Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال سوم، شماره اول، بهار و تابستان ۱٤۰٤، صفحات ۱۲٦ - ۱۱۱

ارائه یک روش ترکیبی داده‌افزایی برای تحلیل احساسات با تمرکز بر رفع چالش‌های متون غیررسمی فارسی

غزل مهرپرور[۱]، سمیرا نوفرستی[۲*]

[۱]دانشجوی کارشناسی ارشد، گروه مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران

[۲]دانشیار، گروه مهندسی فناوری اطلاعات، دانشکده مهندسی برق و کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران

مقاله پژوهشی

تاریخ دریافت:
۱٤۰٤/٤/۱۸
تاریخ پذیرش:
۱٤۰٤/٦/۱۵

کلیدواژه‌ها:
داده‌افزایی، تحلیل احساسات، چالش‌های زبان فارسی، تحلیل احساسات بین‌دامنه‌ای، یادگیری عمیق

نویسنده‌مسئول:
snoferesti@ece.usb.ac.ir

چکیده

با افزایش روزافزون حجم نظرات کاربران در شبکه‌های اجتماعی و بسترهای خدماتی، تحلیل احساسات به یکی از ابزارهای کلیدی در استخراج بینش از داده‌های متنی تبدیل شده است. با این حال، چالش‌هایی نظیر کمبود داده‌های برچسب خورده و پیچیدگی‌های زبانی به‌ویژه در زبان فارسی، عملکرد مدل‌های یادگیری عمیق را با محدودیت‌هایی مواجه کرده است. در این مقاله، روشی برای داده‌افزایی با هدف بهبود تحلیل احساسات متون فارسی ارائه شده است که تمرکز آن بر تولید داده‌های مصنوعی با تکیه بر چالش‌های مطرح در تحلیل نظرات شبکه‌های اجتماعی از جمله کوتاه بودن، زبان عامیانه و اشتباهات املایی و دستوری فراوان در این نوع متون است. همچنین روشی مبتنی بر داده‌افزایی برای بهبود تحلیل احساسات بین‌دامنه‌ای پیشنهاد شده است. نتایج ارزیابی‌های انجام‌گرفته نشان می‌دهد که روش پیشنهادی در دامنه‌های توئیتر و فیلم معیار F1 طبقه‌بند شبکه عصبی پیچشی برای تحلیل احساسات را به ترتیب ٦/۳ و ۸/۲ درصد بهبود داده است. همچنین دو روش پیشنهادی برای داده‌افزایی بین‌دامنه‌ای (غنی‌سازی با واژه‌های دامنه مقصد و داده‌افزایی با ChatGPT)، توانسته‌اند معیار F1 مدل آموزش دیده بر روی دامنه توئیتر را در تحلیل احساسات دامنه هتل به میزان ۹/۸ و ۱٤/٦ درصد و در دامنه فیلم به میزان ۲/٦ و ۰/٤ درصد افزایش دهند.

: 10.22034/ABMIR.2025.23361.1141

E-ISSN: 2821-2037

۱- مقدمه

در سال‌های اخیر، شبکه‌های اجتماعی به بخشی جدایی‌ناپذیر از زندگی روزمره انسان‌ها تبدیل شده‌اند. این بسترها به کاربران امکان می‌دهند تا افکار، احساسات، تجربیات و نظرات خود را به‌صورت عمومی یا خصوصی به اشتراک بگذارند. این روند باعث شکل‌گیری حجم عظیمی از داده‌های متنی شده است که تحلیل آن‌ها می‌تواند بینش‌های ارزشمندی دربارهٔ رفتار کاربران، روندهای اجتماعی و حتی نیازهای بازار فراهم کند.

علاوه بر شبکه‌های اجتماعی، وب‌سایت‌های خدماتی نظیر سایت‌های خرید اینترنتی، رستوران‌یاب‌ها و نرم‌افزارهای خدمات شهری نیز میزبان حجم زیادی از نظرات کاربران هستند. این محتواها می‌توانند اطلاعات ارزشمندی درباره احساسات، نیازها و تجربیات افراد ارائه دهند. با این حال، تحلیل دستی این حجم از داده‌ها هم بسیار زمان‌بر و پرهزینه است و هم از دقت و انسجام لازم برخوردار نیست، از این رو، نمی‌تواند برای تصمیم‌گیری‌های برخط مورد استفاده قرار گیرد.

در این میان، تحلیل احساسات[1] به‌عنوان یکی از زیرشاخه‌های مهم پردازش زبان طبیعی[2]، نقش کلیدی در تفسیر و ارزیابی خودکار محتوای متنی نظرات کاربران ایفا می‌کند. هدف تحلیل احساسات، شناسایی و تعیین حالت عاطفی یا نگرشی است که در متن نظرات، پیام‌ها یا محتوای شبکه‌های اجتماعی وجود دارد. این احساسات می‌توانند مثبت، منفی یا خنثی باشند و در برخی موارد می‌توان آن‌ها را به دسته‌بندی‌های دقیق‌تری مانند خوشحالی، ناراحتی، عصبانیت و غیره نیز تقسیم کرد [۱].

در سال‌های اخیر، تحلیل احساسات با استفاده از الگوریتم‌های یادگیری عمیق کانون توجه محققان بوده و پیشرفت‌های خیره‌کننده‌ای در این زمینه حاصل شده است [۲, ۳]. با این وجود، مسیر پیش رو در این حوزه همچنان با چالش‌های مهمی مواجه است. یکی از این چالش‌ها، کمبود داده‌های برچسب خورده باکیفیت و متوازن است که آموزش مدل‌های یادگیری ماشین را دشوار می‌سازد. علاوه بر این، به دلیل کوتاه بودن متن نظرات کاربران و همچنین زبان عامیانه و اشتباهات رایج املایی و دستوری در این گونه نظرات، مدل‌های یادگیری عمیق در درک و تحلیل احساسات آن‌ها با مشکل مواجه می‌شوند.

در زمینه حل چالش کمبود داده‌های برچسب خورده، روش‌های داده‌افزایی[3] با هدف تولید داده‌های مصنوعی بر اساس داده‌های موجود، می‌توانند به بهبود عملکرد مدل‌های یادگیری عمیق کمک شایانی نمایند. روش‌های موجود برای داده‌افزایی را می‌توان به سه دسته کلی تقسیم کرد: دسته اول مبتنی بر بازنویسی[4] متن هستند و با روش‌هایی مانند ترجمه متن به زبانی دیگر و ترجمه مجدد به زبان مبدأ جملات جدید می‌سازند [٤, ٥]. دسته دوم که مبتنی بر تغییر کلمات جمله هستند با کمک عملیات ساده مانند حذف، درج، جابجایی یا جایگزینی تصادفی کلمات یا علائم نگارشی در یک جمله سعی در تولید جملات جدید دارند [٦–٨]. دسته سوم مبتنی بر مدل‌های زبانی بزرگ هستند و از روش‌هایی نظیر اصلاح جملات و یا مخفی کردن برخی از کلمات متن و پیش‌بینی آن‌ها برای تولید جملات جدید استفاده می‌کنند][۹–۱۲].

روش‌های سنتی داده‌افزایی عمدتاً متکی بر عملیات تصادفی و غیرهوشمندانه هستند. هدف این مقاله، ارائه روشی برای داده‌افزایی در کاربرد تحلیل احساسات است که ضمن غلبه بر مشکل کمبود داده‌های آموزشی، با تمرکز بر اصلاح اشتباهات املایی و دستوری و غنی‌سازی معنایی متن، جملات مصنوعی با کیفیت‌تری تولید کند. همچنین، در این مقاله نشان می‌دهیم که چگونه داده‌افزایی می‌تواند به بهبود دقت الگوریتم‌های تحلیل احساسات، حتی در دامنه‌هایی فراتر از مجموعه داده آموزشی، کمک کند.

نتایج ارزیابی‌های انجام‌گرفته نشان می‌دهد حالت هشتم روش پیشنهادی برای داده‌افزایی در مقایسه با روش شناخته شده EDA [٦]، روش مبتنی بر تولید و اصلاح جملات با کمک ChatGPT [۱۳] و روش پیشنهادی در [٥]، در دامنه توئیتر به ترتیب ۳/۸، ۱/۸ و ۳/۲ درصد و در دامنه فیلم به ترتیب ٤/۲، ۳/٦ و ۳ درصد معیار F1 طبقه‌بند شبکه عصبی پیچشی[5] (CNN) برای تحلیل احساسات

[1] Sentiment Analysis
[2] Natural Language Processing
[3] Data Augmentation
[4] Paraphrasing
[5] Convolutional Neural Network

را افزایش داده است. همچنین، دو روش پیشنهادی برای داده‌افزایی بین‌دامنه‌ای[1]، به نام‌های غنی‌سازی با واژه‌های دامنه مقصد و داده‌افزایی با ChatGPT، معیار F1 شبکه عصبی آموزش دیده برای تحلیل احساسات توئیت‌ها را در تعیین قطبیت نظرات کاربران در دامنه هتل به میزان قابل توجه ۹/۸ و ۱۴/۶ درصد بهبود بخشیده است. این بهبود در دامنه فیلم به ترتیب ۲/۶ و ۰/۴ درصد گزارش شده است.

