علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۷، شماره ۱، ۱۳۹۸
صفحات ۵۹-۴۴
مقاله عادی

ارائه روشی برای تحلیل احساسات با استفاده از معیار وزن‌دهیTF-IGM و میدان تصادفی شرطی

مریم عموعلی[۱]، فرساد زمانی بروجنی*[۲]

*نویسنده مسئول، دریافت: ۹۷/۰۸/۱۷، بازنگری: ۹۸/۰۱/۲۹، پذیرش: ۹۸/۰۶/۰۹

[۱] دانشجوی کارشناسی ارشد، مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد اصفهان (خوراسگان)، اصفهان، ایران
[۲] استادیار، علوم کامپیوتر، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد اصفهان (خوراسگان)، اصفهان، ایران

چکیده

رشد چشم‌گیر شبکه‌های اجتماعی باعث ایجاد سطح بالایی از اطلاعات در فضای اینترنت می‌شود که این حجم بالای اطلاعات می‌تواند شامل ثبت نظرات کاربران، نیازهای کاربران و یا احساسات آن‌ها باشد. تحلیل نیازهای کاربران سبب معرفی حوزه‌ای بنام تحلیل احساس شد که هدف آن شناسایی احساسات (مثبت، منفی،خنثی) کاربران بر اساس نظرات ثبت شده آن‌ها می‌باشد. در این روش‌ها عموماً از یک الگوریتم دسته‌بندی و معیار وزن‌دهی سنتی TF-IDF استفاده می‌شود که در فرآیند وزن‌دهی به کلمات از اطلاعات کلاس داده‌های آموزشی استفاده نمی‌کند و این اطلاعات را در فرآیند وزن دهی دخیل نمی‌کند، ازاین‌رو نتایج حاصل شده به اندازه کافی مطلوب نمی‌باشد. در این پژوهش از معیار وزن دهی جدیدی تحت عنوان TF-IGM برای وزن‌دهی کلمات استفاده شده است که یک معیار وزن‌دهی با ناظر می‌باشد. علاوه بر این برخلاف روش‌های پیشین در این پژوهش از ترکیب دو روش مدل مخفی مارکوف و میدان تصادفی شرطی که حاصل ترکیب این دو، میدان تصادفی شرطی مخفی می‌باشد برای تحلیل احساس استفاده شده است. نتایج حاصل از اجرای روش پیشنهادی بر روی پایگاه داده نظرات کاربران شبکه توئیتر که شامل ۱۲۰۰۰ توئیت می‌باشد، حاکی از آن است که صحت مدل پیشنهادی در مقایسه با روش سنتی مبتنی بر TF-IDF دارای ۵/۸۲٪ بهبود می‌باشد. درواقع نتایج نشان می‌دهند که استفاده از الگوریتم‌های دسته‌بندی ترکیبی در کنار روش وزن‌دهی با ناظر نتایج بهتری در مقایسه با روش‌های ساده ارائه می‌دهد.

کلمات کلیدی: تحلیل احساس، میدان تصادفی شرطی، میدان تصادفی شرطی مخفی، معیار وزن دهی با ناظر

۱- مقدمه

امروزه عوامل مختلفی همچون توسعه کسب‌وکارهای اینترنتی، تنوع بسیار محصولات و خدمات و همچنین رشد شبکه‌های اجتماعی و ارتباطات مجازی افراد، باعث تشکیل مخزنی از داده‌های غیر ساخت‌یافته از نظرات کاربران در رابطه با محصولات و خدمات مختلف شده است. تشکیل چنین منبع عظیمی از نظرات افراد باعث گردیده که بررسی عقاید کاربران در راستای استخراج دانش کاربردی از اهمیت ویژه‌ای برخوردار باشد [۱]. ازیک‌طرف سازمان‌ها و شرکت‌های تولیدی با بررسی نظرات کاربران در رابطه با محصولات خود به‌راحتی می‌توانند آن‌ها را توسعه داده و یا ارتقاء ببخشند. از طرف دیگر به دلیل تنوع بسیار محصولات، افراد می‌تواند با بررسی نظرات پیشین دیگر مصرف‌کنندگان، در کوتاه‌ترین زمان بهترین انتخاب را داشته باشند. تحلیل و بررسی این قبیل از اطلاعات با اهداف تجاری ، سیاسی و اجتماعی اخیراً توجه بسیاری از محققان را به خود جلب کرده است که منجر به معرفی حوزه جدیدی تحت عنوان تحلیل احساس۱ شده است [۲]. فرآیند تحلیل احساس سعی در تعیین قطبیت نظرات کاربران به‌صورت منفی مثبت و یا خنثی دارد. اما با توجه به اینکه در سال‌های اخیر رشد شبکه‌های اجتماعی و سهولت در اشتراک‌گذاری نظرات افراد به‌واسطه چنین رسانه‌هایی به‌راحتی امکان‌پذیر شده است، بررسی این حجم داده به‌راحتی امکان‌پذیر نمی‌باشد. ازاین‌رو پژوهشگران برای

بررسی و تحلیل داده‌های غیر ساخت‌یافته نظرات کاربران از تکنیک‌های داده‌کاوی بهره برده‌اند. یکی از کارآمدترین تکنیک‌های داده‌کاوی که در حوزه تحلیل احساس موفق بوده است تکنیک کلاس‌بندی است [۳،۴]. که از الگوریتم‌های مختلف آن برای رسیدن به اهداف فوق استفاده شده است. در این پژوهش نیز یک روش تحلیل احساس مبتنی بر دو الگوریتم کلاس‌بندی میدان تصادفی شرطی۲ [۲۶] و مدل مخفی مارکوف و یک معیار وزن دهی با ناظر نوین ارائه شده است .

تحلیل احساس که به‌اختصار به آن SA نیز گفته می‌شود فرآیندی است که طی آن قطبیت یک متن تعیین می‌گردد. منظور از قطبیت در اینجا احساسی است که می‌توان از متن موردنظر دریافت نمود [۵]. به‌واسطه اعمال تحلیل احساس بر روی نظرات کاربران یک فروشگاه اینترنتی می‌توان دریافت که نظرات کاربران در رابطه با محصولات مختلف مثبت است یا خیر. به‌راحتی قابل درک است که نتیجه فرآیند فوق برای فروشگاه موردنظر بسیار کاربردی می‌باشد. از طرف دیگر نتایج این فرآیند برای کاربران نیز مفید است؛ کاربران با اتکا به نظرات ثبت شده می‌توانند انتخاب مناسب‌تری در زمان کوتاه‌تری داشته باشند؛ در نقطه مقابل عاملان فروش هستند که می‌توانند به‌واسطه دانش استخراج شده از فرآیند تحلیل احساس فروش بیشتر و درنتیجه آن سود بیشتری داشته باشند و در سوی دیگر تولیدکنندگان محصولات می‌توانند با استفاده از این دانش محصولات خود را به‌روزرسانی کرده و ارتقاء بخشند [۶،۷].

با توجه به مطالب فوق می‌توان دریافت که دانش حاصل از فرآیند تحلیل احساس به‌قدری سودمند است که نیاز سازمان‌ها و افراد به دریافت آن را نمی‌توان نادیده گرفت؛ اما یکی از مشکلاتی که امروزه در فرآیند تحلیل احساس مطرح است، حجم نظرات ثبت شده توسط کاربران است که رشد شبکه‌های اجتماعی و سهولت در بهره‌گیری از آن‌ها باعث شده به‌صورت روزافزون به حجم اطلاعات ذخیره شده افزوده گردد. این معضل باعث گردیده که محققان به دنبال روشی باشند که به‌صورت خودکار و با حداقل دخالت کاربر و با حداکثر دقت ممکن دانش مفید را از این حجم انبوه استخراج نماید. روش‌های تحلیل احساس مبتنی بر روش‌های یادگیری ماشین یکی از پرکاربردترین روش‌های تحلیل احساس هستند که با حداقل دخالت کاربر فرآیند تحلیل احساس را انجام می‌دهد و از دقت مطلوبی نیز برخوردار هستند [۸،۹]. در روش‌های مبتنی بر یادگیری ماشین دو مرحله اساسی مطرح می‌باشد:

۱) مرحله وزن دهی به کلمات و انتخاب ویژگی

وزن‌دهی به کلمات یکی از مراحل اصلی در دسته‌بندی متون می‌باشد و تأثیر مستقیم در نتایج حاصل از کلاس‌بندی و دقت کلاس‌بندی دارد. اگرچه غالباً در روش‌های دسته‌بندی متون از معیار وزن دهی TF-IDF استفاده می‌شود، این معیار برای کلاس‌بندی متون کاملاً مؤثر نمی‌باشد. طبقه‌بندی متن به‌عنوان یادگیری ماشین با ناظر مطرح می‌باشد که نیاز به مجموعه‌ای از متن با کلاس‌هایی دارد که برای آموزش مدل یادگیری استفاده می‌شوند. اما معیار سنتی TF-IDF در فرآیند وزن‌دهی به کلمات از اطلاعات کلاس داده‌های آموزشی استفاده نمی‌کند و این اطلاعات را در فرآیند وزن دهی دخیل نمی‌کند ، بنابراین وزن محاسبه شده نمی‌تواند به‌طور کامل نشان‌دهنده اهمیت اصطلاح در طبقه‌بندی متن باشد. با توجه به ضعف‌های معیار وزن‌دهی TF-IDF در این پژوهش از یکی از نسخه‌های بهبودیافته آن، تحت عنوان معیار TF-IGM[۳] که یک معیار وزن‌دهی با ناظر می‌باشد، استفاده می‌شود، که در سال ۲۰۱۶ توسط چن و همکاران ارائه شده است [۱۰]. علاوه بر این، در فرآیند تحلیل احساس به دلیل حجم بالای نظرات کاربران مشکل ابعاد بالا در مسئله مطرح می‌شود که استفاده از کل ابعاد و ویژگی‌ها مسئله باعث افزایش زمان فرآیند و درعین‌حال کاهش دقت نتایج خواهد شد، ازاین‌رو برای رفع این مشکل نیز در این پژوهش از الگوریتم انتخاب ویژگی Information Gain استفاده خواهد شد که از بین کلیه ویژگی‌ها تعدادی از مناسب‌ترین ویژگی‌ها را انتخاب کرده و ارائه می‌دهد.

۲) الگوریتم یادگیری و کلاس‌بندی

اغلب روش‌های مبتنی بر یادگیری ماشین برای دسته‌بندی متون از یک الگوریتم کلاس‌بندی استفاده می‌کنند که ازجمله می‌توان به الگوریتم مانک و همکاران ارائه شده در سال ۲۰۱۷ اشاره نمود [۱۱]، اگرچه الگوریتم‌های ارائه شده از نتایج مناسبی برخوردارند اما نتایج آن‌ها به اندازه کافی مطلوب و مناسب نمی‌باشد. ازاین‌رو در این پژوهش از ترکیب دو الگوریتم کلاس‌بندی میدان تصادفی شرطی و مدل مخفی مارکوف استفاده شده است. همچنین، در این پژوهش از میدان تصادفی شرطی مخفی برای تحلیل احساس بهره برده شده است .

با توجه به مطالب فوق به‌طور خلاصه می‌توان گفت، با الهام از روش مانک و همکاران [۱۱]، هدف در این تحقیق ارائه یک روش تحلیل احساس مبتنی بر یادگیری ماشین است که با حداکثر دقت فرآیند تحلیل احساس را انجام دهد. برای این منظور در این پژوهش از معیار وزن دهی با ناظر TF-IGM برای وزن دهی به کلمات استفاده می‌گردد. همچنین در راستای کاهش ابعاد و ویژگی‌ها مسئله از الگوریتم انتخاب ویژگی Information Gain بهره برده می‌شود و درنهایت ترکیب دو الگوریتم طبقه‌بندی میدان تصادفی شرطی و مدل مخفی مارکوف که نتیجه ترکیب این دو میدان تصادفی شرطی مخفی می‌باشد برای طبقه‌بندی متون استفاده می‌شود.

در ادامه در بخش دوم به پیشینه تحقیق، بخش سوم ارائه روش پیشنهادی ، بخش چهارم معیارها و نحوه پیاده‌سازی روش و در بخش پنجم نتیجه‌گیری پرداخته شده است.

۲- کارهای مرتبط

تورنی و همکاران در پژوهشی در سال ۲۰۰۲ الگوریتمی مبتنی بر واژگان برای تحلیل احساس در مجموعه داده‌ای شامل نظراتی در رابطه با محصولات مختلفی همچون اتومبیل، بانک و غیره ارائه نمودند. در پژوهش آن‌ها از الگوهای نحوی برای استخراج عبارات حسی و تعیین گرایش سند استفاده شده است. ازجمله مزایای این روش سادگی آن می‌باشد؛ اما در مقابل دقت پایین راهکار ارائه شده ازجمله معایب آن محسوب می‌شود [۱۲].

پنگ و همکاران در پژوهشی در سال ۲۰۰۴ مدلی مبتنی بر یادگیری ماشین با ناظر با استفاده از الگوریتم‌های Minimum cut، SVM و Naive Bayes در مجموعه دادگانی شامل نظرات درباره فیلم ارائه نمودند. در این پژوهش در مرحله پیش پردازش از الگوریتم برش کمینه و حذف جملات عینی از سند استفاده شده است. ازجمله مزایای این روش تأثیر بالای فرایند پیش پردازش بر روی کارایی طبقه بند سطح سند می‌باشد [۱۳]. در پژوهش دیگری در سال ۲۰۰۵ چویی و همکاران از الگوریتم میدان تصادفی شرطی برای تحلیل احساس استفاده کرده‌اند، ازجمله معایب این روش دقت پایین آن می‌باشد [۱۴].

کندی و همکاران در سال ۲۰۰۶ در تحقیقی برای تحلیل احساس مدلی مبتنی بر یادگیری ماشین با استفاده از الگوریتم SVM در زمینه فیلم ارائه نمودند. در این الگوریتم از کلمات خاصی در متن ازجمله منفی‌کننده‌ها و شدت دهنده‌ها برای بهبود و افزایش دقت طبقه‌بندی استفاده شده است. اگرچه این روش در دسته روش‌های ساده قرار می‌گیرد و این مورد مزیت آن محسوب می‌شود اما ازجمله معایب آن استفاده از ویژگی‌های کم برای الگوریتم ماشین بردار پشتیبان است که باعث کاهش دقت آن شده است [۱۵]. در پژوهش دیگری در سال ۲۰۰۷ هانان و همکاران تحلیل احساس را با استفاده از الگوریتم یادگیری ماشین با ناظر SVM و ترکیب آن با الگوریتم انتخاب ویژگی Information Gain انجام دادند. در این مقاله از دو مجموعه داده Corpora شامل ۳۰۵ نظر مثبت و ۳۰۷ نظر منفی در مورد دوربین دیجیتال و مجموعه داده Blitzer [۱۶] شامل نظرات در رابطه با محصولات متنوع، استفاده کردند. نتایج حاصل در این پژوهش بیانگر این است که نتایج بر روی

مجموعه دادهBlitzer که دارای نظرات مختلف در رابطه با محصولات مختلف می‌باشد و در مقایسه با مجموعه داده Corpora بهتر می‌باشد و دارای دقت ۸۴/۱۵ ٪ است. هر دو مجموعه داده استفاده شده در این پژوهش در دسته مجموعه داده‌های کوچک قرار می‌گیرد این مورد از معایب روش فوق است [۱۷].

