علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۲۳، شماره ۱، ۱۴۰۴
صفحات ۲۲-۳۰
مقاله پژوهشی

ارائه یک رویکرد مبتنی بر نمایش ویژگی‌های متنی برای تحلیل احساس متن فارسی

میثم روستائی[1*]

[*] نویسنده مسئول، دریافت: ۱۴۰۱/۱۱/۲۵ ، بازنگری: ۱۴۰۲/۲/۰۱ ، پذیرش: ۱۴۰۲/۰۲/۱۰

[1] استادیار دانشکده مهندسی و فناوری، دانشگاه مازندران، بابلسر، ایران

چکیده

با رشد گسترده شبکه‌های اجتماعی و سیستم‌های پرسش و پاسخ انجمنی، کاربران قادرند به راحتی نظرات و عقاید خود را در مورد اخبار، محصولات و رویدادهای اجتماعی بیان کنند. در سال‌های اخیر، میزان داده‌های غیرساخت‌یافته نظیر متن بسیار افزایش یافته و نیاز به ابزارهای خودکار برای پردازش آن‌ها بیش از پیش حس می‌شود. تشخیص احساس متن به فرآیند درک و طبقه‌بندی حالت عاطفی نویسنده از روی متن نوشته شده نظیر نظرات داده شده در فروشگاه‌های اینترنتی و شبکه‌های اجتماعی اشاره دارد و در تحلیل بازار و ارزیابی کیفیت خدمات نقش اساسی ایفا می‌کند. در این پژوهش، حدود ۱۲ هزار جمله فارسی که شامل نظرات کاربران در مورد محصولات دیجیتال است مورد بررسی قرار می‌گیرد. برای این منظور، با تولید بردارهای نمایش کلمات و اطلاعات نحوی آن‌ها با استفاده از تکنیک skip-gram و فراهم کردن شکل‌های متفاوتی از نمایش ویژگی‌ها، اطلاعات کامل‌تری به معماری پیشنهادی که بر اساس شبکه‌های عصبی کانولوشنی است تزریق می‌شود. با استفاده از این رویکرد، ویژگی‌های عملکردی کلمات بهتر ثبت می‌شود. ارزیابی‌های انجام شده نشان می‌دهند نتایج مدل پیشنهادی بر اساس معیار F_1 بر روی پیکره استاندارد به ترتیب برای حالت‌های دو کلاسه و چند کلاسه ۹۲.۱۳ و ۷۰.۵۱ است که حاکی از کارایی بالاتر رویکرد پیشنهادی در مقایسه با مدل‌های سنتی یادگیری ماشین و مدل‌های بروز در تعیین احساس متن فارسی می‌باشد.

کلمات کلیدی: پردازش زبان طبیعی، طبقه‌بندی متن، شبکه‌های عصبی کانولوشنی، یادگیری نمایش ویژگی‌ها، تحلیل احساس، متن فارسی.

۱- مقدمه

تجزیه و تحلیل احساس که به آن عقیده کاوی نیز گفته می‌شود یکی از مسائل پردازش زبان طبیعی است که لحن احساسی پشت متن را مشخص می‌کند. این یک روش محبوب برای سازمان‌ها جهت تعیین و دسته‌بندی نظرات درباره یک محصول، خدمات یا ایده است. به عنوان مثال، یک شرکت تولید محصولات نرم افزاری می‌تواند با استفاده از ابزار تحلیل احساس، یک تحلیل مختصر و مفید از نظرات کاربران در مورد بروزرسانی جدید نرم افزار خود دریافت کند. همچنین یک مدل تحلیل احساس متن می‌تواند به کاربران نیز در یافتن محصول مناسب برای خرید کمک کند. به جای اینکه کاربر مجبور باشد نظرات دیگران در مورد یک محصول خاص را یک به یک بررسی کرده و در نهایت تصمیم بگیرد که آیا انتخاب درستی هست یا خیر، متد تحلیل احساس با بررسی و تحلیل نظرات کاربران، اعتقاد دیگران نسبت به آن محصول را مشخص می‌کند [۱]. بنابراین نظرات افراد بسیاری در خرید تاثیر گذار خواهد بود [۲]. جدول ۱ نمونه‌ای از تعیین احساس کاربران نسبت به محصولات یک فروشگاه اینترنتی را نشان می‌دهد. بنابراین، ابزار تحلیل احساس در بسیاری از کاربردها قابل نگاشت بوده و منجر به پیشرفت کسب و کار و بهبود عملکرد می‌گردد.

با توجه به تغییرات و پیچیدگی بسیار زیاد زبان طبیعی، توسعه یک مدل خودکار برای تحلیل احساس، یک مسئله غیربدیهی و البته چالشی جذاب و دشوار می‌باشد [۳]. در ساده‌ترین حالت، یک مدل تحلیل احساس، به متن داده شده دو قطب مثبت یا منفی را نسبت می‌دهد؛ به این معنا که کاربران نظر مثبتی نسبت به آن دارند یا خیر. یا در سطوح بالاتر، تعداد کلاس‌های قابل استنتاج از متن بسیار بیشتر می‌باشد.

جدول ۱ نمونه‌ای از مسئله تحلیل احساس متن فارسی- متن از [۴] DeepSentiPers

جمله ورودی	احساس نویسنده
متاسفانه این گوشی مکالمه تصویری مبتنی بر شبکه نسل سوم رایتل رو پشتیبانی نمیکنه.	منفی
رزولوشن این دوربین برای یک گوشی در این سطح خیلی خوب است.	مثبت
کیفیت این دو مدل در مقایسه با هم واقعا متفاوته.	خنثی

زبان فارسی یک زبان هندواروپایی است که بیشتر در کشورهای ایران، افغانستان و تاجیکستان استفاده می‌شود. پیچیدگی زبان فارسی نسبت به زبان‌های دیگر از جمله انگلیسی بسیار بیشتر می‌باشد. به عنوان مثال، وجود فاصله و رعایت نشدن دقیق نیم فاصله‌ها در نوشتار و وجود کلمات غیررسمی در نوشتار، پردازش آن را دشوارتر کرده و در نتیجه ابزارهای پردازش زبان کمتری برای آن ارائه شده است. اگرچه در سال‌های اخیر ابزارهایی برای پردازش زبان فارسی نظیر هضم ارائه شده است که فرآیند پیش‌پردازش بر روی متن را امکان‌پذیر کرده و پژوهشگران قادر شدند بر روی سایر وظیفه‌های پردازش زبان طبیعی نظیر تحلیل احساس نیز تمرکز کنند.

به طور کلی تحلیل احساس می‌تواند در سه سطح متفاوت انجام شود [۵]: ۱) سطح سند ۲) سطح جمله ۳) سطح جنبه (ویژگی). در تحلیل احساس در سطح سند، اطلاعات موجود در تمام جملات برای تعیین احساس متن اهمیت دارد؛ به عبارت دیگر کل سند مورد بررسی و تحلیل قرار می‌گیرد تا کلاس صحیح آن مشخص شود. در این سطح انتظار می‌رود که نظرات بیان شده در مورد یک موجودیت یا رویداد باشد. تحلیل احساس در سطح جمله، همانطور که از نام آن پیداست با توجه به کلمات موجود در جمله انجام می‌شود. با توجه به کوتاه بودن طول جملات، اطلاعات زمینه در این حالت کمتر بوده و در نتیجه تعیین احساس متن دشوار می‌باشد. تحلیل احساس در سطح جنبه که تحلیل در سطح ویژگی نیز نامیده می‌شود، متفاوت با دو مورد قبلی می‌باشد. در این سطح به دنبال تعیین احساس نویسنده نسبت به یک رویداد یا موجودیت خاص نبوده بلکه بر روی جنبه‌های مختلف متن تمرکز داشته و برآیند جنبه‌های مختلف را اعلام می‌کند. در این پژوهش بر روی تحلیل احساس در سطح جمله تمرکز داشته و مدل پیشنهادی را بر روی داده‌های در سطح جمله مورد ارزیابی قرار خواهیم داد.

کلمات به عنوان ویژگی‌های اصلی متن می‌توانند برای نمایش آن استفاده شوند. مدل‌های قدیمی از رویکرد کیف کلمات برای نمایش متون استفاده می‌کردند که تنها تطابق در سطح کلمه را در نظر می‌گرفتند [۶]. این تطابق البته تنها به صورت تطابق دقیق میان کلمات بوده و اطلاعات معنایی عمیق و وابستگی کلمات را کنار می‌گذاشتند. ایده استفاده از رویکردهای مبتنی بر یادگیری عمیق برای نمایش کلمات به صورت بردارها بهتر توانست اطلاعات معنایی را دریافت کند به طوری که فاصله بردار کلمات هم معنا در فضای چند بعدی بسیار کم می‌باشد.