به صورت خلاصه نوآوری‌های روش پیشنهادی عبارت‌اند از:

- ارائه روشی جدید برای داده‌افزایی با تمرکز بر مشکلات متون عامیانه نوشته شده به زبان فارسی
- بهبود دقت تحلیل احساسات بین‌دامنه‌ای با استفاده از داده‌افزایی

ادامه مقاله به‌صورت زیر سازمان‌دهی شده است: در بخش دوم، مروری بر کارهای پیشین مرتبط با داده‌افزایی در حوزه تحلیل احساسات ارائه می‌شود. بخش سوم به معرفی روش پیشنهادی اختصاص دارد. در بخش چهارم، نتایج حاصل از ارزیابی روش پیشنهادی با استفاده از معیارهای متعدد روی چند مجموعه تست مختلف ارائه می‌گردد. در نهایت، بخش پنجم به نتیجه‌گیری می‌پردازد.

۲- مرورکارهای پیشین

روش‌های داده‌افزایی موجود به دسته‌های مختلفی تقسیم‌بندی می‌شوند که عبارت‌اند از: روش‌های مبتنی بر بازنویسی، روش‌های مبتنی بر تغییرات در سطح کلمه، روش‌های مبتنی بر مدل‌های زبانی بزرگ و رویکردهای ترکیبی. در ادامه هر یک از این دسته‌ها معرفی می‌گردد.

در روش‌های مبتنی بر بازنویسی، تغییرات در سطح جمله یا کل متن اعمال می‌شود و داده‌های مصنوعی از نظر معنایی تفاوت بسیار کمی با داده‌های اصلی دارند. از رایج‌ترین روش‌های این دسته ترجمه برگشتی است. در این روش، ابتدا هر یک از جملات برچسب‌گذاری شده در مجموعه داده اصلی به یک زبان دیگر ترجمه می‌شوند. سپس هر یک از جملات ترجمه‌شده دوباره به زبان اصلی بازگردانده می‌شوند. در آخر نمونه‌های تکراری حذف می‌شوند [٤].

در روش‌های داده‌افزایی مبتنی بر تغییرات در سطح کلمه، با ایجاد اصلاحات جزئی در متن، ضمن حفظ معنای اصلی، تنوع لازم نسبت به داده‌های اولیه ایجاد می‌شود. یکی از مشهورترین روش‌های این دسته، افزایش آسان داده‌ها (EDA[2]) است [٦]. این روش شامل چهار عمل ساده می‌شود: جایگزینی مترادف با کمک منابعی مانند وردنت[3] [١٤]، حذف تصادفی، جابجایی تصادفی و درج تصادفی کلمات. در [٧]، به جای انتخاب تصادفی، کلمات و عملیات داده‌افزایی به صورت هوشمند و بر اساس الگوریتم ژنتیک انتخاب می‌شوند. در [٨] نیز، دو استراتژی برای داده‌افزایی در کاربرد تحلیل احساسات مبتنی بر جنبه[4] پیشنهاد شده است: جایگزینی مترادف با در نظر گرفتن نقش نحوی کلمات و جابجایی کلمه بر اساس روابط وابستگی و دانش دامنه.

روش‌های مبتنی بر مدل‌های زبانی بزرگ را می‌توان به دو زیرگروه روش‌های مبتنی بر ماسک‌گذاری و روش‌های مبتنی بر تولید متون جدید تقسیم کرد. روش‌های مبتنی بر ماسک‌گذاری، با مخفی کردن کلمات خاص و پیش‌بینی آن‌ها با کمک مدل‌هایی مانند BERT، DistilBERT و RoBERTA می‌توانند نسخه‌های جدیدی از متن ایجاد کنند [٩, ١٠]. به صورت دقیق‌تر، این مدل‌ها توکن‌های <mask> را در برخی از موقعیت‌های متن وارد می‌کنند یا برخی از کلمات را با توکن‌های <mask> جایگزین می‌کنند و سپس یک مدل MLM[5] پیش‌بینی می‌کند که چه کلماتی باید در این موقعیت‌های پنهان شده قرار گیرند. از آنجا که MLMها بر روی تعداد زیادی متن پیش‌آموزش داده شده‌اند، این روش معمولاً قادر به تولید متون جدید معنی‌دار است.

برای تولید متون جدید، مدل‌های زبانی مانند GPT می‌توانند متن‌های جدیدی مشابه داده‌های موجود ایجاد کنند که به افزایش تنوع و دقت داده‌ها کمک می‌کند. در [١١] یک روش داده‌افزایی به نام AugGPT معرفی شده است که به ویژه در یادگیری با نمونه‌های کم کاربرد دارد. این روش، هر جمله در مجموعه آموزش

[1] Cross-domain
[2] Easy Data Augmentation
[3] WordNet
[4] Aspect-based Sentiment Analysis
[5] Masked Language Model

را به چندین نمونه مشابه از نظر مفهومی ولی متفاوت از نظر ساختاری بازنویسی می‌کند. در حین بازنویسی، AugGPT تلاش می‌کند تا اطمینان حاصل کند که داده‌های مصنوعی تولید شده با برچسب جمله اولیه همخوانی دارند.

در [۱۲]، روشی مبتنی بر شبکه‌های مولد متخاصم (GAN) برای داده‌افزایی متون فارسی ارائه شده است. معماری پیشنهادی دارای دو بخش اصلی است: ۱) مولد مبتنی بر ParsBERT که کلمات دارای نقش نحوی مهم برای تحلیل احساسات یعنی صفات، اسامی و قیدها را به صورت تصادفی انتخاب و پنهان می‌کند. سپس کلمات پنهان را با تکنیک‌هایی مانند درج کلمه جدید، جایگزینی با مترادف یا جایجایی کلمات تغییر می‌دهد. ۲) متمایز کننده که شباهت کسینوسی جمله جدید را با جمله اولیه می‌سنجد و جملاتی که شباهت آن‌ها از آستانه ۰/۸ بیشتر است را انتخاب می‌کند.

در راستای بهبود عملکرد تحلیل احساسات مبتنی بر جنبه، در [۱۳] سه رویکرد داده‌افزایی مبتنی بر پرسش[۱] از ChatGPT معرفی شده است: ۱) بازنویسی زمینه‌ای که در آن کلمات زمینه‌ای جمله بدون تغییر جنبه و احساس آن بازنویسی می‌شوند؛ ۲) تغییر جنبه که در آن جنبه با واژه‌ای مشابه از نظر معنایی جایگزین می‌گردد؛ و ۳) ترکیب دو رویکرد قبل که بیشترین بهبود عملکرد را در مدل‌های پایه مانند BERT نشان داده است. این رویکردها حفظ ثبات معنایی و قطبیت جملات را با طرح پرسش‌های هدفمند از ChatGPT تضمین می‌کنند. نتایج این پژوهش نشان می‌دهد که استفاده از مدل‌های زبانی بزرگ در تولید داده‌های مصنوعی می‌تواند به شکل مؤثری دقت و تعمیم‌پذیری مدل‌های تحلیل احساسات را افزایش دهد.

در [۱۵] نیز یک روش سه مرحله‌ای برای داده‌افزایی پیشنهاد شده است. در مرحله اول، سه روش برای بازسازی متن معرفی شده است که با کمک آن‌ها تعدادی از جملات مجموعه آموزش که به صورت تصادفی انتخاب شده‌اند، بازسازی شده و زوج‌های (متن اصلی، متن بازنویسی شده) ساخته می‌شوند. در مرحله دوم، این زوج‌ها به عنوان نمونه به مدل‌های زبانی بزرگ ارائه می‌شود و با طراحی پرسش‌های مناسب از مدل خواسته می‌شود که داده‌های جدید تولید کند. در مرحله پایانی نیز داده‌های نادرست حذف می‌شوند.

در برخی از کارهای پیشین نیز، داده‌افزایی با رویکرد ترکیبی انجام شده است. برای مثال، در [۵] برای تولید نمونه‌های مصنوعی ابتدا هر جمله از مجموعه آموزش که به زبان فارسی است با کمک مترجم گوگل به انگلیسی ترجمه می‌شود. سپس برای نقش‌های نحوی تاثیرگذار در تحلیل احساسات یعنی صفات، افعال و اسامی با کمک وردنت کلمات مترادف استخراج می‌شوند. این کلمات استخراج شده با احتمالی جایگزین کلمه اصلی در جمله می‌شوند. در پایان جمله انگلیسی مجدد به فارسی ترجمه می‌شود. در این مقاله نشان داده شده است که ترکیب دو روش ترجمه برگشتی و جایگزینی با کلمات مترادف باعث می‌شود تا حد زیادی بر مشکل تولید داده‌های تکراری که در ترجمه برگشتی برای جملات کوتاه به کرات رخ می‌دهد غلبه شود.

تحقیق [۱۶] یک رویکرد ترکیبی مبتنی بر توضیح را برای داده‌افزایی پیشنهاد می‌کند که به محدودیت‌های روش‌های موجود می‌پردازد. در این رویکرد، ابتدا اهمیت کلمات در تخصیص برچسب حسی مشخص می‌شود و سپس با کمک یک روش ترکیبی برای جایگزین کردن کلمات در جملات مختلف، متون جدید ایجاد می‌شوند. برخلاف روش‌های سنتی که جمع‌بندی وزنی ساده را برای تخصیص برچسب انجام می‌دهند، این روش قدرت معنایی کلمات کلیدی را در نظر می‌گیرد و اطمینان می‌دهد که حس و معنای جملات در حین داده‌افزایی حفظ می‌شود. این روش به‌ویژه در کاربرد تحلیل احساسات تأثیر بسزایی در بهبود عملکرد داشته است زیرا در این کاربرد تغییر کلمات خاص می‌تواند قطبیت یک جمله را به طور قابل‌توجهی تغییر دهد.