پنگ و لی در سال ۲۰۰۸ در تحقیقی از الگوریتم‌های یادگیری ماشین بیز ساده[۴]، ماشین بردار پشتیبان و آنتروپی بیشینه[۵] برای تحلیل احساس در زمینه نظرات درباره فیلم استفاده کردند. در این مقاله مجموعه دادگان سایت IMDB.com استفاده شده است. نتایج حاصل در این پژوهش نشان می‌دهد که الگوریتم ماشین بردار پشتیبان با دقت ۸۲/۹٪ در مقایسه با دو الگوریتم دیگر عملکرد مطلوب‌تری دارد [۶]. شریفی و همکاران در سال ۲۰۰۸ در مطالعه‌ای از میدان تصادفی شرطی برای استخراج کلمات قطبی و تحلیل احساس در متن استفاده کرده‌اند، این الگوریتم از دقت مطلوبی برخوردار نیست و این مورد از معایب آن محسوب می‌شود [۱۸].

ناکاگاوا و همکاران در سال ۲۰۱۰ مدلی جهت تحلیل احساس بر مبنای روش‌های درخت محور و ترکیب آن با الگوریتم میدان تصادفی شرطی با متغیرهای پنهان ارائه نمودند. در این روش تحلیل احساس در دو زبان انگلیسی و ژاپنی انجام شده است. ازجمله معایب این روش پیچیدگی آن است [۱۹]. صالح و همکاران در سال ۲۰۱۱ برای تحلیل احساس از الگوریتم ماشین بردار پشتیبان و الگوریتم‌های انتخاب ویژگی مختلف استفاده کرده‌اند، هدف در این پژوهش بررسی تأثیر هر یک از الگوریتم‌های انتخاب ویژگی بوده است. در این پژوهش از چندین مجموعه داده مختلف ازجمله مجموعه داده مقاله لی و پنگ [۲۰] و مجموعه داده Corpora [۲۱] استفاده شده است. بیشترین دقت حاصل شده در این پژوهش ۹۱٪ گزارش شده است [۲۲].

بایی و همکاران در سال ۲۰۱۱ در پژوهشی برای تحلیل احساس از الگوریتم جستجوی ممنوعه[۶] و MB[۷] استفاده کرده‌اند. در این پژوهش الگوریتم MB وابستگی کلمات را کشف کرده و آن را به یک گراف غیرمدور[۸] تبدیل می‌کند. سپس از الگوریتم جستجوی ممنوعه برای بهبود ساختار گراف استفاده شده است. بیشترین دقت گزارش شده ۹۲/۷۰٪ برای مجموعه داده پنگ و لی [۱۳] است. ژانگ و همکاران در سال ۲۰۱۱ از دو الگوریتم بیز ساده و ماشین بردار پشتیبان برای تحلیل نظرات در مورد یک رستوران به زبان چینی استفاده کرده‌اند. نتایج بر روی ۱۵۰۰ نظر مثبت و ۱۵۰۰ نظر منفی انجام شد. بهترین دقت گزارش شده در این مقاله ۹۵/۶۷٪ است. ازجمله معایب این الگوریتم استفاده از مجموعه دادگان کوچک می‌باشد [۲۳].

کلومپیس و همکاران در سال ۲۰۱۱ مجدداً از الگوریتم ماشین بردار پشتیبان برای تحلیل احساس بر روی داده‌های شبکه اجتماعی توئیتر استفاده کرده‌اند. در مدل فوق استفاده از الگوریتم انتخاب ویژگی و همچنین استفاده از مجموعه داده مناسب باعث افزایش دقت طبقه‌بندی شده است. ازجمله مزایای این روش بررسی تأثیر انتخاب ویژگی بر نتایج می‌باشد [۷]. در سال ۲۰۱۳ مورایس و همکاران در پژوهشی برای تحلیل احساس از سه الگوریتم یادگیری ماشین بیز ساده، ماشین بردار پشتیبان و شبکه عصبی مصنوعی در مجموعه داده‌هایی شامل نظرات در رابطه با فیلم و نظرات در رابطه با محصولات مختلف، استفاده نمودند. در این پژوهش تحلیل احساس در سطح سند انجام شده است. ازجمله مزایای این روش توجه دقیق به ویژگی‌های محاسباتی می‌باشد ازجمله معایب این روش عدم بیان نحوه انتخاب ویژگی است [۲۴]. بصری و همکاران در سال ۲۰۱۳ مدلی برای تحلیلی احساس با استفاده از دو الگوریتم ازدحام ذرات[۹] و ماشین بردار پشتیبان ارائه کرده‌اند. در روش فوق از الگوریتم تکاملی ازدحام ذرات برای انتخاب ویژگی استفاده شده است. نتایج حاصل بر روی مجموعه دادهEMOT دقت ۷۶/۲۰٪ را نشان می‌دهد [۲۵].

در سال ۲۰۱۴ پاترا و همکاران با استفاده از الگوریتم میدان تصادفی شرطی به تحلیل احساس پرداختند. در مدل فوق فقط از تعداد محدودی از کلمات موجود در هر جمله استفاده شده است که درواقع کلمات جنبه محور [۲۶]. در سال ۲۰۱۴ ادنان و همکاران از تکنیک روش تحلیل معنایی صریح[۱۰] برای خوشه‌بندی سلسله مراتبی اسناد استفاده کردند. برای بررسی روش مذکور از مجموعه داده NEWS20 استفاده شده و دقت مدل بررسی شده است. این مدل در مقایسه با روش‌های مشابه دارای عملکرد بهتری است [۲۷].

در سال ۲۰۱۵ کالیوانی و همکاران از الگوریتم تکاملی ژنتیک برای انتخاب ویژگی استفاده کرده‌اند و از الگوریتم‌های یادگیری ماشین بیز ساده، ماشین بردار پشتیبان و رگرسیون منطقی[۱۱]برای تحلیل احساس استفاده نمودند. نتایج این پژوهش نشان می‌دهد که الگوریتم ماشین بردار پشتیبان با دقت ۷۴/۹۵٪ نتایج بهتری در مقایسه با دو روش دیگر داشته است [۲۸]. در سال ۲۰۱۵ اُمام و همکاران نیز از الگوریتم تکاملی ازدحام ذرات و ماشین بردار پشتیبان برای تحلیل احساس در مجموعه داده‌ای شامل نظرات در رابطه با فیلم ارائه کرده‌اند. در این روش نیز از الگوریتم تکاملی ازدحام ذرات برای کاهش ابعاد ویژگی و انتخاب ویژگی استفاده شده است. دقت روش ارائه شده در حدود ۸۱٪ ارائه شده است، ازجمله معایب این روش زمان اجرای بالای آن است [۲۹].

سورین و همکاران در سال ۲۰۱۵ از الگوریتم شبکه‌های عصبی کانولوشن برای تحلیل احساس در نظرات کاربران شبکه اجتماعی توئیتر استفاده کرده‌اند. در این روش از الگوریتمی برای وزن‌دهی به پارامترهای الگوریتم شبکه عصبی استفاده شده است. در این مطالعه از چندین مجموعه داده مختلف استفاده شده است که بیشترین دقت ۷۳٪ ارائه شده است [۳۰]. در سال ۲۰۱۷ جیاتسگولو و همکاران یک روش سریع، انعطاف‌پذیر و عمومی برای تشخیص احساس برای تکه‌های کوچک متن که نظرات مردم به زبان انگلیسی و یونانی را بیان می‌کند، ارائه نمودند. روش مذکور یک روش یادگیری ماشین بوده و از الگوریتم ماشین بردار پشتیبان و از چندین روش نمایش برداری ازجمله Lexicon-based، Embedding-based و Hybrid vectorization استفاده می‌کند. روش فوق به حداقل منابع محاسباتی نیاز دارد؛ در روش فوق برای ارزیابی از چندین پایگاه مختلف استفاده شده است که بیشترین دقت ۸۳٪ بوده است [۳۱].

مانک و همکاران در سال ۲۰۱۷ در پژوهشی یک مدل تحلیل احساس از ترکیب الگوریتم ماشین بردار پشتیبان و الگوریتم انتخاب ویژگی Gini ارائه نمودند. روش مذکور بر روی داده‌های نظرات در رابطه با فیلم اجرا شده است. نتایج روش مذکور در بهترین حالت برابر با ۸۳٪ گزارش شده است. روش‌هایی که از یک الگوریتم ماشین یادگیری استفاده می‌کنند اگرچه نتایج نسبتاً مناسبی ارائه می‌دهند اما نتایج آن‌ها کاملاً مطلوب نیست. این مورد ازجمله معایب روش مذکور است [۱۱].

جدول ۱- بررسی روش‌های انجام‌شده

روش‌های پیشنهادی	مشخصات روش	مزایا و معایب
پنگ و همکاران [۱۳]	ترکیب N B, MINIMUMCUT, SVM	•تأثیر بالای فرآیند پیش پردازش بر روی کارایی طبقه بند سطح
هاتان و همکاران [۱۷]	ترکیب SVM,INFORMATION GAIN	•سادگی روش •ویژگی‌ها کم برای الگوریتم‌های یادگیری
پنگ و لی [۶]	ترکیب NB,SVM,ME	•دقت بالا
بایی و همکاران [۲۰]	MB,TB	•ساخت و بهبود ساختار گراف
ژانگ و همکاران [۲۳]	SVM,NB	•بررسی تأثیر انتخاب ویژگی بر نتایج •عدم بحث در رابطه با الگوریتم طبقه‌بندی
مورایس و همکاران [۲۴]	ترکیبANN,SVM,NB	•توجه دقیق به ویژگی محاسباتی •عدم نحوه انتخاب ویژگی‌ها

ناکاگاوا و همکاران [۱۹]	مدلی بر مبنای روش‌های درخت محور و ترکیب آن با CRF	
ادنان و همکاران [۲۷]	ESA	•خوشه‌بندی اسناد،عملکرد بهتر در مقایسه با روش‌های مشابه
مصلی نژاد و همکاران [۴۷]	ترکیب TF-IDF,ESA	•کاهش ابعاد بردار اسناد با استفاده از تشکیل بردار مفاهیم اسناد
صدر و همکاران [۴۸]	ESA	•بهبود ضریب مطابقت از ۷۴٪ به ۸۱٪

۳- روش پیشنهادی

در این بخش مدل تحلیل احساس پیشنهادی در این پژوهش به‌طور کامل شرح داده می‌شود. پس از مطالعه در خصوص الگوریتم‌های مختلف تحلیل احساس و مقایسه نتایج آن‌ها با یکدیگر می‌توان دریافت که مرحله انتخاب ویژگی و همچنین الگوریتم طبقه‌بندی مورد استفاده در نتایج حاصل از تحلیل احساس تأثیر بسزایی دارند. طبق پژوهش‌های مقایسه‌ای مختلفی که در زمینه انتخاب ویژگی انجام شده‌اند می‌توان دریافت که الگوریتم انتخاب ویژگی Information Gain در مقایسه با روش‌های دیگر ازجمله Gini، Chi-Squared Statistic،Correlation موفق‌تر می‌باشد [۳۲-۳۴]. در مطالعه مانک و همکاران در سال ۲۰۱۷ نیز ۴ روش انتخاب ویژگی مختلف با معیار Gini مورد مقایسه قرارگرفته‌اند و در کلیه آزمایش‌ها غیر از یک آزمایش، الگوریتم Information Gain نتیجه مطلوب‌تری ارائه نموده است [۱۱]. در روش پیشنهادی این مقاله نیز از الگوریتم Information Gain استفاده می‌شود. همچنین مطالعات پژوهشگران مختلف نشان می‌دهد که در مسائل طبقه‌بندی، استفاده از روش‌های ترکیبی و استفاده از دو الگوریتم یادگیری ماشین به‌طور همزمان و ترکیبی نتایج مطلوب‌تری در مقایسه با روش‌هایی که فقط از یک الگوریتم یادگیری ماشین استفاده می‌کنند، ارائه می‌دهد [۳۵-۳۸].

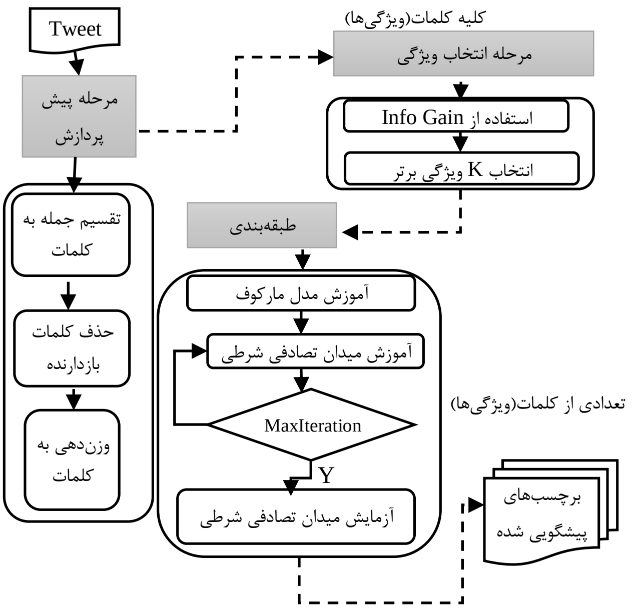
۳-۱- مرحله پیش‌پردازش

همان‌طور که در شکل ۱ مشاهده می‌شود اولین مرحله در مدل پیشنهادی فرآیند پیش پردازش است که از دو بخش اصلی تشکیل شده است:

Tokenization: در این مرحله جملات کاربران به کلمات مجزا تقسیم می‌شوند. پس از تقسیم جملات به کلمات تشکیل دهنده آن‌ها علائم نگارشی و اعداد از مجموعه کلمات تولید شده حذف می‌شوند، لیست علائم حذف شده در این بخش در جدول ۲ ارائه شده است.پس از حذف علائم لیستی از کلمات غیر تکراری که به حالت "حروف کوچک[۱۲]" تبدیل شده‌اند، ذخیره می‌شوند. هدف از اعمال فرآیند پیش پردازش افزایش دقت طبقه‌بندی است. اطلاعاتی که در این بخش حذف می‌شوند ازجمله علائم نگارشی و یا علامت فراخوانی افراد هیچ‌گونه بار احساسی ندارد؛ ازاین‌رو حذف آن‌ها نه‌تنها مخرب نمی‌باشد بلکه در کاهش حافظه مصرفی و زمان اجرای الگوریتم و افزایش دقت فرآیند تحلیل احساس بسیار مؤثر می‌باشد.

جدول ۲- لیست علائم حذف‌شده از توئیت کاربران در مرحله پیش‌پردازش

علامت فراخوانی افراد: @
علامت #
لینک‌های اینترنتی
اعداد : ۱-۹
فضای خالی
شکلک‌ها (ایموجی‌ها)
علائم نگارشی: "؛", "،", ".", "-", "_", "^", "(", ")", "[", "]", "&", "?", "٪", "+", ":",



شکل ۱- دیاگرام مراحل مدل پیشنهادی

Removing Stop Words: پس از فرآیند فوق دومین بخش در مرحله پیش پردازش، حذف کلمات بازدارنده از مجموعه کلمات تشکیل شده در مرحله قبل است؛ این کلمات شامل کلمات ربطی و اضافه هستند که هیچ‌گونه بار احساسی ندارند. غالباً در متن‌کاوی لیست کلمات بازدارنده برای هر زبان به‌صورت آماده قابل دریافت است. در این مدل نیز از لیست کلمات بازدارنده دریافت شده از آدرس XPO6[۱۳] استفاده شده است؛ که این لیست بالغبر ۵۰۰ کلمه را شامل می‌شود. در شکل ۲ تعدادی از کلمات بازدارنده موجود در لیست استفاده شده در تقسیم جملات این پژوهش ارائه شده است.

