
بسم الله الرحمن الرحيم

دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی هوش مصنوعی

تحلیل احساسات در شبکه‌های اجتماعی با استفاده از روش‌های آماری

نگارنده: حمیده شیخ

استاد راهنما

دکتر مرتضی زاهدی

استاد مشاور

دکتر مرضیه رحیمی

شهریور ۱۳۹۸

شماره:
تاریخ:

باسمه‌تعالی

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای .....

با شماره دانشجویی ..... رشته ..... گرایش .....

تحت عنوان .....

که در تاریخ ..... با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

قبول (با درجه: .....) ☐ **مردود** ☐

نوع تحقیق: **نظری** ☐ **عملی** ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول			
۲- استادراهنمای دوم			
۳- استاد مشاور			
۴- نماینده تحصیلات تکمیلی			
۵- استاد ممتحن اول			
۶ـ استاد ممتحن دوم			

نام و نام خانوادگی رئیس دانشکده:

تاریخ و امضاء و مهر دانشکده:

تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می تواند از پایان نامه خود دفاع نماید (دفاع مجدد نباید زودتر از ۴ ماه برگزار شود).

تقدیم به

همسر عزیزم

که سایه مهربانیش سایه سار زندگیم می باشد.

تشکر و قدردانی

سپاس خدای بزرگ را که مرا یاری رساند تا بتوانم این مقطع تحصیلی را به پایان رسانده و گامی در راستای اعتلای علم بردارم. از استاد راهنمای گران‌قدرم جناب آقای دکتر مرتضی زاهدی که وجودشان همیشه قوتی برای انجام کارهایم بوده است و بدون شک انجام این پایان نامه بدون کمک و راهنمایی‌های ارزنده ایشان امکان‌پذیر نبوده است، کمال تشکر را دارم. همچنین از سرکار خانم دکتر مرضیه رحیمی که استاد مشاور این پایان‌نامه بوده‌اند نیز قدردانی می‌نمایم.

از داوران گرامی سرکار خانم دکتر هدی مشایخی و جناب آقای دکتر منصور فاتح که زحمت داوری و تصحیح این پایان‌نامه را به عهده داشتند کمال سپاس را دارم. خالصانه از تمامی اساتید و معلمان و مدرسانی که در مقاطع مختلف تحصیلی به من علم آموخته و مرا از سرچشمه دانایی سیراب کرده‌اند متشکرم.

تعهدنامه

اینجانب حمیده شیخ دانشجوی دوره کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود نویسنده پایان‌نامه تحلیل احساسات در شبکه های اجتماعی با استفاده از روش های آماری تحت راهنمائی دکتر مرتضی زاهدی متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان‌نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

چکیده

امروزه شبکه‌های اجتماعی توجه ویژه‌ای را به خود جلب نموده‌اند. در شبکه‌های اجتماعی گوناگون، کاربران دائما در حال ابراز نظرات عمومی و همچنین خصوصی خود درباره‌ی موضوعات مختلف هستند. توییتر یکی از این شبکه‌های اجتماعی است که در دهه اخیر محبوبیت بسیاری یافته است. این شبکه اجتماعی روشی سریع و موثر برای تحلیل احساسات، دیدگاه‌ها و انتقادات مشتریان برای موفقیت در بازار را به سازمان‌ها ارائه می‌دهد. تحلیل احساسات یا عقیده کاوی فرآیندی است که در آن نظرات، احساسات و نگرش افراد در ارتباط با موضوعی خاص استخراج می‌شود. تحلیل احساسات در حوزه بلاگ‌نویسی کوچک موضوعی نوپا بوده و کماکان پتانسیل تحقیق و توسعه بسیاری دارد.

پژوهش‌های زیادی در رابطه با تحلیل احساسات بر روی نظرات کاربران، مستندات و مقالات انجام شده است. تحلیل بر روی موارد بیان شده تفاوت عمده‌ای با داده‌های توییتر دارد، به این سبب که توییت‌های توییتر محدودیت ۲۸۰ کاراکتری دارند و کاربران را وادار به بیان احساسات خود به صورت فشرده و کوتاه می‌نمایند. بهترین نتایج به‌دست آمده در طبقه‌بندی احساسات از تکنیک‌های یادگیری‌ماشین مثل بیز ساده و ماشین بردار پشتیبان حاصل شده است.

در این پژوهش به ارائه روشی برای تحلیل احساسات در شبکه‌های اجتماعی پرداخته می‌شود. در این راستا سعی شده با تمرکز بر مراحل پیش‌پردازش داده‌ها و انتخاب ویژگی، طبقه‌بندی متن توسط روش بیز را تا حدودی بهبود بخشیم. روش پیشنهادی صحت تشخیص ۸۳/۳۷ درصدی دارد. همچنین نتایج نشان می‌دهد که استفاده از ویژگی‌های n-gram بهترین نتیجه را در سیستم ارائه نموده و با ترکیب آن با وزن‌دهی به صفات براساس نسبت بهره، بهبود چشم‌گیری در نتایج به دست آمده حاصل می‌شود.

کلمات کلیدی: تحلیل احساسات، متن کاوی، کاوش نظرات، طبقه بندی متن، یادگیری ماشین، انتخاب ویژگی.

فهرست مطالب

فهرست جداول

فهرست اشکال

فصل ۱: مقدمه

۱-۱- مقدمه

انتشار گسترده شبکه‌های اجتماعی[۱] و نقش آنها در جامعه‌ی مدرن، یکی از جالب‌ترین موارد نوآوری در سال‌های اخیر است که علاقه‌ی محققان، روزنامه‌نگاران، شرکت‌ها و دولت‌ها را به خود جلب کرده است. روابط میان کاربران فعال موجب ایجاد انگیزه و جذب افرادی با اهداف مشترک و همچنین تسهیل در تنوع فعالیت‌های اجتماعی می‌شود؛ بنابراین شبکه‌های اجتماعی انقلابی در دنیای دیجیتال ایجاد می‌کنند.

انقلاب[۲] در شبکه‌های اجتماعی ما را قادر می‌سازد تا در رابطه با احساسات[۳] و ایده‌های خود و اطرافیان به گفتگو بپردازیم، با آن‌ها تعامل داشته‌باشیم و در نهایت همه و یا اکثر اتفاقات زندگی خود را بازگو کنیم، به عنوان مثال شادی حاصل از تولد یک نوزاد، خشم حاصل از تاخیر قطار و خریدهای عید. جای تعجب نیست که محققان در حال بررسی روش‌های گردآوری این حجم بالای اطلاعات هستند. اگر اطلاعات موجود در اینترنت به درستی جمع‌آوری‌شده و مورد بررسی قرارگیرند، نه تنها به ما اجازه می‌دهند برای بسیاری از پدیده‌های اجتماعی پیچیده پاسخی بیابیم؛ بلکه این امکان را فراهم می‌سازند که آن‌ها را پیش‌بینی نماید. در این بین تحلیل احساسات[۴] در تلاش برای آشکارسازی افکار مردم توسط ارائه نمادها، مدل‌ها و الگوریتم‌ها، از یک «متن ساده‌ی بدون ساختار[۵]» به سوی یک «درک پیچیده[۶]» است [۱].

تحلیل احساسات و به بیان دیگر نظرکاوی از سال ۲۰۰۰ میلادی تا کنون، به یکی از فعال‌ترین حوزه‌های پژوهشی در زمینه پردازش زبان‌های طبیعی تبدیل شده‌است [۲]. هدف از تحلیل احساسات، معرفی یک ابزار اتوماتیک برای استخراج اطلاعات از یک متن به زبان طبیعی مثل نظرات و احساسات است؛ به طوری که در

---

۱ social media

۲ revolution

۳ sentiment

۴ sentiment analysis

۵ simple unstructured text

۶ complex insight

نهایت دانشی ساختاریافته و عملی ایجاد کند تا بتوان از آن در یک سیستم پشتیبانی و تصمیم‌گیرنده استفاده کرد [۱].

با رشد و توسعه روزافزون شبکه‌های اجتماعی نقش تحلیل احساسات بسیار پررنگ‌تر گردیده‌است. امروزه اکثر وب‌سایت‌ها بخشی را برای نظردهی کاربران در رابطه با محصولات و خدمات ارائه‌شده و به اشتراک گذاری آن‌ها با دوستانشان در توییتر، فیس‌بوک و دیگر شبکه‌های اجتماعی در نظر گرفته‌اند.

۱-۲- تعریف احساس

مجموعه ای از حالات مانند، خشم، تنفر، شادی، عشق، اندوه، آرزو، ترس و... احساسات هر فرد را تشکیل می‌دهند. حرکات کسی که می‌خندد، برای هر فردی قابل فهم است. ایما و اشاره در سرتاسر جهان معتبر و بیانگر احساسات و عواطف انسان و از ویژگیهای ذاتی وی هستند. وجود احساسات به انسانها کمک می‌کند که با هم ارتباط برقرار کنند.

احساسات همواره از دیرباز جنبه مرموز و ناشناخته انسانها بوده و جایگاه مهمی در حیات اجتماعی افراد دارا می‌باشد. پرداختن به حالات احساسی و عاطفی همچون خشم، شادی، ترس، غم، کینه، بغض، عصبانیت، شرم، گناه و امثال آن به ضرورتی بنیادین در کنش‌های انسانی تبدیل شده و تحلیل رفتار افراد بدون در نظر گرفتن احساسات و عواطف ناقص بوده و ارزش چندانی ندارد. همیشه بخش مهمی از رفتاری که بشر در راستای گردآوری اطلاعات انجام می‌دهد برای فهمیدن این است که سایر انسان‌ها چه فکری می‌کنند. با افزایش دسترسی‌پذیری و محبوبیت منابع غنی از عقاید مانند سایت‌های نقد و بررسی آنلاین، وبلاگ‌های شخصی و شبکه‌های اجتماعی، فرصت‌ها و چالش‌های جدیدی در این حوزه ایجاد شده است.

۱-۳- تحلیل احساسات

تحلیل احساسات یا عقیده‌کاوی شاخه‌ای از پردازش زبان طبیعی است که سعی در تشخیص و استخراج نظرات از یک متن، عبارت و یا یک جمله دارد [۳]. به بیان دیگر تحلیل احساسات فرایندی است که در آن مثبت، منفی و یا خنثی بودن بخشی از متن مشخص می‌شود. در حال حاضر تجزیه و تحلیل احساسات با توجه به کاربردهای فراوانی که دارد، مورد توجه بسیاری قرارگرفته و توسعه فراوانی یافته است. این تکنیک در بسیاری از زمینه‌ها از جمله تجارت الکترونیک، مراقبت‌های پزشکی و در موارد اندکی در سرگرمی و سیاست مورد استفاده قرار می‌گیرد. از آنجا که داده‌های موجود در اینترنت به طور مداوم در حال افزایش است، تعداد زیادی متن حاوی نظرات مختلف در سایت‌ها، انجمن‌ها، بلاگ‌ها و شبکه‌های اجتماعی در دسترس هستند. به عنوان مثال، تحلیل احساسات برای شرکت‌هایی که قصد نظارت بر نظرات مشتریان از هر یک از محصولات خود دارند و همچنین برای مشتریانی که با توجه به نظرات دیگران نسبت به محصولات سعی در انتخاب بهترین محصول دارند، می‌تواند مفید واقع شود [۳].

۱-۴- شبکه اجتماعی توییتر[۱]

توییتر یکی از پرطرفدارترین بسترهای شبکه‌های اجتماعی است که برای بیان ایده‌ها، افکار، و نظرات کاربران استفاده می‌شود و در سال‌های اخیر بسیار مورد توجه آن‌ها قرارگرفته است. این میکروبلاگینگ[۲] در سال ۲۰۱۱ بیش از ۱۰۰ میلیون کاربر داشت که در حال حاضر روزانه ۳۴۰ میلیون توییت[۳] در آن فرستاده می‌شود و بیش از ۱/۶ میلیارد جست‌وجو در آن انجام می‌گردد. توییتر به کاربران اجازه ارسال متن‌های کوتاه شامل ۲۸۰

[۱] twitter

[۲] microblogging

[۳] tweet

کاراکتر را می‌دهد که توییت نامیده می‌شوند. توییت‌ها می‌توانند حاوی متن، ویدئو، تصویر، لینک[1] و... باشند. توییتی که توسط کاربر پست می‌شود در پروفایل شخصی او به نمایش گذاشته می‌شود که توسط دنبال‌شوندگان شخص قابل مشاهده است. توییتر همچنین قابلیت ارسال پیام خصوصی را داراست. یک بازتوییت[2] توییتی است که توسط کاربری ساخته شده و توسط کاربر دیگری مجدد ارسال می‌گردد [۴].

به دلیل سهولت دسترسی به حجم زیادی از پیام‌ها که به صورت آنی تولید می‌شوند، داده‌های توییتر معمولا به عنوان دیتاستی مناسب برای تحلیل احساسات استفاده می‌شود. افراد مختلف از قشرهای متفاوت مثل دانشجویان، شرکت‌ها، افراد سیاسی و... از این شبکه اجتماعی به عنوان محلی برای گردهمایی استفاده می‌کنند. شهرت توییتر در بین افراد سبب شده روزانه حجم وسیعی از اطلاعات در موضوعات متفاوت، از صحبت‌های روزمره کاربران گرفته تا نظرات آن‌ها درباره‌ی کالاها، برندها و رویدادهای اجتماعی و سیاسی مبادله شود. از این رو توییتر ابزاری قدرتمند برای پیش‌بینی‌ها به‌شمار می‌رود [۵].

۱-۵- کاربردهای تحلیل احساسات

یکی از مهم‌ترین نیازهای سازمان‌ها و کسب و کارها در دنیای واقعی در وحله‌ی اول یافتن مشتری و سپس تحلیل و بررسی نظرات عموم مشتریان درباره‌ی محصولات و خدمات آن‌هاست (به عنوان مثال چرا مشتریان لپ‌تاپ‌های ما را نمی‌خرند؟!). همچنین آگاهی از نظرات مشتریان حول یک محصول خاص برای دیگر مشتریان جالب توجه است. آن‌ها از این اطلاعات برای تصمیم‌گیری درباره‌ی خرید و یا عدم‌خرید یک محصول استفاده می‌کنند. این مسئله نشان می‌دهد فرایندهای تصمیم گیری در زندگی روزمره امر رایجی هستند. بنابراین مصرف‌کنندگان برای دانستن نظرات دیگران محدود به دوستان و آشنایان خود نمی‌شوند و سازمان‌ها نیز مجبور به گردآوری نظرات نیستند.

---

[1] link

[2] retweet

تحلیل احساسات کاربردهای فراوان و جالبی در همه‌ی زمینه‌های ممکن دارد. به عنوان مثال در امتیازدهی به محصولات ممکن است کاربران به طور تصادفی رتبه‌ی کمی به یک محصول بدهند درحالی‌که نظرات آن‌ها حاکی از ارزیابی مثبت آن‌ها دارد. علاوه بر این نظرات نقش مهمی را در سیاست ایفا می‌کنند. در برخی موارد تمرکز بر روی افکار رای‌دهندگان است. مثلا در سال ۲۰۰۸ میلادی باراک اوباما، رئیس‌جمهور آمریکا از تحلیل احساسات برای ارزیابی احساسات رای‌دهندگان اصلی در طول انتخابات ریاست جمهوری استفاده کرد. به عنوان مثالی دیگر می‌توان تقویت سیستم‌های توصیه‌گر را نام برد که در آن‌ها سیستم مواردی که چندین بار بازخورد منفی دریافت کرده‌است را توصیه نمی‌کند. در بسیاری از سیستم‌های آنلاین، تبلیغات در منوی کناری صفحه نمایش داده می‌شوند. اگر صفحاتی که محتوای آن‌ها برای تبلیغات مناسب نیستند تشخیص داده‌شوند، سیستم می‌تواند تبلیغاتی را که نظرات مثبت دریافت کرده‌اند برجسته‌تر و تبلیغاتی را که نظرات منفی دریافت کرده‌اند مخفی کند.

۱-۶- روش‌های تحلیل احساسات

در زمینه تحلیل احساسات سه روش مختلف ارائه گردیده‌است: روش‌های مبتنی بر یادگیری ماشین[۱]، روش‌های مبتنی بر فرهنگ لغات[۲] و روش‌های حاصل از ترکیب این دو روش. در بخش‌های بعد به تفصیل به بررسی هر یک از این روش‌ها پرداخته می‌شود.

۱-۶-۱- روش‌های مبتنی بر فرهنگ لغات

در این رویکرد از یک فرهنگ لغت[۳] استفاده می‌شود که در واقع معرف هر توییت خواهد بود. این فرهنگ لغت توسط انسان ساخته می‌شود و شامل همه لغات مثبت، منفی و خنثای توییت‌ها است. لغاتی که نه احساس

---

۱ machine learning (ML)

۲ lexicon-based

۳ lexicon

مثبت و نه منفی دارند به صورت خنثی در نظر گرفته می‌شوند. پس از تشکیل فرهنگ لغت، برچسب گذاری توییت‌ها انجام می‌شود. به این صورت که وجود هر یک از لغات توییت در فرهنگ لغت بررسی می‌شود و کلاسی که دارای بیشترین لغت در توییت مورد نظر بود، نشان دهنده‌ی کلاس توییت است. اگر تعداد کلمات مثبت و منفی یکسان باشد کلاس توییت خنثی در نظر گرفته می‌شود.

۱-۶-۲- روش‌های مبتنی بر یادگیری ماشین

رویکردهای یادگیری ماشین را می‌توان به‌طورکلی در دو بخش تقسیم‌بندی نمود:

۱. رویکردهای یادگیری ماشین نظارت شده[۱]
۲. رویکردهای یادگیری ماشین بدون نظارت[۲]

رویکردهای یادگیری ماشین نظارت شده نوعی از یادگیری است که در آن ورودی و خروجی مشخص است و ناظری وجود دارد که اطلاعاتی را در اختیار یادگیرنده قرار می‌دهد و به این ترتیب سیستم سعی می‌کند تا تابعی را از ورودی به خروجی فراگیرد. رویکرد یادگیری ماشین بدون نظارت دریادگیری بدون نظارت برخلاف یادگیری نظارت‌شده، داده‌های مشخصی از قبل وجود ندارد و هدف، ارتباط ورودی و خروجی نیست بلکه تنها دسته‌بندی آنها مهم است و این یادگیرنده است که بایستی در داده‌ها به دنبال ساختاری خاص بگردد.

۱-۷- چالش‌های تحلیل احساسات

چالش اصلی تحلیل احساسات این است که یک روش طبقه‌بندی[۳] ساده به درستی کار نمی‌کند. دلیل آن این است که در میان نظرات و متن‌های غافلگیرکننده‌ی موجود، نظراتی وجود دارند که اگرچه بیان‌کننده‌ی یک

---

۱ supervised machine learning approaches

۲ unsupervised machine learning approaches

۳ classification

نظر محکم و خاص هستند اما مستقیما دارای کلماتی نیستند که وابسته به طرز تفکر شخص[۱] باشند. همچنین نظراتی وجود دارند که دارای کلمات فریبکارانه بوده و به ظاهر حامل نظر مثبت یا منفی هستند، در صورتی که درواقع این‌طور نیست [۶].

