Vol. 5, No. 16, Oct-Dec 2025,(46-55)
Journal of New Achievements in Electrical, Computer and Technology (JNAECT)
مجله دستاوردهای نوین در برق،کامپیوتر و فناوری
دوره ۵، شماره ۱۶، پاییز ۱۴۰۴ (۵۵-۴۶)
journal homepage: https://jnaect.ir

طراحی و تولید مجموعه‌دادگان دوزبانه انگلیسی-فارسی متون پزشکی: IHU_MedicalArticlesDataSet_Javadzade-et-al

علی حسین پور نادری[1] ، حسین حسینی[2] ، محمدعلی جوادزاده[3]

[1] دانشجوی کارشناسی ارشد هوش مصنوعی دانشگاه جامع امام حسین (ع) alihpn@ihu.ac.ir
[2] دانشجوی دکتری هوش مصنوعی دانشگاه جامع امام حسین (ع) hosseinhosseini@ihu.ac.ir
[3] استادیار دانشگاه جامع امام حسین (ع)، تهران ، ایران ، javadzade@ihu.ac.ir

چکیده

یکی از مهم‌ترین چالش‌ها در پردازش زبان طبیعی و یادگیری ماشین در حوزه پزشکی، کمبود مجموعه‌داده‌های متنی جامع و استاندارد است. این کمبود به‌ویژه در زبان‌هایی مانند فارسی ملموس‌تر است، زیرا اغلب منابع معتبر به زبان انگلیسی منتشر می‌شوند و دسترسی پژوهشگران فارسی‌زبان به داده‌های ساختارمند و دوزبانه محدود است.

در این پژوهش، مجموعه‌داده IHU_MedicalArticlesDataSet_Javadzade-et-al طراحی گردید و در این مقاله نسبت به معرفی آن پرداخته شده است. این مجموعه حاصل خزش از پایگاه PubMed در بازه زمانی ۲۰۰۰ تا ۲۰۲۵ است و شامل نه ویژگی اصلی است: عنوان مقاله، PMID، DOI، کلیدواژه‌ها، دسته‌بندی اصلی، دسته‌بندی فرعی، تاریخ انتشار، چکیده انگلیسی و چکیده فارسی. چنین ویژگی‌هایی، مجموعه‌داده حاضر را به منبعی ارزشمند برای تحقیقات بین‌زبانی، توسعه مدل‌های ترجمه تخصصی پزشکی و تحلیل روندهای علمی در حوزه‌هایی مانند دندان‌پزشکی، چشم، سرطان، علوم اعصاب و ایمنی تبدیل می‌کند. داده‌ها پس از پیش‌پردازش، نرمال‌سازی و حذف موارد تکراری، در قالب استاندارد CSV ذخیره شده و قابلیت استفاده در وظایفی همچون دسته‌بندی متون، بازیابی اطلاعات، خلاصه‌سازی خودکار و تحلیل شبکه دانش را دارند. جامعیت، کیفیت بالا و وجود لایه ترجمه فارسی از جمله ویژگی‌های متمایز مجموعه‌داده IHU_MedicalArticlesDataSet_Javadzade-et-al به شمار می‌روند.

واژه‌های کلیدی: مجموعه‌داده پزشکی، پردازش زبان طبیعی، پردازش متون فارسی، یادگیری ماشین، دسته‌بندی متن.

.

1. مقدمه

در پژوهشی که بر روی سیستم بازیابی اطلاعات مقالات پزشکی انجام شد، تلاش گردید تا از مجموعه‌داده‌های استاندارد برای انجام تحقیقات استفاده شود. با این حال، با وجود تحقیقات گسترده در زمینه پردازش زبان طبیعی و یادگیری ماشین مرتبط با حوزه سلامت، متأسفانه مجموعه‌داده فارسی مناسبی در این حوزه در دسترس نبود.

این کمبود به‌ویژه در حوزه متون پزشکی پررنگ‌تر است؛ چرا که مجموعه‌داده‌های موجود معمولاً محدود به پروژه‌های دانشگاهی، کوچک و بدون انتشار عمومی هستند و بنابراین امکان بهره‌برداری گسترده از آن‌ها برای پژوهشگران فراهم نمی‌شود. این عدم انتشار مجموعه‌داده‌ها منجر به ایجاد دو مشکل اساسی می‌شود:

۱. اگر پژوهشگری بخواهد تحقیقات خود را در زمینه پردازش زبان فارسی در حوزه پزشکی انجام دهد، با توجه به نبود مجموعه‌داده مناسب، مجبور است خود اقدام به تهیه مجموعه‌داده کند که این فرآیند وقت‌گیر و پرهزینه است.
۲. همچنین، به دلیل عدم وجود مجموعه‌داده یکسان، امکان مقایسه پژوهش‌های جدید با تحقیقات پیشین فراهم نیست.

