ساخت و بکارگیری یک پیکره برای تعیین بار نظرات غیرمستقیم

سمیرا نوفرستی[۱]، مهرنوش شمس فرد[۲]

[۱] دانشکده علوم و مهندسی کامپیوتر، دانشگاه شهیدبهشتی، تهران
snoferesti@ece.usb.ac.ir

[۲] دانشکده علوم و مهندسی کامپیوتر، دانشگاه شهیدبهشتی، تهران
m-shams@sbu.ac.ir

چکیده

نظرکاوی یکی از مسائل شناخته شده در حوزه پردازش زبان طبیعی است که در سال‌های اخیر بسیار مورد توجه قرار گرفته است. اگرچه تلاش‌های متعددی جهت تعیین بار نظرات در دامنه‌های مختلف انجام گرفته، تمرکز روش‌های موجود بر روی نظرات مستقیم بوده است و اغلب این روش‌ها از نظرات غیرمستقیم صرفنظر کرده‌اند. این در حالی است که در برخی از دامنه‌ها از جمله پزشکی نظرات غیرمستقیم به کرات رخ می‌دهند و نادیده گرفتن آنها باعث کاهش دقت سیستم نظرکاوی می‌شود. بنابراین ارائه روش‌های جدید به منظور مدیریت نظرات غیرمستقیم ضروری به نظر می‌رسد.

در این مقاله روشی خودکار برای ساخت یک پیکره از نظرات غیرمستقیم در دامنه دارو ارائه می‌شود. سپس از این پیکره در روش‌های یادگیری ماشین به منظور تعیین بار نظرات مطرح شده درباره داروها استفاده می‌گردد. نتایج آزمایشات انجام گرفته نشان می‌دهد که روش پیشنهادی در تعیین بار مجموعه تست به نتایج بهتری در مقایسه با یک روش برجسته موجود می‌رسد.

کلمات کلیدی

نظرکاوی، نظرات غیرمستقیم، تحلیل احساسات، ساخت خودکار پیکره، یادگیری ماشین، ویژگی‌های معنایی

۱- مقدمه

نظرکاوی[۱] که در برخی از تحقیقات تحلیل احساسات[۲] نیز نامیده می‌شود به ارائه روش‌های خودکار جهت کاوش، تحلیل، دسته‌بندی و خلاصه‌سازی نظرات می‌پردازد. نظرات به دو دسته کلی تقسیم می‌شوند [1]: نظرات مستقیم[۳] و نظرات غیرمستقیم[۴]. در یک نظر مستقیم، یک موجودیت یا جنبه‌ای از آن به صورت مستقیم مورد توصیف قرار می‌گیرد. به عنوان مثال جمله "این دارو بسیار مفید است." یک نظر مستقیم با بار مثبت را نشان می‌دهد. در مقابل در یک نظر غیرمستقیم اثر یک موجودیت بر روی موجودیت‌های دیگر بیان می‌شود. به عنوان مثال جمله "پس از مصرف این دارو دچار درد مفاصل شدم." یک نظر غیر مستقیم است که بیانگر اثری منفی از دارو بر روی مفاصل است.

یکی از وظایف اصلی در نظرکاوی تعیین بار[۵] یک نظر است. هدف این وظیفه دسته‌بندی نظرات در گروه‌های از پیش تعیین شده (اغلب در دو گروه مثبت و منفی) است. تمرکز روش‌های موجود نظرکاوی بر روی نظرات مستقیم بوده است. این در حالی است که در برخی از دامنه‌ها نظیر اقتصاد و پزشکی نظرات غیرمستقیم به کرات رخ می‌دهند. به طور خاص در دامنه دارو، اغلب کاربران بجای بیان یک نظر مستقیم درباره دارو به توصیف اثرات مثبت و عوارضی که دارو برای آن‌ها داشته است می‌پردازند. بررسی‌های ما بر روی سایت *www.druglib.com* که یک سایت نظرسنجی محبوب در رابطه با داروها است، نشان می‌دهد که بیش از نیمی از نظرات، غیرمستقیم هستند. بنابراین روش‌های موجود تحلیل نظرات که تنها نظرات مستقیم را مورد توجه قرار می‌دهند، بخش عمده‌ای از اطلاعات مفید در نظرات کاربران را نادیده

می‌گیرند که این امر خود سبب کاهش دقت الگوریتم نظرکاوی می‌شود. بنا به دلایل مذکور این مقاله به مسأله تعیین بار نظرات غیرمستقیم پرداخته است.

به منظور تعیین بار نظرات غیرمستقیم از روش‌های یادگیری ماشین استفاده شده است. پیش از این روش‌های یادگیری ماشین در تحلیل نظرات مستقیم بکار گرفته شده‌اند و به نتایج خوبی دست یافته‌اند. مشکل اصلی این روش‌ها ساخت یک مجموعه آموزش از نظرات است که بار مثبت و منفی آن‌ها برچسب خورده باشد. ساخت دستی مجموعه آموزش امری زمان‌بر و هزینه‌بر است. به منظور رفع این مشکل در این مقاله روشی برای ساخت خودکار مجموعه آموزش پیشنهاد می‌شود که از روش‌های نظارت از راه دور[6] [2] الهام گرفته است. در این روش‌ها بجای بکارگیری یک مجموعه آموزش دقیق، از یک مجموعه بزرگ از داده‌های نویزی که به صورت خودکار بدست آمده‌اند، استفاده می‌شود.