در [۱۷]، یک روش ترکیبی داده‌افزایی با هدف بهبود عملکرد مدل‌های تحلیل احساسات مبتنی بر جنبه ارائه شده است. در این روش، برای این که جملات مصنوعی تولید شده معنا و احساسات را نسبت به جنبه مورد توصیف حفظ کنند، از ترجمه برگشتی مبتنی بر درج کاراکتر ویژه (SCI) استفاده شده است. درج SCI از عبارت توصیف کننده جنبه در طول فرآیند ترجمه محافظت می‌کند و

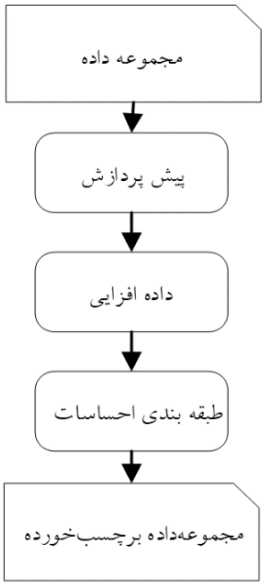
[۱] Prompt

باعث می‌شود عبارت جنبه شکل و مکان اصلی خود را در جملات ترجمه شده حفظ کند. همچنین، برای تنوع‌بخشی به جملات حاصل از ترجمه برگشتی، تکنیک جایگزینی تصادفی موجودیت‌های نامدار پیشنهاد شده است.

علی‌رغم پژوهش‌های متعدد صورت گرفته در زمینه به‌کارگیری تکنیک‌های داده‌افزایی برای بهبود عملکرد الگوریتم‌های تحلیل احساسات در زبان‌های مختلف به‌خصوص زبان انگلیسی، این شاخه تحقیقاتی در زبان فارسی به خوبی مورد مطالعه قرار نگرفته است و بر اساس مطالعات نویسندگان این مقاله، تاکنون به حل مسئله داده‌افزایی با تمرکز بر چالش‌های اصلی متون عامیانه نوشته شده به زبان فارسی پرداخته نشده است.

۳- روش پیشنهادی

مراحل یک سیستم تحلیل احساسات مبتنی بر داده‌افزایی در شکل (۱) نشان داده شده است. همان‌طور که مشاهده می‌شود ابتدا جملات مجموعه آموزش پیش‌پردازش شده و سپس با روش پیشنهادی برای داده‌افزایی گسترش می‌یابند. در پایان نیز با الگوریتم CNN به طبقه‌بندی احساسات در دو دسته مثبت و منفی پرداخته می‌شود. در ادامه جزئیات هرکدام از این مراحل تشریح می‌گردد.



شکل (۱): مراحل تحلیل احساسات مبتنی بر داده‌افزایی

۳-۱ پیش‌پردازش

در اولین مرحله، پیش‌پردازش‌های رایج تقطیع واژه‌ها[1]، نرمال‌سازی و حذف ایست‌واژه‌ها[2] بر روی جملات مجموعه‌داده انجام می‌گیرد. به این صورت که پس از جداسازی واژه‌های یک جمله، با جایگزین کردن کاراکترهای استاندارد در متن، نرمال‌سازی انجام می‌شود. سپس ایست‌واژه‌ها یعنی واژه‌های پرتکراری که از لحاظ معنایی ارزش کمی دارند مانند "اگر" و "به" از متن حذف می‌شوند. ایست‌واژه‌ها در تعیین حس یک جمله نقش موثری ندارند و حذف آن‌ها باعث کاهش بار محاسباتی و افزایش سرعت الگوریتم‌های پردازش متن می‌شود. برای حذف ایست‌واژه‌ها از لیست ایست‌واژه‌های رایج زبان فارسی شامل ۷۰۰ کلمه استفاده شده است [۵]. لازم به ذکر است که کلمات منفی‌کننده مانند "نه" و "هرگز" که نقش بسزایی در تحلیل احساسات دارند از این لیست حذف شده‌اند. در پایان این مرحله نیز با هدف کاهش تنوع متن و افزایش دقت طبقه‌بندی، ریشه‌یابی انجام می‌گردد.

۳-۲ داده‌افزایی

هدف روش پیشنهادی برای داده‌افزایی، تولید نمونه‌های آموزشی جدید با تمرکز بر چالش‌های متون عامیانه نوشته شده توسط کاربران شبکه‌های اجتماعی است. از جمله رایج‌ترین مشکلات، اشتباهات املایی و دستوری مکرر، کوتاه بودن متون و استفاده از عبارات محاوره‌ای است. برای غلبه بر این چالش‌ها، مراحل زیر برای داده‌افزایی پیشنهاد شده است به این صورت که به ازای هر جمله اولیه در مجموعه داده، به کمک روش‌های زیر تعدادی جمله مصنوعی تولید شده و با برچسب مشابه جمله اولیه به مجموعه آموزش اضافه می‌شوند:

۱- اصلاح املای کلمات: در این مرحله، تمامی جملات موجود در مجموعه‌داده بررسی شده و با استفاده از واژه‌نامه و الگوریتم‌های تصحیح املا، کلمات نادرست شناسایی و به شکل صحیح جایگزین می‌شوند. این فرآیند باعث افزایش انسجام زبانی داده‌ها و در نتیجه بهبود عملکرد مدل یادگیری ماشین در مرحله بعد می‌شود. برای مثال عبارت "چند وخ پیش" به صورت " چند وقت پیش" اصلاح می‌شود.

[1] Tokenization

[2] Stopwords

۲- حذف حروف تکراری: در متون غیررسمی، به‌ویژه در شبکه‌های اجتماعی، کاربران معمولاً برای تاکید یا جلب توجه از تکرار حروف استفاده می‌کنند. برای مثال، کلمهٔ «عاااااالی» به جای «عالی» نوشته می‌شود. این نوع نگارش، هرچند در محاوره رایج است، اما باعث افزایش نویز در داده‌ها و کاهش دقت مدل‌های زبانی می‌شود، زیرا این کلمات از نظر نگارشی با فرم اصلی خود تفاوت دارند و در واژه‌نامه‌های استاندارد یافت نمی‌شوند. برای رفع این مشکل، الگوریتمی طراحی و پیاده‌سازی شده است که در آن تکرار حروف شناسایی و حذف می‌شود.

۳- غنی‌سازی متن با کلمات مترادف: به منظور ایجاد تغییر و همچنین غنی‌سازی معنایی جملات، از یک واژه‌نامه از کلمات مترادف کمک گرفته شده است. به این صورت که برای هر جمله، بررسی می‌شود که اگر کلمه‌ای از آن در واژه‌نامه موجود باشد، یکی از کلمات مترادف آن به‌صورت تصادفی جایگزین خواهد شد. این مرحله ۲ بار تکرار می‌شود تا تعداد جملات ایجاد شده بیشتر شود.

٤- غنی‌سازی با واژه‌های حسی: برای این منظور، ابتدا کلمات حسی مجموعه‌داده با استفاده از یک واژه‌نامه حسی که در آن برچسب قطبیت کلمات مشخص است، استخراج می‌شوند. سپس با استفاده از معیار TF-IDF کلمات کم‌رخداد این لیست تعیین می‌شوند. سپس به منظور غنی‌سازی جملات حاوی واژه‌های کم رخداد، با توجه به برچسب جمله سه کلمه حسی به تصادف انتخاب و اضافه می‌گردد. به صورت دقیق‌تر، اگر جمله مثبت باشد، سه کلمه حسی با برچسب مثبت که کم‌رخداد هستند یعنی در مجموعه‌داده کم‌تر دیده شده‌اند، اضافه می‌شود. برای جملات منفی نیز به طور مشابه سه کلمه حسی با برچسب منفی که کم‌رخداد هستند اضافه می‌شود. برای مثال جملهٔ "ای جونم تولدت خیلی مبارک" به "ای جونم تولدت خیلی مبارک و عزت و یادبود و همکاری" تبدیل شده است.

۵- تبدیل عبارات محاوره‌ای به متون رسمی و عاری از اشتباه و برعکس: برای هر جمله از مجموعه داده، با کمک ChatGPT و با استفاده از سه پرسش زیر سه جمله جدید تولید شده است.

- پرسش اول: جمله "..." را بدون غلط املایی و با ساختار دیگری بنویس بدون اینکه قطبیت مثبت یا منفی جمله تغییر کند.
- پرسش دوم: جمله "..." را بدون غلط املایی و به صورت رسمی بنویس بدون اینکه قطبیت مثبت یا منفی جمله تغییر کند.
- پرسش سوم: جمله "..." را بدون غلط املایی و به صورت عامیانه بنویس بدون اینکه قطبیت مثبت یا منفی جمله تغییر کند.