تکنیک‌های متعددی برای تولید نمایش پیوسته کلمات ارائه شده است که از جمله آنها می‌توان به skip-gram [۷] اشاره کرد که با در اختیار داشتن کلمه مرکزی، مدل به پیش‌بینی کلمات همسایه آن در یک پنجره با طول مشخص می‌پردازد. این رویکردها اگرچه منجر به بردارهایی می‌شوند که امکان تطابق تقریبی را برای مسائل متعددی فراهم می‌کنند اما آنها متون را به صورت ساختار متوالی در نظر می‌گیرند. با توجه به ابهام و وابستگی‌های طولانی که میان کلمات موجود در جملات وجود دارد و از آنجا که این وابستگی‌ها می‌تواند خارج از محدوده پنجره ثابت کوچک باشد احتمال خطا به دلیل ظهور تصادفی وجود دارد. همچنین با توجه به اینکه اطلاعات نحوی شامل نوع وابستگی نیز می‌شود، بردارهای تولید شده به این روش کاربردی‌تر هستند.

رویکردهای قبلی ارائه شده برای مسئله تعیین احساس متن از بردارهای تعبیه شده از پیش آموزش دیده و یا بردارهایی که با رویکرد مبتنی بر پنجره ثابت تولید می‌شوند برای نمایش کلمات و در نتیجه بردارهای ویژگی جملات استفاده کردند. در این پژوهش ما نشان خواهیم داد که استفاده از رویکرد بهتری برای نمایش ویژگی‌ها به مدل‌های با کارایی بالاتر منجر می‌شود. پژوهش جاری از اطلاعات نحوی برای یادگیری بردارهای کلمات و وابستگی‌های آنها استفاده می‌کند. در مدل پیشنهادی، اطلاعات نحوی استخراج شده از جملات به عنوان اطلاعات زمینه کلمات در فرآیند یادگیری مدل به جای پنجره ثابت استفاده می‌شود. در نتیجه مدل پیشنهادی در سطح بالاتری وابستگی‌های طولانی را در نظر گرفته و موجب کاهش مشکل ابهام در فرآیند یادگیری می‌شود.

ما همچنین مدل را به صورت محلی بر روی داده‌های آموزشی مسئله آموزش می‌دهیم که منجر به بردارهای خاص وظیفه مورد بررسی می‌شود که عملکرد مدل را بهبود می‌دهد. علاوه بر آن، شکل‌های متفاوت ترکیب بردارهای کلمات و وابستگی‌ها که منجر به نمایش‌های متفاوتی برای هر یک از کلمات جمله می‌شود بررسی شده و عملکرد آنها مورد مقایسه قرار می‌گیرد. در این کار ما مسئله تعیین احساس متن را به عنوان یک مسئله طبقه‌بندی در نظر گرفته (با این فرض که هر اطلاعات زمینه ، نشانه‌ای از احساس را در خود دارد) و تلاش کردیم تا مسئله را با شبکه عصبی کانولوشنال که یک معماری مبتنی بر تکنیک یادگیری عمیق است حل کنیم.

ادامه مقاله به صورت زیر سازماندهی شده است. بخش ۲ به پیشینه پژوهش در زمینه تحلیل احساس می‌پردازد. در این بخش همچنین پژوهش‌های انجام شده در زمینه تحلیل احساس متن فارسی نیز مورد بررسی قرار می‌گیرد. روش پژوهش که شامل یادگیری تعبیه کلمات، تعبیه اطلاعات نحوی، نمایش ویژگی‌ها و مدل طبقه‌بند است در بخش ۳ به طور کامل بیان می‌شود. یافته‌های پژوهش در بخش ۴ ارائه می‌شود. در این بخش ابتدا اطلاعات و آمارها در مورد دیتاست مورد استفاده و معیارهای ارزیابی معرفی شده و سپس نتایج آزمایشات برای پاسخ به پرسش‌های پژوهش ارائه می‌شود. در نهایت، بخش ۵ به نتیجه‌گیری و برخی جهت‌گیری‌ها برای کارهای آینده اختصاص دارد.

۲- کارهای مرتبط

با توجه به اهمیت و کاربرد بسیار تحلیل احساس متون، افراد زیادی بر روی این مسئله متمرکز شدند. در جدول ۲ دسته‌بندی از کارهای انجام شده بالقوه و بروز در زمینه تحلیل احساس متن ارائه شده است. این دسته‌بندی از چند جنبه مختلف می‌باشد. ابتدا پژوهش‌های انجام شده بر روی متون انگلیسی که بیشترین تعداد کارهای انجام شده در این حوزه است نشان داده شده است. سپس کارهای انجام شده برای زبان عربی که از نظر الفبا به زبان فارسی بسیار نزدیک است مشخص شده است. در ادامه برخی از بهترین کارهای انجام شده برای زبان فارسی قابل مشاهده است. ما همچنین برخی از پژوهش‌هایی که از خروجی سیستم تحلیل احساس متن استفاده می‌کنند را نیز در این جدول نشان داده‌ایم. اطلاعات کامل‌تر در مورد هر یک از این پژوهش‌ها در ادامه بیان شده است.

جدول ۲ دسته‌بندی رویکردهای اخیر ارائه شده برای مسئله تحلیل احساس متون

کاربرد	منابع
تحلیل احساس متن انگلیسی	[۸]، [۱۰]، [۹]، [۳] ، [۳۳]، [۱۱] و [۱۲]
تحلیل احساس متن عربی	[۱۳]، [۵] و [۱۴]
تحلیل احساس متن فارسی	[۱۵]، [۲]، [۲۱]، [۱۶]، [۲۰]، [۴]، [۱۷] و [۱۹]
استفاده از ماژول تحلیل احساس	[۲۲] و [۱]

کارهای زیادی بر روی مسئله تحلیل احساس متون انگلیسی انجام شده است. در پژوهش [۸] از رویکرد تحت نظارت برای تعیین احساس پیام‌های ارسال شده در توییتر استفاده کردند. در [۹] مدلی را بر اساس نمایش‌های مختلف برداری مبتنی

بر Word2Vec، FastText و Glove برای ۱۵۷ زبان مختلف ارائه دادند. این مدل بر روی داده‌های ویکی‌پدیا آموزش دیدند. اخیرا در پژوهش [۳] کارهای انجام شده در زمینه تحلیل احساس را مورد بررسی و مطالعه قرار دادند. آنها کارهای انجام شده را از نظر رویکرد مورد استفاده در ده دسته مختلف گروه‌بندی نموده و کارایی آنها را مقایسه کردند. در تحقیق [۱۰] رویکری بر اساس معماری یادگیری عمیق برای مسئله تحلیل احساس داده‌های توییتر ارائه دادند. متد آنها از ترکیب مدل زبان و الگوریتم ماشین بردار پشتیبان تشکیل شده است. عملکرد مدل بر روی دیتاست Twitter US Airlines نسبت به کارهای مورد مقایسه بهتر بوده است. در [۳۳] مسئله تحلیل احساس را به صورت مسئله بهینه سازی چند هدفه در نظر گرفته و از یک رویکرد تطبیق یافته مبتنی بر شبکه بهینه ساز Grey Wolf برای حل آن استفاده کردند. نتایج آنها بر روی سه دیتاست مختلف قابل قبول بوده است. در [۱۱] تحلیل احساس را به صورت مسئله دو کلاس در نظر گرفته و شبکه‌های عصبی بازگشتی را برای حل آن به کار گرفتند. رویکرد آنها، مکانیزم توجه (برای تقویت شبکه) را با یک سیستم محلی سازی دوطرفه ترکیب می‌کند (در هر دو سطح سلول‌های حافظه و سطح شبکه).

استفاده از رویکردهای غیرنظارتی نیز برای حل مسئله تحلیل احساس مورد بررسی قرار گرفت. در پژوهش [۱۲] از یک رویکرد غیرنظارتی مبتنی بر خوشه‌بندی مفهومی و سلسله مراتبی برای حل مسئله تحلیل احساس توییتر استفاده کردند. در این مدل، الگوریتم‌های رایج خوشه بندی پیوند تکی، پیوند کامل و پیوند میانگین به صورت خطی با هم ترکیب شده و کاربرد نمایش‌های متفاوت بولین و TF-IDF مورد بررسی قرار گرفتند. بر اساس ارزیابی انجام شده، عملکرد مدل غیرنظارتی نزدیک به مدل‌های نظارتی است.

زبان عربی نیز یکی از زبان‌هایی است که مسئله تحلیل احساس در آن بسیار مورد توجه قرار گرفت. از آنجا که این زبان نسبت به زبان انگلیسی به زبان فارسی نزدیک‌تر است برخی از کارهای بروز در زمینه تحلیل احساس بر روی این زبان را نیز مورد مطالعه قرار می‌دهیم. در [۱۳] یک مدل مبتنی بر BERT برای درک متن عربی ارائه دادند که یکی از وظایف مورد ارزیابی مدل در کار آن‌ها، تحلیل احساس بود. نتایج کار آن‌ها نشان داد که استفاده از مدل BERT منجر به کارایی بالاتری نسبت به Bert چندزبانه می‌شود.

