Journal of New Achievements in Electrical, Computer and Technology
Vol. 2, No. 3, Jun-Sep 2022,(134-144)
WWW.JNAECT.IR

مجله دستاوردهای نوین در برق،کامپیوتر و فناوری
دوره ۲، شماره ۳، تابستان ۱۴۰۱ (۱۴۴-۱۳۴)

تحلیل احساسات توییت‌های فارسی چند کلاسه مبتنی بر ترکیب روابط معنایی و ماشین بردار پشتیبان

حسین رعیت پرور[۱] ، محمدعلی جوادزاده[۲]

[۱] دانشجوی کارشناسی ارشد دانشگاه جامع امام حسین(ع)، hosseinrayatparvar@ihu.ac.ir
[۲] استادیار دانشگاه جامع امام حسین(ع)، javadzade@ ihu.ac.ir

چکیده

با رشد تکنولوژی، استفاده از شبکه‌های اجتماعی محبوبیت بیشتری پیدا کرده است. یکی از محبوب‌ترین شبکه‌های اجتماعی پلتفرم توییتر می‌باشد. تجزیه و تحلیل احساسات توییت‌های کاربران نقش مهمی در نمایش احساسات کاربران از شرایط موجود جامعه دارد. در سال‌های اخیر باتوجه به محاوره‌ای شدن متن‌های توییت‌های کاربران تحلیل احساسات را دچار مشکل و باعث کاهش دقت آن شده است؛ همچنین باعث سختی عملیات پردازش زبان طبیعی شده است. در این پژوهش، روشی برای تحلیل احساسات توییت‌های فارسی مبتنی بر ترکیب روابط معنایی و طبقه‌بندی ماشین بردار پشتیبان ارائه شده است. برای استخراج ویژگی‌ها از رابطه معنایی FastText استفاده می‌شود. باتوجه به اینکه تعداد ویژگی‌های زیادی استخراج شده است؛ باید آن‌ها را کاهش داد که این عمل با استفاده از شبکه عصبی حافظه طولانی کوتاه مدت (LSTM) صورت می‌پذیرد. در بخش آخر روش پیشنهادی برای طبقه‌بندی احساسات درون توییت‌ها از مدل ماشین بردار پشتیبان بهره گرفته شده است. معیارهای ارزیابی مورد استفاده در این پژوهش دقت، صحت، فراخوان و معیار F بوده است که نتایج ارزیابی به ترتیب مقادیر ۸۳.۹، ۸۴.۳، ۸۳.۹ و ۸۴ به دست آمده است. نتایج آزمایش‌ها نشان‌دهنده کاربردی بودن روش پیشنهادی در تحلیل احساسات توییت‌های فارسی به شش کلاس عصبانیت، ناراحتی، شادی، انزجار، تعجب‌برانگیز و ترس است.

واژه‌های کلیدی: تحلیل احساسات چندکلاسه، پردازش زبان طبیعی، شبکه¬های عصبی LSTM، ماشین بردار پشتیبان.

۱- مقدمه

وب‌سایت‌های بلاگ‌نویسی کوچک[1] از انواع مختلف داده برای تکامل خودشان استفاده کرده‌اند که این باعث شده است که کاربرپسند شوند و کاربران به آن‌ها جذب شوند. ماهیت وجودی میکروبلاگ‌ها دارای ویژگی‌های مشخصی هستند که عبارتند از کاربران در آن نظراتی را درباره موضوعات مختلف ارسال می‌کنند، در مورد مسائل روز بحث می‌کنند، اعتراض خود را نشان می‌دهند و احساسات مختلف خود را نسبت به زندگی روزمره ابراز می‌کنند. یکی از چالش‌ها ایجاد رویکردی است که بتوان با دقت بالا احساسات درون متن را تشخیص داد. یکی از این میکروبلاگ‌های معروف پلتفرم توییتر می‌باشد. در توییتر کاربران احساسات درونی خود را توییت می‌کنند، یعنی نظر خود را نسبت به مسائل مختلف اعم از سیاسی، اقتصادی، فرهنگی و غیره را با کلمات بیان می‌کنند و تجزیه و تحلیل احساسات از آن مفید است[۱]. حوزه تجزیه و تحلیل احساسات که در آن احساسات از متن جمع‌آوری، تحلیل و تجمیع می‌شود، در چند سال اخیر مورد توجه بسیاری قرار گرفته است. تجزیه و تحلیل این احساسات کاربران توییتر به صورت دستی بسیار وقت‌گیر و پرهزینه است. همچنین، تجزیه و تحلیل اظهار نظرهای اتوماتیک و به صورت رویکردهای ماشینی هم موانع و مشکلاتی دارد: این مشکلات اعم از استفاده از جملات با دستور زبان نادرست، استفاده از اصطلاحات محاوره‌ای و عدم پیروی از نگارش صحیح، بخش جدایی ناپذیری از تحلیل متن هستند[۲][۳]. زمانی‌که برای حل این مشکلات صحبت می‌شود، تکنیک‌های تجزیه و تحلیل احساسات نقش اساسی دارند. این تکنیک‌ها به طور خودکار احساسات کاربران توییتر را به کلاس‌های مثبت، منفی و حتی خنثی تخمین می‌زنند یا اینکه احساسات کاربران را به چند کلاس مختلف تقسیم‌بندی می‌کند[۴]. بنابراین، تجزیه و تحلیل احساسات برای استفاده از این توییت‌ها بسیار ارزشمند و کاربردی است زیرا تجزیه و تحلیل احساسات می‌تواند به عنوان زمینه‌ای از تحلیل متن‌های توییت‌ها در نظر گرفته شود که به طور خاص به دنبال تعیین احساس درونی متن یا کاربر است[۳]. این کار براساس ارتباط و وابستگی احساسی بین کلمات و جملات متون انجام می‌شود [۵][۶].

