نشریه پژوهش‌های زبان‌شناسی تطبیقی
سال دوازدهم- شماره ۲۳- بهار و تابستان ۱۴۰۱

برچسب‌گذاری خودکار فراگفتمان موضع‌گیری مقالات علمی فارسی با استفاده از یادگیری عمیق

مسعود قیومی[۱*]
محمدمهدی عبداله‌پور[۲]

مقاله پژوهشی

چکیده

در تحلیل فراگفتمانی روابط بین نویسنده، خواننده و خود متن بررسی می‌شود. هایلند (۲۰۰۵) شیوه‌ای از تحلیل را مطرح کرد که نوعی رابطه تعاملی بین این سه رکن متن ایجاد می‌کند. روابط فراگفتمان تعاملی به دو دسته موضع‌گیری و مشارکتی تقسیم می‌شود. فراگفتمان موضع‌گیری با پنج نوع نشانگر نشانه‌گذاری می‌شود. هدف از انجام این پژوهش این است که به مدل رایانشی دست یافت تا به‌صورت خودکار امکان برچسب‌گذاری فراگفتمانی نشانگرها، خواه واژه‌ها یا عبارات، میسر گردد. برای رسیدن به هدف، ابتدا به‌واسطه خزش پیکره‌ای از چکیده مقالات موجود در پرتال جامع علوم انسانی به‌دست آمد و کار نشانه‌گذاری فراگفتمان موضع‌گیری چکیده‌ها که متعلق به ۱۶ حوزه علوم انسانی بود بر اساس نظر فراگفتمان تعاملی هایلند (۲۰۰۵) نشانه‌گذاری شد. در این فرایند ۵۰ چکیده برای هر حوزه نشانه‌گذاری شد. از این داده برای آموزش مدل پردازشی استفاده شد. در این پژوهش، یک مدل با استفاده از بازنمایی معنایی واژه‌ها در فضای برداری ساخته شده توسط ورد۲وک تهیه شده است و در دو مدل دیگر از بازنمایی معنایی مبتنی بر برت به‌نام‌ها پارس‌برت و ایکس.ال.ام-روبرتا استفاده شده است تا بافت‌های جایگاهی متنوع‌تری از واژه‌ها در بردارها لحاظ گردد. مدل‌ها در سطح واژه یا عبارت کار برچسب‌گذاری را انجام داده است. براساس نتایج عملی به‌دست‌آمده، مدل مبتنی‌بر بازنمایی برت با تفاوت معناداری بهتر از مدل مبتنی بر بازنمایی ورد۲وک عمل نمود. از میان دو مدل مبتنی بر بازنمایی برت، در مجموع برچسب‌گذاری در سطح واژه و عبارت، مدل ایکس.ال.ام-روبرتا با ۸۲/۸۲ درصد امتیاز F در سطح واژه و ۵۱/۸۱ درصد امتیاز F در سطح عبارت کارایی بهتری نسبت‌به مدل پارس‌برت به‌دست آورد.

کلیدواژه‌ها: فراگفتمان موضع‌گیری، بازنمایی معنایی، ورد۲وک، برت، نشانه‌گذاری پیکره زبانی.

۱- استادیار پژوهشگاه علوم انسانی و مطالعات فرهنگی* ✉| m.ghayoomi@ihcs.ac.ir
۲- دانشجوی کارشناسی دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر ✉| mabdollahpour@aut.ac.ir

تاریخ دریافت: ۱۴۰۰/۰۶/۰۱ **تاریخ پذیرش:** ۱۴۰۱/۰۱/۱۶

شناسه دیجیتال (DOI): 10.22084/RJHLL.2022.24787.2170

۱- مقدمه

نگارش متن و نویسندگی یکی‌از شیوه‌های مرسوم انتقال اندیشه و نظر از نویسنده به خواننده است. نویسنده برای انتقال اندیشه خود، اطلاعات را در چارچوب ساختار یک زبان طبیعی کدگذاری کرده و خواننده نیز با کدگشایی اطلاعات مفهوم آن اندیشه را درک می‌نماید. بنابراین، میان متن، نویسنده و خواننده نوعی تعامل وجود دارد و از تعامل بین این سه درک زبانی متن حاصل می‌گردد. ضروری است برای درک اطلاعات تبادل شده اطلاعات نویسنده کدگشایی گردد. شایان ذکر است شیوه کدگذاری اطلاعات در موضوعات مختلف متفاوت است. در این مقاله می‌کوشیم به کدگشایی این اطلاعات بپردازیم.

اصطلاح فراگفتمان[1] به‌عنوان شیوه‌ای برای درک زبان کاربردی و بررسی رویکرد گوینده یا نویسنده راجع به یک موضوع توسط هریس (۱۹۷۰) در سال ۱۹۵۹ معرفی شد. زبان که به‌عنوان یک ابزار ارتباطی برای تبادل اطلاعات و اندیشه میان تولیدکننده (گوینده یا نویسنده) و دریافت‌کننده (شنونده یا خواننده) است در تحلیل فراگفتمانی، ارتباط فراتر از این ارتباط دیده می‌شود به‌گونه‌ای که در این کانال ارتباطی سبک تفکر، رویکرد افراد و نوع نگرش آنها براساس جایگاه اجتماعی و مانند آن وجود دارد و در درک نقش بازی می‌کند. برای درک عمیق‌تر اطلاعات، ساختاری در فراتحلیل فراهم می‌آید و تلاش می‌شود با توجه به بافت و این ویژگی‌ها، فضای گفتمانی بین تولیدکننده و دریافت‌کننده ایجاد شود. هریس (۱۹۹۸: ۴۳۷) فراگفتمان را این گونه تعریف کرده است که فراگفتمان «یک ساختار کاربردی اصلی است که به ما اجازه می‌دهد ببینیم چگونه نویسندگان با توجه به متن بر درک خوانندگان تأثیر می‌گذارند و همچنین چگونه نگرش‌شان نسبت به محتوا و مخاطب اثربخش است». کریسمور و همکاران (۱۹۹۳) فراگفتمان را این گونه تعریف می‌کنند که «مطالب زبانی موجود در متون، اعم از نوشتاری یا گفتاری، چیزی به محتوای گزاره‌ای اضافه نمی‌کند، اما به شنونده یا خواننده کمک می‌کند تا کار سازماندهی، تفسیر و ارزیابی اطلاعات داده شده را انجام دهد». در این تعاریف فراگفتمان، تأکید بیشتر بر روی ساختار است که به «نظریه ساختار بلاغت» (من و تامسون، ۱۹۸۸) و توضیح روابط بین اجزای یک متن برای تولید متن توسط رایانه نزدیک است.

در نوع دیگری از تحلیل فراگفتمانی که در چارچوب نظری هایلند (۲۰۰۵) قرار می‌گیرد به‌جای استفاده از عبارات در به‌دست آوردن ساختار متن، از واژه‌ها یا عبارت به‌عنوان نشانگرهای[2] فراگفتمان استفاده می‌شود. این شیوه تحلیل در این پژوهش مورد استفاده قرار می‌گیرد و تلاش می‌شود با تخصیص برچسب‌هایی به واژه‌ها، عبارارت، نقش واژه‌ها در تحلیل فراگفتمانی مشخص شود و براساس آنها ساختار متن به‌دست آید.

یکی از شیوه‌های برچسب‌دهی داده‌ها استفاده از روش‌های رایانشی است که اخیراً بسیار متداول شده است. هدف از انجام این پژوهش تهیه یک مدل برچسب‌گذاری است تا بتواند در چارچوب رویکرد هایلند (۲۰۰۵) کار برچسب‌گذاری نشانگرهای فراگفتمانی در مقالات علمی نوشته شده به زبان فارسی در رشته علوم انسانی را انجام داد. ویژگی برچسب‌گذاری داده‌ها به‌صورت ماشینی در مقایسه با برچسب‌گذاری دستی

1. metadiscourse
2. marker

این است که می‌توان یک الگوریتم محاسباتی تهیه نمود و با کمک آن حجم زیادی از داده‌ها را در زمان کوتاهی برچسب‌گذاری نمود. ویژگی دیگر این شیوه تحلیل این است که قابلیت توسعه و بهبود این الگوریتم وجود دارد تا بتوان کار برچسب‌گذاری را با دقت بالاتر انجام نمود.

ساختار این پژوهش به این صورت است که پس از مقدمه، در بخش ۲ پیشینه مطالعات انجام شده در حوزه برچسب‌گذاری فراگفتمان بررسی می‌گردد. در بخش ۳، رویکرد هایلند (۲۰۰۵) در تحلیل فراگفتمانی به‌صورت فشرده معرفی می‌شود. روش پژوهش و الگوریتم برچسب‌گذاری در بخش ۴ معرفی می‌گردد. در بخش ۵، داده‌های تهیه شده برای این پژوهش معرفی شده و نتایج تجربی به‌دست‌آمده از مدل توضیح داده می‌شود. در بخش آخر، بخش ۶، جمع‌بندی و نتیجه‌گیری از مقاله انجام می‌پذیرد.

