Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال سوم، شماره اول، بهار و تابستان ۱۴۰۴، صفحات ۶۷ - ۵۱

تجزیه و تحلیل احساسات چندوجهی در زبان فارسی با استفاده از مدل‌های مبتنی بر ترانسفورمر و کپسول‌های تشخیص احساس

پریسا رحمانیان[۱]، سید محمدرضا موسوی[۲*]، محمدهادی صدرالدینی[۳]

[۱]کارشناس ارشد نرم‌افزار، بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات، دانشکده مهندسی برق و کامپیوتر، دانشگاه شیراز، شیراز، ایران

[۲]دانشیار بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات، دانشکده مهندسی برق و کامپیوتر، دانشگاه شیراز، شیراز، ایران

[۳]استاد بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات، دانشکده مهندسی برق و کامپیوتر، دانشگاه شیراز، شیراز، ایران

چکیده

تحلیل احساسات، فرآیند شناسایی یا طبقه‌بندی عقاید و احساسات افراد در مورد یک موضوع است. کاربردهای تحلیل احساسات در زمینه‌هایی همچون بازاریابی و تحلیل شبکه‌های اجتماعی به‌طور گسترده مورد توجه قرار گرفته است. اگرچه تحقیقات اولیه در زمینه تحلیل احساسات عمدتاً تنها بر داده‌های متنی متکی بودند، اما مطالعات اخیر نشان داده‌اند که سیستم‌های چندوجهی که انواع مختلفی از داده‌ها را در برمی‌گیرند، می‌توانند عملکرد بهتری داشته باشند. در این پژوهش، به تحلیل احساسات چندوجهی در زبان فارسی پرداخته شده است و برای اولین بار در این زمینه، روشی براساس مدل‌های مبتنی بر ترانسفورمر پیشنهاد شده است. برای استخراج ویژگی‌های متنی، از مدل پارس‌برت و یک مدل مبتنی بر ترانسفورمرهای بینایی به نام مدل DINOv2 برای استخراج ویژگی‌های بصری به کار گرفته شده است. تشخیص احساسات در هر وجه از طریق کپسول‌های تشخیص احساسات صورت پذیرفته و درنهایت به‌منظور پیش‌بینی احساسات در حالت چندوجهی، تکنیک همجوشی دیرهنگام در لایه نهایی اعمال گردیده است. افزون بر این، از یک تکنیک هوش مصنوعی تفسیرپذیر مستقل از مدل، به نام لایم، برای به دست آوردن بینشی در مورد پیش‌بینی‌های انجام شده توسط شاخه‌های تک‌وجهی کمک گرفته شده است. آزمایش‌ها انجام شده نشان داد که مدل چندوجهی ارائه شده توسط این پژوهش بر روی مجموعه داده عکس نظر به‌دقت ۹۶/۵ درصد و امتیاز F1 معادل ۹۶/۴۸ درصد دست یافته است.

مقاله پژوهشی

تاریخ دریافت:
۱۴۰۴/۴/۱۵
تاریخ پذیرش:
۱۴۰۴/۵/۲۳

کلیدواژه‌ها:
تحلیل احساسات چندوجهی، تحلیل احساسات فارسی، ترانسفورمر، مدل کپسول، هوش مصنوعی تفسیرپذیر

نویسنده‌مسئول:
smmosavi@shirazu.ac.ir

: 10.22034/ABMIR.2025.23347.1137

E-ISSN: 2821-2037

۱- مقدمه

احساسات انسانی نقش مؤثری در شکل دادن به تصمیم‌ها و رفتارها ایفا می‌کنند. این احساسات از طریق روش‌های کلامی و غیرکلامی بیان می‌شوند و منعکس‌کننده شیوه‌های متفاوتی هستند که افراد، مسائل گوناگون را تجربه و تفسیر می‌کنند. تجزیه‌وتحلیل احساسات فرآیند شناسایی و یا طبقه‌بندی احساسات و نظرات کاربران درباره خدمات، محصولات، وقایع یا هر موضوع دیگر است [۱]. تحلیل احساسات به کسب‌وکارها، نهادها و سازمان‌ها کمک می‌کند تا عقاید مشتریان و یا افراد جامعه را درک کنند، روندهای نوظهور را رصد کنند و فرآیندهای تصمیم‌گیری و کیفیت خدمات خود را بهبود ببخشند. به همین دلیل است که در طول دو دهه گذشته، پژوهشگران حوزه هوش مصنوعی در تلاش بوده‌اند تا چگونگی تشخیص احساسات انسانی را به‌طور مؤثر به ماشین‌ها آموزش دهند.

در ابتدا، پژوهش‌های تحلیل احساسات فقط بر داده‌های متنی متمرکز بودند. بااین‌حال، متن اغلب به تنهایی نمی‌تواند معنای کامل آن‌چه که فرد بیان می‌کند را منتقل کند. تحلیل احساسات چندوجهی، فرایند شناسایی احساسات یا عقاید افراد از طریق چندین نوع داده مانند متن، تصویر، صوت و ویدیو به‌طور هم‌زمان است [۲]. مطالعات نشان داده‌اند که سیستم‌هایی که از انواع متفاوت داده‌ها (چندوجهی) برای تحلیل احساسات استفاده می‌کنند، نسبت به سیستم‌هایی که تنها با یک نوع داده (تک‌وجهی) کار می‌کنند، عملکرد بهتری دارند [۳]. تحلیل احساسات چندوجهی نیاز به مدل‌سازی مناسب ارتباط میان وجه‌های مختلف دارد. این موضوع باعث می‌شود که تحلیل احساسات چندوجهی نسبت به تحلیل احساسات متنی سنتی چالش‌برانگیزتر باشد [٤]. تعداد قابل‌توجهی از پژوهش‌های تحلیل احساسات چندوجهی برای زبان انگلیسی انجام شده است اما در زبان فارسی این نوع پژوهش‌ها محدود به [۵-۷] است.

با توجه به رشد روزافزون تولید محتوای چندرسانه‌ای به زبان فارسی در شبکه‌های اجتماعی، تحلیل احساسات چندوجهی در این بستر از اهمیت ویژه‌ای برخوردار شده است. تشخیص دقیق احساسات موجود در محتوای منتشرشده، نقشی کلیدی در درک صحیح بازخورد کاربران و پایش افکار عمومی در شبکه‌های اجتماعی ایفا می‌کند. چالش اصلی مطرح‌شده در این مقاله، بهبود دقت در تحلیل احساسات چندوجهی زبان فارسی است. با وجود دستاوردهای قابل‌توجه در پژوهش‌های پیشین، تمرکز حوزه تحلیل احساسات همچنان بر ارتقای مداوم دقت و کارایی مدل‌ها است. پژوهشگران تلاش می‌کنند با بهبود عملکرد سیستم‌های تحلیل احساسات، مسیر توسعه سامانه‌های دقیق و هوشمند تشخیص احساس را هموار سازند، این موضوع یکی از انگیزه‌های اصلی نگارش مقالات متعدد در حوزه تحلیل احساسات به شمار می‌رود. در همین راستا، مقاله حاضر با هدف بهبود دقت، یک مدل ترکیبی مبتنی بر یادگیری عمیق برای تحلیل احساسات چندوجهی فارسی ارائه می‌دهد.

در مطالعات پیشین تحلیل احساسات چندوجهی در زبان فارسی [۵-۷]، عمدتاً از استخراج گرهای ویژگی سنتی مانند شبکه‌های عصبی پیچشی[1] و شبکه‌های بازگشتی بهره گرفته شده است. همچنین، لایه طبقه‌بندی در این مدل‌ها اغلب بر پایه ساختار ساده شبکه کاملاً متصل[2] طراحی شده است. در پژوهش حاضر، به‌جای استفاده از معماری‌های سنتی نظیر شبکه‌های پیچشی و شبکه‌های بازگشتی، از مدل‌های پیش‌آموزش‌دیده پارس‌برت[3] و DINOv2 که مبتنی بر معماری ترنسفورمر هستند، استفاده شده است. این مدل‌ها به‌دلیل بهره‌مندی از دانش پیش‌زمینه (حاصل از مرحله پیش‌آموزش[4])، حتی با داده محدود نیز قادر به استخراج ویژگی‌های عمیق و زمینه‌محور هستند. ساختار ترنسفورمر و سازوکار خود-توجه چند سر[5]، امکان مدل‌سازی وابستگی‌های محلی و سراسری را به‌طور هم‌زمان فراهم می‌کنند [۸]؛ قابلیتی که در روش‌های سنتی با محدودیت همراه است. مدل پارس‌برت معنای واژه‌ها را با توجه به اطلاعات زمینه‌ای درک می‌کند و مدل DINOv2 ویژگی‌هایی تعمیم‌پذیر و چندمنظوره از تصاویر استخراج می‌کند. ترکیب این

[1] Convolutional Neural Network
[2] Fully Connected
[3] ParsBERT
[4] Pretraining
[5] Multi Head Self-Attention

