دوفصلنامه مدیریت مهندسی و رایانش نرم/سال چهارم، شماره۱، بهار و تابستان ۱۳۹۷
صفحات ۷۶- ۵۹ (DOI): 10.22091/jemsc.2018.1272
JEMSC.QOM
مدیریت مهندسی و رایانش نرم ۱۳۹۴

ارائه رویکرد ترکیبی نوین جهت متن‌کاوی تحلیل احساسات در توییتر با استفاده از درخت تصمیم CART *

نصیر طیرانی نجاران [۱]
مهرداد جلالی [۲]

چکیده

با گسترش شبکه‌های اجتماعی به عنوان اجتماعات مجازی و استفاده روزافزون از آن‌ها، حجم انبوهی از نظرات کاربران در ارتباط با موضوعات مختلف پدید می‌آید. در نتیجه، به کارگیری تکنیک‌های علمی نوین جهت تحلیل این شبکه‌ها ضروری به نظر می‌رسد. متن‌کاوی به عنوان یک راهکار مؤثر، به دنبال کشف دانش از متون می‌باشد. در این مقاله، رویکردی نوین از ترکیب همزمان دو روش یادگیری ماشین و مبتنی بر واژگان جهت متن‌کاوی تحلیل احساسات در توییتر ارائه شده است. برای بهبود متن‌کاوی تحلیل احساسات، و دسته‌بندی داده‌ها از درخت تصمیم CART به عنوان روش یادگیری ماشین، و برای کاوش دقیق‌تر در نوع احساسات بیان شده در توییت‌ها از لیست الگوریتم SentiStrength به عنوان روش مبتنی بر واژه، استفاده شده است. ویژگی منحصر به فرد CART، تحلیل ساختار داده پیچیده است که با توجه به ورودی مسئله می‌تواند عملیات مربوط به رگرسیون، همچنین دسته‌بندی داده‌ها را انجام دهد. توانمندی الگوریتم SentiStrength در تشخیص احساسات، موجب تحلیل دقیق احساسات موجود در توییت‌ها گردیده است. نتایج پیاده‌سازی جهت تشخیص احساسات توییت‌ها، در اغلب شاخص‌ها بهبود دسته‌بندی را نشان می‌دهد.

واژه‌های کلیدی: *تحلیل احساسات، درخت تصمیم CART، شبکه‌های اجتماعی، لیست الگوریتم SentiStrength، متن‌کاوی*

---

* تاریخ دریافت: ۹۷/۷/۴؛ تاریخ پذیرش: ۹۷/۱۰/۲۹.

۱. کارشناس ارشد، مهندسی کامپیوتر نرم افزار، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد مشهد، مشهد، ایران (نویسنده مسئول)
nr.tayarani@gmail.com
۲. استادیار گروه مهندسی کامپیوتر نرم افزار، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد مشهد، مشهد، ایران
Jalali@mshdiau.ac.ir

مقدمه

تحلیل شبکه‌های اجتماعی[1] که گاهی به اختصار SNA هم گفته می‌شود به معنای فرآیند بررسی و ارزیابی ساختارهای یک شبکه اجتماعی به عنوان یک زمینه مهم تحقیقاتی در داده‌کاوی و استخراج اطلاعات مفید برای به دست آوردن یک الگوی رفتاری و بهره‌برداری از آن جهت کاربردهای علمی، تجاری، تحقیقاتی، توسعه و غیره می‌باشد. تحلیل شبکه‌های اجتماعی با استفاده از دانش ریاضی و نظریه گراف‌ها، شکلی بسیار علمی و ساختار یافته به خود گرفته است.

احساسات همواره از دیرباز جنبه مرموز و ناشناخته انسان‌ها بوده و جایگاه مهمی در حیات اجتماعی افراد دارا می‌باشد. پرداختن به احساسات و عواطف به ضرورتی بنیادین در تعاملات انسانی تبدیل شده و تحلیل رفتار کاربران بدون در نظر گرفتن احساسات و عواطف آن‌ها ناقص بوده و ارزش چندانی ندارد. بنابراین، اهمیت تحلیل احساسات کاربران شبکه‌های اجتماعی بیش از پیش نمایان شده است. وبسایت‌های شبکه‌های اجتماعی، پنجره‌ای رو به اجتماعات مجازی جدیدی می‌گشاید که در آن کاربران عقاید خود را به راحتی مطرح می‌کنند (بامر و سینکلار و تاملینسون، ۲۰۱۰). این وبسایت‌ها یک رسانه قدرتمند را برای ارتباطات و اشتراک عقاید فراهم می‌آورد (سورنسن، ۲۰۰۹). محبوب‌ترین این وبسایت‌ها توییتر، فیس‌بوک، مای‌اسپیس و لینکداین می‌باشند که در آن کاربران با پیوستن به اجتماعات مجازی به بحث و تبادل نظر با یکدیگر می‌پردازند (اوانز و کایرم و پیرولی، ۲۰۱۰). یکی از بزرگترین مزایای استفاده از شبکه‌های اجتماعی حذف فاصله جغرافیایی بین آن‌ها است (لی و لی و خان و قانی، ۲۰۱۱).