۲- پیشینه مطالعاتی

در حوزه تحلیل‌های فراگفتمان پژوهش‌های متعددی انجام شده است که تقریباً تمامی آنها از پیکره زبانی استفاده شده است و تحلیل‌های خود را براساس شواهد زبانی انجام داده‌اند که نمونه بارز آن پژوهش هایلند (۲۰۰۵) بر روی مقالات علمی است. در این بخش تعدادی از پژوهش‌های انجام شده در حوزه تهیه پیکره برچسب‌گذاری شده فراگفتمان و همچنین تهیه مدل پردازشی توضیح داده می‌شود.

در مورد پیکره‌های موجود درحوزه فراگفتمان می‌توان به «دادگان درختی گفتمانی پن»[1] (مارکوس و همکاران، ۱۹۹۳) اشاره کرد که در آن مقالات مجله وال‌استریت[2] تحلیل شده و ضمن فراهم آمدن نمودار درختی نحوی جملات، چهار نوع روابط گفتمانی در این داده مشخص شده است (وبرو و جوشی، ۱۹۹۸). مارکو (۲۰۰۰) پیکره دیگری را براساس مقالات مجله وال‌استریت تهیه کرده است که با نام «دادگان درختی گفتمان مبتنی بر نظریه ساختار بلاغت»[3] معرفی شده است و در آن برچسب‌های گفتمانی متون براساس نظریه ساختار بلاغت تعیین شده است.

سوریکات و مارکو (۲۰۰۳) سامانه‌ای را طراحی کرده‌اند که در سطح جمله، درخت تجزیه نحوی جمله را به‌همراه تحلیل گفتمان جمله ارائه می‌نماید. داده آموزش استفاده شده در این پژوهش «دادگان درختی گفتمان مبتنی بر نظریه ساختار بلاغت» (مارکو، ۲۰۰۰) است.

هنگ و تان (۲۰۱۰) دو پیکره را از نوشتار انگلیسی دانشجویان مقطع کارشناسی سال اولی و سال دوم و سومی در کشور اندونزی را تهیه کرده‌اند و در چارچوب رویکرد هالند (۲۰۰۵) از دو جنبه موضع‌گیری[4] و مشارکتی[5]، کار برچسب‌گذاری فراگفتمان تعاملی[6] را انجام داده‌اند. پیکره اول شامل ۱۴۵۴۲۵ واژه بوده و

1. Penn Discourse Treebank
2. Wall Street Journal
3. Rhetorical Structure Theory Discourse Treebank
4. interactional
5. interpersonal
6. interactive metadiscourse

حاوی ۴۶۴۴ برچسب موضع‌گیری و ۵۱۵۱ برچسب مشارکتی است. پیکره دوم شامل ۸۰۸۶۴۲ واژه بوده و حاوی ۳۰۶۴۶ برچسب موضع‌گیری و ۱۹۵۷۱ برچسب مشارکتی است.

یان (۲۰۱۵) در ژانر خبری به‌صورت تصادفی ۲۰ خبر گزارشی و تفسیری را از دو روزنامه روسی انتخاب کرده و دو پیکره کوچک تهیه کرده است. پیکره خبری گزارشی حاوی ۴۱۴۶۸ واژه و پیکره خبری تفسیری حاوی ۱۱۴۰۹ واژه است. در این پژوهش، پیکره تهیه‌شده براساس فراگفتمان تعاملی هایلند (۲۰۰۵) برچسب‌گذاری شده است.

ویلسون (۲۰۱۰؛ ۲۰۱۲؛ ۲۰۱۳) با استفاده از پیکره و روش‌های دسته‌بندی و با رویکرد فرامعنایی به توصیف و تحلیل معنایی زبان پرداخته است. در این تحلیل‌ها صیغگان کاربرد- اشارت[1] معرفی شده توسط لاینز (۱۹۷۷) به‌کار برده شده است. در تهیه پیکره از جملات منابع مختلف از جمله ویکی‌پدیای انگلیسی استفاده شده‌است.

مدننی و همکاران (۲۰۱۲) موضوعی با نام زبان پوششی در گفتمان استدلالی را مطرح کرده‌اند که می‌تواند برای ادعاها و اثبات‌ها و سازماندهی گفتمان کاربرد داشته باشد. آنان دو مدل پیشنهاد کرده‌اند که مدل اول قاعده- بنیان بوده و از ۲۵ قاعد منظم و توالی n-تایی[2] (توالی ۱ تا ۹) واژه‌ها تشکیل شده است. مدل دوم آماری بوده و از «میدان تصادفی شرطی»[3] (لافرتی و همکاران، ۲۰۰۱) برای تهیه یک مدل دنباله‌ای[4] احتمالاتی مبتنی بر بسامد واژه‌ها استفاده شده است.

الحاربی (۲۰۱۶) اقدام به برچسب‌گذاری ساختاری دو پیکره متشکل از سخنرانی‌های علمی در حوزه فیزیک و اقتصاد کرده است. در این برچسب‌گذاری عبارات مربوط به شروع سخنرانی یا تأکید بررسی شده است. در تهیه این دو پیکره، از یک سامانه سنتز گفتار برای تبدیل گفتار به نوشتار استفاده شده است. سپس با پیشنهاد دو مدل پردازشی، این داده برای آموزش و ارزیابی مدل‌ها به‌کار رفته است. در مدل اول از ماشین بردار پشتیبان[5] (وپنیک، ۱۹۹۸) برای دسته‌بندی[6] برچسب‌های فراگفتمان استفاده شده است. برای کاهش مشکل تُنُک‌بودن داده‌ها از توالی n-تایی واژه‌ها استفاده شده است. در مدل دوم از تعبیه واژگانی[7] بر اساس مدل کیسه‌واژه پیوسته[8] (پنینگتون و همکاران، ۲۰۱۴) استفاده شده است و این نوع بازنمایی در یک مدل دنباله‌ای در یک «شبکه عصبی پیچشی»[9] به‌منظور برچسب‌گذاری فراگفتمان به‌کار رفته است. براساس نتایج عملی، مدل شبکه عصبی بهتر از دسته‌بند[10] ماشین بردار پشتیبان عمل کرده است.

1. use-mention
2. n-gram
3. Conditional Random Field (CRF)
4. sequence
5. Support Vector Machine (SVM)
6. classification
7. word embedding
8. Continuous Bag Of Word (CBOW)
9. Convolutional Neural Network (CNN)
10. classifier

بکتیک (۲۰۱۷) به واشکافی ابزار «تجزیه‌گر افزایشی زیراکس»[1] (ایت‌مختار و همکاران، ۲۰۰۲) که برای دسته‌بندی فراگفتمان نوشتار دانشجویان و برچسب‌زنی کنش بلاغت تهیه شده است پرداخته است.

دوسانتوکوریا (۲۰۱۸) پژوهش خود را به ارائه دو مدل دسته‌بندی فراگفتمان گفتاری متمرکز کرده است. پس‌از تهیه داده، وی از دو دسته‌بند ماشین بردار پشتیبان و میدان تصادفی شرطی در دو مرحله استفاده کرده است. در مرحله اول با استفاده از دسته‌بندهای درخت تصمیم[2] و ماشین بردار پشتیبان مشخص می‌شود کدام جملات حاوی فراگفتمان است. در مرحله دوم با استفاده از میدان تصادفی شرطی واژه‌های خاص فراگفتمان مشخص می‌شود. ویژگی‌هایی که در این مدل به‌کار برده شده است عبارت است از مقوله دستوری واژه، توالی n–تایی بن‌واژه‌ها و واژه‌ها.

در تحلیل‌های فراگفتمان تعاملی انجام شده برای فارسی در چارچوب نظری هایلند (۲۰۰۵) عمدتاً از پیکره زبانی استفاده شده است؛ ولی این پیکره‌های برچسب‌گذاری شده به‌صورت عمومی در دسترس نیست. از این‌رو، در هر پژوهش به‌صورت جداگانه داده‌های متنوعی تحلیل شده است، مانند شکوهی و طلعتی باغ‌سیاهی (۲۰۰۹)، عبدی و احمدی (۲۰۱۵)، رضایی و همکاران (۲۰۱۵)، رضاقلی‌فامیان (۱۳۹۳)، طارمی و همکاران (۱۳۹۷؛ ۱۳۹۸) و مانند آن. تا جایی‌که نویسندگان مقاله می‌دانند در حوزه برچسب‌گذاری خودکار فراگفتمان داده‌های فارسی پژوهشی انجام نشده است و تنها پژوهش انجام شده کاربرد نظر هایلند (۲۰۰۵) در خلاصه‌سازی خودکار بوده است که توسط تاجر و همکاران (۱۳۹۸) انجام شده است.