"am", "amid", "amidst", "an", "and", "anent", "any", "as", "aslant", astride", "at", "athwart", "away", "back", "bar", "betwixt", "beyond", "by", "circa", "close", "concerning",”

شکل ۲- تعدادی از کلمات بازدارنده استفاده شده

فرآیند وزن‌دهی به کلمات یکی از مراحل اصلی در دسته‌بندی متون می‌باشد و تأثیر مستقیم در نتایج حاصل از کلاس‌بندی و دقت کلاس‌بندی دارد. یکی از متداول‌ترین معیارهای وزن‌دهی که در مدل فضای بردار استفاده می‌شود معیار وزن دهی TF-IDF است. این معیار که از رابطه (۱) حاصل می‌شود یک روش بدون ناظر می‌باشد و دارای نسخه‌های مختلفی ازجمله لگاریتمی و بولین می‌باشد [۴۵].

$$TF\_IDF(t,d,D) = tf \times \log\frac{N}{n_t + 1} \qquad (۱)$$

در رابطه فوق N تعداد کل اسناد است و nt بیانگر تعداد اسنادی است که حاوی کلمه t هستند. در این معیار TF[۱۴] تعداد حضور یک کلمه در یک سند می‌باشد وIDF[۱۵] یا فراوانی معکوس تعداد حضور کلمه موردنظر در کل اسناد را ارائه می‌دهد [۴۴]. معیار سنتی TF-IDF برای کلاس‌بندی متون کاملاً مؤثر نمی‌باشد. ازاین‌رو در پژوهش‌های مختلفی نسخه‌های مختلفی از آن ارائه شده است. طبقه‌بندی متن به‌عنوان یادگیری ماشین تحت نظارت مطرح می‌باشد که نیاز به مجموعه‌ای از متن با کلاس‌هایی دارد که برای آموزش مدل طبقه‌بندی استفاده می‌شوند. اما معیار TF-IDF در فرآیند وزن‌دهی به کلمات از اطلاعات کلاس داده‌های آموزشی استفاده نمی‌کند و این اطلاعات را در فرآیند وزن دهی دخیل نمی‌کند ، بنابراین وزن محاسبه

شده نمی‌تواند به‌طور کامل نشان دهنده اهمیت کلمه در کلاس‌های مختلف متن باشد. در شکل ۳ شبه کد روش پیشنهادی آمده است.

HCRF and TF-IGM Sentiment Analysis

Input: Train Corpus, Labels, Test Corpus, Labels-test, Top k Features
Output: Confusion Matrix, Accuracy, Precision, Recall, F.Score.
Begin
Read Test and Train data
// Preprocess
Tokenization: list<words>= Split sentence to words.
Remove Stop word: delete some words from list<words>
Delete redundant Words from list<word>
Calculate weights of words using eq.3-1, eq.3-2
//Features Selection
Select top k Features using Info Gain with eq.3-3
//Create Models
Create Markov models
Teach= Hidden Markov Classifier (Baum Welch Learning)
HMM= Teach.Learn (train data, labels)
Base_Func= Hidden Conditional Random Field from Hidden Markov (HMM)
Rprop = Hidden Resilient Gradient Learning (Base_Func);
HCRF = Rprop.Learn (train data, labels)
Predicted Labels = HCRF.Decide (Test data)
// Evaluation
Confusion matrix= compare (Predicted Labels, Labels-test)
Return (Accuracy, Precision, Recall, F.Score)
End

شکل ۳- شبه کد مراحل اصلی مدل پیشنهادی

به‌عنوان‌مثال، اگر فرض شود که دو کلمه با فرکانس‌های سند[۱۶] یکسان وجود دارند، که یکی که در کلاس‌های مختلف ظاهر می‌شود درحالی‌که دیگری فقط در یک کلاس از متن ظاهر می‌شود. این‌طور به نظر می‌رسد کلمه دوم دارای قدرت بیشتری برای تفکیک کلاس‌ها در مقایسه با کلمه اول می‌باشد، اما معیار وزن سراسری آن‌ها، یا به‌اختصار IDF، یکسان می‌باشد. از سوی دیگر، TF-IDF اهمیت اصطلاح را در یک سند بیش‌ازحد مورد توجه قرار می‌دهد، اما سهم آن را در طبقه‌بندی متن نادیده می‌گیرد. در این روش یک کلمه خاص در یک کلاس از متن، به احتمال زیاد یک وزن TF-IDF پایین‌تر از یک کلمه نادر می‌گیرد. با توجه به ضعف‌های معیار وزن‌دهی TF-IDF در این پژوهش از یکی از نسخه‌های بهبود یافته آن، تحت عنوان معیار TF-IGM[۱۷] استفاده می‌شود، که در سال ۲۰۱۶ توسط چن و همکاران ارائه شده است [۱۰]. در این معیار وزن‌دهی تعداد حضور کلمات در کلاس‌های مثبت و منفی به‌صورت مجزا محاسبه شده و بیشترین مقدار برای محاسبه وزن استفاده می‌شود. پس از حذف کلمات بازدارنده و علائم نگارشی و دیگر علائم غیرضروری فرآیند وزن دهی به کلمات با استفاده از دو رابطه (۲) و (۳) انجام می‌شود.

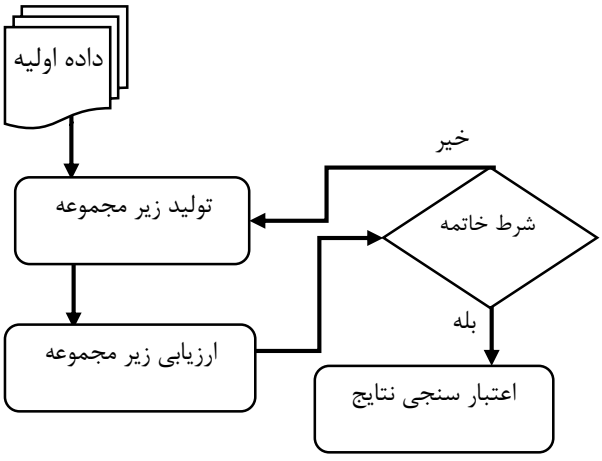
$$TF_{IGM(t\,k\,,d)} = tf_{kd} \cdot \left(1 + \lambda \cdot \frac{f_{k1}}{\sum_{r=1}^{m} f_{kr} \cdot r}\right) \quad r = 1, 2, \dots, m \qquad (۲)$$

$$RTF_{IGM(t\,k\,,d)} = \sqrt{tf_{kd}} \cdot \left(1 + \lambda \cdot \frac{f_{k1}}{\sum_{r=1}^{m} f_{kr} \cdot r}\right) \quad r = 1, 2, \dots, m \qquad (۳)$$

در روابط فوق f_{kr} تعداد رخداد کلمه موردنظر در کلاس r ام است. m در رابطه فوق تعداد کلاس‌ها می‌باشد و tf_{kd} تعداد رخداد کلمه موردنظر در سند d می‌باشد. در رابطه فوق λ متغیری قابل تنظیم بین ۰ تا ۱ است که یک تعادل نسبی بین وزن سراسری و محلی برقرار می‌کند [۱۰]. تنها تفاوت این دو معیار در مقدار TF است که در رابطه دوم ریشه آن در فرمول در نظر گرفته شده است. پس از انجام فرآیند پیش پردازش لیستی از کلمات وزن دار که درواقع ویژگی‌ها مسئله می‌باشند تشکیل می‌شوند. ویژگی‌ها اگرچه تا حد زیادی در مرحله پیش پردازش پالایش شده‌اند و کلمات بی ارزشی که بار احساسی ندارند حذف شده است، اما بازهم ویژگی‌ها و یا کلماتی در بین آن‌ها وجود دارند که از دیگر کلمات دارای بار احساسی بیشتری بوده و در دسته‌بندی جملات مؤثرتر هستند؛ برای یافتن این قبیل از کلمات یا ویژگی‌ها که در حقیقت ویژگی‌ها برتر محسوب می‌شوند از روش‌های انتخاب ویژگی استفاده می‌شود. در این پژوهش نیز از روش انتخاب ویژگی فیلتر و الگوریتم Information Gain برای رسیدن به هدف فوق و انتخاب ویژگی‌ها برتر استفاده شده است که در ادامه به تشریح این مرحله از مدل پرداخته می‌شود.

۳-۲- مرحله انتخاب ویژگی با استفاده از الگوریتم Information Gain

یک روش انتخاب ویژگی فارغ از نوع آن به‌طورکلی از اجزای اصلی، فرایند تولید یا تابع جستجو ، ارزیابی زیرمجموعه ، معیار توقف و تابع تعیین اعتبار تشکیل شده است. که تعامل هر یک از این اجزا با یکدیگر در شکل ۴ نمایش داده شده است. در فرآیند انتخاب ویژگی تابع جستجو یا تولید، زیرمجموعه‌های مختلف از ویژگی‌ها را تولید می‌کند، سپس هر یک از زیرمجموعه‌های تولید شده با استفاده از تابع ارزیابی بررسی می‌شوند و درصورتی‌که نتیجه مطلوب باشد به‌عنوان بهترین ویژگی‌ها انتخاب می‌شوند [۳۰]. تابع تولیدکننده زیرمجموعه‌ای از ویژگی‌ها، در سه حالت، پیشرو، پسرو و تولید تصادفی به تشکیل زیرمجموعه‌ای از ویژگی‌ها می‌پردازد. در حالت پیشرو، تولید زیر مجموعه ویژگی جدید با حداقل تعداد ویژگی آغاز می‌گردد و در هر مرحله به تعداد ویژگی‌ها افزوده می‌شود و زیر مجموعه جدید تولید می‌شود؛ این روند تا رسیدن به بهترین زیر مجموعه ادامه می‌یابد [۳۰].



شکل ۴- چهار جزء اصلی در یک روش انتخاب ویژگی

در حالت پسرو برخلاف روش قبل، ابتدا کل مجموعه ویژگی‌ها انتخاب شده و سپس در هر مرحله، ویژگی‌ها با ارزش کمتر حذف می‌شوند تا جایی که بهترین زیر مجموعه حاصل شود [۳۴]. و در آخرین روش زیرمجموعه‌های تصادفی تولید شده و برای هر یک، با استفاده از تابع ارزیابی مقدار برازش محاسبه می‌شود و زیر مجموعه‌ای که دارای بهترین برازش باشد جایگزین مجموعه قبلی خواهد شد [۳۵]. شرط خاتمه در فرآیند انتخاب ویژگی یا بر اساس تابع تولید است یا بر اساس تابع ارزیابی است و به‌طورکلی باعث کوتاه‌تر شدن فرآیند انتخاب ویژگی می‌شود [۳۰،۳۳،۳۵].

پس از فرآیند پیش پردازش و وزن‌دهی به کلمات، فرآیند انتخاب ویژگی برای بهبود دقت نتایج انجام می‌شود. درواقع هدف از این بخش این است که مشخص شود، کدام ویژگی‌ها در مجموعه بردارهای ویژگی آموزش برای ایجاد تمایز بین کلاس‌ها بسیار مفید است. برای این کار در این بخش از الگوریتم Information Gain استفاده شده است. این معیار از آنتروپی[۱۸] هر یک از ویژگی‌ها استفاده کرده و به میزان اطلاعاتی که می‌توان از یک ویژگی به دست آورد، اطلاق می‌شود.

اگر فرض شود T مجموعه داده‌های آموزشی باشد؛ که هر عضو آن به‌صورت $(X,y)=(x_1,x_2,\dots x_n,y)$ باشد به‌طوری‌که در آن $x_i \in val(i)$ بیانگر وزن ویژگی i در نمونه X باشد و y برچسب نمونه موردنظر باشد آنگاه مقدار Information Gain برای یک ویژگی i ازلحاظ آنتروپی $H()$ از رابطه (۴) حاصل می‌شود[۳۹]:

$$IG(T,i)=H(T)-\sum_{v\in val(i)}\frac{|\{\boldsymbol{X}\in T|x_i=v\}|}{|T|}\cdot H(\{\boldsymbol{X}\in T|x_i=v\}) \qquad (۴)$$

در رابطه فوق H تابع آنتروپی است که از رابطه (۵) حاصل می‌شود.

$$H(X)=\sum_i -P_i\log_2 P_i \qquad (۵)$$

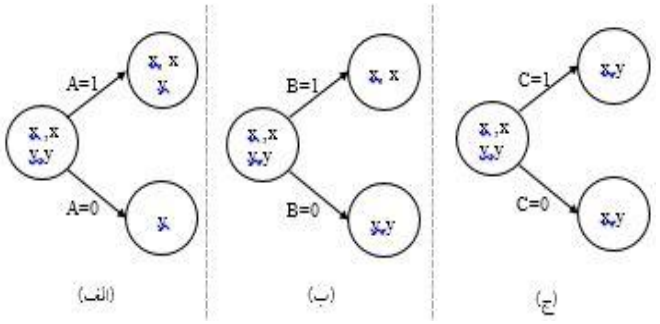
در رابطه فوق P_i احتمال کلاس i است که آن را به‌عنوان نسبت کلاس i در مجموعه محاسبه می‌کند. با ذکر یک مثال نحوه محاسبه Information Gain بررسی می‌گردد.

اگر فرض شود مجموعه داده‌ای متشکل از دو کلاس x و y با سه ویژگی A وB وC به‌صورت ارائه شده در جدول ۳ وجود داشته باشد. آنگاه آنتروپی و Information Gain برای هر ویژگی به‌صورت زیر محاسبه می‌گردد.

ابتدا ویژگی A بررسی می‌شود. تفکیک کلاس‌ها بر اساس ویژگی A در شکل ۵ بخش (الف) نمایش داده شده است. آنتروپی برای هر گره در شکل ۵ محاسبه شده و سپس مقدار Gain ویژگی A با استفاده از مقادیر آنتروپی محاسبه می‌گردد. سپس فرآیند فوق برای دو ویژگی دیگر تکرار می‌شود.

جدول ۳- مجموعه داده متشکل از سه ویژگی متعلق به دو کلاس

A	B	C	Class
1	1	1	X
1	1	0	X
0	0	1	Y
1	0	0	y



شکل ۵- تفکیک کلاس‌ها بر اساس هر ویژگی به‌صورت مجزا

$$Entropy_{child1}=-\left(\frac{1}{3}\right)log_2\left(\frac{1}{3}\right)-\left(\frac{2}{3}\right)log_2\left(\frac{2}{3}\right)$$
$$=0.5284+0.39=0.9184$$
$$Entropy_{Child2}=0$$
$$Entropy_{Parent}=1$$
$$Info\ Gain(A)=1-\left(\frac{3}{4}\right)\cdot(0.9184)-\left(\frac{1}{4}\right)\cdot(0)=0.3112$$

برای ویژگی B با توجه به شکل ۵ بخش (ب) مقادیر آنتروپی و Gain به‌صورت زیر می‌شود.

$$Entropy_{child1}=0$$
$$Entropy_{Child2}=0$$
$$Entropy_{Parent}=1$$
$$Info\ Gain(B)=1-\left(\frac{1}{2}\right)\cdot(0)-\left(\frac{1}{2}\right)\cdot(0)=1 \qquad Best$$

حال با توجه به شکل ۵ بخش (ج) مقادیر آنتروپی و Gain برای ویژگی C محاسبه می‌شود.

$$Entropy_{child1}=1$$
$$Entropy_{Child2}=1$$
$$Entropy_{Parent}=1$$
$$Info\ Gain(C)=1-\left(\frac{1}{2}\right)\cdot(1)-\left(\frac{1}{2}\right)\cdot(1)=0 \qquad Worst$$

از مقادیر حاصل شده می‌توان دریافت که از بین سه ویژگی فوق ویژگی B دارای بهترین Gain و ویژگی C دارای بدترین Gain است.