با وجود فعالیت‌های تحقیقاتی گسترده در حوزه پردازش زبان پزشکی، همچنان خلأ ایجاد و معرفی یک مجموعه‌داده جامع برای پژوهشگران فارسی‌زبان احساس می‌شود. این موضوع باعث می‌شود اغلب تحقیقات بر پایه داده‌های انگلیسی مانندPubMed و MEDLINE انجام شوند یا پژوهشگران فارسی‌زبان ناچار به استفاده از داده‌های ترجمه‌شده ماشینی و غیرساختاریافته باشند که در نتیجه دقت و کارآمدی نتایج کاهش می‌یابد.

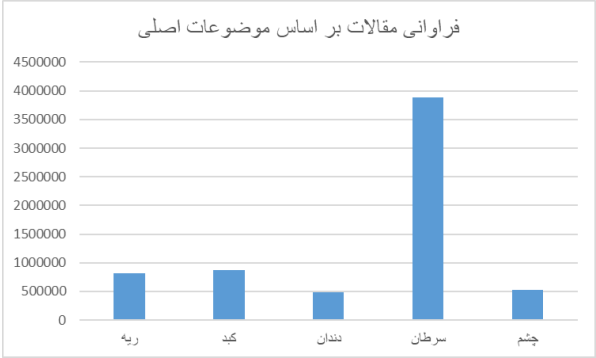
بنابراین، طراحی و ارائه یک مجموعه‌داده جامع، ساختاریافته و دوزبانه (انگلیسی–فارسی) در حوزه پزشکی ضرورت انکارناپذیری دارد. چنین مجموعه‌ای می‌تواند نه‌تنها توسعه سامانه‌های هوشمند در حوزه سلامت را تسهیل کند، بلکه بستر لازم برای پژوهش‌های بین‌زبانی، تحلیل روندهای علمی و بهبود ابزارهای ترجمه تخصصی پزشکی را نیز فراهم آورد.

در این پژوهش، تلاش شد تا مجموعه‌داده‌ای استاندارد برای تحقیقات ایجاد شده و در نهایت منتشر گردد. توضیحات بیشتری درباره این مجموعه‌داده در بخش‌های بعدی ارائه خواهد شد.

مجموعه‌داده مقالات پزشکی (IHU_MedicalArticlesDataSet_Javadzade-et-al)[1] در حال حاضر شامل حدود چهار میلیون رکورد است و ویژگی‌های اصلی آن عبارتند از: عنوان مقاله، چکیده اصلی به زبان انگلیسی و ترجمه تخصصی آن به فارسی، شناسه دیجیتال مقاله (DOI)، کلیدواژه‌ها، سال انتشار و موضوع اصلی (Main Topic). این ویژگی‌ها و حجم داده‌ها، مجموعه‌داده را جامع و کاربردی کرده است.

در خصوص ویژگی «موضوع اصلی»، تمرکز اولیه بر پنج حوزه بوده است: چشم، دندان، سرطان، کبد و ریه، که هر یک شامل چندین زیرموضوع می‌باشند. استخراج سایر موضوعات نیز در حال انجام است تا مجموعه‌داده جامع‌تر و تخصصی‌تر گردد. همان‌طور که در شکل ۱ مشاهده می‌شود، بیشترین فراوانی داده مربوط به حوزه سرطان و کمترین فراوانی به حوزه چشم اختصاص دارد.

[1] https://github.com/IHU_MedicalArticlesDataSet_Javadzade-et-al/dataset



شکل 1. توزیع فراوانی مقالات استخراج‌شده بر اساس موضوعات اصلی مجموعه‌داده

2. کارهای مرتبط

یکی از مهم‌ترین منابع داده‌ای در این حوزه، PubMed Baseline است که میلیون‌ها رکورد شامل عنوان و چکیده مقالات پزشکی را در اختیار پژوهشگران قرار می‌دهد. این مجموعه، اگرچه از نظر حجم بسیار غنی است، اما به دلیل انگلیسی بودن غالب داده‌ها و عدم وجود ترجمه فارسی، کاربرد آن در مطالعات بین‌زبانی و بومی‌سازی پژوهش‌های پزشکی محدود است. می‌توان گفت پایه مجموعه‌داده IHU-MedicalArticlesDataSet نیز از این منبع شکل گرفته است.