۳- چارچوب نظری

هایلند (۲۰۰۵) نگرشی را در تحلیل فراگفتمان فراهم آورده است که میان نویسنده متن، خواننده متن و خود متن نوعی رابطه تعاملی ایجاد می‌شود. در این تحلیل، زبان که ابزار ارتباطی بین این سه ضلع است فراتر از نقش ارتباطی خود رفته و به‌مثابه یک کنش اجتماعی تحلیل می‌شود. به اعتقاد هایلند، نویسنده خواسته یا ناخواسته تلاش می‌کند با بکارگیری ابزارهای مختلف زبانی با خواننده در تعامل باشد. از این‌رو، این رویکرد تعاملی به دو دسته تقسیم شده است: دسته اول فراگفتمان موضع‌گیری و دسته دوم فراگفتمان تعاملی مشارکتی است. در فراگفتمان دسته اول، گزاره‌ها و اطلاعات به‌کار رفته در متن به‌گونه‌ای سازماندهی می‌شود که سبب می‌شود متن برای خواننده منسجم و جلوه‌کننده به‌نظر برسد. وی معتقد است واژه‌ها و عبارت می‌تواند نقش نشانگرهای فراگفتمان را بازی کنند. نشانگرهای فراگفتمان موضع‌گیری عبارت است از *گذار*[3] مانند «و» که سبب ایجاد ارتباط میان پاره‌گفتارهای مختلف متن می‌شود؛ *قالب‌نما*[4] مانند «نخست» که بیانگر نوعی حد و مرز در متن است؛ *درون‌متنی*[5] مانند «در بخش بعدی» که خواننده به بخش‌های دیگر متن

1. Xerox incremental parser
2. decision tree
3. transition
4. frame marker
5. endophoric marker

ارجاع داده می‌شود؛ *گواه‌نما*[1] مانند «چامسکی (۱۹۵۷)» که بیانگر نوعی گواه و شاهد از منبعی دیگر است و *ابهام‌زدا*[2] مانند «به عبارتی دیگر» که نویسنده قصد بازگویی مطلبی را دارد که قبلاً مطرح شده و قرار است شرح و تفصیل به آن اضافه گردد.

در فراگفتمان دسته دوم خواننده در متن مشارکت کرده و چشم‌انداز نویسنده نسبت به اطلاعات گزاره‌ای و خواننده مورد توجه است. نشانگرهای فراگفتمان تعاملی مشارکتی عبارت است از *تردیدنما*[3] مانند «به نظر می‌رسد» که بیانگر تردید نویسنده در مورد یک پاره‌گفتار است؛ *یقین‌نما*[4] مانند «قطعاً» که بیانگر قطعیت نویسنده در مورد یک پاره‌گفتار است؛ *نگرش‌نما*[5] مانند «خوشبختانه» یا «متأسفانه» که بیانگر نگرش مثبت یا منفی نویسنده در مورد یک پاره‌گفتار است؛ *خوداظهاری*[6] مانند «نگارنده» که به‌طور مستقیم یک پاره‌گفتار به خود نویسنده ارجاع داده می‌شود و *مشارکت‌نما*[7] مانند «فرض کنید» که بین متن و خواننده ارتباط آشکاری برقرار می‌گردد.

در این پژوهش تلاش می‌شود در چارچوب نظر هایلند (۲۰۰۵) نشانگرهای فراگفتمان موضع‌گیری مقالات فارسی متعلق به رشته علوم انسانی مشخص شود و یک پیکره برچسب‌خورده به‌دست آید. سپس، از این داده برای آموزش یک مدل برچسب‌گذاری آماری دنباله‌ای استفاده گردد.

٤- مدل پیشنهادی

برای پیاده‌سازی مدل پیشنهادی برچسب‌زنی نشانگرهای فراگفتمانی نیاز به یک ساختار مبتنی بر یادگیری ماشین است تا بتواند برچسب‌زنی دنباله‌ها را انجام دهد. مدل پیشنهادی باید بتواند یک دنباله از واژه‌های متن را دریافت نماید و یک دنباله از برچسب‌های مناسب برای واژه‌ها را تولید نماید. برای این هدف، روش‌های مختلفی مبتنی بر مدل‌های احتمالاتی گرافیکی و مدل‌های مبتنی بر یادگیری عمیق[8] ارائه شده است. در این میان مدل‌های عمیق مبتنی بر «شبکه‌های عصبی انتقالی»[9]، مانند برت[10]، توانسته است به بهترین نتایج برسد.

1. evidential
2. code gloss
3. hedge
4. booster
5. attitude marker
6. self-mention
7. engagement marker
8. deep learning
9. Transformer neural network model
10. Bidirectional Encoder Representations from Transformers (BERT)

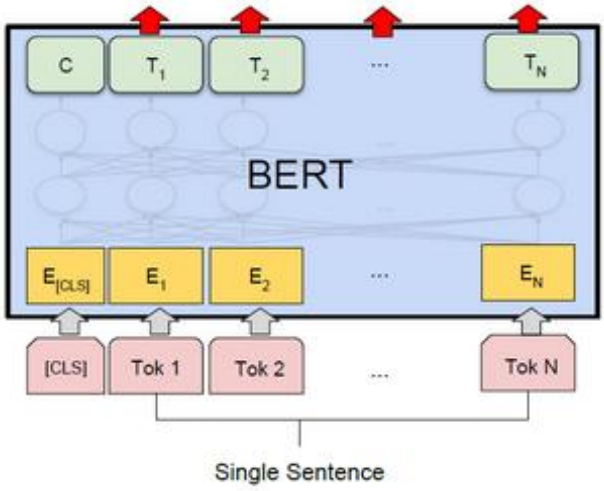
٤-١- برچسب‌زنی دنباله با مدل زبانی برت

برت (دولین و همکاران، ۲۰۱۹) مدل پردازشی است که از آموزش مدل انتقال دو طرفه که به مدل توجه[1] نیز معروف است برای مدل‌سازی زبان استفاده شده است. این مدل پردازشی که توسط گروه پژوهشی در شرکت گوگل تهیه شده است با ارائه نتایج فوق‌العاده در طیف گسترده‌ای از حوزه‌های پردازش زبان طبیعی باعث ایجاد حرکتی نوینی در یادگیری ماشین و پردازش زبان طبیعی شده است. فناوری استفاده شده در برت برخلاف روش‌های قبلی است که در آنها دنباله متن چپ‌به‌راست یا ترکیبی از چپ‌به‌راست و راست‌به‌چپ بررسی می‌شود. نتایج عملی کاربرد برت نشان داده است که یک مدل زبانی که به‌صورت دو طرفه آموزش می‌بیند می‌تواند نسبت‌به مدل‌های زبانی یک طرفه، درک عمیق‌تری از زبان داشته باشد. در مدل برت، از روشی به‌نام مدل‌سازی زبان پوشانده شده[2] استفاده شده است که امکان آموزش دو طرفه در مدل‌هایی که قبلاً غیرممکن بود فراهم می‌گردد.

برت از مدل انتقال استفاده می‌کند و شامل یک فرایند توجه است که در آن روابط متنی بین واژه‌ها در یک متن یاد می‌گرفته می‌شود. مدل انتقال شامل دو فرایند جداگانه است: الف) کدگذار که ورودی متن را می‌خواند و ب) رمزگشا که پیش‌بینی را انجام می‌دهد. از آنجا که هدف برت تولید یک مدل زبانی است، فقط فرایند کدگذار لازم است (وسوانی و همکاران، ۲۰۱۷).

برخلاف مدل‌های جهت‌دار، که ورودی متن را به‌صورت متوالی (از چپ‌به‌راست یا راست‌به‌چپ) می‌خواند، کدگذار مدل انتقال کل دنباله واژه‌ها را یکجا می‌خواند. بنابراین این شیوه مدل‌سازی دوجهته تلقی می‌شود. این ویژگی به مدل اجازه می‌دهد تا بافت جایگاهی یک واژه را بر اساس واژه‌های همسایه خود در سمت چپ و راست واژه هدف بیاموزد. شکل (۱) نمای کلی از برت را نشان داده است. ورودی این مدل دنباله‌ای از واژه‌ها است که در شکل با tok_n نمایش داده شده است. این دنباله‌ها ابتدا به‌صورت بردارهای اولیه بازنمایی شده (E_n) و سپس در شبکه عصبی پردازش می‌شود. خروجی این مدل، دنباله‌ای از بردارهای پنهان (T_n) است که هر بردار آن با واژه ورودی (tok_n) مطابقت دارد.

1. Attention Model
2. Masked Language Modeling



شکل ۱: نمایی از مدل برت (دولین و همکاران، ۲۰۱۹)

یکی‌از مشکلات آموزش مدل‌های زبانی، تعیین نوع فعالیت پردازشی هدف است. بسیاری از این مدل‌ها استفاده از دنباله واژه‌های دیده شده قبلی برای پیش‌بینی واژه‌های بعدی را به‌عنوان فعالیت پردازشی هدف انتخاب کرده است. اما پیش‌بینی واژه بعدی به‌دلیل استفاده یک‌جهته از متن، برای مثال راست‌به‌چپ در فارسی یا چپ‌به‌راست در انگلیسی، سبب محدود شدن یادگیری کامل متن می‌شود. برای رفع این محدودیت، برت از دو روش آموزش استفاده می‌کند:

- مدل‌سازی زبان پوشانده‌شده:

قبل از ورود دنباله‌ای از واژه‌ها به برت، پانزده درصد واژه‌ها در هر دنباله با یک نشان [MASK] جایگزین می‌شود. سپس مدل سعی می‌کند بر اساس متن ارائه شده توسط واژه‌های پوشانده نشده دیگر در دنباله، این واژه‌های پوشانده شده را پیش‌بینی کند. از نظر فنی، پیش‌بینی واژه‌های خروجی مستلزم اضافه کردن یک لایه طبقه‌بندی پس از خروجی کدگذار و سپس استفاده از تابع «بیشینه نرم»[1] است. تابع هزینه[2] برت فقط پیش‌بینی مقادیر پوشانده شده را در نظر می‌گیرد و پیش‌بینی واژه‌های غیر از آنها را نادیده می‌گیرد. اگر چه همگرایی در فرایند بهینه‌سازی این مدل از مدل‌های جهت‌دار آهسته‌تر است، مدل‌سازی انجام شده نسبت به سایر مدل‌سازی‌ها بهتر است.

