Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال سوم، شماره اول، بهار و تابستان ۱٤۰٤، صفحات ۱۹۱ - ۱۷۹

مدیریت وجوه گم‌شده مبتنی بر یادگیری تقابلی در ادغام داده‌های چندوجهی برای تشخیص افسردگی در شبکه‌های اجتماعی

حامد مروی[۱]، ابوالفضل نادی [۲*]، محمد مهدی کیخا[۳]

[۱] دانشجوی کارشناسی ارشد علوم کامپیوتر، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران، تهران، ایران

[۲] استادیار گروه علوم کامپیوتر، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران، تهران، ایران

[۳] استادیار گروه علوم‌کامپیوتر، دانشگاه سیستان وبلوچستان، زاهدان، ایران

مقاله پژوهشی

تاریخ دریافت:
۱٤۰٤/٦/۳

تاریخ پذیرش:
۱٤۰٤/۸/۷

کلیدواژه‌ها:
مدل‌های چندوجهی، وجه گم‌شده، ادغام داده‌های چندوجهی، یادگیری تقابلی، شبکه‌های اجتماعی

نویسنده‌مسئول:
a.nadi@ut.ac.ir

چکیده

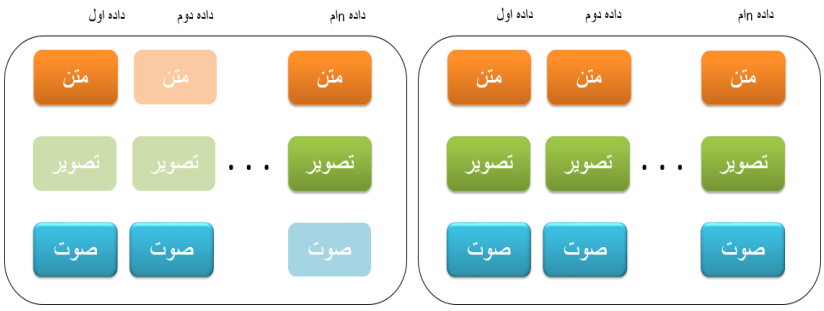
تحلیل داده‌های شبکه‌های اجتماعی اهمیت بنیادینی در استخراج الگوهای رفتاری کاربران دارد. مدل‌های چندوجهی که ترکیبی از اطلاعات متنی، تصویری و سایر منابع را به‌کار می‌گیرند، ابزارهای مناسبی برای این دست از تحلیل‌ها هستند. بااین‌حال، چالشی اساسی در این مدل‌ها، فقدان برخی وجوه در بخشی از نمونه‌های داده است؛ برای نمونه، کاربری ممکن است تنها متن منتشر کند و هیچ تصویری به اشتراک نگذارد. این مسئله سبب می‌شود مدل‌های چندوجهی نتوانند از تمامی اطلاعات موجود به‌طور کامل بهره‌برداری کنند. در این مقاله، روشی برای بهره‌گیری از داده‌های ناقص در مدل‌های چندوجهی ارائه شده است. ابتدا، مدل‌های تک‌وجهی برای پردازش هر وجه به‌صورت مستقل آموزش داده شدند. سپس، یک رمزگذار مبتنی بر یادگیری تقابلی طراحی و آموزش داده شد که هدف آن، برآورد بردار ویژگی وجوه گم‌شده با اتکا به ویژگی‌های وجوه موجود است. در نهایت، داده‌های متنی و تصویری) واقعی یا بازسازی‌شده (در یک مدل چندوجهی ادغام شده و برای تحلیل رفتار کاربران مورد استفاده قرار گرفت. نتایج آزمایش‌ها بر روی دادگان MDDL و بر پایه‌ی معیارهای دقت و F1 نشان می‌دهند که مدل چندوجهی پیشنهادی، با دقت ۹۰/۱۷ درصد و امتیاز F1 برابر با ۹۰/٦٤درصد، عملکرد بهتری نسبت به مدل‌های تک‌وجهی مبتنی بر متن (۸۷/۸۷درصد دقت) و تصویر (۷۳/۳۷درصد دقت) دارد. این نتایج مؤید آن است که مدل پیشنهادی، بدون نیاز به حذف داده‌های ناقص، از اطلاعات موجود به‌طور مؤثر بهره‌برداری می‌کند.

: 10.22034/ABMIR.2025.23578.1160

E-ISSN: 2821-2037

۱- مقدمه

افسردگی یکی از شایع‌ترین اختلالات روانی در جهان است که حدود ۵/۷ درصد از بزرگسالان را درگیر می‌کند و در مواردی می‌تواند منجر به خودکشی شود [۱]. با گسترش شبکه‌های اجتماعی، پژوهش‌های متعددی برای شناسایی نشانه‌های افسردگی از رفتار و محتوای کاربران انجام شده است. در این میان، مدل‌های چندوجهی[۱] به‌دلیل توانایی ترکیب داده‌های متنی و تصویری، نقش مهمی در بهبود دقّت و جامعیت تحلیل احساسات و پیش‌بینی وضعیت روانی کاربران دارند [۳][۲]. استفاده از داده‌های چندوجهی می‌تواند دقت پیش‌بینی و درک جامع از اطلاعات را بهبود بخشد، زیرا ترکیب وجه‌های مختلف به نمایش بهتری از یک پدیده منجر می‌شود. در ادغام داده‌های[۲] چندوجهی، سه روش اصلی برای ترکیب داده‌ها وجود دارد: ادغام اولیه [۳]، ادغام نهایی[۴] و ادغام ترکیبی.[۵] در ادغام اولیه، داده‌های خام پیش از هرگونه استخراج ویژگی با هم ترکیب می‌شوند. به‌عنوان مثال، اگر وجه‌های متن و تصویر داشته باشیم، جریان‌های ورودی متن و تصویر پس از هم‌ترازی و پیش‌پردازش اولیه به یک معماری مشترک داده می‌شوند تا مدل ویژگی‌ها را به‌صورت مشترک بیاموزد. در روش ادغام نهایی، بردارهای تعبیه‌شده‌ی هر وجه به‌دست می‌آیند و روی هر کدام مدل مستقلی اجرا می‌شود؛ سپس در انتها نتایج این مدل‌ها با هم ترکیب می‌شود تا به پاسخ نهایی برسیم. در روش ادغام ترکیبی، ابتدا بردار تعبیه‌شده‌ی هر وجه استخراج می‌شود. سپس روی آن‌ها پردازش‌های مستقلی انجام می‌گیرد و درنهایت این بردارها با هم ترکیب می‌شوند [۶][۵][۴]. با این حال، یکی از چالش‌های اصلی در این مدل‌ها که استفاده از برخی از وجه‌ها را دچار مشکل می‌کند این است که در بعضی از مواقع یک یا چند وجه در دسترس نبوده و یا ازدست‌رفته است؛ به این معنا که ممکن است یک یا چند وجه در برخی از نمونه‌های داده در دسترس نباشد. برای مثال، در سیستم‌های تشخیص احساسات یا تحلیل رفتار کاربران در شبکه‌های اجتماعی، ترکیب اطلاعات متنی، تصویری و صوتی می‌تواند به تشخیص دقیق‌تر و کامل‌تر کمک کند، امّا لزوماً پست‌های کاربران همه‌ی وجه‌ها را ندارند و ممکن است کاربر فقط متنی بنویسد و تصویر یا صوتی در پستش ارسال نکند. در دسترس نبودن هر یک از این وجوه یعنی بخشی از اطلاعاتی که می‌توانست به تصمیم‌گیری بهتر کمک کند، از دست می‌رود. به‌عنوان نمونه، در دسترس نبودن تصویر ممکن است منجر به ازدست‌رفتن ویژگی‌های بصری مهمی شود که به‌طور غیرمستقیم بر دقت تشخیص یا پیش‌بینی مدل تأثیرگذار باشد. در روش‌های گذشته، برای این‌که در زمان اجرا بتوان از نمونه داده‌هایی‌که یک یا چند وجه آن‌ها گم شده است استفاده کرد، آن وجه‌ها در مدل نادیده گرفته می‌شد و در نتیجه اطلاعات زیادی از دست می‌رفت. در شکل ۱ می‌بینیم که در شرایطی با سه وجه، نمونه‌های داده می‌توانند به دو دسته تقسیم شوند: داده‌های کامل (که اطلاعات را از هر سه وجه دارند) و داده‌های ناقص (که اطلاعات از یک یا چند وجه در دسترس نیستند)