در این مقاله، ما به دنبال تجزیه و تحلیل احساسات کاربران در توییت‌های فارسی کاربران در پلتفرم توییتر هستیم. برای انجام این کار در این پژوهش از یک فرآیند چند مرحله‌ای بهره خواهیم گرفت بطوری که در مرحله اول عملیات پیش‌پردازش داده‌ها صورت خواهد گرفت. در ادامه مراحل روش پیشنهادی، عملیات استخراج ویژگی‌ها از متن‌های توییت‌ها می‌باشد که براساس روابط معنایی مدل FastText انجام می‌گیرد. در بخش سوم به دلیل تعداد زیاد ویژگی استخراج شده عملیات کاهش ویژگی صورت می‌گیرد که عملیات کاهش ویژگی توسط معماری شبکه عصبی حافظه طولانی کوتاه مدت[2] صورت خواهد گرفت و در نهایت در بخش آخر طبقه‌بندی توییت‌ها با استفاده از الگوریتم ماشین بردار پشتیبان[3] صورت گرفته خواهد شد. تفاوت کار ما با دیگر پژوهش‌ها در این است که کلاس‌ها را به سه دسته مثبت، منفی و خنثی تقسیم نمی‌کنیم بلکه به شش کلاس عصبانیت، ناراحتی، شادی، انزجار، تعجب‌برانگیز و ترس تقسیم می‌کنیم.

۲- کارهای مرتبط

روش‌های تجزیه و تحلیل احساسات را می‌توان به هفت دسته کلی تقسیم کرد که عبارتند از: مدل‌های مبتنی بر فرهنگ لغت، مدل‌های مبتنی بر واژگان، مدل‌های مبتنی بر نحو، یادگیری ماشین نظارت شده، یادگیری ماشین بدون نظارت [۵] و مدل‌های یادگیری عمیق [۷]. دسته اول، مینگ هو و بینگ لیو [۸] بیان می‌کنند که یک فرهنگ لغت از احساسات، با انتشار احساساتی که یافت شده‌اند چند کلمه اولیه این فرهنگ لغت را به‌وسیله نمودار مترادف و متضاد که درمدل WordNet است، به دست

[1] Microblogging

[2] Long short-term memory (LSTM)

[3] Support-vector networks (SVM)