- مدل پیش‌بینی جمله بعدی:

در فرایند آموزش برت، مدل دو جمله را به‌عنوان ورودی دریافت می‌کند و از این داده یاد می‌گیرد که پیش‌بینی کند آیا جملات ورودی دو جمله متوالی در پیکره است یا خیر. برای آموزش مدل پیش‌بینی جمله

1. Softmax function
2. Cost function

بعدی، نیمی از ورودی‌ها دو جمله متوالی در پیکره است، و نیم دیگر ورودی‌ها دو جمله تصادفی از پیکره انتخاب می‌شود. برای کمک به مدل در تشخیص دو جمله در آموزش، ورودی ابتدا به ساختاری تغییر می‌کند که در ابتدای جمله اول، نشانه [CLS] و در انتهای هر جمله نشانه [SEP] درج می‌شود. سپس کل دنباله ورودی از مدل انتقال عبور می‌کند. خروجی نشانه [CLS] با استفاده از یک لایه طبقه‌بندی ساده به یک بردار ۲ بُعدی تبدیل می‌شود که هر بُعد مشخص می‌کند آیا دو جمله متوالی است یا خیر. سپس با استفاده از تابع بیشینه نرم محاسبه احتمال این که دو جمله متوالی است یا خیر انجام می‌شود.

همان‌طور که گفته شد می‌توان مدل‌های زبانی مبتنی بر انتقال را برای فعالیت‌های مختلف پردازش متن تطبیق داد. برای این کار کافی است با توجه به فعالیت مدنظر، یک لایه به مدل اصلی اضافه نمود. در این پژوهش از مدل برت برای برچسب‌گذاری فراگفتمان تعاملی استفاده می‌کنیم. به‌منظور تشخیص برچسب‌های فراگفتمان، مدل برت متن ورودی را به‌صورت دنباله‌ای از واژه‌ها دریافت می‌کند و به هر واژه یک برچسب تخصیص می‌دهد و به این صورت متن را نشانه‌گذاری می‌کند. با استفاده از برت، ابتدا بازنمایی برداری هر واژه به‌دست می‌آید و سپس این بازنمایی به‌عنوان داده ورودی به یک لایه دسته‌بندی تمام متصل یا میدان تصادفی شرطی منتقل می‌شود تا برچسب آن واژه را پیش‌بینی کند. به این ترتیب، یک مدل تشخیص برچسب فراگفتمان آموزش داده می‌شود. همچنین در ساختاری دیگر از لایه شبکه عصبی بازگشتی همراه با میدان شرطی تصادفی استفاده می‌کنیم و نتایج را برای هر مدل ارزیابی می‌کنیم.

مدل برت برای زبان انگلیسی آموزش داده شده است. برای بکارگیری از این مدل با داده‌های زبان فارسی باید از مدل‌هایی که برای زبان فارسی تهیه شده است استفاده نمود. «مدل بین زبانی مدل روبرتا»[1] (کونیو و همکاران، ۲۰۲۰) و پارس‌برت[2] (فراهانی و همکاران، ۲۰۲۰) مدل‌های زبانی توسعه‌یافته مدل برت بوده و قابلیت پردازش داده‌های فارسی را دارد و در این پژوهش از این دو مدل استفاده می‌کنیم.

٤-۲- مدل‌های زبانی مبتنی بر برت برای زبان فارسی

٤-۲-۱- مدل بین زبانی مدل روبرتا

کونیو و همکاران (۲۰۲۰) مدلی با نام «مدل بین زبانی مدل روبرتا» را معرفی کرده‌اند که بازنمایی میان‌زبانی را برای حدود ۱۰۰ زبان در برگرفته است. این مدل ایکس.ال.ام-روبرتا نیز نامیده می‌شود. مدل ایکس.ال.ام-روبرتا توانسته است با استفاده از داده‌های آموزش غنی‌تر، مدل‌های چندزبانه پیشین، مانند برت چندزبانه (ام.برت)[3]، را بهبود دهد. همچنین، این مدل، زبان‌های بیشتر از جمله زبان‌هایی که داده‌های برچسب‌گذاری شده بسیار محدود دارد را نیز پوشش می‌دهد. نتایج عملی گزارش شده توسط کونیو و همکاران (۲۰۲۰) نشان داده‌است ایکس.ال.ام-روبرتا اولین مدل چندزبانه است که از مدل‌های تک‌زبانه بهتر عمل می‌کند.

1. proposed Cross Language Model RoBERTa (XLM-RoBERTa)
2. ParsBERT
3. https://github.com/google-research/bert

٤-۲-۲- پارس برت

پارس برت یک مدل تک‌زبانه مبتنی بر معماری برت است که توسط فراهانی و همکاران (۲۰۲۰) تهیه شده است. این مدل با استفاده از یک پیکره زبانی ترکیبی فارسی که دربرگیرنده سبک‌ها و ژانرهای مختلفی، از جمله علمی، رمان، خبری و مانند آن، است آموزش داده شده است. این پیکره ترکیبی متشکل‌از حدود ۳/۹ میلیون متن فارسی است که حاوی ۷۳ میلیون جمله و ۱/۳ میلیارد واژه می‌باشد.

۵- نتایج عملی

۵-۱- داده‌های آزمایش‌ها

همان‌طور که در بخش ۴ مطرح شد، هدف از انجام این پژوهش برچسب‌گذاری خودکار عناصر فراگفتمانی تعاملی موضع‌گیری در فارسی براساس رویکرد هایلند (۲۰۰۵) است. برای این هدف نیاز است مجموعه داده‌ای برای فارسی در چارچوب این نظریه تهیه شود. در راستای اهداف این پژوهش، چکیده مقالات موجود در پرتال جامع علوم انسانی[1] به‌واسطه خزش گردآوری شد. ویژگی اسناد خزش‌شده این است که مقوله محتوایی هر یک از چکیده‌ها از سمت درشت‌دانگی به ریزدانگی مشخص شده است. تعداد کل اسناد علمی خزش‌شده ۳۷۴۹۱۵ چکیده می‌باشد. از آنجا که اطلاعات بعضی از اسناد علمی، مانند چکیده، سال انتشار، عنوان و نام نویسنده، مشخص نبود، این موارد پالایش شد. علاوه بر این موارد، چکیده‌های نوشته شده به زبانی جز فارسی، مانند زبان عربی یا انگلیسی، به‌صورت الگوریتمی در مجموعه داده خزش‌شده حذف شد تا فقط داده فارسی بماند و بتواند به‌عنوان یک پیکره در این پژوهش مورد استفاده قرار گیرد. آنچه از پالایش‌های انجام شده باقی ماند، تعداد ۱۱۴۱۷۰ چکیده است که در جدول (۱) اطلاعات آماری پیکره به‌دست آمده براساس تفکیک هر مقوله موجود است. این مجموعه داده به‌طور کلی حاوی ۱۰۹۲۲۶۰۶ واژه و ۲۳۷۷۲۸ واژه نماینده در چکیده بدون احتساب عنوان مقاله است. تنوع واژگانی مجموعه داده گردآوری شده ۰/۰۲۱۸ است که خیلی زیاد نیست. این نکته بیانگر این است که در متون علمی علی‌رغم کاربرد واژه‌های تخصصی، از تنوع واژگانی کمی برای خلق یک متن استفاده می‌شود.

همان‌طور که در این جدول مشخص است، بیشترین مقالات خزش‌شده از پرتال جامع علوم انسانی به حوزه «مدیریت و حسابداری» و کمترین تعداد به حوزه «مطالعات زنان» متعلق است. بیشترین طول متوسط چکیده‌ها از نظر تعداد واژه‌های به‌کار رفته به حوزه «جغرافیا» تعلق داشته و کمترین طول متوسط چکیده‌ها به حوزه «علوم سیاسی و روابط بین‌الملل» تعلق دارد. به‌طور کلی، تنوع واژگانی به‌کار رفته در متون علمی خیلی بالا نیست و متن‌ها به نسبت ساده به نظر می‌رسد.