دو مدل، چارچوبی دقیق و پیشرفته برای تحلیل احساسات چندوجهی فراهم کرده است که نسبت به روش‌های کلاسیک از عملکرد برتری برخوردار است. افزون بر این، در لایه طبقه‌بندی نیز به‌جای ساختار سنتی شبکه کاملاً متصل، از مدل کپسول بهره گرفته شده است. ساختار اصلی مدل کپسول مبتنی بر کپسول‌هایی است که خروجی آن‌ها بردارهای چندبُعدی هستند؛ این بردارها نسبت به خروجی‌های اسکالر نورون‌ها در شبکه‌های عصبی سنتی، توانایی بیشتری در نگهداری ویژگی‌های متنوع دارند [۹]. در این پژوهش، از مدل کپسولی معرفی‌شده توسط وانگ و همکاران [۱۰] استفاده شده است. اگرچه مدل کپسول طراحی‌شده توسط وانگ و همکاران در ابتدا برای تحلیل احساسات تک‌وجهی متنی توسعه یافته بود، پژوهش حاضر نخستین مطالعه‌ای است که آن را در چارچوب تحلیل احساسات چندوجهی فارسی به‌کار گرفته است. پژوهش‌های دیگری نیز از جمله [۱۱] و [۱۲] در معماری پیشنهادی خود از مدل کپسول وانگ استفاده کرده‌اند. چِنگ و همکارانش به‌منظور تحلیل احساسات داده‌های متنی این مدل کپسولی را در ترکیب با سازوکار خود-توجه چند سر، شبکه‌های عصبی پیچشی و GRU[1] دوسویه به کار بردند [۱۱]. سان و همکارانش برای نخستین بار مدل کپسول را برای وظیفه تشخیص موضع استفاده کردند [۱۲]. علاوه بر این‌ها، به‌منظور انجام وظیفه تحلیل احساسات چند دامنه‌ای برای داده‌های متنی، نسخه‌ای از مدل کپسول در ترکیب با معماری برت مورداستفاده قرار گرفته است [۱۳]. بدین ترتیب که مدل کپسول وانگ به‌گونه‌ای سفارشی‌سازی شد تا بتواند علاوه بر احساس متن، دامنه را نیز پیش‌بینی کند. نکته شایان ذکر این است که برخلافِ [۱۳]، این پژوهش مدل استاندارد کپسول وانگ را با رویکرد همجوشی دیرهنگام[2] ترکیب می‌کند و آن را برای تحلیل احساسات چندوجهی تطبیق می‌دهد. همچنین برای تحلیل احساسات تصاویر، مدل کپسول وانگ را با ترانسفورمرهای بینایی (مدل DINOv2) ترکیب کردیم که این موضوع پیش‌تر مورد مطالعه قرار نگرفته است. بنابراین اقدامات اصلی این پژوهش شامل موارد زیر است:

- استفاده از مدل مبتنی بر ترانسفورمر پارس‌برت برای استخراج ویژگی از داده‌های متنی فارسی که امکان تولید نمایش‌های متنی غنی‌تری را در مقایسه با روش‌های سنتی فراهم می‌کند.
- به‌کارگیری مدل مبتنی بر ترانسفورمر DINOv2 برای استخراج ویژگی‌های تصویری. مدل‌های مبتنی بر ترانسفورمرهای بینایی از جمله دینو می‌توانند جایگزینی قدرتمند برای شبکه‌های عصبی پیچشی معمولی باشند [۱۴].
- تطبیق‌دهی و به‌کارگیری مدل کپسول ارائه شده توسط وانگ و همکاران [۱۰] برای وظیفه تحلیل احساسات چندوجهی فارسی، که در پژوهش‌های پیشین به بررسی تأثیر این مدل پرداخته نشده است.
- استفاده از یک تکنیک هوش مصنوعی تفسیرپذیر به‌منظور تفسیر پیش‌بینی‌های انجام شده توسط شاخه‌های تک‌وجهی مدل.

ادامه مقاله به این صورت سازمان‌دهی شده است: در بخش دوم، پیشینه تحقیق در زمینه زبان‌های کم‌منبع مانند فارسی مرور می‌شود. بخش سوم به معرفی معماری مدل پیشنهادی اختصاص دارد. در بخش چهارم آزمایش‌ها و یافته‌ها مورد بررسی قرار می‌گیرد و درنهایت، بخش پنجم به نتیجه‌گیری می‌پردازد.

۲- پیشینه پژوهش

در این بخش، پژوهش‌های حائز اهمیتی که به وظیفه تحلیل احساسات چندوجهی پرداخته‌اند، با تأکید بر تحقیقات مربوط به زبان‌های کم‌منبع، بررسی شده است.

نخستین مجموعه داده تحلیل احساسات چندوجهی در زبان فارسی توسط دشتی‌پور و همکارانش معرفی شده است [۵]. این مجموعه داده شامل ویدیوهای فارسی جمع‌آوری شده از وب‌سایت یوتیوب است. مطالعه آن‌ها بر سه نوعِ داده، شامل داده‌های متنی، ویدیویی و صوتی متمرکز بود. برای استخراج ویژگی از این داده‌ها، به ترتیب از شبکه عصبی LSTM[3] دوسویه، شبکه عصبی پیچشی سه‌بعدی و نرم‌افزار OpenSMILE استفاده شد. افزون بر این، دو تکنیک همجوشی دیرهنگام و همجوشی زودهنگام[4] برای ادغام اطلاعات

[1] Gated Recurrent Unit
[2] Late Fusion
[3] Long Short-Term Memory
[4] Early Fusion

بین‌وجهی به کار برده شد. آن‌ها چندین آزمایش را به‌منظور بررسی اثربخشی هر نوعِ داده و اثربخشی نوع همجوشی بر عملکرد مدل انجام دادند. این آزمایش‌ها نشان داد که ترکیبی از هر سه نوع داده متن، ویدیو و صوت به همراه روش همجوشی زودهنگام می‌تواند به بالاترین دقت دست یابد.

علاوه بر مجموعه داده معرفی شده توسط دشتی‌پور و همکاران [۵]، مجموعه داده فارسی دیگری به نام MPerSocial توسط ناصری کریموند و همکاران جمع‌آوری شده است [۶]. این مجموعه داده شامل جفت‌های متن-تصویر استخراج‌شده از پست‌های شبکه‌های اجتماعی اینستاگرام و تلگرام است. آن‌ها یک مدل یادگیری عمیق چندوجهی را پیشنهاد دادند که شامل دو شاخه بود، یک شاخه برای استخراج ویژگی‌های معنایی از داده‌های متنی که توسط یک شبکه LSTM دوسویه انجام می‌شد و شاخه دیگری برای استخراج ویژگی‌های بصری از داده‌های تصویری که با استفاده از مدل VGG-16 پیاده‌سازی شده بود. ویژگی‌های به‌دست‌آمده به هم متصل شدند و سپس برای طبقه‌بندی احساسات به یک لایه کاملاً متصل ارسال شدند. آن‌ها مدل پیشنهادی خود را با طبقه‌بندهای سنتی از جمله ماشین بردار پشتیبان، بیز ساده و مدل‌های یادگیری عمیق مشابه مانند 2CNN-GRU مقایسه کردند و برتری روش ارائه شده را نشان دادند.

زارعی‌نژاد و بهشتی‌فرد مجموعه داده‌ای به نام مجموعه داده عکس نظر را معرفی کردند [۷]. این مجموعه داده بزرگ‌ترین مجموعه داده تحلیل احساسات چندوجهی در زبان فارسی به شمار می‌رود که شامل ده هزار جفت متن-تصویر است. آن‌ها به‌منظور استخراج ویژگی از تصاویر، مدل از پیش‌آموزش‌دیده ResNet-50 را به کار بردند و برای استخراج ویژگی‌های متنی از شبکه‌های عصبی پیچشی، LSTM دوسویه و GRU دوسویه استفاده کردند. سپس یک تکنیک همجوشی میانی جهت ادغام ویژگی‌های متنی و بصری اعمال شد و درنهایت طبقه‌بندی احساسات از طریق لایه کاملاً متصل صورت گرفت. نتایج آزمایش‌های انجام شده حاکی از آن بود که رویکرد چندوجهی مبتنی بر LSTM دوسویه و ResNet-50 با دقت ۹۳/۳۹ توانست عملکرد بهتری نسبت به مدل‌های تک‌وجهی از خود نشان دهد.

تِیلور و فوزی یک مجموعه داده تحلیل احساسات چندوجهی جدید برای زبان کم منبع مالایی در حوزه اجتماعی-سیاسی معرفی کردند [۱۵]. این مجموعه داده شامل سه نوع داده متن، صوت و ویدیو بود. برای استخراج ویژگی از داده‌های متنی و ویدیویی، از مدل‌های مبتنی بر شبکه عصبی پیچشی استفاده شد، درحالی‌که ویژگی‌های صوتی با استفاده از کتابخانه LibROSA به دست آمد. همه ویژگی‌های استخراج‌شده به هم متصل شدند و سپس طبقه‌بندی انجام شد. نتایج به دست آمده از آزمایش‌های آنان نشان داد که استفاده از داده‌های چندوجهی می‌تواند عملکرد مدل را در مقایسه با استفاده از داده‌های تک‌وجهی بهبود ببخشد.

چوداری و همکاران یک روش آگاه از هیجان[1] برای تحلیل احساسات چندوجهی در زبان مَراتی پیشنهاد کردند [۱۶]. مراتی یک زبان کم‌منبع در حوزه تحقیقات پردازش زبان طبیعی است. آن‌ها از روش ترجمه معکوس برای افزایش اندازه مجموعه داده آموزشی بهره بردند. به‌منظور استخراج ویژگی‌های متن و تصویر، به ترتیب مدل Marathi RoBERTa و یک مدل VGG-16 از پیش آموزش‌دیده، پیاده‌سازی شدند. علاوه بر این، برای داده‌های متنی، یک فاز استخراج هیجان را در نظر گرفتند. همچنین برای غنی‌سازی ویژگی‌های متنی، نوعی سازوکار توجه[2] را در سطح کلمه اعمال کردند. یکی از مشاهدات آنان، این بود که افزودن فاز استخراج هیجان به مدل می‌تواند عملکرد آن را بهبود ببخشد.