نوآوری‌های این مقاله را می‌توان به صورت زیر خلاصه کرد:

- ارائه روشی نوین برای ساخت خودکار یک مجموعه برچسب خورده از نظرات غیرمستقیم
- معرفی، بکارگیری و ارزیابی ویژگی‌های لغوی، نحوی و معنایی مختلف در روش‌های یادگیری ماشین به منظور تعیین بار نظرات غیرمستقیم

ادامه این مقاله به صورت زیر سازمان‌دهی شده است. در بخش ۲ کارهای مرتبط انجام گرفته در زمینه تعیین بار نظرات مرور و مشکلات آنها در تعیین بار نظرات غیرمستقیم بررسی می‌شود. سپس در بخش‌های ۳ و ۴ جزئیات روش پیشنهادی برای ساخت خودکار پیکره‌ای برچسب خورده از نظرات غیرمستقیم و بکارگیری آن به منظور تعیین بار نظرات جدید شرح داده می‌شود. بخش ۵ نتایج آزمایشات انجام گرفته به منظور ارزیابی روش پیشنهادی را ارائه می‌کند. در پایان بخش ۶ به نتیجه‌گیری می‌پردازد.

۲- مرور کارهای پیشین

کارهای انجام گرفته در زمینه تعیین بار نظرات را می‌توان به دو رده کلی تقسیم کرد: روش‌های مبتنی بر واژگان و روش‌های یادگیری ماشین.

تمرکز روش‌های مبتنی بر واژگان بر ساخت و بکارگیری یک واژگان سنجمانی[7] عمومی [3,4] یا خاص دامنه [5,6] بوده است. یک واژگان سنجمانی حاوی مجموعه‌ای از لغات و عبارات با بار سنجمانی مشخص است. این روش‌ها در تحلیل نظرات غیرمستقیم از دو مشکل اصلی رنج می‌برند. مشکل اول این است که در دامنه پزشکی اغلب واژه‌ها مانند "بیماری"، "آلزایمر"، "تب" و "درد" منفی هستند. با این وجود، این واژه‌ها به مراتب در جملات مثبت رخ می‌دهند. به مثال‌های زیر توجه کنید[8]:

"This drug eliminated my acne completely."
"This drug reduced my pain."

روش‌های مبتنی بر واژگان از تعیین بار درست این گونه مثال‌ها قاصرند زیرا واژه‌هایی نظیر "آکنه" و "درد" در واژگان‌های موجود مانند SentiWordNet [3] منفی هستند. این امر باعث می‌شود دو جمله فوق منفی در نظر گرفته شوند. در حالی که بار کلی هر دو جمله مثبت است.

مشکل دوم در رابطه با جملاتی است که حاوی واژه سنجمانی صریح نیستند. به مثال زیر توجه کنید:

"This drug decreased my vision."

در SentiWordNet واژه "بینایی" بدون بار است. این امر باعث می‌شود جمله منفی فوق بدون بار در نظر گرفته شود.

روش‌های یادگیری ماشین با بکارگیری ویژگی‌های مختلف از جمله کیسه لغات[9] [7]، ترکیب دوتایی لغات[10] [7]، برچسب‌های نحوی[11] [8]، شکلک‌ها[12] [9] و مسیر بین واژه‌ها در درخت وابستگی [10]، به نتایج قابل قبولی در تعیین بار نظرات مستقیم دست یافته‌اند. همچنین این روش‌ها با در نظر گرفتن اطلاعات زمینه و روابط بین واژه‌ها تا حدودی قادر به تعیین بار نظرات ضمنی (فاقد واژه سنجمانی) هستند. با این وجود این روش‌ها از مشکل ساخت پیکره آموزش رنج می‌برند. اغلب کارهای پیشین پیکره آموزش را به صورت دستی ساخته‌اند [11,12] که امری زمان‌بر و هزینه‌بر است. بدین دلیل تلاش‌هایی جهت ساخت خودکار پیکره آموزش انجام شده است که عمدتاً مبتنی بر شکلک‌های مورد استفاده در نظرات هستند [8].

اغلب کارهای پیشین تنها نظرات مستقیم را مورد توجه قرار داده‌اند و روش خاصی برای مدیریت نظرات غیرمستقیم ندارند. تعداد معدودی از کارهای پیشین به نظرات غیرمستقیم توجه کرده‌اند. در [13] یک پیکره از نظرات مستقیم و غیرمستقیم در دامنه اخبار به صورت دستی برچسب زده شده است. در [14,15] روشی برای تعیین بار نظرات غیرمستقیم در دامنه اقتصاد معرفی شده است که در ابتدا با استفاده از یک فرهنگ لغت خاص دامنه اصطلاحات تخصصی متن را استخراج می‌کند. شاخص‌های اقتصادی به دو گروه مثبت و منفی تقسیم شده‌اند. همچنین تعدادی اصلاح کننده مثبت و منفی تعریف شده است. سپس با بکارگیری یک فرمول ساده بار شاخص در بار اصلاح‌کننده آن ضرب شده و بار کل نظر بدست می‌آید.

در این مقاله روشی جدید برای ساخت خودکار یک پیکره از نظرات غیرمستقیم ارائه می‌شود و سپس پیکره ساخته شده برای تعیین بار نظرات غیرمستقیم مورد استفاده قرار می‌گیرد.