در مطالعه [۱۴] از قوانین مبتنی بر وابستگی و یادگیری عمیق برای تحلیل احساس متن عربی استفاده کردند. قوانین مبتنی بر وابستگی با استفاده از الگوهای زبانی برای نگاشت معنای کلمات به مفاهیم ایجاد می‌شوند. با بررسی کلمات وابسته در یک جمله، احساس کلی آشکار می‌شود. در این مدل، در مرحله اول طبقه‌بندی احساس، از قواعد گرامر وابستگی استفاده می‌شود. اگر قوانین در طبقه‌بندی احساس ناموفق باشند، سپس شبکه‌های عصبی عمیق اعمال می‌شود. پژوهشگران در [۵] مطالعه ای از کارهای انجام شده بر روی تحلیل احساس متن عربی انجام داده و سپس مدل‌های زبانی مبتنی بر ترنسفورمر را برای این وظیفه مورد بررسی قرار داد. نتایج کار آنها نشان داد که این مدل‌های زبانی نسبت به سایر رویکردها در زبان عربی عملکرد بهتری دارد.

محققان در [۱۵] با معرفی SPerSent و CNRC که به ترتیب یک مجموعه داده در سطح جمله برای تحلیل احساس متن فارسی و یک مجموعه واژگان فارسی است به مشکل کمبود منابع در تحلیل احساس متن فارسی پرداختند. SPerSent شامل ۱۵۰۰۰ جمله است که قطبیت آنها با استفاده از متد مبتنی بر واژگان و به شیوه رأی اکثریت تعیین شده است. در [۱۶] از یک رویکرد مبتنی با ناظر برای ایجاد واژه نامه تحلیل احساس فارسی استفاده کردند. مدل پیشنهادی آنها از ویژگی‌های زبانشناسی و اطلاعات آماری به منظور تعیین جهت‌گیری احساس واژه‌گان بهره می‌برد. در [۱۷] محققان از تحلیل احساس در سطح ویژگی برای حل مسئله خلاصه سازی بر اساس ویژگی استفاده کردند. مدل آنها به صورت بدون ناظر و با استفاده از بردارهای تعبیه کلمات عمل می‌کند.

همچنین در [۱۸]، رویکرد مبتنی بر شبکه‌های عصبی خودرمزگذار عمیق را برای تحلیل احساس متن فارسی ارائه شد. نتایج ارزیابی مدل بر روی مجموعه‌ای از نقد فیلم‌ها حاکی از عملکرد بهتر مدل نسبت به شبکه‌های عصبی چندلایه کم عمق است. آنها در ادامه پژوهش خود در [۱۹]، یک مدل ارزیابی چندوجهی برای تعیین احساس ویدئوها زبان فارسی ارائه دادند. مدل پیشنهادی آنها از نشانه‌های صوتی، بصری و متنی برای تعیین احساس استفاده می‌کند. آنها همچنین پیکره‌ای شامل ۸۰۰ نمونه را برای ارزیابی تحلیل احساس ویدیوها زبان فارسی ارائه دادند.

در [۲] با جمع آوری حدود سه میلیون نظر از وب سایت دیجی کالا و آموزش مدل FastText بر روی آن، یک نمایش تعبیه شده ایجاد و از الگوریتم‌های یادگیری ماشین سنتی و یادگیری عمیق برای طبقه‌بندی استفاده شد. محققان در [۲۰] از ترکیب یادگیری مکاتبات ساختاری و یادگیری عمیق برای تحلیل احساس استفاده کردند. همچنین در [۴] از معماری شبکه عصبی بازگشتی حافظه کوتاه مدت طولانی مدت دو جهته ترکیب شده با FastText برای مسئله تحلیل احساس متن فارسی استفاده کردند. مدل پیشنهادی آن بر روی پیکره DeepSentiPers که یک نسخه متعادل شده از پیکره SentiPers [۲۱] است ارزیابی شده است.

استفاده از متدهای تعیین احساس در کاربردهای دیگر نیز مورد بررسی قرار گرفت. محققان در پژوهش [۲۲] به بررسی تاثیر قطبیت پیام ارسال شده در توییتر بر روی بازار سهام پرداختند. برای این فرآیند تحلیل احساسی بر روی پیام‌ها انجام شد و وابستگی بسیار زیاد بین این متغیرها نشان داده شد. در [۱] بر روی مسئله مشابهی کار کرده و از LSTM برای تعیین قطبیت پیام نسبت به قیمت بازار سهام استفاده کردند. برای بهبود عملکرد مدل در پیش‌بینی از بهینه ساز Adam استفاده شده است. همچنین جهت تنظیم پارامترهای خروجی مدل، بهینه سازی مبتنی بر یادگیری مورد استقاده قرار گرفته است. نتایج این پژوهش نیز نشان می‌دهد که تحلیل احساس کمک بسیاری در پیش‌بینی بازار سهام دارد.

رویکردهای قبلی تحلیل احساس متن فارسی اکثرا یا از روش‌های سنتی نظیر کیف کلمات و یا از روش‌های یادگیری عمیق مبتنی بر بردارهایی که به روش پنجره با طول ثابت تولید شده بودند برای نمایش ویژگی‌ها استفاده کردند. اگر چه این نمایش برخی اطلاعات نحوی را نیز ارائه می‌دهد اما با توجه به پیچیدگی زبان طبیعی و وابستگی که میان کلمات وجود دارد، استفاده از پنجره با طول ثابت قادر به پوشش وابستگی‌های طولانی میان کلمات نیست. برای این منظور مدل پیشنهادی از یک رویکرد ترکیبی مبتنی بر کلمات و همچنین وابستگی‌های نحوی آنها برای نمایش ویژگی‌ها استفاده می‌کند. ارزیابی مدل پیشنهادی نشان می‌دهد رویکرد ارائه شده با استفاده از این شیوه نمایش ویژگی‌ها و معماری شبکه عصبی کانولوشن، کارایی به مراتب بالاتری از رویکردهای بروز دارد.

۳- رویکرد پیشنهادی

با در اختیار داشتن مجموعه‌ای از نمونه‌های برچسب زده شده، فرآیند رویکرد پیشنهادی به صورت زیر می‌باشد:

از آنجا که در مدل پیشنهادی هدف در نظر گرفتن هر دو شباهت معنایی و نحوی است، تلاش می‌شود تا کلمات موجود در جمله به عنوان ویژگی‌های اصلی آن، به صورت بردارهایی با مقادیر پیوسته مبتنی بر کلمه و وابستگی نمایش داده شوند. برای این منظور پیش پردازش بر روی داده‌ها به کمک کتابخانه هضم انجام می‌شود و روابط وابستگی هر یک از نمونه‌ها با پیاده‌سازی توابعی که از خروجی توابع موجود

در کتابخانه هضم استفاده می‌کند استخراج می‌شود. از داده‌های آموزش (بدون توجه به برچسب آنها) برای آموزش نسخه توسعه یافته تعبیه کلمات استفاده می‌شود تا یک مدل وابسته به مسئله برای تولید بردارها بدست آید. سپس، هر نمونه به صورت ماتریسی که ردیف‌های آن مربوط به بردارهای کلمات تشکیل دهنده آن است نشان داده می‌شود. از ماتریس جملات بدست آمده به عنوان مجموعه ورودی برای آموزش مدل طبقه‌بند شبکه عصبی کانولوشن استفاده می‌شود. همچنین داده‌های ارزیابی برای تخمین تعداد تکرارهای صحیح برای آموزش طبقه‌بند مورد استفاده قرار می‌گیرند. با استفاده از وزن‌های لایه آخر مدل که یک لایه کاملا متصل است که همراه با یک لایه سیگموئید می‌باشد، قطبیت نمونه‌ها تعیین می‌شود. از آنجا که فرآیند آموزش مدل نمایش بردارهای کلمات و وابستگی و همچنین فرآیند آموزش مدل طبقه‌بند شبکه عصبی کانولوشن به صورت افلاین انجام می‌شود، مدل پیشنهادی از پیچیدگی زمانی پایینی در فاز آنلاین برخوردار است. در ادامه راه‌حل پیشنهادی با جزئیات بیشتری بیان می‌شوند.

۳-۱- یادگیری تعبیه کلمات

تعداد کلمات زیادی در زبان طبیعی وجود دارد و این کلمات لزوماً نامرتبط نیستند. به عنوان مثال هم معناهای بسیاری در زبان وجود دارند که می‌توانند جایگزین یکدیگر شوند. از سوی دیگر، می‌دانیم که در نظر گرفتن کلمات به صورت رشته‌های کاراکتری به سیستم کامپیوتری امکان تعیین شباهت میان کلمات را در محاسبات نخواهد داد. در نتیجه تحقیقات به سمت نمایش کلمات به صورت بردارهای چند بعدی که قابل خواندن توسط ماشین هستند سوغ پیدا کرده است. ساده‌ترین حالت نمایش کلمات به صورت بردارها، نمایشOne-Hot می‌باشد. در این نمایش، هر کلمه به یک بردار با طول $|V|$ نگاشت می‌شود که $|V|$ به تعداد کلمات موجود در لغت‌نامه اشاره دارد. در این بردار، تمام المنت‌ها به جز اندیس کلمه مورد نظر که یک است، صفر می‌باشد.