داس و سینگ مجموعه داده‌ای را برای تحلیل احساسات چندوجهیِ اخبار در زبان کم‌منبع آسامی جمع‌آوری کردند [۱۷]. چارچوب تشخیص احساس پیشنهادی آنان از سه واحد شامل واحد تشخیص احساس متنی، واحد تشخیص احساس بصری و واحد تشخیص احساس متن-تصویر تشکیل شده بود. واحد تشخیص احساس متنی براساس یک تکنیک استخراج ویژگی مبتنی بر لغت‌نامه در کنار الگوریتم‌های یادگیری ماشین سنتی، طراحی شده بود و واحد تشخیص احساس بصری شامل مدلی برای تولید زیرنویس متنی برای تصاویر اخبار بود. همچنین در واحد تشخیص

[1] Emotion-Aware

[2] Attention Mechanism

احساس متن-تصویر، از متون زیرنویس تصاویر و متن اصلی اخبار در کنار هم برای استخراج ویژگی استفاده کردند. درنهایت، با استفاده از روش همجوشی دیرهنگام، احساس نهایی براساس همه وجه‌ها به دست آمد. نتایج نشان داد که ادغام زمینه‌ای ویژگی‌های چندوجهی منجر به عملکرد بهتری نسبت به ویژگی‌های تک‌وجهی شده است.

کوتاه سخن اینکه بررسی ادبیات تحقیق نشان می‌دهد عملکرد مدل‌های چندوجهی به‌طورکلی بهتر از مدل‌های تک‌وجهی بوده است که این موضوع اثربخشی و توجیه‌پذیری استفاده از داده‌های چندوجهی را در تحلیل احساسات تأیید می‌کند.

۳- روش تحقیق

مدل پیشنهادی این تحقیق که کلیت آن در شکل ۱ آمده است، از چهار واحد اصلی تشکیل شده است: واحد استخراج ویژگی‌های متنی با استفاده از مدل پارس‌برت، واحد استخراج ویژگی‌های تصویری از طریق مدل DINOv2، واحد تشخیص احساسات تک‌وجهی با استفاده از کپسول‌های تشخیص احساسات و واحد تشخیص احساسات چندوجهی از طریق همجوشی دیرهنگام.

در روش ارائه شده ابتدا، پیش‌پردازش‌های ساده‌ای بر روی داده‌های متنی اعمال می‌شود که شامل حذف کاراکترهای خطوط جدید (\n)، حذف فضاهای اضافی میان کلمات، نرمال‌سازی حروف عربی (مثلاً تبدیل "ك" به "ک") و تبدیل شکلک‌ها به متن هستند. سپس متنِ پیش‌پردازش شده و تصاویر مربوطه به واحدهای استخراج ویژگی ارسال می‌شوند. پارس‌برت نمایش‌های برداری از توکن‌های متن ایجاد می‌کند که این نمایش‌های برداری از طریق سازوکار خود-توجه[1] با اطلاعات زمینه‌ای غنی شده‌اند. این بردارهای پنهان، به‌جز بردار مربوط به توکن خاص [CLS]، به‌عنوان ویژگی‌های متنی به مدل کپسول ارسال می‌شوند تا کپسول مرتبط با احساس متن فعال شود. به‌طور مشابه، در وجه تصویری، مدل DINOv2 برای استخراج ویژگی‌های بصری به کار برده می‌شود. بردارهای خروجی DINOv2، نمایش‌های پنهان برداری از قطعات[2] تصویر هستند. این بردارهای پنهان به مدل کپسول داده می‌شود تا کپسول مربوط به احساس تصویر فعال شود. درنهایت، خروجی‌های کپسول‌های تشخیص احساسات تصاویر و کپسول‌های تشخیص احساسات متون با استفاده از همجوشی دیرهنگام ترکیب می‌شوند و تشخیص احساسات چندوجهی امکان‌پذیر می‌شود. همچنین به‌منظور تفسیر پیش‌بینی‌های انجام شده توسط شاخه‌های تک‌وجهی مدل از تکنیک لایم[3] استفاده شده است. در ادامه این بخش، در مورد هر یک از اجزای سازنده مدل پیشنهادی توضیحات تفصیلی ارائه خواهد شد.

۳-۱ استخراج ویژگی‌های متنی و تصویری

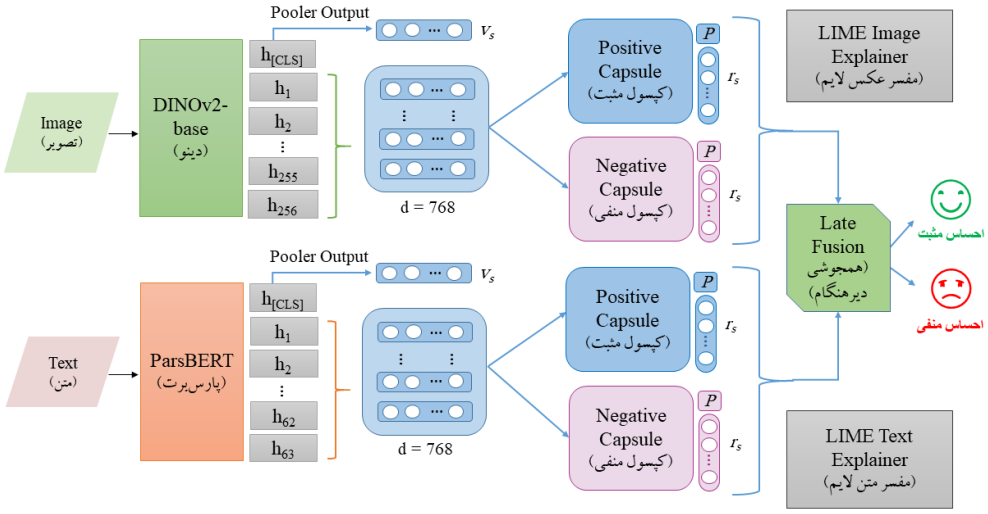
مدل پارس‌برت

مدل برت، یک مدل یادگیری عمیق پیش‌آموزش دیده است که در سال ۲۰۱۸ توسط محققان گوگل معرفی شد [۱۸]. برت براساس بخش رمزگذار (اِنکُدر) معماری ترانسفورمر معرفی شده توسط واسوانی و همکاران [۸]، بنا شده است. یک نوآوری مهم برت، ماهیت دوطرفه آن است که به مدل اجازه می‌دهد برخلاف مدل‌های سنتی یک طرفه، متن را از هر دو سمت چپ و راستِ یک توکن به‌طور هم‌زمان در نظر بگیرد. این مدل بر روی مجموعه داده‌های بزرگ متنی با استفاده از دو وظیفه یادگیری خودنظارتی که شامل مدل زبانی ماسک شده[4] و پیش‌بینی جمله بعدی[5] بودند، از قبل آموزش دیده است و سپس بر روی وظایف مختلف پردازش زبان طبیعی از جمله استنتاج زبان طبیعی، تشخیص موجودیت نامدار[6] و تحلیل احساسات ریز تنظیم[7] شده است.

توکن‌ساز برت بر پایه روش توکن‌سازی WordPiece عمل می‌کند که می‌تواند کلمات را به واحدهای زیرکلمه تجزیه کند.

یکی از قابلیت‌های مهم برت ایجاد تعبیه‌های کلمات با در نظرگرفتن اطلاعات زمینه‌ای و توانایی آن در مدیریت چندمعنایی واژگان است.

[1] Self-Attention Mechanism
[2] Patches
[3] LIME
[4] Masked Language Model (MLM)
[5] Next Sentence Prediction (NSP)
[6] Named Entity Recognition (NER)
[7] Fine-Tune



شکل (۱): روش پیشنهادی پژوهش حاضر

پارس‌برت نسخه‌ای تک زبانه از برت است که بر روی حجم عظیمی از متون فارسی از قبل آموزش دیده است. این مدل در تعدادی از وظایف پردازش زبان طبیعی از جمله تحلیل احساسات توانسته است عملکرد بهتری نسبت به برت چندزبانه از خود نشان دهد. پارس‌برت براساس معماری برت پایه با ۱۲ لایه یا بلوک ترانسفورمر، ۱۲ سرِ خود-توجه[1] و اندازه بردار پنهان ۷۶۸ است [۱۹].

در این پژوهش، به‌منظور استخراج ویژگی از متن، مدل از پیش‌آموزش‌دیده پارس‌برت که توسط گروه تحقیقاتی هوشواره [۱۹] بر روی مجموعه داده تحلیل احساسات دیجی‌کالا ریزتنظیم شده است، در حالت بدون گرادیان (با ثابت نگه‌داشتن پارامترهای مدل) بهره گرفته شده است.