مقدار Gain هر چه بیشتر باشد نشان دهنده مفیدتر بودن ویژگی موردنظر برای تفکیک کلاس‌ها از یکدیگر است. پس از محاسبه مقدار Information Gain برای هر یک از ویژگی‌ها، ویژگی‌ها وزن دهی شده بر اساس مقدار Information Gain به‌صورت نزولی مرتب می‌شوند. سپس به تعداد K مشخص شده توسط کاربر برترین ویژگی‌ها از ابتدای لیست مرتب شده انتخاب می‌شوند و به‌عنوان ویژگی‌ها نهایی به مرحله طبقه‌بندی ارسال می‌شوند. ابعاد مجموعه داده آموزشی و مجموعه داده آزمایشی با توجه به ویژگی‌ها منتخب کاهش می‌یابد.

۳-۳- مرحله طبقه‌بندی با استفاده از الگوریتم میدان تصادفی شرطی مخفی

مدل مارکوف مخفی را می‌توان به‌عنوان نسخه‌ی دنباله‌ای از یک مدل بیز ساده در نظر گرفت: به‌جای یک تصمیم‌گیری مستقل، مدل مارکوف مخفی یک توالی خطی از تصمیمات را مدل می‌کند. بر این اساس، میدان تصادفی شرطی را نیز می‌توان، به‌عنوان نسخه دنباله‌ای (توالی) مدل‌های حداکثر آنتروپی در نظر گرفت. میدان‌های تصادفی شرطی نیز از دسته مدل‌های تفکیکی می‌باشند. علاوه بر این، در مقایسه با مدل‌های مارکوف مخفی، میدان‌های تصادفی شرطی، ملزم به داشتن ساختار دنباله خطی نمی‌باشند و می‌توانند به‌طور دلخواه ساختاربندی شوند.

میدان‌های تصادفی شرطی در سال ۲۰۰۱ توسط لافرتی و همکارانش [۴۹] ارائه شد، CRFها مدل‌های احتمالی برای برچسب‌زنی داده‌های متوالی می‌باشند. درواقع CRFها مدل‌های احتمالی برای محاسبه‌ی $p(\vec{y}|\vec{x})$ از خروجی ممکن $\vec{x}=(x1, \cdots, xn) \in \mathfrak{x}n$، با توجه به ورودی یا مشاهده‌ی $\vec{x}=(y1, \cdots, yn) \in \mathfrak{Y}n$ می‌باشند CRF .به‌طورکلی می‌تواند از فرمول (۶) حاصل شود.

$$P(\vec{v}) = \frac{1}{z}\prod_{c\in C}\Psi_c(\vec{v}_c) \qquad (۶)$$

احتمال شرطی $p(\vec{y}|\vec{x})$ را می‌توان به شکل زیر نوشت:

$$P(\vec{y}|\vec{x}) = \frac{p(\vec{x},\vec{y})}{p(\vec{x})} = \frac{p(\vec{x},\vec{y})}{\sum_{\vec{y'}} p(\vec{y'},\vec{x})} = \frac{\frac{1}{z}\prod_{c\in C}\psi_c(\vec{x}_c,\vec{y}_c)}{\frac{1}{z}\sum_{\vec{y'}}\prod_{c\in C}\psi_c(\vec{x}_c,\vec{y'}_c)} \qquad (۷)$$

از این طریق، فرمول کلی CRF ها به دست می‌آید:

$$P(\vec{y}|\vec{x}) = \frac{1}{z(\vec{x})}\prod_{c\in C}\psi_c(\vec{x}_c\,,\vec{y}_c) \qquad (۸)$$

Ψc عامل‌های مختلف متناظر با کلیک‌های ماکسیمال در گراف استقلال می‌باشند. به‌عنوان‌مثال در شکل ۶ نمونه‌ای از میدان تصادفی شرطی زنجیره‌ی-خطی نشان داده شده است. هر عامل مربوط به یک تابع پتانسیل است که ویژگی‌ها متفاوت fi را بر اساس قسمت در نظر گرفته شده از مشاهده و خروجی باهم ترکیب می‌کند. مخرج کسر معادله (۸) به‌عنوان ضریب نرمال‌سازی شناخته می‌شود و برابر معادله (۸) می‌باشد.

$$Z(\vec{x}) = \sum_{\vec{y'}}\prod_{c\in C}\psi_c(\vec{x}_c\,,\vec{y'}_c) \qquad (۹)$$

در حقیقت، در طول هر دو مرحله‌ی آموزش و استنتاج، برای هر نمونه یک گراف جداگانه مورد استفاده قرار می‌گیرد که از قالب‌های کلیک ساخته می‌شوند. قالب‌های (الگوهای) کلیک بر اساس ساختار داده‌های پایه از طریق تعریف ترکیب کلیک‌ها شکل می‌گیرند. هر کلیک مجموعه‌ای از متغیرهای وابسته به هم هستند، یعنی آن‌هایی که در تابع پتانسیل مربوطه قرار دارند.

CRFهای زنجیره‌ای خطی، شکل خاصی از یک CRF، که دارای ساختار زنجیره خطی می‌باشد و متغیرهای خروجی را به‌صورت یک دنباله مدل می‌کند. شکل۵ گراف استقلال و گراف عامل معادل با میدان تصادفی شرطی زنجیره‌ی خطی را نشان می‌دهد CRFهای معرفی شده در معادله (۸) را می‌توان به شکل زیر نیز فرمول‌بندی کرد:

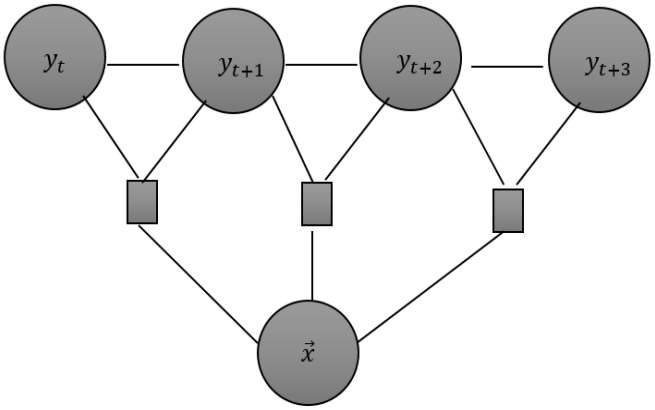
$$P(\vec{y}|\vec{x}) = \frac{1}{z(\vec{x})}\prod_{j=1}^{n}\psi_j(\vec{x}\,,\vec{y}) \qquad (۱۰)$$

در معادله (۱۰)، $Z(x)$ برابر است با:

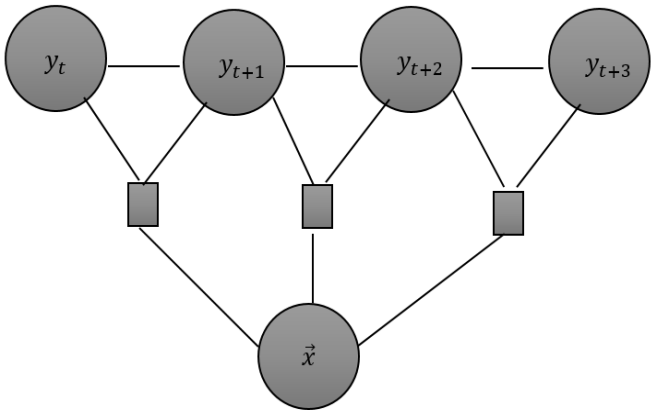
$$Z(\vec{x}) = \sum_{\vec{y'}}\prod_{j=1}^{n}\psi_j(\vec{x}_c\,,\vec{y'}) \qquad (۱۱)$$

در میدان‌های تصادفی شرطی زنجیره‌ی خطی، عامل‌های $\psi j(\vec{x}\,,\vec{y})$ به‌صورت معادله (۱۲) در نظر گرفته می‌شوند.

$$\psi_j(\vec{x},\vec{y}) = \exp\left(\sum_{i=1}^{m}\lambda_i f_i(\vec{y}_{j-1},\vec{y}_j,\vec{x},j)\right) \qquad (۱۲)$$



الف) گراف استقلال



ب) گراف عامل

شکل ۶- یک میدان تصادفی شرطی زنجیره‌ی خطی

با توجه به معادله و با فرض n+1 به‌عنوان طول دنباله مشاهده، CRF زنجیره خطی به‌صورت زیر نوشته می‌شود:

$$p_{\vec{\lambda}}(\vec{y}|\vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})}\exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i(\vec{y}_{j-1},\vec{y}_j,\vec{x},J)\right) \qquad (۱۳)$$

اندیس J در مقایسه با مدل حداکثر آنتروپی مورد نیاز است زیرا در میدان تصادفی شرطی یک دنباله برچسب به‌جای یک برچسب واحد برای پیش‌بینی در نظر گرفته می‌شود (در معادله (۱۳) اندیس J موقعیت در دنباله ورودی $\vec{x}$ را مشخص می‌کند. باید توجه شود که که وزنهای λi به موقعیت J وابسته نیست. این تکنیک، به نام گره پارامتر شناخته می‌شود، برای اطمینان از یک مجموعه مشخص از متغیرها برای داشتن مقدار یکسان، استفاده می‌شود. در رابطه (۱۳)، n طول دنباله، mتعداد توابع ویژگی، fiها توابع ویژگی، λiها ضریب تأثیر توابع ویژگی وZ عامل نرمال‌کننده تابع احتمال است که فقط به دنباله مشاهدات وابسته است و نرمال‌سازی در بازه‌ی [۰،۱] به‌صورت (۱۴) انجام می‌شود:

$$Z_{\vec{\lambda}}(\vec{x}) = \sum_{y\in Y}\exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i(\vec{y}_{j-1},\vec{y}_j,\vec{x},J)\right) \qquad (۱۴)$$

جمع (مجموع) روی y، مجموعه‌ای از تمام دنباله برچسب‌های ممکن، برای رسیدن به یک احتمال محتمل (عملی) می‌باشد. برای به‌کارگیری روش میدان تصادفی شرطی، متناسب با مسئله باید تعدادی تابع ویژگی تعریف کرد. نقش توابع ویژگی، تشویق الگوهای مناسب و تنبیه الگوهای نامناسب برچسب‌گذاری است. درواقع سعی می‌کنیم برچسب‌گذاری‌های دلخواه در دنباله را در قالب توابع ویژگی

بیان کنیم. از رابطه‌ی بالا مشخص است که توابع ویژگی، تابعی از نمونه i-ام، برچسب احتمالی آن و برچسب احتمالی نمونه قبل از آن می‌باشد. معمولاً به جهت سادگی، توابع ویژگی دودویی طراحی می‌شوند اما در حالت کلی می‌توانند هر مقدار حقیقی داشته باشند. به‌عنوان‌مثال در مسئله برچسب‌گذاری واژگان در زبان انگلیسی، توابع زیر با توجه به دستور زبان انگلیسی قابل تعریف می‌باشند:

$$f_1(x_i, y_i, y_{i-1}) = \begin{cases} 1 & \text{if } y_i - 1 = \text{ADJECTIVE and } y_i = \text{NOUN} \\ 0 & else \end{cases} \quad (۱۵)$$

$$f_1(x_i, y_i, y_{i-1}) = \begin{cases} 1 & \text{if } y_i - 1 = \text{ADJECTIVE and } y_i = \text{NOUN} \\ 0 & else \end{cases} \quad (۱۶)$$

$$f_3(x_i, y_i, y_{i-1}) = \begin{cases} 1 & \text{if } y_i = ADVERB \text{ and } x_i \text{ends with 'ly'} \\ 0 & else \end{cases} \quad (۱۷)$$

تابع f1 به برچسب‌گذاری ای که اسم پس از صفت بیاید امتیاز مثبت می‌دهد. تابع f2 به برچسب‌گذاری‌ای که در آن دو نمونه پشت سرهم «حرف اضافه» تشخیص داده شوند امتیاز منفی می‌دهد) منفی بودن این امتیاز از طریق ضرایب λ در گام یادگیری لحاظ خواهد شد. به چنین توابعی که تنها به برچسب نمونه‌ها وابسته هستند و نه خود نمونه ویژگی لبه‌ای (جفتی) گفته می‌شود که در حالت دودویی در قالب زیر تعریف می‌شوند:

$$f^{edge}(x_i, y_i, y_{i-1}) = \delta(y_i, l) * \delta(y_{i-1}, l') \quad (۱۸)$$

$$\delta(y_i, l) = \begin{cases} 1 & \text{if } y_i = l \\ 0 & else \end{cases} \quad (۱۹)$$

واضح است که هر چه تعداد این ویژگی‌ها بیشتر باشد، دسته‌بندی بیشتر وابسته به دانش قبلی به‌دست‌آمده از مسئله و مستقل از داده‌های حاضر می‌شود.

تابع f3، یک نمونه تابع ویژگی را نشان می‌دهد که در آن از ویژگی‌ها نمونه‌ی i-ام استفاده شده است. در این تابع برچسب‌گذاری «قید» برای کلماتی که پسوند «ly» دارند، تشویق شده است. به چنین توابعی که تنها به یک نمونه و برچسب آن وابسته هستند، ویژگی گره‌ای (تکی) گفته می‌شود که در حالت دودویی به شکل زیر تعریف می‌شوند:

$$f^{edge}(x_i, y_i, y_{i-1}) = \delta(y_i, l) * g(x_i) \quad (۲۰)$$

که در آن g تابعی دودویی (یا حقیقی) است که ویژگی‌ها خود نمونه را مدنظر قرار می‌دهد. این توابع درواقع متناظر با همان نوع نگاه مستقل به هر داده در الگوشناسی آماری می‌باشند. این‌ها، تنها مثال‌های از انواع توابع ویژگی می‌باشند که در طول زمان به‌صورت الگو درآمده‌اند. اما به‌طورکلی توابع ویژگی طبق رابطه (۱۳) می‌توانند در یک لحظه به هر دو برچسب نمونه‌های حال و قبلی و همین‌طور ویژگی‌ها نمونه حاضر وابسته باشند.

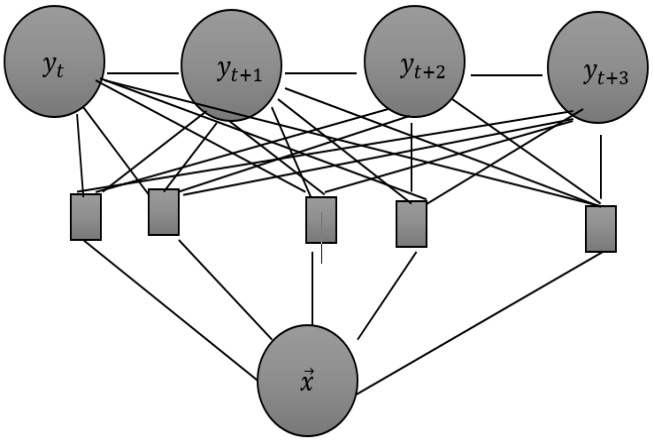
در معادله (۲۰) یک فرمول‌بندی از CRF زنجیره‌ی خطی داده شده است. با جابه‌جایی جمع‌های انجام شده بر روی موقعیت‌های دنباله در جلوی تابع نمایی، عامل بندی واقعی که برای CRF به‌طورمعمول مورد استفاده قرار می‌گیرد به‌صورت معادله به دست می‌آید:

$$p_{\vec{\lambda}}(\vec{y}|\vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \prod_{j=1}^{n} \exp\left(\sum_{i=1}^{m} \lambda_i f_i(\vec{y}_{j-1}, \vec{y}_j, \vec{x}, J) \right) \quad (۲۱)$$

گراف عامل متناظر با این عامل بندی در شکل ۶(ب) نشان داده شده است. همچنین با جابه‌جایی جمع انجام شده بر روی ویژگی‌ها متفاوت در جلوی تابع نمایی، می‌توان عامل بندی دیگری را برای CRF به دست آورد که به‌صورت معادله (۲۲) خواهد بود،

$$p_{\vec{\lambda}}(\vec{y}|\vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \prod_{i=1}^{m} \exp\left(\sum_{j=1}^{n} \lambda_i f_i(\vec{y}_{j-1}, \vec{y}_j, \vec{x}, J) \right) \quad (۲۲)$$

در این تفسیر از میدان‌های تصادفی شرطی، عامل‌ها در عوض اجرا شدن بر روی دنباله، بر روی ویژگی‌ها اجرا می‌شوند. گراف عامل با عامل‌های $\psi_i = \exp\left(\sum_{j=1}^{n} \lambda_i f_i(\vec{y}_{j-1}, \vec{y}_j, \vec{x}, J)\right)$ متناظر با ویژگی‌ها fi در شکل ۷ نشان داده شده است.