به عنوان مثال جمله‌های "این فیلم بسیار درخشان است. به نظر طرح بزرگی می‌رسد، بازیگران درجه یک و دست‌اندرکاران نیز خوب هستند و استالون نیز سعی داشت اجرای خوبی ارائه دهد اما فیلم موفقی نبود." اگرچه در این نظر، جملات مثبت غالب بودند اما منعکس کننده‌ی نظر کلی نبودند. این دست نظرات طبقه‌بند را فریب داده و دقت[۲] روش را کاهش می‌دهند [۶]. یکی از اهداف تحلیل احساسات ارائه روش‌های جدید برای غلبه بر چنین مشکلاتی است.

علاوه بر چالش‌هایی که تحلیل احساسات بنابر تعریف کلی خود دارد، تحلیل احساسات در توییتر نیز به نوبه‌ی خود چالش‌هایی را در بر دارد. همانطور که می‌دانید، توییت‌های توییتر توسط افراد مختلف نوشته می‌شود، لذا این توییت‌ها به صورت غیررسمی و به زبان عامیانه هستند. بنابراین کلمات هر توییت می‌تواند از توییت دیگر متمایز بوده و حاوی غلط املایی، تکرار حروف و... باشد. چالش بعدی که در تحلیل احساسات در توییتر وجود دارد، کوتاه بودن جملات توییتر است که سبب محدودیت‌هایی در ابراز احساسات در چند جمله‌ی کوتاه و چند کلمه خواهد شد. به این معنی که یافتن نشانه‌ای از احساس در کلمات کمتر، مشکل تر از متون طولانی‌تر است. چالش سوم در تحلیل احساسات در توییتر استفاده از اختصارات و مخففها توسط کاربران است که تحلیل و تشخیص احساس از این اختصارات مشکل است. به خصوص زمانی که این چالش در کنار چالش اول قرار گرفته و ممکن است شکل متفاوتی داشته باشند اما به یک موضوع خاص اشاره کنند [۳].

---

[۱] subjective

[۲] accuracy

۱-۸- هدف تحقیق

تاکنون تحقیقات و پژوهش‌های فراوانی در زمینه تحلیل احساسات و بر روی مجموعه‌داده‌های مختلف انجام شده است. تحلیل احساسات در توییتر به دلیل محدودیت در تعداد کاراکترها و چالش‌هایی که این مسئله به دنبال دارد از تحلیل احساسات در سایر متون متفاوت بوده و نیاز به تحلیل و بررسی بیشتری دارد.

هدف از این تحقیق ارائه روشی برای تحلیل احساس در مجموعه‌داده توییتر می‌باشد که برای این منظور توییت‌ها به دو دسته مثبت و منفی طبقه‌بندی می‌شوند. همان‌طور که در بخش‌های پیشین به آن اشاره شد، از جمله مشکلاتی که برای تحلیل احساسات در توییتر وجود دارد کوتاه بودن جملات کاربران است که برای رسیدن به نتیجه‌ی بهتر علاوه بر انتخاب بهترین روش طبقه‌بندی، مراحل پیش‌پردازش داده‌ها و انتخاب خصیصه اهمیت بسیاری دارند. در این تحقیق سعی گردیده تا علاوه بر انجام پیش‌پردازش‌های لازم بر روی توییت‌ها، با استفاده از وزن‌دهی به صفات در مرحله انتخاب خصیصه، الگوریتم طبقه‌بند بیز بهبود یابد.

۱-۹- ساختار پایان‌نامه

در فصل دوم این تحقیق به طبقه‌بندی و بررسی روش‌های موجود در تحلیل احساسات پرداخته شده و همچنین کارهای مرتبط و شیوه‌های مورد استفاده در آن‌ها بیان گردیده است. در قصل سوم، روش‌های پیشنهادی تحلیل احساسات در توییتر طبق تحقیق فوق، در راستای افزایش صحت عملکرد طبقه‌بند ارائه گردیده است. در فصل چهارم، روش پیشنهادی پیاده سازی شده و جزئیات پیاده‌سازی و پارامترها و متغیرهای مورد استفاده بیان شده است. در فصل پنجم، نتایج حاصل از پیاده‌سازی مورد ارزیابی قرار گرفته است و در فصل ششم، نتیجه‌گیری کلی و برخی از پیشنهادات کارهایی که در آینده در جهت توسعه هدف تحقیق می‌توان انجام داد، بازگو شده است.

فصل ۲: پیشینه تحقیق

۲-۱- مقدمه

در این فصل به بررسی کارهای پیشین در حوزه تحلیل احساسات پرداخته می‌شود. هرچند تعداد آثار موجود در زمینه تحلیل احساسات در متن فراوان است اما در این بخش تعدادی از آثار مرتبط به صورت منتخب بیان خواهد شد. سپس کارهای پیشین را از دیدگاه‌های مختلف بررسی و طبقه‌بندی خواهد شد.

۲-۲- سطوح تحلیل احساسات

تحلیل احساسات در سه سطح متفاوت مورد بررسی قرار می‌گیرد:

۱. در سطح متن
۲. در سطح جمله
۳. در سطح ویژگی

هدف از تحلیل احساسات در سطح متن تعیین قطبیت احساسی نشان داده‌شده در کل متن است. تحلیل احساسات در سطح جمله، جملات را به دسته‌های مورد نظر (مثبت یا منفی و...) طبقه‌بندی می‌کند و تحلیل احساسات در سطح ویژگی یا موجودیت، قطبیت یک موجودیت خاص از یک شئ را تشخیص می‌دهد. در تحلیل احساسات در سطح متن و جمله، تنها از یک شئ استفاده شده و تنها یک نظر از آن شئ استخراج می‌شود اما این قبیل فرضیات در بسیاری از موارد مناسب نیست. استخراج احساس از کل یک متن یا بلاگ به اندازه استخراج احساس از یک جمله کارآمد و مناسب نخواهد بود زیرا استخراج احساس از یک جمله تمام جوانب هر یک از موضوعات را در جمله مد نظر قرار می‌دهد.

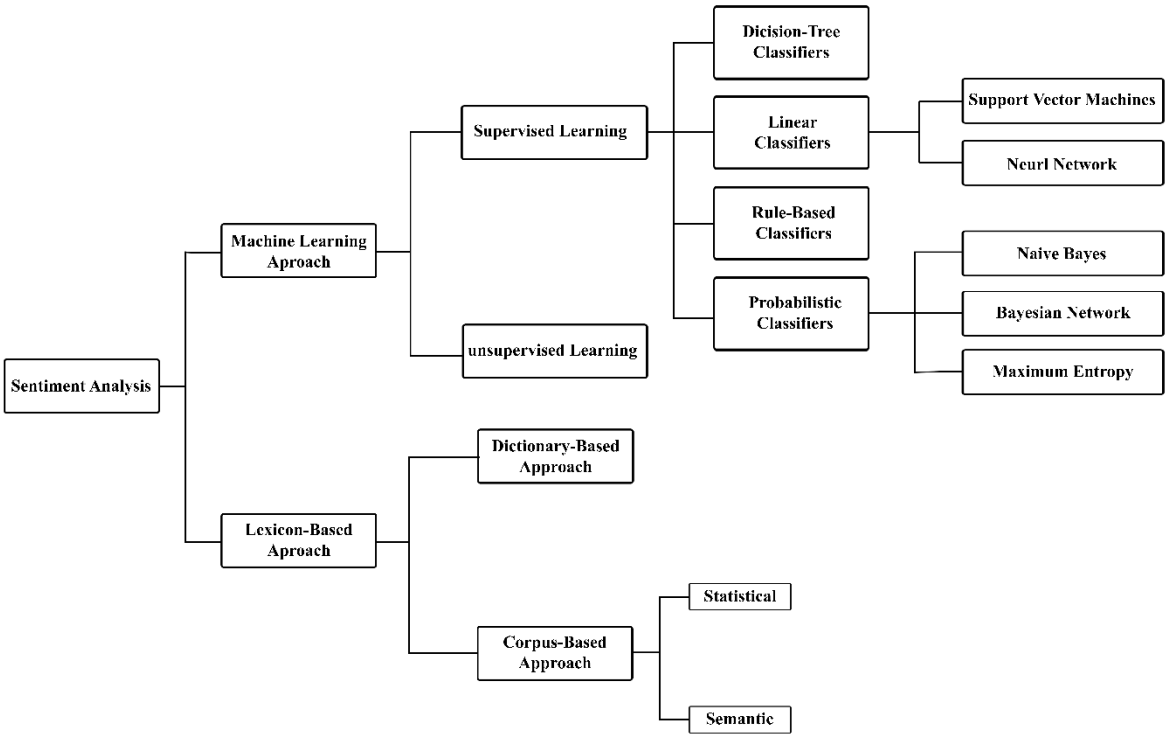
۲-۳- طبقه‌بندی کارهای پیشین بر اساس رویکردهای مورد استفاده

روش‌ها و الگوریتم‌های زیادی در زمینه تحلیل احساسات وجود دارد که می‌توانند در سه گروه طبقه‌بندی شوند:

۱. رویکردهای مبتنی بر یادگیری ماشین

۲. رویکردهای مبتنی بر لغت نامه

۳. رویکردهای ترکیبی



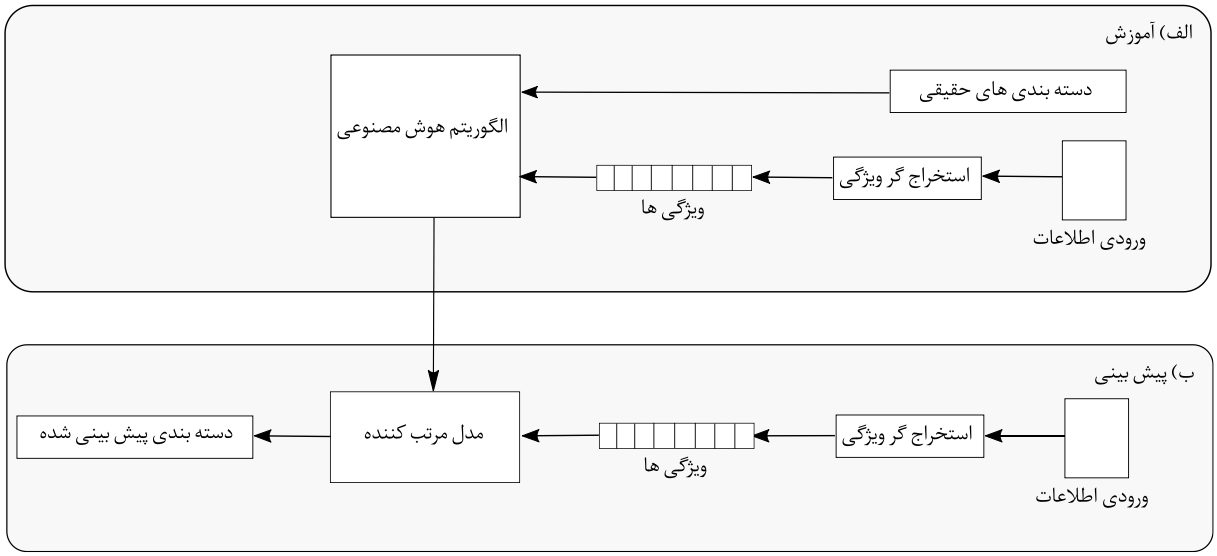
شکل ۲-۱. تقسیم‌بندی کلی روش‌های تحلیل احساسات

تقسیم‌بندی کلی روش‌های تحلیل احساسات را در شکل ۱-۲ نشان داده شده‌است. در ادامه هر یک از این رویکردها به تفصیل شرح داده‌شده و سپس برخی از پژوهش‌های انجام‌شده در این زمینه مورد تحلیل و بررسی قرار خواهدگرفت.

۲-۳-۱- رویکردهای مبتنی بر یادگیری ماشین

رویکردهای مبتنی بر یادگیری ماشین عموما به رویکردهای یادگیری با ناظر اطلاق می‌شود. در این رویکردها دو مجموعه داده نیاز است: مجموعه‌داده آموزش و مجموعه‌داده تست. مجموعه‌داده آموزش توسط یک طبقه‌بند و به منظور یادگیری ویژگی‌های متن‌ها و مجموعه‌داده تست برای ارزیابی کارایی و صحت عملکرد طبقه‌بند

مورد استفاده قرار می‌گیرد. در واقع وظیفه‌ی تحلیل احساسات معمولا مدل‌کردن یک مسئله‌ی طبقه‌بندی است، به این صورت که متنی به عنوان ورودی به طبقه‌بند داده می‌شود و طبقه‌بند دسته‌ی مطابق با آن را به صورت مثبت، منفی و خنثی (در صورتی که تعیین قطبیت مدنظر باشد) برمی‌گرداند.



شکل ۲-۲. (الف) فرایند آموزش و (ب) فرایند تست در الگوریتم‌های مبتنی بر یادگیری ماشین

در شکل ۲-۲ گام‌ها و اجزای اصلی پیاده‌سازی الگوریتم‌های مبتنی بر یادگیری ماشین نشان داده شده‌است. قسمت ۲-۲ (الف) مرحله آموزش الگوریتم‌های یادگیری ماشین است. در این بخش، فرایند یاد می‌گیرد هر ورودی از مجموعه آموزشی را به برچسب دسته‌ی متناظر با آن (مثبت، منفی و خنثی) منطبق کند. استخراج ویژگی متن ورودی را به یک بردار ویژگی تبدیل می‌کند. هر جفت عنصر بردار ویژگی و برچسب مربوط به آن که دسته‌ی آن ویژگی است منجر به تولید یک مدل توسط الگوریتم یادگیری ماشین می‌شوند. در قسمت ۲-۲ (ب) مرحله تست الگوریتم‌های یادگیری ماشین به تصویر کشیده شده‌است. در مرحله تست، استخراج ویژگی ورودی‌های دیده نشده از مجموعه‌داده تست را به بردار ویژگی تبدیل می‌کند. این بردارها به عنوان ورودی به مدل ساخته‌شده در مرحله قبل داده‌شده و دسته‌ی متناظر با آن‌ها پیش‌بینی می‌شود.

از جمله الگوریتم‌های یادگیری ماشین که در زمینه تحلیل احساسات مورد بررسی قرار می‌گیرند عبارتند از الگوریتم ماشین بردار پشتیبان[۱]، الگوریتم بیز[۲]، الگوریتم حداکثر آنتروپی[۳]. اولین گام در رویکردهای مبتنی بر یادگیری ماشین جمع‌آوری مجموعه داده آموزشی و سپس آموزش یک طبقه‌بند روی مجموعه‌داده است. یکی از مهم‌ترین مراحل در تحلیل احساسات مرحله انتخاب ویژگی است. رایج‌ترین ویژگی‌های مورد استفاده در زیر ارائه گردیده‌است.

وجود لغات و تعداد تکرار آن‌ها

این ویژگی‌ها که در سطح وسیعی در تحلیل احساسات به‌کاربرده شده شامل uni-gram ها و n-gram ها و همچنین تعداد تکرار آن‌ها می‌شود. پنگ و همکارانش [۷] معتقدند uni-gram ها نسبت به bi-gram ها در تحلیل احساسات از میان نظرات کاربران درباره‌ی فیلم‌ها بهتر عمل می‌کنند، اما دیو و همکارانش [۸] در تحلیل احساسات و طبقه‌بندی نظرات کاربران در رابطه با محصولات نتایج بهتری در مورد bi-gram ها و tri-gram ها گزارش کرده‌اند.

اطلاعات اجزای جمله[۴] (POS)

POS به گروهی از کلمات گفته می‌شود که از نظر ویژگی‌های گرامری به هم شبیه هستند. معمولا کلماتی که به یک POS یکسان نسبت داده می‌شوند، رفتار مشابهی از لحاظ نحوی از خود نشان می‌دهند. به عبارت دیگرُ این کلمات نقش مشابهی را از لحاظ ساختار گرامری جمله دارند. معروف‌ترین POS ها در زبان انگلیسی عبارتند از اسم، فعل، صفت، قید، ضمیر، پیشوند و حرف ربط. POS معمولا در تحلیل احساسات و نظرکاوی

[۱] Support Vector Machine (SVM)

[۲] Naïve Bayes (NB)

[۳] Maximum Entropy (ME)

[۴] Part Of Speech

مورد استفاده قرارمی‌گیرد. این که برچسب‌گذاری POS به عنوان شکل خامی از ابهام زدایی معنایی کلمه[۱] درنظرگرفته می‌شود، یک دلیل ساده نه تنها برای نظرکاوی بلکه برای تجزیه و تحلیل متن نیز می‌باشد [۹]. در برچسب‌گذاری POS به هر عبارت در جمله یک برچسب انتساب داده می‌شود که نشان دهنده‌ی نقش و جایگاه آن کلمه در ساختار گرامری جمله است. به عنوان مثال، با برچسب‌گذاری POS می‌توان صفت‌ها و قیدها را که شاخص‌های اصلی احساسات هستند تشخیص داد [۱۰].

کلمات نفی[۲]

کلمات نفی یکی از مهم‌ترین ویژگی‌ها در تحلیل احساسات محسوب می‌شوند زیرا توانایی این را دارند که احساس متن را معکوس کرده و جملات را در دسته‌های مخالف طبقه‌بندی کنند [۹].

کلمات و عبارات نظر

کلمات و عبارات نظر، آن دسته از کلمات و عباراتی هستند که احساسات مثبت و منفی را بیان می‌کنند. از رویکردهای اصلی برای تشخیص جهت‌گیری معنایی یک کلمه نظر، رویکردهای آماری یا رویکردهای برمبنای لغت‌نامه هستند. هو و همکارانش [۱۱] از Wordnet برای تشخیص مثبت یا منفی بودن قطبیت صفات استخراج شده استفاده کرده‌اند.

پنگ و همکارانش [۷] کارایی الگوریتم‌های بیز، حداکثر آنتروپی و ماشین بردار پشتیبان را در سطح متن و با درنظرگرفتن ویژگی‌هایی مثل uni-gram ها، bi-gram ها و ترکیب آن دو، همچنین ترکیب uni-gram ها با POS، با درنظرگرفتن صفت‌ها به تنهایی و ترکیب uni-gram ها و اطلاعات مکانی آن‌ها مورد بررسی قرارداده و با هم مقایسه کردند. نتایج به‌دست آمده نشان داده است که حضور ویژگی‌ها مهم‌تر از فراوانی آن‌هاست و زمانی که مجموعه ویژگی کوچک است الگوریتم بیز نسبت به ماشین بردار پشتیبان عملکرد بهتری دارد.