شکل (۱): نمونه‌های با وجه‌های کامل (راست (و نمونه‌های با وجه‌های گم‌شده)چپ)

۲- پیشینه پژوهش

پژوهش‌های پیشین در زمینه تشخیص افسردگی از داده‌های شبکه‌های اجتماعی را می‌توان به سه دسته اصلی تقسیم کرد: (۱) روش‌های تک‌وجهی که معمولاً بر تحلیل داده‌های متنی متمرکز هستند، (۲) روش‌های چندوجهی که از ترکیب داده‌های متنی، تصویری و صوتی برای بهبود عملکرد استفاده می‌کنند، و (۳) روش‌های مدیریت داده‌های دارای وجه گم‌شده که اگرچه الزاماً در حوزه تشخیص افسردگی نیستند، اما از نظر فنی در زمینه‌ی یادگیری چندوجهی و مقابله با نبود داده در یک یا چند وجه مرتبط هستند. در ادامه، هر یک از این دسته‌ها به‌صورت جداگانه بررسی می‌شوند تا روند تکامل پژوهش‌ها و شکاف‌های موجود در این حوزه مشخص شود.

۲-۱ روش‌های تک‌وجهی

امروزه کاربران احساسات و فعالیت‌های خود را از طریق پست‌هایشان در شبکه‌های اجتماعی به اشتراک می‌گذارند. این امر فرصتی را برای ارزیابی وضعیت سلامت روان، از جمله افسردگی، فراهم کرده. پس می‌توانیم با استفاده از پست‌های کاربران، وضعیت روانی آن‌ها را ارزیابی کنیم و نشانه‌های مرتبط با سلامت روان آن‌ها را شناسایی کنیم. توییتر به‌عنوان منبعی ارزشمند برای درک افکار و احساسات افراد مطرح شده است، زیرا کاربران معمولاً تجربیات شخصی خود را به‌صورت عمومی به اشتراک می‌گذارند. این ویژگی، امکان تحلیل متن پیام‌های عمومی آن‌ها
را برای شناسایی الگوهای زبانی مرتبط با افسردگی فراهم می‌کند [۷] [۸].

روش‌های اولیه تحلیل احساسات تک‌وجهی بودند و عمدتاً بر روی متن تمرکز داشتند. با این حال، تکیه صرف بر داده‌های متنی برای استخراج احساسات بیان‌شده توسط انسان‌ها کافی نیست، زیرا معنای آنچه یک کاربر می‌گوید اغلب بر اساس رفتارهای غیرکلامی تغییر می‌کند [۹].

در مطالعه‌ای که توسط کوپراسمیت و همکاران [۱۰] انجام شد، داده‌های شبکه‌های اجتماعی تحلیل شد و به همبستگی معناداری بین زبان استفاده‌شده در توییت‌ها و شیوع علائم افسردگی پی برده شد. آن‌ها دریافتند که افرادی با سطح بالاتر افسردگی بیشتر تمایل دارند احساسات منفی را در توییت‌های خود بیان کنند.

دشپاند و رائو [۱۱] روشی مبتنی بر پردازش زبان طبیعی پیشنهاد کردند تا اطلاعات متنی موجود در توییتر را مدل‌سازی کنند. برخلاف روش‌های پیشین، آن‌ها از کیسه واژگان[1] برای نمایش متن توییت به‌عنوان یک بردار تنک استفاده کردند، که این روش به دسته‌بند امکان می‌دهد ویژگی‌های نهان را به‌صورت خودکار یاد بگیرد. طبقه‌بند بیز ساده[2] آموزش‌دیده آن‌ها برای معیار F1 به امتیاز ۰/۸۳ دست یافت، درحالی که ماشین بردار پشتیبان[3] به امتیاز ۰/۷۹ رسید. در مطالعه‌ای توسط دی‌چودری و همکاران [۱۲]، ارتباط بین شاخص‌های زبانی و افسردگی در پلتفرم‌های شبکه‌های اجتماعی بررسی شد. آن‌ها دریافتند که افراد دارای علائم افسردگی بیشتر از زبان منفی استفاده می‌کنند که نشان‌دهنده همبستگی بین بیان منفی و افسردگی است.

موراركا و همکاران یک مدل مبتنی بر روبرتا[4] برای دسته‌بندی بیماری‌های روانی در شبکه‌های اجتماعی توسعه دادند. دادگان آن‌ها شامل ۱۵۹/۱۷ پست از سایت ردیت[5] است. نتایج پژوهش آن‌ها نشان داد که مدل‌های مبتنی بر روبرتا نسبت به روش‌های سنتی پردازش زبان طبیعی مانند حافظه طولانی‌مدت کوتاه‌مدت[6] یا ماشین بردار پشتیبان دقت بالاتری در تشخیص بیماری‌های روانی از محتوای متنی شبکه‌های اجتماعی دارند [۱۳].

در مجموع، روش‌های تک‌وجهی اگرچه در استخراج نشانه‌های زبانی مؤثرند، اما قادر به بهره‌گیری از اطلاعات غیرکلامی نیستند و در نتیجه درک جامع از وضعیت روانی کاربر ارائه نمی‌دهند.

۲-۲ روش‌های چندوجهی

با وجود پیشرفت‌هایی که در تشخیص افسردگی با استفاده از داده‌های تک‌وجهی حاصل شده، داده‌های چندوجهی دارای مزایای بیشتری هستند. پژوهش در زمینه تشخیص افسردگی به‌کمک

[1] Bag of Words
[2] Naive Bayes
[3] Support Vector Machine
[4] RoBERTa
[5] Reddit
[6] Long Short-Term Memory

داده‌های چندوجهی بسیار مورد توجه قرار گرفته شده و تعداد قابل توجهی از پژوهشگران را به خود جذب کرده است [۱۴]. مدل‌های چندوجهی با چالش‌هایی روبه‌رو هستند که در سه دسته‌ی اصلی جای می‌گیرند: «ترجمه و هم‌ترازی داده‌ها[1]»، «پیچیدگی محاسباتی»، و «ادغام مؤثر وجوه مختلف». هم‌ترازی داده‌های چندوجهی به معنی پیدا کردن ارتباط‌های معنادار بین انواع داده‌های مختلف مانند تصویر و متن است که به دلیل تفاوت‌های ساختاری، یکی از چالش‌های بزرگ محسوب می‌شود. ترجمه چندوجهی نیز به تبدیل داده‌ها از یک وجه به وجه دیگر اشاره دارد. همچنین، پردازش داده‌های متنوع به‌طور هم‌زمان نیازمند منابع محاسباتی زیاد است، به‌خصوص در مدل‌های عمیق و مبدل‌ها[2] که با داده‌های بزرگ سروکار دارند. در نهایت، چگونگی ادغام مؤثر داده‌های چندوجهی به‌گونه‌ای که اطلاعات کلیدی حفظ شوند، یکی از مسائل مهمی است که می‌تواند بر دقت نهایی مدل تأثیرگذار باشد [۱۵] [۱۶].