برای رسیدن به جواب‌های دقیق‌تر، پرسش‌ها به‌طور صریح محدود شدند تا صرفاً بازنویسی جمله اصلی با حفظ قطبیت را انجام دهند. همچنین، نمونه‌هایی از جملات تولید شده به‌صورت دستی بازبینی گردید تا از حفظ قطبیت و طبیعی بودن نگارش آن‌ها اطمینان حاصل شود. در پایان نیز، جملات تکراری حذف شدند.

برای مثال برای جمله "اینمی سری از عروسکاش چقد قشنگ نگه میداره بازی میکنه!" جملات زیر ساخته شدند:

- با پرسش اول: "عروسک‌های او در بازی کردن بسیار زیبا و دوست‌داشتنی هستند، همه روحیه و شور و شوق خود را در نگهداری از آن‌ها نشان می‌دهد."
- با پرسش دوم: "این هم یک سری از عروسک‌هایش است که به زیبایی نگه داشته می‌شود و با آن‌ها بازی می‌کند!"
- با پرسش سوم: "این همه عروسک قشنگ داره بازی میکنه! بهت میگم!"

دو مرحله نخست بر اصلاح املای کلمات تمرکز دارند، در حالی که مراحل سوم و چهارم به افزایش طول جملات و تنوع بخشیدن به واژه‌ها می‌پردازند. مرحله پنجم نیز بر تبدیل جملات محاوره‌ای به رسمی و برعکس تأکید دارد. دو مرحله زیر نیز برای داده‌افزایی بین‌دامنه‌ای پیشنهاد شده‌اند:

٦- غنی‌سازی با واژه‌های دامنه مقصد: در این روش ابتدا واژه‌های مهم در دامنه مقصد با استفاده از معیار TF-IDF استخراج می‌شوند. همچنین با کمک یک واژه‌نامه حسی، کلمات حسی

دامنه مقصد استخراج می‌شوند. برای هر جمله در مجموعه‌داده دامنه مبدأ، یک کلمه از کلمات مهم و یک کلمه از کلمات حسی استخراج شده از دامنه مقصد به تصادف انتخاب شده و اضافه می‌شود. کلمه حسی با توجه به قطبیت جمله اضافه می‌شود یعنی اگر جمله مثبت باشد، یک کلمه حسی با قطبیت مثبت اضافه می‌شود و بالعکس.

۷- استفاده از ChatGPT: در این روش برای هر جمله از مجموعه‌داده مبدأ، با پرسش زیر از ChatGPT خواسته شده است که جمله‌ای معادل در دامنه مقصد (برای مثال هتل) تولید کند.

- پرسش: برای جمله "..." در دامنه توئیتر، یک جمله معادل در دامنه هتل و با استفاده از واژه‌های رایج در دامنه هتل بنویس بدون اینکه قطبیت مثبت یا منفی جمله تغییر کند.

۳-۳ طبقه‌بندی

پس از افزایش تعداد نمونه‌های آموزشی، مجموعه‌داده گسترش‌یافته برای آموزش مدل یادگیری عمیق استفاده می‌شود. در این مقاله از شبکه عصبی پیچشی (CNN) برای تعیین قطبیت جملات استفاده شده است. معماری مورد استفاده برای CNN، معماری پیشنهادی در [۵]است که مشخصات آن در جدول (۱) نشان داده شده است.

جدول (۱): معماری مورد استفاده برای شبکه عصبی پیچشی

تعداد لایه‌ها	٦
لایه اول	لایه تعبیه‌شده با بردار ۱۰۰ بعدی برای نمایش واژه‌ها
لایه دوم	لایه پیچشی با ۱۲۸ فیلتر با اندازه ۳
لایه سوم	لایه حذف تصادفی با نرخ ۰/۵
لایه چهارم	لایه ادغام بیشینه با اندازه ۲×۲
لایه پنجم	لایه مسطح‌سازی
لایه ششم	لایه کاملاً متصل
بهینه‌ساز	آدام
نرخ یادگیری	۰,۰۰۱
تابع زیان	sparse_categorical_crossentropy
اندازه دسته	۵۰
تعداد دوره‌ها	۲۰

معماری مورد استفاده دارای ٦ لایه است. لایه اول که لایه تعبیه‌شده[1] است، وظیفه دارد کلمات را به بردارهای عددی با ابعاد ثابت تبدیل کند. بردارهای تعبیه‌شده، بازنمایی معنایی واژه‌ها را فراهم می‌کنند و شباهت‌های معنایی بین کلمات را برای مدل قابل درک می‌سازند.

لایه دوم، یک لایه پیچشی است که در آن، فیلترهایی با اندازه مشخص بر روی توالی کلمات اعمال می‌شوند تا ویژگی‌های محلی مانند عبارات کلیدی را شناسایی کنند. این ویژگی‌ها در تشخیص قطبیت جمله بسیار مؤثرند. لایه سوم لایه حذف تصادفی[2] است که به‌صورت تصادفی بخشی از نورون‌ها را در هر مرحله آموزش غیرفعال می‌کند تا بیش‌برازش[3] را کاهش و توان تعمیم‌دهی مدل را افزایش دهد. در لایه چهارم یعنی ادغام بیشینه[4]، ویژگی‌های استخراج‌شده فشرده‌سازی می‌شوند. با انتخاب بیشینه مقدار در هر ناحیه، مدل قادر است مهم‌ترین اطلاعات را حفظ کند و پیچیدگی محاسباتی را کاهش دهد. لایه مسطح‌سازی[5] خروجی چندبعدی حاصل از لایه‌های قبلی را به یک بردار یک‌بعدی تبدیل می‌کند تا بتوان آن را به لایه خروجی متصل کرد. لایه کاملاً متصل[6] نورون‌های ورودی را به نورون‌های خروجی متصل می‌کند و با استفاده از تابع بیشینه نرم[7] برچسب نهایی (مثبت یا منفی) را پیش‌بینی می‌کند.

٤- نتایج

این بخش شامل معرفی اجمالی مجموعه‌داده‌های مورد استفاده و گزارش نتایج حاصل از ارزیابی‌هایی است که با هدف سنجش کارایی روش پیشنهادی برای داده‌افزایی انجام شده‌اند. پیاده‌سازی و تحلیل داده‌ها با استفاده از زبان برنامه‌نویسی پایتون انجام شده است. به‌منظور ارزیابی عملکرد مدل CNN در تشخیص قطبیت جملات، از معیارهای متداول دقت[8]، فراخوانی[9] و معیار F1استفاده شده است. دقت نشان می‌دهد چه نسبتی از جملات مثبت پیش‌بینی شده، واقعاً مثبت هستند. فراخوانی نشان می‌دهد چه نسبتی از جملات مثبت واقعی، توسط مدل به درستی شناسایی شده‌اند. معیار

[1] Embedding Layer
[2] Dropout
[3] Overfitting
[4] MaxPooling
[5] Flatten layer
[6] Fully Connected Layer
[7] Softmax
[8] Precision
[9] Recall

F1 میانگین هارمونی دقت و فراخوانی است و از رابطه زیر محاسبه می‌شود:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (۱)$$

٤-١ مجموعه‌داده‌های مورد استفاده

در این مقاله، از دو مجموعه‌داده نظرکاوی و تحلیل احساسات در دامنه‌های توئیتر و فیلم استفاده شده است. از مجموعه‌داده توئیت‌های فارسی [۱۸]، ۱۰۰ توئیت مثبت و ۱۰۰ توئیت منفی و از مجموعه‌داده فیلم [۱۹]، ۲۵۰ نظر مثبت و ۲۵۰ نظر منفی، به عنوان مجموعه آموزش طبقه‌بند CNN در نظر گرفته شده است. مجموعه‌های آموزش کوچک درنظر گرفته شده است تا اثربخشی الگوریتم‌های داده‌افزایی زمانی که داده‌های برچسب خورده بسیار اندکی وجود دارند ارزیابی شود.

به منظور ارزیابی کارایی روش پیشنهادی برای داده‌افزایی، ۲۵۰ رکورد از مجموعه‌داده توئیتر و ۱۰۰۰ رکورد از مجموعه داده فیلم به صورت تصادفی به عنوان مجموعه‌های تست انتخاب شده‌اند. همچنین از یک مجموعه تست دیگر شامل ۲۳٤ جمله در دامنه متمایز هتل جهت سنجش اثربخشی روش پیشنهادی برای تحلیل احساسات بین‌دامنه‌ای استفاده شده است [۲۰].

جزئیات مجموعه داده‌های مورد استفاده در جدول (۲) نشان داده شده است. تفاوت اصلی مجموعه‌داده‌های انتخاب شده از دو دامنه هتل و توئیتر در سبک نگارش و طول نظرات است. جملات در دامنه هتل معمولاً طولانی‌تر، رسمی‌تر و دارای ساختار نحوی درست‌تری هستند. در مقابل، توئیت‌ها سرشار از اشتباهات املایی و دستوری، اصطلاحات عامیانه و نمادهای خاص مانند شکلک و هشتگ هستند. جملات دامنه فیلم در مقایسه با دامنه توئیتر رسمی‌تر و در مقایسه با دامنه هتل محاوره‌ای‌تر هستند. همچنین، متوسط طول نظرات در دامنه فیلم بسیار بیشتر از دامنه‌های هتل و توئیتر است.