۳- ساخت یک پیکره از نظرات غیرمستقیم

در این بخش در ابتدا مجموعه داده‌ای مورد استفاده برای ساخت پیکره‌ای از نظرات غیرمستقیم درباره داروها توصیف می‌گردد. سپس جزئیات روش پیشنهادی برای ساخت این پیکره شرح داده می‌شود.

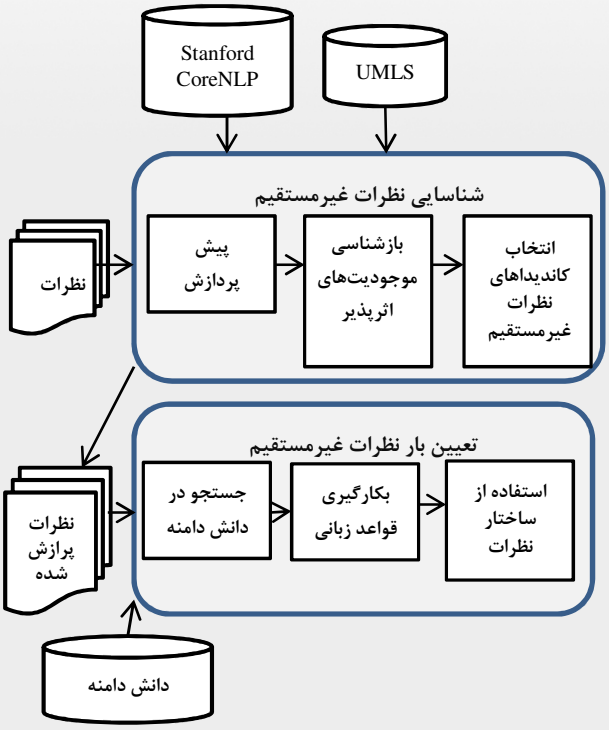
۳-۱- مجموعه داده‌ای

مجموعه داده‌ای مورد استفاده جهت ساخت پیکره از سایت *www.druglib.com* جمع‌آوری شده است. بدین منظور نظرات مطرح شده درباره ۸۳ دارو گردآوری شده است. این مجموعه دارای ۱۳۱۳ نظر و ۷۰۱۷ جمله است. متن هر نظر شامل سه بخش است: فواید، عوارض دارویی و توضیحات. پیش‌فرض این است که کاربر در بخش فواید درباره اثرات مثبت دارو و در بخش عوارض درباره اثرات منفی دارو بنویسد. اگر چه این فرض الزاماً همیشه درست نیست. جملات بخش توضیحات نیز محدودیتی ندارند.

۳-۲- ساخت پیکره

به منظور ساخت پیکره یک روش دو مرحله‌ای معرفی می‌شود. در مرحله اول جملات بیانگر یک نظر غیرمستقیم شناسایی می‌شوند و در مرحله دوم بار

مثبت و منفی آن‌ها تعیین می‌شود. نمای کلی روش پیشنهادی در شکل (۱) نشان داده شده است. در ادامه مراحل کار با جزئیات تشریح می‌گردد.



شکل (۱): نمای روش پیشنهادی برای ساخت پیکره

۳-۲-۱- شناسایی نظرات غیرمستقیم

نظر غیرمستقیم اثر یک موجودیت (اثرگذار) را بر موجودیتی دیگر (اثرپذیر) بیان می‌کند. در ابتدا فرض شد هر جمله‌ای که شامل یک موجودیت اثرپذیر و یک موجودیت اثرگذار است یک نظر غیرمستقیم را نشان می‌دهد. اگر چه این روش دقت خوبی در شناسایی نظرات غیرمستقیم دارد (جملات استخراج شده با این روش اغلب به درستی یک نظر غیرمستقیم را نشان می‌دهند) اما بسیاری از نظرات غیرمستقیم را استخراج نمی‌کند. علت این امر این است که اغلب نظرات غیرمستقیم فاقد موجودیت اثرگذار هستند. از آنجا که کاربران سایت‌های نظرسنجی اغلب درباره یک موجودیت مشخص (موجودیتی که نام آن در عنوان پست ذکر شده است) نظر می‌دهند، بندرت نام آن موجودیت را در متن نظر خود ذکر می‌کنند؛ بلکه تنها به بیان اثر آن موجودیت می‌پردازند. بدین دلیل فرض شده است هر جمله حاوی یک موجودیت اثرپذیر یک نظر غیرمستقیم را نشان می‌دهد. با این فرض به منظور شناسایی نظرات غیرمستقیم گام‌های زیر طی می‌شود (شکل ۱).

پیش‌پردازش: برای هر نظر در مجموعه نظرات، ابتدا با توجه به ساختار نظر تنها جملاتی که در بخش فواید یا عوارض مطرح شده‌اند استخراج می‌شوند. سپس با بکارگیری Stanford CoreNLP[۱۳] جمله‌ها از یکدیگر جدا شده و هر جمله مرکب به تعدادی واحد کوچکتر شکسته می‌شود. در پایان برای هر جمله قطعه‌بندی[۱۴]، لم‌یابی[۱۵]، برچسب‌زنی نحوی[۱۶] و تشخیص مرجع ضمایر انجام می‌گیرد [16].