توییتر به عنوان یک شبکه اجتماعی فراگیر و در دسترس برای عموم، نقش بسیار پررنگی در بیان احساسات و عقاید طیف وسیعی از کاربران اینترنت و شبکه‌های اجتماعی در سالیان اخیر ایفا کرده است. به تازگی، تحلیل احساسات در توییتر به یکی از حیطه‌های

1 Social Network Analysis

جذاب تحقیقاتی تبدیل شده است. توییتر یکی از محبوب‌ترین بسترهای میکروبلاگ است که در آن کاربران می‌توانند افکار و عقاید خود را انتشار دهند. تحلیل احساسات در توییتر بر اساس تحلیل توییت‌ها بر مبنای عقاید مطرح شده در آن صورت می‌پذیرد. دسته‌بندی و تحقیق بر روی نظرات استخراج شده درباره یک موضوع مشترک، نتایج قابل توجهی را در بر خواهد داشت. علاوه بر این، فیلدهای مرتبط با موضوع تحلیل احساسات در توییتر مانند بازیابی عقاید در توییتر، دنبال کردن احساسات در طول زمان، تشخیص کنایه‌ها و طنز، تشخیص احساسات، تفسیر مفاهیم با توجه به موقعیت جغرافیایی کاربران و تعریف توییت احساسی نیز از اهداف تحلیل احساسات در توییتر است. موضوعی که امروزه مورد بحث قرار گرفته است این است که تحلیل بازخوردهای مربوط به احساسات منتشر شده در توییتر رابطه مستقیمی با موضوعات اجتماعی روز در جامعه دارد. بنابراین دقت در انتخاب جامعه هدف و کلمات احساسی مرتبط حتماً باید مد نظر باشد. در این بخش، یک توییت به عنوان نمونه جهت نشان دادن اجزاء موجودیت، ویژگی و بار احساسی یک توییت مورد بررسی مختصر قرار گرفته است تا بتوان با مفهوم موجودیت و بار احساسی در تحلیل احساسات در توییتر آشنا شد.

یک تعریف کامل از تحلیل احساسات که در (لیو و ژانگ، ۲۰۱۲) بیان شده است، احساسات را در قالب یک موجودیت معرفی می‌کند. یک عقیده به صورت یک پنج گانه $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ نمایش داده می‌شود که در آن e_i نام موجودیت، a_{ij} جنبه مورد اشاره به e_i، s_{ijkl} احساسات مربوط به جنبه a_{ij} از موجودیت e_i، h_k شخص نظر دهنده و t_l زمانی که نظر توسط h_k مطرح شده است را نشان می‌دهد. برای فهم ساده‌تر بخش‌های این پنج‌گانه، توییت زیر را که توسط کاربری به نام Helen در تاریخ ۲۰۱۵/۰۶/۱۰ منتشر شده است، تشریح می‌شود.

"The picture quality of my new Nikon V3 camera is great."

در این مثال، «Nikon V3» موجودیتی است که نظر بر آن اعمال شده، « picture quality» جنبه‌ای است که بر موجودیت اشاره می‌کند، احساسات مطرح شده در این

توییت «مثبت» است. شخصی که احساسات خود را نشان داده «Helen» می‌باشد و در نهایت زمان توییت «2015.06.10» است. پنج‌گانه‌ای که این نظر را نشان می‌دهد به صورت زیر می‌باشد که بعد از تحلیل احساسات بدست می‌آید.

(Nikon_V3, picture_quality, positive, Helen, 2015.06.10)

تشخیص احساسات در توییتر کار ساده‌ای به شمار نمی‌رود و تفاوت‌های قابل توجهی با تحلیل احساسات در دیگر متن‌های معمولی موجود در وب، مانند وبسایت‌ها، وب‌نوشت‌ها، اخبار و انجمن‌ها دارد (سرانو-گوئررو و اولیواس و رومر و هررا-ویدما، ۲۰۱۵). مهم‌ترین چالش‌های تحلیل احساسات در توییتر، در ادامه بررسی می‌گردند. یکی از اصلی‌ترین چالش‌های تحلیل توییت‌ها، طول متن (تعداد کاراکتر توییت) است. یکی از ویژگی‌های منحصر به فرد توییتر کوتاه بودن طول پیام‌های آن است که حداکثر می‌تواند ۲۸۰ کاراکتر باشد. این امر موجب می‌شود تا تحلیل احساسات در توییتر با تحلیل احساسات در دیگر بسترهای تحت وب مانند وبسایت‌ها و وبلاگ‌ها متفاوت باشد. چالش بعدی استفاده از زبان مخفف و اصطلاحات خلاصه شده می‌باشد. با توجه به محدودیت طول متن و غیررسمی بودن ارتباطات در توییتر، کاربران از مختصات متنی خاصی استفاده می‌کنند؛ مانند تأکید بینهایت، تأکید طولی، اختصارنویسی، همچنین اصطلاحات عامیانه و واژه‌های جدید (جیاچانو و کریستانی، ۲۰۱۶). از دیگر چالش‌های عمده در تحلیل احساسات در توییتر می‌توان به محتوای چندزبانه[1] اشاره کرد. محتوای چندزبانه به این شکل بیان می‌شود که توییت‌ها در طیف وسیعی از زبان‌ها نوشته می‌شوند، و گاهی اوقات نیز چند زبان با هم ترکیب می‌شوند (مانند حالت به اصطلاح فینگلیش که زبان فارسی را با حروف انگلیسی می‌نویسند). سختی تشخیص یک زبان به علت کوتاه بودن برخی از این توییت‌ها افزایش می‌یابد.

1. Multilingual Content

هدف تحلیل احساسات در شبکه‌های اجتماعی، استفاده از الگوریتم‌های کامپیوتری به منظور استخراج احساسات از داده‌های بزرگ[1] موجود در وب می‌باشد. برای این منظور از تکنیک‌های یادگیری ماشین برای دسته‌بندی احساسات، همچنین از روش‌های مبتنی بر واژه برای تحلیل دقیق‌تر احساسات دسته‌بندی شده استفاده می‌گردد.