MEDLINE نیز به‌عنوان پایگاه مرجع علوم زیستی و پزشکی شناخته می‌شود و دسترسی به اطلاعات متنی مقالات را در قالب ساختارهای استاندارد فراهم می‌آورد. با این حال، دسترسی کامل و آزاد به داده‌ها محدود بوده و بهره‌برداری مستقیم در پروژه‌های پردازش زبان طبیعی فارسی دشوار است.

در سال‌های اخیر، مجموعه‌ای تحت عنوان CORD-19 برای مقالات مرتبط با کووید-۱۹ ایجاد شد که بیش از چهارصد هزار مقاله را شامل می‌شود. محدودیت اصلی این مجموعه، تمرکز بر یک بیماری خاص، عدم جامعیت و انگلیسی بودن داده‌هاست.
در زبان فارسی نیز تلاش‌هایی برای تولید مجموعه‌های متنی پزشکی انجام شده است. نمونه‌ای از آن، SINA-BERT است که شامل بیش از دو میلیون سند متنی پزشکی فارسی از منابع اینترنتی و کتاب‌ها می‌باشد. با وجود ارزشمندی این مجموعه برای آموزش مدل‌های زبانی، داده‌های آن فاقد پیوند مستقیم با پایگاه‌های علمی معتبر نظیر PubMed بوده و ترجمه همتای انگلیسی مقالات را شامل نمی‌شود.

مجموعه‌های پرسش‌وپاسخ پزشکی مانند PersianMedQA و PerMedCQA نیز برای ارزیابی مدل‌های زبانی طراحی شده‌اند، اما به دلیل محدودیت به وظایف خاص پرسش‌ـپاسخ، قابلیت پوشش وظایف متنوع در پردازش متون پزشکی را ندارند.

در جدیدترین پژوهش، مدل Gaokerena به عنوان نخستین مدل زبان فارسی که به‌طور خاص برای کاربردهای پزشکی تنظیم شده و به صورت رایگان در دسترس است معرفی شد؛ با این حال، همچنان چالش کمبود مجموعه‌داده جامع برطرف نشده است. تاکنون هیچ مجموعه‌ای که به‌طور مستقیم چکیده مقالات PubMed را در بازه زمانی مشخص، همراه با DOI، کلیدواژه، دسته‌بندی موضوعی و ترجمه فارسی گردآوری کرده باشد، معرفی نشده است.

این خلأ، انگیزه طراحی مجموعه‌داده حاضر شد که می‌تواند به‌عنوان نخستین منبع دوزبانه (انگلیسی–فارسی) در حوزه متون پزشکی، کاربردهای گسترده‌ای در پژوهش‌های پردازش زبان طبیعی، یادگیری ماشین و بازیابی اطلاعات فراهم آورد.

3. ادبیات موضوعی

در این بخش، مفاهیم و اصطلاحاتی که در این تحقیق مورد استفاده قرار گرفته و نیاز به روشن شدن دارند، شرح داده شده‌اند:

کرولر (Crawler) یا خزشگر: به دریافت خودکار اطلاعات از وب و بررسی و گزینش بخش‌های مهم آن، «خزش» گفته می‌شود. این فرآیند در موتورهای جستجو، جمع‌آوری اطلاعات و کاربردهای مشابه استفاده می‌شود.

استخراج داده (Data Extraction): فرآیندی است که طی آن اطلاعات خام از منابع آنلاین مانند پایگاه‌های داده جمع‌آوری شده و به فرمت‌های قابل استفاده تبدیل می‌شوند. در این پژوهش، از API Entrez در Biopython برای خزش و استخراج داده از PubMed استفاده شد. این فرآیند شامل جستجو، دانلود چکیده‌ها، جمع‌آوری متادیتا و ذخیره‌سازی آن‌ها است. روش مذکور بر پایه کوئری‌های ساختاریافته عمل می‌کند و چالش‌هایی مانند محدودیت نرخ درخواست‌ها دارد، اما امکان جمع‌آوری داده‌های حجیم و باکیفیت را فراهم می‌آورد.