شکل ۷- تفسیر جایگزین از یک CRF زنجیره‌ی خطی

این تفسیر درک شهودی کمتری دارد، اما نشان‌دهنده‌ی ارتباط با مدل حداکثر آنتروپی می‌باشد. همچنین با جابه‌جایی هر دو جمع انجام شده در جلوی تابع نمایی، می‌توان تفسیر جدیدی از CRF ارائه کرد، که دارای تعداد بیشتری عامل می‌باشد.

$$p_{\vec{\lambda}}(\vec{y}|\vec{x}) = \prod_{i=1}^{m} \prod_{j=1}^{n} \exp\left(\lambda_i f_i(\vec{y}_{j-1}, \vec{y}_j, \vec{x}, J)\right) \quad (۲۳)$$

در اینجا از نمایش گراف عامل مربوط به این تفسیر به دلیل تعداد زیاد عامل‌ها در گراف صرف‌نظر شده است. عامل بندی مبتنی بر کلیک‌های ماکسیمال معادله‌ی (۱۹) معمولاً برای CRF زنجیره خطی استفاده می‌شوند. دو روش عامل بندی دیگر معادله (۲۰ و ۲۱) به این بیشینگی پایبند نمی‌باشند. به‌طورکلی، عامل بندی با توجه به کلیک‌هایی که شامل گره‌های متغیر کمتر از کلیک ماکسیمال می‌باشند، منجر به عدم دقت در عامل بندی می‌گردد، زیرا تمام وابستگی‌های موجود به‌درستی در نظر گرفته نمی‌شوند. بااین‌حال، در این حالت، همان‌طور که در معادله (۲۳) دیده می‌شود، این امر منجر به محاسبات اضافی می‌شود.

آخرین مرحله و اصلی‌ترین مرحله در فرآیند تحلیل احساس طبقه‌بندی هر نمونه داده به کلاس موردنظر است. در این بخش ابتدا الگوریتم طبقه‌بندی که در این پژوهش میدان تصادفی شرطی مخفی می‌باشد، با استفاده از بخشی از داده‌ها که به نام داده‌های آموزشی هستند، آموزش دیده و مدلی بر اساس آن‌ها تولید می‌کند. سپس مدل تولید شده برای پیشگویی کلاس در بخش دیگری از داده‌ها تحت عنوان داده‌های آزمایشی، استفاده می‌شود. پس از پایان فرآیند پیشگویی مدل، صحت عملکرد آن با استفاده از معیارهای مختلفی بررسی می‌گردد.

ترکیب دو مدل مخفی مارکوف و میدان تصادفی شرطی برای تولید مدل میدان تصادفی شرطی مخفی به‌صورت زیر انجام شده است:در این بخش ابتدا یک مدل مخفی مارکوف با ورودی‌های، تعداد حالت‌های[۱۹] مدل و توپولوژی مدل تعریف می‌گردد. توپولوژی‌های مختلفی برای سازماندهی حالات در مدل تعریف شده وجود دارد. به‌عبارت‌دیگر، این امر نشان می‌دهد که کدام نوع از توالی‌ها مجاز هستند و کدام‌یک از آن‌ها غیرممکن هستند. ازجمله توپولوژی‌های مطرح در مدل مارکوف Forward وErgodic که در این پژوهش از مدل اول استفاده شده است. هر توپولوژی تعداد حالات مدل مارکوف را به‌عنوان ورودی دریافت می‌کند.

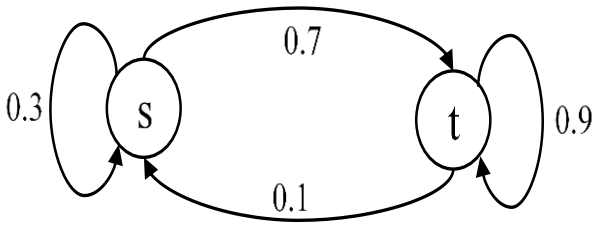
پس از تعریف مدل مارکوف، الگوریتمی برای آموزش آن باید تعریف گردد. در مدل پیشنهادی برای آموزش مدل مارکوف تعریف شده از الگوریتم بدون ناظر باوم-ولچ[۲۰] [۴۰] استفاده شده است، که فرآیند آموزش مدل را تا جایی که تغییرات مقدار لگاریتم درست‌نمایی[۲۱] به زیر حداقل آستانه تحمل[۲۲] برسد ادامه می‌دهد. مقدار لگاریتم درست‌نمایی نشان می‌دهد که هر نمونه داده ورودی به محتمل‌ترین کلاس خود تخصیص یافته است این تابع با استفاده از رابطه (۲۴) برای طبقه‌بندی باینری حاصل می‌شود.

$$likelihood = \sum_{i=1}^{n} y_i \, log\, p(x_i) + (1 - y_i)\, log\big(1 - p(x_i)\big) \quad (۲۴)$$

در رابطه فوق $p(x_i)$ احتمال پیش بینی شده برای نمونه i ام از داده که برابر با ۱ است و y_i برچسب نمونه i ام است.

الگوریتم باوم-ولچ در دسته الگوریتم‌های تکرار شونده قرار می‌گیرد و مقدار درست‌نمایی بیشینه را برای پارامترهای انتقال و انتشار مدل مارکوف محاسبه می‌کند. در این روش ابتدا پارامترهای مدل به‌صورت تصادفی تعیین می‌شوند و سپس در هر تکرار پارامترها به نحوی تغییر می‌یابند که مدل به داده‌های آموزشی نزدیک شود. الگوریتم باوم-ولچ دو ورودی تعداد تکرار و حداقل آستانه دریافت می‌کند. درصورتی‌که تعداد تکرار برای این الگوریتم مشخص نشود تا جایی که تغییرات درست‌نمایی به زیر حداقل آستانه تحمل برسد فرآیند آموزش هر مدل مارکوف را تکرار می‌کند. در روش پیشنهادی مقدار تحمل برابر با $1e-3$ در نظر گرفته شده است. نحوه عملکرد الگوریتم بام-ولچ در این بخش از مدل ارائه شده با ذکر یک مثال کوچک بررسی می‌شود [۴۱].

اگر فرض شود Y مجموعه‌ای از کلمات تحلیل نشده به‌صورت $\{ABBA, BAB\}$ باشد؛ و مجموعه داده c دارای ۱۰ کلمه (ABBA) و ۲۰ کلمه (BAB) و درمجموع دارای ۳۰ کلمه باشد، و اولین مدل مارکوف با نام h_1 همانند شکل ۸ تعریف شود.



شکل ۸- مدل مارکوف h_1

اگر احتمال شروع درs برابر با 0.85 و از t برابر با 0.15 باشد. و احتمال A وB در s به ترتیب برابر با، $P_r(B) = 0.6$,$p_r(A) = 0.4$ باشد. و همچنین در t $P_r(B) = 0.5$,$p_r(A) = 0.5$ باشد. حال اگر فرض شود $y = A_1, \dots A_n$ و $y \in Y$ باشد و n به‌عنوان طول y در نظر گرفته شود، در حالت s و برای $1 \le j \le n$ ،عبارت $\alpha(y, j, s)$ احتمال تعریف شده در فضای واژه‌های تحلیل شده است، که در آن اولین نمادهای j با y مطابقت می‌کنند، و حالت پایانی s است. با توجه به مطالب فوق دو احتمال (۲۵) و (۲۶) به‌صورت زیر مطرح است:

$$\alpha(y, 1, s) = start(s) out(s, A_1) \quad (۲۵)$$

$$\alpha(y, j+1, s) = \sum_{t \in S} \alpha(y, j, t)\, go(t, s)\, out(s, A_{j+1}) \quad (۲۶)$$

حال می‌توان احتمال عبارت ABBA را به‌صورت زیر محاسبه نمود:

$$\begin{aligned}
\alpha(ABBA,1,s) &= (0.85)(0.4) = 0.34\\
\alpha(ABBA,1,t) &= (0.15)(0.5) = 0.08\\
\alpha(ABBA,2,s) &= (0.34)(0.3)(0.6) + (0.08)(0.1)(0.6)\\
&= 0.06120 + 0.00480 = 0.066\\
\alpha(ABBA,2,t) &= (0.34)(0.7)(0.5) + (0.08)(0.9)(0.5)\\
&= 0.119 + 0.036 = 0.155\\
\alpha(ABBA,3,s) &= (0.066)(0.3)(0.6) + (0.155)(0.1)(0.6)\\
&= 0.01188 + 0.00930 = 0.02118\\
\alpha(ABBA,3,t) &= (0.066)(0.7)(0.5) + (0.155)(0.9)(0.5)\\
&= 0.02310 + 0.06975 = 0.09285\\
\alpha(ABBA,4,s) &= (0.02118)(0.3)(0.4)\\
&+ (0.09285)(0.1)(0.4)\\
&= 0.0025 + 0.0037 = 0.00625\\
\alpha(ABBA,4,t) &= (0.02118)(0.7)(0.5)\\
&+ (0.09285)(0.9)(0.5)\\
&= 0.0074 + 0.0417 = 0.04919
\end{aligned}$$

با توجه به مقادیر فوق مجموع احتمال ABBA به‌صورت زیر حاصل می‌شود:

$$\begin{aligned}
Total\ probability\ (ABBA) &= 0.00625 + 0.04919\\
&= 0.05544
\end{aligned}$$

احتمال عبارت BAB به‌صورت زیر محاسبه می‌شود:

$$\begin{aligned}
\alpha(BAB,1,s) &= (0.85)(0.6) = 0.51\\
\alpha(BAB,1,t) &= (0.15)(0.5) = 0.08\\
\alpha(BAB,2,s) &= (0.51)(0.3)(0.4) + (0.08)(0.1)(0.4)\\
&= 0.0612 + 0.0032 = 0.0644\\
\alpha(BAB,2,t) &= (0.51)(0.7)(0.5) + (0.08)(0.9)(0.5)\\
&= 0.1785 + 0.0360 = 0.2145\\
\alpha(BAB,3,s) &= (0.066)(0.3)(0.6) + (0.155)(0.1)(0.6)\\
&= 0.01188 + 0.0093 = 0.0209\\
\alpha(BAB,3,t) &= (0.0644)(0.7)(0.5) + (0.2145)(0.9)(0.5)\\
&= 0.0225 + 0.0965 = 0.1190\\
Total\ probability\ (BAB) &= 0.0209 + 0.1190\\
&= 0.1399
\end{aligned}$$

مقدار لگاریتم درست‌نمایی برای مجموعه داده c با استفاده از مدل مارکوف h_1 به‌صورت زیر محاسبه می‌شود:

$$\begin{aligned}
\mathrm{L(c, h_1)} &= \Pr(ABBA)^{c(ABBA)} \cdot \Pr(BAB)^{c(BAB)}\\
&= 0.05544^{10} 0.1399^{20}\\
\log \mathrm{L(c, h_1)} &= (10 \times \log 0.05544) + (20\\
&\times \log 0.1399) = -68.2611
\end{aligned}$$

فرآیند فوق برای مدل‌های مارکوف h_2, h_3, h_4 نیز تکرار می‌شود و الگوریتم درست‌نمایی هر یک محاسبه می‌شود و فرآیند آموزش هر یک از مدل‌ها تا جایی ادامه می‌یابد که تغییرات در مقدار درست‌نمایی به حداقل ممکن برسد.

پس از تکمیل فرآیند آموزش مدل مارکوف از آن برای ساخت مدل میدان تصادفی مخفی استفاده می‌شود. تابع اولیه در میدان تصادفی شرطی مخفی، مدل آموزش دیده مارکوف است. پس از تعریف مدل میدان تصادفی مخفی با استفاده از مدل آموزش دیده مارکوف، میدان تصادفی شرطی به‌صورت مجزا آموزش داده می‌شود. برای آموزش میدان تصادفی شرطی از الگوریتم‌های مختلفی ازجمله Conjugate-Gradient، Quasi-Newton Methods و Resilient Backpropagation [۴۲] که به‌اختصار به آن Rprop گفته می‌شود، می‌توان استفاده نمود. ازآنجایی‌که میدان‌های تصادفی شرطی مخفی در دسته مدل‌های متمایز کننده قرار می‌گیرند، می‌توان آن‌ها را به‌طور مستقیم با استفاده از روش‌های گرادیان[۲۳] بهینه کرد که مشابه آنچه معمولاً با شبکه‌های عصبی انجام می‌شود، صورت می‌گیرد. ازاین‌رو در مدل پیشنهاد شده از الگوریتم اکتشافی Rprop استفاده شده است که یکی از بهترین الگوریتم‌ها برای بهینه‌سازی میدان تصادفی شرطی مخفی است [۴۳]. الگوریتم Pprop نیز دو ورودی تعداد تکرار و حداقل آستانه تحمل دریافت می‌کند. و فرآیند آموزش مدل میدان تصادفی شرطی مخفی را تا زمانی که تغییرات درست‌نمایی به زیر حداقل آستانه تحمل برسد ادامه می‌دهد. برای آموزش مدل میدان تصادفی شرطی مخفی مقدار تحمل برابر با $1e-5$ در نظر گرفته شده است.