1. http://ensani.ir

جدول ۱: اطلاعات آماری داده‌های خزش‌شده از پرتال جامع علوم انسانی

مقوله	تعداد اسناد علمی	تعداد واژه	تعداد واژه نماینده	متوسط واژه در سند	تنوع واژگانی
ادبیات	۸۸۲۶	۸۰۶۲۳۹	۵۲۱۸۰	۹۱/۳۵	۰/۰۶۴۷
اقتصاد	۹۹۶۵	۹۳۲۶۵۸	۳۹۳۱۵	۹۳/۵۹	۰/۰۴۲۲
تاریخ	۵۰۰۴	۴۶۲۲۶۷	۳۵۰۶۳	۹۲/۳۸	۰/۰۷۵۹
تربیت بدنی	۴۴۵۴	۵۰۴۶۶۰	۲۹۰۳۳	۱۱۳/۳۰	۰/۰۵۷۵
جغرافیا	۱۰۳۷۰	۱۲۴۳۸۴۲	۵۳۴۷۶	۱۱۹/۹۵	۰/۰۴۳۰
حقوق	۵۲۴۰	۴۸۵۰۶۳	۲۶۸۲۶	۹۲/۵۷	۰/۰۵۵۳
روانشناسی و علوم تربیتی	۱۳۵۴۳	۱۳۷۳۰۳۱	۴۹۲۷۱	۱۰۱/۳۸	۰/۰۳۵۹
زبان‌شناسی	۲۵۹۱	۲۳۶۶۱۶	۲۳۳۵۸	۹۱/۳۲	۰/۰۹۸۷
علوم اجتماعی و ارتباطات	۸۸۷۲	۸۵۴۴۳۱	۴۲۸۹۳	۹۶/۳۱	۰/۰۵۰۲
علوم اسلامی	۱۱۴۰۱	۹۶۵۲۱۵	۴۹۸۹۱	۸۴/۶۶	۰/۰۵۱۷
علوم سیاسی و روابط بین‌الملل	۶۵۹۶	۵۴۹۸۱۳	۳۰۹۱۰	۸۳/۳۶	۰/۰۵۶۲
علوم کتابداری	۲۶۰۰	۲۴۸۳۶۵	۲۰۴۷۲	۹۵/۵۳	۰/۰۸۲۴
فلسفه و منطق	۴۸۶۳	۴۱۳۶۰۷	۳۰۷۴۱	۸۵/۰۵	۰/۰۷۴۳
مدیریت و حسابداری	۱۴۵۶۹	۱۳۰۷۸۵۹	۵۰۲۵۱	۸۹/۷۷	۰/۰۳۸۴
مطالعات زنان	۲۰۷۷	۱۸۰۹۴۵	۱۸۱۳۷	۸۷/۱۲	۰/۱۰۰۲
مطالعات هنر	۳۱۹۹	۳۵۷۹۹۵	۲۸۹۸۲	۱۱۱/۹۱	۰/۰۸۱۰

از ۱۶ حوزه متعلق به رشته علوم انسانی در پرتال جامع علوم انسانی، به‌صورت تصادفی تعداد ۵۰ چکیده به‌ازای هر حوزه و جمعاً تعداد ۸۰۰ چکیده از مجموعه داده خزش‌شده انتخاب شد و واژه‌های این مجموعه داده براساس رویکرد هایلند (۲۰۰۵) و فراگفتمان موضع‌گیری برچسب‌گذاری شد. در چارچوب این نظریه، تعداد ۵ برچسب «گذار»، «قالب‌نما»، «درون‌متنی»، «گواه‌نما» و «ابهام‌زدا» تعریف شد و به‌صورت دستی توسط فرد خبره در حوزه تحلیل فراگفتمانی کار برچسب‌گذاری چکیده‌ها انجام شد.

در برچسب‌زنی داده‌ها از شیوه برچسب‌زنی استاندارد IOB استفاده شده است که پیش‌تر رامشاو و مارکوس (۱۹۹۵) آن را برای تجزیه سطحی[1] و قلاب‌گذاری گروه اسمی در داده‌های زبان انگلیسی معرفی استفاده کرده‌اند. در این استاندارد، برچسب‌گذاری در سطح واژه است؛ ولی قابلیت گسترش در سطح عبارت را داد. در این استاندارد، B به معنی شروع برچسب برای یک واژه یا عبارت است؛ I به معنی واژه یا واژه‌ها درونی یک عبارت حاوی برچسب بوده و O به معنی سایر واژه‌هایی است که برچسب‌های پنج‌گانه فراگفتمان موضع‌گیری به آنها تعلق نمی‌گیرد. در مثال (۱) نمونه‌ای از برچسب‌گذاری فراگفتمانی موضع‌گیری یک

1. shallow parsing

چکیده مربوط به حوزه «تربیت بدنی» نمایش داده شده است. برچسب‌های «گذار»، «قالب‌نما»، «درون‌متنی»، «گواه‌نما» و «ابهام‌زدا» به‌ترتیب با کدهای ۱ تا ۵ براساس استاندارد IOB تعیین شده‌است.

(۱)

هدفB2 اینI2 پژوهشI2 سنجشO مدیریتO آشوبO وO هنجارسنجیO ابزارO اندازه‌گیریO مدیریتO آشوبO برO اساسO الگویO تئوریO آشوبO درO سازمانO تربیتO بدنیO جمهوریO اسلامیO ایرانO استO .O اینB2 تحقیقI2 درO دوO مرحله‌O انجامO گرفتO ،O درB2 مرحله‌I2 اولI2 باO بررسیO پیشینه‌O تحقیقO وO نظریه‌O آشوبO وO باO استفاده‌O ازO روشO تحلیلO محتوایO کیفیO ازO طریقO مصاحبه‌O باO ده‌O متخصصO مدیریتO وO مدیریتO ورزشیO، O مفهومO مدیریتO آشوبO تبیینO شدO .O یافته‌هایO حاصلO ازO بخشO اولO مدلO مدیریتO آشوبO خودمولدO راO معرفیO کردO .O برO اساسO مدلO طراحیO شده‌O وO باO استفاده‌O ازO نظریه‌O آشوبO ،O عناصرO وO ابعادO اصلیO اینO مفهومO باO عنوانO درونمایه‌هایO چهارگانه‌O (B5 اثرI5 پروانه‌ایI5 ،I5 سازگاریI5 پویاI5 ،I5 جاذبه‌هایI5 عجیبI5 ،I5 خودمانایی‌I5) I5 درO نظرO گرفته‌O شدندO .O درB2 مرحله‌I2 دومI2 برO اساسO مدلO مدیریتO آشوبO خودمولدO ،O گویه‌هایO مقیاسO مدیریتO آشوبO استخراجO شدO .O به‌منظورO بررسیO رواییO سازه‌O یاO محتوایO پرسشنامه‌O ازO تحلیلO عاملیO اکتشافیO استفاده‌O شدO کهB1 بیانگرO وجودO چهارO عاملO به‌عنوانO عواملO تشکیل‌دهنده‌O سازه‌O مدیریتO آشوب‌گونه‌O بودO کهB1 عبارتندB5 ازI5 ساختارO سازمانیO ،O مهارتO مدیرانO ،O شرایطO وO فضایO سازمانO وO کارکنانO .O پایاییO پرسشنامه‌O نیزO ازO طریقO روشO آلفایO کرونباخO (O ۰/۹۲O) O کهO مؤیدO همسانیO درونیO مناسبO ابزارO استO ،O تأییدO شدO .O نتایجB2 نشانO دادO سطحO تحصیلاتO مدیرانO وO سابقه‌O مدیریتO باO میزانO مدیریتO آشوب‌گونه‌O ارتباطO مستقیمO داردO .O سپسB2 باO استفاده‌O ازO اینO مقیاسO مدیریتO آشوبO درO سازمانO تربیتO بدنیO بررسیO شدO وB1 یافته‌هاO حاکیO ازO آنO بودO کهB1 مدیریتO آشوبO درO سازمانO به‌طورO معناداریO درO حدO متوسطیO قرارO داردO وB1 باتوجه‌O بهO اینکه‌O ارتباطO معناداریO بینO مقدارO هرO کدامO ازO ابعادO مدیریتO آشوبO وO اهمیتO ابعادO وجودO نداشتO ،O می‌توانO استنباطO کردO کهB1 مدیرانO سازمانO تربیت بدنیO درO وضعیتO موجودO ازO مزایاO وO آثارO مثبتO مهارتO مدیریتO آشوبO به‌منظورO تصمیم‌گیریO ،O بهبودO کیفیتO وO ...O بهره‌مندO نیستندO .O

در جدول (۲) اطلاعات آماری مربوط به داده برچسب‌خورده استفاده شده در این پژوهش ارائه شده است. همان‌گونه که مشخص است، کمترین تعداد نشانگرهای فراگفتمان با ۵/۳۹ درصد در چکیده‌های حوزه «تاریخ» استفاده شده و بیشترین نشانگرهای فراگفتمان با ۸/۶۷ درصد در چکیده‌های حوزه «اقتصاد» به‌کار برده شده است.

در جدول (۳) تنوع برچسب‌های فراگفتمان تعاملی به‌کار رفته در تحلیل داده‌های علمی منتخب گزارش شده است. براساس برچسب‌های تخصیص داده شده در چکیده‌های منتخب، سه فراگفتمان موضع‌گیری «گذار»، «قالب‌نما» و «ابهام‌زدا» بیشترین و «گواه‌نما» کمترین کاربرد را در داده‌های مربوط به حوزه‌های مختلف علوم انسانی داشته است. حوزه‌های «تربیت بدنی»، «حقوق»، «علوم اجتماعی و ارتباطات»، «علوم کتابداری»، «فلسفه و منطق» و«مطالعات زنان» حاوی نشانگرهای پنج‌گانه فراگفتمان تعاملی بوده است.