مدل DINOv2

مدل DINOv2 که در سال ۲۰۲۳ توسط محققان هوش مصنوعی مِتا معرفی شد [۲۰]، مدلی از پیش آموزش دیده و مبتنی بر ترانسفورمرهای بینایی است. مدل DINOv2 بر روی حجم متنوع و عظیمی از داده‌های تصویری بدون برچسب با استفاده از روش یادگیری خود نظارتی آموزش داده شده است. برای پیش آموزش این مدل، از معماری شبکه معلم-دانش‌آموز استفاده شده است. همچنین، تکنیک‌های مختلفی برای مقیاس‌کردن فرآیند پیش‌آموزش از نظر اندازه مدل و داده‌ها، به کار برده شد. مدل از پیش آموزش‌دیده DINOv2 می‌تواند ویژگی‌های بصریِ همه‌منظوره تولید کند و استفاده از آن بدون انجام ریز تنظیم به سادگی امکان‌پذیر است. این مدل برای وظایف مختلف بینایی ماشین مانند طبقه‌بندی تصاویر، بازیابی تصویر و تقسیم‌بندی تصویر کاربرد دارد.

در این پژوهش، به‌منظور استخراج ویژگی از تصاویر، از مدل از پیش‌آموزش‌دیده DINOv2-base به‌صورت بدون گرادیان (با ثابت نگه‌داشتن پارامترهای مدل) استفاده شده است. پیش از ارسال تصاویر به مدل DINOv2، از ماژول AutoImageProcessor در کتابخانه Transformers استفاده شد. این ماژول به‌صورت خودکار مراحل پیش‌پردازش موردنیاز را مطابق با تنظیمات مدل از پیش‌آموزش‌دیده انجام می‌دهد.

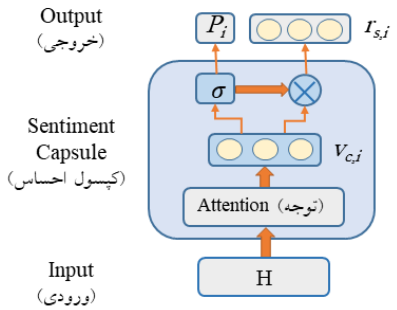
۳-۲ کپسول‌های تشخیص احساسات

این لایه، وظیفه طبقه‌بندی و تشخیص احساسات را به عهده دارد. شکل ۲ ساختار داخلی یک کپسول تشخیص احساس مورد استفاده در روش ارائه شده را نشان می‌دهد. تعداد این کپسول‌ها با تعداد کلاس‌های مجموعه داده سازگار است. به بیانی دیگر، وجود دو کپسول با دو کلاس مثبت و منفیِ مجموعه داده، مرتبط است.

[1] Self-Attention Head

بردارهای ویژگی H به این کپسول‌های احساس ارسال می‌شود. بردارهای تشکیل‌دهنده H (رابطه ۱) همان بردارهای پنهان ایجاد شده توسط مدل پارس‌برت و یا DINOv2 در مدل پیشنهادی هستند.

$$H = [h_1, h_2, ..., h_{N_s}] \quad (۱)$$



شکل (۲): ساختار داخلی یک کپسول

مقدار P_i در شکل ۲ نشان‌دهنده احتمال فعال‌سازی کپسول i ام است. از میان کپسول‌ها، کپسولی فعال می‌شود که مقدار P آن بیشینه باشد و به‌عنوان مثال در صورتی که کپسول احساس مثبت فعال شود، به این معنی است که کلاس پیش‌بینی شده توسط مدل برای داده ورودی، کلاس مثبت است. در روش ارائه شده هر کپسول از سه واحد تشکیل شده است که شامل واحد بازنمایی[1]، واحد احتمال[2] و واحد بازسازی[3] است. در ادامه، در خصوص نقش و جزئیات هر یک از واحدهای سازنده کپسول بحث خواهد شد.

واحد بازنمایی

به‌طورکلی این واحد، از سازوکار توجه به‌منظور ایجاد نمایش ویژگی کپسولی (بردار $v_{c,i}$ در شکل ۲) استفاده می‌کند. سازوکار توجه در مورد اهمیت کلمات سازنده متن قضاوت می‌کند و به‌طور مشابه برای داده‌های تصویری، این سازوکار می‌تواند به مدل کمک کند تا بر مهم‌ترین قطعات (پَچ‌های) سازنده تصویر تمرکز نماید. چگونگی محاسبه و اعمال وزن‌های توجه به صورت فرمول‌های ۲ تا ٤ است:

$$u_{i,t} = \tanh(W_w H + b_w) \quad (۲)$$

$$a_{i,t} = \frac{\exp\left(u_{i,t}^T u_w\right)}{\sum_t \exp\left(u_{i,t}^T u_w\right)} \quad (۳)$$

$$v_{c,i} = \sum_t a_{i,t} H \quad (٤)$$

بردارهای ویژگی H به لایه کاملاً متصل با تابع فعال‌ساز تانژانت هیپربولیک ارسال می‌شود تا بردار $u_{i,t}$ طبق فرمول ۲ به دست آید. در فرمول‌های ۲ و ۳ پارامتر b_w مقدار بایاس و ماتریس‌های W_w و u_w ماتریس‌های وزنی هستند که پارامترهایی آموزش‌پذیر محسوب می‌شوند. فرمول ٤، فرمول سافت‌مکس است که از طریق آن، وزن‌های توجه $(a_{i,t})$ تولید می‌شود. درنهایت، وزن‌های توجه بر بردارهای H اعمال می‌گردد و مجموع آن طبق فرمول ٤ محاسبه می‌شود.

واحد احتمال

در این واحد مطابق با فرمول ۵، احتمال فعال شدن کپسول، براساس بردار ویژگی $v_{c,i}$ که پیش‌تر از طریق واحد بازنمایی به دست آمده است، محاسبه می‌شود.

$$P_i = \sigma\left(W_{p,i} v_{c,i} + b_{p,i}\right) \quad (۵)$$

که P_i نمایانگر احتمال فعال‌سازی کپسول i ام است. نماد σ، تابع فعال‌ساز سیگموید را نشان می‌دهد و $W_{p,i}$ و $b_{p,i}$ به ترتیب ماتریس‌های وزنی و ماتریس‌های بایاس هستند.

واحد بازسازی

واحد بازسازی، بردار ویژگی معنایی $v_{c,i}$ را در ماتریس احتمال P_i ضرب می‌کند تا نمایش ویژگی بازسازی‌شده $(r_{s,i})$ طبق فرمول ٦ حاصل شود.

$$r_{s,i} = P_i v_{c,i} \quad (٦)$$

بردار v_s که در شکل ۱ و فرمول ۹ قابل مشاهده است، نمایش تجمیع‌شده‌ای از بردارهای پنهان H است که در روش ارائه شده مقدار آن طبق رابطه ۷ برابر با خروجی pooler_output (یا h_{pooler}) تولید شده توسط مدل پارس‌برت و یا مدل DINOv2 در نظر گرفته شده است.

$$v_s = h_{pooler} \quad (۷)$$

[1] Representation Module
[2] Probability Module
[3] Reconstruction Module

در توضیحات پیشین، به‌طور مفصل در خصوص واحدهای تشکیل‌دهنده کپسول صحبت شد. این واحدها در کنار هم و به صورت مکمل هستند. مطلب شایان توجه در مورد کپسول‌های تشخیص احساس این است که در درجه اول، هدف از آموزش چنین کپسول‌هایی، بیشینه‌سازی احتمال فعال‌سازی کپسول مرتبط با احساس داده ورودی است در حالی که خطای میان بردار بازسازی ($r_{s,i}$) و بردار v_s کمینه شود و دومین هدف کمینه‌سازی احتمال فعال شدن سایر کپسول‌های غیرمرتبط با احساس داده ورودی است، در حالی که خطای میان بردار بازسازی ($r_{s,i}$) و بردار v_s بیشینه شود. برای دستیابی به این اهداف، از یک تابع زیان مبتنی بر زیان لولا[1] استفاده می‌شود که در روابط ۸ تا ۱۰ تعریف شده است.

$$J(\theta)=\sum \max\left(0,1+\sum_{i=1}^{N} y_i P_i\right) \quad (۸)$$

$$U(\theta)=\sum \max\left(0,1+\sum_{i=1}^{N} y_i v_s r_{s,i}\right) \quad (۹)$$

$$L(\theta)=J(\theta)+U(\theta) \quad (۱۰)$$

به بیانی ساده، تابع J تلاش می‌کند تا احتمال فعال‌سازی کپسول مرتبط با برچسب داده ورودی را بیشینه کند، در حالی‌که این مقدار برای سایر کپسول‌ها (کپسول‌های غیرفعال) کاهش یابد. همچنین، تابع U تلاش می‌کند تا شباهت (ضرب نقطه‌ای) میان بردار بازسازی‌شده $r_{s,i}$ و بردار v_s در کپسول فعال بیشینه شود؛ به‌عبارت دیگر، خطای میان این دو بردار کاهش یابد. متقابلاً، برای کپسول‌های غیرفعال، این شباهت تا حد امکان کمینه شود، یعنی خطای میان بردار بازسازی‌شده $r_{s,i}$ و v_s افزایش یابد. در مجموع، تابع زیان نهایی L از جمع دو تابع J و U به دست می‌آید. همچنین، برای کپسول فعال مقدار y_i=-1 و برای سایر کپسول‌های غیرفعال مقدار y_i=+1 در نظر گرفته می‌شود.

۳-۳ همجوشی چندوجهی

طبق تعریف ارائه شده توسط بالتروشایتیس و همکارانش [۲۱]، همجوشی چندوجهی به مفهومی اشاره دارد که در آن اطلاعاتی از چندین نوع داده مانند تصویر، متن و صدا با یکدیگر ترکیب می‌شوند تا بتوان یک نتیجه نهایی را پیش‌بینی کرد. این نتیجه می‌تواند یک برچسب طبقه‌بندی باشد (مثل قطبیت مثبت یا قطبیت منفی) یا یک مقدار پیوسته باشد (در مسائل رگرسیون).