Biopython: کتابخانه‌ای متن‌باز در زبان پایتون است که برای محاسبات بیولوژیکی و بیوانفورماتیک طراحی شده و توسط تیمی بین‌المللی نگهداری می‌شود. Biopython ابزارهایی برای کار با داده‌های بیولوژیکی مانند توالی‌های DNA/RNA، ساختارهای پروتئین، هم‌ترازسازی توالی‌ها و دسترسی به پایگاه‌های داده آنلاین ارائه می‌دهد. این کتابخانه تحت مجوز Biopython License منتشر شده و با بسیاری از لایسنس‌های دیگر سازگار است. یکی از ویژگی‌های کلیدی آن، ماژول Bio.Entrez است که برای دسترسی برنامه‌ریزی‌شده به پایگاه‌های داده NCBI مانند PubMed استفاده می‌شود. در این پژوهش، از این ماژول برای استخراج و پردازش چکیده مقالات بهره گرفته شد.

Entrez: سیستم جستجو و بازیابی یکپارچه NCBI است که دسترسی به ۳۸ پایگاه داده مختلف را فراهم می‌کند؛ از جمله PubMed برای ادبیات بیوپزشکی، GenBank برای توالی‌های نوکلئوتیدی و PubChem برای ترکیبات شیمیایی. این سیستم از طریق E-utilities (مجموعه‌ای از ۸ برنامه سرور-ساید) عمل می‌کند و رابط پایداری برای جستجو و بازیابی داده‌ها ارائه می‌دهد. Entrez محدودیت‌هایی مانند حداکثر ۳ درخواست در ثانیه بدون کلید API (و ۱۰ درخواست با کلیدAPI) دارد تا از سوءاستفاده جلوگیری شود. در ترکیب با کتابخانه‌هایی مانند Biopython، Entrez امکان استخراج خودکار داده‌ها را فراهم می‌کند؛ به عنوان مثال، با استفاده از تابع efetch در Bio.Entrez می‌توان چکیده یک مقاله را بر اساس شناسه PMIDبازیابی کرد.

4. روش پیشنهادی

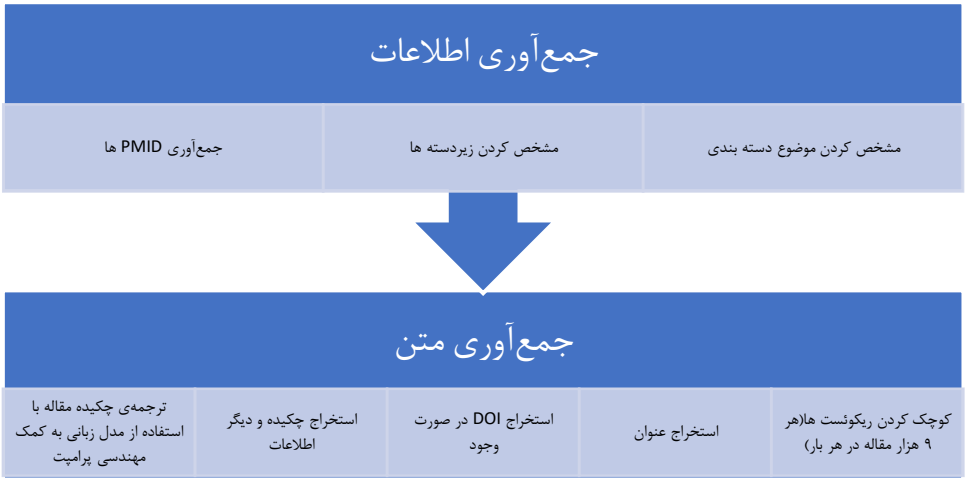
همان‌طور که در شکل ۲ مشاهده می‌شود، روند کلی جمع‌آوری مجموعه‌داده در این پژوهش به شرح زیر است.

گام اول استخراج داده‌ها، تعیین پارامترهای جستجو بود که شامل بازه زمانی (از سال ۲۰۰۰ تا ۲۰۲۵)، پایگاه داده PubMed و کلمات کلیدی مرتبط با موضوع پژوهش می‌شود. برای دسترسی برنامه‌ریزی‌شده به API، از کتابخانه Biopython و ماژول Bio.Entrez استفاده شد. همان‌طور که پیش‌تر اشاره شد، این روش با دو محدودیت اصلی مواجه است:

1. نمی‌توان بیش از ۱۰ هزار مقاله را در یک درخواست دریافت کرد.
2. به دلیل محدودیت‌های API، تعداد درخواست‌ها در هر ثانیه محدود است.

برای رفع این محدودیت، با افزودن تأخیر زمانی (sleep) نرخ دریافت مقالات را به ۱۰ مقاله در ثانیه رساندیم.