جدول (۲): مشخصات مجموعه‌داده‌های مورد استفاده

دامنه	کاربرد	اندازه	رکوردهای مثبت	رکوردهای منفی	متوسط طول نظرات
توئیتر	آموزش	۲۰۰	۱۰۰	۱۰۰	۱۰٫٤٥
	تست	۲٥۰	۱۱۲	۱۳۸	۱٥٫٥٥
فیلم	آموزش	٥۰۰	۲٥۰	۲٥۰	٤۹٫۸۰
	تست	۱۰۰۰	٥۰۰	٥۰۰	۳۳٫۹٦
هتل	تست	۲۳٤	۱۲٤	۱۱۰	۱۲٫۲۰

٤- ۲ ارزیابی روش پیشنهادی برای داده‌افزایی

در اولین آزمایش، به ارزیابی کارایی هر یک از مراحل روش پیشنهادی برای داده‌افزایی به صورت جداگانه پرداختیم. برای سنجش اثربخشی روش پیشنهادی، برای هر یک از دو دامنه توئیتر و فیلم، دقت مدل CNN که جزئیات آن در جدول (۱) آمده است را بر روی مجموعه آموزش اولیه و مجموعه‌داده گسترش‌یافته با روش داده‌افزایی پیشنهادی مقایسه کردیم.

جدول (۳) درصد افزایش داده‌ها و جدول (٤) کارایی هر بخش از روش پیشنهادی را بر روی دامنه‌های توئیتر و فیلم نشان می‌دهد. در حالت اول که داده‌افزایی انجام نشده، کارایی CNN در طبقه‌بندی احساسات بر اساس معیار F1 در دامنه توئیتر ٦۰/۷ درصد و در دامنه فیلم ٦۸/٤ درصد است. دلیل اصلی عملکرد ضعیف CNN، افت دقت مدل‌های یادگیری عمیق در شرایط کمبود داده است. مطابق جداول (۳) و (٤)، در حالت دوم، با تصحیح غلط‌های املایی و افزودن جملات اصلاح شده به مجموعه‌داده اولیه، اندازه مجموعه‌های آموزش در دامنه‌های توئیتر و فیلم به ترتیب ۷٤ و ۳٥ درصد و معیار F1 به ترتیب ٥/۱ و ۳ درصد افزایش داشته است.

جدول (۳): درصد افزایش مراحل مختلف روش پیشنهادی در دو دامنه توئیتر و فیلم

حالت	روش	دامنه توئیتر			دامنه فیلم		
		رکوردهای مثبت	رکوردهای منفی	درصد افزایش	رکوردهای مثبت	رکوردهای منفی	درصد افزایش
۱	بدون داده‌افزایی	۱۰۰	۱۰۰	–	۲۵۰	۲۵۰	–
۲	حالت ۱+ تصحیح اشتباهات املایی	۱۶۱	۱۸۷	۷٤	۳٤۵	۳۳۰	۳۵
۳	حالت ۱+ حذف تکرار حروف در کلمه	۱۲۱	۱۲۰	۲۰,۵	۲۵٤	۲۵۲	۱,۲
٤	حالت ۱+ تصحیح اشتباهات املایی+ حذف تکرار حروف در کلمه	۱۸۱	۲۰٦	۹۳,۵	۳٤۹	۳۳۲	۳٦
۵	حالت ٤+ غنی‌سازی با کلمات مترادف	۲٦٦	۲۸٤	۱۷۵	۷۳٦	۷۳۹	۱۹۵
٦	حالت ۵+ غنی‌سازی با کلمات حسی	٤۳۰	۳۷۷	۳۰۳,۵	۷۹٤	۷۸۰	۲۱٤,۸
۷	حالت ۱+ استفاده از ChatGPT با سه پرسش	۳٦٤	۳٤۲	۲۵۳	۹۹۷	۹۸٦	۲۹٦,٦
۸	حالت ٦ + حالت ۷	٦۹٤	٦۲۰	۵۵۷	۱۵٤۱	۱۵۱٦	۵۱۱,٤

جدول (٤): کارایی بخش‌های مختلف روش پیشنهادی برای داده‌افزایی در دو دامنه توئیتر و فیلم

حالت	روش	دامنه توئیتر			دامنه فیلم		
		فراخوانی (٪)	دقت (٪)	F1 (٪)	فراخوانی (٪)	دقت (٪)	F1 (٪)
۱	بدون داده‌افزایی	۵۱,٦	۷۳,۸	٦۰,۷	٦۷,۸	٦۹,۱	٦۸,٤
۲	تصحیح اشتباهات املایی	٦۰,٤	۷۲,۵	٦۵,۸	۷۱,٤	۷۱,۵	۷۱,٤
۳	حذف تکرار حروف در کلمه	۵۵,٦	۷۲,٦	٦۲,۹	٦۸,٦	٦۸,۷	٦۸,٦
٤	تصحیح اشتباهات املایی+ حذف تکرار حروف در کلمه	۵۹,٦	۷۲,۹	٦۵,۵	۷۳	۷۳,٤	۷۲,۲
۵	حالت ٤+ غنی‌سازی با کلمات مترادف	۵٦,۸	۷۳,٤	٦٤	۷۲	۷۳	۷۲,۵
٦	حالت ۵+ غنی سازی با کلمات حسی	٦۲	۷٤,۳	٦۷,۵	۷۳	۷۳,۹	۷۳,٤
۷	حالت ۱+ استفاده از ChatGPT با سه پرسش	۵۸	۷۱	٦۳,۸	۷٦,۳	۷٦,۵	۷٦,٤
۸	حالت ٦ + حالت ۷	٦۰,٤	۷۵,۳	٦۷	۷٦,۳	۷٦,۹	۷٦,٦

در حالت سوم یعنی حذف تکرار حروف در کلمات و افزودن جملات اصلاح شده به مجموعه‌داده اصلی، معیار F1 در مقایسه با حالت بدون داده‌افزایی در دامنه‌های توئیتر و فیلم به ترتیب ۲/۲ و ۰/۲ درصد بهبود داشته است. این نتایج نشان می‌دهد داده‌افزایی بر اساس حذف حروف تکراری به مدل CNN در تشخیص بهتر قطبیت جملات کمک می‌کند اما به اندازه تصحیح غلط‌های املایی مؤثر نیست. دلیل اصلی تأثیر اندک روش حذف تکرار حروف، تعداد جملات کمتر تولید شده در مقایسه با روش تصحیح املای کلمات است. برای مثال، در دامنه فیلم تنها ٦ رکورد مجموعه آموزش حاوی تکرار حروف بوده‌اند.

در حالت چهارم، پس از اعمال هر دو روش تصحیح املا و حذف حروف تکراری بر روی جملات اصلی، داده‌های بیشتری تولید شده است (افزایش ۹۳/۵ و ۳٦ درصد در دامنه‌های توئیتر و فیلم) اما معیار F1 برای مجموعه تست توئیتر در مقایسه با حالت دوم که تنها متکی بر تصحیح املای کلمات است، ۰/۳ درصد کاهش داشته است. در مقابل، حالت چهارم منجر به بهبود کارایی مدل CNN بر روی مجموعه تست فیلم شده است، به گونه‌ای که معیار F1 در مقایسه با حالات دوم و سوم به ترتیب ۰/۸ و ۳/٦ درصد افزایش داشته است.

تحلیل نتایج چهار حالت اول نشان می‌دهد، در متن‌های عامیانه مانند توئیت‌ها که در مقایسه با جملات رسمی‌تر دامنه فیلم دارای غلط‌های املایی و تکرار حروف به مراتب بیشتری هستند، دو روش داده‌افزایی مبتنی بر اصلاح املا و داده‌افزایی مبتنی بر حذف تکرار حروف، عملکرد مدل طبقه‌بندی احساسات را به میزان بیشتری افزایش داده‌اند. اما ترکیب این دو روش گاهی باعث از بین رفتن تاکیدهای رایج در زبان عامیانه می‌شود، بنابراین کارایی اندکی افت می‌کند. در مقابل، در متن‌های رسمی‌تر که غلط‌های املایی کمتری دارند، ترکیب اصلاح املا با حذف حروف تکراری منجر به پاکسازی دقیق‌تر داده‌ها و بهبود عملکرد مدل شده است. در نتیجه، هر بخش روش پیشنهادی زمانی بیشترین اثربخشی را دارد که با نوع خطاها و ویژگی‌های زبانی دامنه همخوان باشد.

در حالت پنجم، مرحله غنی‌سازی با کلمات مترادف بر روی جملات مجموعه داده حاصل از حالت چهارم اعمال شده است. همان‌طور که در جدول (٤) مشاهده می‌شود، در دامنه توئیتر، این مرحله در مقایسه با حالت چهارم عملکرد ضعیف‌تری دارد (۱/۵ درصد کاهش معیار F1)، اما در دامنه فیلم منجر به بهبود اندک (۰/۳ درصد) معیار F1 شده است.

تمرکز حالت ششم بر غنی‌سازی متن با کلمات حسی است که نقش بسزایی در تعیین قطبیت جمله دارند. مطابق جدول (٤)، این روش در مقایسه با حالت چهارم، معیار F1 در مجموعه تست‌های توئیتر و فیلم را به ترتیب ۲ و ۱/۲ درصد افزایش داده است.