---

[۱] Word Sense Disambiguation (WSD)

[۲] Negation

الگوریتم ماشین بردار پشتیبان زمانی عملکرد بهتری دارد که فضای ویژگی افزایش پیدا می‌کند. زمانی که فضای ویژگی افزایش می‌یابد الگوریتم حداکثر آنتروپی ممکن است از الگوریتم بیز بهتر عمل کند اما ممکن است دچار بیش برازش[۱] شود.

عباسی و همکارانش [۱۲] روشی برای طبقه‌بندی احساسات در نوشته‌های گروه‌های افراطی انجمن‌های وب به زبان‌های عربی و انگلیسی پیشنهاد دادند و از ویژگی‌های نحوی و ظاهری استفاده نمودند. آن‌ها الگوریتم جدیدی به نام الگوریتم ژنتیک آنتروپی وزن دار[۲] را معرفی کردند و از الگوریتم ماشین بردار پشتیبان با اعتبارسنجی متقابل ۱۰ برابر[۳] و بوت استرپ[۴] برای طبقه‌بندی در هر آزمایش استفاده‌کردند. وقتی از هردو ویژگی نحوی و ظاهری استفاده کردند به دقت ۹۹/۵۵ درصدی در اعتبارسنجی متقابل ۱۰ برابر رسیدند.

آنکیت و همکارانش [۳] علاوه بر اقداماتی که در مراحل پیش‌پردازش و انتخاب ویژگی انجام دادند، چهار الگوریتم شامل الگوریتم بیز ساده، جنگل تصادفی[۵]، ماشین بردار پشتیبان و رگرسیون لجستیک[۶] را ترکیب نمودند. آن‌ها با ترکیب این روش‌ها روشی را ارائه دادند که از هر یک از این الگوریتم‌ها به تنهایی عملکرد بهتری دارد. آن‌ها روش فوق را بر روی چهار مجموعه‌داده مختلف مورد ارزیابی قرار دادند. نتایج ارائه شده نشان می‌دهد که روش پیشنهادی آنیک و همکارانش در همه موارد بهترین عملکرد را داراست.

۲-۳-۲- رویکردهای مبتنی بر لغت‌نامه

رویکردهای مبتنی بر لغت‌نامه از یک فرهنگ‌لغت یا فرهنگ‌واژگان که از قبل برچسب‌گذاری شده استفاده می‌کنند. فرهنگ واژگان احساسی حاوی لیستی از کلمات و عبارات است که نظرات و احساسات ذهنی افراد

---

۱ overfitting

۲ Entropy Weighted Genetic Algorithm (EWGA)

۳ 10-fold cross validation

۴ bootstrap

۵ Random forest

۶ Logistic regression

در آن بیان شده است. اگر در رویکردهای مبتنی بر لغت‌نامه بخواهیم احساس یک متن را تحلیل کنیم، در صورتی که هرکلمه موجود در متن در لغت‌نامه هم موجود باشد، با لغت‌نامه مقایسه می‌شود و مقدار[۱] قطبیت[۲] آن به مقدار قطبیت کل متن اضافه می‌شود. به عنوان مثال اگر کلمه‌ی "عالی" موجود در متن در لغت‌نامه نیز بوده و دارای قطبیت مثبت باشد، قطبیت کل متن افزایش می‌یابد. درصورتی که قطبیت کل متن مثبت باشد آن متن در دسته مثبت و در غیر این صورت در دسته منفی قرار می‌گیرد. به بیان دیگر درصورتی‌که متن تعداد بیشتری کلمه مثبت داشته‌باشد، احساس کل متن مثبت و در غیر این صورت منفی در نظر گرفته می‌شود [۱۳].

رویکردهای مبتنی بر لغت‌نامه در تحلیل احساسات، یادگیری بدون ناظر هستند زیرا برای طبقه‌بندی اطلاعات نیاز به آموزش قبلی ندارند.

گام‌های اصلی رویکردهای مبتنی بر لغت‌نامه به صورت زیر است [۱۳]:

۱. پیش پردازش هریک از متون شامل حذف برچسب‌گذاری‌ها، کاراکترهای زاید، حذف اعداد و ...
۲. مقداردهی اولیه به قطبیت یا احساس کل متن: $s \to 0$
۳. شکستن متن به کلمات و چک کردن وجود یا عدم وجود هر کلمه در لغت‌نامه
   a. اگر در لغت‌نامه بود، در صورتی که قطبیت کلمه مثبت بود: $s \to s + w$
   b. اگر در لغت‌نامه بود، در صورتی که قطبیت کلمه منفی بود: $s \to s - w$
۴. بررسی قطبیت کل متن (s)
   a. اگر s بزرگتر از حد آستانه بود، متن در دسته مثبت قرار می‌گیرد.
   b. اگر s کوچکتر از حد آستانه بود، متن در دسته منفی قرار می‌گیرد.

[۱] value

[۲] polarity

سه روش برای ساخت یک لغت‌نامه احساسی وجود دارد: ساخت دستی[۱]، روش‌های مبتنی بر فرهنگ‌لغت[۲]، روش‌های مبتنی بر پیکره[۳] که از بین آن‌ها روش ساخت دستی لغت‌نامه فرایندی مشکل و زمان‌بر است.

در تکنیک‌های مبتنی بر فرهنگ‌لغت به این صورت عمل می‌شود که ابتدا یک مجموعه کوچک از کلمات نظر همراه با جهت قطبیت هریک جمع‌آوری کرده و سپس این مجموعه با جستجو در فرهنگ‌لغت WordNet و پیدا کردن مترادف‌ها و متضادها گسترش داده‌می‌شود. کلمات جدید به لیست اضافه می‌شوند و تکرار بعدی شروع می‌شود. این تکرارها تا زمانی ادامه پیدا می‌کنند که کلمه جدیدی یافت نشود [۱۴]. قطبیت کلمات به کلمات مترادف نسبت داده می‌شود و قطبیت کلمات متضاد نیز مخالف قطبیت کلمه است.

هو و همکارانش [۱۱] این روش را برای پیداکردن قطبیت صفت‌ها به‌کاربردند. آن‌ها از ۳۰ صفت به عنوان لیست اولیه استفاده‌کردند. یکی از محدودیت‌های روش‌های مبتنی بر فرهنگ‌لغت این است که قادر به پیداکردن کلمات احساسی برای موضوعات خاص نیست [۱۴].

تکنیک‌های مبتنی بر پیکره در سطح وسیع با تکیه بر الگوهای نحوی عمل می‌کنند. این روش‌ها می‌توانند کلمات احساسی با دقت نسبتا بالایی را تولید کنند. بسیاری از این روش‌ها نیاز به یک مجموعه داده آموزشی بسیار بزرگ دارند. یکی از مهم‌ترین مزیت‌های این روش‌ها نسبت به روش‌های مبتنی بر لغت‌نامه این است که می‌توانند لغات احساسی مرتبط با موضوعات خاص را پیدا کنند.

یکی از نخستین و مهم‌ترین کارهای انجام شده با استفاده از روش‌های بدون ناظر در زمینه تحلیل احساسات و نظرکاوی توسط ترکی و همکارانش [۱۰] انجام شده است. در روش‌های تکامل یافته‌تر از لغت‌نامه‌های پیشرفته برای تحلیل متن استفاده می‌شود. از جمله این لغت‌نامه‌ها می‌توان به WordNet اشاره‌کرد که برای پردازش‌های مربوط به پردازش زبان ساخته شده‌است. این لغت‌نامه مجموعه‌ای به نسبت بزرگ از لغات و ارتباط

---

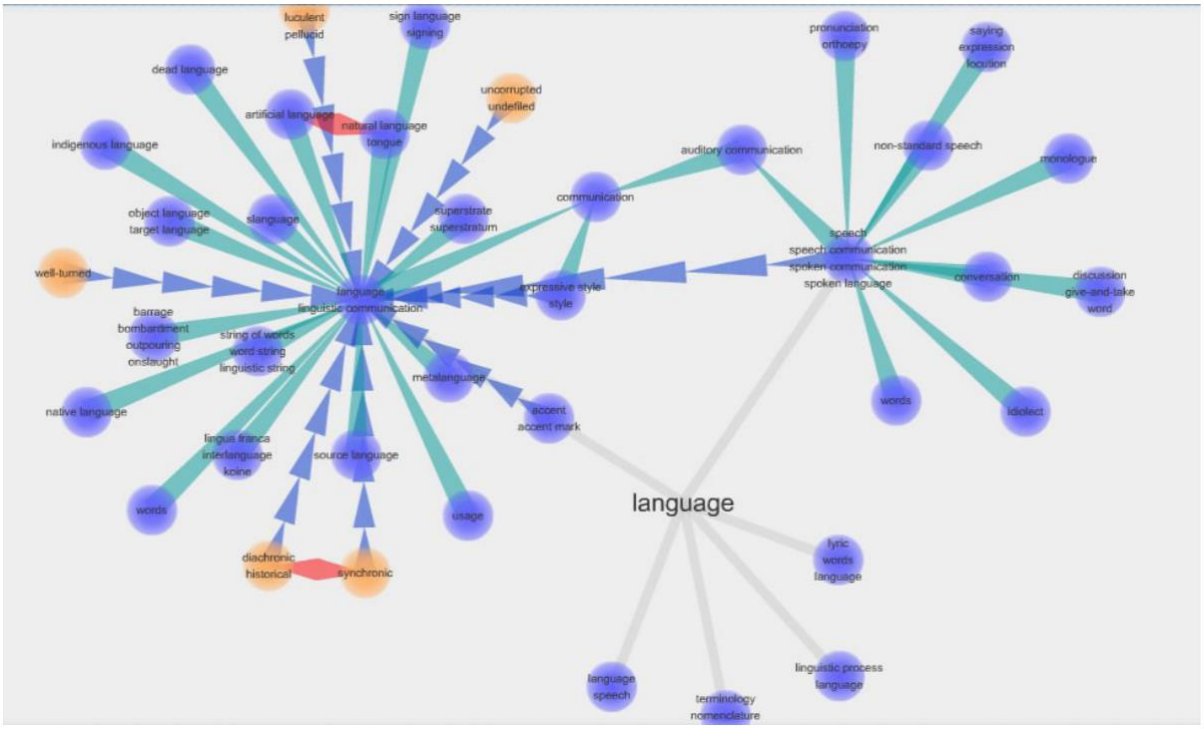
۱ manual

۲ dictionary based

۳ corpus based

آن ها در ساختار متون است. WordNet در حقیقت رابطه معنایی بین دو کلمه را بیان می‌کند. به عنوان مثال کبوتر از نوع پرنده است و پرنده یک نوع حیوان است و حیوان یک نوع جاندار است. این لغت‌نامه تمام کلمات انگلیسی را طبقه‌بندی کرده به طوری‌که هر کدام از دسته‌ها می‌توانند با دسته‌های دیگر رابطه‌ای مثل تضاد یا تناسب داشته باشند. در شکل ۳-۲ نمایی از WordNet نشان داده شده‌است.

SentiWordNet نسخه‌ی ویرایش شده‌ی WordNet است. در SentiWordNet به تمام لغات دو مقدار مثبت و منفی نسبت داده شده‌است که این مقدار عددی بین ۰ و ۱ است.



شکل ۲-۳. نمای کلی از WordNet

۲-۳-۳- رویکردهای ترکیبی

ایده روش‌های ترکیبی بسیار نوآورانه و خلاقانه است. بسیاری از پژوهشگران رویکردهای مبتنی بر یادگیری ماشین و رویکردهای مبتنی بر لغت‌نامه را ترکیب کردند. یکی از جالب‌ترین مطالعات در این زمینه توسط ژانگ

و همکارانش [۱۵] انجام شده است. آن‌ها روشی ترکیبی ارائه دادند که در آن از تحلیل احساسات توییتر[۱] در سطح موجودیت[۲] استفاده کردند. آن‌ها برای هر یک از موجودیت‌های "Obama"، "Harry Potter"، "Tangled"، "Ipad" و "Parkers" بر اساس کلمات مجاور و از یک لغت‌نامه احساسی، مقدار قطبیت یا احساس را محاسبه کردند. آن‌ها یک الگوریتم مبتنی بر لغت‌نامه ارائه دادند که در آن نظرات مقایسه‌ای، کلمات نفی و عباراتی که امکان تغییر جهت قطبیت را دارند، مورد توجه قرار گرفتند. دقت این روش بالاست اما بازخوانی کم است. به منظور بالابردن معیار بازخوانی، با انجام آزمون Chi-Square برروی نتایج حاصل از روش‌های مبتنی بر لغت‌نامه تعدادی شاخص معتبر اضافی استخراج کردند. سپس از یک طبقه‌بند ماشین بردار پشتیبان برای تشخیص قطبیت احساس استفاده نمودند. در واقع در این روش به جای برچسب‌گذاری دستی، مثال‌های آموزشی توسط رویکرد مبتنی بر لغت‌نامه داده‌شدند.

مطالعه جالب توجه دیگری در زمینه رویکردهای ترکیبی توسط غیاثی و همکارانش [16] انجام شده‌است. آن‌ها شبکه عصبی مصنوعی پویا[۳] را با n-gram ترکیب کردند. شکلک‌ها و توییت‌هایی که حاوی کلمات "Love" و "Hate" و مترادف‌های آن‌ها هستند را به عنوان ویژگی برای ساخت دو طبقه‌بند ماشین بردار پشتیبان و معماری پویای شبکه عصبی مصنوعی[۴] (DAN2) استفاده کردند. رویکرد پیشنهادی را روی مجموعه‌ای از توییت‌های حاوی کلمه "Justin Bieber" تست کردند. نتایج حاکی از آن است که الگوریتم DAN2 از ماشین بردار پشتیبان عملکرد بهتری دارد.

یکی از پژوهش‌های اخیر که در آن از روش ترکیبی استفاده نمودند توسط اپل و همکارانش [۱۷] انجام شده‌است. آن‌ها روشی برای تحلیل احساسات در توییتر و در سطح جمله ارائه دادند که در آن دو الگوریتم حداکثر آنتروپی و بیز را ترکیب کردند. روش ارائه شده بر روی دو مجموعه‌داده Movie Review و Sentiment

---

[۱] Twitter Sentiment Analysis (TSA)

[۲] entity level

[۳] dynamic artificial neural network

[۴] dynamic architecture for artificial neural etworks (DAN2)

Twitter مورد بررسی قرارگرفت. روش‌های حداکثر آنتروپی و بیز به طور جداگانه بر روی این دو مجموعه‌داده اعمال شدند. مقایسه نتایج نشان داد که در روش ترکیبی به دقت بالاتری نسبت به بیز و حداکثر آنتروپی رسیدند که نشان دهنده برتری روش ترکیبی در تحلیل احساسات توییتر در سطح جمله است.

پژوهش‌هایی در زمینه تحلیل احساسات با تمرکز بر مرحله انتخاب ویژگی و به ویژه در قسمت وزن‌دهی به صفات صورت گرفته است. از جمله این پژوهش‌ها می‌توان به ژانگ و همکارانش [18] اشاره کرد. آن‌ها در تحقیق خود دو روش وزن‌دهی به صفات برای طبقه‌بند بیز ارائه نمودند. آن‌ها رویکردهای وزن‌دهی به صفات بر اساس نسبت بهره و وزن‌دهی به صفات بر اساس درخت تصمیم را مورد بررسی قرار دادند. نتایج به دست آمده نشان می‌دهد که دقت الگوریتم بیز با رویکرد پیشنهادی افزایش می‌یابد.

۲-۳-۴- مقایسه سه رویکرد تحلیل احساسات

رویکردهای باناظر مبتنی بر یادگیری ماشین کارایی تقریبا بالایی نسبت به رویکردهای بدون‌ناظر مبتنی بر لغت‌نامه نشان دادند. اگرچه رویکردهای بدون‌ناظر هم بسیار مهم هستند زیرا رویکردهای باناظر نیاز به حجم زیادی داده آموزشی برچسب‌گذاری شده دارند که جمع آوری آن‌ها بسیار زمان‌بر و پرهزینه است. در حالی که جمع‌آوری داده‌های بدون برچسب بسیار ساده است. بسیاری از مجموعه‌داده‌ها فاقد برچسب‌گذاری هستند که در این موارد رویکردهای بدون‌ناظر پرکاربردتر هستند.

در مورد رویکردهای مبتنی بر لغت‌نامه، کلمات نظر یا احساسی که در فرهنگ‌لغت وجود دارد حائز اهمیت هستند. اگر فرهنگ‌لغت تعداد کمی کلمه داشته‌باشد، منجر به کاهش کارایی خواهد شد. چالش مهم دیگری که در این رویکرد وجود دارد این است که قطبیت بسیاری از کلمات مستقل از زمینه و محتوا است. به عنوان مثال، "فیلم مضحک" در زمینه فیلم مثبت است اما "مزه‌ی مضحک" در زمینه غذا منفی است. لغت‌نامه‌های فعلی این قبیل حساسیت‌ها به زمینه و محتوای کلمات احساسی را دربرنمی‌گیرند. بدون یک لغت‌نامه کامل و جامع تحلیل احساسات به درستی عمل نخواهد کرد.

یکی از مزیت‌های رویکردهای ترکیبی بااستفاده از ترکیب دو روش یادگیری ماشین و لغت‌نامه، دستیابی به بهترین‌های هردو است، دقت بالا از یک الگوریتم یادگیری باناظر قوی و ثبات از یک رویکرد مبتنی بر لغت‌نامه.

جدول ۲-۱. مقایسه سه رویکرد تحلیل احساسات

ردیف	رویکرد	مزایا	معایب
۱	یادگیری ماشین	دقت بالا	نیاز به مجموعه داده برچسب گذاری شده – هزینه بالا و زمان بر بودن فرایند آموزش و تست
۲	لغت نامه	عدم نیاز به مجموعه داده برچسب گذاری شده – ثبات در کارایی	نیاز به مجموعه داده بزرگ – حساسیت به معنی کلمات در جملات مختلف
۳	ترکیبی	دقت و ثبات بالا	معایب هر دو روش

از جدیدترین کارها در زمینه تحلیل احساسات، توسط جاشی و همکارانش [۱۹] انجام شده است. آن‌ها کارایی الگوریتم‌های بیز و ماشین بردار پشتیبان را با درنظر گرفتن ویژگی‌هایی از قبیل uni-gram ها، bi-gram ها و همچنین تمرکز فراوان بر مرحله پیش‌پردازش داده‌ها مثل درنظر گرفتن احساسات شکلک‌ها، حذف اعداد و کلمات غیرضروری و... پیاده‌سازی نمودند. آن‌ها علاوه بر استفاده از روش‌های یادگیری ماشین، روش پایه فرهنگ‌لغات یعنی همان شمارش تعداد کلمات مثبت و منفی در متن از طریق استفاده از یک فرهنگ لغت را نیز پیاده سازی کردند. نتایج به‌دست آمده نشان داده است که با بردار ویژگی یکسان در هر دو روش، روش‌های یادگیری ماشین عملکرد بسیار بهتری از خود نشان دادند.