ژو و همکاران [۲۲] مروری جامع بر انواع روش‌های همجوشی در وظیفه تحلیل احساسات چندوجهی انجام دادند و این روش‌ها را به هشت دسته کلی تقسیم کردند. یکی از روش‌های همجوشی، همجوشی دیرهنگام است که تحت عنوان همجوشی در سطح تصمیم نیز شناخته می‌شود. در همجوشی دیرهنگام، تحلیل احساسات براساس هر وجه به صورت جداگانه انجام می‌شود. سپس سازوکارهای مختلفی اتخاذ می‌شود تا پیش‌بینی‌های محلیِ مربوط به هر وجه در تصمیم‌گیری و طبقه‌بندی نهایی احساسات، گنجانده شود. برخی از این سازوکارها شامل میانگین‌گیری، رأی‌گیری اکثریت و جمع وزنی هستند. یک حالت ساده برای انجام همجوشی دیرهنگام، محاسبه میانگین میان احتمال‌های به دست آمده از طبقه‌بندهای مربوط به هر وجه است. گونه‌ای دیگر از همجوشی، همجوشی زودهنگام است. در این نوع همجوشی، ابتدا ویژگی‌های استخراج شده از همه وجه‌ها به یکدیگر متصل می‌گردند و سپس به یک طبقه‌بند واحد ارسال می‌شوند.

در این پژوهش از روش همجوشی دیرهنگام استفاده شد، زیرا این رویکرد ضمن حفظ استقلال پردازش وجه‌ها و سادگی پیاده‌سازی، از افزایش بیش از حد بُعد بردار ویژگی که در همجوشی زودهنگام ناشی از الحاق مستقیم بردارها رخ می‌دهد نیز جلوگیری می‌کند. با توجه به این‌که در روش ارائه شده از همجوشی دیرهنگام با سازوکار میانگین‌گیری استفاده شده است، محاسبه تابع زیان مدل چندوجهی پیشنهادی براساس فرمول‌های ۸ تا ۱۰ به صورت فرمول‌های ۱۱ تا ۱۳ بازنویسی می‌شود:

$$J(\theta)=\sum \max \left(0,1+\sum_{i=1}^{N} y_i \left(P_{image,i}+P_{text,i}\right)/2\right) \quad (۱۱)$$

$$U(\theta)=\sum \max \left(0,1+\sum_{i=1}^{N} y_i \left(v_{image,s} r_{image,s,i}+v_{text,s} r_{text,s,i}\right)/2\right) \quad (۱۲)$$

$$L(\theta)=J(\theta)+U(\theta) \quad (۱۳)$$

[1] Hinge Loss

۳-٤ تکنیک لایم

لایم تکنیکی است که برای تفسیر پیش‌بینی‌های مدل‌های جعبه سیاه یادگیری ماشین کاربرد دارد. این تکنیک مستقل از مدل است، بدین معنی که می‌تواند برای هر مدل پیش‌بینی‌کننده‌ای صرف‌نظر از ساختار داخلی مدل اعمال شود [۲۳]. لایم توضیحات محلی تولید می‌کند و بر استدلال نهفته در پیش‌بینی مدل برای یک نمونه خاص تمرکز می‌کند. به بیانی دیگر، لایم می‌تواند مشخص کند که چرا یک نمونه خاص در یک کلاس خاص (به‌عنوان مثال کلاس ۰ یا کلاس ۱) طبقه‌بندی شده است. فرایند تولید تفسیرها با استفاده از تکنیک لایم اغلب شامل مراحل زیر است:

- انتخاب نمونه موردنظر: در این مرحله، نمونه داده‌ای که قرار است پیش‌بینی مدل برای آن تفسیر شود، انتخاب می‌گردد.
- ایجاد نمونه داده‌های جدید: با اعمال تغییرات جزئی در نمونه داده اصلی، مجموعه‌ای از داده‌های جدید تولید می‌شود و پیش‌بینی‌های مدل جعبه‌سیاه برای این نمونه‌ها دریافت می‌گردد.
- وزن‌دهی به نمونه‌های جدید: نمونه‌های تولید شده براساس شباهت و نزدیکی آن‌ها به نمونه اصلی، وزن‌دهی می‌شوند.
- آموزش یک مدل محلی: در این مرحله، یک مدل ساده و تفسیرپذیر (مانند طبقه‌بند خطی) بر روی داده‌های وزن‌دهی شده از مرحله قبل، آموزش داده می‌شود.
- تفسیر نتایج: مدل محلی برای تفسیر و توضیح این‌که چرا نمونه اصلی به یک کلاس خاص نسبت داده شده است، مورد استفاده قرار می‌گیرد.

این مراحل در فرمول‌های ۱٤ و ۱۵ خلاصه شده است:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \qquad (۱٤)$$

$$L(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z)\left(f(z) - g(z')\right)^2 \qquad (۱۵)$$

این دو فرمول (۱٤ و ۱۵) سازوکار اصلی لایم را برای تفسیر یک نمونه خاص x بیان می‌کنند. در این روش، هدف یافتن یک مدل ساده محلی g از میان مجموعه‌ای از مدل‌های تفسیرپذیر G است؛ به‌گونه‌ای که این مدل ساده بتواند رفتار مدل پیچیده اصلی f را در اطراف x به خوبی تقریب بزند. در فرمول ۱٤، تابع زیان $L(f, g, \pi_x)$ میزان اختلاف میان خروجی مدل پیچیده و مدل ساده را بر روی نمونه‌های مصنوعی وزن‌دارشده اندازه‌گیری می‌کند. تابع وزن π_x به هر نمونه مصنوعی براساس میزان شباهتش به x وزنی اختصاص می‌دهد، به‌طوری که نمونه‌های نزدیک‌تر به x تأثیر بیش‌تری بر یادگیری مدل ساده داشته باشند. از سوی دیگر، عبارت $\Omega(g)$ در همین فرمول به‌منظور کنترل میزان پیچیدگی مدل محلی، به تابع هدف اضافه می‌شود تا مدل‌های ساده‌تر (مانند مدل‌هایی با منطق قابل‌فهم‌تر یا تعداد ویژگی‌های کمتر) ترجیح داده شوند.

در مجموع، لایم یک مسئله بهینه‌سازی را حل می‌کند تا مدلی ساده و تفسیرپذیر بیابد که رفتار مدل پیچیده را در ناحیه اطراف نمونه مورد نظر شبیه‌سازی کند.

٤- آزمایش‌ها و یافته‌ها

٤-۱ مجموعه داده

در این پژوهش، از یک مجموعه داده تحلیل احساسات چندوجهی فارسی به نام مجموعه داده عکس نظر استفاده شده است که توسط زارعی‌نژاد و بهشتی‌فرد از توییتر و تلگرام جمع‌آوری شده است و شامل جفت‌های متن-تصویر است [۷]. در جدول ۱، دو نمونه از این مجموعه داده آورده شده است.

مجموعه داده عکس نظر با ده هزار نمونه برچسب‌گذاری‌شده، در حال حاضر بزرگ‌ترین مجموعه داده برای تحلیل احساسات چندوجهی در زبان فارسی است و به‌طور قابل‌توجهی از مجموعه داده‌های فارسی دیگر مانند مجموعه داده MPerSocial ارائه شده توسط ناصری کریموند و همکاران [٦] که تنها حاوی هزار نمونه داده است، بزرگ‌تر است. عکس نظر، مجموعه داده‌ای متوازن است و شامل تعداد برابری از نمونه‌ها در دو کلاس مثبت و منفی است. افزون بر این، برای پشتیبانی از پژوهش‌های چندزبانه، ترجمه متون فارسی به زبان انگلیسی نیز در این مجموعه داده گنجانده شده است. مجموعه داده عکس نظر در وب‌سایت کگل[۱] قابل دسترس

[۱] https://www.kaggle.com/datasets/penhangara1/aksnazar5

است. جدول ۲ آمار کلی این مجموعه داده را نشان می‌دهد.

جدول (۱): دو جفتِ متن-تصویر از مجموعه داده عکس نظر

تصویر		
متن فارسی	چمدان دستِ تو و ترس به چشمان من است این غم‌انگیزترین حالت غمگین شدن است	هر روز یک آغاز جدید است. پس با عشق، انرژی و امید به زندگی ادامه دهید.
متن انگلیسی	You have the suitcase in your hand, and fear is in my eyes; this is the saddest state of sadness.	Every day is a new beginning. So, continue to live with love, energy, and hope.
برچسب	۰ (قطبیت منفی)	۱ (قطبیت مثبت)

جدول (۲): آمار کلی مجموعه داده عکس نظر [۷]

گروه‌بندی	احساس مثبت	احساس منفی	جمع
تعداد جمله‌ها	۸۱۷۷	۷۸۲۶	۱۶۰۰۳
تعداد واژه‌ها	۱۳۲۱۶۳	۱۱۷۵۴۸	۲۴۹۷۱۱
تعداد تصاویر	۵۰۰۰	۵۰۰۰	۱۰۰۰۰
تعداد پیام‌های متنی	۵۰۰۰	۵۰۰۰	۱۰۰۰۰

۴-۲ معیارهای ارزیابی

ماتریس درهم‌ریختگی ابزاری ساده و مفید برای ارزیابی عملکرد یک مدل طبقه‌بند دودویی است. این ماتریس، شامل چهار عنصر کلیدی است: مثبت واقعی (TP)، منفی واقعی (TN)، مثبت کاذب (FP) و منفی کاذب (FN) [۲۴]. این چهار عنصر کلیدی برای محاسبه معیارهای ارزیابی مورد نیاز ضروری هستند.