می‌آید. در اینجا فقط صفت‌ها به عنوان کلمات احساسی در نظر گرفته می‌شوند. به هر صفت در یک جمله یک کلاس احساس (یعنی مثبت یا منفی) از فرهنگ لغت احساسات تولید شده، اختصاص داده می‌شود. وقتی یک کلمه نفی در فاصله چند کلمه‌ای از کلمه‌ای که شامل احساس است ظاهر می‌شود، قطبیت آن تغییر می‌کند. سپس با استفاده از رای اکثریت برای هر جمله یک کلاس از احساسات تعیین می‌شود. از این رو، همان کلاس احساس به هر بعد در آن جمله اختصاص داده می‌شود. با این حال، زمانی که تعداد کلمات مثبت و منفی یکسان باشد یعنی جمله در حالت خنثی باشند، از روش متفاوتی استفاده می‌شود. در آن صورت، هر صفت حامل احساس با نزدیک‌ترین بعد در جمله، از نظر فاصله کلمه، همراه است. سپس رأی اکثریت در میان تمام کلمات احساسی که با یک بعد مرتبط هستند استفاده می‌شود. دسته دوم، مبتنی بر احساسات یک واژه می‌باشد. در این دسته، از شباهت‌های احساسی و وابستگی معنایی کلمات و جملات استفاده می‌شود و دقت آن به وزن‌های آموزش دیده بسیار وابسته است[۹]. دسته سوم، مدل مبتنی بر نحو یک تجزیه‌کننده کم عمق و مجموعه گسترده‌ای از قوانین برای تشخیص و تحلیل احساسات استفاده می‌شود. در این مدل در کنار توصیف احساس یک کلمه، بلکه الگوهای انتقالی را نیز ارائه می‌دهد که بیان می‌کند کدام کلمات تحت تأثیر این احساس قرار می گیرند[۱۰]. دسته بعدی مدل‌های مبتنی بر یادگیری ماشین هستند که به دو دسته نظارت شده و بدون نظارت تقسیم‌بندی می‌شوند. دقت این روش‌ها به شدت تحت تأثیر ویژگی‌های استخراج شده از متن است. رایج‌ترین روش‌های یادگیری ماشین نظارت شده مانند بیز، ماشین بردار پشتیبان و رگرسیون لجستیک هستند[۱۱][۱۲]. روش‌های بدون نظارت، مناسب مواقعی هستند که تخصیص برچسب برای داده‌ها غیرممکن است[۱۳]. با پیشرفت تکنولوژی و روی کار آمدن شبکه‌های عصبی و یادگیری عمیق، استفاده از این روش‌ها به دلیل بهبود دقت افزایش پیدا کرده است [۱۴]؛ ترکیب یادگیری عمیق با یادگیری ماشین برای تجزیه و تحلیل احساسات نهفته در متن مورد استفاده قرار گرفته است که نتایج خوبی را به نمایش گذاشته است[۱۵]. علاوه بر تفاوت در نتیجه به دست آمده با شبکه عصبی نسبت به مابقی روش‌ها، تفاوت اصلی آن در مدل و طریقه استخراج ویژگی‌ها است. یکی از مزایای دیگر رویکردهای شبکه عصبی این است که نیازی به دخالت و ورود کاربر در استخراج ویژگی نیست و خودکار ویژگی‌ها استخراج می‌شوند. شبکه عصبی کانولوشنال[۴]، شبکه عصبی بازگشتی[۵] و حافظه کوتاه بلند مدت پرکاربردترین رویکردهای شبکه عصبی در تحلیل احساسات هستند[۱۶][۱۷]. از میان مدل‌های شبکه عصبی، پرکاربردترین آن شبکه عصبی کانولوشنال است. کیم در سال ۲۰۱۴ رویکرد استفاده از شبکه عصبی کانولوشنال برای تحلیل احساسات در سطح یک جمله را ارائه کرده‌است[۱۸] [۱۹]. باتوجه به گذر زمان پیشرفت‌های زیاد و چشمگیری روی بهبود عملکرد مدل‌های تجزیه و تحلیل احساسات مبتنی بر شبکه عصبی کانولوشنال انجام شده است.

در سال ۲۰۱۹ بوکایی نژاد و همکارانش،پژوهش‌هایی روی تحلیل احساسات انجام دادند که ثمره این پژوهش‌ها ارائه معماری یادگیری عمیق ترکیبی از CNN و LSTM شد. معماری پیشنهادی‌شان به این صورت می‌باشد که برای استخراج ویژگی از CNN و برای یادگیری در زمان بلند مدت از LSTM استفاده کرده‌اند. ترکیب این دو مدل با دقتی در حدود ۸۵ درصد می‌باشد.[۲۰].

در سال ۲۰۲۱ صابری و همکارانش به جمع آوری و برچسب گذاری توییت‌های شامل انگلیسی و فارسی مشغول شدند که دستاورد این کار، جمع‌آوری مجموعه داده شامل توییت‌های انگلیسی و فارسی شد. همچنین یک رویکردی با استفاده از مدل BERT از پیش آموزش داده شده ارائه دادند. دلیل استفاده از مدل BERT این است که به طور خودکار امتیازات قطبیت توئیت‌های مجموعه داده را یاد بگیرد. نتایج مدل پیشنهادی آن‌ها نشان داد که از مدل‌های پایه که از بیز و روش جنگلی تصادفی استفاده می‌کنند، بهتر عمل می‌کند [۲۱].

[4] Convolutional neural network (CNN)