علاوه بر اینکه ابعاد بردارهای One-Hot بسیار بزرگ می‌باشد که محاسبات را با چالش مواجه می‌کند. این نمایش کلمات را نیز مستقل فرض می‌کند و در تعیین شباهت آنها ناتوان می‌باشد. در ادامه تکنیک تعبیه کلمات Word2Vec که منجر به تولید بردارهایی با ابعاد کوچک‌تر می‌شود تشریح می‌گردد. هر بعد این بردار می‌تواند جنبه خاصی از معنا را پوشش دهد و در نتیجه ترکیب این ابعاد می‌تواند به کدگذاری معناها کمک کند. مطالعات قبلی بر روی مسئله‌های مختلف نشان داده‌اند که یادگیری بردارها بر روی داده‌های خاص مسئله منتج به نتایج بهتری می‌گردد [۲۳-۲۴]. از این رو در این پژوهش، یادگیری نمایش برداری بر روی داده‌های آموزش مسئله انجام می‌شود.

نمایش فضای برداری کلمات این امکان را فراهم می‌کند که کلمات به صورت مستقیم مقایسه شده و شباهت معنایی آنها تعیین شود. یکی از مشخصه‌های اصلی این نمایش این است که کلمات مشابه در فضای نمایش بسیار به هم نزدیک هستند [۹]. این نمایش بر اساس این تئوری می‌باشد که معنای کلمات بر اساس زمینه‌ای که در آن ظاهر شده است قابل تعیین می‌باشد [۲۵]. با استفاده از نمایش برداری، شباهت معنایی کلمات با محاسبه شباهت بردار مرتبط با آنها (نظیر محاسبه cosine بردارها) محاسبه می‌شود. یکی از تکنیک‌های خوش نام برای یادگیری بردارها Skip-gram [۷] است که با عنوان Word2Vec نیز صدا زده می‌شود (که البته Skip-gram یکی از دو الگوریتم این ابزار می‌باشد). با توجه به کارایی بالای بردارهای تولید شده توسط این تکنیک، این مدل به عنوان یک روش رایج برای تولید نمایش برداری کلمات تبدیل شده است.

با در اختیار داشتن کلمه ورودی (کلمه مرکزی)، متد Skip-gram برای پیش‌بینی کلمات زمینه در یک پنجره با طول n استفاده می‌شود (n کلمه در سمت راست و n کلمه در سمت چپ کلمه هدف). به عنوان مثال با در نظر گرفتن جمله "چینش دکمه‌ها و کنترل‌های دوربین واقعا عالیست" در پیکره، کلمه ورودی "دوربین" و با پنجره به طول دو، هدف مدل پیش‌بینی کلمات زمینه آن یعنی "دکمه‌ها"، "و"، "واقعا" و "عالیست" است. این رویکرد یک شبکه عصبی است که شامل یک لایه مخفی می‌باشد. ورودی به شبکه یک بردار one-hot از کلمه ورودی است و خروجی آن احتمال ظهور به عنوان کلمه زمینه کلمه ورودی می‌باشد. با در اختیار داشتن جفت $(w, c) \in D$ که در آن w به کلمه ورودی، c به کلمه زمینه و D به مجموعه جفت کلمات و زمینه‌ها در پیکره اشاره دارد، هدف حداکثر کردن احتمال $p(c|w;x)$ است (x پارامترهای مدل است):

$$\underset{x}{\operatorname{argmax}} \prod_{(w,\,c)\in D} p(c|w;x) \qquad (1)$$

با در دست داشتن v_w و v_c به عنوان نمایش برداری w و c، و C به عنوان مجموعه تمام زمینه‌ها (تمام کلمات پیکره)، احتمال $p(c|w;x)$ توسط تابع بیشینه هموار به صورت زیر تعریف می‌شود:

$$p(c|w;x) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}} \qquad (2)$$

و در نتیجه تابع هدف کلی به صورت زیر می‌باشد:

$$\underset{x}{argmax} \sum_{(w,c)\in D} \log p(c|w) \qquad (3)$$

یکی از مشکلات عمده در استفاده از تابع بیشینه هموار، بار محاسباتی بالای آن می‌باشد که با افزایش تعداد کلمات لغت نامه، پیچیدگی نیز افزایش زیادی می‌یابد. برای کاهش بار محاسباتی از تکنیک نمونه برداری منفی استفاده می‌شود که در آن به جای محاسبه احتمال توزیع تمام کلمات، متد تلاش می‌کند تا احتمال جفت نمونه استخراج شده از پیکره $p(D = 1|w, c)$ و احتمال جفت نمونه تولید شده به صورت تصادفی که عضو پیکره نیست $p(D = 0|w, c)$ را حداکثر کند (با مدل کردن $p(D = 1|w, c)$ با استفاده از تابع سیگموئید). و در نتیجه تابع هدف به صورت زیر تعریف می‌شود:

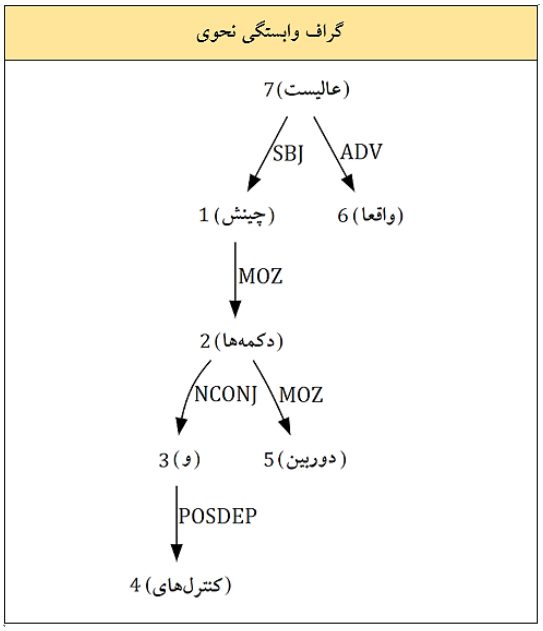
$$\underset{x}{\operatorname{argmax}} \sum_{(w,c)\in D} \log \frac{1}{1 + e^{-v_w \cdot v_c}} + \sum_{(w,c)\in \widetilde{D}} \log \frac{1}{1 + e^{v_w \cdot v_c}} \qquad (4)$$

که در آن جفت‌های $(w, c_k) \in \widetilde{D}$ با انتخاب تصادفی c_k از مجموعه D بدست می‌آید.

۳-۲- یادگیری تعبیه اطلاعات نحوی

اگرچه بردار کلمات با در نظر گرفتن کلمات همجوار تولید می‌شود که تا حدودی اطلاعات نحوی نیز در این مقادیر لحاظ می‌شود، اما با توجه به پیچیدگی زبان طبیعی و وابستگی طولانی میان کلمات، نیاز به در نظر گرفتن نقش نحوی کلمات و وابستگی‌ها آنها بیشتر حس می‌شود. با توجه به اینکه روش Skip-gram با نمونه‌برداری منفی با پیش‌بینی زمینه به تولید بردار می‌پردازد می‌توان از آن برای تولید بردارهای پیوسته دیگری نظیر بردارهای نحوی نیز استفاده کرد. در این روش پنجره‌ای به طول n در سمت راست و چپ کلمه ورودی به عنوان زمینه آن در نظر گرفته می‌شوند. با تعیین نقش نحوی کلمه ورودی با ایجاد ساختار نحوی جمله، وابستگی‌های بدست آمده برای آن را می‌توان به عنوان زمینه کلمه ورودی در فرآیند ایجاد بردار در نظر گرفت.

به عنوان مثال، گراف وابستگی نحوی جمله نمونه "چینش دکمه‌ها و کنترل‌های دوربین واقعا عالیست" در شکل ۱ نشان داده شده است. این گراف با توجه به اطلاعات نحوی مربوط به جمله ایجاد شده است. در رویکرد پیشنهادی، اطلاعات نحوی زمینه هر کلمه موجود در جمله استخراج شده و برای ایجاد نمایش برداری به کار گرفته می‌شود. با بیان دقیق‌تر، برای تولید اطلاعات نحوی، عامل تغییر دهنده و نوع وابستگی با استفاده از خط زیر به هم متصل شده‌اند. همچنین والد یک نود در گراف وابستگی، با در نظر گرفتن ارتباط معکوس و اضافه کردن 'inv' به وابستگی مربوطه، پوشش داده شده است. شکل ۲ اطلاعات نحوی استخراج شده برای هر یک از کلمات موجود در جمله نمونه مطرح شده را نشان می‌دهد.