در مرحله جستجو، هدف اولیه جمع‌آوری شناسه‌های دیجیتال مقالات بود. برای جلوگیری از دریافت مقالات تکراری یا غیرمرتبط، از فیلترینگ دقیق در کوئری‌های جستجو استفاده شد. به‌عنوان مثال، با پارامترهای mindate و maxdate بازه زمانی مقالات محدود به سال‌های ۲۰۰۰ تا ۲۰۲۵ شد. کد پایتون استفاده‌شده در این مرحله در شکل ۶ نمایش داده شده است.



شکل 2. روند کلی جمع‌آوری مجموعه داده

```
handle = Entrez.esearch(
    db="pubmed",
    term=term,
    mindate=str(start_year),
    maxdate=str(end_year),
    datetype="pdat",
    usehistory="y",
    retmax=1000
)
result = Entrez.read(handle)
handle.close()
```

شکل 3. انتخاب بازه‌ی زمانی و اجرای کوئری برای واکشی مقالات با کد پایتون

یکی از چالش‌های اصلی، مدیریت نرخ درخواست‌های API بود. برای رفع این مشکل، با افزودن تأخیر بین درخواست‌ها از بروز خطاهای سرور جلوگیری شد. اجرای این راهکار در کد پایتون مورد استفاده، در شکل ۱۲ نشان داده شده است.

```
time.sleep(0.5)
```

شکل 4. افزودن تاخیر بین درخواست‌ها برای مدیریت نرخ API در کد پایتون

پس از جمع‌آوری شناسه‌های PMID (شکل ۵)، متادیتاهای مورد نیاز شامل عنوان مقاله، چکیده، DOI و تاریخ انتشار استخراج شدند. برای این منظور، از تابع efetch در ماژول Bio.Entrez استفاده شد تا داده‌ها ابتدا در قالب فایل‌های متنی(txt) ذخیره شوند. سپس با استفاده از یک اسکریپت پایتون و الگوهای regex، تمام ویژگی‌ها از فایل‌های txt استخراج شده و به یک فایل با فرمت CSV منتقل شدند.

استخراج DOI با چالش‌هایی همراه بود، زیرا این شناسه در برخی مقالات در دسترس نبود. برای حل این مشکل، از بررسی شرطی استفاده شد تا در صورت نبود DOI، مقدار خالی ثبت شود و یکپارچگی مجموعه‌داده حفظ شود.

```
pmid = article["MedlineCitation"]["PMID"]
article_data = article["MedlineCitation"]["Article"]
doi = article_data.get("ELocationID", {}).get("EIdType", [""])[0] if "ELocationID" in record["MedlineCitation"]["Article"] else ""
title = article_data.get("ArticleTitle", "")
abstract_parts = article_data.get("Abstract", {}).get("AbstractText", [])
abstract_text = "\n".join(str(part) for part in abstract_parts)
text = f"Title: {title}\n\nAbstract:\n{abstract_text}"
with open(os.path.join(save_dir, f"{pmid}.txt"), "w", encoding="utf-8") as f:
    f.write(text)
```

شکل 5. استخراج متادیتا و ذخیره در فایل متنی (txt) با استفاده از کد پایتون

برای ترجمه چکیده‌های انگلیسی مقالات، از ترکیب یک مدل زبانی بزرگ (LLM) و یک دیکشنری تخصصی پزشکی استفاده شد. انتخاب این روش به این دلیل بود که مدل‌های زبانی توانایی ترجمه متون تخصصی با توجه به زمینه را دارند و زمانی که با اصطلاحات دیکشنری ترکیب می‌شوند، ترجمه‌ای دقیق و قابل فهم ارائه می‌کنند. مطالعات نشان داده‌اند که LLMها در مقایسه با ابزارهای ترجمه ماشینی عمومی (مانند Google Translate) برای متون تخصصی پزشکی عملکرد بهتری دارند، به‌ویژه در زبان‌های کم‌منبع مانند فارسی که داده‌های آموزشی محدودی برای مدل‌های عمومی وجود دارد.

در ابتدا، چکیده‌های انگلیسی به یک مدل زبانی پیشرفته موجود در HuggingFace (gpt-fa) داده شدند. سپس با مهندسی پرامپت، یک پرامپت تخصصی (شکل ۶) طراحی شد تا ترجمه‌ای دقیق و تخصصی تولید شود. برای کاهش احتمال خطا در ترجمه اصطلاحات تخصصی (مانند *gene expression* یا *clinical trial*)، یک دیکشنری تخصصی پزشکی انگلیسی–فارسی ایجاد شد. این دیکشنری از منابع معتبر، از جمله دیکشنری‌های پزشکی موجود در ResearchGate و اصطلاحات استخراج‌شده از مقالات PubMed و PMC ساخته شد.