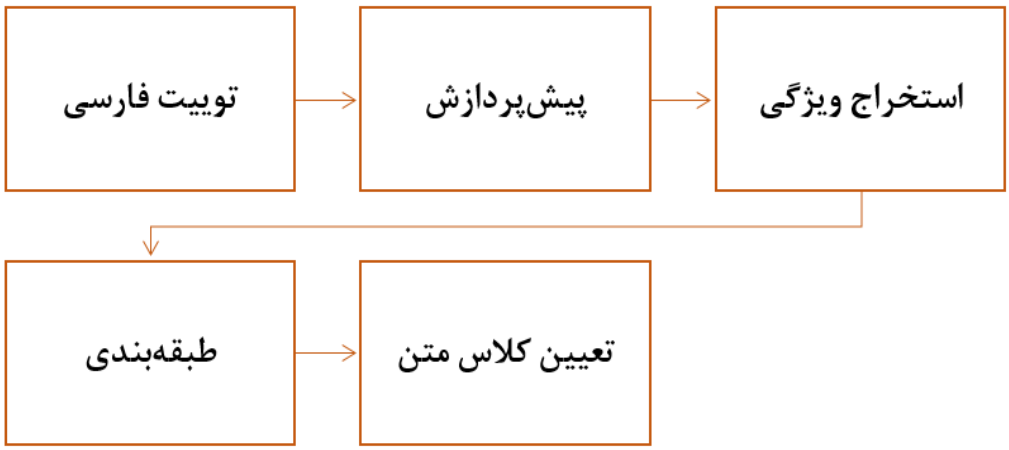
[5] Recurrent neural network (RNN)

در سال ۲۰۲۱ عثمان نسیم و همکارانش توییت‌هایی حاوی متن بیماری همه‌گیر کرونا را یک مجموعه داده به تعداد ۹۰ هزار توییت جمع‌آوری و برچسب گذاری کردند. این مجموعه داده شامل سه کلاس مثبت، منفی و خنثی می‌باشد و آن را مجموعه داده را تجزیه و تحلیل نموده‌اند که متوجه شدند نظر منفی در توییت‌ها نقش مهمی در شرطی کردن احساسات عمومی ایفا کرد[۲۲].

در سال ۲۰۲۱ تام و همکارانش یک معماری یکپارچه از CNN و مدل Bi-LSTM پیشنهاد دادند و مدل ConvBiLSTM پیاده‌سازی کردند. معماری این مدل شامل، یک مدل جاسازی کلمه که توییت‌ها را به مقادیر عددی تبدیل می‌کند، لایه CNN تعبیه ویژگی را به عنوان ورودی دریافت می‌کند و ابعاد کوچکتری از ویژگی‌ها را تولید می‌کند؛ مدل Bi-LSTM ورودی را از لایه CNN می‌گیرد و نتیجه طبقه‌بندی را تولید می‌کند. ConvBiLSTM با توییت‌های بازیابی شده و مجموعه داده‌های SST-2 اعمال شد. مدل ConvBiLSTM در با دقت ۹۱ درصد از مدل های دیگر بهتر عمل کرد[۲۳].

۳- روش پژوهش

روش پژوهش تحلیل احساسات کاربران از روی توییت‌های فارسی از بخش‌هایی که در شکل ۳-۱ نشان داده شده است، پیروی می‌کند. به طور بهتر نقشه راه روش پیشنهادی برای تحلیل احساسات توییت‌هایی فارسی می‌باشد.

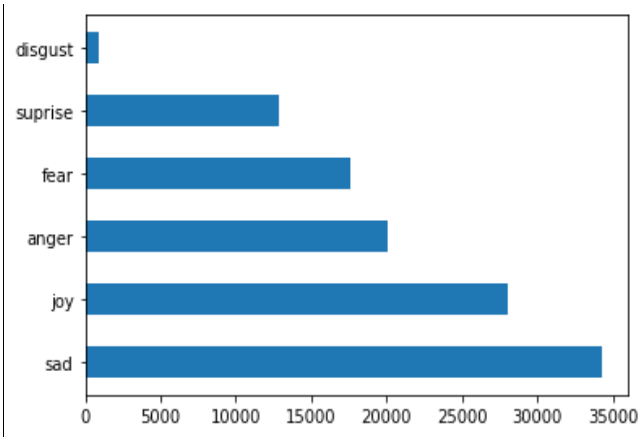


شکل ۳-۱- روش کلی پژوهش

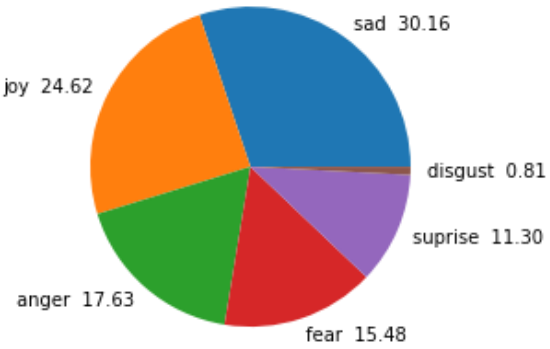
تحلیل و تشخیص احساسات در این روش دارای مراحل مختلفی است. در مرحله اول جمع‌آوری داده است که در این مرحله مجموعه داده توییت‌های فارسی برچسب گذاری شده، تشکیل شده است. در مرحله دوم باید متن توییت‌های فارسی را مورد پیش‌پردازش قرار گرفته شود تا یک داده خالص برای استخراج ویژگی وجود داشته باشد. بعداز مرحله پیش‌پردازش، مرحله استخراج ویژگی است. این ویژگی‌های استخراج شده برای طبقه‌بندی وارد مدل ماشین بردار پشتیبان می‌شود؛ در پایان کلاس توییت‌ها به دست می‌آید.