شکل ۱ گراف وابستگی نحوی جمله نمونه "چینش دکمه‌ها و کنترل‌های دوربین واقعا عالیست"

کلمه هدف	اطلاعات نحوی
چینش	SBJ_INV_عالیست MOZ_دکمه‌ها
دکمه‌ها	MOZ_INV_چینش NCONJ_و MOZ_دوربین
و	NCONJ_INV_دکمه‌ها POSDEP_کنترل‌های
کنترل‌های	POSDEP_INV_و
دوربین	MOZ_INV_دکمه‌ها
واقعا	ADV_INV_عالیست
عالیست	SBJ_چینش ADV_واقعا

شکل ۲ اطلاعات نحوی هر یک از کلمات جمله نمونه "چینش دکمه‌ها و کنترل‌های دوربین واقعا عالیست"

در روش Skip-gram با نمونه‌برداری منفی پیش فرض با پنجره دو، برای واژه "چینش" وابستگی آن با واژه "عالیست" در نظر گرفته نمی‌شود در حالی که این واژه در مجموعه اطلاعات زمینه واژه "چینش" با روش نحوی قرار دارد. برای جملات با طول بیشتر، تاثیر در نظر گرفتن ساختار نحوی بسیار بیشتر مشاهده می‌شود. همچنین این رویکرد از اضافه شدن کلمات نامرتبط به مجموعه اطلاعات زمینه کلمه هدف تنها به دلیل اینکه در پنجره آن قرار دارد اجتناب می‌کند که منجر به بردارهای با دقت بیشتری می‌گردد.

برای تولید بردارهای پیوسته، از نسخه بهبود یافته مدل Skip-gram با نمونه برداری منفی، اقتباس شده از [۲۶] استفاده شده است. مدل شبکه عصبی بر روی داده‌های آموزشی پیش پردازش مسئله، آموزش می‌بینند. پیش پردازش شامل نرمال سازی کلمات (حذف فاصله‌های اضافی، تنظیم نیم‌فاصله‌ها، حذف کشیدگی‌های اضافی) و تبدیل متن غیررسمی به رسمی می‌باشد. سپس ساختار نحوی هر یک از جملات استخراج می‌شود. برای آموزش، هر کلمه (نود موجود در گراف نحوی) به عنوان کلمه هدف و وابستگی‌های آن به عنوان اطلاعات نحوی آن در نظر گرفته می‌شوند. مدل شبکه عصبی آموزش دیده شده شامل دو ماتریس می‌باشد که ماتریس وزن مربوط به لایه embeddings به نمایش برداری تعبیه شده کلمات و ماتریس وزن مربوط به لایه prediction به نمایش برداری تعبیه شده اطلاعات نحوی اشاره دارد.

۳-۳- نمایش ویژگی‌ها

در رویکرد پیشنهادی، هر نمونه توسط کلمات موجود در آن که ویژگی‌های اصلی آن محسوب می‌شوند نمایش داده می‌شود. در واقع یک ماتریس تعبیه شده به طبقه‌بند داده خواهد شد که هر ردیف آن، ویژگی‌های نمونه را نمایش می‌دهد. برای نمایش ویژگی‌ها نیز از بردارهای کلمه و نحوی بدست آمده در گام قبلی استفاده می‌شود. برای این منظور، همانند گام آموزش، هر جمله ورودی ابتدا پیش پردازش شده و ساختار نحوی آن ایجاد می‌شود. ما هر کلمه را با w و وابستگی‌های آن را با c_i نشان می‌دهیم. به عنوان مثال برای واژه (w_i) "چینش"، اطلاعات زمینه (c_i) آن شامل $c_i = \{'SBJ_INV_عالیست'; 'MOZ_دکمه‌ها'\}$ می‌باشد. اگر $\overrightarrow{v_w}$ بردار استخراج شده با استفاده از تعبیه کلمات و $\overrightarrow{v'_{c_i}}$ بردار استخراج شده با استفاده از تعبیه نحوی برای وابستگی c_i باشد، حالت‌های مختلف نمایش w عبارتند از:

- نمایش مبتنی بر تعبیه کلمات (WEbR)

$$WEbR(w) = \overrightarrow{v_w} \qquad (5)$$

در این حالت برای نمایش کلمه w، بردار $\overrightarrow{v_w}$ از تعبیه کلمات استخراج شده بدست می‌آید. فرضیه اساسی در اینجا این است که نمایش کلمه به تنهایی بدون توجه به نمایش وابستگی‌های آن می‌تواند کافی باشد یا خیر.

- نمایش مبتنی بر تعبیه نحوی (SEbR)

$$SEbR(w) = \frac{1}{|C|}\sum_{c_i \in C} \overrightarrow{v'_{c_i}} \qquad (6)$$

این نمایش فرض می‌کند که استفاده از بردارهای نحوی کلمه مورد نظر به تنهایی برای نمایش کلمه موثرتر است و نیازی به word embeddings آن نمی‌باشد. در این حالت برای نمایش کلمه، میانگین بردارهای وابستگی‌های کلمه مورد نظر بدست می‌آید.

- میانگین نمایش مبتنی بر تعبیه کلمات و نمایش مبتنی بر تعبیه نحوی (AWSE)

$$AWSE(w) = \frac{1}{2}\big(WEbR(w) + SEbR(w)\big) \qquad (7)$$

برای در نظر گرفتن کلمه با استفاده از هر دو نمایش مبتنی بر بردار کلمه و نمایش نحوی، می‌توان میانگین نمایش‌های فوق را در نظر گرفت. به عبارت دیگر در این نمایش به هر دو نمایش مبتنی بر بردار کلمه و نمایش نحوی اهمیت یکسانی داده می‌شود.

- ترکیب نمایش‌های مبتنی بر تعبیه کلمات و نحوی (CSWE)

$$CWSE(w) = WEbR(w) \,.\, SEbR(W) \qquad (8)$$

مطالعات قبلی نشان داده‌اند که ترکیب تعبیه‌ها می‌تواند بسیار موثر باشد [۲۷-۲۸]. فرضیه اساسی در اینجا این است که نگه داشتن هر دو نمایش مبتنی بر کلمه و نحوی در بردار ویژگی، قدرت تفکیک‌پذیری بالاتری را ارائه می‌دهد. در این حالت، ابعاد نمایش هر کلمه افزایش می‌یابد.

۴-۳- طبقه‌بندی با شبکه عصبی کانولوشنی

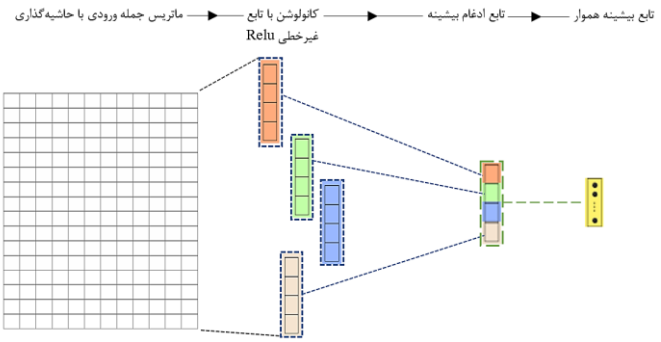
به منظور تعیین احساس جمله ورودی به یکی از دسته‌های تعیین شده از معماری یادگیری عمیق شبکه عصبی کانولوشنی استفاده شده است. شمای کلی مدل در شکل ۳ قابل مشاهده است. مدل از چهار لایه اصلی تشکیل شده است. لایه اول، لایه ورودی است. هر جمله نمونه در لایه ورودی به صورت ترکیبی از بردارهای ویژگی آن می‌باشد. به عبارت دیگر، اگر $\overrightarrow{v_{w_i}}$ بردار کلمه w_i موجود در جمله S باشد، آنگاه جمله S به صورت ترکیبی از بردارهای کلمات آن نشان داده می‌شود:

$$S = \overrightarrow{v_{w_i}} \,.\, \overrightarrow{v_{w_2}} \;\ldots\; \overrightarrow{v_{w_n}} \qquad (9)$$

همچنین باید توجه کرد که هر یک از بردارهای $\overrightarrow{v_{w_i}}$ با توجه به شکل‌های مختلف نمایش تشریح شده در بخش ۳-۳ می‌باشد. دقت شود با توجه به طول متغیر جملات، فرآیند حاشیه گذاری بر روی ورودی‌ها انجام می‌شود.

لایه دوم لایه کانولوشن است که در آن بر روی هر دنباله از کلمات جمله، ابتدا یک فیلتر کانولوشنی اعمال شده و با بایاس جمع می‌شود. سپس تابعی غیر خطی Relu بر روی آن اعمال می‌شود تا مجموعه جدیدی از ویژگی‌ها بدست آید.