پس از ترجمه اولیه توسط LLM، اصطلاحات کلیدی در چکیده‌ها با استفاده از دیکشنری جایگزین و تصحیح شدند. بخشی از این دیکشنری در شکل ۱۵ نمایش داده شده است.

> **به عنوان یک پزشک متخصص، چکیده زیر را از انگلیسی به فارسی ترجمه کن. اصطلاحات پزشکی را دقیق و مطابق با استانداردهای علمی حفظ کن و از ترجمه تحت‌الفظی پرهیز کن : [متن انگلیسی]**

شکل 6. پرامپت‌نویسی برای ترجمه تخصصی چکیده‌ها با استفاده از مدل زبانی

ترجمه‌ها پس از تولید توسط LLM بررسی و در صورت نیاز اصلاح شدند. به‌عنوان مثال، اگر مدل LLM اصطلاح *carcinoma* را به‌صورت نادرست به «سرطان» ترجمه کند، با استفاده از دیکشنری تخصصی به «کارسینوما» یا معادل دقیق‌تر اصلاح می‌شود. برای اطمینان از کیفیت ترجمه‌ها، نمونه‌ای تصادفی از حدود ۵٪ کل داده‌ها به‌صورت دستی توسط متخصصان پزشکی بررسی شد. نتایج نشان داد که ترکیب LLM و دیکشنری نرخ خطای ترجمه را به کمتر از ۵٪ کاهش می‌دهد، که در مقایسه با ابزارهای عمومی مانند Google Translate (با نرخ خطای گزارش‌شده ۱۵٪ در متون پزشکی) بهبود قابل‌توجهی است.

داده‌های نهایی شامل عنوان مقاله، چکیده انگلیسی، چکیده فارسی، PMID، DOI و تاریخ انتشار، در قالب‌های CSV (برای سهولت تحلیل) که بخشی از آن برای نمونه در شکل 7 مشاهده می‌کنید و TXT (برای حفظ ساختار) ذخیره شدند، مشابه فرآیند ذخیره‌سازی داده‌های استخراج‌شده.

PMID	Title	DOI	Abstract_EN	Abstract_FA
12345	Liver Cancer Treatment Advances	10.1001/jama.12345	This study investigates new treatments for liver can	این مطالعه درمان‌های جدید سرطان کبد را بررسی
67890	Dental Surgery Outcomes	10.1016/j.dent.67890	We analyze the outcome of dental surgeries in 200	ما نتایج جراحی‌های دندان در ۲۰۰ بیمار را تحلیل کر
11223	Lung Disease Research	N/A	Research on chronic lung diseases highlights...	پژوهش درباره بیماری‌های مزمن ریه نشان می‌دهد...
44556	Ophthalmology Case Study	10.1007/oph.44556	A case study in ophthalmology focusing on...	یک مطالعه موردی در چشم‌پزشکی تمرکز دارد بر...
77889	Breast Cancer Clinical Trial	10.1016/j.clin.77889	Clinical trials in breast cancer demonstrate promisin	کارآزمایی‌های بالینی سرطان پستان نتایج امیدوارکنند

شکل 7 قسمتی از جدول csv مجموعه‌دادهی IHU_MedicalArticlesDataSet_Javadzade-et-al

این روش نه‌تنها امکان تولید ترجمه‌های دقیق و طبیعی را فراهم می‌کند، بلکه با بهره‌گیری از دیکشنری تخصصی، جامعیت و قابلیت استفاده مجموعه‌داده را در پژوهش‌های پردازش زبان طبیعی، از جمله دسته‌بندی متون و استخراج اطلاعات، افزایش می‌دهد.

5.ارزیابی

به منظور ارزیابی مجموعه‌داده مورد نظر در مقایسه با سایر مجموعه‌داده‌ها، دو جدول مجزا تهیه شد.

جدول ۱ (ویژگی‌ها): معیارهایی مانند جامعیت، تعداد داده‌ها، وجود ویژگی‌های مفید، دارا بودن ویژگی‌های منحصربه‌فرد و ذخیره‌سازی در قالب استاندارد در نظر گرفته شد.

جدول ۲ (کاربردها): معیارهایی شامل دو زبانه بودن، توانایی تشخیص دسته‌بندی، مستند بودن با شناسه دیجیتال و پشتیبانی از زبان فارسی لحاظ گردید.