۳-۱- مجموعه داده

مجموعه داده استفاده شده برای روش پیشنهادی از سایت Kaggle می‌باشد[۲۴]، این مجموعه داده شامل ۱۱۳هزار و ۸۹۲ توییت فارسی است. همچنین هشت ستون اطلاعاتی از توییت‌ها دارد که ستون‌های tweet، replyCount، retweetCount، likeCount، quoteCount، hashtags، sourceLabel وemotion می‌باشد. این توییت‌ها توسط یک خزش‌گر وب از توییتر استخراج شده‌اند. همچنین توییت‌ها براساس شش احساس اصلی عصبانیت، ناراحتی، شادی، انزجار، تعجب برانگیز و ترس تقسیم‌بندی می‌شود. در شکل ۳-۱-۱ تعداد توییت‌های موجود در هر کلاس را مشاهده می‌کنید. همچنین شکل ۳-۱-۲ درصد اختصاص دهی نمونه‌ها به کلاس‌ها می‌باشد.



شکل ۳-۱-۱- تعداد توییت‌های موجود در هر کلاس



شکل ۳-۱-۲- درصد اختصاص‌دهی نمونه‌ها به کلاس‌ها

۳-۲- پیش‌پردازش

هنگام کار با متن در پردازش زبان طبیعی باتوجه به ساختار محاوره‌ای توییت‌ها مدل دچار مشکل می‌شود پس باید داده‌های خالصی را به ورودی مدل داد. به همین دلیل عمل پیش‌پردازش انجام می‌شود. پیش‌پردازش پاک‌سازی متون از ناخالصی می‌باشد. در ماهیت تحلیل احساسات وزن‌دهی به کلمات یک متن می‌باشد؛ پس باید متن را از کلماتی که هیچ بار معنایی برای تحلیل احساسات آن متن ندارد پاک‌سازی کرد. گام‌های پیش‌پردازش در زبان طبیعی به این صورت می‌باشد که در ابتدا باید متن را نرمال‌سازی کرد با این کار متن از حالت تصادفی خارج می‌شود و به متن استاندارد نزدیک می‌شود. در گام بعدی باید علائم نگارشی، ایموجی‌ها، اعداد را با مورد مناسبی جایگزین کرد. همچنین باید کلمات زائد را از متون حذف نمود.

۳-۳- استخراج ویژگی

اساسی‌ترین بخش در تحلیل احساسات کاربران از روی متون فارسی بخش استخراج ویژگی می‌باشد. دلیل اهمیت این بخش این است که مرحله استخراج ویژگی تاثیر مستقیمی بر روی کارایی الگوریتم‌های طبقه‌بندی دارد به همین خاطر ویژگی‌های استخراج شده باید توانایی جداسازی احساسات را از یکدیگر را داشته باشند. یکی از بهترین روش‌هایی که در استخراج ویژگی وجود دارد روش روابط معنایی می‌باشد. در این پژوهش برای استخراج ویژگی‌های متن از مدل FastText استفاده شده است.در این مدل برای استخراج ویژگی‌ها از روابط معنایی موجود در بین کلمات بهره می‌برد. به دلیل این که مدل FastText براساس روابط معنایی به هر کلمه وزن اختصاص می‌دهد. در این پژوهش برای هر کلمه یک بردار ۳۰۰ دارای وزن در نظر

گرفته می‌شود. بنابراین برای داشتن وزن یک عبارت از میانگین وزن‌ها استفاده می‌شود. باتوجه به این که یک بردار ۳۰۰ تایی از وزن برای هر توییت می‌باشد، این مقدار زیادی می‌باشد و حافظه زیادی اشغال می‌شود. پس باید ویژگی‌ها را کاهش داد، برای کاهش ویژگی‌ها از یک شبکه LSTM استفاده شده است. در این معماری از سه لایه استفاده شده است که لایه اول وظیفه الگویابی از بردار ویژگی‌ها، لایه دوم کاهش الگوها و لایه آخر هم وظیفه تبدیل الگوها به بردار را دارد. شکل ۱-۳-۳ ابرکلمات از متن توییت‌ها می‌باشد. به طور بهتر هر کلمه‌ای که به تعداد بیشتری در متن توییت‌ها تکرار شده باشد دارای وزن و بار معنایی بیشتری در آن توییت می‌باشد. با دقت به شکل ۳-۳-۱ می‌توان به آسانی متوجه شد که چه کلماتی بیشتر در متن توییت‌ها تکرار شده‌اند.