- دقت (ACC): نسبت تعداد نمونه‌هایی که به طور صحیح طبقه‌بندی شده‌اند به تعداد کل نمونه‌ها است.
- این معیار برای مجموعه داده‌های متوازن مناسب است و براساس فرمول ۱۶ محاسبه می‌شود.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (۱۶)$$

- حساسیت (REC): حساسیت یا بازخوانی، نسبت تعداد نمونه‌های مثبتی که به‌طور صحیح طبقه‌بندی شده‌اند به کل نمونه‌های مثبت است. این معیار در فرمول ۱۷ نمایش داده شده است.

$$REC = \frac{TP}{TP + FN} \quad (۱۷)$$

- صحت (PREC): نسبت مثبت‌های واقعی (TP) به تعداد کل نمونه‌هایی که مدل به‌عنوان مثبت پیش‌بینی کرده است. این معیار طبق فرمول ۱۸ محاسبه می‌گردد.

$$PREC = \frac{TP}{TP + FP} \quad (۱۸)$$

- امتیاز F1 (F1-score): میانگین هارمونیک میان حساسیت و صحت است که در فرمول ۱۹ آمده است.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (۱۹)$$

۴-۳ ارزیابی عملکرد مدل پیشنهادی

توسعه یک مدل یادگیری عمیق معمولاً به سه فاز تقسیم می‌شود. در فاز اول، ابتدا مدل بر روی یک مجموعه داده متناسب با صورت مسئله مورد نظر آموزش داده می‌شود. در فاز دوم، مدل به صورت پیوسته بر روی مجموعه داده اعتبارسنجی ارزیابی می‌شود تا کیفیت تعمیم مدل بر داده‌های دیده نشده بررسی شود. در فاز سوم، زمانی که فرایند آموزش به اتمام رسید، مدل بر روی یک مجموعه داده تست ارزیابی می‌شود تا معیارهای ارزیابی نهایی مدل محاسبه شود [۲۴]. به همین منظور در این پژوهش، ۱۰ درصد از کل مجموعه داده به‌عنوان مجموعه داده تست جدا شده است و از ۹۰ درصد باقی‌مانده، ۲۰ درصد آن برای فاز اعتبارسنجی و مابقی به‌عنوان مجموعه داده آموزشی در نظر گرفته شده است. به‌منظور حفظ نسبت نمونه‌ها در هر یک از کلاس‌ها و جلوگیری از بروز عدم‌توازن در تقسیم‌بندی داده‌ها، در فرآیند جداسازی داده‌ها از روش نمونه‌گیری افراز شده[1] استفاده شد.

[1] Stratified Sampling

کلیه آزمایش‌ها و پیاده‌سازی مدل‌ها در این پژوهش، با استفاده از چارچوب یادگیری عمیق پایتورچ[1] انجام شده است. فراپارامترهای مدل پیشنهادی به‌طور خلاصه در جدول ۳ ارائه شده است.

انتخاب بسیاری از فراپارامترهای جدول ۳ براساس تجربه قبلی در کار با مدل‌های مشابه و بازخوردهای به‌دست‌آمده از عملکرد مدل در چند اجرای اولیه انجام شده است. در این میان، طول بردارهای پنهان کپسول برابر با ۷۶۸ تعیین شد، زیرا بُعد بردارهای ورودی کپسول (بردارهای تولیدشده توسط پارس‌برت و DINOv2) نیز ۷۶۸ بود. این انتخاب به‌منظور حفظ سازگاری ابعادی در معماری مدل پیشنهادی صورت گرفته است.

جدول (۳): فراپارامترهای مدل پیشنهادی

فراپارامتر	مقدار
حداکثر طول دنباله متنی	٦٤
طول بردارهای پنهان کپسول	۷٦۸
Capsule Dropout	۰/۳
اندازه دسته	۳۲
حداکثر تعداد دوره‌ها	۱۵
بهینه‌ساز	آدام
نرخ یادگیری	۰/۰۰۰۵
Random Seed	۲۰۱۸

حداکثر تعداد دوره‌های[2] آموزشی برابر با ۱۵ دوره در نظر گرفته شده است. به‌منظور جلوگیری از بیش‌برازش، از تکنیک توقف زودهنگام[3] با مقدار شکیبایی[4] برابر با ۳ استفاده شده است، به این معنی که اگر در طی سه دوره متوالی معیار امتیاز F1 برای داده‌های اعتبارسنجی بهبود نیابد، آموزش مدل زودتر از موعد متوقف شود. درنهایت برای ارزیابی در فاز تست، بهترین مدلِ ذخیره‌شده که بیشترین مقدار امتیاز F1 را در طول فرایند آموزش کسب کرده است، مورد استفاده قرار گرفت.

نتایج جدول‌های ٤ و ۵ و همچنین شکل ٦ نشان می‌دهد که مدل‌های چندوجهی از نظر دقت نسبت به مدل‌های تک‌وجهی بهتر عمل می‌کنند که بیانگر مزیت استفاده از چندین وجه برای تحلیل احساسات است. به‌علاوه، نسخه‌ای از پارس‌برت که بر روی مجموعه داده تحلیل احساسات دیجی‌کالا توسط گروه تحقیقاتی هوشواره ریز تنظیم شده است [۱۹]، عملکرد بهتری را نسبت به ParsBERTv3 نشان داده است. زیرا مدل Digikala ParsBERT انطباق بیش‌تری با دامنه مسئله و وظیفه تحلیل احساسات دارد. بااین‌حال، نتایج مدل ParsBERTv3 نیز نشان می‌دهد که حتی بدون ریز تنظیم، می‌توان اطلاعات معنادار مناسبی را استخراج کرد. افزون بر این، اگرچه انتظار می‌رود مدل‌های تک‌زبانه‌ای مانند پارس‌برت عملکرد بهتری نسبت به مدل چندزبانه mBERT داشته باشند، این فرضیه به صورت تجربی نیز مورد بررسی قرار گرفت.

با توجه به این‌که تنها پژوهش انجام شده بر روی مجموعه داده عکس نظر، مقاله ارائه‌دهندگان این مجموعه داده است [۷]، روش پیشنهادی پژوهش حاضر با آن روش در جدول ٦ مقایسه شده است. ماتریس درهم‌ریختگی نرمال‌شده در شکل ۳، نشان می‌دهد که مدل پیشنهادی حداقل ۹٦ درصد نمونه‌های هر کلاس را به‌درستی طبقه‌بندی می‌کند. به‌علاوه، مدل نسبت به کلاس خاصی سوگیری نداشته است و به‌طور مؤثر بین دو کلاس تمایز قائل شده است. همچنین در شکل ٤ و ۵ ماتریس‌های درهم‌ریختگی نرمال‌شده به تفکیک وجه قابل ملاحظه است.

جدول (٤): نتایج آزمایش‌ها برای مجموعه داده عکس نظر

استخراج گر ویژگی‌های بصری	استخراج گر ویژگی‌های متنی	مدل طبقه‌بند	دقت	صحت	حساسیت	F1-score
DINOv2 base [6]	Digikala ParsBERT [5]	Sentiment Capsules	۹٦/۵۰	۹٦/۹٦	۹٦/۰۰	۹٦/٤۸
DINOv2 base	ParsBERTv3 [7]		۹٤/٦۹	۹٤/٦۱	۹٤/۸۰	۹٤/۷۰
DINOv2 base	mBERT [8]		۹٤/۳۹	۹٤/۳۹	۹٤/۳۹	۹٤/۳۹

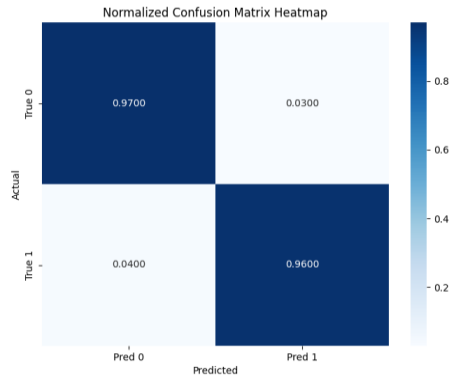
[1] PyTorch Version 2.6.0+cu124
[2] Epoch
[3] Early Stopping
[4] Patience
[5] https://huggingface.co/HooshvareLab/bert-fa-base-uncased-sentiment-digikala
[6] https://huggingface.co/facebook/dinov2-base
[7] https://huggingface.co/HooshvareLab/bert-fa-zwnj-base
[8] https://huggingface.co/google-bert/bert-base-multilingual-cased