الگوریتم Rprop برای تخمین وزن هر یک از پارامترهای مدل، از مقدار بروز رسانی $\Delta_{i,j}$ که به‌تنهایی مقدار به‌روزرسانی را برای هر یک مشخص می‌کند استفاده می‌کند. این مقدار به‌روزرسانی تطبیقی در طول فرآیند یادگیری بر اساس تابع خطا E، مطابق با قانون یادگیری (۲۷) تکامل می‌یابد [۴۴]:

$$\Delta_{i,j}^{(t)}=\begin{cases}\eta^{+}*\Delta_{i,j}^{(t-1)}, & if \quad \frac{\delta E^{(t-1)}}{\delta wij}*\frac{\delta E^{(t)}}{\delta wij}>0\\ \eta^{-}*\Delta_{i,j}^{(t-1)}, & if \quad \frac{\delta E^{(t-1)}}{\delta wij}*\frac{\delta E^{(t)}}{\delta wij}<0\\ \Delta_{i,j}^{(t-1)}, & else\end{cases}$$

$$where\begin{cases}0<\eta^{-}<1<\eta^{+}\\ \frac{\delta E}{\delta w_{ij}}=\frac{\delta E}{\delta S_i}\frac{\delta S_i}{\delta net_i}\frac{\delta net_i}{w_{ij}}\end{cases} \qquad (۲۷)$$

در رابطه فوق η^{+} و η^{-} به ترتیب فاکتور افزایش و کاهش نرخ یادگیری هستند. به عبارت ساده مقدار $\Delta_{i,j}$ با استفاده از مقدار دو فاکتور مذکور کاهش و یا افزایش می یابد. در روش پیشنهادی مقدار دو فاکتور افزایش و کاهش به ترتیب برابر با $\eta^{-}=1.2$,$\eta^{+}=0.5$ در نظر گرفته شده است. همچنین در رابطه فوق w_{ij} وزن از حالت j به حالت i است، و S_i خروجی است و net_i مجموع وزن ورودی‌های حالت i می‌باشد. حداقل سازی مقدار تابع خطا در رابطه فوق با استفاده از تابع gradient descent و از رابطه(۲۸) حاصل می‌شود[۴۴]:

$$w_{i,j}(t+1)=w_{i,j}(t)-\Delta w_{i,j}(t) \qquad (۲۸)$$

انتخاب نرخ یادگیری، تأثیر مهمی در زمان مورد نیاز تا رسیدن به هم‌گرایی دارد. اگر خیلی کوچک باشد، برای رسیدن به یک راه‌حل قابل‌قبول گام‌های زیادی لازم است؛ برخلاف آن نرخ یادگیری زیاد، منجر به نوسان خواهد شد.

پس پایان فرآیند آموزش میدان تصادفی شرطی مخفی، از مدل تولید شده برای پیشگویی برچسب داده‌های آزمایشی استفاده می‌شود. پس از تولید برچسب‌های پیش بینی شده، می‌توان نحوه عملکرد مدل را با استفاده از ماتریس درهم‌ریختگی و معیارهای ارزیابی مختلف مورد ارزیابی قرار داد.

مدل تحلیل احساس معرفی شده در این پژوهش، یک مدل طبقه‌بندی کننده باینری می‌باشد. به این معنی که نمونه‌های داده را در دو کلاس مثبت و منفی دسته‌بندی می‌کند، به عبارت ساده مدل در مجموعه داده‌های آزمایشی کلاس و یا برچسب یک نظر را پیشگویی کرده و مشخص می‌کند که نظر مذکور مثبت است یا منفی. ارزیابی مدل‌های پیشگویی کننده با استفاده از لیست برچسب‌های پیشگویی شده و برچسب‌های حقیقی داده‌ها انجام می‌گیرد. برای انجام این بخش از ماتریس درهم‌ریختگی[۲۴] استفاده می‌شود. این ماتریس لیست برچسب‌های حقیقی داده‌های آزمایشی و لیست برچسب‌های پیشگویی شده توسط مدل برای داده‌های آزمایشی را دریافت کرده و با مقایسه دو لیست مذکور تعداد نمونه‌های مثبت و منفی که به‌درستی برچسب‌گذاری شده‌اند، را مشخص می‌کند.

یک نمونه از لیست برچسب نهایی مدل ارائه شده به شرح زیر است:

۱) مجموعه دادگان آموزشی نظرات کاربران شبکه اجتماعی توئیتر شامل ۱۲۰۰۰ کامنت دارای برچسب مثبت و منفی (شکل ۹).

۲) مجموعه دادگان آزمایش نظرات کاربران توئیتر شامل ۵۰۰ توئیت دارای برچسب مثبت و منفی (شکل ۱۰).

۳) تعداد TOP K ویژگی برتر که توسط الگوریتم انتخاب ویژگی تفکیک می‌شوند.

۴) مقدار متغیر آستانه تحمل، حداکثر تعداد تکرار ، تعداد حالت ، توپولوژی که به ترتیب برای میدان تصادفی شرطی و مدل مخفی مارکوف تعیین می‌شوند

```
Class,Index,Date,Unknown,UserID,Comment
0,1467810369,Mon Apr 06 22:19:45 PDT 2009,NO_QUERY,_TheSpecialOne_,"@switchfoot http://tw
0,1467810672,Mon Apr 06 22:19:49 PDT 2009,NO_QUERY,scotthamilton,is upset that he can't up
0,1467810917,Mon Apr 06 22:19:53 PDT 2009,NO_QUERY,mattycus,@Kenichan I dived many times t
0,1467811184,Mon Apr 06 22:19:57 PDT 2009,NO_QUERY,ElleCTF,my whole body feels itchy and I
0,1467811193,Mon Apr 06 22:19:57 PDT 2009,NO_QUERY,Karoli,"@nationwideclass no, it's not b
0,1467811372,Mon Apr 06 22:20:00 PDT 2009,NO_QUERY,joy_wolf,@Kwesidei not the whole crew
0,1467811592,Mon Apr 06 22:20:03 PDT 2009,NO_QUERY,mybirch,Need a hug
0,1467811594,Mon Apr 06 22:20:03 PDT 2009,NO_QUERY,coZZ,"@LOLTrish hey  long time no see!
0,1467811795,Mon Apr 06 22:20:05 PDT 2009,NO_QUERY,2Hood4Hollywood,@Tatiana_K nope they di
0,1467812025,Mon Apr 06 22:20:09 PDT 2009,NO_QUERY,mimismo,@twittera que me muera ?
0,1467812416,Mon Apr 06 22:20:16 PDT 2009,NO_QUERY,erinx3leannexo,spring break in plain ci
0,1467812579,Mon Apr 06 22:20:17 PDT 2009,NO_QUERY,pardonlauren,I just re-pierced my ears
```

شکل ۹- ساختار داده‌های آموزشی

```
Class,Index,Date,Unknown,UserID,Comment
1,3,Mon May 11 03:17:40 UTC 2009,kindle2,tpryan,"@stellargirl I loooooooovvvvvveee my k
1,4,Mon May 11 03:18:03 UTC 2009,kindle2,vcu451,Reading my kindle2...  Love it... Lee
1,5,Mon May 11 03:18:54 UTC 2009,kindle2,chadfu,"Ok, first assesment of the #kindle2 .
1,6,Mon May 11 03:19:04 UTC 2009,kindle2,SIX15,@kenburbary You'll love your Kindle2. I
1,7,Mon May 11 03:21:41 UTC 2009,kindle2,yamarama,@mikefish  Fair enough. But i have th
1,8,Mon May 11 03:22:00 UTC 2009,kindle2,GeorgeVHulme,@richardebaker no. it is too big
0,9,Mon May 11 03:22:30 UTC 2009,aig,Seth937,Fuck this economy. I hate aig and their n
1,10,Mon May 11 03:26:10 UTC 2009,jquery,dcostalis,Jquery is my new best friend.
1,11,Mon May 11 03:27:15 UTC 2009,twitter,PJ_King,Loves twitter
1,12,Mon May 11 03:29:20 UTC 2009,obama,mandanicole,how can you not love Obama? he mak
1,13,Mon May 11 03:32:42 UTC 2009,obama,jpeb,Check this video out -- President Obama a
0,14,Mon May 11 03:32:48 UTC 2009,obama,kylesellers,"@Karoli I firmly believe that Obam
```

شکل ۱۰- ساختار داده‌های آزمایشی

CLASS: مثبت و منفی بودن یک نظر را مشخص می‌کند که ۰ نشانگر نظر منفی و ۱ نظر مثبت است.

Index : شماره شاخص نظرات است.

Date : تاریخ و ساعت ثبت نظر توسط کاربر.

Unknown: ستون ناشناخته که در نظر گرفته نشده است.

UserId: شناسه کاربران توئیتر.

Comment : نظرات کاربران.

۴- معیارها و پیاده‌سازی

۴-۱- معیارهای ارزیابی

در زمینه مدل‌های مبتنی بر یادگیری ماشین معیارهای ارزیابی بسیار مختلفی در پژوهش‌های مختلف مطرح شده است که ازجمله می‌توان به Accuracy، Precision، Recall، F.Score، Sensitivity، Specificity و غیره اشاره نمود. در این پژوهش نیز از ۴ معیار اول برای ارزیابی و مقایسه مدل ارائه شده استفاده شده است که در ادامه به معرفی آن‌ها پرداخته و نحوه محاسبه هر یک ارائه می‌گردد.

- صحت Accuracy: این معیار از نسب تعداد پیشگویی‌های صحیح مدل به کل پیشگویی‌ها حاصل می‌گردد.
- دقت Precision: این معیار از نسب تعداد پیشگویی‌های صحیح در کلاس مثبت به کل پیشگویی‌های کلاس مثبت حاصل می‌شود.
- فراخوان Recall: این معیار از نسبت تعداد پیشگویی‌های صحیح در کلاس مثبت به کل تعداد حقیقی کلاس مثبت حاصل می‌شود.
- میانگین F: این معیار از میانگین دو معیار فراخوانی و دقت حاصل می‌شود.

معیارهای مورد ارزیابی در این پژوهش در روابط (۲۹) تا (۳۲) ارائه شده‌اند [۴۵،۴۶].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (۲۹)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (۳۰)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (۳۱)$$

$$\text{F\_Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (۳۲)$$

متغیرهای موجود در روابط فوق را می‌توان به‌صورت زیر تعریف نمود:

- True Positive یا به‌اختصار TP: بیانگر تعداد نمونه‌هایی است که کلاس حقیقی آن‌ها مثبت بوده و مدل به‌درستی کلاس آن‌ها را مثبت پیش بینی نموده است [۴۵].
- False Positive یا به‌اختصار FP: بیانگر نمونه‌هایی است که کلاس حقیقی آن‌ها منفی بوده ولی مدل به‌اشتباه آن‌ها را مثبت پیش بینی کرده است به عبارت ساده آن‌ها را به کلاس مثبت تخصیص داده است [۴۵].
- True Negative یا به‌اختصار TN: بیانگر نمونه‌هایی است که کلاس حقیقی آن‌ها منفی بوده و مدل به‌درستی آن‌ها را منفی پیش بینی کرده است [۴۵].

False Negative یا به‌اختصار FN: بیانگر نمونه‌هایی است که کلاس حقیقی آن‌ها مثبت بوده است اما مدل به‌اشتباه کلاس منفی برای آن‌ها پیشگویی کرده است [۴۵].

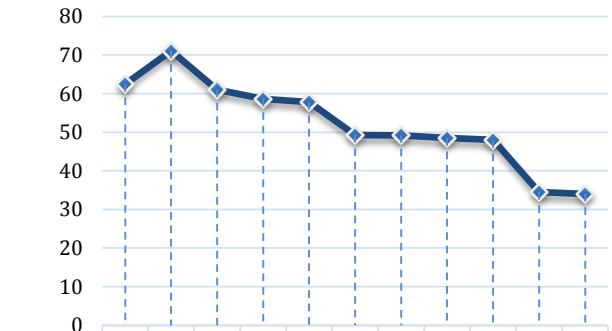
۴-۲- مشخصات پایگاه‌داده مورد استفاده

در این پژوهش از مجموعه دادگان STS [۲۵] دریافت شده از آدرس Stanford[۲۶] استفاده شده است. این مجموعه داده که متشکل از نظرات کاربران شبکه اجتماعی توئیتر می‌باشد، دارای دو ویژگی اصلی است: ۱) بخش آموزش و آزمایش در دو فایل CSV[۲۷] به‌صورت مجزا ارائه شده است. ۲) به‌صورت دستی هر دو مجموعه آموزش و آزمایش توسط تهیه‌کنندگان آن برچسب‌گذاری شده است.

۴-۳- طراحی آزمایش‌ها

در این بخش برای ارزیابی و مقایسه مدل ارائه شده سه دسته اصلی آزمایش انجام شده است. درواقع هدف از آزمایش‌های انجام شده در این بخش بررسی تأثیر متغیرهای مستقل تحقیق بر روی متغیرهای وابسته است. این تحقیق دارای یک متغیر مستقل تعداد ویژگی‌ها و چهار متغیر وابسته دقت، صحت، فراخوان و میانگین F می‌باشد که دسته‌ای از آزمایش‌ها به بررسی تأثیر تغییرات این متغیر بر روی متغیرهای وابسته تحقیق است. در این دسته از آزمایش‌ها تعداد ویژگی‌ها دریافتی در مرحله انتخاب ویژگی را تغییر داده و در هر مرحله نتایج هر چهار متغیر وابسته ذخیره می‌گردد. در دسته دیگری از آزمایش‌ها، با توجه به اینکه در این پژوهش از دو معیار وزن دهی جدید استفاده شده است، تأثیر معیار وزن‌دهی بر نتایج مدل ارزیابی شده است. در دسته سوم از آزمایش‌ها با بهترین تعداد ویژگی حاصل شده در آزمایش قبل، نتایج مدل پیشنهادی و مدل پایه ارائه شده در سال ۲۰۱۷ مورد مقایسه قرار می‌گیرد و میزان بهبودها در بین دو الگوریتم مشخص می‌گردد.

دسته اول: بررسی تأثیر تعداد ویژگی منتخب بر نتایج



	1000	2000	3000	4000	5000	6000	7000	10000	12000	15000	17000
Accuracy	62.5	71.1	61	58.6	57.8	49.3	49.2	48.5	48	34.5	34

شکل ۱۱- بررسی تأثیر مرحله انتخاب ویژگی در نتایج مدل پیشنهادی

در نمودار فوق مشاهده می‌شود که بهترین نتایج مدل در تعداد ۲۰۰۰ ویژگی حاصل شده است. این در شرایطی است که اگر کل ویژگی‌ها مجموعه داده یعنی ۱۷۰۰۰ ویژگی برای طبقه‌بندی در نظر گرفته شود دقت مدل به دلیل غیر متراکم و خلوت[۲۸] بودن داده‌ها به‌طور چشم‌گیری کاهش می‌یابد. از نتایج این بخش می‌توان دریافت که مرحله انتخاب ویژگی با استفاده از الگوریتم Info Gain در افزایش دقت مدل بسیار مؤثر است. لازم به ذکر است که کاهش تعداد ویژگی و ابعاد مسئله در زمان طبقه‌بندی مدل و همچنین حافظه مصرفی بسیار تأثیرگذار می‌باشد. با توجه به نتایج فوق در کلیه آزمایش‌های بعدی از تعداد ۲۰۰۰ ویژگی برای طبقه‌بندی و آموزش مدل استفاده شده است.

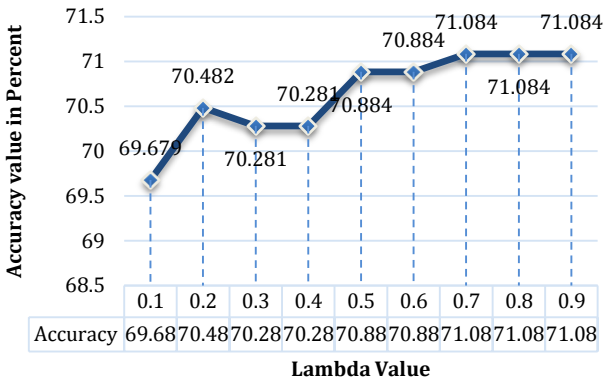
دسته دوم: بررسی تأثیر معیار وزن‌دهی بر نتایج

شکل ۱۲ نشان می‌دهد که بهترین مقدار برای متغیر λ برابر با ۰/۷ می‌باشد. این مقدار مشابه با مقاله چن [۵] حاصل گردید.