پیشینه پژوهش

به دلیل محدودیت طول متن در توییتر، توییت‌ها عمدتاً شامل یک جمله هستند در نتیجه تحلیل احساسات در توییتر فقط در دو سطح جمله و موجودیت بررسی می‌شود. این روش‌ها به طور کلی در چهار کلاس یادگیری ماشین، مبتنی بر واژه، مبتنی بر گراف و ترکیبی معرفی می‌شوند که به عنوان «کلاس‌بندی تحلیل احساسات در توییتر» مشخص شده‌اند.

یادگیری ماشین

رویکردهای یادگیری ماشین با به کارگیری روش‌های یادگیری ماشین و چند ویژگی متفاوت دیگر، یک کلاس‌بند می‌سازند تا بتوانند توییت‌هایی که یک ایده یا یک احساس خاص را بیان کرده‌اند، شناسایی نمایند.

مرجع (کارالامپاکیس و اسپاثیس و کوئیزلیس و کرمانتیس، ۲۰۱۶) یک طرح طبقه‌بندی برای تشخیص طنزپردازی در توییت‌های سیاسی را بررسی می‌نماید. فرضیه‌ای مطرح می‌شود که ثابت می‌کند توییت‌های سیاسی طنزآمیز می‌تواند نتایج انتخابات را پیش‌بینی کند. مفهوم تشخیص طنز بر پایه برداشت‌های ذهنی است لذا صرفاً اتکا به روش‌های انسانی نمی‌تواند یک راهکار مناسب باشد. روش پیشنهادی بر پایه داده‌های محدودی است که برچسب‌گذاری شده‌اند و از راهکاری نیمه نظارتی استفاده می‌کند. این راهکار از روش‌های یادگیری جمعی استفاده کرده و شامل داده‌های برچسب‌گذاری شده

1. Big Data

و برچسب‌گذاری نشده می‌باشد. نتایج دریافت شده از روش نیمه نظارتی با روش‌های قبلی که از راهکارهای نظارتی استفاده می‌کردند، مقایسه شده‌اند. رویکرد اصلی این روش، شناسایی طنز به وسیله کلاس‌بندی متون می‌باشد و نحوه مشخص شدن طنز بودن یا نبودن یک متن کاملاً دودویی است. در این روش به هر توییت ۵ ویژگی بر اساس ساختار جمله تخصیص داده شده است. بعضی از این ویژگی‌ها برای تشخیص عدم توازن و برخی دیگر برای موارد غیرمنتظره طراحی شده‌اند. ویژگی‌های دیگری نیز برای یافتن الگوهای رایج در ساختار توییت‌های طنز طراحی شده‌اند (مانند نوع نقطه گذاری، شکلک‌ها یا طول متن). نتایج این مقاله نشان می‌دهد دقت این روش در الگوریتم‌های با ناظر ۸۲ درصد، و نیز امتیاز «F» آن ۷۹ درصد می‌باشد، همچنین در الگوریتم‌های نیمه نظارتی میزان دقت ۸۳ درصد و امتیاز «F» در آن ۷۴ درصد بوده است.

مبتنی بر واژه

رویکردهای مبتنی بر واژه نیز از یک لیست از پیش ساخته شده دستی یا خودکار استفاده می‌کنند تا واژگان مثبت یا منفی در یک توییت را تشخیص داده و قطبیت پیام را شناسایی کند.

مرجع (دلوال و باکلی و پاتوگلو و کای و کاپاس، ۲۰۱۰) به عنوان یکی از معروف‌ترین و قدرتمندترین رویکردهای مبتنی بر واژه معرفی شده است. الگوریتم‌هایی که به شناسایی احساسات و قدرت احساسات می‌پردازند، نیازمند درک نقش شکلک‌ها در ارتباطات غیر رسمی هستند. در مرجع (دلوال و باکلی و پاتوگلو و کای و کاپاس، ۲۰۱۰) بیانی برای حل این مشکل پیشنهاد شده است. الگوریتم SentiStrength برای استخراج قدرت احساسات از متن‌های غیر رسمی انگلیسی با استفاده از یک روش جدید بر اساس گرامر و ساختار هجی ارائه شده است. SentiStrength با استفاده از چند روش نوین به شکل همزمان، به استخراج احساسات مثبت و منفی می‌پردازد. SentiStrength از یک واژه‌نامه کلمات احساسی به همراه یک مقیاس قدرت احساسی به شکل عددی، استفاده