بازشناسی موجودیت‌های اثرپذیر: همان‌طور که گفته شد، فرض شده است هر جمله دارای یک موجودیت اثرپذیر یک نظر غیرمستقیم است. بنابراین به منظور شناسایی نظرات غیرمستقیم در ابتدا نیاز است انواع ممکن برای موجودیت‌های اثرپذیر تعیین شود. بدین منظور یک روش خودکار پیشنهاد شده است که نوع موجودیت‌های پررخداد را به عنوان کاندیدا در نظر می‌گیرد. دلیل درست بودن این ایده این است که معمولاً موجودیت‌هایی که اغلب کاربران درباره آن نظر می‌نویسند موجودیت‌های مهم هستند. برای استخراج موجودیت‌های پررخداد، ۲۰۰ نظر از سایت *druglib* به عنوان مجموعه توسعه[۱۷] انتخاب شده است. برای هر جمله در این مجموعه مفاهیم پزشکی با کمک نرم‌افزار MetaMap[۱۸] [17] برچسب زده شده‌اند. این نرم‌افزار مفاهیم جمله را به مفاهیم تعریف شده در UMLS[۱۹] نگاشت می‌کند. UMLS یک فرهنگ جامع شامل ۱.۷ میلیون مفهوم است. این مفاهیم در ۱۳۰ نوع معنایی[۲۰] و ۱۵ گروه معنایی[۲۱] دسته‌بندی شده‌اند.

مشکل در این است که UMLS علاوه بر مفاهیم پزشکی مفاهیم عمومی نظیر زمان را نیز دارا است. برای حل این مشکل باید مفاهیم مهم خاص دامنه انتخاب و مفاهیم عمومی حذف شوند. مفاهیم عمومی بر خلاف مفاهیم خاص دامنه در اغلب دامنه‌ها پررخداد هستند. بنابراین می‌توان مفاهیمی را که در دامنه‌های مختلف پررخداد هستند به عنوان مفاهیم عمومی در نظر گرفت و آن‌ها را حذف کرد. بدین منظور معیاری با نام STF-IOF[۲۲] (تعداد رخداد نوع معنایی–معکوس تعداد رخداد نظر) تعریف شده که اصلاح شده معیار CF-IOF است [18].

$$STF\_IOF_{sg,d} = \frac{n_{st,d}}{\sum_k n_{k,d}} log \sum_k \alpha . \frac{n_k}{n_{st}} \qquad (۱)$$

$n_{st,d}$ تعداد رخداد نوع معنایی st در مجموعه نظرات دارویی d است. n_k تعداد کل رخدادهای انواع معنایی و n_{st} تعداد رخداد نوع معنایی st در کل مجموعه نظرات است. α نیز یک پارامتر ثابت است که اهمیت نسبی STF در مقابل IOF را نشان می‌دهد. در این مقاله مقدار α، ۰/۱ در نظر گرفته شده است.

برای استفاده از این معیار نیاز به یک مجموعه از نظرات است که بخشی از نظرات (به نام d) مربوط به دامنه مدنظر و بخش دیگر شامل نظرات در دیگر دامنه‌ها باشد. یک مجموعه از نظرات از سه دامنه دارو، رستوران[۲۳] و محصولات [11] گردآوری شده است. مجموعه نظرات در دامنه داروها همان مجموعه توسعه است که پیش‌تر توصیف شد. با اعمال معیار STF_IOF مفاهیم با اهمیت دامنه دارو (مفاهیمی که مقدار STF-IOF آنها از یک آستانه از پیش تعیین شده بیشتر است) استخراج شده‌اند.

انتخاب کاندیداها: پس از تعیین انواع ممکن برای موجودیت‌های اثرپذیر هر جمله‌ای که دارای حداقل یک موجودیت اثرپذیر باشد به عنوان کاندیدا برای نظر غیرمستقیم در نظر گرفته می‌شود.

۳-۲-۲- تعیین بار نظرات غیرمستقیم

به منظور تعیین بار نظرات از دانش دامنه، الگوهای زبانی و ساختار نظرات استفاده شده است. در ادامه به توصیف هر کدام از این موارد می‌پردازیم.

دانش دامنه: بسیاری از فواید و عوارض دارویی شناخته شده هستند. به عنوان مثال "کاهش درد" از فواید شناخته شده "استامینوفن" و "خشکی لب" از عوارض شناخته شده "آکوتان" است. در این مقاله دانش دامنه که شامل فواید و عوارض شناخته شده داروهای مختلف است از سایت *www.dailymed.com* استخراج شده است. اگر در یک نظر غیر مستقیم درباره فایده‌ای شناخته شده از دارو صحبت شده باشد بار آن نظر را مثبت در

نظر می‌گیریم. در مقابل اگر نظر عارضه‌ای شناخته شده از دارو را بیان کند بار آن منفی در نظر گرفته می‌شود. البته برای تعیین بار نظرات منفی‌کننده‌ها را نیز مد نظر قرار داده‌ایم. منفی‌کننده‌ها واژه‌هایی مانند "no" و "never" هستند که بار یک جمله را معکوس می‌کنند.