شکل ۳-۳-۱- ابرکلمات توییت‌ها

۳-۴- طبقه‌بندی

در این بخش هر توییت به یکی از کلاس‌های عصبانیت، ناراحتی، شادی، انزجار، تعجب برانگیز و ترس تخصیص داده می‌شود. برای طبقه‌بندی توییت‌ها از مدل ماشین بردار پشتیبان استفاده می‌شود. این مدل در سال‌های اخیر کارایی خوبی نسبت به

روش‌های قدیمی‌تر برای طبقه‌بندی نشان داده‌است. مبنای کاری SVM دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی می‌شود خطی را انتخاب کند که حاشیه اطمینان بیشتری داشته باشد. با فرض اینکه دسته‌ها به صورت خطی جداپذیر باشند، ابرصفحه‌هایی با حداکثر حاشیه[6] را به دست می‌آورد که دسته‌ها را جدا کنند.

۴- ارزیابی

در این پژوهش جهت ارزیابی روش پیشنهادی از معیارهای طبقه‌بندی استفاده می‌شود. این معیارها شامل معیارهای دقت، صحت، فراخوان و معیار F می‌باشد.

$$Accuracy=(TP+TN)/(TP+FP+FN+TN) \quad (1)$$

$$Precision=TP/(TP+FP) \quad (2)$$

$$Recall=TP/(TP+FN) \quad (3)$$

$$F1=(2*Precision*Recall)/(Precision+Recall) \quad (4)$$

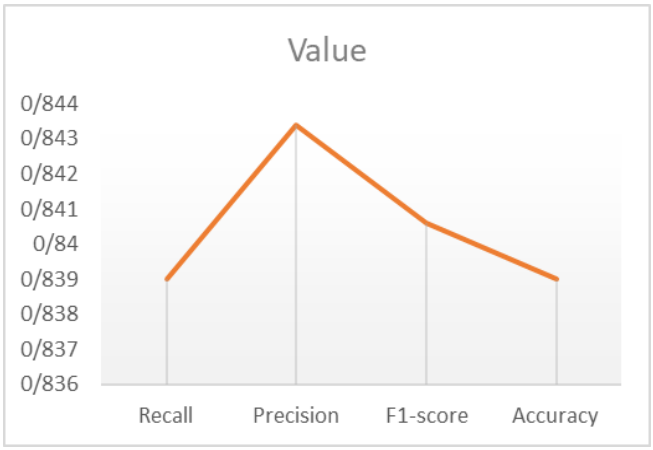
در روابط ذکر شده در بالا، نماد TP بیانگر تعداد نمونه‌هایی از مجموعه است که کلاس واقعی آن‌ها مثبت بوده و روش پیشنهادی نیز کلاس آن‌ها را به درستی مثبت تشخیص داده است. نماد FP بیانگر تعداد نمونه‌هایی است که کلاس واقعی آن‌ها منفی بوده ولی روش پیشنهادی به اشتباه کلاس آن را مثبت تشخیص داده است. نماد FN بیانگر تعداد نمونه‌هایی است که کلاس واقعی آن‌ها مثبت بوده ولی روش پیشنهادی به اشتباه کلاس آن‌ها را منفی تشخیص داده است. نماد TN بیانگر تعداد نمونه‌هایی است که کلاس واقعی آن‌ها منفی بوده و روش پیشنهادی نیز به درستی کلاس آن را منفی تشخیص داده است.

در این پژوهش برای تقسیم‌بندی داده‌های مجموعه داده مورد استفاده به دسته‌های آموزش و آزمایش از روش تصادفی بهره گرفته شده است. در روش تصادفی داده‌ها با نسبت ۸۰ به ۲۰ به ترتیب به داده‌های آموزشی و آزمایشی تقسیم‌بندی شده‌اند. در جدول ۴-۱ نتایج ارزیابی براساس معیارهای دقت، صحت، فراخوان و F1 آورده شده است.