می‌کند. این الگوریتم از یک واژه‌نامه کامل از کلمات احساساتی تشکیل شده است که به هر کلمه احساسی بسته به میزان شدت احساسات آن لغت، به آن امتیاز ۵+ الی ۵- می‌دهد. یعنی به هر واژه امتیازی مابین ۵+ (بیشترین امتیاز برای احساسات مثبت) الی ۵- (بیشترین امتیاز برای احساسات منفی) اختصاص می‌دهد. الگوریتم SentiStrength از نظرات شبکه MySpace توسعه پیدا کرده است. رویکرد این مقاله بهبود و توسعه نوین تحلیل احساسات نسبت به رویکردهای یادگیری ماشین برای بهینه‌سازی وزن‌دهی به احساسات، روش‌هایی برای استخراج احساسات از کلمات غیراستاندارد در متن‌های غیررسمی و یک روش اصلاح ساختار کلمات می‌باشد. هسته اصلی این الگوریتم لیست کلمات احساسی بر مبنای قدرت آن‌هاست که شامل ۲۹۸ کلمه مثبت، و ۴۶۵ کلمه منفی، از قدرت ۲ الی ۵ می‌باشد. سپس قدرت تشخیص لیست مذکور به وسیله یک الگوریتم آموزش‌پذیر بهبود پیدا کرده است. همچنین این الگوریتم چندین بار تمام کلمات را بررسی کرده، و تا زمانی که میزان احساسات هیچ کلمه‌ای تغییر نکرده باشد، این روال را ادامه می‌دهد. به عنوان مثال برای استفاده از شکلک قلب (♥) قدرت احساسی کمتری نسبت به استفاده از کلمه «*Love*» در نظر گرفته است (به این علت که استفاده از شکلک نسبت به کلمه احساسی بسیار رایج‌تر است). همچنین به کلمه «Miss» احساس مثبت و منفی با قدرت ۲ تخصیص داده شده است؛ زیرا با وجود این که این کلمه به طور ذاتی یک واژه منفی است اما می‌تواند بار احساسی مثبت نیز داشته باشد، به عنوان مثال در جمله « *I Miss You!* » هم بار احساسی غمگین و منفی و هم بار احساسی عاشقانه و مثبت نیز دارد. یک الگوریتم اصلاح هجی نیز برای اصلاح کلمات کشیده و کلماتی که یک یا چند حرف آن حذف شده‌اند، استفاده شده است. برای مثال کلمه «*Hello*» هم به صورت «*Helllooo*» و هم به صورت «*Hli*» نوشته می‌شود. همچنین، یک لیست نیز وجود دارد از کلمات تشدید کننده که احساسات یک کلمه منفی یا مثبت را زیاد می‌کند. مانند کلمه «*Very*» که بار مثبت کلمه را یک واحد زیاد می‌کند و کلمه «*Some*» که از بار منفی کلمه یک واحد می‌کاهد. یک لیست از کلمات نفی نیز وجود دارد که بار احساسی جمله را معکوس می‌کند. به عنوان مثال کلمه

«*Very Happy*» بار احساسی مثبت با قدرت ۴ دارد که کلمه نفی «*Not*» بار احساسی آن را منفی با قدرت ۴ می‌کند. همچنین حروف تکرارشونده نیز در بار احساسی مثبت و منفی کلمه به میزان یک واحد تأثیر دارد. برای مثال بار احساسی کلمه «*Niice*» یک واحد تأثیر مثبت بیشتری نسبت به کلمه «*Nice*» دارد. در کلمات سوالی نیز تأثیر وجود یک کلمه منفی در نظر گرفته نشده است. به عنوان مثال در مورد سوال «*Are you Angry?*» با وجود این که کلمه احساسی «Angry» را در خود دارد اما به علت سوالی بودن جمله بار احساسی منفی آن حذف شده است. در نهایت با اضافه شدن موارد فوق به الگوریتم SentiStrength دقت کلاس‌بند به ۶۰/۶ درصد در تشخیص کلمات مثبت و ۷۳/۵ درصد در کلمات منفی رسیده است که در مقایسه با الگوریتم‌های پرکاربرد یادگیری ماشین و کلاس‌بندی از جمله SVM، J48 و NB از دقت بیشتری برخوردار می‌باشد.

مبتنی بر گراف

در مقابل روش‌های یادگیری ماشین و مبتنی بر واژه که به هر نوع از متنی اعمال می‌شوند، روش‌های مبتنی بر گراف به دنبال یافتن راهکاری برای پیدا کردن روابط بین کاربران در شبکه‌های اجتماعی هستند. رویکردهای مبتنی بر گراف با تشکیل گرافی از ارتباط کاربران و کاوش در ویژگی‌های شبکه‌های اجتماعی سعی در بالا بردن کارایی تحلیل احساسات در توییتر دارند.

مرجع (لوئترو-اوترو و کوردرو-گوتیئرز، ۲۰۱۶) به بررسی گراف نفوذ کاربران در توییتر برای کشف مشخصات توییت‌ها توسط PIAR می‌باشد. PIAR یک ابزار تحقیقاتی داده‌کاوی مخصوص است که توسط دانشگاه سالامانکای اسپانیا معرفی گردیده است. در واقع این ابزار، تلفیقی از تئوری گراف و تئوری نفوذ می‌باشد. در این مدل پیشنهادی، مقایسه ضریب نفوذ، بر مبنای استفاده از برچسب‌های بیشتر و تعداد کلمات کمتر در توییت‌ها، بوده است. نتایج منتشر شده نشان می‌دهد چگونه افراد تأثیرگذار در شبکه‌های اجتماعی نظرات خود را ابراز کرده‌اند و چه رفتاری از خود نشان می‌دهند. همچنین بیانگر

این است که توییت‌های این افراد چه تأثیری بر رفتار و احساسات جمعی دارد. یکی از زمینه‌های تحقیقاتی در حوزه میکروبلاگ‌ها، برندسازی و بازاریابی ویروسی است که به آن بازاریابی شبکه‌های اجتماعی یا اصطلاحاً e-WoM [1] گفته می‌شود. در این رویکرد پنج فرضیه برای شناسایی افراد بانفوذ و اثرگذار در توییتر مطرح می‌شود که می‌تواند گراف بزرگی از کاربران را تشکیل دهند و بر عقاید آن‌ها تأثیر بگذارند. پنج فرضیه برای تشخیص افراد پرنفوذ ارائه شده است. در نهایت برای محاسبه ضریب نفوذ یک شخص در توییتر از ابزار Klout استفاده می‌شود. Klout یک ابزار شناخته شده در این زمینه است که از تمام شبکه‌های اجتماعی برای امتیازدهی به میزان نفوذ یک شخص استفاده می‌کند. این امتیاز به شکل عددی بین ۰ تا ۱۰۰ داده می‌شود. از ابزار PIAR استفاده شده تا به کمک تئوری گراف، یک جامعه مجازی پیرامون موضوعات مربوط به دو شرکت خودروسازی مطرح ژاپنی ایجاد نماید. زمانی که یک کاربر توییتی در این زمینه را بازتوییت نموده و یا به آن واکنش نشان می‌دهد، این ابزار یک لینک ارتباطی بین این دو برقرار می‌نماید. در این رویکرد تمام توییت‌هایی که در آن‌ها کلمه تویوتا و یا نیسان وجود داشتند، استخراج گردیده است. در مجموع از ۳۸۵۳ کاربر توییتر که در بازه زمانی ۱۳ الی ۲۵ آوریل سال ۲۰۱۵ توییت کرده بودند، بیش از ۳۰۰۰۰ توییت استخراج شد. نتایج نشان می‌دهد که فرضیات مطرح شده، در تشخیص افراد بانفوذ کاملاً موثر عمل کرده است.