ساختار نظرات: بسیاری از نظرات کاربران بیانگر اثراتی هستند که در دانش دامنه به وضوح و با آن الفاظ دیده نمی‌شوند. برای تعیین بار این‌گونه نظرات از ساختار نظرات استفاده شده است. در ابتدا فرض شد جملات بخش فواید مثبت و جملات بخش عوارض جانبی منفی هستند. با این وجود این فرض در برخی از موارد درست نیست. به منظور رفع این مشکل به هر اثر یک درجه اطمینان نسبت داده می‌شود که از رابطه زیر محاسبه می‌گردد:

$$confidence(e) = \frac{|PF-NF|}{PF+NF} \quad (۲)$$

متغیر e یک اثر را نشان می‌دهد. PF و NF نیز به ترتیب تعداد رخداد آن اثر در بخش فواید و عوارض هستند. هرچه درجه اطمینان یک اثر به یک نزدیک‌تر باشد میزان اطمینان ما به مثبت یا منفی بودن آن اثر بیشتر است.

به منظور استفاده از رابطه (۲) به این صورت عمل شده است که در مجموعه جملاتی که به عنوان نظر غیرمستقیم شناسایی شده‌اند اگر نظر حاوی یک فعل باشد دوتایی (ریشه فعل، نوع معنایی موجودیت اثرپذیر) و اگر نظر یک عبارت بدون فعل باشد دوتایی (اصلاح کننده، نوع معنایی موجودیت اثرپذیر) به عنوان اثر استخراج می‌شود که البته اصلاح کننده می‌تواند مقدار نداشته باشد. مولفه دوم یک دوتایی نوع معنایی موجودیت اثرپذیر است که از UMLS بدست می‌آید. نوع معنایی در واقع کلاس معنایی یک موجودیت را نشان می‌دهد. سپس برای هر دوتایی درجه اطمینان محاسبه شده است. تنها دوتایی‌هایی که درجه اطمینان آن‌ها از یک آستانه از پیش تعیین شده (در این مقاله ۰.۵) بیشتر باشد انتخاب می‌شوند. درجه اطمینان را تنها برای دوتایی‌هایی که حداقل سه بار در مجموعه نظرات رخ داده‌اند محاسبه می‌کنیم و از سایر دوتایی‌ها صرفنظر می‌شود. جدول (۱) مثال‌هایی از نظرات غیرمستقیم و دوتایی‌های استخراج شده از آنها را نشان می‌دهد.

جدول (۱) : مثال‌هایی از دوتایی‌های استخراج شده

نظر غیرمستقیم	دوتایی استخراج شده
I experienced severe headache.	(experience, sosy)
Accutane cleared my acne.	(clear, dsyn)
Dry lips	(null, fndg)
Reduction of acne	(reduce, dsyn)

در پایان برای هر دوتایی انتخاب شده اگر تعداد رخداد آن در بخش فواید بیشتر از تعداد رخداد آن در بخش عوارض باشد بار آن را مثبت و در غیر این صورت بار آن را منفی در نظر می‌گیریم. در این مرحله نیز منفی‌کننده‌ها مد نظر قرار گرفته‌اند. به عنوان مثال اگرچه دوتایی (clear,dsyn) دارای بار مثبت است اما اگر این دوتایی در جمله‌ای که حاوی یک منفی‌کننده است ظاهر شود بار جمله منفی خواهد بود.

به منظور استخراج دوتایی (ریشه فعل، نوع معنایی موجودیت اثرپذیر) از یک روش فعل-محور که به منظور استخراج رابطه از متون پزشکی معرفی شده [19] استفاده شده است. در این روش ابتدا فعل اصلی تعیین می‌شود. سپس با استفاده از درخت وابستگی جمله اگر موجودیت اثرپذیر با فعل جمله رابطه‌ای نظیر مفعول، فاعل و اصلاحگر قیدی[۲۴] داشته باشد دوتایی مذکور استخراج می‌شود. مولفه اول این دوتایی ریشه فعل است که از ریشه‌یاب موجود در Stanford CoreNLP بدست می‌آید و مولفه دوم نوع معنایی موجودیت اثرپذیر است که توسط MetaMap تعیین می‌شود.

به طور مشابه برای استخراج دوتایی (اصلاح کننده، نوع معنایی موجودیت اثرپذیر) از عبارات فاقد فعل، در ابتدا موجودیت اثرپذیر مشخص می‌شود. سپس با اعمال تعدادی قاعده زبانی بر روی درخت وابستگی، اصلاح کننده عبارت اسمی بیانگر موجودیت اثرپذیر بدست می‌آید. ریشه اصلاح کننده با استفاده از تابع getDerivationallyRelatedForms در JAWS[۲۵] و نیز ریشه‌یاب Staford بدست می‌آید.

قواعد زبانی: در این مرحله بار برخی از نظرات که تابحال برچسب بار نخورده‌اند با استفاده از قواعد زبانی تعیین می‌شود. در این مقاله دو قاعده زبانی مورد استفاده قرار گرفته است [1]:

- قاعده "و": جملات و عباراتی که با حروف ربط مشابه "و" به یکدیگر متصل می‌شوند دارای بار یکسانی هستند.
- قاعده "اما": جملات و عباراتی که با حروف ربط مشابه "اما" مرتبط می‌شود اغلب دارای بار متضاد هستند.

با بکارگیری دو قاعده فوق اگر بار یک جمله مشخص باشد (در گام‌های پیشین بار آن تعیین شده باشد) می‌توان بار جمله دیگر را تعیین کرد. در پایان از نظراتی که بار آنها نامشخص باقی‌مانده است صرفنظر می‌کنیم.