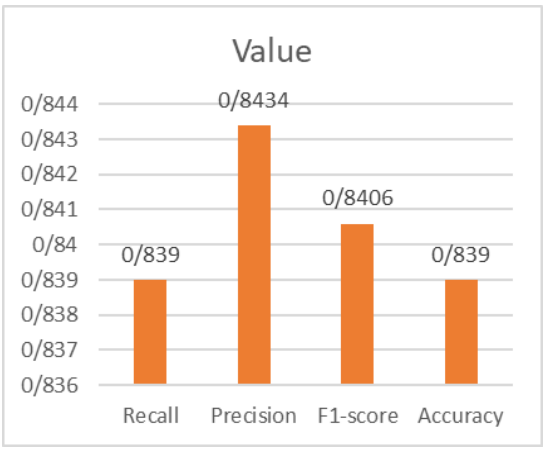
جدول۴-۱- نتایج ارزیابی مدل

دقت	صحت	فراخوان	F1
۸۳.۹	۸۴.۳	۸۳.۹	۸۴

[6] Maximum margin



شکل۴-۱- نمودار خطی نتایج



شکل۴-۲- نمودار میله‌ای نتایج

همانطور که در جدول ۴-۱ و اشکال ۴-۱ و ۴-۲ نشان داده شده است، دقت روش پیشنهادی ۸۳.۹ درصد می‌باشد که دقت قابل توجهی برروی توییت‌های فارسی می‌باشد. روش معنایی FastText از کارایی بالایی برخوردار است زیرا در این روش فقط خود کلمه از نظر وجود داشتن در متن بررسی نمی‌شود و بلکه بار و رابطه معنایی هرکلمه در متن با دیگر کلمات موجود در متن محاسبه می‌شود. همچنین در این روش وابستگی بین کلمات موجود در یک متن هم در نظر گرفته می‌شود. به دلیل اینکه بیشتر کار دیگران برروی مجموعه داده‌های انگلیسی در تحلیل احساسات چندکلاسه می‌باشد و همچنین کارهای انجام شده روی مجموعه داده‌های فارسی با تعداد کلاس‌های ثابت(مثبت، منفی و خنثی) می‌باشد، نمی‌توان با کارهای دیگران مقایسه کرد و می‌توان این پژوهش را پیشرو در این زمینه نام برد.

۵- نتیجه‌گیری

در این پژوهش تحلیل احساسات توییت‌های فارسی مبتنی بر ترکیب روابط معنایی و ماشین بردار پشتیبان می‌باشد. این روش پیشنهادی از یک فرآیند چند مرحله‌ای بهره گرفته شده است که در گام اول پیش‌پردازش توییت‌ها می‌باشد. در گام دوم عملیات استخراج ویژگی براساس روابط معنایی FastText انجام گرفته است. هدف از این بخش استخراج روابط معنایی بین کلمات و بررسی وجود وابستگی بین کلمات بوده است. در گام سوم عملیات کاهش ویژگی توسط معماری شبکه عصبی LSTM صورت گرفته و در نهایت در گام آخر طبقه‌بندی توییت‌ها به وسیله مدل ماشین بردار پشتیبان صورت گرفته شده است. ارزیابی‌ها نشان دهنده کاربردی بودن روش پیشنهادی و روش استخراج عملیات ویژگی براساس روابط معنایی و همچنین طبقه‌بندی ماشین بردار پشتیبان می‌باشد . معیارهای ارزیابی دقت، صحت، فراخوانی و F1 روش پیشنهادی به ترتیب برابر ۸۳.۹، ۸۴.۳، ۸۳.۹ و ۸۴ می‌باشد. نتایج آزمایش‌ها نشان‌دهنده کاربردی بودن روش پیشنهادی در تحلیل احساسات چندکلاسه توییت‌های فارسی به شش کلاس عصبانیت، ناراحتی، شادی، انزجار، تعجب‌برانگیز و ترس است.

۶- منابع

[1] S. A. Phand and J. A. Phand, "Twitter sentiment classification using stanford NLP," 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), 2017, pp. 1-5, doi: 10.1109/ICISIM.2017.8122138.

[2] R. Irfan et al., “A survey on text mining in social networks,” Knowl. Eng. Rev., vol. 30, no. 2, pp. 157–170, 2015.

[3] N. Oscar, P. A. Fox, R. Croucher, R. Wernick, J. Keune, and K. Hooker, “Machine learning, sentiment analysis, and tweets: An examination of Alzheimer’s disease stigma on Twitter,” Journals Gerontol. - Ser. B Psychol. Sci. Soc. Sci., vol. 72, no. 5, pp. 742–751, 2017.

[4] M. Bouazizi and T. Ohtsuki, "Multi-Class Sentiment Analysis in Twitter: What if Classification is Not the Answer," in IEEE Access, vol. 6, pp. 64486-64502, 2018, doi: 10.1109/ACCESS.2018.2876674.

[5] K. Schouten and F. Frasincar, "Survey on Aspect-Level Sentiment Analysis," in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 3, pp. 813-830, 1 March 2016, doi: 10.1109/TKDE.2015.2485209.

[6] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: LIWC and computerized text analysis methods,” J. Lang. Soc. Psychol., vol. 29, no. 1, pp. 24–54, 2010.

[7] A. Yadav and D. K. Vishwakarma, “Sentiment analysis using deep learning architectures: a review,” Artif. Intell. Rev., vol. 53, no. 6, pp. 4335–4385, Aug. 2020.

[8] Hu, Minqing and Bing Liu. “Mining and summarizing customer reviews.” Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (2004): n. pag.