فانگ و همکاران [۱۷] مدلی چندوجهی برای تشخیص افسردگی پیشنهاد دادند که آر مکانیزم توجه چندسطحی[3] استفاده کرده که از سه وجه داده صوتی، بصری و متنی استفاده می‌کند. این مدل برای استخراج ویژگی‌های هر وجه، از حافظه طولانی‌مدت کوتاه‌مدت و نوع دو طرفه آن[4] همراه با مکانیزم توجه استفاده می‌کند. در مرحله بعدی، خروجی‌ها را از طریق شبکه ادغام توجه[5] ترکیب می‌کند تا اطلاعات معنادارتر استخراج شود. نتایج آزمایش‌ها روی مجموعه داده DAIC-WOZ نشان داد که این مدل نسبت به روش‌های قبلی عملکرد بهتری دارد و توانسته با کمک مکانیزم توجه چندسطحی، ویژگی‌های کلیدی افسردگی را با دقت بیشتری شناسایی کند.

وانگ و همکاران [۱۸] روشی چندوجهی برای شناسایی افسردگی در شبکه‌های اجتماعی پیشنهاد دادند که از متن، تصاویر و ویژگی‌های رفتاری کاربران در پلتفرم ویبو[6] استفاده می‌کند. در این پژوهش، برای پردازش متن از مدل XLNet جهت استخراج بردارهای معنایی پست‌ها و پروفایل کاربران استفاده شد. تحلیل ویژگی‌های تصویری نیز بدون بهره‌گیری از شبکه‌های عصبی عمیق و صرفاً براساس توزیع رنگ‌های سرد و گرم انجام گرفت. علاوه بر این، ویژگی‌های رفتاری کاربران از جمله تعداد پست‌ها، فعالیت شبانه و میزان تعاملات اجتماعی مورد بررسی قرار گرفتند. مدل پیشنهادی آن‌ها، از Bi-GRU همراه با مکانیزم توجه برای پردازش متن و پرسپترون چندلایه[7] برای ادغام داده‌های چندوجهی استفاده کرد. نتایج این مدل نشان داد که ترکیب داده‌های متنی، تصویری و رفتاری، نسبت به روش‌های تک‌وجهی، دقت بیشتری در شناسایی افسردگی در شبکه‌های اجتماعی دارد.

با وجود بهبود عملکرد مدل‌های چندوجهی، این مطالعات فرض کرده‌اند که تمام وجوه داده در دسترس هستند و به مسئله داده‌های ناقص توجه نکرده‌اند.

۲-۳ روش‌های مدیریت وجه گم‌شده

بسیاری از دادگان چندوجهی شامل نمونه‌هایی هستند که همه وجه‌ها را در اختیار ندارند. این مشکل در داده‌های جمع‌آوری شده از شبکه‌های اجتماعی بسیار اتفاق می‌افتد زیرا ممکن است کاربر، متنی را پست کند اما تصویری را به آن پست الحاق نکند و این نمونه داده دارای وجه تصویر نیست. روش‌هایی که در ادامه بررسی می‌کنیم مستقیماً در حوزه تشخیص افسردگی متمرکز نیستند، اما از نظر فنی در زمینه‌ی یادگیری چندوجهی و مدیریت داده‌های ناقص کاربرد دارند. در پژوهش‌های قبلی، چند روش برای مدیریت داده‌های ناقص ارائه شده است. یکی از روش‌ها، حذف داده‌های ناقص است. در این روش، نمونه‌هایی که یک یا چند وجه را ندارند کنار گذاشته می‌شوند. این روش اگرچه ساده‌ترین رویکرد است، اما باعث از دست رفتن اطلاعات مفید و کاهش اندازه مجموعه داده‌ها می‌شود [۱۹]. روش دیگر، نادیده گرفتن وجوه گم‌شده[8] است. در این روش تمام داده‌های موجود در هر نمونه استفاده می‌شود، بدون آنکه برای وجه‌های غایب جایگزینی ایجاد کند. به عنوان مثال، اگر یک نمونه داده دارای دو وجه از سه وجه باشد، مدل تنها از همان دو وجه استفاده خواهد کرد. این رویکرد در

[1] Translation and Alignment
[2] Transformer
[3] Multi-Level Attention Mechanism
[4] Bidirectional Long Short-Term Memory
[5] Attention Fusion Network
[6] Weibo
[7] Multi-Layer Perceptron
[8] Missing Modality

مقایسه با حذف کامل داده‌ها، اطلاعات بیشتری را حفظ می‌کند. با این حال، مشکل اصلی یعنی بهره‌نبردن از وجه‌های گم‌شده همچنان پابرجاست [۱۹].

کای و همکاران [۲۰] در پژوهش خود از مدل‌های مولد[1] مانند خودرمزگذارها[2] و شبکه‌های مولد تخاصمی[3] برای بازسازی وجه گم‌شده استفاده کرده‌اند. نتایج آن‌ها نشان می‌دهد که این روش می‌تواند امکان استفاده از تمام داده‌ها را فراهم کند، اما در عین حال، پیچیدگی محاسباتی بالایی دارد و کیفیت داده‌های بازسازی‌شده همواره ثابت نیست [۲۰].

در پژوهشی دیگر روشی برای بازسازی بردار ویژگی[4] وجه گم‌شده ارائه شد که در آن، به جای بازسازی کامل وجه، تنها بردار ویژگی آن تخمین زده می‌شود. در این روش شبکه‌ای طراحی شد که با یادگیری ویژگی‌های مشترک بین وجوه مختلف، قادر است ویژگی‌های گم‌شده را از طریق اطلاعات موجود بازسازی کند. مدل ابتدا با داده‌هایی که تمام وجوه را دارند آموزش داده می‌شود و سپس، در هر مرحله یک یا چند وجه به‌صورت تصادفی حذف می‌شوند تا مدل یاد بگیرد ویژگی‌های ازدست‌رفته را بازسازی کند. نتایج این پژوهش نشان می‌دهد که این روش در مقایسه با بازسازی مستقیم داده‌ها، بهینه‌تر بوده و امکان تخمین ویژگی‌های ازدست‌رفته را با دقت بیشتری فراهم می‌کند [۲۱].

وانگ و همکاران روشی برای مدل‌سازی هم‌زمان ویژگی‌های مشترک و اختصاصی در داده‌های چندوجهی پیشنهاد داده‌اند. در این رویکرد، ویژگی‌های هر وجه به دو دسته ویژگی‌های مشترک (که بین تمام وجوه مشابه هستند) و ویژگی‌های اختصاصی (که مختص همان وجه‌اند) تقسیم می‌شوند. در صورت گم‌شدن یک وجه، این مدل به‌جای بازسازی مستقیم بردار ویژگی، از ویژگی‌های مشترک موجود در سایر وجوه برای استخراج اطلاعات استفاده می‌کند. به عبارت دیگر، مدل فرض می‌کند ویژگی‌های مشترک بین وجوه مختلف حاوی اطلاعات کافی برای پردازش هستند و در نتیجه، می‌تواند بدون نیاز به بازسازی مستقیم وجه گم‌شده، عملکرد خود را حفظ کند. نتایج این پژوهش نشان می‌دهد که این روش علاوه بر کاهش پیچیدگی مدل، امکان استفاده مؤثر از داده‌های ناقص را بدون نیاز به بازسازی مستقیم فراهم می‌کند [۲۲].