جدول ۲: اطلاعات آماری داده‌های برچسب‌خورده

مقوله	تعداد واژه	تعداد واژه نماینده	تنوع واژگانی	تعداد برچسب در سطح عبارات	تعداد برچسب در سطح واژه	درصد واژه برچسب‌دار
ادبیات	۱۰۹۱۶	۳۰۳۳	۰/۲۷۷۸	۴۷۳	۸۰۳	۷/۳۶
اقتصاد	۱۰۵۵۶	۲۱۹۲	۰/۲۰۷۷	۵۰۲	۹۱۵	۸/۶۷
تاریخ	۹۸۱۲	۲۹۱۶	۰/۲۹۷۲	۳۲۵	۵۲۹	۵/۳۹
تربیت بدنی	۱۱۸۲۶	۲۳۰۸	۰/۱۹۵۲	۳۸۸	۸۱۲	۶/۸۷
جغرافیا	۱۴۴۵۸	۳۱۱۰	۰/۲۱۵۱	۵۱۵	۹۰۸	۶/۲۸
حقوق	۹۹۳۱	۲۶۲۱	۰/۲۶۳۹	۳۹۹	۶۷۶	۶/۸۱
روانشناسی و علوم تربیتی	۱۲۰۷۷	۲۲۷۰	۰/۱۸۸۰	۴۱۳	۸۶۷	۷/۱۸
زبان‌شناسی	۱۰۴۸۴	۲۶۷۳	۰/۲۵۵۰	۴۳۶	۸۴۹	۸/۱۰
علوم اجتماعی و ارتباطات	۱۱۵۷۶	۲۶۲۶	۰/۲۲۶۸	۵۱۱	۹۳۴	۸/۰۷
علوم اسلامی	۹۰۵۵	۲۴۶۵	۰/۲۷۲۲	۳۵۲	۵۱۹	۵/۷۳
علوم سیاسی و روابط بین‌الملل	۹۲۲۸	۲۴۰۸	۰/۲۶۰۹	۳۴۶	۵۵۹	۶/۰۶
علوم کتابداری	۱۰۷۶۷	۲۱۱۰	۰/۱۹۶۰	۴۳۴	۷۵۸	۷/۰۴
فلسفه و منطق	۱۰۰۷۴	۲۵۷۹	۰/۲۵۶۰	۴۵۲	۷۲۳	۷/۱۸
مدیریت و حسابداری	۱۰۱۵۸	۲۵۶۳	۰/۲۵۲۲	۳۸۲	۶۸۰	۶/۶۹
مطالعات زنان	۹۷۳۹	۲۲۹۳	۰/۲۲۵۴	۳۹۴	۶۹۰	۷/۰۸
مطالعات هنر	۱۲۲۰۱	۳۲۲۹	۰/۲۶۴۷	۴۵۳	۷۶۰	۶/۲۳

جدول ۳: توزیع آماری برچسب‌های فراگفتمان موضع‌گیری به‌کاررفته در تحلیل داده‌های علمی

مقوله	گذار	قالب‌نما	درون‌متنی	گواه‌نما	ابهام‌زدا
ادبیات	۲۹۰	۳۴۴	۰	۶	۱۶۴
اقتصاد	۳۲۴	۲۸۷	۲	۰	۳۰۵
تاریخ	۱۹۹	۱۸۶	۱۵	۰	۱۲۸
تربیت بدنی	۱۸۳	۲۹۷	۵	۲	۳۳۴
جغرافیا	۲۸۸	۳۲۶	۵	۰	۲۹۲
حقوق	۲۸۹	۲۱۴	۱۶	۲	۱۶۵
روانشناسی و علوم تربیتی	۲۰۰	۳۱۴	۱۰	۰	۳۴۴
زبان‌شناسی	۲۶۵	۲۵۹	۱۳	۰	۳۱۸
علوم اجتماعی و ارتباطات	۲۵۹	۳۷۸	۶	۱۲	۲۷۸
علوم اسلامی	۲۶۷	۱۹۱	۲	۰	۶۱
علوم سیاسی و روابط بین‌الملل	۲۳۴	۲۱۳	۸	۰	۱۰۵
علوم کتابداری	۲۲۸	۳۵۴	۹	۸	۱۵۹
فلسفه و منطق	۳۲۸	۲۱۰	۱۱	۶	۱۷۳
مدیریت و حسابداری	۲۲۷	۲۶۴	۷	۰	۱۸۵
مطالعات زنان	۲۳۶	۲۶۲	۷	۱۸	۱۷۲
مطالعات هنر	۲۶۱	۳۲۰	۱۲	۰	۱۷۱
جمع کل	۴۰۷۸	۴۴۱۹	۱۲۸	۵۴	۳۳۵۴

۵-۲- نتایج به‌دست‌آمده

در انجام این پژوهش سه مدل برچسب‌گذاری فراگفتمان موضع‌گیری مقالات فارسی در رشته علوم انسانی معرفی می‌شود. در مدل اول از یک شبکه عصبی «حافظه بلند-کوتاه‌مدت دوطرفه»[1] و دسته‌بند میدان تصادفی شرطی به‌عنوان مدل پایه جهت مقایسه با دو مدل دیگر استفاده می‌شود. در این مدل از بازنمایی ورد۲وک معرفی شده توسط بلای و همکاران (۲۰۱۳) به‌عنوان یکی از شیوه‌های معروف برای بازنمایی معنایی واژه‌ها در فضای برداری استفاده می‌شود. در این نوع بردارسازی، برای هر واژه یک بردار تهیه شده است. برای آموزش مدل، از بردارهای از پیش آموزش‌دیده مبتنی بر ورد۲وک که توسط هادی‌فر و ممتازی (۲۰۱۸) تهیه شده است استفاده می‌شود. بردار واژه‌های این داده که در ۱۰۰ بُعد تهیه شده است از یک پیکره ۱/۱۳ میلیارد واژه‌ای متشکل از منابعی مانند توییتر فارسی، ویکی‌پدیای فارسی، پیکره همشهری (آل‌احمد و همکاران، ۲۰۰۹) و وبلاگ ایران (آل‌احمد و همکاران، ۲۰۱۶) به‌دست آمده است.

1. Bidirectional Long-Short Term Memory (BiLSTM)

در مدل دوم، از بازنمایی برت (دولین و همکاران، ۲۰۱۹) به‌عنوان یکی از شیوه‌های «آخرین وضعیت روز»[1] استفاده می‌شود. در بخش ۴-۲ دو مدل برت که قابلیت پردازش داده‌های فارسی را دارد، یعنی پارس برت (فراهانی و همکاران، ۲۰۲۰) و ایکس.ال.ام-روبرتا (کونیو و همکاران، ۲۰۲۰)، معرفی شد که در این پژوهش از این دو نسخه استفاده می‌شود. یکی از ویژگی‌های تهیه بردار واژه‌ها توسط برت این است که در هنگام بازنمایی، براساس بافت‌های متفاوتی که یک واژه در آن ظاهر می‌شود، یک بردار برای واژه تهیه می‌گردد. در این دو مدل، از یک دسته‌بند شبکه عصبی متصل که لایه آخر آن از میدان تصادفی شرطی استفاده شده است بهره برده می‌شود. از آنجا که محدودیت در مقدار داده برچسب‌خورده برای ۱۶ حوزه رشته علوم انسانی وجود دارد، از روش ارزیابی متقاطع[2] ۵ بخشی استفاده کرده‌ایم به این صورت که تمامی داده‌های موجود به ۵ قسمت متوازن از تمامی حوزه‌ها تقسیم می‌شود؛ سپس، چهار بخش برای آموزش مدل و یک بخش برای ارزیابی مدل استفاده می‌گردد. در جدول (۴) نتایج میانگین آزمایش‌های عملی برای برچسب‌گذاری فراگفتمان موضع‌گیری گزارش شده است. نتایج گزارش شده در سطح واژه و عبارت می‌باشد.

جدول ٤: نتایج مدل‌های برچسب‌گذاری فراگفتمان تبادلی

بازنمایی		سطح واژه			سطح عبارت		
		دقت	فراخوانی	امتیاز F	دقت	فراخوانی	امتیاز F
مدل اول	وردتو‌وک	۷۹/۱۴	۵۹/۴۵	۶۷/۷۶	۶۵/۷۰	۶۲/۲۱	۶۳/۸۵
مدل دوم	پارس برت	۸۰/۵۴	۸۵/۵۷	۸۲/۹۶	۷۷/۱۹	۸۵/۶۰	۸۱/۱۶
مدل سوم	ایکس.ال.ام-روبرتا	۷۸/۵۷	۸۷/۵۷	۸۲/۸۲	۷۶/۳۶	۸۷/۴۱	۸۱/۵۱

همان‌گونه که از نتایج مشخص است، مدل‌های دوم و سوم که از بازنمایی برت بهره می‌برد در سطح واژه و عبارت براساس آزمون تی[3] با تفاوت معناداری نسبت به مدل اول که از بازنمایی وردتو‌وک استفاده کرده است عمل کرده است ($p<۰/۰۱$). این تفاوت بیانگر این نکته است که در مدل‌هایی که در بازنمایی معنایی واژه، بافت جایگاهی واژهٔ هدف مورد توجه قرار گرفته است، درک مدل بیشتر شده و به افزایش کارایی مدل انجامیده است. از میان دو مدل مبتنی بر برت، مدل دوم اندک کارایی بهتر در حدود ۰/۱۶ درصد در امتیاز F را نسبت به مدل سوم در سطح واژه به‌دست آورده است که این تفاوت معنادار نبود. در سطح عبارت مدل سوم نسبت به مدل دوم در حدود ۰/۳۵ درصد در امتیاز F کارایی بالاتر به‌دست آورده است که این تفاوت نیز معنادار نبود. در ارزیابی در سطح واژه، خروجی مدل‌ها با شش برچسب (پنج برچسب فراگفتمان موضع‌گیری و یک برچسب «سایر») طلایی در داده استاندارد طلایی مقایسه شده است. در ارزیابی در سطح عبارت، باید ضمن تشخیص درست برچسب برای هر واژه، بایست برچسب‌های تمام یک عبارت از ابتدا تا انتها درست

1. state-of-the-art
2. cross validation
3. t-test

تشخیص داده شده باشد. از این‌رو، معیار ارزیابی سخت‌تری محسوب می‌شود. هرگونه اشتباه موجب کاهش امتیاز در ارزیابی برچسب‌دهی عبارت می‌شود. تفاوت مدل دوم در سطح واژه و عبارت از نظر آماری معنادار نبود، در حالی‌که تفاوت مدل سوم در سطح واژه و عبارت ازنظر آماری معنادار بود ($p<۰/۰۵$). از آنجا که میانگین کارایی مدل سوم در سطح واژه و عبارت در مقابل با مدل دوم بالاتر بود، تحلیل نتایج را متمرکز مدل پردازشی مدل سوم، یعنی مدل ایکس.ال.ام-روبرتا، می‌کنیم.