به منظور ترکیب بردارهای حاصل از لایه کانولوشن در یک بردار تکی و مقابله با طول متفاوت جملات، عملگر ادغام بیشینه مورد استفاده قرار می‌گیرد. استفاده از این عملگر موجب می‌شود بردار حاصل شامل مهترین ویژگی‌ها باشد. بردار ویژگی حاصل به یک لایه کاملا متصل ارسال می‌شود تا توضیح احتمال هر یک از کلاس‌های مسئله با استفاده از تابع بیشینه هموار مشخص شود.



شکل ۳ معماری طبقه‌بند شبکه عصبی کانولوشنی برای مسئله تعیین احساس متن فارسی

۴- یافته‌های پژوهش

در این بخش ابتدا مشخصات مجموعه داده مورد استفاده بیان شده و سپس معیارهای ارزیابی تشریح می‌گردد. سپس کارایی رویکرد پیشنهادی با توجه به پرسش‌های پژوهش با آزمایش‌های مختلف نشان داده می‌شود.

۴-۱- مجموعه داده‌های آزمون

تحلیل احساس متن می‌تواند در بسیاری از کاربردها مورد استفاده قرار گیرد. بسته به نوع سیستم، سطحی که در آن نیاز به تحلیل احساس داریم نیز می‌تواند متفاوت باشد. تعیین احساس متن در سطح جمله موجب می‌شود سیستم مربوط به اطلاعات زمینه به طور دقیق اهمیت داده و بررسی را به صورت محلی و با ریزدانگی خوب انجام دهد. از آنجا که رویکرد پیشنهادی نیز در سطح جمله کار می‌کند، به منظور ارزیابی آن از پیکره DeepSentiPers [۴] استفاده می‌شود که در سایر پژوهش‌های که بر روی این سطح از ریزدانگی تمرکز کردند نیز به کار گرفته شد. این مجموعه داده نسخه متعادل شده و تقویت شده پیکره SentiPers [۲۱] است که شامل نظرات کاربران در مورد محصولات دیجیتال است که از دیجی‌کالا گردآوری شده است. به هر یک از جملات موجود در پیکره به صورت دستی عددی در بازه 2- تا 2+ نسبت داده شده است که میزان منفی بودن تا مثبت بودن جمله را مشخص می‌کند. این مقادیر به صورت {-2,-1,0,+1,+2} است که شامل دو کلاس منفی (خشمگین و ناراحت)، دو کلاس مثبت (شاد و مسرور) و یک کلاس خنثی است. نمونه‌ای از جملات موجود در این پیکره در جدول ۱ نشان داده شده است. همچنین در جدول ۳، آمارهای مربوط به این مجموعه داده ارائه شده است.

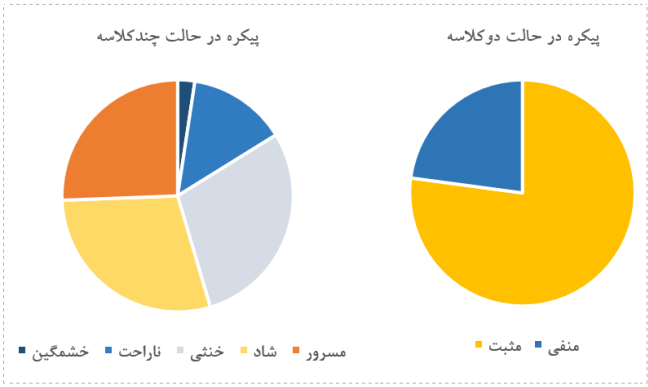
جدول ۳ آمارهای مربوط به پیکره DeepSentiPers

تعداد کل جملات	۱۲۱۳۸
میانگین طول جملات	۱۹.۲
تعداد جملات خشمگین	۲۹۱
تعداد جملات ناراحت	۱۶۷۵
تعداد جملات خنثی	۳۵۴۸
تعداد جملات شاد	۳۵۱۷
تعداد جملات مسرور	۳۱۰۷

با توجه به شیوه برچسب گذاری داده‌ها، می‌توان از این مجموعه داده در هر دو حالت دو کلاسه و چندکلاسه استفاده کرد. در حالت دو کلاسه، داده‌های خنثی کنار گذاشته می‌شوند و برچسب‌های بیشتر از صفر به عنوان کلاس مثبت و کمتر از صفر به عنوان کلاس منفی در نظر گرفته می‌شود. شکل ۴ توزیع داده‌ها را در هر دو حالت مسئله نشان می‌دهد.

یکی از مواردی که الگوریتم‌های طبقه‌بندی را با مشکل مواجه می‌کند بحث نامتوازن بودن داده‌ها است. در این شرایط اکثر الگوریتم‌ها به سمت کلاس غالب تمایل پیدا می‌کنند و به کلاس با توزیع بالاتر اولویت بالاتری نسبت به کلاس‌های دیگر میدهند.

برای دستیابی به یک مدل با کیفیت بالا، مدیریت داده‌های نامتوازن یکی از گام‌هایی است که در بحث آنالیز داده‌ها می‌بایست انجام شود. در مدل پیشنهادی برای کاهش این مشکل، نمونه‌های کلاس‌های با توزیع کمتر تکرار می‌شوند تا تاثیر کلاس غالب کمتر شود.



شکل ۴ توزیع کلاس‌ها در پیکره DeepSentiPers

۴-۲- معیارهای ارزیابی آزمایش

به منظور ارزیابی رویکرد پیشنهادی و مقایسه با روش‌های بروز، ما از معیارهای ارزیابی مشابه با کارهای قبلی در زمینه تحلیل احساس متن استفاده کردیم. معیارهای ارزیابی استفاده شده شامل پوشش ، صحت و معیار F می‌باشد که به صورت زیر تعریف می‌شوند. در رابطه‌های زیر، tp، fn، fp و tn به ترتیب به تعداد مثبت صادق ، منفی کاذب ، مثبت کاذب و منفی صادق اشاره دارد.

- معیار صحت: نسبت تعداد مثبت‌های درست (مثبت صادق) بر کل تعداد مثبت‌های پیش‌بینی (مثبت صادق و مثبت کاذب) شده است:

$$precision = \frac{tp}{tp + fp} \quad (10)$$

- معیار پوشش: نسبت تعداد مثبت‌های درست (مثبت صادق) بر تعداد کل موارد مورد انتظار (مثبت صادق و منفی کاذب):

$$recall = \frac{tp}{tp + fn} \quad (11)$$

- معیار F: از آنجا که مدل مورد استفاده می‌بایست از هر دو صحت و پوشش بالایی برخوردار باشد، معیارهای فوق در قالب یک معیار به صورت زیر ترکیب می‌شوند:

$$F_1 - Mesaure = \frac{2 * recall * precision}{recall + precision} \quad (12)$$

۴-۳- نتایج ارزیابی

با توجه به پرسش‌های مطرح شده برای پژوهش جاری، آزمایشات متفاوتی انجام شده است. در این بخش کارایی رویکرد پیشنهادی برای تعیین احساس متن بر روی مجموعه داده استاندارد معرفی شده ارزیابی می‌شود. همچنین به منظور مقایسه، از نتایج ارائه شده در مقاله [۴] استفاده می‌شود.

۴-۳-۱- پرسش اول: آیا شکل‌های متفاوت نمایش ویژگی‌ها تاثیر متفاوتی بر عملکرد مدل پیشنهادی تعیین احساس متن دارد؟

در مرحله اول ارزیابی، مدل پیشنهادی با شکل‌های متفاوت نمایش ویژگی‌ها مورد بررسی قرار می‌گیرد. این آزمایش جهت تعیین بهترین نمایش بر اساس شکل‌های متفاوت نمایش ویژگی‌ها انجام می‌شود. نتایج ارزیابی که بر روی پیکره DeepSentiPers انجام شده در جدول ۴ نشان داده شده است. جزئیات مربوط به نمایش‌های مختلف مدل در بخش ۳-۳ بیان شده است. نتایج مربوطه با توجه به دو کلاسه بودن یا چند کلاسه بودن مسئله متفاوت می‌باشد. ردیف دوم که مربوط به نتایج مدل پیشنهادی با نمایش نحوی ویژگی‌ها است نشان می‌دهد که تنها اتکا به اطلاعات نحوی کافی نیست.

مدل پیشنهادی با نمایش مبتنی بر میانگین بردارهای مبتنی بر کلمه و نحو و ترکیب آنها تحقیقات قبلی را تایید می‌کند [۲۸-۲۷] و حاکی از بهبود نتایج است. بهترین نتایج مدل پیشنهادی با استفاده از نمایش مبتنی بر ترکیب بردارها بدست آمده است. این مدل از نظر معیار F1 بالاترین امتیاز را در هر دو مساله دوکلاسه و چندکلاسه کسب کرده است که حاکی از امتیاز پوشش و صحت بالاتر نسبت به سایر شکل‌های نمایش ویژگی‌ها می‌باشد.