ترکیبی

روش‌هایی با رویکرد ترکیبی نیز وجود دارند که با به کارگیری روش یادگیری ماشین و رویکرد مبتنی بر واژه قصد دارند به کارایی بالاتری در تحلیل احساسات دست پیدا کنند. محققین زیادی در سالیان اخیر روش‌های یادگیری ماشین و رویکرد مبتنی بر واژه را ترکیب کرده‌اند.

1. Electronic Word-Of-Mouth

مرجع (سیف و فرناندز و آلانی، ۲۰۱۶) الگوریتم SentiCircles را معرفی می‌کند که یک تحلیلگر ترکیبی احساسات در توییتر است. بر خلاف رویکردهای عمومی که از روش‌های ایستا استفاده می‌کنند و برای هر کلمه قطبیتی مشخص دارند، SentiCircles با استفاده از یک الگوی رخداد کلمات در متن توییت‌ها، احساسات آن‌ها را تشخیص داده، قدرت و قطبیت هر کلمه را مشخص می‌کند. نتایج نشان داده که این روش توانسته شاخص‌های دقت و امتیاز «F» را در زمینه تشخیص قطبیت احساسات بهبود ببخشد. مزیت این روش شامل، معرفی یک رویکرد ترکیبی جدید با به کارگیری SVM به عنوان روش یادگیری ماشین و رویکرد مبتنی بر واژگان استفاده شده است. این الگوریتم با استفاده از نمایش متنی کلمات، به نام SentiCircles نامگذاری شده است. این رویکرد قابلیت گرفتن معنای پنهان کلمات را با توجه به الگوی رخداد آن‌ها دارد. از مجموعه داده STS-Gold، که یک مجموعه داده برای مقایسه تحلیل احساسات می‌باشد به عنوان مجموعه داده پایه و بدون عنوان استفاده گردیده است.

نتایج پیاده سازی در دو سطح موجودیت و توییت بررسی شده است، که در هر سطح به کشف قطبیت احساسات در توییت می‌پردازد. در سطح موجودیت‌ها، نتایج در قسمت تشخیص قطبیت، بهبود ۱/۵ درصدی در امتیاز «F» و ۲/۵ درصدی در دقت، و همچنین در سطح توییت، و در بخش تشخیص قطبیت بهبود ۴ درصدی در امتیاز «F» و ۳ درصدی در دقت را نشان می‌دهد.

در مرجع (ژانگ و گوش و دهیل و هسو و لیو، ۲۰۱۱) یک روش ترکیبی برای تحلیل احساسات در توییتر ارائه شده است. در این روش به هر موجودیت بر اساس شباهت به کلمه و احساساتی که نسبت به آن توییت شده است، یک امتیاز تخصیص می‌دهد. در این مقاله یک الگوریتم مبتنی بر قانون[1] استفاده شده است که قضاوت، نفی و اصطلاحات را نیز در مقایسه‌ها در نظر می‌گیرد. علاوه بر این از مربع کای[2] برای افزایش داده‌های

1. Rule-Based
2. Chi-Square

مرتبط و بهبود میزان بازخوانی بهره گرفته شده است. در انتها از کلاس‌بند SVM برای تشخیص و تعیین میزان قطبیت احساسات استفاده گردیده است.

یک روش جالب دیگر در (قیاسی و اسکینر و زیمبرا، ۲۰۱۳) ارائه شده است. در این روش شبکه‌های عصبی مصنوعی پویا با n-gram ترکیب شده است. هدف آن شناسایی شکلک‌ها و توییت‌هایی است که شامل کلمات «*Love*» و «*Hate*»، و مترادف‌های آنان می‌باشد. این هدف با به کارگیری کلاس‌بند SVM و روش شبکه‌های عصبی مصنوعی پویا دنبال می‌گردد. نتایج مقایسه نشان داده است که روش شبکه‌های عصبی مصنوعی پویا از SVM کارایی بهتری دارد.

روش‌شناسی پژوهش

تحلیل احساسات با استفاده از CART

در این بخش به معرفی رویکرد پیشنهادی و بیان مراحل آن پرداخته می‌شود. ابتدا موضوع پیش‌پردازش داده‌ها و اهمیت آن در پردازش نهایی بیان می‌گردد. سپس روش‌ها و الگوریتم‌های مورد استفاده جهت متن‌کاوی تحلیل احساسات توضیح داده می‌شوند. در انتها جزئیات مربوط به روش پیشنهادی به صورت کامل معرفی شده و فازهای عملیاتی آن نیز به همراه شبه کد مربوطه، ارائه می‌گردند.