در تحلیل انجام‌گرفته بر روی نتایج حالات پنجم و ششم، دو دلیل اصلی برای افت کارایی حالت پنجم در دامنه توئیتر مشخص شد. اولین دلیل وابستگی معنایی برخی از واژه‌ها به زمینه و بافتار متن است. در این مورد، درج مترادف کلمات از واژه‌نامه‌های عمومی بدون در نظر گرفتن زمینه متن و دامنه کاربرد باعث تغییر معنایی جمله اولیه و به تبع آن تغییر شدت یا جهت حس جمله می‌شود. دلیل دوم، وجود کلمات چندنقشی (مانند "دانشمند" که می‌تواند اسم یا صفت باشد) در زبان فارسی است. درج کلمات مترادف بدون در نظر گرفتن نقش نحوی یک کلمه چندنقشی منجر به تغییر ساختار و معنای جمله می‌شود. با این وجود، ترکیب این روش با روش غنی‌سازی با واژه‌های حسی (حالت ششم) منجر به بهبود عملکرد مدل تحلیل احساسات شده است. در واقع، افزودن واژه‌های حسی هم‌جهت با قطبیت جمله، علاوه بر تقویت بار عاطفی و شفاف کردن قطبیت جمله و در نتیجه تمایز بهتر کلاس‌ها، نقش نوعی سیگنال تقویتی برای تشخیص دقیق‌تر داده‌های مبهم حالت پنجم را ایفا می‌کند.

در حالت هفتم، با کمک سه پرسش از ChatGPT که در بخش ۳-۲ معرفی شدند، جملات جدید تولید شدند. مقدار F1 این حالت در مقایسه با حالت بدون داده‌افزایی به ترتیب ۳/۱ و ۸ درصد در دامنه‌های توئیتر و فیلم بهبود داشته است که نشان از عملکرد مؤثر مدل‌های زبانی بزرگ در تولید جملات بامعنا دارد.

در حالت هشتم، جملات ایجاد شده از حالت ششم و هفتم شامل جملات اولیه، جملات حاصل از تصحیح اشتباهات املایی و حذف حروف تکراری، جملات تولید شده توسط روش‌های غنی‌سازی با کلمات مترادف و واژه‌های حسی و جملات ساخته

شده با ChatGPT ادغام شدند. مقدار F1 این حالت در دامنه‌های توئیتر و فیلم به ترتیب به ٦٧ و ٧٦/٦ درصد است که بهترین نتیجه حاصل شده در دامنه فیلم است. برای دامنه توئیتر نیز بهترین کارایی بر اساس معیارهای فراخوانی و F1 در حالت ششم و بیشترین دقت در حالت هشتم حاصل شده است.

به طور خلاصه، حالت ششم روش پیشنهادی که بر رفع چالش‌های متون عامیانه فارسی تمرکز دارد، در دامنه توئیتر بهبود چشمگیری در عملکرد مدل تحلیل احساسات ایجاد کرده است. بر اساس نتایج جدول (٤)، در این دامنه، حتی گسترش مجموعه‌داده حاصل از حالت ششم با روش‌های پیشرفته‌تر مانند استفاده از ChatGPT منجر به ارتقای کارایی مدل تحلیل احساسات نشده است. در مقابل، در دامنه فیلم که متون ساختار رسمی‌تری دارند و خطاهای املایی در آن بسیار کمتر از توئیتر است، حالت ششم در مقایسه با ChatGPT اثربخشی کمتری داشته است. بنابراین می‌توان نتیجه گرفت، حالت ششم روش پیشنهادی برای داده‌افزایی در دامنه‌هایی که متون آن کاملاً عامیانه و سرشار از اشتباهات املایی است، بیشترین اثربخشی را دارد.

در آزمایش دوم، کارایی روش پیشنهادی برای داده‌افزایی با سه روش EDA [٤]، روش پیشنهادی [٥] و ChatGPT که در مقالات متعدد نظیر [١٣] مورد استفاده قرار گرفته، مقایسه شده است. این روش‌ها در بخش ٢ معرفی شده‌اند. نتایج ارزیابی‌ها در جدول (٥) نشان داده شده است.

جدول (٥): مقایسه کارایی روش پیشنهادی با روش‌های موجود

روش	تعداد رکوردهای مثبت	تعداد رکوردهای منفی	درصد افزایش (٪)	دامنه توئیتر			دامنه فیلم		
				فراخوانی (٪)	دقت (٪)	F1 (٪)	فراخوانی (٪)	دقت (٪)	F1 (٪)
EDA با ضریب ٣	٣٩٢	٣٩٩	٢٩٥	٥٩,٢	٧٠,٢	٥٥,٦	٧٢,٥	٧٢,٦	٧٢,٥
EDA با ضریب ٦	٦٦٧	٦٨٧	٥٧٧	٥٨,٤	٦٨,٩	٦٣,٢	٧٢,٤	٧٢,٤	٧٢,٤
روش پیشنهادی [٥]	١٩٩	١٩٩	٩٩	٥٧,٦	٧١,٦	٦٣,٨	٧٢,٨	٧٤,٤	٧٣,٦
ChatGPT با ضریب ٣	٣٥٨	٣٤٤	٢٥١	٦١,٢	٧٠,٩	٦٥,٦	٧٦	٧٦	٧٦
ChatGPT با ضریب ٦	٦٦٥	٦٤١	٦٥٣	٥٧,٦	٧٥,١	٦٥,٢	٧١,٣	٧٤,٨	٧٣
روش پیشنهادی (حالت ٦)	٤٣٠	٣٧٧	٣٠٣,٥	٦٢	٧٤,٣	٦٧,٥	٧٣	٧٣,٩	٧٣,٤
روش پیشنهادی (حالت ٨)	٦٩٤	٦٢٠	٥٥٧	٦٠,٤	٧٥,٣	٦٧	٧٦,٣	٧٦,٩	٧٦,٦

در EDA یک بار برای حصول درصد افزایش تقریباً برابر با حالت ششم روش پیشنهادی، ضریب داده‌افزایی ٣ و بار دیگر برای رسیدن به تعداد جملات تقریباً برابر با حالت هشتم روش پیشنهادی، ضریب داده‌افزایی ٦ در نظر گرفته شده است. به طور مشابه، در روش ChatGPT نیز یک بار به ازای هر جمله در مجموعه‌داده اولیه، ٣ جمله و بار دیگر ٦ جمله با قطبیت یکسان تولید شده است.

همان‌طور که در جدول (٥) مشاهده می‌شود در دامنه توئیتر، EDA با ضریب ٦ نتایج بهتری در مقایسه با EDA با ضریب ٣ داشته است. در مقابل، ChatGPT با ضریب ٣، F1 بالاتری در مقایسه با به‌کارگیری همین مدل با ضریب ٦ داشته است. در این دامنه، معیار F1 حالت ششم روش پیشنهادی در مقایسه با EDA با ضریب ٦، ChatGPT با ضریب ٣ و روش پیشنهادی [٥]، به ترتیب حدود ٤/٣، ١/٩ و ٣/٧ درصد بهبود داشته است.

مطابق جدول (٥)، در دامنه فیلم برای دو روش EDA و ChatGPT بهترین ضریب داده‌افزایی ٣ است. در این دامنه، به‌کارگیری مجموعه‌داده گسترش‌یافته با حالت هشتم روش پیشنهادی در مقایسه با EDA با ضریب ٣، ChatGPT با ضریب

۳ و روش پیشنهادی [۵]، منجر به افزایش F1 مدل طبقه‌بندی احساسات به میزان ۴/۱، ۰/۶ و ۳ درصد شده است.

در آزمایش سوم، برای بررسی معنادار بودن تفاوت نتایج روش پیشنهادی با روش‌های مورد بررسی در جدول (۵) و اطمینان از این که اختلاف مشاهده شده فراتر از تغییرات تصادفی است، از آزمون t زوجی[1] استفاده شده است. برای این منظور، روش پیشنهادی بر روی دامنه توئیتر با حالت پایه بدون داده‌افزایی، EDA با ضریب ۶، روش پیشنهادی [۵]و ChatGPT با ضریب ۶ مقایسه شده است. به این صورت که طبقه‌بند CNN آموزش‌دیده بر روی مجموعه‌داده گسترش‌یافته توسط هر کدام از این روش‌ها، پنج بار بر روی مجموعه تست دامنه توئیتر اجرا شده و مقدار F1 حاصل از هر اجرا ثبت شده است. نتایج آزمون t با سطح اطمینان ۹۵٪ در جدول (۶) ارائه شده است. همان‌طور که مشاهده می‌شود، برای حالت ششم روش پیشنهادی، در همه مقایسه‌ها مقدار p کمتر از ۰/۰۵ است که بیانگر تفاوت آماری معنادار بین نتایج روش پیشنهادی با سایر روش‌ها است. در حالت هشتم نیز، مقدار p تنها در مقایسه با روش ChatGPT بیشتر از ۰/۰۵ شده است که دلیل اصلی آن اشتراک زیاد داده‌های تولید شده توسط این دو روش است. در واقع نیمی از جملات تولید شده توسط ChatGPT در حالت هشتم نیز استفاده شده‌اند.