رضا و همکاران برای مدیریت وجه‌های گم‌شده در مدل‌های چندوجهی، روشی براساس حذف تصادفی وجه‌ها در مرحله آموزش ارائه دادند. در این روش، در هر مرحله آموزش، به‌صورت تصادفی یک یا چند وجه از داده‌ها حذف می‌شود. این کار مدل را مجبور می‌کند تا بازنمایی‌های مقاومی یاد بگیرد که بتوانند در زمان تست، حتی در صورت ناقص بودن داده‌های ورودی، عملکرد قابل قبولی داشته باشند. برخلاف روش‌هایی که برای بازسازی وجوه گم‌شده از شبکه‌های مولد تخاصمی یا خودرمزگذارها استفاده می‌کنند، این روش تلاشی برای بازسازی ندارد. بلکه به مدل آموزش می‌دهد تا صرفاً با استفاده از داده‌های در دسترس، پیش‌بینی دقیقی انجام دهد. این رویکرد نه‌تنها هزینه محاسباتی را کاهش می‌دهد، بلکه در شرایط واقعی که برخی از وجوه داده ممکن است همواره در دسترس نباشند عملکرد مطلوبی از خود نشان می‌دهد [۲۳].

وانگ و همکاران روشی برای یادگیری چندوجهی در شرایطی که برخی وجوه در داده‌های ورودی ناقص هستند، ارائه داده‌اند. ابتدا برای هر وجه، مدل‌های معلم[5] تک‌وجهی آموزش داده می‌شوند که ویژگی‌های مربوط به آن وجه را استخراج می‌کنند. سپس از طریق تقطیر دانش[6] برای انتقال اطلاعات از این مدل‌های معلم به یک مدل دانش‌آموز[7] چندوجهی استفاده می‌شود. مدل دانش‌آموز علاوه بر ادغام داده‌های چندوجهی، یاد می‌گیرد که در صورت نبود برخی وجوه، ویژگی‌های وجوه گم‌شده را تخمین بزند. این یادگیری از طریق حذف مصنوعی وجود در مرحله آموزش انجام می‌شود. در نتیجه مدل دانش‌آموز قادر است با استفاده از وجوه موجود، بردار ویژگی وجوه گم‌شده را بازسازی کند و از آن‌ها هنگام ادغام داده استفاده کند [۲۴].

[1] Generative Model
[2] Auto Encoder
[3] Generative Adversarial Network
[4] Feature Vector
[5] Teacher Model
[6] Knowledge Distillation
[7] Student Model

در مجموع، پژوهش‌های پیشین در زمینه‌ی مدیریت وجه گم‌شده به دو رویکرد اصلی تقسیم می‌شوند. گروه اول، روش‌های بازسازی‌محور هستند که با استفاده از شبکه‌های مولد مانند خودرمزگذارها یا شبکه‌های مولد تخاصمی تلاش می‌کنند داده ازدست‌رفته را بازسازی کنند. این روش‌ها هرچند دقیق‌اند، اما از نظر محاسباتی سنگین بوده و نیاز به داده آموزشی زیاد و تنظیم‌گرهای متعدد دارند و در داده‌های واقعی پایداری محدودی دارند. گروه دوم شامل روش‌هایی است که سعی می‌کنند مدل را نسبت به فقدان وجه مقاوم کنند، بدون آن‌که وجه گم‌شده را بازسازی کنند؛ مانند حذف تصادفی وجه‌ها در آموزش یا مدل‌سازی ویژگی‌های مشترک میان وجوه. این روش‌ها سبک‌تر هستند، اما معمولاً فقط از وجوه موجود استفاده می‌کنند و بازنمایی دقیقی از وجه غایب به دست نمی‌دهند. بنابراین هنوز نیاز به روشی وجود دارد که در عین سادگی و کارایی، بتواند در داده‌های شبکه‌های اجتماعی، ویژگی‌های وجه گم‌شده را با استفاده از هم‌ترازی تقابلی بین وجوه مختلف تخمین بزند.

در بخش بعدی، روش پیشنهادی برای مدیریت وجه گم‌شده توضیح داده خواهد شد. ابتدا مدل‌های تک‌وجهی معرفی می‌شوند و سپس نحوه استفاده از یادگیری تقابلی برای تخمین بردار ویژگی وجه گم‌شده بررسی خواهد شد.

۳- روش پیشنهادی

در سال‌های اخیر، برخی پژوهش‌ها به جنبه‌هایی از داده‌های ناقص در مدل‌های چندوجهی اشاره کرده‌اند، اما عموماً این مسئله به‌صورت محدود یا در قالب حذف داده‌های ناقص بررسی شده است. پژوهش حاضر با تمرکز ویژه بر شرایطی که یکی از وجوه (به‌ویژه تصویر) در دسترس نیست، تلاش می‌کند با بهره‌گیری از یادگیری تقابلی میان وجوه متنی و تصویری، بازنمایی مؤثری از وجه گم‌شده به‌دست آورد و این چالش را حل کند.

مدل پیشنهادی بر پایه ساختاری چندمرحله‌ای طراحی شده است که هدف آن تحلیل داده‌های چندوجهی و مدیریت وجوه گم‌شده است. در گام اول، دو مدل تک‌وجهی[1] برای پردازش متن و تصویر آموزش داده می‌شوند تا بتوانند به‌طور مستقل ویژگی‌های مربوط به هر وجه را استخراج کرده و پیش‌بینی انجام دهند. در مرحله بعد، یک رمزگذار[2] مبتنی بر یادگیری تقابلی به‌کار گرفته می‌شود تا بردار ویژگی مربوط به وجه گم‌شده را برآورد کند. در نهایت، مدل چندوجهی با ترکیب ویژگی‌های متنی و تصویری–چه به‌صورت واقعی و چه تخمین‌زده–فرآیند دسته‌بندی را انجام می‌دهد. در ادامه، هر یک از این مراحل به‌تفصیل توضیح داده خواهند شد.

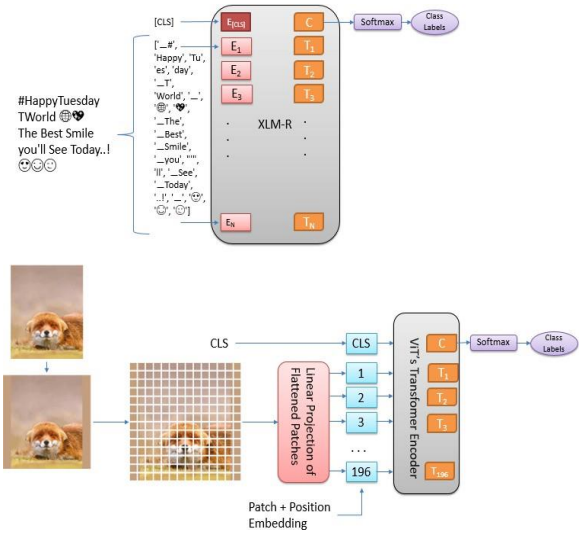
ابتدا مدل‌های تک‌وجهی برای پردازش متن و تصویر آموزش داده شده‌اند. این مدل‌ها به‌طور مستقل ویژگی‌های مربوط به هر وجه را استخراج کرده و پیش‌بینی می‌کنند که آیا یک توییت مربوط به یک کاربر افسرده است یا خیر. ابتدا، مدل‌های تک‌وجهی آموزش داده شدند. برای پردازش متن، از مدل XLM-RoBERTa استفاده شده است. این مدل به دلیل آموزش بر روی داده‌های چندزبانه، قادر به پردازش توییت‌هایی به زبان‌های مختلف است. متن خام مستقیماً به مدل داده می‌شود. خروجی توکن [CLS] برای نمایش ویژگی‌های کلی متن استخراج شده و از آن برای تشخیص افسرده یا غیرافسرده بودن توییت استفاده می‌شود.