عملیات پیش‌پردازش توییت‌ها

هدف از انجام پیش‌پردازش داده‌ها، به دست آوردن ویژگی‌های مناسب از آن‌ها می‌باشد. به منظور تغییر شکل در داده‌های موجود، پردازش‌هایی بر روی توییت‌ها انجام می‌پذیرد. پیش‌پردازش در متن‌کاوی توییتر نیز بسیار حائز اهمیت است زیرا از حجم بسیار زیاد توییت‌هایی که مرتبط نیستند می‌کاهد و به افزایش دقت تحلیل احساسات در یک موضوع خاص کمک شایانی می‌کند. پیش‌پردازش شامل فرآیند تمیز کردن و آماده‌سازی متن توییت‌ها به منظور انجام پردازش نهایی بر روی آن‌ها می‌باشد. متون الکترونیکی از جمله توییتر شامل مقداری زیادی نویز، کلمات بی‌ارزش، کلمات واسط، سمبل‌ها و نشانه‌ها

و حتی تبلیغات است که پاکسازی و پیش‌پردازش آن‌ها تأثیر مستقیمی روی ورودی‌های سیستم پردازشی جهت دسته‌بندی یا طبقه‌بندی داده‌ها خواهند داشت. همچنین در سطح طبقه‌بندی کلمه، بسیاری از کلمات جهت‌گیری معنایی خاصی ندارند که نگه داشتن این کلمات باعث افزایش ابعاد کار و در نتیجه اثرگذاری روی صحت و دقت پردازش نهایی می‌گردند. به منظور کسب اطلاعات مناسب، نگهداری کلمات صحیح و رسیدن به حداقل نویز در داده‌ها، باید از روش‌های پیش‌پردازش صحیح استفاده نمود. ویژگی‌ها در زمینه تجزیه و تحلیل احساسات در توییتر شامل کلمات، عبارات و یا اصطلاحات می‌باشند که به شدت بر جهت‌گیری نظرات و عقاید به سمت مثبت و منفی اثر می‌گذارد و به شکل کلی تأثیر بیشتری روی احساسات کل متن خواهند گذاشت.

روال تعیین توییت‌های مرتبط

پس از انجام مرحله پیش‌پردازش بر روی حجم زیادی از توییت‌ها در مجموعه داده، توییت‌های قابل استفاده جهت متن‌کاوی به دست می‌آیند. مرحله بعدی جداسازی توییت‌های مرتبط (با موضوع مربوطه) از بین توییت‌های موجود در مجموعه داده است. برای دسته‌بندی علاقه‌مندی موضوعات مرتبط با رشته مهندسی کامپیوتر، گرایش‌های این رشته در ۴ دسته پیشنهادی مهندسی نرم‌افزار، هوش مصنوعی، برنامه‌نویسی و شبکه‌های کامپیوتری مورد بررسی قرار گرفته است. در ادامه برای هر کدام از این گرایش‌ها تعدادی زیردسته (حدود ۶۰ زیردسته) در نظر گرفته شده است. این زیردسته‌ها با مطالعه روی کلید واژه‌های موجود در مقالات، بررسی چارت‌های دانشگاهی و اطلاعات عمومی راجع به رشته مهندسی کامپیوتر به دست آمده است. با کمک لیست تهیه شده از علاقه‌مندی‌های مربوط به رشته مهندسی کامپیوتر، استخراج توییت‌های مرتبط با آن، از مجموعه کل توییت‌ها صورت می‌گیرد. در مرحله بعد با استفاده درخت تصمیم CART، تحلیل مربوط به طبقه‌بندی احساسات بر روی توییت‌های انتخابی انجام گرفته و توییت‌ها با احساسات مشابه دسته‌بندی می‌گردند. در واقع روش ترکیبی با استفاده همزمان از هر دو روش یادگیری ماشین و روش مبتنی بر واژه به تشخیص احساسات درون توییت‌ها و دسته‌بندی آن‌ها می‌پردازند.

فازبندی الگوریتم ترکیبی پیشنهادی جهت متن‌کاوی تحلیل احساسات در توییتر به شرح زیر است.

- فاز اول، پیش‌پردازش: به علت حجم بالای توییت‌های موجود در مجموعه داده، همچنین گستردگی طیف زبان‌ها، نظرات و جملات داخل توییت‌ها، انجام یک پیش‌پردازش در جهت بالابردن دقت تحلیل احساسات الزامی است.
- فاز دوم، دسته‌بندی توییت‌های مرتبط: هدف تحلیل احساسات از این مجموعه، تعیین احساسات مربوط به رشته مهندسی کامپیوتر می‌باشد که با تهیه لیست علاقه‌مندی‌های مربوط به آن، توییت‌هایی مرتبط با این رشته از مجموعه کل توییت‌ها استخراج می‌گردد.
- فاز سوم، تحلیل احساسات: پس از انجام دو فاز فشرده اکنون مجموعه داده خالص از توییت‌های مورد نظر به دست آمده است. روش ترکیبی به کمک درخت تصمیم CART و مجموعه کلمات لیست الگوریتم SentiStrength به تحلیل و طبقه‌بندی احساسات در توییت‌ها می‌پردازند.

شکل (۱) شبه کد روش پیشنهادی را نشان می‌دهد.

1	**Phase 1: Preprocessing**
2	**Input**: Dataset
3	**Output**: Optimize Tweets
4	Tweet Tokenization
5	Stop Words Removal
6	Symbol Removal
7	Spell Correction
8	Stemming
9	**Phase 2: Classification**
10	**Input**: Optimize Tweets
11	**Output**: Computer Tweets
12	Classification Tweets With Computer List
13	**Phase 3: Sentiment Analysis**
14	**Input**: Computer Tweets
15	**Output**: Sentiment Analysis Tweets
16	Sentiment Analysis With CART Decision Tree By Using SentiStrength Algorithm List

شکل ۱. تحلیل احساسات با استفاده از درخت تصمیم CART

همان‌طور که بیان شد، پس از انجام این سه فاز، تحلیل احساسات (مرتبط با رشته مهندسی کامپیوتر) از درون توییت‌ها صورت می‌پذیرد.

