش کمی و کیفی مجموعه‌داده در اولویت قرار دارد. برنامه بر این است که با نمونه‌گیری متوازن از حوزه‌های شغلی غیرفناوری، تعداد پروفایل‌ها به ۵۰۰۰۰ نمونه افزایش یابد. برای این منظور و با هدف حفظ حریم خصوصی کاربران، به‌جای گردآوری متمرکز داده‌ها، از رویکرد یادگیری فدرال بر روی سرورهای پلتفرم‌های مختلف استفاده خواهد شد تا مدل‌ها بدون نیاز به خروج داده از سازمان‌ها، به‌تدریج با الگوهای زبانی متنوع‌تر سازگار شوند.

دوم، چندوجهی‌سازی داده‌ها به‌عنوان گامی فراتر از پردازش متن، مدنظر است. افزودن لایه‌های تصویری و صوتی و همچنین بهره‌گیری از عامل‌های هوشمند، امکان استخراج اطلاعات کاری کاربر را از فضای مجازی فراهم می‌کند و به استنباط مهارت‌های عملی‌ای کمک می‌کند که ممکن است در متن پروفایل ذکر نشده باشند. این رویکرد به‌ویژه برای حوزه‌های هنری و فنی که بخش مهمی از مهارت‌ها در عمل و نه در کلام نمود پیدا می‌کنند، کارایی چشمگیری خواهد داشت.

سوم، بهره‌گیری از مدل‌های زبانی بزرگ مولد، ازجمله نسخه‌های فارسی Llama 3، GPT-4o و مدل‌های متن‌باز نظیر Ava، با روش‌های به‌روزرسانی کارآمد پارامتری مانند LoRA در دستور کار است. این اقدام ضمن بهبود کیفیت تولید روایت و کاهش چشمگیر پدیده توهم، امکان ارزیابی قابلیت‌های استنتاج صفر-نمونه را نیز فراهم می‌آورد که برای کاربردهایی با داده‌های آموزشی محدود، بسیار راهگشاست.

و درنهایت، گام چهارم به انتشار عمومی و کاربردی‌سازی نتایج اختصاص دارد. مجموعه‌داده و کدهای ارزیابی با مجوز CC BY-NC-SA 4.0 در مخازن گیت‌هاب و هاگینگ‌فیس در دسترس پژوهشگران و فعالان صنعت قرار خواهد گرفت و هم‌زمان، توسعه یک رابط برنامه‌نویسی کاربردی برای استخراج خودکار مهارت از پروفایل‌های فارسی، تا پایان سال ۱۴۰۵ برنامه‌ریزی‌شده است. افزون بر این، بهره‌گیری از روش‌های افزایش داده مانند جایگزینی

مترادف‌ها و یادگیری نیمه‌نظارتی نیز برای برچسب‌زنی داده‌های برچسب‌نخورده از پلتفرم‌های گوناگون، بخش دیگری از فعالیت‌های آینده را شکل خواهد داد تا مسیر پژوهش از آزمایشگاه به عرصه عمل هموارتر گردد.

امید است که مجموعه‌داده و مدل‌های ارزیابی‌شده در این پژوهش، پایه‌گذار نسل جدیدی از ابزارهای تحلیل سرمایه انسانی در زبان فارسی بوده و بستری مؤثر برای پژوهش‌های آینده در حوزه سیستم‌های توصیه‌گر شغلی و مدیریت منابع انسانی هوشمند فراهم آورند.

References

[1] M. Zhang, K. N. Jensen, R. van der Goot, and B. Plank, "Skill extraction from job postings using weak supervision," in *Proc. 2nd Workshop Recommender Syst. Human Resources (RecSys-in-HR)*, Seattle, WA, USA, Jul. 2022, pp. 1–6. [Online]. Available: https://aclanthology.org/2022.recsyshr-1.1/

[2] LinkedIn Economic Graph, "Global Talent Trends 2024," 2024. [Online]. Available: https://business.linkedin.com/talent-solutions/global-talent-trends

[3] I. Khaouja, I. Kassou, and M. Ghogho, "A survey on skill identification from online job ads," *IEEE Access*, vol. 9, pp. 118134–118153, 2021, doi: 10.1109/ACCESS.2021.3106120.

[4] K. Qin, K. Cunningham, and B. C. M. Fung, "SkillBERT: 'Skilling' the BERT to classify skills!," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, Nov. 2020, pp. 1116–1126. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.100/

[5] A. Gugnani and H. Misra, "Implicit skills extraction using document embedding and its use in job recommendation," in *Proc. AAAI Conf. Artif. Intell.*, New York, NY, USA, Feb. 2020, vol. 34, no. 8, pp. 13286–13293, doi: 10.1609/aaai.v34i08.7038.

[6] M. Zhang, K. N. Jensen, S. D. Sonniks, and B. Plank, "SkillSpan: Hard and soft skill extraction from English job postings," in *Proc. 2022 Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol. (NAACL-HLT)*, Seattle, WA, USA, Jul. 2022, pp. 4962–4984, doi: 10.18653/v1/2022.naacl-main.366.

[7] T. A. F. Green, D. Maynard, and C. Lin, "Development of a benchmark corpus to support entity recognition in job descriptions," in *Proc. Thirteenth Lang. Resources Eval. Conf. (LREC)*, Marseille, France, Jun. 2022, pp. 1201–1208. [Online]. Available: https://aclanthology.org/2022.lrec-1.128/

[8] A.-S. Gnehm, "Evaluation of transfer learning and domain adaptation for analyzing German-speaking job advertisements," in *Proc. Thirteenth Lang. Resources Eval. Conf. (LREC)*, Marseille, France, Jun. 2022, pp. 3891–3899. [Online]. Available: https://aclanthology.org/2022.lrec-1.127/

[9] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, "ParsBERT: Transformer-based model for Persian language understanding," *Neural Process. Lett.*, vol. 53, pp. 3831–3847, Dec. 2021, doi: 10.1007/s11063-021-10528-4.