٤- بکارگیری پیکره برای تعیین بار نظرات

برای تعیین بار نظرات غیرمستقیم از پیکره ساخته شده به عنوان مجموعه آموزش در روش‌های یادگیری ماشین استفاده شده است. بدین منظور در ابتدا سه پیش‌پرازش رایج به نام‌های حذف هرزواژه‌ها، تبدیل متن به حروف کوچک و ریشه‌یابی اجرا شده است. سپس از ویژگی‌های رایج در نظرکاوی یعنی تک‌واژه‌ها، زوج‌واژه‌ها و برچسب نحوی در کنار تعدادی ویژگی نحوی و معنایی برای دسته‌بندی نظرات استفاده شده است. این ویژگی‌ها، نماد مورد استفاده برای آنها و توصیف هر کدام در جدول (۲) آورده شده است.

جدول (۲): ویژگی‌های مورد استفاده در یادگیری ماشین

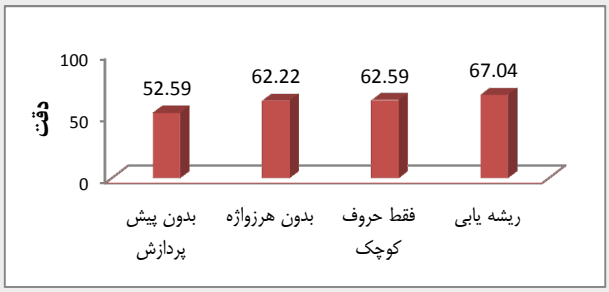
نماد	ویژگی	توصیف
F1	unigram	تک واژه‌ها
F2	bigram	زوج واژه‌ها
F3	POS	برچسب نحوی
F4	mainVerb	فعلی که اثر بیان شده بر روی موجودیت اثرپذیر را بیان می‌کند.
F5	affected	چانک حاوی موجودیت اثرپذیر
F6	type	نوع معنایی موجودیت اثرپذیر
F7	group	گروه معنایی موجودیت اثرپذیر
F8	negation	آیا جمله دارای منفی کننده است؟

٥- ارزیابی روش پیشنهادی

تا آنجا که نگارندگان این مقاله می‌دانند مجموعه تستی عمومی برای مسأله تعیین بار نظرات غیرمستقیم در دامنه دارو وجود ندارد. به هیمن دلیل اقدام به ساخت یک مجموعه دستی از نظرات غیرمستقیم شده است. بدین منظور بخشی از نظرات مطرح شده درباره ۱۹ دارو گردآوری شده است. سپس از دو شخص خواسته شده است که نظرات غیرمستقیم این مجموعه را استخراج کرده و به آنها برچسب بار اختصاص دهند. تنها جملاتی که هر دو شخص به

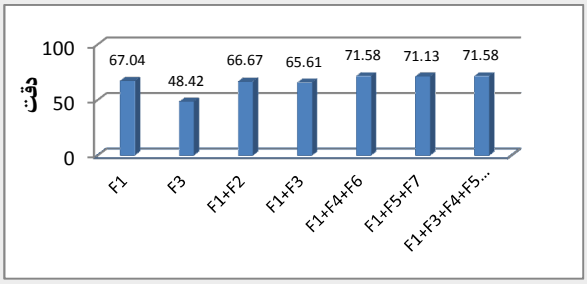
آنها بار یکسانی داده‌اند، در مجموعه تست قرار گرفته‌اند. بدین ترتیب مجموعه‌ای از ۲۸۵ نظرغیرمستقیم ساخته شده است. سپس به منظور ارزیابی کارایی روش پیشنهادی تعدادی آزمایش طراحی شده است. در این آزمایشات از نرم‌افزار وکا[۲۶] با مقادیر پیش‌فرض پارامترها برای شبیه‌سازی الگوریتم‌های یادگیری ماشین استفاده شده است.

در آزمایش اول اثر هر یک از پیش‌پردازش‌های معرفی شده در بخش ۴ بررسی شده است (شکل ۲). در این آزمایش تنها از ویژگی تک‌واژه‌ها و از دسته‌بند بیزین[۲۷] برای تعیین بار نظرات مجموعه تست استفاده شده است. ابتدا تعیین بار بدون هیچ پیش‌پردازشی انجام شده است. سپس به ترتیب پردازش‌های حذف هرزواژه‌ها، تبدیل به حروف کوچک و ریشه‌یابی اضافه شده است. همان‌طور که در شکل (۲) دیده می‌شود بکارگیری این سه پیش‌پردازش دقت الگوریتم تعیین بار را به میزان ۱۴/۵ درصد افزایش می‌دهد. دقت عبارت است از درصدی از مثال‌های مجموعه تست که به درستی دسته‌بندی شده‌اند.



شکل (۲): اثر پیش‌پردازش‌های انجام گرفته بر روی دقت دسته‌بندی

در آزمایش بعدی اثر ویژگی‌های مختلف بررسی شده است (شکل ۳). دو حالت پایه در نظر گرفته شده است: زمانی که تنها از تک‌واژه‌ها استفاده می‌شود و زمانی که تنها از برچسب نحوی استفاده می‌شود. همچنین ترکیب تک‌واژه‌ها با برچسب نحوی و با زوج‌واژه‌ها که در نظرکاوی پرکاربرد است نیز استفاده شده است. علاوه بر آن ترکیب تک‌واژه‌ها با ویژگی‌های معنایی معرفی شده نیز بکار گرفته شده است. همان‌طور که در شکل (۳) مشاهده می‌شود بکارگیری ویژگی‌های معنایی تاثیر بسزایی در افزایش دقت دسته‌بند بیزین در مقایسه با حالات پایه دارد.