شاخص‌های ارزیابی تحلیل احساسات

تحلیل احساسات در توییتر را می‌توان به عنوان یک مسئله کلاس‌بندی در نظر گرفت که هدف اصلی آن، کلاس‌بندی عقاید و احساسات منتشر شده در توییتر به دو دسته مثبت و منفی است. برای تحلیل احساسات شاخص‌های ارزیابی متداول به صورت صحت (Accuracy)/ دقت (Precision)/ بازخوانی (Recall)/ و امتیاز «F» (F-Score) می‌باشد. همچنین امتیاز «F»، یک نوع میانگین بین پارامتر P (دقت) و پارامتر R (بازخوانی) است. P میزان دقت سیستم در میان دادهای پیش‌بینی شده و R نسبت تعداد داده‌های پیش بینی شده، به تعداد کل داده‌های مورد انتظار برای پیش‌بینی می‌باشد. مقدار امتیاز «F» را می‌توان از طریق رابطه (۱) به‌دست آورد.

رابطه ۱)
$$F - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

دقت و بازخوانی شاخص‌های کاربردی در حوزه بازیابی اطلاعات هستند که میزان تناسب اسناد بازیابی شده توسط سیستم را با نیاز کاربر تعیین می‌کنند. این شاخص‌ها به صورت زیر تعریف می‌شوند.

- دقت = تعداد اسناد بازیابی شده مرتبط / تعداد کل اسناد بازیابی شده
- بازخوانی = تعداد اسناد بازیابی شده مرتبط / تعداد کل اسناد مرتبط واقعی در پایگاه داده

یافته‌های پژوهش

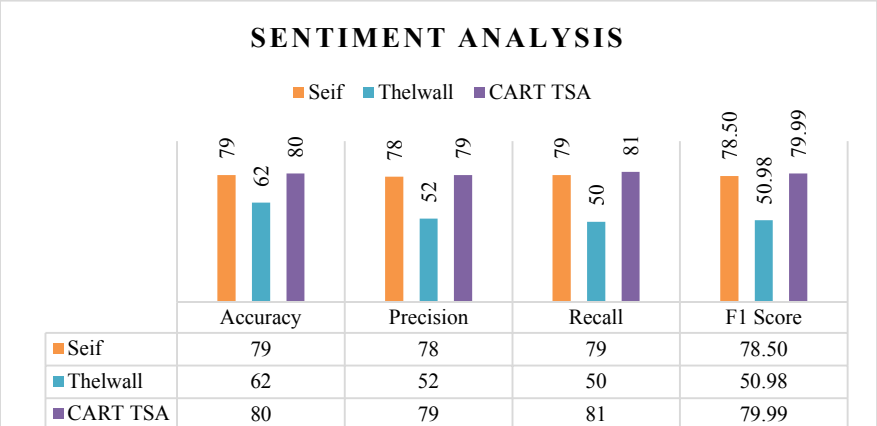
برای مقایسه پیاده‌سازی روش پیشنهادی از STS-Gold که مجموعه داده استاندارد توییتری می‌باشد، استفاده شده است.

ارزیابی و مقایسه روش پیشنهادی

از آنجا که رویکرد پیشنهادی به شکل ترکیبی به تحلیل احساسات می‌پردازد، به منظور مقایسه نتایج پیاده‌سازی روش پیشنهادی با مقالات مشابه، از (دلوال و باکلی و پاتوگلو و کای و کاپاس، ۲۰۱۰) و (سیف و فرناندز و آلانی، ۲۰۱۶) به عنوان مقالات پایه استفاده شده است.

عملکرد روش پیشنهادی در زمینه تشخیص قطبیت توییت‌ها

عملکرد حاصل از پیاده‌سازی روش پیشنهادی، در زمینه تشخیص وجود قطبیت در توییت‌ها سنجیده شده است، یعنی کارایی روش پیشنهادی در تشخیص وجود قطبیت در توییت‌ها. شاخص‌های ارزیابی حاصل از اجرای روش پیشنهادی در زمینه تشخیص قطبیت توییت‌ها به همراه جدول و نمودار مقایسه‌ای با روش‌های پایه در شکل (۱) نمایش داده شده است. بهبود در شاخص‌های دقت، صحت و به طبع آن بازخوانی در نتایج به دست آمده است. میزان بالای شاخص بازخوانی به دست آمده، بیانگر اعتماد بیشتر به نتایج است. همچنین شاخص امتیاز «F» بهبود قابل قبولی را در نتایج نشان می‌دهد.



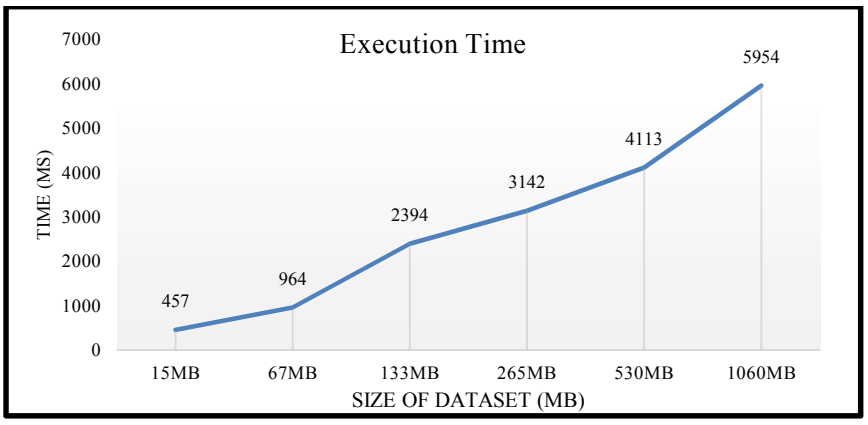
	Accuracy	Precision	Recall	F1 Score
Seif	79	78	79	78.50
Thelwall	62	52	50	50.98
CART TSA	80	79	81	79.99