برای پردازش تصویر، از مدل ViT[3] استفاده شده است. ورودی این مدل باید تصاویری با اندازه ۲۲٤×۲۲٤ باشد؛ به همین منظور، ابتدا تصویر مقیاس‌بندی می‌شود تا یکی از ابعاد (طول یا عرض) به ۲۲٤ برسد، سپس–در صورتی که بعد دیگر کوچک‌تر از ۲۲٤ باشد–نواحی خالی با میانگین رنگ پیکسل‌های تصویر پر می‌شود[4]. تصویر ورودی در مدل ViT به ۱۹٦ بخش ۱٦×۱٦ پیکسلی تقسیم می‌شود و از هر بخش، یک بردار ویژگی استخراج می‌گردد. این بردارها وارد مبدل می‌شوند و مدل بر اساس آن‌ها، نتیجه‌گیری نهایی را انجام می‌دهد.

در تصویر ۲ معماری کلی مدل‌های متنی و تصویری نشان داده شده است: بخش بالا مربوط به پردازش متن توسط مدل XLM-RoBERTa است–ابتدا، ورودی به واحدهای زبانی شکسته می‌شود[5] و سپس پردازش شده و ویژگی‌های آن استخراج می‌گردد؛ بخش پایین نیز به پردازش تصویر اختصاص دارد که با استفاده از

[1] Single Modal Model
[2] Encoder
[3] Vision Transformer
[4] Padding
[5] Tokenize

مدل ViT انجام می‌شود؛ این مدل ابتدا تصویر را به قطعات کوچک‌تر تقسیم می‌کند و سپس با بهره‌گیری از مکانیسم توجه[1]، ویژگی‌های آن را استخراج می‌کند.



شکل (۲): معماری مدل‌های تک‌وجهی: تصویر بالا مدل متنی XLM-RoBERTa و تصویر پایین مدل تصویری ViT

در این مرحله، مدل‌های تک‌وجهی برای متن و تصویر توانسته‌اند به‌صورت مستقل ویژگی‌های خود را استخراج کرده و بر اساس آن تصمیم‌گیری کنند. این مدل‌ها در مرحله بعدی به‌عنوان بخش‌های پایه برای مدل چندوجهی مورد استفاده قرار خواهند گرفت. راه‌حل ما برای مسئله وجه از دست‌رفته، تخمین بردار ویژگی آن وجه است. برای این کار، از یادگیری تقابلی[2] استفاده می‌کنیم؛ روشی که در مدل‌های شناخته‌شده‌ای مانند SimCLR [۲۵] و CLIP [۲۶] به‌کار رفته است. در یادگیری تقابلی، تلاش می‌شود نمونه‌هایی که به هم شباهت دارند، بردارهای ویژگی نزدیکی داشته باشند و در مقابل، نمونه‌های غیرمشابه دارای بردارهای دور از هم باشند [۲۷]. در مواردی که وجه تصویری در دسترس نیست، می‌کوشیم با استفاده از بردار ویژگی متن، بردار متناظر تصویر را تخمین بزنیم. برای این کار، یک رمزگذار[3] به نام T2V[4] طراحی و آموزش داده می‌شود؛ این مدل، بردار ویژگی متن را به‌عنوان ورودی دریافت کرده و می‌کوشد بردار ویژگی متناظر تصویر را برای همان نمونه برآورد کند [۲۹]. برای آموزش مدل T2V از یادگیری تقابلی استفاده می‌شود تا تصویر مجازی (تخمین زده) تا حد امکان به بردار ویژگی تصویر واقعی نزدیک باشد. معیار نزدیکی، تشابه کسینوسی[5] بین دو بردار است. در یادگیری تقابلی، هدف این است که بردار ویژگی مجازی هر نمونه به تصویر خودش نزدیک باشد، و با تصاویر سایر نمونه‌ها فاصله داشته باشد. ابتدا با رمزگذارهای متن و تصویر که به‌صورت تک‌وجهی آموزش دیدند، بردار ویژگی‌های نمونه داده‌هایی که در آن‌ها هر دو وجه موجودند محاسبه می‌شوند. فرض می‌کنیم که در هر گروه[6]، مدل n نمونه به‌عنوان ورودی دریافت می‌کند. در این حالت، n بردار ویژگی متن با نمادهای $T_1,\ldots,T_n$ و n بردار ویژگی تصویر با نمادهای $I_1,\ldots,I_n$ در اختیار خواهیم داشت. تابع هزینه[7] به‌صورت زیر است:

$$L = \frac{1}{n}\sum_{i=1}^{n}(1-\cos(V_i, I_i) + \max(0, x_i - \text{margin})) \qquad (۱)$$

$$x_i = \frac{1}{(n-1)}\sum_{j\neq i}\cos(V_i, I_j) \qquad (۲)$$

که در آن:

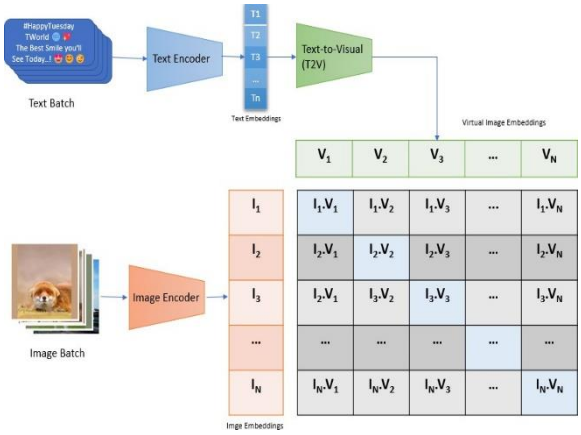
- V_i بردار ویژگی مجازی (تخمین‌زده) برای نمونه i است.
- I_i بردار ویژگی تصویر برای نمونه i است.
- $\cos(V_i, I_i)$ میزان تشابه کسینوسی بین بردار ویژگی‌های مجازی و واقعی نمونه i است. پس برای نمونه i، تابع هزینه از جمع دو بخش تشکیل شده است:

$\max(0, x_i\text{-margin})$
$1-\cos(V_i, I_i)$

بخش $1-\cos(V_i, I_i)$ موجب می‌شود که بردار ویژگی مجای i تا حد امکان شبیه به بردار ویژگی تصویر i شود؛ بخش دوم نیز باعث می‌شود بردار ویژگی مجازی i از بردار ویژگی تصویر نمونه‌های دیگر دور شود. پارامتر margin تضمین می‌کند اگر میزان تشابه کسینوسی بین بردار ویژگی مجازی i با بردار ویژگی تصویر j بیشتر از مقدار margin از هم دور باشند، دیگر نیازی به «دورتر کردن» این دو بردار نباشد و قسمت دوم برابر با صفر شود. مدل T2V

[1] Attention Mechanism
[2] Contrastive Learning
[3] Encoder
[4] Text-toVisual Encoder
[5] Cosine Similarity
[6] Batch
[7] Loss Function

یک پرسپترون چندلایه ساده بدون لایه‌ی پنهان[1] است که اندازه لایه‌ی ورودی ۷۶۸ و اندازه لایه‌ی خروجی نیز ۷۶۸ است. در شکل ۳ نیز نشان داده شده است: اگر ماتریسی بسازیم که در سطرهای آن بردار ویژگی مجازی تصویر را قرار دهیم و در ستون‌ها بردار ویژگی تصویر واقعی را، آن‌گاه درایه‌های ماتریس میزان تشابه کسینوسی بین بردارهای مجازی و واقعی تصویر خواهند بود؛ می‌کوشیم قطر اصلی ماتریس (که با رنگ آبی مشخص شده) به ۱ نزدیک شود و سایر درایه‌ها نزدیک به صفر شوند. معیار ارزیابی T2V را میانگین تشابه کسینوسی بین جفت بردارهای تصویر مجازی با تصویر واقعی‌اش قرار داده‌ایم که در بخش نتایج بررسی می‌شود.