۲-۴- تحلیل احساسات با استفاده از شکلک[۱] ها

[۱] emoticon

بسیاری از پژوهش ها در زمینه تحلیل احساسات، شکلک های موجود در متن های شبکه‌های اجتماعی از قبیل توییتر را نادیده گرفته و یا توجه چندانی به شکلک‌ها ندارند. امروزه شکلک‌ها بسیار معروف و پر استفاده هستند و می توان گفت جزء جدانشدنی ارتباطات متنی هستند. در بین تحقیقات انجام‌شده در تحلیل احساسات، پژوهش هایی با استفاده از شکلک‌ها نیز انجام گرفته‌است. یکی از پژوهش های اخیر در این زمینه توسط بهری و همکارانش [۲۰] در سال ۲۰۱۸ صورت گرفته است. آن ها الگوریتمی ارائه‌دادند که با استفاده از یادگیری مبتنی بر امتیاز شکلک ها[۱] قطبیت جمله را تشخیص می‌دهند. روش کار به این صورت است که ابتدا لیستی از پراستفاده ترین شکلک‌ها را تهیه‌کردند که برای این منظور از رابط کاربردی برنامه[۲]ی توییتر استفاده نمودند. سپس به هر یک از شکلک‌ها بر اساس درجه مثبت و منفی بودن امتیازی اختصاص می‌دهند. اختصاص امتیاز در این مرحله به صورت دستی نبوده و بر اساس محتوای متن‌هایی است که شکلک در آن‌ها استفاده شده‌است. مرحله‌ی آخر، مرحله‌ی امتیازدهی به کل جمله است که با استفاده از فرهنگ لغت AFFIN-111 و امتیازدهی به هر یک از کلمات انجام می‌گیرد. سپس مجموع امتیازات کل کلمات توییت‌ها محاسبه می‌شود و با توجه به مثبت یا منفی بودن امتیاز کلمات و شکلک‌ها، دسته‌ی مورد نظر متن مشخص می‌شود. نتایج این پژوهش نشان داد که الگوریتم ارائه‌شده در زمینه تحلیل احساسات بسیار کاربردی بوده و دارای دقت ۹۱/۱ درصد است.

۲-۵- طبقه‌بندی کارهای پیشین بر اساس زبان متن

یکی از بزرگترین چالش‌های تحلیل احساسات، وابسته بودن رویکردهای ارائه شده به زبان‌های مختلف است. ترکیب کلمات، فرهنگ لغات و حتی داده‌های موجود، متعلق به زبان‌های خاصی هستند و به همه‌ی زبان‌ها در دسترس نمی‌باشند. علاوه بر آن، بهبود روش‌ها و مدل‌ها برای زبان‌های مختلف به ویژه برای شبکه‌های عصبی

[۱] emoticon score learning

[۲] API

کاری زمان‌بر بوده و نیازمند نیرو و منابع بیشتری است. همچنین جمع آوری داده در زبان‌های متفاوت کاری بسیار مشکل و چالش برانگیز است. کارهای پیشین در این رابطه در دو دسته‌ی مستقل از زبان و وابسته به زبان انجام گرفته‌اند که در ادامه به بررسی هر کدام پرداخته می‌شود.

۲-۵-۱- مستقل از زبان

از کارهای اخیر در این رابطه می توان به توماس و همکارانش [۲۱] در سال ۲۰۱۹ اشاره کرد. آن ها مدلی مستقل از زبان را برای تحلیل احساسات ارائه دادند که با استفاده از ویژگی‌های همسایگی هر متن کارایی طبقه بند را بهبود می‌بخشد. از چالش های این تحقیق این است که محدود به چند زبان مشخص بوده و نیازمند فرایندی خودکار برای شناسایی محیط اطراف متن مورد بررسی است. همچنین طبقه بند ماشین بردار پشتیبان به رغم کند بودن به دلیل پیچیدگی‌های محاسباتی، کارایی بهتری نسبت به روش‌های دیگر دارد. .

۲-۵-۲- وابسته به زبان

معمولا کارهای پیشین انجام شده در تحلیل احساسات وابسته به زبان هستند و اکثرا مربوط به زبان انگلیسی می باشند. زبان هایی مثل فارسی و عربی که بیشترین تفاوت را از لحاظ نوع نوشتار، حروف، سمت نوشتن و از این قبیل دارند کمتر مورد توجه قرار گرفته‌اند، به ویژه در استفاده از روش‌هایی که معنی لغات و ساختار جملات حائز اهمیت است.

مازن و همکارانش [۲۲] روشی برای طبقه‌بندی توییت‌های عربی ارائه داده‌اند. آن‌ها روش پیشنهادی خود را هم بر اساس روش‌های مبتنی بر فرهنگ لغت و هم براساس روش‌های یادگیری ماشین توسعه دادند. آن‌ها برای الگوریتم‌های یادگیری ماشین از الگوریتم بیز ساده و ماشین بردار پشتیبان بهره بردند که در نهایت الگوریتم بیز عملکرد بهتری را دربرداشت. در رابطه با روش‌های مبتنی بر فرهنگ لغت نیز با وجود فرهنگ لغات کوچکی که برای این زبان به کار بردند به نتیجه‌ی مطلوبی رسیدند.

۲-۶- طبقه‌بندی کارهای پیشین بر اساس تعداد کلاس‌ها

هدف عمده از تحلیل احساسات طبقه‌بندی متن به کلاس‌ها است. با توجه به مجموعه داده و هدف از طبقه‌بندی، طبقه‌بندی احساسات می تواند به صورت باینری (دو کلاسی) و چندکلاسی باشد. در این بخش به بررسی کارهای پیشین بر اساس تعداد کلاس‌های مورد نظر جهت طبقه‌بندی پرداخته می‌شود.

۲-۶-۱- طبقه‌بندی دو کلاسی[۱]

در طبقه‌بندی دو کلاسی یا طبقه‌بندی باینری، هر یک از متن‌های مورد نظر به دو کلاس اصلی تقسیم می‌شوند که می‌تواند به صورت مثبت و منفی و یا صفر و یک و هر دو صورت ممکن دیگری باشد.

۲-۶-۲- طبقه‌بندی چند کلاسی[۲]

در طبقه‌بندی چند کلاسی، هر یک از متون به بیشتر از دو کلاس تقسیم می‌شوند. به عنوان مثالی از این دست می‌توان به امتیازدهی اشاره کرد که از ۱ تا ۵ ستاره می‌تواند متغیر باشد. همچنین طبقه‌بندی به صورت مثبت، منفی و خنثی که در اکثر مقالات و پژوهش‌ها مورد توجه قرارمی‌گیرد از عمده طبقه‌بندی‌های احساسات می‌باشد. از جمله کارهای انجام شده در این زمینه می توان به سانگ و همکارانش [۲۳] و توماس و همکارانش [۲۱] اشاره کرد.

۲-۷- نتیجه‌گیری

در این فصل به بررسی و دسته‌بندی کارهای پیشین در حوزه‌های مختلف پرداخته شد. هر یک از روش‌های پیش گرفته شده در کارهای گذشته دارای نقاط قوت و ضعفی بودند که بعضا در کارها به آن‌ها اشاره شد. همانطور که بررسی شد تحلیل احساسات با استفاده از روش‌های مبتنی بر الگوریتم‌های یادگیری ماشین دقت بالاتری نسبت به روش‌های مبتنی بر لغت‌نامه دارند و در صورت داشتن مجموعه‌داده‌ی برچسب‌گذاری شده

---

۱ binary classification

۲ multi-way classification

بهترین انتخاب می‌باشند. اگرچه الگوریتم‌های ترکیبی نیز از اعمال دو روش ذکرشده به تنهایی، عملکرد بهتری خواهد داشت اما وجود محدودیت در انتخاب فرهنگ‌لغت و مجموعه‌داده متناسب با آن از جمله معایب این روش است.

فصل ۳: روش پیشنهادی

۳-۱- مقدمه

هدف اصلی این نوشتار ارائه مدلی مناسب برای تحلیل احساسات در متن می‌باشد. در این تحقیق به مطالعه بر روی روش‌های پیش‌پردازش داده‌ها و انتخاب مناسب ویژگی‌ها و وزن‌دهی به آن‌ها پرداخته شده‌است. سپس به ارائه مدلی برای حصول نتیجه‌ی بهتر در طبقه‌بندی مجموعه‌داده انتخابی انجامیده است.

در این فصل به تشریح روش پیشنهادی پرداخته خواهدشد. پس از بررسی مقالات و تحقیقات ارائه‌شده در زمینه تحلیل احساسات در متن درمی‌یابیم که یکی از مهم‌ترین مراحل تحلیل احساسات مرحله پیش‌پردازش داده‌ها و انتخاب ویژگی است. در مرحله پیش‌پردازش هدف اصلی حذف تعداد بیشتری از کلمات و عبارات زاید است، در نتیجه‌ی این مرحله کلماتی که تاثیری در نتیجه‌ی تحلیل احساسات ندارند، حذف شده و فقط کلمات احساسی باقی می‌مانند. این کاهش کلمات سبب افزایش سرعت و همچنین دقت طبقه‌بندی می‌شود.

از جمله چالش‌های دیگر پردازش متن که در انتخاب ویژگی به جهت برطرف‌کردن آن گام برداشته می‌شود، تعداد زیاد ویژگی‌ها است. زیاد بودن تعداد ویژگی ها باعث وجود ویژگی‌های غیرمرتبط می‌شود. بنابراین انتخاب روش مناسب برای مرحله انتخاب ویژگی می‌تواند در نتیجه طبقه‌بندی تاثیر به‌سزایی داشته باشد.

از دیگر مراحل پردازش متن وزن‌دهی به صفات است. هدف از این مرحله اختصاص وزن بالاتر به کلمات پرارزش‌تر احساسی و اختصاص وزن کمتر به کلمات کم اهمیت‌تر است. به این ترتیب شانس کلمات مهم‌تر برای تعیین دسته‌ی متن بیشتر شده و در نتیجه طبقه‌بندی دقیق‌تری خواهیم‌داشت.

در این تحقیق پس از انتخاب مجموعه‌داده، پیش‌پردازش‌هایی روی آن‌ها انجام می‌گیرد. در مرحله بعد ویژگی‌های موردنیاز استخراج شده و سپس وزن‌دهی می‌شوند. در نهایت متن‌ها با داشتن ویژگی‌ها طبقه‌بندی می‌شوند. در ادامه به تشریح دقیق‌تر هر یک از مراحل پرداخته می‌شود.

۳-۲- مجموعه‌داده

هدف از تحلیل احساسات در واقع مطالعه بر روی متون از جمله پست‌ها و نظراتی است که توسط افراد مختلف در میکروبلاگ‌ها و فروم‌ها و شبکه‌های اجتماعی منتشر می‌شود. نظرات افراد در قشرهای مختلف درباره‌ی محصولات، خدمات، افراد و رویدادها متفاوت بوده و به این وسیله تحلیل این نظرات برای تصمیم گیری‌های سازمان‌ها و مراکز تجاری و ... می‌تواند مفید واقع شود.

یکی از مهم‌ترین قسمت‌ها در تحلیل احساسات انتخاب مجموعه‌داده‌ی مناسب است. در این مرحله هم می‌توان از مجموعه‌های داده‌های گردآوری شده و موجود استفاده نمود و هم می‌توان با توجه به زمینه‌ی مورد مطالعه و عدم دسترسی به مجموعه‌داده مورد نظر به جمع‌آوری مجموعه‌داده پرداخت. به منظور اینکه کار انجام‌شده در این تحقیق قابل مقایسه با کارهای دیگران در سطح جهانی باشد و همچنین جهت اطمینان از اثر نتایج تحقیق انجام شده در پیش‌برد علم، روشی که در بخش‌های بعد به توضیح آن خواهیم پرداخت، براساس یک مجموعه‌داده استاندارد ارائه‌شده در سطح جهانی تست خواهد شد.

۳-۳- پیش‌پردازش

یکی از مراحل اولیه در پردازش متن پیش‌پردازش آن است، زیرا همه‌ی کلمات متن در تحلیل احساسات مفید نخواهند بود. ثابت شده‌است که تنها ۳۳ درصد از کلمات در یک متن مفید هستند و می‌توان از آن‌ها برای استخراج اطلاعات استفاده نمود. اغلب کلمات در راستای رساندن منظور و هدف اصلی استفاده می‌شوند و بعضا تکراری هستند. بنابراین پیش پردازش، اولین گام در جهت تطابق متون با نمایش آن‌ها در یک قالب مناسب می‌باشد. قالب و فرمت متن باید به گونه‌ای تغییر یابد تا علاوه بر این که بتواند مدل مناسبی از متن را ارائه دهد، به قالب ساده و مناسب برای پردازش در مراحل بعدی تبدیل شود [۲۴].

توییت‌های توییتر حاوی کلمات و جملات بسیار زیادی هستند و از آنجایی که توسط کاربران متفاوتی نوشته شده‌اند به روش‌های مختلفی نیز بیان شده و شامل کلمات زائد بسیاری هستند. در تحلیل احساسات باید کلمات و عبارات غیرضروری را حذف کنیم. مرحله‌ی حذف عبارات زائد و علائم اضافی از مجموعه داده مرحله پیش پردازش نامیده می‌شود و در نهایت مجموعه داده نسبتا خالصی خواهیم داشت که تا حد ممکن عبارات احساسی و مهم داراست.

۳-۴- استخراج ویژگی[۱]

یکی از مشکلات در طبقه‌بندی متون ابعاد بالای فضای ویژگی است. به بیان ساده‌تر تعداد ویژگی‌ها بسیار زیاد هستند. بسیاری از این ویژگی‌ها نامربوط و غیرمفید بوده و تاثیری در نتیجه طبقه‌بندی ندارند. حتی بسیاری از ویژگی‌ها نویز به شمار رفته و موجب کاهش دقت طبقه‌بندی می‌شوند. علاوه بر این، تعداد زیاد ویژگی‌ها فرایند طبقه‌بندی را زمان‌بر کرده و یا حتی پیاده‌سازی بسیاری از الگوریتم‌های طبقه‌بندی را غیرممکن می‌سازند. استخراج ویژگی فرایندی است که در آن با انجام عملیاتی بر روی داده‌ها، ویژگی‌های بارز و تعیین‌کننده‌ی آن مشخص می‌شود تا ابعاد فضای ویژگی کاهش یافته و کارایی و دقت طبقه‌بند افزایش یابد [۲۵]. هدف از استخراج ویژگی این است که داده‌های خام به شکل قابل استفاده‌تری برای پردازش‌های آماری بعدی درآیند. روش‌های مختلف استخراج ویژگی بنا به هدف تحقیق ممکن است انجام شود. از آنجا که الگوریتم یادگیری ماشین به کار گرفته شده در این تحقیق الگوریتم بیز ساده است و الگوریتم بیز به مرحله استخراج ویژگی بسیار حساس است در این تحقیق از دو روش برای استخراج ویژگی بهره گرفته شده که در ادامه شرح داده خواهد شد.

۳-۴-۱- ویژگی های n-gram

[۱] feature extraction

n-gram درواقع n کلمه‌ی متوالی موجود در متن است. این مدل برای پیش‌بینی مورد بعدی در یک متن کارایی دارد. در تحلیل احساسات از n-gram ها برای کمک به تحلیل احساس دقیق‌تر مورد استفاده قرار می‌گیرد. Unigram همان n-gram با اندازه یک، Bigram همان n-gram با اندازه دو و Trigram نیز n-gram با اندازه سه است [۲۶].

در بسیاری از پژوهش‌ها از ویژگی‌های unigram برای طبقه‌بندی متن استفاده شده‌است. اگرچه به‌کاربردن این ویژگی‌ها نتایج خوبی دربردارد اما در بسیاری موارد منجر به شکست می‌شود. به عنوان مثال در جمله‌ی "the item is not good" در صورتی که تنها از unigram استفاده شود متن با احتمال برابر می‌تواند در دسته مثبت یا منفی قرار گیرد زیرا شامل یک لغت مثبت "good" و یک لغت منفی"not" است. اگر جمله فوق با ویژگی‌های bigram مورد بررسی قرارداده شود شامل ترکیب‌های "the item"، "item is"، "is not" و "not good" می‌شود. از آنجا که کلمه منفی "not" تاثیر بیشتری در ترکیب‌ها می‌گذارد در نتیجه جمله فوق در دسته منفی طبقه‌بندی می‌شود. همین اتفاق در trigram نیز تکرار می‌شود. اگرچه افزایش مقدار n الزاما موجب بهبود نتایج نمی‌شود و بستگی به شرایط دارد.

۳-۲-۴- TF-IDF[۱]

وزن یک کلمه در یک متن به روش‌های مختلفی می‌تواند تعیین شود. یکی از این روش‌ها TF-IDF است.

TF نسبت تعداد تکرارهای یک کلمه در یک متن به تعداد کل کلمات آن متن است. در نتیجه با افزایش تعداد تکرار یک کلمه در متن میزان TF افزایش می‌یابد. هر متن دارای TF یکتاست و توسط رابطه ۳-۱ محاسبه می‌گردد که در واقع به معنی تعداد تکرارهای کلمه i در متن j است.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \qquad (۱-۳)$$

[۱] Term Frequency – Inverse Data Frequency

IDF برای محاسبه وزن کلمات کم‌تکرار در همه‌ی متن‌ها استفاده می‌شود. هر چه کلمات کم‌تکرارتر باشند دارای IDF بالاتری خواهند بود. IDF توسط رابطه ۳-۲ محاسبه می‌شود. در این رابطه N تعداد کل متن‌ها و df_i تعداد متن‌های حاوی کلمه i است.

$$idf(w) = \log\left(\frac{N}{df_i}\right) \qquad (۳-۲)$$

ترکیب دو رابطه بالا به صورت رابطه ۳-۳ تعریف می‌شود.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \qquad (۳-۳)$$

جدول ۳-۱. مقادیر TF-IDF برای کلمات دو جمله نمونه

کلمه	TF		IDF	TF*IDF	
	جمله اول	جمله دوم		جمله اول	جمله دوم
The	1/7	1/7	Log (2/2) =0	0	0
Car	1/7	0	Log (2/1) =0.3	0.043	0
Truck	0	1/7	Log (2/1) =0.3	0	0
Is	1/7	1/7	Log (2/2) =0	0	0
Driven	1/7	1/7	Log (2/2) =0	0	0
On	1/7	1/7	Log (2/2) =0	0	0
The	1/7	1/7	Log (2/2) =0	0	0
Road	1/7	0	Log (2/1) =0.3	0.043	0
Highway	0	1/7	Log (2/1) =0.3	0	0.043

به عنوان مثال دو جمله زیر را در نظر بگیرید:

جمله اول: The car is driven on the road

جمله دوم: The truck is driven on the highway

در این مثال هر جمله را یک متن در نظر گرفته می‌شود. TF-IDF این دو متن در جدول ۳-۱ قابل مشاهده است.