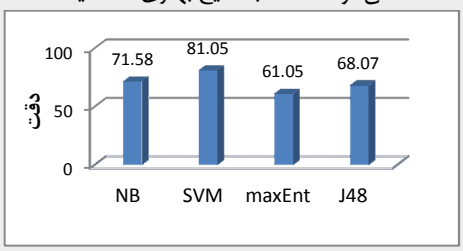
شکل (۳): اثر بکارگیری ویژگی‌های معرفی شده در دسته‌بند بیزین

آزمایش فوق با دسته‌بند ماشین بردار پشیبان (SVM)[۲۸] هم تکرار شده است. در این آزمایش نیز ترکیب ویژگی‌های معنایی دقت دسته‌بند را از ۷۸/۸۹ برای حالت پایه تک‌واژه‌ها به ۸۱/۰۵ افزایش داده است (جدول ۳).

جدول (۳) : اثر بکارگیری ویژگی‌های معرفی شده در دسته‌بند SVM

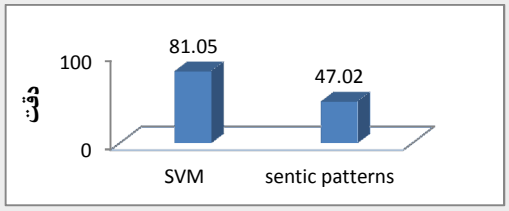
نظر غیرمستقیم	ویژگی
۷۸/۸۹	F1
۸۱/۰۵	F1,F3,F4,F5,F6,F7,F8

در آزمایش بعدی دقت دسته‌بندهای بیزین، ماشین بردار پشتیبان، J48 و بیشینه آنتروپی[۲۹] با یکدیگر مقایسه شده است (شکل ۴). دلیل انتخاب این دسته‌بندها استفاده مکرر آنها در تعیین بار نظرات مستقیم و رضایت بخش بودن نتایج حاصل از آنها است. این دسته‌بندها با استفاده از ترکیب همه ویژگی‌های نامبرده در جدول (۲) بجز ویژگی زوج‌واژه‌ها اجرا شده‌اند. همان‌طور که مشاهده می‌شود SVM به نتایج بهتری دست یافته است.



شکل (٤): مقایسه دقت دسته‌بندهای مختلف

در آزمایش پایانی نتایج بدست آمده توسط دسته‌بند SVM با نتایج حاصل از یکی از روش‌های برجسته موجود به نام sentic patterns [20] مقایسه شده است. این روش که متکی برقواعد زبانی، محاسبات عرفی و یادگیری ماشین است، از SenticNet[۳۰] که یک پایگاه دانش از مفاهیم عرفی دارای بار است استفاده می‌کند. اگرچه در SenticNet بسیاری از مفاهیم رایج در دامنه پزشکی مانند "آکنه"، "معده درد"، "افزایش وزن" و "کاهش استرس" وجود دارند که می‌توانند به درستی برخی از نظرات غیرمستقیم را تعیین بار کنند، با این وجود این روش در تعیین بار اغلب نظرات غیرمستقیم با شکست مواجه می‌شود. دلیل اصلی این امر نیز این است که sentic patterns در اغلب موارد اثر فعل جمله را نادیده می‌گیرد. در شکل (۵) مشاهده می‌شود که روش پیشنهادی در مقایسه با sentic patterns به نتایج بهتری دست یافته است. همچنین این شکل نشان می‌دهد روش‌های موجود نظرکاوی در زمینه تعیین بار نظرات غیرمستقیم ضعیف عمل می‌کنند. به بیانی دیگر مسأله تعیین بار نظرات غیرمستقیم نیازمندی‌های خاص خود را دارد و برای رسیدن به نتایج قابل قبول باید این نیازمندی‌ها مورد توجه قرار گیرند.



شکل (۵): مقایسه روش پیشنهادی با sentic patterns

٦- نتیجه‌گیری

در این مقاله در ابتدا روشی برای ساخت خودکار یک پیکره از نظرات غیرمستقیم ارائه شد. با بکارگیری این روش یک پیکره شامل ۳۸۵۶ نظر غیرمستقیم دارای برچسب مثبت/منفی ایجاد شد. سپس از این پیکره به عنوان مجموعه آموزش در روش‌های یادگیری ماشین برای دسته‌بندی نظرات یک مجموعه تست استفاده شد. بدین منظور اثر پیش‌پردازش‌های مختلف بر روی پیکره مورد بررسی قرار گرفت. همچنین مجموعه‌ای از ویژگی‌های لغوی، نحوی و معنایی در چهار دسته‌بند بیزین، ماشین بردار پشتیبان، J48 و بیشینه آنتروپی بکار گرفته شد.

نتایج آزمایشات انجام گرفته به منظور ارزیابی روش پیشنهادی نشان می‌دهد که بکارگیری ترکیب این ویژگی‌ها دقت الگوریتم تعیین بار را به میزان قابل توجهی در مقایسه با حالات پایه (استفاده از تک‌واژه‌ها و استفاده از برچسب نحوی به تنهایی) افزایش می‌دهد. همچنین روش پیشنهادی در تعیین بار نظرات مجموعه تست به نتایج بسیار بهتری در مقایسه با یک روش برجسته موجود به نام sentic patterns می‌رسد.