شکل (۳): آموزش مدل T2V

پس از آموزش مدل T2V، گام بعدی ادغام[2] ویژگی‌های متنی و تصویری برای انجام دسته‌بندی چندوجهی است. در این ساختار، اگر تصویر موجود باشد، بردار ویژگی واقعی آن به‌کار می‌رود؛ در غیر این صورت، بردار ویژگی تصویر با کمک مدلT2V و براساس بردار متنی تخمین زده شده و جایگزین می‌شود. در این مرحله، ابتدا بردار ویژگی‌های متنی برای تمام داده‌ها استخراج می‌شود؛ سپس برای نمونه‌هایی که تصویر در دسترس دارند، بردار ویژگی تصویری واقعی نیز به‌دست می‌آید؛ اگر در داده‌ای تصویرش موجود نبود، بردار ویژگی‌اش تخمین زده می‌شود–این ساختار در شکل ۳ دیده می‌شود. در ادامه، بردار ویژگی متن و تصویر (واقعی یا تخمینی) با هم ترکیب‌شده[3] شده و به یک شبکه پرسپترون چندلایه داده می‌شود تا عمل دسته‌بندی انجام گیرد. ورودی این لایه، بردارهای ادغام‌شده‌ی متنی و تصویری است که دارای اندازه‌ی ۷۶۸+۷۶۸ برابر ۱۵۳۶ بعد است. در این معماری، لایه مخفی اول بعد را به ۵۱۲ کاهش می‌دهد و در لایه بعدی به ۲۵۶؛ سپس لایه خروجی –۲ بعدی است که به‌کمکش کلاس افسرده/غیرافسرده معلوم می‌شود.

جدول (۱): پارامترهای مدل‌های مورد استفاده

مدل	پارامتر	توضیحات
XLM RoBERTa	مدل متن حداکثر طول دنباله ابعاد خروجی	Xlm-roberta-base 128 768
ViT	مدل تصویر شکل ورودی نرمال‌سازی ابعاد خروجی	google/vit-base-patch16-224 3 × 224 × 224 Normalize(mean=[0.5, 0.5, 0.5],std=[0.5, 0.5, 0.5]) 768
T2V	مدل متن به تصویر ورودی خروجی تابع هزینه نرخ یادگیری تعداد تکرار اندازه دسته	T2V Encoder 768 768 Contrastive Learning 0.004 5 256
Fusion	ابعاد ادغام	768 + 768 = 1536
Multimodal Classification	معماری شبکه	Dense | 1536 → 512 → 256 → 2

در جدول ۱ توضیحاتی درباره مدل‌های تک‌وجهی متن و تصویر، مدل T2V و مدل چندوجهی آمده است. همچنین نرخ‌یادگیری[4]، تعداد تکرار[5] و اندازه دسته[6] مدل T2V گزارش شده.

پس از معرفی روش پیشنهادی، لازم است عملکرد آن را در شرایط مختلف ارزیابی کنیم. در بخش بعدی، نتایج آزمایش‌ها ارائه شده و نتایج بررسی می‌شوند.

[1] Hidden Layer
[2] Fussion
[3] Concatenate
[4] Learning Rate
[5] Epoch
[6] Batch Size

٤- آزمایش و نتایج

در این پژوهش، از دادگاه MDDL که در مقاله [۲۸] معرفی شده است، استفاده شده است. این دادگان شامل مجموعه‌ای از توییت‌های کاربران شبکه اجتماعی توییتر است که برای تحلیل و تشخیص افسردگی طراحی شده است. داده‌های این مجموعه شامل اطلاعات متنی و تصویری است که از پست‌های کاربران استخراج شده‌اند. این مجموعه داده شامل دو گروه «کاربر افسرده» و «کاربر غیر افسرده» را در بر می‌گیرد. کاربران افسرده کسانی هستند که در محتوای منتشرشده‌ی خود به افسردگی اشاره کرده‌اند؛ مثلاً عباراتی مانند «من دچار افسردگی شده‌ام» یا «پزشک به من گفته که افسرده‌ام» را در توییت‌هایشان نوشته‌اند. در مقابل، کاربران غیر افسرده کسانی هستند که در محتوای متنی خود اشاره‌ای به افسردگی نداشته‌اند و از کلمه » افسرده«استفاده نکرده‌اند.

از هر کاربر، مجموعه‌ای از ویژگی‌ها در سه حوزه متن، تصویر و تعاملات اجتماعی استخراج شده است. این ویژگی‌ها شامل اطلاعاتی درباره ساختار شبکه اجتماعی کاربر) مانند تعداد دنبال‌کنندگان و تعاملات (، محتوای متنی پست‌ها، تصویر به اشتراک گذاشته‌شده و همچنین برچسب افسرده بودن یا نبودن کاربر است. در مجموع، ۳۵۷/۲۱۵ نمونه داده در اختیار داریم که از این تعداد، ۳٤/۳۹۳ نمونه دارای برچسب مثبت) افسرده (هستند. این آمار بیانگر نامتوازن بودن مجموعه داده است، به‌طوری که تنها حدود ٦/۹ درصد از نمونه‌ها مربوط به کاربران افسرده بود و باقی موارد به کاربران غیرافسرده تعلق دارد. برای مدیریت این نامتوازنی، در مرحله آموزش مدل‌های تک‌وجهی و همچنین هنگام آموزش مدل چندوجهی، داده‌ها را متوازن کرده‌ایم (۵۰ درصد افسرده، ۵۰ درصد غیرافسرده) تا مدل دچار سوگیری[1] نسبت به کلاس اکثریت نشود. داده‌ها به سه بخش آموزش، اعتبارسنجی و آزمون با نسبت ۷۰، ۱۵ و ۱۵ درصد تقسیم شدند تا از تعمیم‌پذیری مناسب مدل اطمینان حاصل شود.

در مراحل ابتدایی، تنها از داده‌هایی استفاده شده که تمامی وجوه اطلاعاتی–از جمله تصویر– را دارا بودند. برای شبیه‌سازی شرایط نبود وجه تصویری، ۳۰ درصد از نمونه‌ها به‌صورت تصادفی از میان داده‌های کامل انتخاب و به‌عنوان نمونه‌های فاقد تصویر علامت‌گذاری شدند. به دلیل تصادفی بودن انتخاب، نسبت نمونه‌ها در هر کلاس تقریباً ثابت باقی ماند. این نمونه‌ها به گونه‌ای تنظیم شدند که مدل در حین آموزش و در فرایند ادغام، بتواند شرایط فقدان وجه تصویری را نیز تجربه کرده و آن را در یادگیری خود لحاظ کند. این مجموعه‌داده برای آموزش یک مدل چندوجهی باهدف تشخیص افسردگی در شبکه‌های اجتماعی به کار گرفته شده است. افزون بر آن، مدل برای سنجش توانایی خود در شرایطی که یکی از وجوه) تصویر(در دسترس نیست، نیز ارزیابی شد تا مشخص شود آیا می‌تواند در چنین موقعیت‌هایی نیز عملکرد مؤثری ارائه دهد یا نه.