۳-۵- مدل‌سازی احتمالی

اگر n متغیر ورودی داشته باشیم یعنی $x = (x_1, \dots, x_n)$ و خروجی y از یک مجموعه K عضوی باشد، هدف از مدل سازی پیدا کردن احتمال شرطی هر کدام از این K دسته است یعنی $p(C_k|x_1, \dots, x_n)$. طبق قانون بیز این احتمال برابر است با:

$$p(C_k|x) = \frac{p(C_k, x)}{p(x)} \propto p(C_k, x) \quad (۴-۳)$$

فرمول ۳-۴ را می‌توان به زبان ساده این طور شرح داد که احتمال قرار گرفتن یک داده در یک دسته برابر است با تعداد داده‌های مشابه که قبلا در آن دسته قرار داشته‌اند. از آنجایی که داده‌ی ما شامل برداری از ویژگی‌ها است برای به دست آوردن احتمال وجود داده در دسته‌ی مورد نظر، طبق قانون زنجیره‌ای به رابطه‌های زیر می‌رسیم:

$$p(C_k, x_1, \dots, x_n) = p(x_1, \dots, x_n, C_k) \quad (۵-۳)$$

$$p(C_k, x_1, \dots, x_n) = p(x_1|x_2, \dots, x_n, C_k)\, p(x_2, \dots, x_n, C_k) \quad (۶-۳)$$

$$p(C_k, x_1, \dots, x_n) = p(x_1|x_2, \dots, x_n, C_k)\, p(x_2|x_3, \dots, x_n, C_k)\, p(x_3, \dots, x_n, C_k) \quad (۷-۳)$$

$$p(C_k, x_1, \dots, x_n) = \cdots \quad (۸-۳)$$

$$p(C_k, x_1, \dots, x_n) = p(x_1|x_2, \dots, x_n, C_k)\, p(x_2|x_3, \dots, x_n, C_k) \dots p(x_{n-1}|x_n, C_k)\, p(x_n|C_k)p(C_k) \quad (۹-۳)$$

حال اگر فرض کنیم هر متغیری نسبت به متغیرهای دیگر به شرط داشتن دسته C_k مستقل است یعنی $(x_i|x_{i+1}, \dots, x_n, C_k) = p(x_i|C_k)$ به نتیجه می‌رسیم:

$$p(C_k|x_1, \dots, x_n) \propto p(C_k, x_1, \dots, x_n) \quad (۱۰-۳)$$

$$p(C_k, x_1, \dots, x_n) = \ p(C_k)\ p(x_1|C_k)\ p(x_2|C_k)\ p(x_3|C_k) \dots \quad (۱۱-۳)$$

$$p(C_k, x_1, \dots, x_n) = p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \quad (۱۲-۳)$$

با نرمال‌سازی عبارت قبلی می‌توان توزیع احتمال شرطی را پیدا کرد، در معادله ۳-۱۳ $Z = \ p(C_k) = \sum_k p(C_k)\ p(x|C_k)$ همان ضریب نرمال‌سازی است.

$$p(C_k|x_1, \dots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \quad (۱۳-۳)$$

اگر هدف پیدا کردن دسته با بالاترین احتمال باشد، به ضریب نرمال‌سازی یعنی Z نیازی نیست و می‌توان گفت احتمال قرارگرفتن داده x در کلاس C از رابطه ۳-۱۴ محاسبه می شود.

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}}\, p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \quad (۱۴-۳)$$

۳-۶- محاسبه احتمال با روش بیز

بیز ساده یکی از معروف‌ترین و پرکاربردترین روش‌های طبقه‌بندی با استفاده از تکنیک های یادگیری ماشین است که بر گرفته از قضیه بیز می‌باشد. بیز ساده به دلیل سادگی، کارایی محاسباتی بالا و طبقه‌بندی با دقت مناسب به خصوص برای داده‌های با ابعاد بالا مانند متن، یک ابزار محبوب یادگیری ماشین به شمار می‌رود [۲۷].

به طور کلی در مدل‌های بیزی، طبقه‌بندی با استفاده از وابستگی (یا وابستگی شرطی) بین متغیرهای تصادفی انجام می‌گیرد [۳] [۲۳]. این فرایند معمولا فرایندی زمان بر است زیرا بررسی رابطه‌ی بین تمام متغیرهای

تصادفی یک عمل بهینه‌سازی ترکیبی است. روش بیز با فرض اینکه صفات به صورت شرطی مستقل هستند، راه حلی برای این ساختار وابستگی بین صفات است [۲۷].

در تحلیل احساسات و به طور کلی پردازش متن، صفت همان کلمات معنی‌دار موجود در متن مورد نظر جهت طبقه‌بندی هستند. به عنوان مثال در جمله "He plays football"، صفات "He"، "plays" و "football" هستند و در الگوریتم بیز با داشتن کلاس‌ها، صفات به صورت شرطی مستقل فرض می‌شوند. بر اساس این فرض طبقه بند به صورت ۳-۱۵ تعریف می‌شود.

$$c(d) = \arg\max_{c \in C} P(c) \prod_{i=1}^{m} P(a_i|c) \qquad (۱۵-۳)$$

که در آن d متن تست، $c(d)$ کلاس d و c نوع کلاس است. مجموعه‌ی همه ی کلاس‌ها C و m تعداد صفات و a_i مقدار هر صفت A_i $(1,2,\dots,m)$ است. $P(c)$ احتمال پیشین کلاس c و $P(a_i|c)$ احتمال شرطی هر صفت از مجموعه آموزشی است.

بیز چند جمله‌ای[۱] یک مدل بر پایه فراوانی[۲] برای طبقه بندی متن است که تعداد تکرارهای کلمات موجود در متن مورد بررسی قرار می‌گیرند. در طبقه‌بندی متن با بیز چند جمله‌ای، متن مجموعه تست d که با بردار $< w_1, w_2, \dots, w_m >$ نشان داده می‌شود، توسط رابطه ۳-۱۶ طبقه‌بندی می‌شود.

$$c(d) = \arg\max_{c \in C} \left[\log P(c) + \sum_{i=1}^{m} f_i \log P(w_i|c) \right] \qquad (۱۶-۳)$$

که m تعداد کلمات مختلف موجود در متن، w_i $(i = 1,2,\dots,m)$ نیز i-امین کلمه ظاهر شده در متن و f_i $(i = 1,2,\dots,m)$ تعداد w_i ها در d است. در بیز چند جمله‌ای احتمال پیشین $P(c)$ توسط رابطه ۳-۱۷ محاسبه می‌گردد.

---

[۱] MNB (Multinomial Naïve Bayes)

[۲] frequency based

$$P(c) = \frac{\sum_{j=1}^{n} \delta(c_j, c) + 1}{n + 1} \quad (۱۷-۳)$$

علاوه‌براین، احتمال شرطی $P(w_i|c)$ از رابطه ۳-۱۸ تخمین زده می‌شود.

$$P(w_i|c) = \frac{\sum_{j=1}^{n} f_{ji}\delta(c_j, c) + 1}{\sum_{j=1}^{m}\sum_{j=1}^{n} f_{ji}\delta(c_j, c) + m} \quad (۱۸-۳)$$

که در آن n تعداد متن‌های آموزشی است، l تعداد کلاس‌ها، f_{ji} تعداد w_i در j-امین متن مجموعه آموزشی و c_j کلاس j-امین متن مجموعه آموزشی است. تابع باینری $\delta(c_j, c)$ از رابطه ۳-۱۹ به دست می‌آید.

$$\delta(c_j, c) = \begin{cases} 1, & if\ c_j = c \\ 0, & otherwise \end{cases} \quad (۱۹-۳)$$

بیز چند متغیره‌ی برنولی[۱] مدل آماری معروف دیگری است که برای طبقه‌بندی متن ارائه شده‌است. برخلاف بیز چند جمله‌ای، در بیز چند متغیره‌ی برنولی فرض بر این است که هر ویژگی در متن به عنوان یک متغیر باینری مستقل تعریف می‌شود. بیز چند متغیره‌ی برنولی تنها وجود و یا عدم وجود یک کلمه را بدون در نظر گرفتن تعداد تکرار آن در متن مورد بررسی قرار می‌دهد.

۳-۷- وزن‌دهی به صفات[۲]

روش‌های بهبود کارایی بیز از طریق الگوریتم‌های یادگیری ماشین به پنج گروه طبقه‌بندی می‌شوند [۲۳]:

۱. توسعه ساختاری
۲. انتخاب ویژگی
۳. وزن دهی به صفات
۴. یادگیری محلی

[۱] BNB (Multivariate Bernoulli Naïve Bayes)

[۲] attribute weighting

۵. گسترش اطلاعات

از بین این روش‌ها وزن دهی به صفات که به هر صفت وزن‌های متفاوتی اختصاص می‌دهد و انتخاب ویژگی که زیر مجموعه‌ای از صفات را بر اساس وزن‌های آن ها انتخاب می‌کند، به طور گسترده به کار برده می‌شوند.

بیز ساده به همراه وزن‌دهی به صفات به صورت رابطه ۳-۲۰ تعریف می‌شود.

$$c(d) = \arg\max_{c \in C} P(c) \prod_{i=1}^{m} P(a_i|c)^{WT_i} \quad (۲۰-۳)$$

که WT_i وزن A_i $(i = 1,2, \dots, m)$ است. بیز چند جمله‌ای همراه با وزن‌دهی به صفات به صورت معادله ۳-۲۱ مدل‌سازی می‌شود:

$$c(d) = \arg\max_{c \in C} \left[\log P(c) + \sum_{i=1}^{m} WT_i f_i \log P(w_i|c) \right] \quad (۲۱-۳)$$

که WT_i وزن هر کلمه w_i $(i = 1,2, \dots, m)$ است.

۳-۸- وزن‌دهی به صفات بر اساس نسبت بهره[۱]

چگونگی تعریف نسبت بهره‌ی هر یک از صفات (کلمات) مجموعه داده‌ی آموزش، یکی از نکات مهم در وزن‌دهی به صفات بر اساس نسبت بهره است. یک مجموعه داده استاندارد برای طبقه‌بندی متن مجموعه‌ای از متون است، به طوری که هر متن به صورت برداری از کلمات به همراه تعداد تکرار هر کلمه در آن متن نمایش داده می‌شود. به دلیل وسعت دایره لغات، داده‌های طبقه‌بندی متن معمولا یک ماتریس اسپارس است. هر ویژگی در این ماتریس دارای مقدار صفر یا عدد صحیح مثبت است. علاوه‌بر این بیشتر این مقادیر صفر بوده و مقادیر

[۱] gain ratio-based

بالاتر از یک بسیار اندک هستند. بنابراین در تعریف نسبت بهره اطلاعات فرض بر این است که ویژگی‌ها فقط دارای دو مقدار صفر و غیر صفر هستند [۱۸].

برای مجموعه آموزشی D داده شده، نسبت بهره اطلاعات $IGR(C, w_i)$ برای هر کلمه w_i به صورت رابطه ۳-۲۲ تعریف می‌شود:

$$IGR(C, w_i) = \frac{IG(C, w_i)}{H(w_i)} \quad (۲۲-۳)$$

که در آن C مقدار نهایی، $IG(C, w_i)$ بهره اطلاعات هر کلمه w_i و $H(w_i)$ جداکننده اطلاعات هر کلمه w_i است. براساس مشاهدات و فرضیات روی داده‌های متنی $IG(C, w_i)$ به صورت رابطه ۳-۲۳ تعریف می‌شود:

$$IG(C, w_i) = H(C) - H(C|w_i) \quad (۲۳-۳)$$

که در این رابطه $H(C)$ آنتروپی D و $H(C|w_i)$ آنتروپی شرطی D به شرط داشتن w_i است که می‌تواند توسط دو رابطه ۳-۲۴ و ۳-۲۵ محاسبه شود:

$$H(C) = -\sum_{c} P(c) \log_2 P(c) \quad (۲۴-۳)$$

$$H(C|w_i) = -\sum_{v} \frac{|D_v|}{|D|} P(c|v) \log_2 P(c|v) \quad (۲۵-۳)$$

که در آن $P(c)$ احتمال کلاس c در D و $|D_v|$ تعداد متن‌هایی است که مقدار w_i در آن ها v است. ($v \in \{0, \bar{0}\}$)

$H(w_i)$ نیز با توجه به مقدار هر کلمه w_i آنتروپی D است و به صورت رابطه ۳-۲۶ تعریف می‌شود:

$$H(w_i) = -\sum_{v} \frac{|D_v|}{|D|} \log_2 \frac{|D_v|}{|D|} \quad (۲۶-۳)$$

هنگامی که نسبت بهره $IGR(C,w_i)$ هر کلمه w_i $(i = 1,2,\dots,m)$ به دست آمد، می توان مجموع نسبت بهره اطلاعات همه کلمات را به دست آورد و وزن W_i هر کلمه w_i $(i = 1,2,\dots,m)$ طبق رابطه ۳-۲۷ تعریف می‌شود:

$$W_i = \frac{IGR(C,w_i) \times m}{\sum_{i=1}^{m} IGR(C,w_i)} \qquad (۲۷-۳)$$

با دست آوردن مقادیر وزن W_i هر یک از کلمات w_i $(i = 1,2,\dots,m)$ با استفاده از رابطه بالا، وزن دهی به صفات با استفاده از نسبت بهره در بیز ساده به دست می‌آید.

طبق تحقیقات ونگ و همکارانش [۲۸] در هم آمیختگی وزن ویژگی‌ها باعث بهبود عملکرد طبقه‌بندی خواهد شد. این بهبود هم در طبقه‌بندی متن با استفاده از بیز ساده و هم در تخمین‌های احتمالات شرطی آن‌ها قابل ملاحظه خواهد بود. در نتیجه احتمال شرطی بیز چندجمله‌ای $P(w_i|c)$ به صورت رابطه ۳-۲۸ اصلاح شده و تعریف می‌شود:

$$P(w_i|c) = \frac{\sum_{j=1}^{n} WT_i f_{ji} \delta(c_j, c) + 1}{\sum_{j=1}^{m}\sum_{j=1}^{n} WT_i f_{ji} \delta(c_j, c) + m} \qquad (۲۸-۳)$$

۳-۹- اصلاح وزن‌ها

بسیاری از رویکردهای موجود برای تحلیل احساسات از تعداد ثابتی صفت (کلمه) استفاده می‌کنند. این درحالی است که تعداد کلمات موجود در توییت های توییتر بسیار زیاد و غیرقابل شمارش هستند. بنابراین این رویکردها نمی توانند برای چنین مجموعه داده‌هایی موثر واقع شوند. علاوه بر آن توییت‌های توییتر حاوی تعداد زیادی کلمات بی معنی مثل خطاهای تایپی، اسامی خاص و... هستند. در روش ارائه شده در این تحقیق کلمات بی‌معنی وزن بالاتری نسبت به کلمات مهم و پرتکرار می‌گیرند. ابتدا اختلاف وزن WD_i هر کلمه w_i $(i = 1,2,\dots,m)$ از رابطه ۳-۲۹ محاسبه می‌شود:

$$WD_i = \begin{cases} |WT_{i,p} - WT_{i,n}|, & if\ W_i = D_p and\ W_i = D_n \\ WT_{i,c}, & otherwise \end{cases} \quad (۲۹-۳)$$

میانگین WD_i ها نیز طبق رابطه ۳-۳۰ به دست می‌آید:

$$Avg_{WD} = \frac{\sum_{i=1}^{m} WD_i}{m} \quad (۳۰-۳)$$

Avg_{WD} و وزن همه کلمات در هر مجموعه آموزشی D_c با هم مقایسه می‌شود. در صورتی که مقدار وزن بزرگتر از میانگین باشد، مقدار آن به صفر تغییر می‌کند. در غیر این صورت بدون تغییر باقی می‌ماند. رابطه ۳-۳۱ بیانگر توضیحات داده شده‌است:

$$WT_{i,c} = \begin{cases} 0, & if\ WT_{i,c} > Avg_{WD} \\ WT_{i,c}, & otherwise \end{cases} \quad (۳۱-۳)$$

در نهایت، کلاس متن تست بر اساس مقادیر وزن غیرصفر و با استفاده از رابطه ۳-۲۱ پیش‌بینی می‌شود.

۳-۱۰- ارزیابی کارایی[۱]

برای بیان کمی چگونگی عملکرد سیستم، باید از معیارهای اندازه‌گیری درستی عملکرد طبقه‌بندی استفاده نمود. معیارهای ارزیابی مختلفی وجود دارند که ارزیابی نهایی سیستم بر اساس آن ها انجام می‌شود. مهم‌ترین این معیارها دقت[۲]، یادآوری[۳] و صحت[۴] می‌باشد که توسط روابط زیر تعریف می‌شوند:

$$Precision = \frac{TP}{TP + FP} \quad (۳۲-۳)$$

$$Recall = \frac{TP}{TP + FN} \quad (۳۳-۳)$$

[۱] performance evaluation

[۲] precision

[۳] recall

[۴] accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (۳۴-۳)$$

در این روابط، TP تعداد مثبت‌های صحیح، TN تعداد منفی‌های صحیح، FP تعداد مثبت‌های غلط، FN تعداد منفی‌های غلط هستند. جدول زیر ماتریس درهم ریختگی[۱] موارد ذکر شده را نشان می‌دهد. به بیان دیگر، معیار FP تعداد نمونه‌هایی را مشخص می‌کند که نباید در یک کلاس قرار می‌گرفته‌اند ولی به اشتباه در آن کلاس قرار گرفته‌اند و معیار FN تعداد نمونه‌هایی را مشخص می‌کند که باید در کلاس مشخص شده قرار می‌گرفته‌اند ولی در کلاس دیگری قرار گرفته‌اند.

جدول ۳-۲. ماتریس درهم ریختگی معیارهای اندازه‌گیری عملکرد طبقه‌بندی

واقعی / پیش بینی	مثبت	منفی
مثبت	TP	FN
منفی	FP	TN

Precision دقت، Recall یادآوری و Accuracy صحت خروجی‌های طبقه‌بند را مشخص می‌کنند. به بیان دیگر دقت مشخص می‌کند که چند درصد از نمونه‌هایی که در یک کلاس قرار داده شده‌اند، در واقع به آن کلاس تعلق داشته‌اند. یادآوری مشخص می‌کند که چند درصد از نمونه‌های آزمایشی یک کلاس به درستی در کلاس خود قرار داده شده‌اند. در نهایت صحت به عنوان اصلی‌ترین معیار نشان می‌دهد که چند درصد از کل نمونه‌های ورودی یک سیستم به درستی در کلاس خود قرار داده شده‌اند. می توان گفت صحت همان میانگین یادآوری کلاس‌ها است.