مراجع

[1] Liu, B., *"Sentiment Analysis and Opinion Mining,"*, Synthesis Lectures on Human Language Technologies, vol. 5, pp. 1-167, 2012.

[2] Mintz, M., Bills, S., Snow, R., Jurafsky, D., *"Distant Supervision for Relation Extraction without Labeled Data,"* in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, pp. 1003-1011, 2009.

[3] Esuli A., Sebastiani, F., *"Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining,"* in Proceedings of LREC, pp. 417-42, 2006.

[4] Neviarouskaya, A. Prendinger, H., Ishizuka, M., *"SentiFul: A Lexicon for Sentiment Analysis,"* Affective Computing, IEEE Transactions on, vol. 2, pp. 22-36, 2011.

[5] Huang, S., Niu, Z., Shi, C., *"Automatic Construction of Domain-specific Sentiment Lexicon based on Constrained Label Propagation,"* Knowledge-Based Systems, vol. 56, pp. 191-200, 2014.

[6] Goeuriot, L., Na, J. C., Min Kyaing, W. Y., Khoo, C., Chang, Y.K., Theng, Y. L., et al., *"Sentiment Lexicons for Health-related Opinion Mining,"* in Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, pp. 219-226, 2012.

[7] Pang, B., Lee, L., Vaithyanathan, S., *"Thumbs up?: Sentiment Classification using Machine Learning Techniques,"* in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79-86, 2002.

[8] Go, A., Bhayani, R., Huang, L., *Twitter Sentiment Classification using Distant Supervision,* CS224N Project Report, Stanford, pp. 1-12, 2009.

[9] Habernal, I., Ptáček, T., Steinberger, J., *"Supervised Sentiment Analysis in Czech Social Media,"* Information Processing & Management, vol. 50, pp. 693-707, 2014.

[10] Nakagawa, T., Inui, K., Kurohashi, S., *"Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables,"* in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 786-79, 2010.

[11] Hu M., Liu, B., "*Mining and summarizing customer reviews,"* in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168-177.

[12] Bosco, C., Patti, V., Bolioli, A., *"Developing Corpora for Sentiment Analysis and Opinion Mining: the Case of Irony and Senti-tut,"* IEEE Intelligent Systems, p. 1, 2013.

[13] Yu, B., Diermeier, D., Kaufmann, S., *The Wal - Mart Corpus: A Multi - granularity Corporate Opinion Corpus for Opinion Retrieval, Classification and Aggregation,* in working paper, ed, 2009.

[14] Musat, C., Trausan-Matu, S., *"A Comparative Study of the Relevance of Indirect and Direct Opinions in Economic Texts,"* Annals of DAAAM & Proceedings, 2010.

[15] Musat C., Trausan-Matu, S., *"The Impact of Valence Shifters on Mining Implicit Economic Opinions,"* in Artificial Intelligence: Methodology, Systems, and Applications, ed: Springer, pp. 131-140, 2010.

[16] Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D., *"Deterministic Coreference Resolution based on Entity-centric, Precision-ranked Rules,"* Computational Linguistics, vol. 39, pp. 885-916, 2013.

[17] Aronson, A. R., *"Effective Mapping of Biomedical text to the UMLS Metathesaurus: the MetaMap Program,"* in Proceedings of the AMIA Symposium, 2001, p. 17.

[18] Cambria, E., Hussain, A., *Sentic computing*: Springer, 2012.

[19] Sharma, A., Swaminathan, R., Yang, H., *"A Verb-centric Approach for Relationship Extraction in Biomedical Text,"* in Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on, pp. 377-385, 2010.

[20] Poria, S., Cambria, E., Winterstein, G., Huang, G. B., *"Sentic Patterns: Dependency-based Rules for Concept-level Sentiment Analysis,"* Knowledge-Based Systems, 2014.

---

[1] Opinion mining
[2] Sentiment analysis
[3] Direct opinions
[4] Indirect opinions
[5] Polarity detection
[6] Distant supervision
[7] Sentiment lexicon
[۸] مثال‌های این مقاله از سایت *www.druglib.com* انتخاب شده‌اند. از آنجا که کاربران این سایت از کشورهای مختلف هستند که گاهاً با زبان انگلیسی آشنایی کافی ندارند، در نظرات نوشته شده گاهی غلط املایی و/یا گرامری دیده می‌شود.
[9] Bag-of-words
[10] bigram
[11] Part-of-speech
[12] emoticons
[13] http://nlp.stanford.edu/software/corenlp.shtml
[14] Tokenization
[15] lemmatization
[16] POS tagging
[17] development set
[18] http://metamap.nlm.nih.gov/
[19] www.nlm.nih.gov/research/umls/
[20] Semantic type
[21] Semantic group
[22] Semantic Type Frequency – Inverse Opinion Frequency
[23] http://www.cs.cmu.edu/~mehrbod/RR/
[24] Adverbial modifier
[25] http://lyle.smu.edu/~tspell/jaws/
[26] http://www.cs.waikato.ac.nz/ml/weka/
[27] Naïve Bayes
[28] Support Vector Machine
[29] Maximum Entropy
[30] www.sentic.net