جدول (۵): نتایج آزمایش‌ها به تفکیک وجه

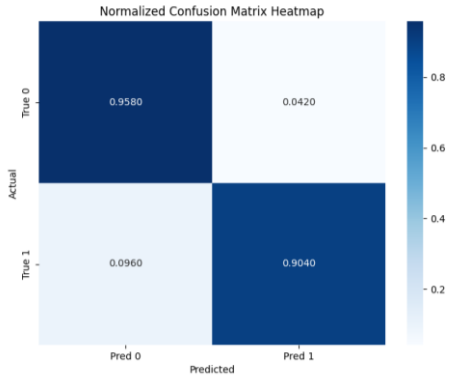
وجه	مدل	دقت	صحت	حساسیت	F1-score
فقط متن	Digikala ParsBERT - Capsules	۹۳/۱۰	۹۵/۵۶	۹۰/۴۰	۹۲/۹۰
	ParsBERTv3 - Capsules	۹۲/۷۰	۹۳/۳۰	۹۲/۰۰	۹۲/۶۴
	mBERT - Capsules	۸۸/۷۰	۸۹/۵۷	۸۷/۶۰	۸۸/۵۷
فقط تصویر	DINOv2 base - Capsules	۹۳/۵۰	۹۳/۹۳	۹۳/۰۰	۹۳/۴۶

جدول (۶): مقایسه روش پیشنهادی با سایر پژوهش‌ها

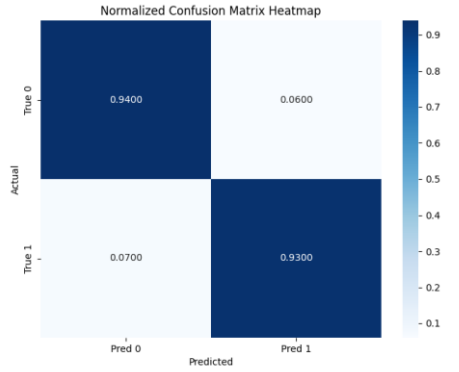
مدل	دقت	صحت	حساسیت	F1-score
روش پیشنهادی پژوهش حاضر	۹۶/۵۰	۹۶/۹۶	۹۶/۰۰	۹۶/۴۸
روش پیشنهادی زارعی‌نژاد و بهشتی‌فرد [۷]	۹۳/۳۹	۹۳/۰۹	۹۳/۲۰	۹۳/۱۴



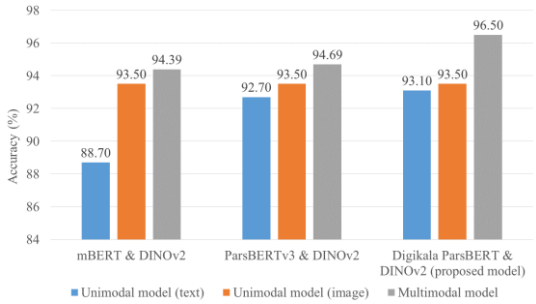
شکل (۳): ماتریس درهم‌ریختگی مدل چندوجهی پیشنهادی



شکل (۵): ماتریس درهم‌ریختگی شاخه تک‌وجهی مدل پیشنهادی (وجه متنی)



شکل (۴): ماتریس درهم‌ریختگی شاخه تک‌وجهی مدل پیشنهادی (وجه تصویری)



شکل (۶): نمودار مقایسه دقت مدل‌های تک‌وجهی و چندوجهی

٤-٤ تفسیر پیش‌بینی‌های تک‌وجهی

در این پژوهش به‌منظور تفسیر پیش‌بینی‌های تک‌وجهی، از تکنیک لایم استفاده شده است. لایم به‌صورت استاندارد برای تفسیرپذیری در داده‌های تک‌وجهی طراحی شده است و اعمال مستقیم آن بر مدل‌های چندوجهی، به دلیل ماهیت متفاوت وجه‌ها و دشواری در تعریف دستکاری[1] هم‌زمان بر چند نوع داده، اغلب امکان‌پذیر نیست. از این‌رو، در این پژوهش، لایم به‌طور جداگانه روی مدل‌های تک‌وجهی اعمال شده است تا علاوه بر ارزیابی نحوه تمرکز مدل بر نشانه‌های کلیدی، از کیفیت آموزش این مدل‌های تک‌وجهی اطمینان حاصل گردد. برای وجه تصویری، پس از آموزش و ذخیره مدل تک‌وجهی DINOv2-Capsules، با استفاده از ابزار LimeImageExplainer پیش‌بینی این مدل بر روی نمونه‌های تصویری تفسیر شد. در وجه متنی نیز پس از آموزش و ذخیره مدل Capsules Digikala ParsBERT-، از LimeTextExplainer بهره گرفته شد تا تفسیری از نحوه توجه مدل به واژگان کلیدی به‌دست آید. دو نمونه شامل یک نمونه مثبت واقعی و یک نمونه منفی واقعی برای هر وجه انتخاب شدند تا رفتار مدل در دو سناریوی موفق و متمایز بررسی شود. اگرچه این روش تعامل میان وجه‌ها را به‌صورت مستقیم نمایش نمی‌دهد، اما چشم‌اندازی از عملکرد مستقل هر شاخه تک‌وجهی فراهم می‌کند. لازم به ذکر است که تفسیرپذیری در مدل‌های چندوجهی تحلیل احساسات، به‌ویژه در مقایسه با وظایف چندوجهی دیگر مانند پرسش و پاسخ بصری[2]، کمتر مورد توجه پژوهش‌ها قرار گرفته است و همچنان به‌عنوان یک چالش مطرح است [۲۵]؛ این محدودیت در بخش ۵ نیز مورد اشاره قرار خواهد گرفت.

در شکل ۷، یک جفت نمونه متن-تصویر ارائه شده است که هر دو مدل (شاخه) تک‌وجهی (متنی و تصویری) توانسته‌اند به‌درستی آن را به‌عنوان کلاس مثبت پیش‌بینی کنند. در وجه متنی، واژگانی که بیش‌ترین تأثیر را در تصمیم مدل داشته‌اند با رنگ نارنجی مشخص شده‌اند. این واژه‌ها اغلب بار معنایی مثبتی دارند و نشان می‌دهند که مدل در تصمیم‌گیری خود بر مفاهیمی مرتبط با احساس مثبت تکیه کرده است. همچنین پیش‌بینی با اطمینان بالایی صورت گرفته است. در بخش تصویری این نمونه، خروجی تفسیر لایم نمایش داده شده است که با رنگ‌های مصنوعی (افزوده‌شده توسط الگوریتم) نواحی مهم تصویر را مشخص کرده است. نواحی سبز به‌عنوان بخش‌هایی از تصویر شناسایی شده‌اند که بیشترین تأثیر مثبت را بر پیش‌بینی مدل داشته‌اند (و شدت رنگ سبز با میزان تأثیر مثبت ارتباط دارد). در مقابل، نواحی که توسط الگوریتم با رنگ قرمز مشخص می‌شوند، بخش‌هایی هستند که تأثیر آن‌ها در پیش‌بینی مدل منفی بوده است. همان‌طور که دیده می‌شود، تمرکز مدل روی شیء گل در تصویر نمایانگر درک صحیح آن نسبت به محتوای مثبت است.

تصویر ورودی تصویر خروجی

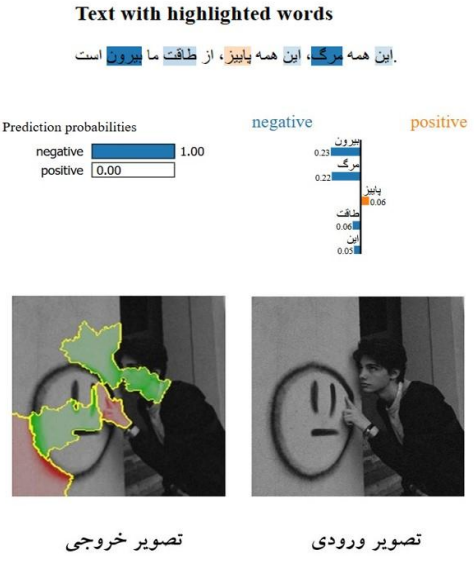
شکل (۷): تفسیرپذیری برای یک نمونه مثبت واقعی

در شکل ۸ نیز نمونه دیگری از یک جفت داده متن-تصویر ارائه شده است که هر دو مدل تک‌وجهی (متنی و تصویری) به‌درستی این نمونه را در کلاس منفی طبقه‌بندی کرده‌اند. در بخش متنی، واژه‌های مؤثر در تصمیم‌گیری مدل برای برچسب منفی با رنگ آبی مشخص شده‌اند؛ کلماتی مانند «مرگ» که دارای بار معنایی منفی هستند. در بخش تصویری نیز، نتایج تفسیر مبتنی بر لایم نشان می‌دهد که نواحی سبز افزوده‌شده در تصویر، نظیر حالت چهره فرد و شکلک غمگین روی دیوار، از دید مدل به‌عنوان شاخص‌هایی مرتبط با احساس منفی در نظر گرفته شده‌اند. در مقابل، نواحی

[1] Perturbation

[2] Visual Question Answering (VQA)

قرمز مانند دست فرد و بخشی از دیوار، با کاهش اطمینان مدل نسبت به پیش‌بینی کلاس منفی، اثر بازدارنده‌ای در تصمیم‌گیری نهایی داشته‌اند. همچنین، فضای تیره کلی تصویر می‌تواند در درک احساس منفی توسط مدل نقش داشته باشد.