[۱] confusion matrix

۳-۱۱- نتیجه‌گیری

همانطور که در این فصل بررسی شد یکی از تاثیرگذارترین بخش‌های تحلیل احساسات پیش از اعمال روش موردنظر برای طبقه‌بندی و در مرحله پیش پردازش صورت می‌گیرد. در این تحقیق سعی می‌شود با تمرکز بر روی مراحل انتخاب ویژگی و وزن‌دهی به آن‌ها دقت طبقه‌بند بیز را تا حد امکان بهبود دهیم. پیاده‌سازی روش‌های مورد بررسی در این فصل با استفاده از منابع داده‌ای و نرم افزاری در فصل آینده انجام خواهد شد. با بررسی نتایج حاصل، عملکرد این روش‌ها مورد ارزیابی قرارگرفته و مناسب‌ترین مقادیر برای پارامترهای استفاده شده مشخص می‌گردد.

فصل ۴: پیاده‌سازی

۴-۱- مقدمه

در این فصل فرایند پیاده‌سازی روش‌های بیان شده در فصل قبل تشریح خواهد شد. برای انجام پیاده‌سازی این تحقیق از زبان برنامه نویسی پایتون بهره گرفته شده است. پایتون یک زبان برنامه نویسی همه منظوره، سطح بالا، شیءگرا، اسکریپتی و متن باز است که فلسفه‌ی ایجاد آن تأکید بر خوانایی بالای برنامه‌های نوشته شده و کوتاهی و بازدهی نسبی بالای آن است. این زبان با بهره‌گیری از کتابخانه‌های قدرتمند ازجمله NLTK، Scikit-learn، SciPy، NumPy و Pandas یکی از قوی‌ترین زبان‌های برنامه‌نویسی در حوزه پردازش زبان‌های طبیعی و هوش مصنوعی می‌باشد. در ادامه به تشریح کامل مجموعه‌داده انتخاب شده، پردازش‌های انجام گرفته برروی آن، انتخاب ویژگی‌ها و سپس انجام طبقه‌بندی با روش بیز پرداخته می‌شود.

۴-۲- انتخاب مجموعه‌داده

مجموعه‌داده مورد استفاده در این سیستم، مجموعه‌داده Sentiment140 شامل حدود ۱,۶۰۰,۰۰۰ نمونه از توییت‌های توییتر است. این مجموعه داده به وسیله رابط برنامه‌نویسی کاربردی[۱] (API) توییتر و برای استفاده در مقاله‌ی نوشته شده توسط گو و همکارانش [۲۹] در سال ۲۰۰۹ جمع آوری گردیده است. مجموعه‌داده ارائه شده حاوی توییت هایی در سه کلاس مثبت، منفی و خنثی هستند. در روش ارائه شده طبقه‌بندی در دو کلاس مثبت و منفی انجام شده‌است، به همین دلیل با حذف توییت‌های خنثی تعداد نمونه‌های مورد استفاده در این پژوهش به حدود ۱,۰۴۸,۰۰۰ توییت کاهش یافت.

ستون‌های مجموعه‌داده Sentiment140 شامل موارد زیر می‌باشد:

۱. هدف (target): قطبیت هر توییت شامل مثبت (۴)، منفی (۰) و خنثی (۲)

[۱] Application Programming Interface

۲. شناسه (ids): شناسه‌ی هر توییت

۳. تاریخ (date): تاریخ انتشار هر توییت

۴. پرچم (flag): پرس‌وجو[۱]، اگر پرس‌وجویی نداشته باشد مقدار NO_QUERY دارد.

۵. کاربر (user): کابر منتشر کننده‌ی توییت

۶. متن (text): متن توییت

۴-۳- پیش‌پردازش

در این پژوهش مرحله پیش‌پردازش شامل حذف موارد زیر از داده‌های توییتر است [۱۹]:

۱. توییت‌های غیر انگلیسی

۲. نقطه‌گذاری‌ها و علائم

۳. اعداد

۴. برچسب‌گذاری‌ها و هشتگ‌ها

۵. آدرس‌های اینترنتی[۲] و پست‌های الکترونیکی[۳]

علاوه بر حذف موارد ذکرشده، برخی از ویژگی‌های توییت‌های توییتر دست‌خوش تغییراتی شدند که در ادامه مورد بررسی قرارخواهندگرفت.

۴-۳-۱- شکلک‌ها

[۱] query

[۲] URLs (Uniform Resource Locator)

[۳] emails

شکلک[1]ها نیز در توییت‌های توییتر نقش قابل توجهی در قطبیت جملات دارند. بنابراین علاوه بر حذف عبارات زائد در مرحله پیش‌پردازش، مدیریت شکلک‌ها نیز می‌تواند در دستیابی به نتیجه‌ی بهتر در طبقه‌بندی توییت‌ها موثر واقع شود. با توجه به این که تعداد شکلک‌های موجود بسیار زیاد است و همچنان نیز در حال افزایش می باشند، در نظر گرفتن همه‌ی آن‌ها مشکل است اما می‌توان تعدادی از رایج‌ترین شکلک‌ها را مورد بررسی قرارداد. با توجه به این که هر شکلک احساس مثبت را می‌رساند یا منفی، با دو کلمه‌ی EMO_POS برای شکلک‌های مثبت و EMO_NEG برای شکلک‌های منفی جابه‌جا می شود. لیست شکلک‌های درنظرگرفته‌شده در این پژوهش در جدول ۴-۱ قابل مشاهده می‌باشد.

جدول ۴-۱. لیست شکلک‌های درنظر گرفته شده در تحقیق و عبارت جایگزین هرکدام

شکلک	نوع شکلک	الگوی نوشتاری	جایگزین
:), :), :-), (:, (:, (-:, :’)	لبخند	(:\s?\)|:-\)|\(\s?:|\(-:|:\’\))	EMO_POS
:D, : D, :-D, xD, x-D, XD, X-D	خنده	(:\s?D|:-D|x- ?D|X-?D)	EMO_POS
;-), ;), ;-D, ;D, (;, (-;	چشمک	(:\s?\(|:-\(|\)\s?:|\)- :)	EMO_POS
<3, :*	عشق	(<3|:*)	EMO_POS
:-(, : (, :(,):,)-:	ناراحت	(:\s?\(|:-\(|\)\s?:|\)- :)	EMO_NEG
:,(, :’(, :"(	گریه	(:,\(|:\’\(|:"\()	EMO_NEG

۴-۳-۲- آدرس‌های اینترنتی

[1] smile

کاربران معمولا در توییت‌های خود لینک‌هایی از صفحات و سایت‌های دیگر استفاده می‌کنند. آدرس‌های اینترنتی نه‌تنها نمی‌توانند برای طبقه‌بندی متن حائز اهمیت باشند بلکه می‌توانند منجر به طبقه‌بندی اشتباه شوند. بنابراین تمام آدرس‌های اینترنتی موجود در توییت‌ها با کلمه URL جابه‌جا می شوند. در قسمت پیش به این موضوع اشاره شد که آدرس‌های اینترنتی جزو حذفیات مرحله پیش‌پردازش است اما حذف به طور کامل صورت نمی‌گیرد و مقصود از حذف جایگزین کردن آدرس‌ها با یک کلمه خاص است تا بتوان در صورت نیاز از وجود و عدم وجود یک آدرس اینترنتی در توییت‌ها اطمینان حاصل کرد.

۴-۳-۳- ذکر کاربر[۱]

هر کاربر توییتر نشان[۲] مرتبط با خود را دارد که به وسیله آن کاربران دیگر می‌توانند آن کاربر را در توییت‌هایشان با نوشتن @handle صدا بزنند[۳]. در این پژوهش تمامی ذکر کاربران با USER_MENTION جابه‌جا می‌شوند.

۴-۳-۴- هشتگ[۴]ها

هشتگ‌ها عبارات بدون فاصله‌ای هستند که پس از نماد # نوشته می‌شوند. کاربران زمانی از هشتگ‌ها استفاده می‌کنند که می‌خواهند به یک موضوع رایج در توییتر اشاره کنند. در این پژوهش تمام نمادهای هشتگ حذف شده و کلمه‌ی پس از نماد باقی می‌ماند. به عنوان مثال #hello با hello جایگزین می‌شود.

۴-۳-۵- بازتوییت[۵]ها

---

۱ user mention

۲ handle

۳ mention

۴ hashtag

۵ retweet

بازتوییت‌ها توییت‌هایی هستند که توسط کاربران منتشر شده‌اند و کاربران دیگر آن‌ها را به اشتراک می‌گذارند. بازتوییت‌ها با RT شروع می‌شوند و از آنجایی که در نتیجه‌ی طبقه‌بندی متن تاثیری ندارند حذف گردیدند.

۴-۳-۶- ریشه‌یابی[۱]

ریشه‌یابی یکی از مهم‌ترین فرایندهای پیش‌پردازش است. معمولا در متون مختلف، کلماتی وجود دارند که از لحاظ معنایی نزدیک به هم هستند ولی به دو یا چند شکل مختلف نوشته می‌شوند و با توجه به محتوای جمله و نقش آن کلمه در جمله دارای املای متفاوتی هستند. به عنوان مثال کلمات "love"، "loving"، "lovingly"، "loved" و "lovely" همه برای اظهار علاقه نویسنده متن به چیزی بیان می‌شوند. الگوریتم‌های متن‌کاوی این نوع کلمات را به دو یا چند شکل مختلف شناسایی می‌کنند و نمی‌توانند رابطه‌ی معنایی بین این کلمات را پیدا کنند. در فرایند ریشه‌یابی همه‌ی لغات موجود در متن به شکل ریشه‌ی زبان‌شناختی اولیه لغت شکسته می‌شوند.

برای ریشه‌یابی کلمات از دو روش Stemming و Lemmatization استفاده می‌شود که هر دو روش قادرند ریشه‌ی یک کلمه را به دست آورند. در هر دو روش شکل ریشه‌ی کلمه‌ی موردنظر به دست می‌آید با این تفاوت که در Stemming ریشه‌ی به‌دست‌آمده کلمه‌ای واقعی نیست اما در Lemmatization ریشه‌ی حاصل کلمه‌ای واقعی در زبان است. در Stemming یک الگوریتم با تعداد گام‌های مشخص روی کلمات به اجرا درمی‌آید که سریع‌تر به نتیجه می‌رسد. درحالی‌که در Lemmatization از دیکشنری WordNet و دیکشنری دیگری برای حذف ایست واژه[۲]ها استفاده می‌شود که این امر سبب کندتر بودن Lemmatization نسبت به Stemming می‌شود. در صورتی که سرعت مد نظر باشد استفاده از روش Stemming مقرون به صرفه‌تر است

[۱] stemming

[۲] stop word

اما در صورتی که زبان از اهمیت بالاتری برخوردار است و از آنجا که Lemmatization از یک دیکشنری برای تطبیق و انتخاب ریشه‌ی مورد نظر استفاده می‌کند، Lemmatization روش مناسب‌تری به شمار می‌رود.

در این تحقیق از روش Stemming استفاده می‌شود زیرا محاسبات و بررسی‌ها به صورت آماری بوده و ساختار زبان مورد توجه قرار نمی‌گیرد.

به عنوان مثالی از Stemming، طی این فرایند هر سه لغت "automatic"، "automation" و "automate" به "automat" تبدیل می‌شوند. طبق آنچه گفته شد، Stemming موجب افزایش سرعت عملکرد الگوریتم می‌شود و در مواردی که دقت الگوریتم مدنظر باشد کاربرد چندانی نخواهد داشت. در جدول ۴-۲ دو نمونه از واژه‌های موجود در مجموعه‌داده Sentiment140 و شکل ریشه‌یابی شده‌ی آن مشاهده می‌شود. همان‌طور که مشخص است شکل ریشه‌یابی شده‌ی لغات کلمه‌های واقعی نیستند.

جدول ۴-۲. دو نمونه از واژه های مجموعه داده و شکل ریشه‌یابی شده‌ی هر یک

ردیف	کلمات	ریشه
۱	Played, playing, player, plays	play
۲	Likes, liked, like, likely, alike	like

جدول ۴-۳ نمونه‌هایی از توییت‌های اولیه و پیش‌پردازش شده را نشان می‌دهد. همانطور که مشخص است قسمت‌های حذف شده و تغییر یافته دربردارنده‌ی هیچ‌گونه اطلاعاتی که در تحلیل احساسات مفید باشد، نیست. بنابراین تغییرات انجام‌شده علاوه بر افزایش سرعت پردازش اطلاعات موجب حصول نتیجه‌ی بهتر و دقیق‌تر در طبقه‌بندی توییت‌ها می‌شود.

جدول ۴-۳. نمونه‌هایی از توییت‌های اولیه و پیش‌پردازش شده

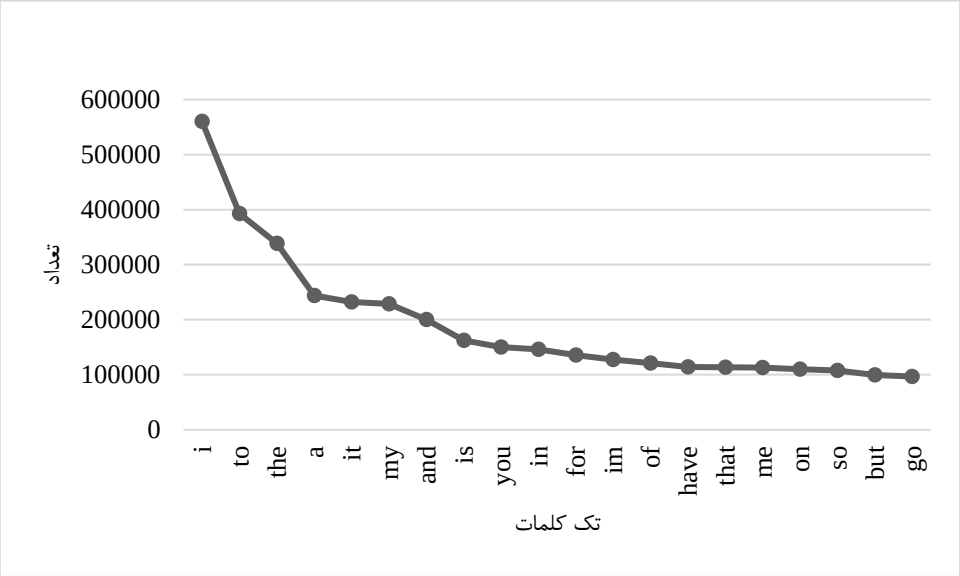
ردیف	جمله خام	جمله پس از اعمال پیش پردازش
۱	help me forget april juli.	help me forget 8th april & 13th july!
۲	i miss watch rocko modern life.	I miss watching Rocko's Modern Life.
۳	car show season ha start without me.	Car show season has started without me.
۴	is tiredd want to sleep but i have an assign to finish and an exam tomorrow to studi for	is tiredddddddd. want to sleep but i have an assignment to finish and an exam tomorrow to study for
۵	user_ment user_ment umm i cant find it	@ddlovato @David_Henrie ummmmm i cant find it.
۶	url cant see the flower fall i dont have a camera just my cellphon	http://twitpic.com/2y36e - cant see the flowers falling i dont have a camera, just my cellphone

الگوریتم‌های Stemming مورد بررسی قرارگرفته بر روی مجموعه‌داده موردنظر الگوریتم Lancaster، Porter و Snowball هستند. الگوریتم Porter از جمله قدیمی‌ترین الگوریتم‌های ریشه‌یابی و بدون شک متداول‌ترین آن‌هاست است. یکی از مزیت‌های این الگوریتم که مورد توجه قرار گرفته پشتیبانی از زبان جاوا است. تمرکز عمده‌ی الگوریتم Porter بر سرعت انجام محاسبات است و از این نظر با اختلاف بسیار، بالاترین رتبه را به خود اختصاص داده است. الگوریتم Snowball در واقع نسخه‌ی بهبود یافته‌ی الگوریتم Porter است و طبق گفته‌ی توسعه‌گر آن بهتر از Porter عمل می‌کند. الگوریتم Lancaster تهاجمی‌ترین الگوریتم Stemming است. زیرا کلمه‌ی ریشه‌ی به‌دست آمده به قدری کوتاه می‌شود که غیر قابل خواندن بوده و تخمین لغت اولیه مشکل است.

۴-۴- استخراج ویژگی[۱]

در مرحله استخراج ویژگی سه نوع ویژگی از مجموعه‌داده‌ی Sentiment140 به نام‌های unigram، bigram و trigram استخراج شده‌است.

۴-۴-۱- Unigram ها



شکل ۴-۱. پرتکرارترین unigram ها پس از اعمال پیش پردازش

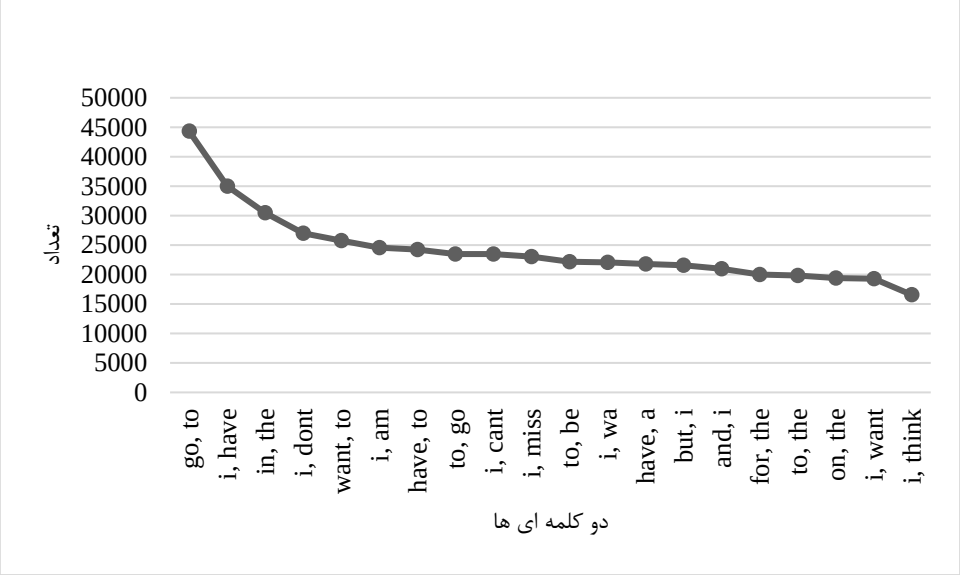
رایج‌ترین و ساده‌ترین ویژگی‌های استفاده‌شده در طبقه‌بندی متن، وجود تک‌کلمات و نشانه[۲]ها در متن است. در این تحقیق تک‌کلمات از مجموعه‌داده آموزشی استخراج گردیده و توزیع فراوانی کلمات استخراج شده ایجاد شد. تعداد ۱۶۹,۱۷۴ تک‌کلمه منحصربه‌فرد از مجموعه‌داده استخراج شد. شکل ۴-۱ نمودار پرتکرارترین unigram های موجود در مجموعه‌داده Sentiment140 پس از انجام پیش‌پردازش است. همانطور که مشاهده

[۱] feature extraction

[۲] token

می‌شود اکثر کلمات انتهای طیف نویز[۱] بوده و تکرار بسیار اندکی دارند، بنابراین در نتیجه طبقه‌بندی تاثیرگذار نخواهند بود. درنهایت تعداد پرتکرارترین N کلمه برای ساخت فرهنگ‌واژگان استفاده شد که N برابر با ۱۵۰۰۰ برای بردار اسپارس طبقه‌بندی در نظر گرفته شده‌است.