جدول (۶): نتایج آزمون t زوجی

روش مورد مقایسه	مقدار p برای حالت ششم	مقدار p برای حالت هشتم
بدون داده‌افزایی	۰/۰۰۰۶	۰/۰۰۰۰۲
EDA با ضریب ۶	۰/۰۱۹۱	۰/۰۰۲۳۱
روش پیشنهادی [۵]	۰/۰۰۱۵	۰/۰۰۲۷۵
ChatGPT با ۶ پرسش	۰/۰۰۳۹	۰/۰۶۴۱۹

در آزمایش چهارم، تأثیر روش پیشنهادی بر کارایی طبقه‌بند ParsBERT در دامنه توئیتر ارزیابی شده است. برای این منظور، از نسخه parsbert-base-uncased با معماری BERT-base استفاده شد. این معماری، شامل ۱۲ لایه ترانسفورمر، ۱۲ هد توجه در هر لایه و در مجموع حدود ۱۱۰ میلیون پارامتر است و از بهینه‌ساز AdamW استفاده می‌کند. بر اساس نتایج جدول (۷)، حالات ششم و هشتم روش پیشنهادی، معیار F1 طبقه‌بند ParsBERT را در مقایسه با حالت بدون داده‌افزایی به ترتیب ۱۱/۳۴ و ۵/۹۸ درصد بهبود داده‌اند که بیانگر اثربخشی روش داده‌افزایی پیشنهادی در بهبود عملکرد مدل‌های پیشرفته طبقه‌بندی احساسات است.

جدول (۷): تأثیر روش پیشنهادی بر عملکرد طبقه‌بند ParsBERT

روش مورد مقایسه	فراخوانی (٪)	دقت (٪)	F1 (٪)
بدون داده‌افزایی	۷۰	۷۶/۳۷	۷۳,۰۵
حالت ششم	۸۴,۴	۸۴,۳۸	۸۴,۳۹
حالت هشتم	۷۸,۴	۷۹,۶۸	۷۹,۰۳

۴-۳ ارزیابی روش پیشنهادی برای تحلیل احساسات بین‌دامنه‌ای

در جدول (۸)، کارایی طبقه‌بند CNN آموزش‌دیده بر روی مجموعه‌داده توئیتر، در تحلیل احساسات دامنه‌های هتل و فیلم نشان داده شده است. در حالت بدون داده‌افزایی، F1 طبقه‌بند CNN بر روی دامنه‌های هتل و فیلم به ترتیب برابر ۵۳/۵ و ۵۳/۷ درصد است که بسیار کمتر از دامنه توئیتر (F1 برابر ۶۰,۷ درصد) است. این نتایج نشان می‌دهد مدل آموزش‌دیده بر روی دامنه توئیتر نمی‌تواند به خوبی الگوها و ویژگی‌های دامنه‌های هتل و فیلم را یاد بگیرد.

پس از داده‌افزایی با حالت هشتم روش پیشنهادی، امتیاز F1 مدل تحلیل احساسات در دامنه هتل به ۶۷,۴ درصد رسیده است که اختلاف اندکی با دامنه توئیتر دارد. این نتایج نشان می‌دهد داده‌افزایی دامنه توئیتر کمک شایانی به بهبود دقت تحلیل احساسات در دامنه هتل کرده است. در مقابل، داده‌افزایی دامنه توئیتر، امتیاز F1 در دامنه فیلم را تنها ۰/۸ درصد بهبود داده است. در بخش ۳-۲، علاوه بر مراحل ۱ تا ۵ پیشنهاد شده برای داده‌افزایی، دو مرحله دیگر (مراحل ۶ و ۷) نیز به منظور گسترش

[1] Paired t-test

داده‌ها در کاربرد تحلیل احساسات بین‌دامنه‌ای پیشنهاد شد که شامل داده‌افزایی مبتنی بر غنی‌سازی با واژه‌های دامنه مقصد و داده‌افزایی با کمک ChatGPT می‌شوند. در جدول (۸)، اثربخشی هر یک از این دو مرحله در بهبود دقت تحلیل احساسات بین‌دامنه‌ای گزارش شده است.

مطابق جدول (۸)، دو روش نوآورانه غنی‌سازی با واژه‌های دامنه مقصد و ChatGPT، معیار F1 مدل تحلیل احساسات بدون داده‌افزایی در دامنه هتل را به ترتیب ۹/۸ و ۱٤/٦ درصد افزایش داده است که جهش چشمگیری در عملکرد این مدل محسوب می‌شود. در دامنه فیلم نیز، این دو رویکرد، در مقایسه با حالت بدون داده‌افزایی، امتیاز F1 مدل طبقه‌بندی احساسات را به میزان ۲/٦ و ۰/٤ درصد بهبود داده‌اند. دلیل اصلی بهبود عملکرد مدل تحلیل احساسات بین‌دامنه‌ای این است که هر دو رویکرد پیشنهادی با افزودن واژه‌های دامنه مقصد به جملات دامنه مبدأ، به مدل امکان آشنایی با فضای واژگانی دامنه مقصد و افزایش توانایی تعمیم آموخته‌ها به دامنه مقصد را می‌دهد.

در دو سطر انتهایی جدول (۸)، نتایج ادغام جملات مصنوعی تولید شده توسط حالت هشتم روش پیشنهادی با خروجی دو روش غنی‌سازی با واژه‌های دامنه مقصد و ChatGPT ارائه شده است. بر اساس نتایج حاصل، در دامنه هتل، ترکیب روش غنی‌سازی با حالت هشتم روش پیشنهادی منجر به بهبود نتایج نشده است. در مقابل، ادغام حالت هشتم روش پیشنهادی با ChatGPT، معیار F1 هر یک از آن‌ها را به ترتیب ۰/۹ و ۰/۲ درصد ارتقا داده است. در دامنه فیلم، بهترین نتیجه توسط روش غنی‌سازی با واژه‌های دامنه مقصد حاصل شده است. در این دامنه، ترکیب دو رویکرد پیشنهادی با حالت هشتم بهبود چشمگیری در نتایج ایجاد نکرده است.

جدول (۸): کارایی روش پیشنهادی برای تحلیل احساسات بین‌دامنه‌ای

روش	دامنه هتل				دامنه فیلم			
	درصد افزایش	فراخوانی	دقت	F1	درصد افزایش	فراخوانی	دقت	F1
بدون داده‌افزایی	–	۵۳٫۸	۵۳٫٤	۵۳٫۵	–	۵۱٫۱	۵٦٫٦	۵۳٫۷
حالت ششم	٦۰٫۲	٦۰٫۲	٦٤	٦۲	٦۰٫۲	۵۲٫۷	۵٦٫٤	۵٤٫۵
حالت هشتم	۵۵۷	٦٤٫۹	۷۰٫۲	٦۷٫٤	۵۵۷	۵۲٫۳	۵۷	۵٤٫۵
غنی‌سازی با واژه‌های دامنه مقصد	۳۰۳٫۵	٦۲٫۸	٦۳٫۹	٦۳٫۳	۳۰۳٫۵	۵۵٫٤	۵۷٫۳	۵٦٫۳
داده‌افزایی با ChatGPT	۱۰۰	٦۷٫۵	٦۸٫۹	٦۸٫۱	۹۹	۵۲٫۳	۵٦٫۱	۵٤٫۱
حالت ۸ + غنی‌سازی با واژه‌های دامنه مقصد	۸٦۱	٦٤٫۱	٦٦٫۵	٦۵٫۲	۹٦۰٫۵	۵۲٫٤	۵۵٫٦	۵٤
حالت ۸ + داده‌افزایی با ChatGPT	٦۵۷	٦٤٫۱	۷۳٫۲	٦۸٫۳	٦۵٦٫۵	۵۳٫۲	۵۸	۵۵٫۵

نتایج جدول (۸) نشان می‌دهد که در دامنه هتل، بر خلاف دامنه فیلم، رویکرد پیشنهادی داده‌افزایی با ChatGPT به نتایج بهتری در مقایسه با رویکرد غنی‌سازی با واژه‌های دامنه مقصد دست یافته است. دلیل اصلی این تفاوت، ابهام کمتر واژه‌های حسی مانند "تمیز" و "آرام" در دامنه هتل و ساختار ساده‌تر و رسمی‌تر جملات آن است. به همین دلیل، بازنویسی جملات توسط ChatGPT، جملات طبیعی‌تر و نزدیک به دامنه مقصد ایجاد کرده است. در مقابل، در دامنه فیلم، جملات طولانی، سبک بیان غیررسمی، استفاده بیشتر از استعاره و کنایه و واژه‌های حسی مبهم مانند "بازی" و "قصه" باعث می‌شود بازنویسی کل جمله باعث از دست رفتن بخشی از ظرافت‌های زبانی شود. در حالی که ایجاد تغییرات کمتر در جملات با افزودن واژه‌های دامنه مقصد علاوه بر حفظ ساختار جمله، باعث آشنایی مدل طبقه‌بندی با واژه‌های دامنه مقصد و به تبع آن بهبود کارایی می‌شود.

همچنین بر اساس نتایج جدول (۸)، اگر چه دو رویکرد پیشنهادی برای تحلیل احساسات بین دامنه‌ای هر کدام به تنهایی به نتایج

موفقیت‌آمیز دست یافته‌اند، اما ترکیب آن‌ها با حالت هشتم روش پیشنهادی تفاوت معناداری در نتایج ایجاد نکرده و گاهی منجر به افت کارایی نیز شده است. دلیل این امر، بیش‌برازش مدل و عدم تعمیم الگوهای آموخته به داده‌های واقعی است. بنابراین، در دامنه‌های مختلف، انتخاب یکی از سه رویکرد پیشنهادی (حالت هشتم، غنی‌سازی با واژه‌های دامنه مقصد و داده‌افزایی با ChatGPT) متناسب با ماهیت کاربرد و ویژگی‌های زبانی و محتوایی داده‌ها، می‌تواند بیشترین اثربخشی را داشته باشد.