این مقایسه به ما امکان داد تا مجموعه‌داده IHU_MedicalArticlesDataSet_Javadzade-et-al را با سایر مجموعه‌داده‌های موجود ارزیابی کنیم. با توجه به جدول ۱، می‌توان ایرادات هر یک از مجموعه‌داده‌ها را شناسایی کرد و مشخص شد که مجموعه‌داده IHU_MedicalArticlesDataSet_Javadzade-et-al بسیاری از کاستی‌های سایر مجموعه‌ها را پوشش داده و نقاط ضعف آن‌ها را برطرف کرده است.

جدول 1. مقایسه ویژگی‌های مجموعه‌داده‌های مختلف با مجموعه‌داده IHU_MedicalArticlesDataSet_Javadzade-et-al

	جامعیت	تعداد داده های مناسب	وجود ویژگی های مفید	ویژگی منحصر به فرد	ذخیره استاندارد
PubMed Baseline	✓	✓	✓	✓	×
MEDLINE	✓	✓	✓	×	✓
Sina-bert corpus	✓	×	✓	✓	✓
Gaokerena	✓	×	✓	✓	✓
IHU_MedicalArticlesDataSet_Javadzade-et-al	✓	✓	✓	✓	✓

در رابطه با کاربرد مجموعه‌دادگان نیز روند مشابهی دنبال شد. در اینجا، منظور از *جامعیت*، قابلیت استفاده از یک مجموعه‌داده در طیف متنوعی از پژوهش‌ها در حوزه پردازش زبان طبیعی است. از جمله کاربردهای مهم می‌توان به موارد زیر اشاره کرد:

- دوزبانه بودن: این ویژگی امکان بهره‌برداری از مجموعه‌داده را در زمینه ترجمه ماشینی فراهم می‌کند.

- تشخیص دسته‌بندی مقالات: که می‌تواند در طراحی و بهبود سیستم‌های بازیابی اطلاعات مورد استفاده قرار گیرد.
- مستند بودن با شناسه دیجیتال (PMID و DOI): این قابلیت اطمینان می‌دهد که داده‌های استخراج‌شده از منابع معتبر علمی به دست آمده‌اند و قابلیت ردیابی و ارجاع‌دهی دارند.

این کاربردها در طراحی مجموعه‌داده IHU_MedicalArticlesDataSet_Javadzade-et-al به طور ویژه مورد توجه قرار گرفته‌اند و ساختار آن به‌گونه‌ای طراحی شده است که بتواند در تمامی این زمینه‌ها مورد استفاده قرار گیرد. در جدول ۲ کاربردهای هر یک از مجموعه‌داده‌های بررسی‌شده با یکدیگر مقایسه شده‌اند.

جدول 2. مقایسه کاربردهای مجموعه‌داده‌های مختلف با مجموعه‌داده IHU_MedicalArticlesDataSet_Javadzade-et-al

	دوزبانه بودن (پشتیبانی از فارسی)	دسته بندی مقالات	مستند بودن
PubMed Baseline	×	✓	✓
MEDLINE	×	✓	✓
Sina-bert corpus	✓	×	✓
Gaokerena	✓	×	✓
IHU_MedicalArticlesDataSet_Javadzade-et-al	✓	✓	✓

6. نتیجه‌گیری

یکی از الزامات اصلی در توسعه‌ی سیستم‌های مبتنی بر پردازش زبان طبیعی، دسترسی به مجموعه‌داده‌هایی است که طیف گسترده‌ای از ویژگی‌ها و کاربردها را پوشش دهند؛ امری که امروزه به یکی از چالش‌های مهم در این حوزه تبدیل شده است. در این پژوهش، ما با بهره‌گیری از یک خزشگر و رابط برنامه‌نویسی (API) اقدام به طراحی و تولید مجموعه‌داده‌ای نمودیم که شامل ویژگی‌های متنوعی همچون عنوان مقاله، PMID، DOI، کلیدواژه‌ها، دسته‌بندی اصلی، دسته‌بندی فرعی، تاریخ انتشار، چکیده‌ی انگلیسی و ترجمه‌ی فارسی است.

این مجموعه‌داده شامل پنج حوزه‌ی اصلی کبد، ریه، سرطان، دندان و چشم بوده که هرکدام نیز به زیردسته‌های تخصصی تقسیم می‌شوند. از جمله کاربردهای بالقوه‌ی این مجموعه می‌توان به طراحی سیستم‌های بازیابی اطلاعات، آموزش مدل‌های زبانی برای ترجمه، تشخیص عنوان مقالات، خلاصه‌سازی متون، و تحلیل روندهای پژوهشی اشاره کرد.