آموزش مدل با متوازن‌سازی داده‌ها آغاز می‌شود تا تعداد نمونه‌های افسرده و غیرافسرده برابر شود. بررسی‌ها نشان داد که انجام پیش‌پردازش[2] بر روی توییت‌ها اثر معناداری بر عملکرد مدل ندارد بنابراین، متن خام توییت‌ها بدون پیش‌پردازش وارد مدل XLM-RoBERTa می‌شود تا مدل متنی آموزش ببیند. در بخش تصویری، چند مدل مختلف از جمله ResNet18، ViT و CLIP مورد بررسی قرار گرفتند و درنهایت ViT به دلیل دقت بهتر انتخاب شد. پس از آموزش مدل‌های تک‌وجهی، بردارهای ویژگی متن و تصویر استخراج می‌شوند و با استفاده از آن‌ها، رمزگذار T2V آموزش داده می‌شود. پارامتر margin در مرحله آموزش به‌صورت تجربی تعیین شد؛ چند مقدار مختلف دربازه ۰/۱ و ۰/۵ بررسی شد و مقدار ۰/۳ براساس عملکرد روی داده اعتبارسنجی انتخاب شد. میانگین تشابه کسینوسی در این مرحله ٦۱ درصد به‌دست آمد. در ادامه، ادغام داده‌ها انجام می‌شود و نتایج آن در ادامه ارائه خواهد شد.

با توجه به نتایج به‌دست‌آمده، مدل چندوجهی عملکرد بهتری نسبت به مدل‌های تک‌وجهی متن و تصویر ارائه داده است. این مدل توانسته دقت و امتیاز F1 بالاتری نسبت به مدل‌های تک‌وجهی به‌دست آورد. همچنین، در میان مدل‌های تک‌وجهی، مدل متنی عملکرد بهتری نسبت به مدل تصویری داشته است. دقّت و فراخوانی بالاتر این مدل نشان می‌دهد که روش‌های پردازش زبان

[1] Bias

[2] Preprocessing

طبیعی در تحلیل محتوای متنی و شناسایی افسردگی مؤثرند. این نتایج–که در جدول ۲ قابل مشاهده است–حاکی از آن است که مدل‌های چندوجهی با ترکیب متن و تصویر می‌توانند ویژگی‌های مکمل را از هر دو منبع استخراج کنند. در نتیجه، استفاده از چنین رویکردی می‌تواند به بهبود عملکرد مدل و دقت بیشتر در تحلیل احساسات و تشخیص افسردگی در شبکه‌های اجتماعی منجر شود.

جدول (۲): نتایج مدل‌های تک‌وجهی و مدل چندوجهی در تشخیص افسردگی در سطح توییت

مدل	دقت روی‌داده آموزش	دقت روی‌داده تست	معیار F1 روی‌داده تست
متن	۹۶/۷۹	۸۷/۸۷	۸۷/۹۴
تصویر	۸۱/۱۵	۷۳/۳۷	۷۴/۳۲
چند وجهی	۹۵/۳۲	۹۰/۱۷	۹۰/۶۴

۵- نتیجه‌گیری

در این پژوهش، از روشی مشابه مقاله [۲۹] به‌منظور مدیریت وجه گم‌شده در مدل‌های چندوجهی جهت تشخیص افسردگی در شبکه‌های اجتماعی استفاده شد. این روش به‌جای حذف نمونه‌های ناقص، می‌کوشد ویژگی‌های وجه گم‌شده را از اطلاعات موجود پیش‌بینی برآورد کند و از این طریق، از داده‌های ناقص نیز در فرآیند آموزش بهره ببرد. نتایج نشان داد که مدل متنی نسبت به مدل تصویری عملکرد بهتری دارد. دلیل این امر آن است که در داده‌های شبکه‌های اجتماعی، محتوای متنی مستقیماً احساسات و افکار کاربران را منعکس می‌کند و بنابراین سرنخ‌های قوی‌تری برای شناسایی افسردگی فراهم می‌آورد، در حالی که تصاویر اغلب متنوع، غیرساختاریافته یا کم‌ارتباط با وضعیت روانی کاربر هستند. این یافته با مطالعات پیشین [۱۳، ۱۲، ۱۰] نیز هم‌راستا است که تأکید داشته‌اند زبان کاربران افسرده ویژگی‌های متمایزی دارد. همچنین، مدل چندوجهی عملکرد بالاتری نسبت به مدل‌های تک‌وجهی کسب کرد. این برتری ناشی از ترکیب اطلاعات متنی و تصویری و ایجاد هم‌افزایی میان آن‌ها است؛ به‌گونه‌ای که ضعف هر وجه توسط دیگری جبران می‌شود. علاوه بر این، بهره‌گیری از رمزگذار T2V امکان جایگزینی وجه گم‌شده را فراهم آورد و سبب شد که حتی نمونه‌های ناقص نیز در فرآیند آموزش نقش داشته باشند. این موضوع نه‌تنها مانع کاهش حجم داده شد، بلکه باعث افزایش پایداری مدل در شرایط واقعی گردید. به‌طور کلی، یافته‌های این پژوهش نشان می‌دهد که ادغام چندوجهی همراه با مدیریت صحیح وجوه گم‌شده می‌تواند موجب بهبود معنادار عملکرد مدل‌های تشخیص افسردگی شود و زمینه را برای تحلیل دقیق‌تر وضعیت روان‌شناختی کاربران در شبکه‌های اجتماعی فراهم کند. در ادامه، چند مسیر پیشنهادی برای بهبود مدل مطرح شده. همچنین استفاده از مکانیزم توجه متقابل[1] می‌تواند به یادگیری مؤثرتر ارتباط بین ویژگی‌های متنی و تصویری کمک کند. وجه‌های دیگری–شامل زمان ارسال توییت، اطلاعات کاربری (تعداد دنبال‌کنندگان و دنبال‌شوندگان) و میزان تعامل کاربران (تعداد لایک‌ها، بازنشرها و پاسخ‌ها)– نیز دردسترس هستند. این اطلاعات می‌تواند بازتابی از میزان اثرگذاری پست‌ها و نوع تعاملات کاربر باشد و در تحلیل بهتر وضعیت احساسی او نقش مهمی ایفا کند. مدل کنونی هر توییت را به‌طور مستقل به‌عنوان افسرده یا غیرافسرده دسته‌بندی می‌کند؛ امّا هدف اصلی، تشخیص وضعیت روانی کاربر است. در ادامه مسیر، مدل به‌گونه‌ای توسعه خواهد یافت که به‌جای تحلیل تک‌توییت‌ها، بتواند وضعیت روانی کاربران را بر اساس مجموعه‌ای از توییت‌های آن‌ها ارزیابی کند. این بهبودها می‌توانند عملکرد مدل را در تشخیص افسردگی افزایش داده و تحلیلی جامع‌تر از وضعیت روان‌شناختی کاربران در شبکه‌های اجتماعی ارائه دهند.

اقرار

در اینجا لازم می‌دانیم تا سپاسگزاری خود را از حمایت‌های بی‌دریغ اپراتور هوش مصنوعی اهورا) شرکت ایران جی‌پی‌یو(در راستای تأمین سخت‌افزار محاسباتی گرافیکی موردنیاز این تحقیق ذیل تفاهم‌نامه به شماره ۰۳۰۹۱۳/۱۰۰ در جهت تعالی حوزه هوش مصنوعی کشور اعلام کنیم.