۵- نتیجه‌گیری

در این مقاله، با هدف بهبود عملکرد مدل‌های یادگیری عمیق در تحلیل احساسات متون عامیانه فارسی، روشی ترکیبی برای داده‌افزایی ارائه گردید. این روش که مبتنی بر چالش‌های مطرح در پردازش متون عامیانه فارسی است طی چند مرحله شامل تصحیح غلط‌های املایی، حذف تکرار حروف در کلمات، غنی‌سازی با کلمات مترادف، غنی‌سازی با واژه‌های حسی و تولید جملات هدفمند با استفاده از ChatGPT به گسترش مجموعه داده می‌پردازد. همچنین روشی برای بهبود عملکرد مدل‌های تحلیل احساسات بین‌دامنه‌ای شامل دو مرحله غنی‌سازی جملات دامنه مبدأ با واژه‌های دامنه مقصد و بازنویسی جملات با واژه‌های دامنه مقصد با کمک ChatGPT پیشنهاد شد.

یافته‌های این پژوهش حاکی از آن است که روش پیشنهادی برای داده‌افزایی تأثیر چشمگیری در بهبود کارایی طبقه‌بند CNN در تحلیل احساسات دامنه‌های توئیتر و فیلم دارد. همچنین در مقایسه با سه روش برجسته موجود، روش پیشنهادی دقت بالاتری را به دست آورده است. افزون بر این، روش پیشنهادی برای تحلیل احساسات بین‌دامنه‌ای قادر است دقت طبقه‌بند CNN آموزش دیده بر روی دامنه توئیتر را برای تحلیل احساسات نظرات کاربران دامنه‌های هتل و فیلم به میزان قابل توجهی بهبود بخشد.

به عنوان کار آتی برآنیم تا با تحلیل جزئی‌تر ساختارهای دستوری، واژگانی و معنایی زبان فارسی مانند افعال پیشوندی، معکوس‌کننده‌های قطبیت و ابهام‌های معنایی واژه‌ها، کارایی روش داده‌افزایی پیشنهادی را بهبود بخشیم. همچنین پژوهش بر روی روش‌های پیشرفته مهندسی پرسش از ChatGPT برای استخراج داده‌های مصنوعی با کیفیت بالاتر و مرتبط‌تر با دامنه مدنظر و نیز اعمال فیلترهایی برای حذف داده‌های نویز و افزایش اعتبار مجموعه داده گسترش‌یافته برای ادامه کار مدنظر است.

References

[1] L. Zhang, and B. Liu. “Sentiment analysis and opinion mining,” Encyclopedia of Machine Learning and Data Science, Springer, New York, pp. 1-13, 2023, doi: 10.1007/978-1-4899-7502-7_907-2.

[2] M. M. Agüero-Torales, J. I. A. Salas, and A. G. López-Herrera, “Deep learning and multilingual sentiment analysis on social media data: An overview,” Applied Soft Computing, vol. 107, p. 107373, 2021, doi: 10.1016/j.asoc.2021.107373.

[3] N. A. Sharma, A. S. Ali, and M. A. Kabir, “A review of sentiment analysis: tasks, applications, and deep learning techniques,” International Journal of Data Science and Analytics, pp. 1–38, 2024, doi: 10.1007/s41060-024-00594-x.

[4] T. Body, X. Tao, Y. Li, L. Li, and N. Zhong, “Using back-and-forth translation to create artificial augmented textual data for sentiment analysis models,” Expert Systems with Applications, vol. 178, p. 115033, 2021, doi: 10.1016/j.eswa.2021.115033.

[5] M. Mir, and S. Noferesti, “Using data augmentation techniques for sentiment analysis of users’ opinions on reopening of schools during the COVID-19 epidemic,” Signal and Data Processing, vol. 21, no. 2, pp. 3–14, 2024, doi: 10.61186/jsdp.21.2.3 [In Persian].

[6] J. Wei, and K. Zou, “EDA: Easy data augmentation techniques for boosting performance on text classification tasks,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), Hong Kong, China, 2019, pp. 6382–6388, doi: 10.18653/v1/D19-1670.

[7] H. Youneszadeh Haghighi, and S. Noferesti, “Text augmentation based on operation weighting using genetic algorithm,” Scientia Iranica, 2025, doi: 10.24200/sci.2025.65358.9440.

[8] G. Li, H. Wang, Y. Ding, K. Zhou, and X. Yan, “Data augmentation for aspect-based sentiment analysis,” International Journal of Machine Learning and Cybernetics, vol. 14, no. 1, pp. 125–133, 2023.

[9] X. Wu, “Conditional BERT contextual augmentation,” in Computational Science – ICCS 2019, Lecture Notes in Computer Science, Y. Shi et al., Eds. Cham, Switzerland: Springer, 2019, pp. 98–113, doi: 10.1007/978-3-030-22747-0_7.

[10] R. Nair, R. P. Singh, D. Gupta, and P. Kumar, "Evaluating the impact of text data augmentation on text classification tasks using DistilBERT," Procedia Computer Science, vol. 235, pp. 102–111, 2024, doi: 10.1016/j.procs.2024.04.013.
[11] H. Dai et al., "Auggpt: Leveraging ChatGPT for text data augmentation," IEEE Transactions on Big Data, 2025, doi: 10.1109/TBDATA.2025.3536934.
[12] M. Mohammadi, M. R. Amin, and S. Tavakoli, "Boosting Sentiment Analysis in Persian through a GAN-Based Synthetic Data Augmentation Method," in Proc. of the 1st Workshop on NLP for Languages Using Arabic Script, 2025, pp. 54-63.
[13] L. Xu, H. Xie, S. J. Qin, F. L. Wang, and X. Tao, "Exploring ChatGPT-based augmentation strategies for contrastive aspect-based sentiment analysis," IEEE Intelligent Systems, vol. 40, no. 1, pp. 69–76, 2025, doi: 10.1109/MIS.2024.3508432.
[14] G. A. Miller, "WordNet: A lexical database for English," Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.
[15] M. Xu, Q. Zhong, and J. Liu, "LLM-as-an-Augmentor: Improving the data augmentation for aspect-based sentiment analysis with large language models," in Poster Volume II, Computational Intelligence and Intelligent Computing (ICIC 2024), Cham, Switzerland: Springer, 2024.
[16] S. Kwon and Y. Lee, "Explainability-based mix-up approach for text data augmentation," ACM Transactions on Knowledge Discovery from Data, vol. 17, no. 1, pp. 1–14, 2023, doi: 10.1145/3533048.
[17] Taheri, A. Zamanifar, and A. Farhadi, "Enhancing aspect-based sentiment analysis using data augmentation based on back-translation," International Journal of Data Science and Analytics, vol. 19, pp. 1-26, 2024. doi:10.1007/s41060-024-00622-w.
[18] "Dataheart". [Online]. Available: (accessed Dec. 2024).
[19] "Dataheart". [Online]. Available: http://dataheart.ir/article/3362/ (accessed Aug. 2025).
[20] R. Dehkharghani. "Labeled-Sentences." [Online].Available: http://myweb.sabanciuniv.edu/rdehkharghani/files/2018/08/Labeled-Sentences.xlsx (accessed April. 2025).

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 3, No.1, Spring & Summer 2025, pp. 111 - 126

A Text Augmentation Approach for Sentiment Analysis with a Focus on Persian Text Processing Challenges

Ghazal Mehrparvar [1], Samira Noferesti [2*]

[1]MS Student, Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Sistan and Baluchestan, Zahedan, Iran

[2]Associate Professor, Department of Information Technology Engineering, Faculty of Electrical and Computer Engineering, University of Sistan and Baluchestan, Zahedan, Iran

Article Information

Original Research Paper

Received:
2025 July 9

Accepted:
2025 September 6

Keywords:
Augmentation, Sentiment analysis, Persian language challenges, Cross-domain sentiment analysis, Deep learning

Corresponding Author*:
snoferesti@ece.usb.ac.ir

Abstract

With the increasing volume of user comments on social networks and service platforms, sentiment analysis has become a key tool for extracting insights from textual data. However, challenges such as the scarcity of labeled data and linguistic complexities, particularly in the Persian language, have imposed limitations on the performance of deep learning models. In this paper, we propose a data augmentation method aimed at improving sentiment analysis of Persian texts. This approach focuses on generating synthetic data by addressing challenges inherent to social media reviews, including their brevity, colloquial language, and frequent spelling and grammatical errors. Additionally, a data augmentation technique for enhancing cross-domain sentiment analysis is introduced. Evaluation results demonstrate that the proposed data augmentation method improves the F1 score of a convolutional neural network classifier for sentiment analysis by 6.8% on the Twitter domain. Furthermore, in cross-domain sentiment analysis, the proposed method increases the F1 score of the CNN model trained on Twitter data by 14.8% when tested on a hotel review dataset.

: 10.22034/ABMIR.2025.23361.1141

E-ISSN: 2821-2037