۴-۴-۲- Bigram ها



شکل ۴-۲. پرتکرارترین bigram ها پس از اعمال پیش پردازش

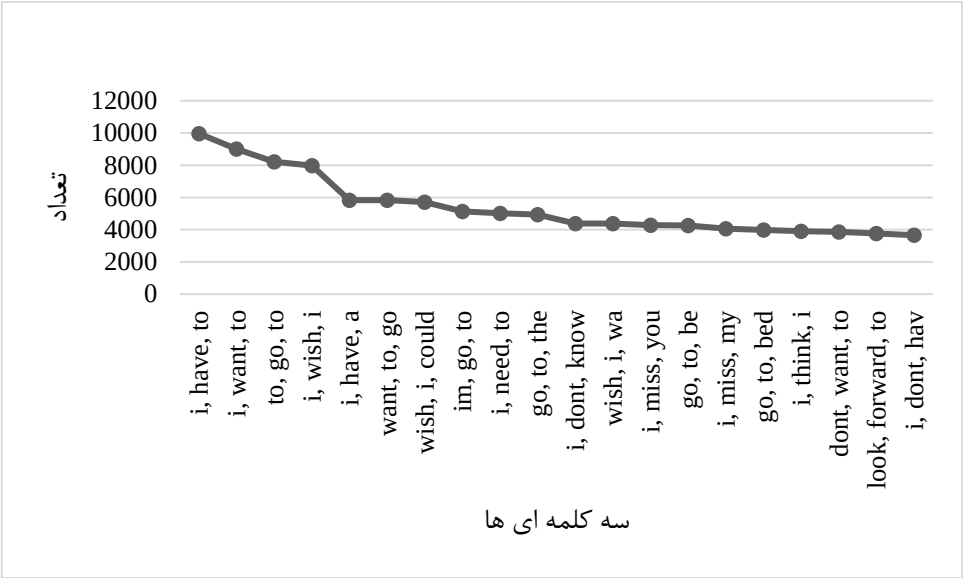
Bigram ها جفت‌کلماتی از مجموعه‌داده هستند که به طور متوالی در متن دیده می‌شوند. این ویژگی‌ها کاربردهای زیادی در مدل‌سازی نفی در زبان طبیعی مثل عبارت "This is not good" را دارند. تعداد bigram ۲,۰۸۱,۴۳۴ منحصربه‌فرد از مجموعه‌داده Sentiment140 استخراج گردیده است. مشابه unigram ها، از بین این جفت‌کلمات نیز اکثر جفت‌کلمات انتهای طیف نویز[۲] هستند و تکرار بسیار اندکی دارند، بنابراین در نتیجه طبقه‌بندی تاثیرگذار نخواهند بود. درنهایت تعداد ۱۰,۰۰۰ جفت کلمه پرتکرارتر برای ساخت

[۱] noise

[۲] noise

فرهنگ‌واژگان استفاده شده است. شکل ۴-۲ نمودار پرتکرارترین unigram موجود در مجموعه‌داده Sentiment140 پس از انجام پیش‌پردازش را نشان می‌دهد.

۴-۴-۳- Trigram ها



شکل ۴-۳. پرتکرارترین trigram ها پس از اعمال پیش پردازش

Trigram ها سه کلمه‌ای‌هایی از مجموعه‌داده هستند که به طور متوالی در متن دیده می‌شوند. تعداد ۵,۸۴۶,۰۲۳ trigram منحصربه‌فرد از مجموعه‌داده استخراج گردیده است. مقدار N سه کلمه‌ای پرتکرار برای این ویژگی‌ها نیز ۱۰,۰۰۰ در نظر گرفته شده‌است. شکل ۴-۳ نمودار پرتکرارترین trigram های موجود در مجموعه‌داده Sentiment140 پس از انجام پیش‌پردازش را نشان می‌دهد.

۴-۵- وزن‌دهی به صفات

همانطور که گفته شد در این تحقیق کلمات بی‌معنی وزن بالاتری نسبت به کلمات مهم و پرتکرار می‌گیرند. به این صورت که اختلاف وزن همه‌ی ویژگی‌های مثبت و منفی موجود در مجموعه‌داده آموزش محاسبه

می‌شود. سپس میانگین این اختلاف‌ها حساب می گردد. به این ترتیب بسیاری از وزن‌ها مقدار صفر می‌گیرند. این فرایند سبب می شود تا کلمات بی‌معنی برای تخمین دسته‌ی متن تست در نظر گرفته نشوند. الگوریتم رویکرد ارائه‌شده به صورت زیر است:

ورودی: یک مجموعه آموزشی $\boldsymbol{D}$، یک متن تست $\boldsymbol{d}$

خروجی: کلاس $\boldsymbol{c(d)}$

شروع

۱. تقسیم $\boldsymbol{D}$ به $\boldsymbol{D_c}$ $(c \in \{\text{positive, negative}\})$

۲. محاسبه $\boldsymbol{IGR(C, w_i)}$ به ازای هر $\boldsymbol{W_{i,c}(i = 1, 2, \dots, m_c)}$ از $\boldsymbol{D_c}$ طبق رابطه ۳-۲۲

۳. محاسبه وزن $\boldsymbol{WT_{i,c}}$ به ازای هر $\boldsymbol{W_{i,c}(i = 1, 2, \dots, m_c)}$ از $\boldsymbol{D_c}$ طبق رابطه ۳-۲۷

۴. محاسبه اختلاف وزن های $\boldsymbol{WD_i}$ به ازای هر $\boldsymbol{w_i}$ $\boldsymbol{(i = 1, 2, \dots, m)}$ از $\boldsymbol{D}$ طبق رابطه ۳-۲۹

۵. محاسبه میانگین $\boldsymbol{WD_i}$ ها با استفاده از رابطه ۳-۲۰

۶. اصلاح وزن $\boldsymbol{WT_{i,c}}$ به ازای هر $\boldsymbol{W_{i,c}(i = 1, 2, \dots, m_c)}$ از $\boldsymbol{D_c}$ طبق رابطه ۳-۳۱

۷. با ازای هر $\boldsymbol{d}$

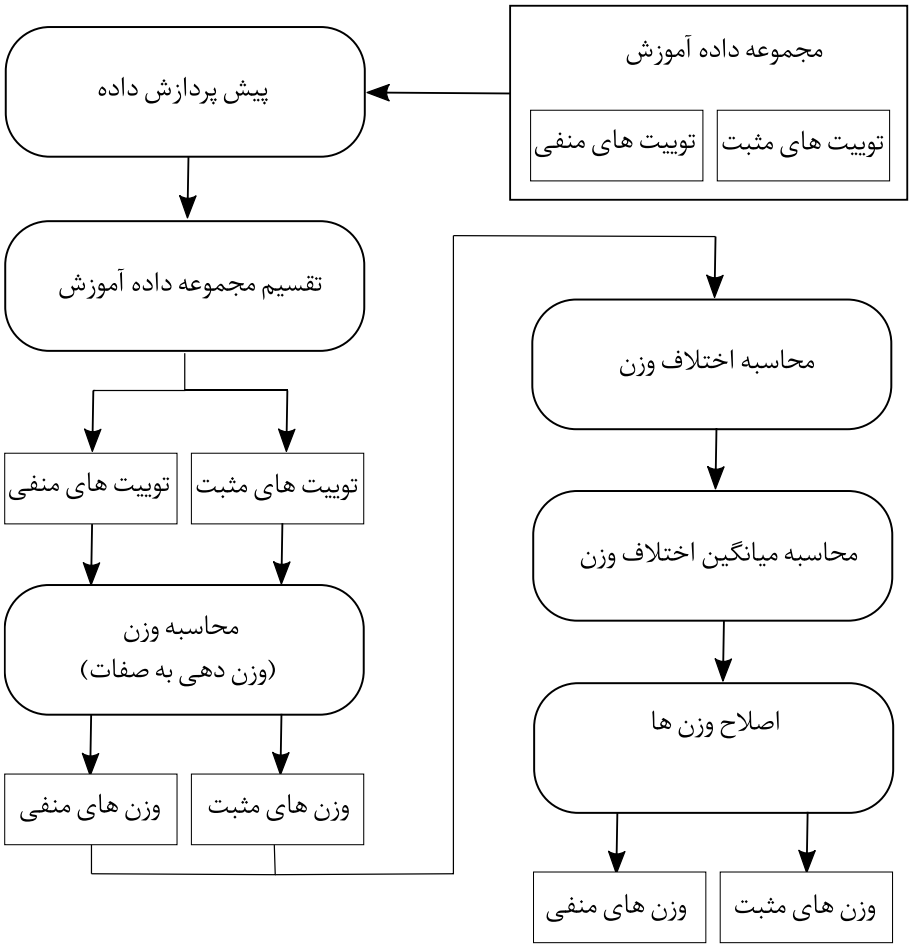
أ محاسبه $\boldsymbol{P(c)}$ با استفاده از رابطه ۳-۱۷

ب محاسبه $\boldsymbol{P(w_i|c)}$ با استفاده از رابطه ۳-۲۸

ج پیش بینی $\boldsymbol{c(d)}$ با استفاده از رابطه ۳-۲۱

۸. بازگرداندن $\boldsymbol{c(d)}$

پایان



شکل ۴-۴. نمای کلی از وزن‌دهی به صفات و انتخاب ویژگی‌ها بر اساس بیز چندجمله‌ای

شکل ۴-۴ [۲۳] نمای کلی از وزن‌دهی به صفات و انتخاب ویژگی‌ها بر اساس بیز چندجمله‌ای را نشان می‌دهد. همان‌طور که مشخص است در مرحله‌ی اول با تقسیم مجموعه‌ی آموزشی به دو دسته‌ی مثبت و منفی وزن‌ها محاسبه می‌شوند و سپس در مرحله‌ی دوم وزن‌ها با استفاده از میانگین اختلاف وزن‌ها اصلاح می‌گردند.

۴-۶- محاسبه احتمال و طبقه‌بندی

پس از پیش پردازش‌داده ها و انتخاب ویژگی‌ها، طبقه‌بندی متن‌ها با استفاده از روش بیز صورت می‌گیرد. در مرحله‌ی آموزش احتمال پیشین هر ویژگی در دسته‌ی آن محاسبه شده و ثبت می‌گردد. سپس در مرحله‌ی

تست، ویژگی‌های هر یک از متن‌ها مشابه مرحله‌ی آموزش استخراج شده و احتمال پسین هر ویژگی در هر دسته محاسبه می‌شود.

فصل ۵: نتایج آزمایشات

۵-۱- مقدمه

در این فصل به بررسی نتایج به دست آمده از آزمایش‌های انجام شده بر روی نمونه‌های مجموعه‌داده مورد استفاده و باتوجه به انتخاب پارامترهای مناسب پرداخته می‌شود. سپس نتایج به دست آمده با کارهای پیشین در این حوزه مقایسه خواهد شد.

۵-۲- نتایج

همان‌طور که پیش‌تر توضیح داده شد، برای ارزیابی کارایی مدل پیشنهادی، از مجموعه‌داده Sentiment140 که حاوی ۱,۶۰۰,۰۰۰ نمونه از توییت‌های توییتر است استفاده شد. با حذف توییت‌های خنثی این تعداد به حدود ۱,۰۴۸,۰۰۰ توییت کاهش یافت.

در این بخش نتایج نهایی مدل پیشنهادی با توجه به پارامترهای موجود ارائه می‌گردد. تغییر هر یک از این پارامترها موجب تغییر در نتایج خواهد شد. با بررسی این تغییرات، بهترین مقادیر برای پارامترها مشخص شده و نتایج به دست آمده با استفاده از این مقادیر، به عنوان نتیجه نهایی تحقیق در نظر گرفته می‌شود. نتایج نشان می‌دهد که روش طبقه‌بندی پیشنهادی بر روی ۱,۰۴۸,۰۰۰ نمونه استخراج شده از مجموعه‌داده Sentiment140 در بهترین حالت با ۸۳/۳۷٪ صحت انجام گرفته است. در ادامه نتایج به دست آمده تحلیل خواهد شد.

۵-۳- تحلیل نتایج

در این بخش نتایج به دست آمده از دیدگاه‌های مختلف تحلیل شده و دستاوردهای آن شرح داده می‌شود. همچنین تاثیر تغییر پارامترهای مختلف بررسی شده و نمودارهای مربوط به آن‌ها ارائه می‌شود.

۵-۳-۱- تاثیر روش‌های ریشه‌یابی

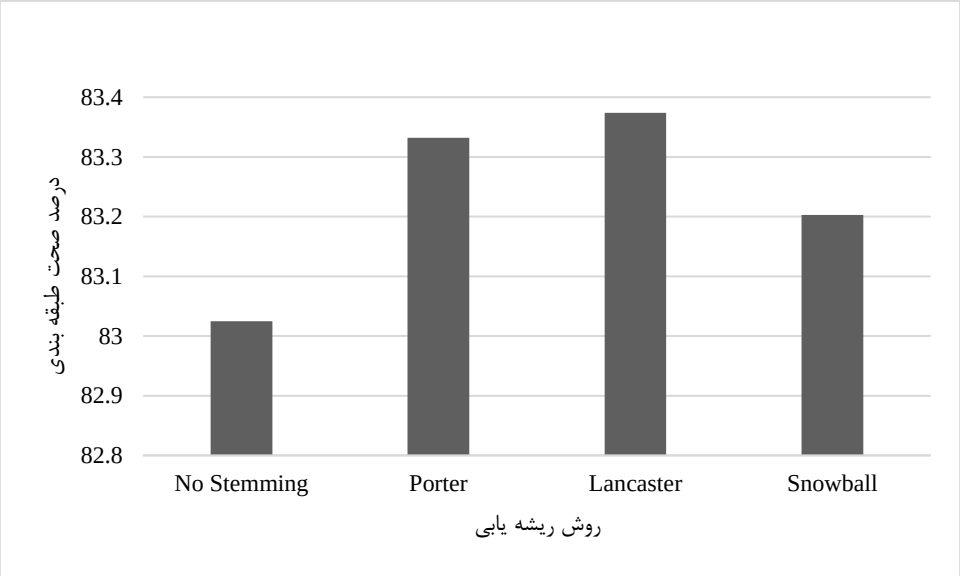
در شکل ۵-۱ نمودار تاثیر روش ریشه‌یابی در نتیجه‌ی طبقه‌بندی ارائه شده‌است. برای این منظور سه الگوریتم معروف‌تر ریشه‌یابی با انتخاب بهترین حالت از پارامترهای دیگر سیستم در نظر گرفته شده‌است. الگوریتم‌های مورد بررسی قرارگرفته بر روی مجموعه‌داده موردنظر شامل الگوریتم Lancaster، الگوریتم Porter و الگوریتم Snowball هستند.

الگوریتم Porter از جمله قدیمی‌ترین الگوریتم‌های ریشه‌یابی و بدون شک متداول‌ترین آن‌ها است. یکی از مزیت‌های این الگوریتم که مورد توجه قرار گرفته پشتیبانی از زبان جاوا است. تمرکز عمده‌ی الگوریتم Porter بر سرعت انجام محاسبات است و از این نظر با اختلاف بسیار، بالاترین رتبه را به خود اختصاص داده است.

الگوریتم Snowball در واقع نسخه‌ی بهبود یافته‌ی الگوریتم Porter است و طبق گفته‌ی توسعه‌گر آن بهتر از Porter عمل می‌کند. به این معنی که زمان انجام محاسبات در این الگوریتم بهبود یافته است. اما همانطور که در نمودار شکل ۵-۱ مشخص است، برای مجموعه داده‌ی مورد استفاده در این تحقیق نتیجه‌ی مناسبی حاصل نگردیده است. دلیل این کاهش صحت طبقه‌بندی این است که در این تحقیق مسئله سرعت محاسبات حائز اهمیت نیست. علاوه بر آن همان‌طور که پیش‌تر گفته شد، توییت‌های توییتر حاوی کلمات و جملاتی است که توسط افراد مختلف نوشته شده و وجود غلط‌های املایی و تایپی بسیار زیاد سبب کاهش دقت الگوریتم ریشه‌یابی خواهد شد. بنابراین این دو الگوریتم اگرچه الگوریتم‌های متداول و مناسبی هستند اما در تحقیق فوق به دلایل ذکر شده نتوانستند به خوبی عمل کنند.

از بین این الگوریتم‌های بررسی شده، از الگوریتم Lancaster به عنوان تهاجمی‌ترین الگوریتم نام برده می‌شود. به این معنی که کلمه‌ی ریشه‌ی به‌دست آمده به قدری کوتاه می‌شود که غیر قابل خواندن بوده و تخمین لغت اولیه مشکل است. اما با توجه به کاربرد و موارد استفاده، این الگوریتم می‌تواند مفید نیز واقع شود، به عنوان مثال در تحقیق فوق و موارد مشابه که مجموعه‌داده بسیار بزرگ است و کلمات موجود در مجموعه‌داده، کلمات

محاوره‌ای و حاوی غلط‌های املایی هستند. همان‌طور که در نمودار نیز مشخص است، روش Lancaster بیشترین میزان صحت را برای طبقه‌بندی مجموعه‌داده ارائه شده داراست.