جدول ۴ ارزیابی عملکرد شمای پیشنهادی با شکل‌های متفاوت نمایش ویژگی‌ها

مدل مورد ارزیابی	حالت دو کلاسه	حالت چند کلاسه
	معیار F_1	معیار F_1
مدل CNN- WEbR	۹۱.۲۶	۶۸.۷۴
مدل CNN- SEbR	۸۹.۹۸	۶۷.۸۹
مدل CNN- AWSE	۹۱.۷۷	۶۹.۸۲
مدل CNN- CSWE	**۹۲.۱۳**	**۷۰.۵۱**

در نتیجه پاسخ پرسش اول پژوهش مثبت است. در ادامه جهت ارزیابی‌های بیشتر و پاسخ به دیگر پرسش‌های پژوهش از مدل معماری یادگیری عمیق شبکه عصبی کانولوشنی با نمایش مبتنی بر ترکیب بردارها (به عنوان مدل پیشنهادی) استفاده می‌شود.

۴-۳-۲- پرسش دوم: آیا یادگیری نمایش ویژگی‌ها در کنار استفاده از معماری شبکه عصبی کانولوشنی عملکرد بهتری نسبت به مدل‌های سنتی نظیر ماشین بردار پشتیبان در تعیین احساس متن دارد؟

به منظور پاسخ به پرسش دوم پژوهش، از سه مدل سنتی که با تبدیل نمونه‌ها به شیوه مدل فضای برداری و با استفاده از مدل‌های یادگیری ماشین خطی شامل بیز ساده، ماشین بردار پشتیبان [۲۹] و نزول شیب تصادفی [۳۰] عمل می‌کنند، استفاده کرده‌ایم. نتایج در جدول ۵ نشان داده شده است. بهترین نتایج میان مدل‌های سنتی با مدل ماشین بردار پشتیبان کسب شده است ($F_1 = 67.62$ برای چند کلاسه و $F_1 = 91.31$ برای دو کلاسه).

نکته دیگری که در نتایج قابل مشاهده است عملکرد مدل بیز ساده است. از آنجا که این مدل ویژگی‌ها (در اینجا کلمات) را مستقل از هم فرض می‌کند نتایج آن در مقایسه با سایر مدل‌ها نشان می‌دهد که این فرض درست نیست و کلمات موجود در جمله بر روی یکدیگر تاثیر می‌گذارند. همانطور که مشاهده می‌شود رویکرد پیشنهادی در مقایسه با ماشین بردار پشتیبان در هر دو حالت دو کلاسه و چندکلاسه معیار F بالاتری کسب کرده است. در نتیجه پاسخ پرسش دوم پژوهش نیز مثبت است.

جدول ۵ ارزیابی عملکرد شمای پیشنهادی در مقایسه با مدل‌های ستنی یادگیری ماشین

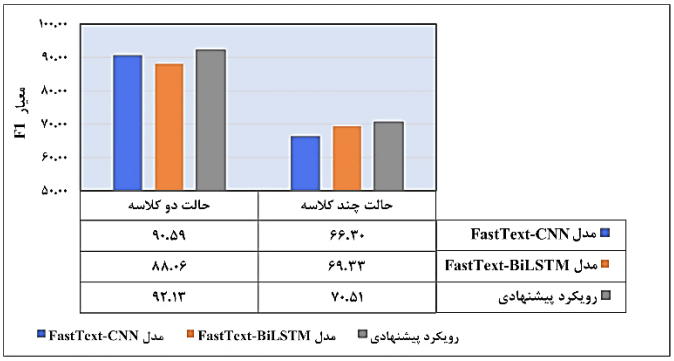
مدل مورد ارزیابی	حالت دو کلاسه	حالت چند کلاسه
	معیار F_1	معیار F_1
مدل Naïve Bayes	۷۶.۶۷	۵۶.۶۵
مدل Support Vector Machine	۹۱.۳۱	۶۷.۶۲
مدل Stochastic Gradient Descent	۷۴.۳۹	۶۲.۰۴
رویکرد پیشنهادی	**۹۲.۱۳**	**۷۰.۵۱**

۴-۳-۳- پرسش سوم: آیا یادگیری نمایش ویژگی‌ها بر روی داده‌های خاص مسئله نسبت به استفاده از بردارهای از قبل آموزش دیده منجر به بهبود نتایج می‌شود؟

همانطور که در بخش پیشینه پژوهش بیان شد اکثریت رویکردهای ارائه شده جدید از بردارهای از قبل آموزش دیده برای نمایش ویژگی‌ها استفاده می‌کنند. پژوهش‌های متعددی نشان دادند [۲۴-۲۳] که یادگیری نمایش ویژگی‌ها بر روی داده‌های خاص مسئله منجر به عملکرد بهتر مدل طبقه‌بندی می‌شود.

به منظور بررسی صحت این مورد برای مسئله تحلیل احساس متن فارسی، عملکرد رویکرد پیشنهادی که در آن بردار ویژگی‌های کلمات و اطلاعات نحوی بر روی داده‌های خاص مسئله آموزش دیده شده‌اند را با مدل‌های به روز در این مسئله مورد مقایسه قرار می‌دهیم. برای این منظور از مدل‌های ارائه شده در پژوهش [۴] که آن‌ها نیز بر روی دیتاست DeepSentiPers ارزیابی شده‌اند استفاده می‌کنیم. در این مدل‌ها از بردارهای از قبل آموزش دیده FastText [۳۲-۳۱] استفاده می‌شود.

نتایج ارزیابی در شکل ۵ نشان داده شده است. همانطور که مشاهده می‌شود عملکرد مدل‌های مورد مقایسه بر روی دو حالت چند کلاسه و دو کلاسه متفاوت می‌باشد. در حالی که مدل FastText-CNN بر روی حالت دو کلاسه عملکرد بهتری دارد، مدل Fast-Text-BiLSTM در حالت چند کلاسه نتایج بهتری کسب کرده است. رویکرد پیشنهادی در هر دو حالت چند کلاسه و دو کلاسه بهتر عمل کرده است. عملکرد مدل علاوه بر اینکه به پرسش سوم پژوهش پاسخ مثبت می‌دهد (استفاده از داده‌های خاص مسئله برای تعیین احساس متن فارسی منجر به نتایج بهتر می‌شود)، همچنین حاکی از کارایی بالاتر مدل در مقایسه با رویکردهای به روز است.



شکل ۵ ارزیابی عملکرد شمای پیشنهادی در مقایسه با رویکردهای به روز

۵- نتیجه‌گیری و کارهای آینده

پژوهش جاری بر روی تعیین احساس متن فارسی در سطح جمله به عنوان یک گام بسیار مهم در بسیاری از سیستم‌های خبره تمرکز کرده است. مدل پیشنهادی با تولید بردارهای نمایش اطلاعات نحوی کلمات و فراهم کردن شکل متفاوتی از نمایش ویژگی‌ها، با هدف کاهش مشکل وابستگی‌های بلند و ابهام ، اطلاعات کامل‌تری را به طبقه‌بند مبتنی بر شبکه عصبی کانولوشنی ارائه می‌دهد. با استفاده از این رویکرد، تاثیر رخداد تصادفی کلمات که موجب اشتباه در نمایش ویژگی‌ها می‌شود کاهش یافته و ویژگی‌های عملکردی کلمات بهتر ثبت می‌شود. ما شکل‌های متفاوتی از نمایش ویژگی‌ها را ارائه دادیم که آزمایشات انجام شده نشان می‌دهد شمای پیشنهادی با ترکیب نمایش بردار کلمات و اطلاعات نحوی منجر به بهترین نتایج می‌گردد. همچنین با توجه به نامتوازن بودن داده‌ها، افزایش تعداد نمونه‌های کلاس با توزیع کمتر نیز انجام شده است. بررسی‌های کامل بر روی پیکره استاندارد DeepSentiPers و مقایسه با مدل‌های سنتی یادگیری ماشین و مدل‌های به روز که عمدتا بر اساس الگوریتم‌های یادگیری عمیق هستند حاکی از عملکرد بهتر رویکرد پیشنهادی می‌باشد. ما همچنین نشان دادیم که یادگیری نمایش ویژگی‌ها بر روی داده‌های خاص مسئله موجب بهبود عملکرد مدل می‌شود.

به عنوان کارهای آینده، از آنجا که خروجی مدل پیشنهادی بسیار امیدوار کننده بود، ما در نظر داریم تا این مدل را در یک سیستم پیش‌بینی وضعیت بازار با توجه به پیام‌های کاربران در شبکه‌های اجتماعی بگنجانیم. ما همچنین در نظر داریم عملکرد مدل با ریزدانگی‌های مختلف (در سطح سند و جنبه) را بیشتر مورد مطالعه قرار دهیم. همچنین مطالعات ما نشان می‌دهد با توسعه اطلاعات زمینه با توجه به جملات همسایه (جملات قبل و بعد)، بهبود بیشتر قابل کسب خواهد بود. ما همچنین تصمیم داریم تا تاثیر این نمایش جدید را بر روی کاربردهای دیگر نیز مورد بررسی قرار دهیم.