۵-۳- بحث و بررسی

با توجه به کارایی بالای مدل‌های مبتنی بر برت، برای بررسی دقیق مدل‌ها، دو بررسی انجام دادیم. در بررسی اول به بررسی وضعیت برچسب‌های شش‌گانه (پنج‌گانه فراگفتمان موضع‌گیری و برچسب «سایر») در دو مدل مبتنی بر بازنمایی برت در مقایسه با برچسب‌های طلایی پرداختیم. نتیجه میانگین ۵ بخشی این دو مدل در ماتریس درهم‌ریختگی[1] (۵) گزارش شده است. در این جدول، ردیف بیانگر برچسب‌های طلایی و ستون بیانگر برچسب‌های پیش‌بینی شده توسط مدل‌ها است. از میان برچسب‌ها، عملکرد مدل ایکس.ال.ام-روبرتا در نشانه‌گذاری برچسب‌های «گذار»، «قالب‌نما»، «ابهام‌زدا» و «سایر» بهتر بود؛ ولی مدل برای برچسب‌های «درون‌متنی» و «گواه‌نما» به‌درستی عمل نکرد. به‌نظر می‌رسد به‌دلیل تُنُک‌بودن داده برای این دو برچسب، مدل به‌خوبی آموزش ندیده است و نتوانسته است پیش‌بینی مناسبی برای این برچسب‌ها ارائه کند. برچسب «سایر» بیشترین خطا را در فرایند برچسب‌گذاری فراهم آورده است که دلیل آن وجود حجم زیادی از داده‌ها با این برچسب است.

جدول ۵: ماتریس درهم‌ریختگی برچسب‌های مدل ایکس.ال.ام-روبرتا

	گذار	قالب‌نما	درون‌متنی	گواه‌نما	ابهام‌زدا	سایر
گذار	۳۸۰۵	۱۵	۰	۰	۲	۲۵۷
قالب‌نما	۴۲	۳۹۶۶	۰	۰	۰	۴۷۳
درون‌متنی	۰	۴۵	۰	۰	۰	۸۴
گواه‌نما	۰	۰	۰	۰	۱۲	۳۱
ابهام‌زدا	۵	۰	۰	۰	۳۱۷۹	۴۴۰
سایر	۹۴۰	۷۱۳	۰	۰	۱۰۴۶	۱۵۲۳۴۳

در بررسی دوم، کارایی میانگین ۵ بخشی مدل‌های سوم را براساس تفکیک حوزه‌های ۱۶گانه علوم انسانی در سطح واژه و عبارت محاسبه کردیم تا عملکرد مدل در حوزه‌های مختلف رشته علوم انسانی مشخص شود. براساس نتایج گزارش شده در جدول (۶)، از میان حوزه‌های ۱۶گانه علوم انسانی، مدل

1. confusion matrix

ایکس.ال.ام-روبرتا در سطح واژه بالاترین کارایی را در حوزه «تاریخ» با ۸۶/۰۸ درصد در امتیاز F و کمترین کارایی را در حوزه «مطالعات زنان» با ۷۸/۱۲ درصد در امتیاز F را به‌دست آورده است. همچنین، این مدل در سطح عبارت بالاترین کارایی را در حوزه «تاریخ» با ۸۳/۸۴ درصد در امتیاز F و کمترین کارایی را در حوزه «مطالعات زنان» با ۷۸/۳۸ درصد در امتیاز F را به‌دست آورده است. دلیل عمده کارایی پایین مدل در این حوزه این است که داده‌های این حوزه حاوی بیشترین تعداد برچسب‌های «درون‌متنی» و «گواه‌نما» است که مدل روی آنها به‌خوبی آموزش ندیده است.

جدول ٦: نتایج مدل ایکس.ال.ام-روبرتا برچسب‌گذاری فراگفتمان تعاملی حوزه‌های ١٦گانه علوم انسانی

مقوله	سطح واژه			سطح عبارت		
	دقت	فراخوانی	امتیاز F	دقت	فراخوانی	امتیاز F
ادبیات	۷۱/۰۳	۸۹/۷۲	۷۹/۲۹	۷۱/۷۹	۸۷/۲۳	۷۸/۷۶
اقتصاد	۷۸/۸۹	۹۰/۸۲	۸۴/۴۳	۷۶/۹۵	۸۹/۶۹	۸۲/۸۳
تاریخ	۸۳/۸۸	۸۸/۴۰	۸۶/۰۸	۷۹/۸۱	۸۸/۲۹	۸۳/۸۴
تربیت بدنی	۸۰/۷۶	۸۶/۶۱	۸۳/۵۸	۷۸/۹۳	۸۷/۳۶	۸۲/۹۳
جغرافیا	۷۹/۰۰	۸۸/۳۷	۸۳/۴۲	۷۷/۱۷	۸۷/۰۹	۸۱/۸۳
حقوق	۸۱/۸۷	۸۹/۲۲	۸۵/۳۹	۷۶/۶۷	۹۰/۰۶	۸۲/۸۳
روانشناسی و علوم تربیتی	۸۲/۳۳	۸۶/۵۰	۸۴/۳۶	۷۸/۱۸	۸۶/۳۰	۸۲/۰۴
زبان‌شناسی	۷۲/۴۸	۸۶/۲۴	۷۸/۷۶	۷۰/۹۵	۸۵/۶۷	۷۷/۶۲
علوم اجتماعی و ارتباطات	۷۹/۹۴	۸۸/۵۹	۸۴/۰۵	۷۵/۷۵	۸۷/۷۰	۸۱/۲۹
علوم اسلامی	۸۲/۱۰	۸۴/۹۹	۸۳/۵۲	۷۹/۵۷	۸۶/۱۲	۸۲/۷۲
علوم سیاسی و روابط بین‌الملل	۷۹/۴۸	۸۹/۰۷	۸۴/۰۰	۷۶/۱۰	۹۰/۲۸	۸۲/۵۸
علوم کتابداری	۷۸/۹۷	۸۵/۲۹	۸۲/۰۱	۷۷/۶۳	۸۶/۰۰	۸۱/۶۰
فلسفه و منطق	۷۸/۱۳	۸۵/۶۶	۸۱/۷۲	۷۷/۹۷	۸۶/۶۲	۸۲/۰۷
مدیریت و حسابداری	۷۶/۲۳	۸۸/۷۷	۸۲/۰۲	۷۴/۴۱	۸۶/۳۸	۷۹/۹۵
مطالعات زنان	۷۳/۴۶	۸۳/۴۲	۷۸/۱۲	۷۱/۴۵	۸۶/۸۰	۷۸/۳۸
مطالعات هنر	۷۸/۲۱	۸۶/۳۹	۸۲/۱۰	۷۵/۰۳	۸۶/۹۴	۸۰/۵۵

٦- جمع‌بندی و نتیجه‌گیری

در این مقاله فرایند برچسب‌گذاری فراگفتمان موضع‌گیری مقالات علمی نگارش شده به زبان فارسی توضیح داده شد. سپس نتایج به‌دست آمده گزارش شده و مورد بررسی قرار گرفت. در چارچوب این پژوهش، تعداد ۸۰۰ مقاله فارسی متعلق به ۱۶ حوزه در رشته علوم انسانی در چارچوب نظری هایلند (۲۰۰۵) بررسی شد و

نشانگرهای فراگفتمان موضع‌گیری نشانه‌گذاری شد. برای این هدف، پنج برچسب فراگفتمان و یک برچسب «سایر» برای موارد دیگر به‌جز برچسب‌های پنج‌گانه فراگفتمان تعاملی تعریف شد. با استفاده از داده‌های تهیه شده سه مدل پردازشی آموزش دید. تفاوت مدل‌های برچسب‌زنی در نحوه بازنمایی معنایی واژه‌ها بود. در مدل اول از بازنمایی برداری ورد۲وک استفاده شد که در آن برای هر واژه فقط یک بردار موجود بود. در دو مدل دیگر از بازنمایی مبتنی‌بر برت به‌نام‌های پارس‌برت و ایکس.ال.ام–روبرتا بهره برده شد. در بازنمایی برت برای هر بافت واژه هدف یک بردار تهیه شد. براساس نتایج به‌دست آمده، مدل‌های مبتنی‌بر برت به‌طور معناداری عملکرد بهتری از مدل ورد۲وک داشت. عملکرد مدل پارس‌برت در سطح واژه و عملکرد مدل ایکس.ال.ام–روبرتا در سطح عبارت بهتر بود. از میان این دو مدل، میانگین عملکرد ایکس.ال.ام–روبرتا بهتر بود و بررسی‌های بعدی روی این مدل متمرکز شد. از میان برچسب‌های پنج‌گانه فراگفتمان تعاملی، به‌دلیل وجود حجم بالای داده آموزش کارایی مدل برای سه برچسب «گذار»، «قالب‌نما» و «ابهام‌زدا» بهتر از دو برچسب «درون‌متنی» و «گواه‌نما» بود. همچنین به‌دلیل کاربرد زیاد برچسب «سایر» در داده، این برچسب سبب ایجاد خطا برای برچسب‌های فراگفتمان موضع‌گیری شد.