شکل (۸): تفسیرپذیری برای یک نمونه منفی واقعی

۵- نتیجه‌گیری و پیشنهادها

در این پژوهش برای نخستین بار مدل‌های مبتنی بر معماری ترانسفورمر به‌منظور انجام تحلیل احساسات چندوجهی در زبان فارسی به کار رفته است. ابتدا، پیش‌پردازش‌هایی مانند تبدیل شکلک‌ها به متن انجام شد. زیرا شکلک‌ها می‌توانند دارای بار احساسی باشند که بر پیش‌بینی احساسات تأثیرگذار هستند. سپس استخراج ویژگی و ایجاد بردارهای تعبیه از طریق مدل‌های پیش‌آموزش‌دیده Digikala ParsBERT و DINOv2 به ترتیب برای داده‌های متنی و داده‌های تصویری انجام شد. سپس جهت تشخیص احساسات برای هر وجه، ویژگی‌های استخراج شده به کپسول‌های تشخیص احساس ارسال شدند. درنهایت پیش‌بینی احساسات چندوجهی از طریق اعمال همجوشی دیرهنگام صورت گرفت. افزون بر این، از یک تکنیک تفسیرپذیری به نام لایم برای تفسیرکردن پیش‌بینی‌های انجام شده توسط مدل‌های تک‌وجهی استفاده شد. نتایج آزمایش‌ها نشان دادند که مدل چندوجهی پیشنهادی به دقت ۹۶/۵ درصد و امتیاز F1 معادل ۹۶/۴۸ درصد دست یافته است و عملکرد بهتری نسبت به مدل‌های تک‌وجهی ارائه داده است. همچنین، نتایج آزمایش‌ها حاکی از آن است که استفاده از مدل Digikala ParsBERT در مقایسه با mBERT و ParsBERTv3، موجب افزایش دقت هم در مدل تک‌وجهی متنی و هم در مدل چندوجهی شده است. دلیل این برتری را می‌توان در فرآیند ریز تنظیم این مدل بر روی داده‌های مرتبط با تحلیل احساسات دانست؛ به‌طور مشخص، این مدل بر پایه نظرات کاربران وب‌سایت دیجی‌کالا ــ که حاوی محتوای احساسی به زبان فارسی است ــ ریز تنظیم شده است، امری که آن را با مسئله پژوهش حاضر سازگارتر می‌سازد.

در این پژوهش، یکی از محدودیت‌های تکنیک لایم این بود که تنها برای مدل‌های تک‌وجهی تصویری یا متنی قابل استفاده است. برای فراهم ساختن امکان تفسیرپذیری چندوجهی (multimodal XAI) لازم است مدل‌های تحلیل احساسات چندوجهی به‌گونه‌ای اختصاصی طراحی شوند تا بتوان درک بهتری از تأثیر و سهم هر وجه در پیش‌بینی نهایی به دست آورد.

در راستای توسعه بیشتر این پژوهش، پیشنهاد می‌شود در آینده وجه‌های جدیدی مانند صوت نیز به مدل افزوده شود. همچنین، بررسی روش‌های پیشرفته‌تر همجوشی می‌تواند موجب بهبود عملکرد مدل گردد. گسترش مدل برای تحلیل احساسات چندزبانه یا سطح‌بندی شدت احساسات نیز از دیگر مسیرهای پژوهشی قابل‌تأمل در آینده است.

References

[1] P. Mehta and S. Pandya, "A review on sentiment analysis methodologies, practices and applications," International Journal of Scientific and Technology Research, vol. 9, no. 2, pp. 601-609, 2020.

[2] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," Information Fusion, vol. 91, pp. 424-444, 2023, doi: 10.1016/j.inffus.2022.09.025.

[3] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," ACM computing surveys (CSUR), vol. 47, no. 3, pp. 1-36, 2015, doi: 10.1145/2682899.

[4] E. Cambria, D. Hazarika, S. Poria, A. Hussain, and R. Subramanyam, "Benchmarking multimodal sentiment analysis," in International conference on computational linguistics and intelligent text processing, 2017: Springer, pp. 166-179, doi: 10.1007/978-3-319-77116-8_13.

[5] K. Dashtipour, M. Gogate, E. Cambria, and A. Hussain, "A novel context-aware multimodal framework for persian sentiment analysis," Neurocomputing, vol. 457, pp. 377-388, 2021, doi: 10.1016/j.neucom.2021.02.020.

[6] A. Naseri Karimvand, S. Nemati, R. Salehi Ghegeni, and M. E. Basiri, "Multimodal Sentiment Analysis of Social Media Posts Using Deep Neural Networks," International Journal of Web Research, vol. 4, no. 1, pp. 1-9, 2021, doi: 10.22133/ijwr.2021.289091.1096.

[7] M. Zareinejad and Z. Beheshtifard, "Aks-Nazar: Introducing a Persian-English Dataset for Multimodal Sentiment Analysis," presented at the International Conference on Artificial Intelligence and Future Civilization, Tehran, 2025. [Online]. Available: https://civilica.com/doc/2195786.

[8] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[9] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in International conference on artificial neural networks, 2011: Springer, pp. 44-51, doi: 10.1007/978-3-642-21735-7_6.

[10] Y. Wang, A. Sun, J. Han, Y. Liu, and X. Zhu, "Sentiment analysis by capsules," in Proceedings of the 2018 world wide web conference, 2018, pp. 1165-1174, doi: 10.1145/3178876.3186015.

[11] Y. Cheng et al., "Sentiment analysis using multi-head attention capsules with multi-channel CNN and bidirectional GRU," IEEE Access, vol. 9, pp. 60383-60395, 2021, doi: 10.1109/ACCESS.2021.3073988.

[12] L. Sun, X. Li, B. Zhang, Y. Ye, and B. Xu, "Learning stance classification with recurrent neural capsule network," in Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I 8, 2019: Springer, pp. 277-289, doi: 10.1007/978-3-030-32233-5_22.

[13] M. Memari, S. M. Nejad, A. P. Rabiei, M. Eisaei, and S. Hesaraki, "BERTCaps: BERT Capsule for Persian Multi-Domain Sentiment Analysis," arXiv preprint arXiv:2412.05591, 2024, doi: 10.48550/arXiv.2412.05591.

[14] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020, doi: 10.48550/arXiv.2010.11929.

[15] S. Taylor and F. Fauzi, "Multimodal Sentiment Analysis for the Malay Language: New Corpus using CNN-based Framework," ACM Transactions on Asian and Low-Resource Language Information Processing, 2024, doi: 10.1145/3703445.

[16] P. Chaudhari, P. Nandeshwar, S. Bansal, and N. Kumar, "MahaEmoSen: Towards Emotion-aware Multimodal Marathi Sentiment Analysis," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 22, no. 9, pp. 1-24, 2023, doi: 10.1145/3618057.

[17] R. Das and T. D. Singh, "Image–text multimodal sentiment analysis framework of assamese news articles using late fusion," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 22, no. 6, pp. 1-30, 2023, doi: 10.1145/3584861.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.0480 2018, doi: 10.48550/arXiv.1810.04805.

[19] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, "Parsbert: Transformer-based model for persian language understanding," Neural Processing Letters, vol. 53, pp. 3831-3847, 2021, doi: 10.1007/s11063-021-10528-4.

[20] M. Oquab et al., "Dinov2: Learning robust visual features without supervision," arXiv preprint arXiv:2304.07193, 2023, doi: 10.48550/arXiv.2304.07193.

[21] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 2, pp. 423-443, 2018, doi: 10.1109/TPAMI.2018.2798607.

[22] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," Information Fusion, vol. 95, pp. 306-325, 2023, doi: 10.1016/j.inffus.2023.02.028.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin, "" Why should i trust you?" Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135-1144, doi: 10.1145/2939672.2939778.

[24] S. A. Hicks et al., "On evaluation metrics for medical applications of artificial intelligence," Scientific reports, vol. 12, no.1, p. 5979, 2022, doi: 10.1038/s41598-022-09954-8.

[25] N. Rodis, C. Sardianos, P. Radoglou-Grammatikis, P. Sarigiannidis, I. Varlamis, and G. T. Papadopoulos, "Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions," IEEe Access, 2024, doi: 10.1109/ACCESS.2024.3467062.

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 3, No. 1, Spring & Summer 2025, pp. 51 - 67

Persian Multimodal Sentiment Analysis Using Transformer-Based Models and Sentiment Detection Capsules

Parisa Rahmanian[1], Mohammad Reza Moosavi [2*], Mohammad Hadi Sadreddini[3]

[1]M.S. in Software Engineering, Department of Computer Science and Engineering and IT, School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran

[2]Associate Professor, Department of Computer Science and Engineering and IT, School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran

[3]Professor, Department of Computer Science and Engineering and IT, School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran

Article Information

Original Research Paper

Received:
2025 July 6

Accepted:
2025 August 14

Keywords:
Multimodal sentiment analysis, Persian sentiment analysis, transformer, capsule model, explainable AI

Corresponding Author*:
smmosavi@shirazu.ac.ir

Abstract

Sentiment analysis is the process of recognizing or classifying people's opinions and sentiments about a topic. Although earlier sentiment analysis research primarily relied only on text data, recent studies have shown that incorporation of multiple types of data in multimodal models improves performance. In this research, we address multimodal sentiment analysis in the Persian language, proposing a method based on transformer-based models for the first time in this context. For text feature extraction, ParsBERT model is used and DinoV2, a Vision Transformer-based model, is employed for extracting visual features. For sentiment recognition in each modality, sentiment detection capsules are utilized. Finally, to predict sentiment in the multimodal setup, we applied a late fusion technique at the final layer. Furthermore, a model-agnostic explainable AI technique, LIME, is used to gain insights into the predictions made by unimodal branches of the proposed multimodal model. Our experiments showed that our proposed model achieved 96.5% accuracy and 96.48% F1-score on the Aks-Nazar dataset.

: 10.22034/ABMIR.2025.23347.1137

E-ISSN: 2821-2037