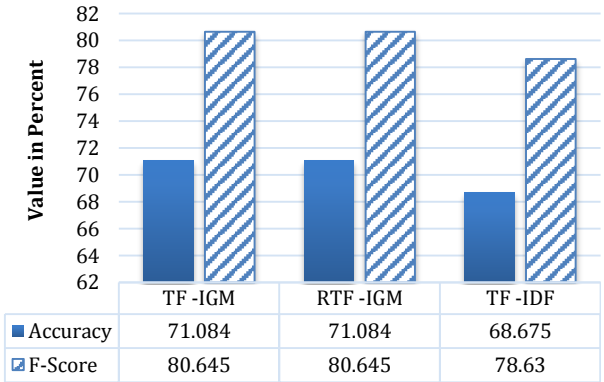
در این دسته از آزمایش‌ها میدان تصادفی شرطی با سه معیار وزن‌دهی مختلف ارزیابی شده است. همان‌طور که در نمودار فوق مشاهده می‌شود نتایج معیار وزن‌دهی TF-IGM و نسخه ریشه آن یعنی RTF-IGM دارای نتایج یکسانی در مسئله تحلیل احساس می‌باشد. با توجه به اینکه در مسئله تحلیل احساس هر جمله به‌عنوان یک

سند محسوب می‌شود، می‌توان گفت علت این امر به این دلیل است که در جملات تکرار کلمات بیش از یک بار تقریباً رخ نمی‌دهد، ازاین‌رو مقدار TF یک کلمه با مقدار ریشه آن یعنی $\sqrt{TF}$ برابر خواهد بود.

اختلاف دو معیار فوق در اسناد طولانی که تکرار کلمات در یک سند بیش از یک بار می‌باشد قابل مشاهده است، در مقاله چن و همکاران که از معیارهای وزن‌دهی فوق برای اسناد طولانی استفاده شده است، نتایج معیار ریشه، مطلوب‌تر از نسخه ساده آن گزارش شده است. نمودار فوق نشان می‌دهد که معیار وزن‌دهی TF-IGM باعث بهبود نتایج در مقایسه با معیار سنتی TF-IDF شده است. چرا که این معیار یک معیار با ناظر بوده و وزن‌دهی به کلمات را در کلاس‌های مختلف در نظر می‌گیرد درنتیجه کلماتی که قدرت تفکیک بیشتری برای تفکیک کلاس‌ها در مقایسه با کلمات دیگر داشته باشند وزن بیشتری دریافت می‌کنند. نتایج فوق نشان می‌دهد که استفاده از معیار TF-IGM و تغییر در مرحله وزن‌دهی به کلمات باعث بهبود ۲/۴٪ در نتایج دقت مدل و همچنین بهبود ۲/۰٪ در نتایج میانگین هارمونیک F شده است. پس از یافتن بهترین تعداد ویژگی و همچنین بهترین معیار وزن‌دهی برای میدان تصادفی شرطی در آخرین دسته از آزمایش‌ها نتایج میدان تصادفی شرطی با نتایج مدل پایه SVM-Gini مورد مقایسه قرار می‌گیرد.



شکل ۱۲- یافتن بهترین مقدار λ در معیار وزن دهی TF-IGM



شکل ۱۳- بررسی تأثیر معیار وزن دهی بر نتایج میدان تصادفی شرطی مخفی

دسته سوم: مقایسه نتایج مدل پیشنهادی و پایه

در آخرین دسته از آزمایش‌ها نتایج حاصل از مدل میدان تصادفی شرطی مخفی و مدل پایه SVM-Gini با یکدیگر مقایسه شده است. در این آزمایش از ۲۰۰۰ ویژگی با استفاده از الگوریتم Information Gain و همچنین معیار وزن‌دهی TF-IGM با مقدار λ برابر با ۰/۷ استفاده شده است. نتایج حاصل در شکل ۱۳ و ماتریس درهم‌ریختگی نیز در جدول‌های ۴ تا ۶ ارائه شده است.

جدول ۴- ماتریس درهم‌ریختگی الگوریتم پایه **SVM-Gini**

کلاس‌های حقیقی			
منفی	مثبت	نام کلاس	**کلاس‌های پیشگویی شده**
۱۳۱	۲۷۳	مثبت	
۵۲	۴۲	منفی	

الگوریتم‌های ترکیبی در طبقه‌بندی نتایج مطلوب‌تری در مقایسه با الگوریتم‌های ساده دارند. از ماتریس‌های درهم‌ریختگی فوق مشاهده می‌شود که مدل پیشنهادی در هر دو کلاس مثبت و منفی تعداد نمونه‌های بیشتری را به‌درستی در مقایسه با روش پایه پیشگویی کرده است. مدل پیشنهادی در کلاس مثبت از ۳۱۵ نمونه واقعی ۳۰۰ نمونه را به‌درستی به دسته مثبت تخصیص داده است. در کلاس منفی تعداد ۵۴ نمونه را از ۱۸۳ نمونه کلاس منفی حقیقی به‌درستی پیشگویی کرده است. در این حالت مدل پیشنهادی در کلاس مثبت ۲۷ نمونه و در کلاس منفی ۲ نمونه را بهتر پیشگویی کرده است. مدل پیشنهادی حتی در شرایطی که از معیار وزن‌دهی مشابه با مقاله پایه یعنی از معیار سنتی TF-IDF نیز استفاده کرده است باز هم نتایج مطلوب‌تری در مقایسه با مدل پایه ارائه داده است. میدان تصادفی شرطی مخفی با استفاده از معیار TF-IDF نیز در مقایسه با مدل پایه در هر دو کلاس مثبت و منفی پیشگویی‌های صحیح بیشتری داشته است. در این حالت در کلاس مثبت ۲۷۸ کلاس را از ۳۱۵ کلاس مثبت به‌درستی پیشگویی نموده است و در کلاس منفی ۵۵ نمونه از ۱۸۳ نمونه را به‌درستی طبقه‌بندی نموده است. در این حالت نیز مدل پیشنهادی به ترتیب در کلاس مثبت و منفی ۱۴ و ۳ نمونه داده را در مقایسه با مدل پایه بهتر پیشگویی کرده است.

نمودار فوق نیز نشان می‌دهد که مدل پیشنهادی در مقایسه با مدل پایه دارای نتایج بهتری در هر ۴ معیار مورد ارزیابی می‌باشد. عوامل مختلفی باعث حصول نتایج فوق و بهبود مدل پیشنهاد شده در مقایسه با مدل پایه در کلیه معیارها شده است.

استفاده از تکنیک کاهش ابعاد Information Gain ازیک‌طرف و همچنین استفاده از معیار وزن‌دهی با ناظر جدید TF-IGM از طرف دیگر و درنهایت استفاده از مدل ترکیبی میدان تصادفی شرطی مخفی که حاصل ترکیب مدل مخفی مارکوف و میدان تصادفی شرطی ساده می‌باشد، باعث بهبود نتایج شده است.

جدول ۵- ماتریس درهم‌ریختگی HCRF-TFIGM **HCRF-TFIGM**

کلاس‌های حقیقی			
منفی	مثبت	نام کلاس	**کلاس‌های پیشگویی شده**
۱۲۹	۳۰۰	مثبت	
۵۴	۱۵	منفی	

جدول ۶- ماتریس درهم‌ریختگی HCRF-TFIDF **HCRF-TFIDF**

کلاس‌های حقیقی		
منفی	مثبت	**نام کلاس**
۱۲۸	۲۸۷	**مثبت**
۵۵	۲۸	**منفی**

الگوریتم‌های ترکیبی در طبقه‌بندی نتایج مطلوب‌تری در مقایسه با الگوریتم‌های ساده دارند. از ماتریس‌های درهم‌ریختگی فوق مشاهده می‌شود که مدل پیشنهادی در هر دو کلاس مثبت و منفی تعداد نمونه‌های بیشتری را به‌درستی در مقایسه با روش پایه پیشگویی کرده است. مدل پیشنهادی در کلاس مثبت از ۳۱۵ نمونه واقعی ۳۰۰ نمونه را به‌درستی به دسته مثبت تخصیص داده است. در کلاس منفی تعداد ۵۴ نمونه را از ۱۸۳ نمونه کلاس منفی حقیقی به‌درستی پیشگویی کرده است. در این حالت مدل پیشنهادی در کلاس مثبت ۲۷ نمونه و در کلاس منفی ۲ نمونه را بهتر پیشگویی کرده است. مدل پیشنهادی حتی در شرایطی که از معیار وزن‌دهی مشابه با مقاله پایه یعنی از معیار سنتی TF-IDF نیز استفاده کرده است باز هم نتایج مطلوب‌تری در مقایسه با مدل پایه ارائه داده است. میدان تصادفی شرطی مخفی با استفاده از معیار TF-IDF نیز در مقایسه با مدل پایه در هر دو کلاس مثبت و منفی پیشگویی‌های صحیح بیشتری داشته است. در این حالت در کلاس مثبت ۲۷۸ کلاس را از ۳۱۵ کلاس مثبت به‌درستی پیشگویی نموده است و در کلاس منفی ۵۵ نمونه از ۱۸۳ نمونه را به‌درستی طبقه‌بندی نموده است. در این حالت نیز مدل پیشنهادی به ترتیب در کلاس مثبت و منفی ۱۴ و ۳ نمونه داده را در مقایسه با مدل پایه بهتر پیشگویی کرده است. نمودار فوق نیز نشان می‌دهد که مدل پیشنهادی در مقایسه با مدل پایه دارای نتایج بهتری در هر ۴ معیار مورد ارزیابی می‌باشد. عوامل مختلفی باعث حصول نتایج فوق و بهبود مدل پیشنهاد شده در مقایسه با مدل پایه در کلیه معیارها شده است. استفاده از تکنیک کاهش ابعاد Information Gain ازیک‌طرف و همچنین استفاده از معیار وزن‌دهی با ناظر جدید TF-IGM از طرف دیگر و درنهایت استفاده از مدل ترکیبی میدان تصادفی شرطی مخفی که حاصل ترکیب مدل مخفی مارکوف و میدان تصادفی شرطی ساده می‌باشد، باعث بهبود نتایج شده است.

نتایج مقایسه‌ای نشان می‌دهد که مدل‌های ترکیبی در مقایسه با مدل‌های ساده عملکرد بهتری از خود نشان می‌دهند. به‌طور خلاصه از نتایج فوق می‌توان به مقادیر زیر دست‌یافت:

- در معیار صحت، مدل پیشنهادی در مقایسه با مدل پایه با استفاده از وزن دهی TF-IGM دارای ۵/۸٪ بهبود می‌باشد. همچنین با استفاده از معیار TF-IDF در مقایسه با مدل پایه دارای بهبود ۳/۴٪ می‌باشد.
- در معیار دقت، مدل پیشنهادی در مقایسه با مدل پایه با استفاده از وزن دهی TF-IGM دارای۲/۳٪ بهبود می‌باشد. همچنین با استفاده از معیار TF-IDF در مقایسه با مدل پایه دارای بهبود ۱/۵٪ می‌باشد.
- در معیار فراخوان، مدل پیشنهادی در مقایسه با مدل پایه با استفاده از وزن دهی TF-IGM دارای ۸/۵٪ بهبود می‌باشد. همچنین با استفاده از معیار TF-IDF در مقایسه با مدل پایه دارای بهبود ۴/۴٪ می‌باشد.
- در معیار میانگین F، مدل پیشنهادی در مقایسه با مدل پایه با استفاده از وزن دهی TF-IGM دارای ۴/۷٪ بهبود می‌باشد. همچنین با استفاده از معیار TF-IDF در مقایسه با مدل پایه دارای بهبود ۲/۶٪ می‌باشد.

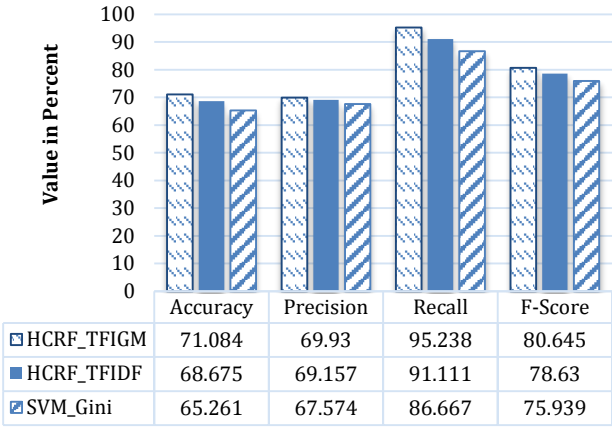
۵- بحث و بررسی

برای بررسی اجمالی نتایج حاصل شده در این بخش در شکل ۱۵ و جدول ۷ خلاصه‌ای از نتایج حاصل شده در این بخش ارائه شده است.

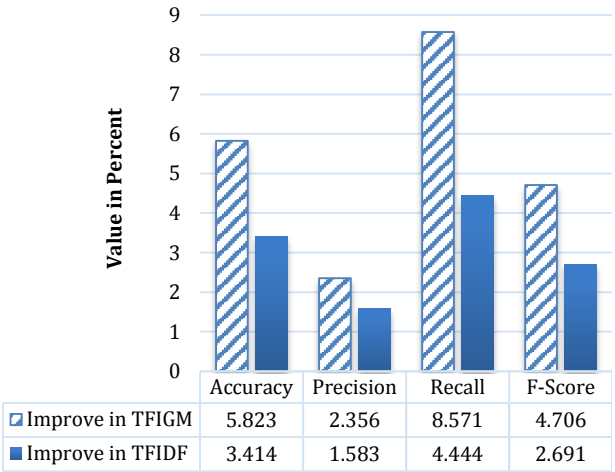
اولین و اصلی‌ترین هدف در این پژوهش ارائه یک مدل تحلیل احساس است که بتواند با موفقیت با استفاده از میدان تصادفی شرطی مخفی و معیار وزن‌دهی TF-IGM فرآیند تحلیل احساس را انجام دهد. همچنین استفاده از الگوریتم انتخاب ویژگی Info Gain برای کاهش ویژگی است. نتایج دسته اول در این بخش نیز نشان می‌دهد که انتخاب ویژگی و کاهش ابعاد با استفاده از الگوریتم Info Gain تأثیر بسیاری در دقت مدل داشته است. نتایج دقت مدل پیش از انتخاب ویژگی چیزی در حدود ۴۰٪ می‌باشد درحالی‌که این نتیجه پس از انتخاب ویژگی به ۷۱٪ افزایش یافته است. از این نتیجه می‌توان دریافت که دومین هدف در این پژوهش نیز به‌خوبی حاصل شده است. همچنین مدل پیشنهادی به ترتیب ۵/۸٪ و ۲/۳٪ و ۸/۵٪ بهبود در صحت و دقت و فراخوان داشته است.

جدول ۷- خلاصه کلیه نتایج حاصل

	HCRF_ TFIGM	HCRF_ TFIDF	SVM_ Gini	Improve in TFIGM	Improve in TFIDF
Accuracy	۷۱/۰۸۴	۶۸/۶۷۵	۶۵/۲۶۱	۵/۸۲۳	۳/۴۱۴
Precision	۶۹/۹۳	۶۹/۱۵۷	۶۷/۵۷۴	۲/۳۵۶	۱/۵۸۳
Recall	۹۵/۲۳۸	۹۱/۱۱۱	۸۶/۶۶۷	۸/۵۷۱	۴/۴۴۴
F-Score	۸۰/۶۴۵	۷۸/۶۳	۷۵/۹۳۹	۴/۷۰۶	۲/۶۹۱



شکل ۱۴- مقایسه نتایج مدل پیشنهاد شده و مدل پایه



شکل ۱۵- خلاصه بهبود مدل پیشنهادی در مقایسه با مدل پایه

در این پژوهش چهار معیار متداول صحت، دقت، فراخوان و میانگین F برای ارزیابی مدل ارائه شده استفاده گردید. نتایج حاصل شده نشان می‌دهد که در هر ۴ معیار نتایج روش ارائه شده در مقایسه با مدل پایه SVM-Gini بهتر می‌باشد.