[10] M. Masumi, S. S. Majd, M. Shamsfard, and H. Beigy, "FaBERT: Pre-training BERT on Persian blogs," 2024, *arXiv:2402.06617*. [Online]. Available: https://arxiv.org/abs/2402.06617

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[12] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguist. (ACL)*, Online, Jul. 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.

[13] European Commission, "ESCO (European Skills, Competences, Qualifications and Occupations) - Version 1.0," 2017. [Online]. Available: https://ec.europa.eu/esco/portal/escopedia/ESCO_version_1.0

[14] European Commission, "ESCO handbook," Publications Office of the European Union, Luxembourg, 2021. [Online]. Available: https://ec.europa.eu/esco/portal/escopedia/ESCO_handbook

[15] Z. Ji et al., "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, Mar. 2023, doi: 10.1145/3571730.

[16] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*. [Online]. Available: https://arxiv.org/abs/1907.11692

[17] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, "Hamshahri: A standard Persian text collection," *Knowl.-Based Syst.*, vol. 22, no. 5, pp. 382–387, Jul. 2009, doi: 10.1016/j.knosys.2009.02.002.

[18] D. Khashabi et al., "ParsiNLU: A suite of language understanding challenges for Persian," 2020, *arXiv:2012.06154*. [Online]. Available: https://arxiv.org/abs/2012.06154

[19] H. M. Khansari and M. Shamsfard, "HmBlogs: A big general Persian corpus," 2021, *arXiv:2111.02362*. [Online]. Available: https://arxiv.org/abs/2111.02362

[20] M. H. Khansari, M. Shamsfard, and H. Beigy, "MELAC: Massive Evaluation of Large Language Models with Alignment of Culture in Persian Language," in Proceedings of the Joint International Conference on Natural Language Processing and the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL), 2025, pp. 1933–1950.

[21] S. Mohtasham, M. Masumi, and M. Shamsfard, "Advancing Persian LLM Evaluation: A Comprehensive Benchmark for Cultural and Linguistic Alignment," in Findings of the North American Chapter of the Association for Computational Linguistics (NAACL), 2025.

[22] M. Zhang, K. N. Jensen, and B. Plank, "SkiLLMo: Normalized ESCO Skill Extraction through Transformer Models," in Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, 2025, pp. 1969–1978..

[23] J. M. Wang and Z. Sun, "LLM-Supervised Multilingual Skill Extraction and Classification from Job Ads," in Natural Language Processing and Information Systems, Springer, 2025.

[24] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, "Matina: A Large-Scale 73B Token Persian Text Corpus," in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[26] A. Conneau et al., “Unsupervised cross-lingual representation learning at scale,” in Proc. ACL, Online, Jul. 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 5, No. 1, Spring & Summer 2026, pp.1 - 22

Construction and Evaluation of a Dataset for Explicit and Implicit Skills from Persian Profiles using Language Models

Ali Hosseinpournaderi [1], MohammadAli Javadzade [2*], Hossein Hosseini[3]

[1]Master Student in Artificial Intelligence and Robotics, Faculty of Artificial Intelligence and Cognitive Sciences, Imam Hossein Comprehensive University, Tehran, Iran

[2]Assistant Professor, Faculty of Artificial Intelligence and Cognitive Sciences, Imam Hossein Comprehensive University, Tehran, Iran

[3]PhD's Student in Artificial Intelligence and Robotics, Faculty of Artificial Intelligence and Cognitive Sciences, Imam Hossein Comprehensive University, Tehran, Iran

Article Information

Original Research Paper

Received:
2026 July 03

Accepted:
2026 August 18

Keywords:
Text Profile Processing, Implicit Skills, Ihuskill Dataset, Pre-trained Language Models, Faithfulness Evaluation

Corresponding Author*:
alihpn@ihu.ac.ir

Abstract

Accurately inferring user skills from textual descriptions in job profiles remains a persistent challenge for employment recommendation systems and job-seeking platforms. The core difficulty arises when a skill is not explicitly stated but is instead embedded within narrative sentences and implicit contextual cues. Despite recent strides in language models, Persian has lacked a standard benchmark capable of capturing both explicit and implicit skills with any meaningful depth. This work introduces Ihuskill, the first Persian dataset of its kind, comprising 10430 annotated profiles sourced from the KarLanser freelancing platform. Beyond explicit user skills, it includes four additional information layers: implicit skills categorized into three difficulty levels, narrative explanations, contextual weights, and a dedicated faithfulness test. The dataset was constructed using a hybrid approach that combines a large language model with a human-in-the-loop refinement process. For validation, four language models—mBERT, ParsBERT, RoBERTa, and FaBERT—were evaluated across five downstream tasks: skill classification, implicit skill detection, narrative generation, context weight classification, and a comprehensive end-to-end scenario. The experimental outcomes indicate that the fine-tuned FaBERT architecture consistently outperforms the other models across all tasks. We argue that releasing this dataset along with the evaluated model provides a robust foundation for advancing skill-oriented search systems, enabling intelligent job matching, and improving textual profile analysis within Persian human resource management ecosystems.

: 10.22034/ABMIR.2026.24873.1268

E-ISSN: 2821-2037

/The Author 2026. Published by Yazd University This is an open access article under the CC BY 4.0 License https://creativecommons.org/licenses/by/4.0/).