شکل ۱. جدول و نمودار مقایسه‌ای تحلیل احساسات

تحلیل زمان اجرا

تحلیل زمان اجرای یک برنامه همواره معیاری برای مقایسه میزان پیچیدگی روش پیشنهادی و بررسی میزان مقیاس‌پذیری الگوریتم‌ها هنگام کار با حجم داده‌های مختلف است. در این بخش نیز، روش پیشنهادی از منظر زمان اجرای برنامه مورد تحلیل قرار گرفته است. در واقع در مبحث داده‌کاوی تحلیل احساسات، زمان اجرا به منظور بیان مقیاس‌پذیری الگوریتم ارائه می‌گردد. هدف از ارائه زمان اجرای روش پیشنهادی نیز، بیان مقایس‌پذیر بودن آن است و همان‌طور که در شکل (۲) مشخص است زمان اجرای برنامه با توجه به افزایش حجم مجموعه داده افزایش نیافته است یعنی نرخ رشد قابل قبولی را نشان می‌دهد که این مهم بیانگر این موضوع می‌باشد.

البته لازم به ذکر است که به علت گستردگی موضوعات موجود در متن‌کاوی تحلیل احساسات و همچنین نوین بودن این زمینه تحقیقاتی، زمان اجرای الگوریتم صرفاً در راستای بیان مقیاس‌پذیری روش پیشنهادی به ویژه در زمینه تحلیل داده‌های بزرگ بیان می‌گردد. مقایسه حجم مجموعه داده و زمان اجرا به منظور متن‌کاوی تحلیل احساسات در توییتر در شکل (۲) بیان شده است.



شکل ۲. تحلیل زمان اجرا

نتیجه‌گیری و پیشنهادها

این مقاله با هدف تجزیه و تحلیل احساسات در توییتر، تهیه شده است. در این مقاله از مجموعه داده از توییت‌های بدون عنوان استفاده شده است. جهت متن‌کاوی در این مجموعه داده، روشی پیشنهاد گردید تا تحلیل احساسات با توجه به متن‌های مرتبط با رشته مهندسی کامپیوتر انتخاب شوند. این رویکرد با استفاده از درخت CART توییت‌ها با موضوعیت گرایش‌های رشته مهندسی کامپیوتر را انتخاب و دسته‌بندی می‌کند. در ادامه به تحلیل احساسات توییت‌های انتخاب شده و نوع احساسات موجود در آن‌ها از طریق لیست الگوریتم SentiStrength پرداخته شده است. نتایج پیاده‌سازی بیان می‌دارد شاخص‌های ارزیابی متن‌کاوی تحلیل احساسات اعم از دقت، صحت، بازخوانی و امتیاز «F» بهبود یافته‌اند. امید است که ارائه این رویکرد ترکیبی منجر به پیشبرد تحقیقات و رویکردهای مبتنی بر متن و تحلیل احساسات شود. پیشنهاد می‌شود جهت پیشبرد تحقیقات آتی با به کارگیری قواعد انجمنی برای بهینه‌سازی لیست مورد جستجو زمان اجرای الگوریتم کاهش یابد. همچنین جهت متن‌کاوی تحلیل احساسات پیشنهاد می‌گردد از ویژگی‌های دیگر شبکه‌های اجتماعی مانند لینکداین، فیس‌بوک، اینستاگرام استفاده گردد.

منابع

- Baumer, E. P., Sinclair, J., & Tomlinson, B. (2010, April). America is like Metamucil: fostering critical and creative thinking about metaphor in political blogs. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1437-1446). ACM.
- Charalampakis, B., Spathis, D., Kouslis, E., & Kermanidis, K. (2016). A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets. *Engineering Applications of Artificial Intelligence*, 51, 50-57.
- Evans, B. M., Kairam, S., & Pirolli, P. (2010). Do your friends make you smarter?: An analysis of social strategies in online information seeking. *Information Processing & Management*, 46(6), 679-692.
- Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, *40*(16), 6266-6282.
- Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2), 28.
- Lahuerta-Otero, E., & Cordero-Gutiérrez, R. (2016). Looking for the perfect tweet. The use of data mining techniques to find influencers on Twitter. *Computers in Human Behavior*, 64, 575-583.
- Li, J., Li, Q., Khan, S. U., & Ghani, N. (2011, June). Community-based cloud for emergency management. *In 2011 6th International Conference on System of Systems Engineering* (pp. 55-60). IEEE.
- Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). *Springer, Boston, MA*.
- Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management*, 52(1), 5-19.
- Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311, 18-38.
- Sorensen, L. (2009, May). User managed trust in social networking-Comparing Facebook, MySpace and Linkedin. In *2009 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology* (pp. 427-431). IEEE.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544-2558.
- Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *HP Laboratories, Technical Report HPL*-2011, 89.

استناد به این مقاله: شناسه دیجیتال (DOI): 10.22091/jemsc.2018.1272

طیرانی نجاران، نصیر، جلالی، مهرداد. (۱۳۹۷). «ارائه رویکرد ترکیبی نوین جهت متن‌کاوی تحلیل احساسات در توییتر با استفاده از درخت تصمیم CART». *مدیریت مهندسی و رایانش نرم*، ۴ (۱)، ۷۶-۵۹.