شکل ۵-۱. نمودار تاثیر هر یک از روش های ریشه یابی در صحت عملکرد طبقه بند

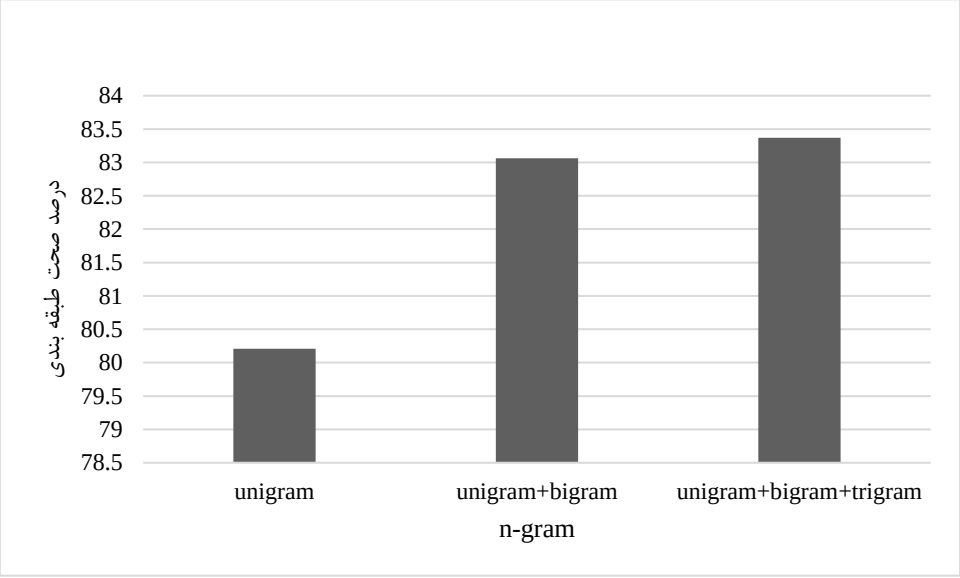
۵-۳-۲- تاثیر عمق‌های مختلف در n-gram

یکی از کاربردهای عمده n-gram ها برای توسعه و استخراج ویژگی در مدل‌های باناظر یادگیری ماشین مثل ماشین بردار پشتیبان و الگوریتم بیز است. در این تحقیق نیز در مرحله انتخاب ویژگی از ویژگی های n-gram بهره گرفته شده است. با انجام آزمایش‌های متعدد و طبق تجربه پژوهشگران مختلف، استفاده از bigram و trigram در بردار ویژگی موجب بهبود قابل توجه و چشمگیری در صحت عملکرد سیستم نمی‌شود. با تجربه و آزمایش در هر پژوهش می‌توان میزان کارا بودن آن را مورد سنجش قرار داد و بسته به نوع الگوریتم طبقه‌بندی و مجموعه داده می‌تواند تاثیرگذار باشد و یا تاثیر کمی داشته باشد.

نمودار شکل ۵-۲ نشان دهنده تاثیر عمق‌های مختلف در n-gram در این تحقیق است. طبق آن‌چه در نمودار مشخص است، با اضافه شدن bigram صحت طبقه‌بندی به طور قابل ملاحظه‌ای افزایش یافت. اگرچه

اضافه‌شدن trigram به اندازه bigram موجب افزایش عملکرد سیستم نشده‌است، اما می‌توان گفت trigram نیز باعث افزایش صحت طبقه‌بندی گردیده است. این افزایش در هر یک از عمق‌های بالاتر، کمتر از عمق ماقبل خود است. بدین ترتیب میزان n انتخابی در بهترین حالت برای این تحقیق عدد ۳ در نظر گرفته می‌شود، زیرا استفاده از n های بالاتر به نسبت بهبودی که در صحت طبقه‌بندی حاصل می‌شود، مقرون به صرفه نیست.

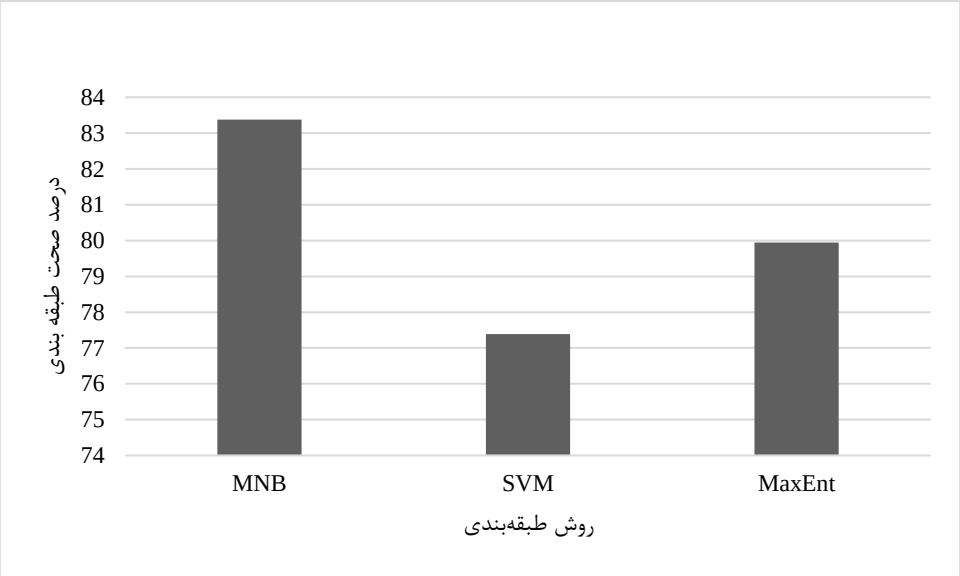
استفاده از bigram ها و trigram ها در طبقه‌بندی متن نسبت به unigram سبب تاثیرگذاری هرچه بیشتر موضوع و هدف کلی جمله در هر ویژگی می‌شود. به عبارت دیگر کلاس و دسته‌ی متن از تک کلمات موجود در متن با دقت بالایی قابل تشخیص نیست، اما کلمات قبل و یا بعد از کلمه که ممکن است دربردارنده‌ی کلاس متن باشند در کلمات همسایه‌ی خود تاثیرگذار هستند. به عنوان مثال در جمله‌ی "It is not clean"، کلمه ی "clean" به تنهایی گویای کلاس مثبت خواهد بود اما با همراه شدن با "not" نتیجه تغییر خواهد کرد.



شکل ۵-۲. نمودار تاثیر عمق های مختلف n-gram در صحت عملکرد طبقه بند

۵-۳-۳- تاثیر روش طبقه‌بندی

همان‌طور که در فصل‌های پیشین بررسی شد روش‌های مبتنی بر یادگیری ماشین تاثیرگذارترین روش‌ها در تحلیل احساسات هستند. در این تحقیق از سعی شده است تا با در نظرگرفتن پارامترهای مختلف و با تمرکز بیشتر بر مرحله پیش‌پردازش عملکرد الگوریتم بیز را بهبود بخشیم. در نمودار ۵-۳ روش پیشنهاد شده در این تحقیق با الگوریتم‌های ماشین بردار پشتیبان، حداکثر آنتروپی مقایسه شده است. همان‌طور که مشخص است الگوریتم بیز عملکرد بهتری دارد.



شکل ۵-۳. نمودار تاثیر روش‌های مختلف طبقه‌بندی در صحت عملکرد طبقه بند

۵-۴- مقایسه نتایج تحقیق با کارهای پیشین

با توجه به آن‌چه در بخش‌های قبل گفته شد، بهترین صحت برای روش پیشنهادی در این تحقیق ۸۳/۳۷٪ به دست آمده است. این دقت با انتخاب بهترین روش ریشه‌یابی با توجه به شرایط مسئله، یعنی استفاده از ریشه‌یابی با روش Lancaster، بهره‌گیری از ویژگی‌های trigram و وزن دهی به صفات بر اساس نسبت بهره و همچنین با به‌کارگیری طبقه‌بند بیز حاصل می‌شود. در جدول ۵-۱ به بررسی روش‌های پیشین و مقایسه آن

با روش پیشنهاد شده در این تحقیق پرداخته شده‌است. همان‌طور که مشخص است صحت روش پیشنهادی ارائه شده از سایر کارهای مشابه بررسی شده بیشتر می‌باشد.

جدول ۵-۱. مقایسه رویکرد ارائه‌شده با مهم‌ترین کارهای پیشین

ردیف	مرجع	سال انتشار	مجموعه‌داده	روش طبقه‌بندی	صحت
۱	[۱۹]	۲۰۱۸	Golden Dataset	مبتنی بر لغت‌نامه بیز ساده حداکثر آنتروپی	٪۷۵.۳۴
۲	[۱۸]	۲۰۱۶	Amazon Commerce Website	بیز ساده	٪۷۸.۲۹
۳	[۲۶]	۲۰۱۶	acl Internet Movie Database (IMDb)	بیز ساده حداکثر آنتروپی گرادیان نزولی تصادفی ماشین بردار پشتیبان	٪۸۱.۴
۴	[۲۳]	۲۰۱۷	Sentiment140	بیز ساده	٪۷۷.۶۵

فصل ۶: نتیجه‌گیری و پیشنهادات

۶-۱- نتیجه گیری

در این تحقیق سیستمی برای تحلیل احساسات در شبکه های اجتماعی با استفاده از روش‌های آماری ارائه شده‌است. مجموعه‌داده مورد استفاده در این سیستم، Sentiment140 درنظر گرفته شده‌است که حاوی بیش از ۱,۶۰۰,۰۰۰ نمونه از توییت‌های شبکه اجتماعی توییتر است.

روش طبقه‌بندی مورد استفاده در این سیستم، طبقه بند بیز است. پس از انجام پیش‌پردازش‌های لازم بر روی داده های مجموعه‌داده و انتخاب ویژگی، با اعمال روش بیز، طبقه‌بندی داده‌ها انجام شد.

مراحل پیش‌پردازش داده‌ها شامل حذف مواردی است که در تحلیل احساسات تاثیری ندارند مثل اعداد، برچسب گذاری‌ها، هشتگ‌ها، علائم، نقطه گذاری‌ها، آدرس‌های اینترنتی و... . از جمله مهم‌ترین مراحل در پیش‌پردازش ریشه‌یابی است. با بررسی‌های لازم و تحلیل نتایج، سه مورد از پرکاربردترین روش ریشه‌یابی مورد ارزیابی قرار گرفته و بهترین روش ریشه‌یابی برای این تحقیق روش Lancaster انتخاب شده است.

در بخش انتخاب ویژگی نیز از ویژگی های n-gram استفاده گردیده است. بررسی نتایج حاصل از مقادیر مختلف برای n نشان می‌دهد که بهترین و مقرون به صرفه‌ترین حالت استفاده از trigram است. استفاده از n-gram در مرحله انتخاب ویژگی و ترکیب آن با وزن‌دهی به صفات براساس نسبت بهره، دستاورد اصلی این تحقیق می باشد.

بهترین صحت برای روش پیشنهادی در این تحقیق ۸۳/۳۷٪ به دست آمده است. این دقت با استفاده از ریشه‌یابی با روش Lancaster، بهره‌گیری از ویژگی‌های trigram و وزن دهی به صفات بر اساس نسبت بهره و همچنین با به‌کارگیری طبقه‌بند بیز حاصل می‌شود.

۶-۲- پیشنهاد کارهای آینده

کارهای آینده‌ای که برای این تحقیق می‌توان متصور شد، استفاده از طبقه‌بند چند کلاسی برای درنظرگرفتن احساسات دقیق‌تر کاربران مثل ناراحتی، شادی، هیجان‌زده، کسل و... است. بسیاری از جملات را نمی‌توان تنها با مثبت و یا منفی طبقه‌بندی کرد. علاوه بر آن با توجه به نیاز و کاربرد طبقه‌بندی، دسته‌ها می‌توانند فراتر از مثبت و منفی بوده و احساسات دقیق‌تر افراد را بیان کنند. شکلک‌ها را نیز می توان در بیش از دو دسته‌ی احساسی جا داد. بنابراین می‌توان با درنظر گرفتن احساسات دقیق‌تر شکلک‌ها و طبقه‌بندی چند کلاسی نتایج خوبی را برای این تحقیق متصور شد.

با پیاده‌سازی انواع دیگر روش‌های یادگیری ماشین از جمله ماشین بردار پشتیبان، درخت تصمیم و... ، همچنین اعمال روش‌های کاهش ویژگی از جمله PCA انتظار می‌رود بتوان به نتایج مهمی در طبقه‌بندی متن رسید.

مراجع

[1] Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, Bing Liu, (2016), "*Sentiment Analysis in Social Networks*", 1st edition, Morgan Kaufmann, pp. 1-10.

[2] Bing Liu, (2012), "*Sentiment Analysis and Opinion Mining*", vol. 5, no. 1, Morgan & Claypool, pp. 1-167.

[3] Ankit, Nabizath Saleena, (2018), "An Ensemble Classification System for Twitter Sentiment Analysis" *International Conference on Computational Intelligence and Data Science (ICCIDS 2018)*, vol. 132, pp. 937–946.

[4] Sitaram Asur, Bernardo A. Huberman, (2010), "Predicting the future with social media",*2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 492-499.

[5] Olga Kolchyna, Th´arsis T. P. Souza, Philip C. Treleaven, Tomaso Aste, (2015), "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination.", *Handbook of Sentiment Analysis in Finance*, .

[6] Ana¨ıs Collomb, Crina Costea, Damien Joyeux, Omar Hasan, Lionel Brunie, (2018), "A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation" *Springer Singapore*, pp. 53-66.

[7] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, (2002), "Thumbs up? Sentiment Classification using Machine Learning Techniques", *ACL-02 conference on Empirical methods in natural language processing*, vol. 71, no. 11, pp. 79-86.

[8] Kushal Dave, Steve Lawrence, David M. Pennock, (2003), "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", *12th international conference on World Wide Web*, pp. 519-528.

[9] Bo Pang, Lillian Lee, (2008), "Opinion Mining and Sentiment Analysis", *Foundations and Trends in Information Retrieval*, vol. 11, no. 11, pp. 1-135.

[10] Peter D. Turney, (2002), "Thumbs up or thumbs down? Semantic orientation applied to

unsupervised classification of reviews", *40th Annual Meeting on Association for Computational Linguistics*, pp. 417-424.

[11] Minqing Hu, Bing Liu, (2004), "Mining and Summarizing Customer Reviews", *10th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168-177.

[12] Ahmed Abbasi, Hsinchun Chen, Arab Salem, (2008), "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums", *ACM Transactions on Information Systems*, vol. 26, no. 3, pp. 1-34.

[13] Michelle Annett, Grzegorz Kondrak, (2008), "A Comparison of Sentiment Analysis Techniques : Polarizing Movie Blogs", *Twenty-First Canadian Conference on Artificial Intelligence*, no. Figure 1, pp. 25–35.

[14] Bing Liu, *Web Data Mining (second edition)*, (2011), 2st edition, Springer.

[15] Lei Johnny Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, Bing Liu, (2011), "Sentiment Analysis Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", *HP Laboratories*.

[16] M. Ghiassi, J. Skinner, and D. Zimbra, (2013), "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network" *Expert Systems with Applications*, Elsevier, vol. 40, no. 16, pp. 6266-6282.

[17] Orestes Appel, Francisco Chiclana, Jenny Carter, Hamido Fujita, (2018), "Successes and challenges in developing a hybrid approach to sentiment analysis", *Applied Intelligence*, Springer Nature, vol. 48, pp. 1176-1188.

[18] Lungan Zhang, Liangxiao Jiang, Chaoqun Li, Ganggang Kong, (2016),"Two feature weighting approaches for naive Bayes text classifiers", *Knowledge-Based Systems*, vol. 100, pp. 137–144.

[19] Shaunak Joshi, Deepali Deshpande, (2018), "Twitter Sentiment Analysis System", *International Journal of Computer Applications*, vol. 180, no. 47, pp. 35–39.

[20] Shivani Bahri, Pranav Bahri, Sangeeta Lal, (2018), "A Novel approach of Sentiment Classification using Emoticons", *Procedia Computer Science*, vol. 132, pp. 669-678.

[21] Michal Novak, Tomas Kincl, Jiri Pribil, (2019), “Improving sentiment analysis performance on morphologically rich languages : Language and domain independent approach”, *Computer Speech & Language, Elsevier*, vol. 56, pp. 36-51.

[22] Mazen El-Masri, Nabeela Altrabsheha , Hanady Mansourb, Allan Ramsay, (2017), “A web-based tool for Arabic sentiment analysis”, *3rd International Conference on Arabic Computational Linguistics*, vol. 117. pp. 38–45.

[23] Junseok Song, Kyung Tae Kim, Byungjun Lee, Sangyoung Kim, Hee Yong Youn, (2017), “A novel classification approach based on Naïve Bayes for Twitter sentiment analysis”, *KSII Transactions on Internet and Information Systems*, Korean Society for Internet Information, vol. 11, no. 6, pp. 2996-3011.

[24] Symeon Symeonidis , Dimitrios Effrosynidis , Avi Arampatzis, (2018), “A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis,”*Expert Systems with Applications*, vol. 110, pp. 298–310.

[25] Jingnian Chen, Houkuan Huang a , Shengfeng Tian a , Youli Qu, (2009),“Feature selection for text classification with Naïve Bayes”, *Expert Systems with Applications, Elsevier*, vol. 36, no. 3, pp. 5432–5435.

[26] Abinash Tripathy , Ankit Agrawal, Santanu Kumar Rath, (2016), “Classification of Sentiment Reviews using N-gram Machine Learning Approach”, Expert Systems with Applications, *Elsevier*, vol. 57, no. October 2017, pp. 117–126.

[27] Jia Wu, Shirui Pan, Xingquan Zhu, Zhihua Cai, Peng Zhang, Chengqi Zhang, (2015), “Self-adaptive attribute weighting for Naive Bayes classification”, *Expert Systems with Applications, Elsevier*, vol. 42, no. 3, pp. 1487-1502.

[28] Shasha Wang, Liangxiao Jiang, Chaoqun Li, (2014), “A CFS-Based Feature Weighting Approach to Naive Bayes Text Classifiers”, *Artificial Neural Networks and Machine Learning*, Springer International Publishing, pp. 555-562.

[29] Alec Go, Richa Bhayani, Lei Huang, (2009), “Twitter Sentiment Classification using Distant Supervision”, p. 6.

Abstract

Nowadays social media have received more attention. Public and private opinions about a wide variety of subjects are expressed by users continually via numerous social media. Twitter is one of the social media that is gaining popularity in recent decade. Twitter offers organizations a fast and effective way to analyze customer's perspectives toward the critical to success in the market place. Sentiment analysis or opinion mining is the process of extracting people's opinions, perspectives and sentiments about a particular topic. Sentiment analysis in the domain of micro-blogging is a new research topic so it has still the potential to research and develop in this field.

Lots of works has been done on sentiment analysis of user reviews, documents and articles. Analysis on the above items differ from twitter mainly because of the limit of 280 characters per tweet which forces the user to express his/her opinion compressed and short. The best results reached in sentiment classification use machine learning techniques such as Naive Bayes and Support Vector Machines.

In this research we proposed a method for sentiment analysis in social networks. In this respect, we have tried to improve Naïve Bayes classification by focusing on preprocessing and feature selection steps. The proposed method has the detection accuracy of 83.37 percent. In addition, results shows that using n-gram features provide the best results and combining it with attribute weighting based on gain ratio obtains a remarkable improvement in the results.

Keywords: sentiment analysis, text mining, opinion mining, text classification, machine learning, feature selection.

Shahrood University of

Technology

Faculty of Computer Engineering

M.Sc. Thesis in Artificial Intelligence Engineering

Sentiment Analysis in Social Networks using Statistical Analysis

By: Hamideh Sheikh

Supervisor:

Dr. Morteza Zahedi

Advisor:

Dr. Marzye Rahimi

September 2019