[9] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-Based Methods for Sentiment Analysis DRAFT DRAFT DRAFT!,” Comput. Linguist. vol. 37, no. 2, pp. 267–307, 2011.

[10] T. Nasukawa and J. Yi, “Sentiment Analysis: Capturing Favorability Using Natural Language Processing,” in Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003). ACM, 2003, pp. 70–77.

[11] A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López, “Ranked WordNet graph for Sentiment Polarity Classification in Twitter,” Comput. Speech Lang., vol. 28, no. 1, pp. 93–107, 2014.

[12] Q. Ye, Z. Zhang, and R. Law, “Sentiment classification of online reviews to travel destinations by supervised machine learning approaches,” Expert Syst. Appl., vol. 36, no. 3 PART 2, pp. 6527–6535, 2009.

[13] G. Paltoglou and M. Thelwall, “Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media,” ACM Trans. Intell. Syst. Technol., vol. 3, no. 4, 2012.

[14] M. Dragoni and G. Petrucci, "A Neural Word Embeddings Approach for Multi-Domain Sentiment Analysis," in IEEE Transactions on Affective Computing, vol. 8, no. 4, pp. 457-470, 1 Oct.-Dec. 2017, doi: 10.1109/TAFFC.2017.2717879.

[15] Z. Jianqiang, G. Xiaolin and Z. Xuejun, "Deep Convolution Neural Networks for Twitter Sentiment Analysis," in IEEE Access, vol. 6, pp. 23253-23260, 2018, doi: 10.1109/ACCESS.2017.2776930.

[16] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov. vol. 8, no. 4, pp. 1–25, 2018.

[17] G. Xu, Y. Meng, X. Qiu, Z. Yu and X. Wu, "Sentiment Analysis of Comment Texts Based on BiLSTM," in IEEE Access, vol. 7, pp. 51522-51532, 2019, doi: 10.1109/ACCESS.2019.2909919.

[18] Y. Kim, “Convolutional neural networks for sentence classification,” EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., pp. 1746–1751, 2014.

[19] T. L. Nguyen, S. Kavuri, and M. Lee, “A fuzzy convolutional neural network for text sentiment analysis,” J. Intell. Fuzzy Syst., vol. 35, no. 6, pp. 6025–6034, 2018.

[20] Z. Bokaee Nezhad and M. A. Deihimi, “A COMBINED DEEP LEARNING MODEL FOR PERSIAN SENTIMENT ANALYSIS,” IIUM Eng. J., vol. 20, no. 1, pp. 129–139, Jun. 2019.

[21] N. Sabri, A. Edalat, and B. Bahrak, “Sentiment Analysis of Persian-English Code-mixed Texts,” 26th International Computer Conference, Computer Society of Iran, CSICC 2021. 2021.

[22] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund and J. Kim, "COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis," in IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 1003-1015, Aug. 2021, doi: 10.1109/TCSS.2021.3051189.

[23] S. Tam, R. B. Said and Ö. Ö. Tanriöver, "A ConvBiLSTM Deep Learning Model-Based Approach for Twitter Sentiment Classification," in IEEE Access, vol. 9, pp. 41283-41293, 2021, doi: 10.1109/ACCESS.2021.3064830.

[24] https://www.kaggle.com/datasets/behdadkarimi/persian-tweets-emotional-dataset

Sentiment analysis of Persian multi-class tweets based on a combination of semantic relations and support vector machine

Hossein Rayatparvar, MohammadAli Javadzade
Master's student of Imam Hossein (PBUH) Comprehensive University,
hosseinrayatparvar@ihu.ac.ir
Assistant Professor of Imam Hossein (PBUH) Comprehensive University,
javadzade@ihu.ac.ir

Abstract— With the development of technology, the use of social networks has become more popular. One of the most popular social networks is the Twitter platform. Sentiment analysis of users' tweets plays an important role in showing users' feelings about the existing conditions of society. In recent years, due to the fact that the text of users' tweets have become more conversational, sentiment analysis has become problematic and reduced its accuracy; It also makes natural language processing difficult. In this research, a method for sentiment analysis of Persian tweets based on the combination of semantic relations and support vector machine classification has been presented. FastText semantic relation is used to extract features. Considering that a large number of features have been extracted; They should be reduced, which is done by using long-short-term memory (LSTM) neural network. In the last part of the proposed method for the classification of sentiments in tweets, the support vector machine model is used. The evaluation criteria used in this research were precision, accuracy, recall and F criterion, and the evaluation results were 83.9, 84.3, 83.9 and 84, respectively. The results of the experiments show the applicability of the proposed method in analyzing the emotions of Persian tweets into six classes of anger, sad, joy, disgust, surprise and fear.

Key words: multi-class sentiment analysis, natural language processing, LSTM neural networks, support vector machine