مراجع

[1] T. Swathi, N. Kasiviswanath, and A. A. Rao, “An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis,” Applied Intelligence, vol. 52, no. 12, pp. 13675–13688, Sep. 2022, doi: 10.1007/S10489-022-03175-2/METRICS.

[2] S. Shumaly, M. Yazdinejad, and Y. Guo, “Persian sentiment analysis of an online store independent of pre-processing using convolutional neural network with fastText embeddings,” PeerJ Comput Sci, vol. 7, pp. 1–22, Mar. 2021, doi: 10.7717/PEERJ-CS.422/SUPP-1.

[3] 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, Mar. 2023, doi: 10.1016/j.inffus.2022.09.025.
[4] J. Pourmostafa, R. Sharami, P. A. Sarabestani, and S. A. Mirroshandel, "DeepSentiPers: Novel Deep Learning Models Trained Over Proposed Augmented Persian Sentiment Corpus," 2020, https://doi.org/10.48550/arXiv.2004.05328
[5] I. Abu Farha and W. Magdy, "A comparative study of effective approaches for Arabic sentiment analysis," Inf Process Manag, vol. 58, no. 2, Mar. 2021, doi: 10.1016/j.ipm.2020.102438.
[6] S. Marchesin, A. Purpura, and G. Silvello, "Focal elements of neural information retrieval models. An outlook through a reproducibility study," Inf Process Manag, vol. 57, no. 6, p. 102109, 2020, https://doi.org/10.48550/arXiv.2004.05328.
[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013, https://doi.org/10.48550/arXiv.1301.3781.
[8] N. Yadav, O. Kudale, A. Rao, S. Gupta, and A. Shitole, "Twitter Sentiment Analysis Using Supervised Machine Learning," Lecture Notes on Data Engineering and Communications Technologies, vol. 57, pp. 631–642, 2021, doi: 10.1007/978-981-15-9509-7_51/COVER.
[9] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning Word Vectors for 157 Languages," Feb. 2018, https://doi.org/10.48550/arXiv.1802.06893
[10] B. AlBadani, R. Shi, and J. Dong, "A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM," Applied System Innovation 2022, Vol. 5, Page 13, vol. 5, no. 1, p. 13, Jan. 2022, doi: 10.3390/ASI5010013.
[11] S. C. Nistor, M. Moca, D. Moldovan, D. B. Oprean, and R. L. Nistor, "Building a Twitter Sentiment Analysis System with Recurrent Neural Networks," Sensors 2021, Vol. 21, Page 2266, vol. 21, no. 7, p. 2266, Mar. 2021, doi: 10.3390/S21072266.
[12] M. Bibi et al., "A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis," Pattern Recognit Lett, vol. 158, pp. 80–86, Jun. 2022, doi: 10.1016/J.PATREC.2022.04.004.
[13] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," Feb. 2020, [Online]. https://doi.org/10.48550/arXiv.2003.00104
[14] A. Diwali, K. Dashtipour, K. Saeedi, M. Gogate, E. Cambria, and A. Hussain, "Arabic sentiment analysis using dependency-based rules and deep neural networks," Appl Soft Comput, vol. 127, Sep. 2022, doi: 10.1016/j.asoc.2022.109377.
[15] M. E. Basiri and A. Kabiri, "Sentence-level sentiment analysis in Persian," 3rd International Conference on Pattern Analysis and Image Analysis, IPRIA 2017, pp. 84–89, Jul. 2017, doi: 10.1109/PRIA.2017.7983023.
[16] F. E. Rashed and N. Abdolvand, "A Supervised Method for Constructing Sentiment Lexicon in Persian Language," Journal of Computer & Robotics, vol. 10, no. 1, pp. 2017–2028, 2017.
[17] S. A. Razavi and M. Asadpour, "Word embedding-based approach to aspect detection for aspect-based summarization of Persian customer reviews," ACM International Conference Proceeding Series, Oct. 2017, doi: 10.1145/3109761.3158403.
[18] K. Dashtipour, M. Gogate, A. Adeel, C. Ieracitano, H. Larijani, and A. Hussain, "Exploiting Deep Learning for Persian Sentiment Analysis," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10989 LNAI, pp. 597–604, 2018, doi: 10.1007/978-3-030-00563-4_58/COVER.
[19] 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, Oct. 2021, doi: 10.1016/J.NEUCOM.2021.02.020.
[20] M. B. Dastgheib, S. Koleini, and F. Rasti, "The application of Deep Learning in Persian Documents Sentiment Analysis," International Journal of Information Science and Management (IJISM), vol. 18, no. 1, pp. 1–15, Feb. 2020, Accessed: Feb. 01, 2023. [Online]. Available: https://ijism.ricest.ac.ir/article_698303.html
[21] P. Hosseini, A. A. Ramaki, H. Maleki, M. Anvari, and S. A. Mirroshandel, "SentiPers: A Sentiment Analysis Corpus for Persian," Jan. 2018, doi: 10.48550/arxiv.1801.07737.
[22] D. Valle-Cruz, V. Fernandez-Cortez, A. López-Chau, and R. Sandoval-Almazán, "Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis During Pandemics: A Comparative Study of the H1N1 and the COVID-19 Periods," Cognit Comput, vol. 14, no. 1, pp. 372–387, Jan. 2022, doi: 10.1007/S12559-021-09819-8/TABLES/17.
[23] Chen, T. L., Emerling, M., Chaudhari, G. R., Chillakuru, Y. R., Seo, Y., Vu, T. H., & Sohn, J. H. (2021). Domain specific word embeddings for natural language processing in radiology. Journal of Biomedical Informatics, 113, 103665. https://doi.org/10.1016/J.JBI.2020.103665
[24] G. Li, "Improving Biomedical Ontology Matching Using Domain-specific Word Embeddings," ACM International Conference Proceeding Series, Oct. 2020, doi: 10.1145/3424978.3425102.
[25] Y. Zhang, H. Zhao, J. Wei, J. Zhang, M. F. Flanagan, and J. Xiong, "Context-Based Semantic Communication via Dynamic Programming," IEEE Trans Cogn Commun Netw, vol. 8, no. 3, pp. 1453–1467, Sep. 2022, doi: 10.1109/TCCN.2022.3173056.
[26] A. Komninos and S. Manandhar, "Dependency based embeddings for sentence classification tasks," in 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference, 2016. doi: 10.18653/v1/n16-1175.
[27] M. Pota, F. Marulli, M. Esposito, G. de Pietro, and H. Fujita, "Multilingual POS tagging by a composite deep architecture based on character-level features and on-the-fly enriched Word Embeddings," Knowl Based Syst, vol. 164, 2019, doi: 10.1016/j.knosys.2018.11.003.
[28] A. Dogra, P. Kaur, G. S. Kohli, and J. Bedi, "Raccoons at SemEval-2022 Task 11: Leveraging Concatenated Word Embeddings for Named Entity Recognition," SemEval 2022 - 16th International Workshop on Semantic Evaluation, Proceedings of the Workshop, pp. 1576–1582, 2022, doi: 10.18653/V1/2022.SEMEVAL-1.217.
[29] B. He, J. X. Huang, and X. Zhou, "Modeling term proximity for probabilistic information retrieval models," Inf Sci (N Y), vol. 181, no. 14, pp. 3017–3031, 2011, https://doi.org/10.1016/j.ins.2011.03.007.
[30] A. B. Prasetijo, R. R. Isnanto, D. Eridani, Y. A. A. Soetrisno, M. Arfan, and A. Sofwan, "Hoax detection system on Indonesian news sites based on text classification using SVM and SGD," in Proceedings - 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2017, 2017, vol. 2018-January. doi: 10.1109/ICITACEE.2017.8257673.
[31] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," Trans Assoc Comput Linguist, vol. 5, pp. 135–146, Dec. 2017, https://doi.org/10.48550/arXiv.1607.04606.
[32] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "FastText.zip: Compressing text classification models," Dec. 2016, doi: 10.48550/arxiv.1612.03651.
[33] G. Yildirim, "A novel grid-based many-objective swarm intelligence approach for sentiment analysis in social media," Neurocomputing, vol. 503, pp. 173–188, Sep. 2022, doi: 10.1016/j.neucom.2022.06.092.

میثم روستائی عضو هیئت علمی گروه مهندسی کامپیوتر دانشگاه مازندران، دکتری خود را در رشته مهندسی کامپیوتر از دانشگاه شیراز دریافت کرده‌اند. زمینه‌های تحقیقاتی اصلی ایشان شامل یادگیری عمیق، پردازش زبان طبیعی و تحلیل شبکه‌های اجتماعی است. در حال حاضر، ایشان در اجرای پروژه تحقیقاتی و کاربردی در حوزه فناوری اطلاعات مشارکت دارند.

آدرس پست الکترونیک ایشان عبارت است از:

m.roostaee@umz.ac.ir