حصول نتایج مطلوب دارای دلایل مختلفی است که در زیر به آن‌ها پرداخته می‌شود:

- اولین مرحله در فرآیند تحلیل احساس مرحله پیش پردازش است. در این مرحله از مدل پیشنهادی به دو روش نتایج بهبود داده شده است. اول اینکه در بحث حذف کلمات بازدارنده لیست‌هایی که عموماً از سایت‌های مختلف قابل دریافت می‌باشند دارای کلماتی هستند که نقش آن‌ها در تحلیل احساس بسیار پررنگ است ازاین‌رو در این پژوهش برای جلوگیری از حذف کلماتی که می‌تواند در تحلیل احساس مؤثر واقع گردد لیست کلمات بازدارنده بازبینی گردید و تعدادی از کلمات آن حذف شد. دومین و اصلی‌ترین مسئله که باعث بهبود نتایج گردید استفاده از معیار جدید وزن‌دهی با نام TF-IGM است. این معیار در روش پیشنهادی به‌جای معیار سنتی TF-IDF استفاده گردید و همان‌طور که در نتایج این بخش گزارش شد باعث بهبود نتایج مدل گردید. با توجه به اینکه این معیار با ناظر می‌باشد، تکرار و حضور کلمات را در هر کلاس به‌صورت مجزا در نظر می‌گیرد و در محاسبه وزن کلمات لحاظ می‌کند که این امر باعث می‌شود کلمات با قدرت تفکیک کننده بالاتر وزن بیشتری بگیرند.
- در فاز انتخاب ویژگی نیز استفاده از الگوریتم انتخاب ویژگی Information Gain که یک روش انتخاب ویژگی پالایش می‌باشد به‌طور چشمگیری باعث بهبود نتایج شد. نتایج طبقه‌بندی مدل قبل از انتخاب ویژگی در حدود ۴۰٪ و پس از انتخاب ۲۰۰۰ ویژگی برتر به ۷۱٪ افزایش یافت.

آخرین و اصلی‌ترین دلیل بهبود نتایج مدل ارائه شده در مقایسه با مدل پایه این است که مدل ارائه شده یک مدل ترکیبی می‌باشد، و نتایج حاصل شده در این پژوهش و دیگر پژوهش‌ها نشان می‌دهد که غالباً روش‌های یادگیری ترکیبی از روش‌های یادگیری ساده عملکرد بهتری دارند. در این پژوهش نیز برای طبقه‌بندی نظرات از ترکیب میدان تصادفی شرطی و مدل مخفی مارکوف استفاده گردید و این باعث شد که نتایج میدان تصادفی شرطی مخفی از نتایج ماشین بردار پشتیبان بهتر باشد.

۶- نتیجه‌گیری و پیشنهادهای آتی

خلاصه بهبود نتایج مدل میدان تصادفی شرطی مخفی در مقایسه با مدل پایه به شرح زیر است:

- در معیار صحت، مدل پیشنهادی در مقایسه با مدل پایه دارای ۵/۸٪ بهبود می‌باشد.
- در معیار دقت، مدل پیشنهادی در مقایسه با مدل پایه دارای ۲/۳٪ بهبود می‌باشد.
- در معیار فراخوان، مدل پیشنهادی در مقایسه با مدل پایه دارای ۸/۵٪ بهبود می‌باشد.
- در معیار میانگین F، مدل پیشنهادی در مقایسه با مدل پایه دارای ۴/۷٪ بهبود می‌باشد.

در راستای ادامه این پژوهش پیشنهادهایی به شرح زیر مطرح می‌شود:

۱) برای بهبود دقت مدل ارائه شده می‌توان از ترکیب الگوریتم BiLSTM[۲۹] و میدان تصادفی شرطی مخفی استفاده نمود، استفاده از تکنیک‌های شبکه عصبی عمیق می‌تواند در بهبود نتایج مؤثر واقع گردد.

۲) برای بهبود نتایج می‌توان به گونه دیگری از ترکیب مدل مارکوف و میدان تصادفی استفاده نمود. می‌توان هر دو را به‌صورت جداگانه آموزش داده و آزمایش نمود و نتایج پیشگویی حاصل را با استفاده از یکی از الگوریتم‌های فرا مکاشفه‌ای همچون ژنتیک ترکیب نمود. الگوریتم ژنتیک برای وزن‌دهی بین پاسخ‌های مشابه و غیرمشابه مؤثر می‌باشد.

۳) همچنین می‌توان مدل ارائه شده را در زبان‌های دیگر همچون زبان فارسی استفاده کرده و نتایج آن را بررسی نمود و یا از مدل‌های یادگیری توالی دیگر برای افزایش دقت طبقه‌بندی استفاده نمود.

۷- مراجع

[1] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, pp. 1093-1113, 2014.

[2] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," *Mining text data*, Springer, pp. 415-463, 2012.

[3] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: a survey," *International Journal*, vol. 2, pp. 282-292, 2012.

[4] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, pp. 14-46, 2015.

[5] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th international conference on World Wide Web*, pp. 519-528, 2003.

[6] B. Pang and L. Lee, *Opinion mining and sentiment analysis Foundations and Trends in Information Retrieval*, vol. 2, 2008.

[7] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!," *In Fifth International AAAI conference on weblogs and social media (Icwsm)*, vol. 11, pp. 538-541, 2011.

[8] W. Wei, "Analyzing text data for opinion mining," in *International Conference on Application of Natural Language to Information Systems*, pp. 330-335, 2011.

[9] A. Z. Khan, M. Atique, and V. Thakare, "Combining lexicon-based and learning-based methods for Twitter sentiment analysis," *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)*, 2011.

[10] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Systems with Applications*, vol. 66, pp. 245-260, 2016.

[11] A. S. Manek, P. D. Shenoy, M. C. Mohan, and K. Venugopal, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier," *World wide web*, vol. 20, pp. 135-154, 2017.

[12] P. D. Turney and M. L. Littman, "Unsupervised learning of semantic orientation from a hundred-billion-word corpus," *arXiv preprint cs/0212012*, 2002.

[13] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 2004, p. 271.

[14] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "Identifying sources of opinions with conditional random fields and extraction patterns," *in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 355-362, 2005.

[15] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational intelligence*, vol. 22, pp. 110-125, 2006.

[16] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," *In Proceedings of the 45th annual meeting of the association of computational linguistics (ACL)*, pp. 440-447, 2007.

[17] R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar, "Structured models for fine-to-coarse sentiment analysis," *in Annual Meeting-Association For Computational Linguistics*, p. 432-439, 2007.

[18] M. Sharifi and W. Cohen, "Finding domain specific polar words for sentiment classification," *in Language Technologies Institute Student Research Symposium*, 2008.

[19] T. Nakagawa, K. Inui, and S. Kurohashi, "Dependency tree-based sentiment classification using CRFs with hidden variables," *in Human Language Technologies: The 2010 Annual Conference of the*

[39] T. M. Mitchell, "Machine learning. 1997," *Burr Ridge, IL: McGraw Hill*, vol. 45, pp. 870-877, 1997.

[40] L. R. Welch, "Hidden Markov models and the Baum-Welch algorithm," *IEEE Information Theory Society Newsletter*, vol. 53, pp. 10-13, 2003.

[41] L. Moss, "Example of the Baum-Welch Algorithm," *Indiana University, Bloomington*, Spring, 2008.

[42] M. Riedmiller and H. Braun, "RPROP-A fast adaptive learning algorithm," *in Proc. of ISCIS VII)*, Universitat, 1992.

[43] M. Mahajan, A. Gunawardana, and A. Acero, "Training algorithms for hidden conditional random fields," *in Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 1, 2006.

[44] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," *in Neural Networks*, pp. 586-591, 1993.

[45] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, 2011.

[46] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," *in Proceedings of the 23rd international conference on Machine learning*, pp. 233-240, 2006.

[47] A Mosalanezhad and M. javad, "Provide an efficient rhythmic template for extracting semantic relationships in documentation, based on Wikipedia's tacit knowledge base." *Presented at the 23rd Iranian Conference on Electrical Engineering*, Sharif University of Technology, 1394.

[48] H Sadr, E Atani and M Yamghani, "Calculating the Semantic Relationship of Texts Using the Improvement of Explicitly Developed Semantic Analysis Algorithm", *The First National Conference on New Approaches in Computer Engineering and Information Retrieval*, 1392.

[49] J. Lafferty, A. McCallum, and F. C. Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, 2001.

مریم عموعلی دانشجوی کارشناسی ارشد مهندسی کامپیوتر گرایش نرم‌افزار در دانشگاه آزاد اسلامی واحد اصفهان (خوراسگان) می باشد. ایشان همچنین به‌عنوان کارشناس فناوری اطلاعات و کارمند رسمی اداره فرهنگ و ارشاد اسلامی استان اصفهان مشغول به کار می‌باشد.

آدرس پست الکترونیکی ایشان عبارت است از:

maryamamooali@gmail.com

فرساد زمانی، در سال ۱۳۹۲ با معدل ۱۹,۳۷ از دانشگاه یو پی ام مالزی فارغ‌التحصیل شد و در همان سال موفق به کسب مدال نقره در مسابقات کشوری دانشگاه‌های مالزی از وزارت علوم این کشور شد. او هم‌اکنون به‌عنوان استادیار تمام‌وقت و مدیرکل آموزش با دانشگاه آزاد اسلامی واحد اصفهان (خوراسگان) همکاری دارد.

آدرس پست الکترونیکی ایشان عبارت است از:

f.zamani@khuisf.ac.ir

[20] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," *in Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Vol. 10, pp. 79-86, 2002.

[21] M. Taboada and J. Grieve, "Analyzing appraisal automatically," *in Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Re# port SS# 04# 07)*, Stanford University, CA, pp. 158q161. AAAI Press, 2004.

[22] M. R. Saleh, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. Ureña-López, "Experiments with SVM to classify opinions in different domains," *Expert Systems with Applications*, vol. 38, pp. 14799-14804, 2011.

[23] Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, "Sentiment classification of Internet restaurant reviews written in *Cantonese," Expert Systems with Applications*, vol. 38, pp. 7674-7682, 2011.

[24] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Systems with Applications*, vol. 40, pp. 621-633, 2013.

[25] A. S. H. Basari, B. Hussin, I. G. P. Ananta, and J. Zeniarja, "Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization," *Procedia Engineering*, vol. 53, pp. 453-462, 2013.

[26] B. G. Patra, S. Mandal, D. Das, and S. Bandyopadhyay, "Ju_cse: A conditional random field (crf) based approach to aspect based sentiment analysis," *in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 370-374, 2014.

[27] M. Adnan and M. Rafi, "Document clustering with explicit semantic analysis (ESA)," *Journal of Independent Studies and Research*, vol. 12, p. 50, 2014.

[28] P. Kalaivani and K. Shunmuganathan, "Feature reduction based on genetic algorithm and hybrid model for opinion mining," *Scientific Programming*, vol. 2015, p. 12, 2015.

[29] K. Umamaheswari, S. Rajamohana, and G. Aishwaryalakshmi, "Opinion Mining using Hybrid Methods," *International Journal of Computer Application*, pp. 18-21, 2015.

[30] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks," *in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 959-962, 2015.

[31] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, "Sentiment analysis leveraging emotions and word embeddings," *Expert Systems with Applications*, vol. 69, pp. 214-224, 2017.

[32] V. Haralampieva and G. Brown, *Evaluation of Mutual information versus Gini index for stable feature selection*, 2016.

[33] A. G. Karegowda, A. Manjunath, and M. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *International Journal of Information Technology and Knowledge Management*, vol. 2, pp. 271-277, 2010.

[34] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Icml*, pp. 412-420, 1997.

[35] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Expert Systems with Applications*, vol. 72, pp. 221-230, 2017.

[36] H. Hamdan, P. Bellot, and F. Bechet, "Lsislif: Crf and logistic regression for opinion target extraction and sentiment polarity analysis," in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pp. 753-758, 2015.

[37] T. Alvarez-López, J. Juncal-Martinez, M. Fernández-Gavilanes, E. Costa-Montenegro, and F. J. González-Castano, "Gti at semeval-2016 task 5: Svm and crf for aspect detection and unsupervised aspect-based sentiment analysis," *in Proceedings of the 10th international workshop on semantic evaluation (SemEval)*, pp. 306-311, 2016.

[38] H. Xu, H. Lu, G. Yang, and C. Zhang, "Sentiment Analysis of Chinese Version Using SVM & RNN," *in Proceedings of the 6th International Conference on Information Engineering*, pp.1-5, 2017.

[1] Sentiment Analysis
[2] CRF: Conditional Random Field
[3] Term Frequency & Inverse Gravity Moment
[4] Naïve Bayes
[5] Maximum Entropy
[6] Tabu Search
[7] Markov-Blanket
[8] Acyclic
[9] Particle Swarm Optimization
[10] Explicit Semantic Analysis
[11] Logistic Regression
[12] Lower Case
[13] http://xpo6.com/download-stop-word-list/
[14] Term Frequency

[15] Inverse Document Frequency
[16] Document frequency
[17] Term Frequency & Inverse Gravity Moment
[18] Entropy
[19] Number of States
[20] Baum-Welch
[21] log-likelihood
[22] Tolerance
[23] gradient methods
[24] Confusion Matrix
[25] Stanford Twitter Sentiment
[26] http://help.sentiment140.com/for-students / Stanford link
[27] Comma Separated Values
[28] Sparse
[29] Bidirectional Long Short-Term Memory

The CSI Journal on Computing Science and Information Technology

The CSI Journal on
Computing Science and
Information Technology
Vol. 17, No.1, 2019

A Method for Sentiment Analysis Using TF-IGM and Conditional Random Field

Maryam Amooali, Farsad Zamani Boroujeni

Faculty of Engineering, Iran Islamic Azad University Isfahan (Khorasgan) Branch, Isfahan, Iran

Abstract

In recent years, the number of social media users has been exponentially rising, and at the same time, their comments on services and products have increased steadily. User feedback is a rich source of information that can be used to gain feedback on how to improve the quality of the products of different companies, as well as reviewing different opinions by users to make informed decisions about choosing the right service or products. Also helps. The numerous benefits of reviewing user feedback on the one hand and the impossibility of manually reviewing this volume of data on the other hand have made the field of sentiment analysis introduced. Emotional analysis is an area whose purpose is to identify emotions (positive, negative or neutral) of users based on their registered opinions. Most of the techniques that have been introduced in this area so far are based on data mining techniques, in particular the technique used for classifying. In these methods, a traditional TF-IDF classification and weighting algorithm is commonly used, which does not use the class information data in the weighting process to the words and does not involve this information in the weighting process, hence the results are sufficiently achieved not desirable. In this regard, a new weighting criterion TF-IGM is used for weighting words, which is a weighting criterion for observer. In addition, unlike the previous methods in this study, the combination of two methods of hidden Markov model and conditional random field, which is the result of combining these two random conditional random fields, is used for sentiment analysis. The results of the implementation of the proposed method on the Twitter account user's database, which contains 12,000 tweets, indicate that the proposed model is up to an improvement of 5.82% compared with the traditional TF-IDF method. In fact, the results show that the use of hybrid classification algorithms, along with the weighting method with observer, provides better results than simple methods.

Keywords: Sentiment Analysis, Conditional Random Field, Hidden Markov Model, Term Frequency and Inverse Gravity Moment (TF-IGM).