نتیجه انجام این پژوهش این است که می‌توان از روش‌های نوین برچسب‌زنی در حوزه زبان‌شناسی رایانشی با دقت قابل قبولی برای نشانه‌گذاری واژه‌ها و عبارات فراگفتمان تعاملی استفاده کرد. یکی از روش‌های افزایش کارایی مدل به‌خصوص برای دو برچسبی که به‌خوبی آموزش ندیده است افزایش حجم داده آموزش است؛ ولی تهیه این نوع داده ساده نیست. این موضوع می‌تواند در پژوهش‌های آتی مورد توجه قرار بگیرد.

تقدیر و تشکر

این پژوهش در چارچوب طرح پژوهشی ۲۸۱۱۱ در مجموعه طرح‌های «طرح جامع اعتلای علوم انسانی معطوف به پیشرفت کشور» در پژوهشگاه علوم انسانی و مطالعات فرهنگی انجام پذیرفته است.

منابع

– تاجر، پگاه؛ جوکار، عبدالرسول؛ فخراحمد، سید مصطفی؛ خرمایی، علیرضا؛ و ستوده، هاجر (۱۳۹۶). «کاربرد تحلیل گفتمان در خلاصه‌سازی خودکار متون علمی،» در *مجموعه مقالات نخستین همایش ملی رویکردهای نوین در مطالعات زبان و ادبیات*، مؤسسه آموزش عالی زند، شیراز، ایران.

– تاجر، پگاه؛ جوکار، عبدالرسول؛ فخراحمد، سید مصطفی؛ ستوده، هاجر؛ و خرمایی، علیرضا (۱۳۹۸). «تحلیل کاربرد الگوی فراگفتمان هایلند در خلاصه‌سازی خودکار استناد مدار: پیشنهاد طرح حاشیه‌نویسی بافتارهای استنادی»، *کتابداری و اطلاع‌رسانی*، ۲۲(۳): ۹۱–۱۱۱.

– طارمی، طاهره؛ تاکی، گیتی و یوسفیان، پاکزاد (۱۳۹۷). «جنسیت در مقالات علمی فارسی زبان: مطالعه پیکره بنیاد نشانگرهای فراگفتمان تعاملی بر اساس انگاره هایلند»، *پژوهش‌های زبان‌شناسی*، ۱۰(۱): ۲۳-۴۱.

– طارمی، طاهره؛ تاکی، گیتی؛ و یوسفیان، پاکزاد (۱۳۹۸). «واکاوی پیکره‌بنیاد فراگفتمان تعاملی در مقالات علمی پژوهشی فارسی: انگاره هایلند (۲۰۰۵)»، *پژوهش‌های زبانی*، ۱۰(۲): ۱۲۹-۱۵۱.

– رضاقلی‌فامیان، علی (۱۳۹۳). «موضع‌گیری و مشارکت‌جویی در مقالات نقد کتاب‌های ادبیات فارسی»، *فصلنامه تخصصی نقد ادبی*، ۲۶: ۴۹-۶۶

– Abdi, R.; & Ahmadi, P. (2015). "Research article introductions and disciplinary influences based on interactive metadiscourse markers", *Journal of Modern Research in English Language Studies*, 2(1): 99-85.

– AleAhmad, A.; Amiri, H.; Darrudi, E.; Rahgozar, M.; & Oroumchian, F. (2009). "Hamshahri: A standard Persian text collection", *Knowledge-based Systems*, 2: 382–387.

– AleAhmad, A.; Zahedi, M. S.; Rahgozar, M.; & Moshiri, B. (2016). "IrBlogs: A standard collection for studying Persian bloggers", *Computers in Human Behavior*, 57: 195–207.

– Alharbi, G. (2016). *Metadiscourse Tagging in Academic Lectures*, PhD Dissertation, University of Sheffield.

– Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; & Stoyanov, V. (2020). "Unsupervised cross-lingual representation learning at scale", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

– Crismore, A; Markkanen R.; & Steffensen, M. (1993). "Metadiscourse in persuasive writing: a study of texts written by American and Finnish university students", *Written Communication*, 10: 39-71.

– Devlin, J.; Chang, M.W.; Lee, K.; & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp: 4171–4186, Minneapolis: Association for Computational Linguistics.

– Dos Santos Correia, R. P. (2018). *Automatic Classification of Metadiscourse*. PhD Dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.

– Farahani, M.; Gharachorloo, M.; Farahani, M.; & Manthouri, M. (2020). "ParsBERT: Transformer-based Model for Persian Language Understanding," arXiv preprint arXiv: 2005. 12515.

– Hadifar, A.; & Momtazi, S. (2018). "The impact of corpus domain on word representation: A study on Persian word embeddings", *Lang Resources & Evaluation*, 52(4): 997-1019.

– Heng, C. S.; & Tan, H. (2010). "Extracting and comparing the intricacies of metadiscourse of two written persuasive corpora", *International Journal of Education and Development Using Information and Communication Technology*, 6 (3): 124-146.

– Harris, Z. S. (1970). "Linguistic transformations for information retrieval," in *Papers in Structural and Transformational Linguistics*. Dordrecht: Springer.

- Hyland, K. (2005). *Metadiscourse: Exploring Interaction in Writing*. London: Continuum.
- Lafferty, J.; McCallum, A.; & Pereira, F. C. N. (2001). "Conditional random fields: Probabilistic models for segmenting and labelling sequence data", in *Proceedings of the 18th International Conference on Machine Learning*. Morgan Kaufmann.
- Lyons, J. (1977). *Semantics*. vol. 2. Cambridge University Press.
- Madnani, N.; Heilman, M.; Tetreault, J.; & Chodorow, M. (2012). "Identifying high-level organizational elements in argumentative discourse", in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp: 20–28. Association for Computational Linguistics.
- Mann, W. C.; & Thompson, S. A. (1988). "Rhetorical structure theory: Toward a functional theory of text organization", *Text: Interdisciplinary Journal for the Study of Discourse*, 8 (3): 243–281.
- Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MITpress.
- Marcus, M. P; Marcinkiewicz, M. A; & Santorini, B. (1993) "Building a large annotated corpus of English: The Penn treebank", *Computational Linguistics*, 19(2): 313–330.
- Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; & Dean, J. (2013). "Distributed representations of words and phrases and their compositionality", in *Advances in Neural Information Processing Systems 26*, eds. Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., Curran Associates, Inc., pp. 3111–3119.
- Pennington, J; Socher, R; & Manning; C.D. (2014). "Glove: Global Vectors for word representation", in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, vol. 14: 1532–1543.
- Rezaei, S; Estaji, M; & Hasanpour, M. (2015). "Examining the interactional metadiscourse markers in Iranian MA applied linguistics theses", *Journal of Modern Research in English Language Studies*, 2(1): 71-43.
- Shokouhi, H; & Talati Baghsiahi, A. (2009). "Metadiscourse functions in English and Persian sociology articles: A study in contrastive rhetoric", *Studies in Contemporary Linguistics*, 45(4): 549-568.
- Vapnik, V. N. (1998). *Statistical Learning Theory*, Wiley-Interscience.
- Vaswani, A; Shazeer, N; Parmar, N; Uszkoreit, J; Jones, L; Gomez, A. N; Kaiser, L & Polosukhin, I. (2017). "Attention is all you need", in *Proceedings of the 31st Conference on Neural Information Processing Systems*, pp: 5998-6008.
- Webber, B; & Joshi, A. (1998). "Anchoring a lexicalized tree-adjoining grammar for discourse", in *Proceedings of the Joint Conference on Computational Linguistics and the Association for Computational Linguistics Workshop on Discourse Relations and Discourse Markers*, pp: 86–92.
- Wilson, S. (2010). "Distinguishing use and mention in natural language", in *Proceedings of the Student Research Workshop at 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology*, pp: 29–33. Association for Computational Linguistics.
- Wilson, S. (2012). "The creation of a corpus of English metalanguage", in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp: 638–646. Association for Computational Linguistics.

– Wilson, S. (2013). “Toward automatic processing of English metalanguage”, in *Proceedings of International Joint Conference on Natural Language Processing*, 760-766.
– Yan, L. (2015). "Comparative analysis of Russian news reporting and news commentary in metadiscourse applications", in *Proceedings of the International Conference on Informatization in Education, Management and Business*. 1089-1089.