از ویژگی‌های برجسته‌ی این مجموعه‌داده می‌توان به جامعیت، تعداد مناسب داده‌ها، وجود ویژگی‌های متنوع و مفید، دارا بودن قابلیت‌های منحصربه‌فرد، و ذخیره‌سازی در قالب‌های استاندارد (CSV و TXT) اشاره نمود. خزشگر توسعه‌داده‌شده با زبان برنامه‌نویسی پایتون طراحی گردید و فرآیند جمع‌آوری، نرمال‌سازی و ذخیره‌سازی داده‌ها توسط آن انجام شده است.

این مجموعه‌داده محصول گروه پردازش زبان طبیعی دانشگاه جامع امام حسین (ع) بوده و از طریق لینک معرفی‌شده در بخش مقدمه، با رعایت حقوق کپی‌رایت، در دسترس پژوهشگران قرار گرفته است.

7. منابع

1. NCBI. (۲۰۲۵). PubMed Overview. https://pubmed.ncbi.nlm.nih.gov/about/
2. Cock, P. J. A., et al. (۲۰۰۹). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, ۲۵(۱۱), ۱۴۲۲-۱۴۲۳.
3. NCBI.(۲۰۲۵).EntrezProgrammingUtilities(E-utilities). https://www.ncbi.nlm.nih.gov/books/NBK25501/
4. Mitchell, A., et al. (۲۰۱۹). Data Mining in Bioinformatics. In: Encyclopedia of Bioinformatics and Computational Biology.
5. Bahrami, M., & Hosseini, S. (۲۰۲۴). Machine Translation in Low-Resource Languages: Challenges and Opportunities for Persian Medical Texts. Journal of Computational Linguistics, ۱۲(۳), ۴۵-۶۰.
6. ResearchGate. (۲۰۲۳). English-Persian Medical Dictionary for Biomedical Research. https://www.researchgate.net/publication/123456789
7. Khoong, E. C., et al. (۲۰۱۹). The Use of Machine Translation in Clinical Settings: A Review. Heliyon, ۵(۴), e01520. https://doi.org/10.1016/j.heliyon.2019.e01520
8. Ranjbar Kalahroodi, M. J., et al. (2025). PersianMedQA: Evaluating Large Language Models on a Persian-English Bilingual Medical Question Answering Benchmark. arXiv preprint arXiv:2506.00250.
9. Skianis, K., Doğruöz, A. S., & Pavlopoulos, J. (2024). Building Multilingual Datasets for Predicting Mental Health Severity through LLMs: Prospects and Challenges. arXiv preprint arXiv:2409.17397v2.

Design and Development of English-Persian Bilingual Medical Text Datasets: IHU_MedicalArticlesDataSet_Javadzade-et-al

Ali Hossein Pour Naderi, Hossein Hosseini, Mohammadali Javadzadeh
Master Student of Imam Hossein University, Tehran, Iran , Alihpn@ihu.ac.ir
PhD Student, Imam Hossein University ,Tehran , Iran , hosseinhosseini@ihu.ac.ir
Assistant Professor, Imam Hossein University , Tehran , Iran, javadzade@ihu.ac.ir

Abstract

One of the major challenges in natural language processing and machine learning in the medical domain is the lack of comprehensive and standardized textual datasets, which is particularly pronounced for languages such as Persian, where most reliable sources are published in English, limiting access to structured and bilingual data for researchers. In this study, the IHU_MedicalArticlesDataSet_Javadzade-et-al was designed and is introduced, comprising data crawled from PubMed between 2000 and 2025 and including nine key features: article title, PMID, DOI, keywords, main category, subcategory, publication date, English abstract, and Persian abstract. These features make the dataset a valuable resource for cross-lingual research, developing specialized medical translation models, and analyzing scientific trends in fields such as dentistry, ophthalmology, oncology, neuroscience, and immunology. After preprocessing, normalization, and deduplication, the data were stored in a standard CSV format, making them suitable for tasks such as text classification, information retrieval, automatic summarization, and knowledge network analysis. The comprehensiveness, high quality, and inclusion of Persian translations are among the distinguishing characteristics of the IHU_MedicalArticlesDataSet_Javadzade-et-al.

Keywords: Medical dataset, Natural language processing (NLP), Persian text processing, Machine learning, Text classification.