[1] Cross-Attention

References

[1] World Health Organization (WHO), "Depressive disorder (depression)," WHO Fact Sheets, Aug. 29, 2025. [Online]. Available:https://www.who.int/news-room/fact-sheets/detail/depression

[2] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: a survey," IEEE Access, vol. 7, pp. 63373–63394, 2019, doi: 10.1109/ACCESS.2019.2916887.

[3] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: a survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 10, pp. 12113–12132, 2023, doi: 10.1109/TPAMI.2023.3275156.

[4] D. Ailyn, "Multimodal data fusion techniques," 2024. [Online]. Available: https://www.researchgate.net/publication/383887675_Multimodal_Data_Fusion_Techniques

[5] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," Information Fusion, vol. 57, pp. 115–129, 2020, doi: 10.1016/j.inffus.2019.12.001.

[6] M. Pawłowski, A. Wróblewska, and S. Sysko-Romańczuk, "Effective techniques for multimodal data fusion: a comparative analysis," Sensors, vol. 23, p. 2381, 2023, doi: 10.3390/s23052381.

[7] D. Lee, S. Park, J. Kang, D. Choi, and J. Han, "Cross-lingual suicidal-oriented word embedding toward suicide prevention," in Findings of the Association for Computational Linguistics: EMNLP 2020, R. Cotterell, S. Eger, and S. Wiseman, Eds., Online, Nov. 2020, pp. 2208-2217, doi: 10.18653/v1/2020.findings-emnlp.200.

[8] B. G. Bokolo and Q. Liu, "Deep learning-based depression detection from social media: comparative evaluation of ML and transformer techniques," *Electronics*, vol. 12, no. 21, p. 4396, 2023, doi: 10.3390/electronics12214396.

[9] 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.

[10] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in Proc. Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Baltimore, MD, USA, Jun. 27, 2014, pp. 51–60, doi: 10.3115/v1/W14-3207.

[11] M. Deshpande and V. Rao, "Depression detection using emotion artificial intelligence," in *Proc. 19th IEEE Int. Conf. Intelligent Sustainable Systems*, Palladam, Tirupur, India, Dec. 2017, pp. 858-862, doi: 10.1109/ISS1.2017.8389299.

[12] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proc. Int. AAAI Conf. Web and Social Media, Limassol, Cyprus, Jun. 5–8, 2013, vol. 7, pp. 128–137, doi: 10.1609/icwsm.v7i1.14432.

[13] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Classification of mental illnesses on social media using RoBERTa," in Proc. 12th Int. Workshop on Health Text Mining and Information Analysis, E. Holderness et al., Eds., Online, 2021, pp. 59–68. [Online]. Available: https://aclanthology.org/2021.louhi-1.7/

[14] S. Yang, L. Cui, L. Wang, T. Wang, and J. You, "Enhancing multimodal depression diagnosis through representation learning and knowledge transfer," *Heliyon*, vol. 10, no. 4, p. e25959, Feb. 2024, doi: 10.1016/j.heliyon.2024.e25959.

[15] T. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: a survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[16] S. Yang, L. Cui, L. Wang, T. Wang, and J. You, "Cross-modal contrastive learning for multimodal sentiment recognition," *Applied Intelligence*, vol. 54, pp. 4260–4276, 2024, doi: 10.1007/s10489-024-05355-8.

[17] M. Fang, S. Peng, Y. Liang, C.-C. Hung, and S. Liu, "A multimodal fusion model with multi-level attention mechanism for depression detection," *Biomedical Signal Processing and Control*, vol. 82, p. 104561, Apr. 2023, doi: 10.1016/j.bspc.2022.104561.

[18] Y. Wang, Z. Wang, C. Li, Y. Zhang, and H. Wang, "Online social network individual depression detection using a multitask heterogeneous modality fusion approach," *Information Sciences*, vol. 609, pp. 727–749, 2022, doi: 10.1016/j.ins.2022.07.109.

[19] R. Wu, H. Wang, H.-T. Chen, and G. Carneiro, "Deep multimodal learning with missing modality: a survey," *arXiv preprint* arXiv:2409.07825, 2024.

[20] L. Cai, Z. Wang, H. Gao, D. Shen, and S. Ji, "Deep adversarial learning for multi-modality missing data completion," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 1158–1166, 2018, doi: 10.1145/3219819.3219963.

[21] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "SMIL: Multimodal learning with

severely missing modality," in *Proc. 35th AAAI Conf. Artificial Intelligence (AAAI '21)*, vol. 35, no. 3, pp. 2302–2310, 2021, doi: 10.1609/AAAI.V35I3.16330.

[22] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multi-modal learning with missing modality via shared-specific feature modelling," in Proc. 2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 15878–15887, 2023, doi: 10.1109/CVPR56347.2023.

[23] M. K. Reza, A. Prater-Bennette, and M. S. Asif, "Robust multimodal learning with missing modalities via parameter-efficient adaptation," *arXiv preprint* arXiv:2310.03986, 2024.

[24] Q. Wang, L. Zhan, P. Thompson, and J. Zhou, "Multimodal learning with incomplete modalities by knowledge distillation," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2020, pp. 1828-1838, doi: 10.1145/3394486.3403234.

[25] X. Chen, S. Kornblith, M. Noroozi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Machine Learning (ICML)*, 2020, pp. 1597–1607, doi: 10.48550/arXiv.2002.05709.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 139, pp. 8748–8763, 2021, doi: 10.48550/arXiv.2103.00020.

[27] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *arXiv preprint* arXiv:2010.05113, 2021.

[28] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: a multimodal dictionary learning solution," in *Proc. 26th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 3838–3844, doi: 10.24963/ijcai.2017/535.

[29] C. Zhang, F. Nian, and J. Lee, "Toward robust multimodal learning using multimodal foundational models," *arXiv preprint* arXiv:2401.13697, 2024, doi: 10.48550/arXiv.2401.13697.

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 3, No. 1, Spring & Summer 2025, pp. 179 - 191

Contrastive-Learning–Based Handling of Missing Modalities in Multimodal Data Fusion for Depression Detection on Social Networks

Hamed Marvi [1], Abolfazl Nadi [2*], Mohammad Mehdi Keikha [3]

[1] MS.c Student, Department of Computer Science, School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran

[2]Assistant Professor, Department of Computer Science, School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran

[3] Assistant Professor, Department of Computer Science, University of Sistan and Baluchestan, Zahedan, Iran

Article Information

Original Research Paper

Received:
2025 August 25

Accepted:
2025 October 29

Keywords:
Multimodal Models, Missing Modality, Multimodal Data Fusion, Contrastive Learning, Social Media

Corresponding Author*:
a.nadi@ut.ac.ir

Abstract

The analysis of social network data plays a fundamental role in uncovering users' behavioral patterns. Multimodal models that combine textual, visual, and other information sources are effective tools for such analyses. However, a major challenge in these models is the absence of certain modalities in parts of the dataset; for instance, a user may post only text without sharing any images. This issue prevents multimodal models from fully exploiting all available information.In this paper, a method is proposed for leveraging incomplete data within multimodal models. First, unimodal models are trained independently to process each modality. Then, a contrastive learning–based encoder is designed and trained to estimate the feature vector of the missing modality using the features of the available modalities.Finally, textual and visual data—either real or reconstructed—are fused in a multimodal model to analyze user behavior. Experimental results on the MDDL dataset, based on accuracy and F1-score metrics, demonstrate that the proposed multimodal model achieves superior performance, reaching an accuracy of 90.17% and an F1-score of 90.64%, outperforming unimodal text-based (87.87% accuracy) and image-based (73.37% accuracy) models. These findings confirm that the proposed model effectively utilizes available information without discarding incomplete data.

: 10.22034/ABMIR.2025.23578.1160

E-ISSN: 2821-2037