ساخت نیمه‌خودکار یک پیکره از نظرات غیرمستقیم در دامنهٔ دارو و به‌کارگیری آن برای تعیین قطبیت نظرات

سمیرا نوفرستی[1*] و مهرنوش شمس‌فرد[2]

[1] دانشکدهٔ مهندسی برق و کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران

[1و2] آزمایشگاه پردازش زبان طبیعی، دانشکدهٔ مهندسی و علوم کامپیوتر، دانشگاه شهیدبهشتی، تهران، ایران

چکیده

مسئلهٔ نظرکاوی در سال‌های اخیر، بسیار مورد توجه بوده است. تمرکز روش‌های موجود نظرکاوی بر نظرات مستقیم بوده است و اغلب آن‌ها از نظرات غیرمستقیم صرف‌نظر کرده‌اند. درحالی‌که در برخی دامنه‌ها از جمله پزشکی نظرات غیرمستقیم به دفعات رخ می‌دهند و نادیده گرفتن آن‌ها باعث کاهش دقت سامانهٔ نظرکاوی می‌شود. در این مقاله روشی نیمه‌خودکار برای ساخت پیکره‌ای از نظرات غیرمستقیم به زبان انگلیسی در دامنهٔ دارو ارائه می‌شود. در مرحلهٔ نخست روش پیشنهادی، جملاتی که بیان‌گر یک نظر غیرمستقیم هستند، شناسایی و در مرحلهٔ دوم قطبیت آن‌ها با کمک دانش دامنه، الگوهای زبانی و ساختار نظرات تعیین می‌شود؛ سپس از این پیکره در روش‌های یادگیری ماشین جهت تعیین قطبیت نظرات مطرح‌شده دربارهٔ داروها استفاده می‌شود. نتایج آزمایش‌ها نشان می‌دهد که روش پیشنهادی در تعیین قطبیت مجموعهٔ آزمون به‌دقت ۸۲/۸۱ درصد می‌رسد و بر یک روش برجسته تحلیل نظرات مستقیم به نام الگوهای سنتیک غلبه می‌کند.

واژگان کلیدی: نظرکاوی، نظرات غیرمستقیم، تحلیل احساسات، ساخت پیکره، یادگیری ماشین.

۱- مقدّمه

از دیرباز معیار تصمیم‌گیری برای خرید یک محصول، انتخاب یک پزشک متبحر، داشتن سفری مفرح و یا برگزیدن یک رستوران خوب، مشورت و کسب نظرات افرادی بوده است که تجربهٔ مشابهی در این زمینه داشته‌اند. پیش از پیدایش وب، تعداد این افراد کم و محدود به اقوام، دوستان و آشنایان بود. وب تغییر بزرگی در روش بیان و دسترسی به نظرات ایجاد کرده است. در سال‌های اخیر، گسترش روزافزون سایت‌های نظرسنجی و وبلاگ‌های شخصی منبع عظیمی از نظرات و تجربیات افراد را فراهم آورده است. پژوهش‌ها نشان می‌دهد که این نظرات هم برای کابران در انتخاب یک محصول و هم برای تولیدکنندگان در ارزیابی محصولاتشان مفید بوده است (موراس و همکاران، ۲۰۱۳).

با توجه به حجم فراوان نظرات بر روی وب امکان خواندن و تحلیل همهٔ آن‌ها توسط کاربران وجود ندارد؛ بدین منظور نیاز به ایجاد ابزارهای خودکار برای تحلیل نظرات احساس می‌شود. نظرکاوی[1] که در برخی از پژوهش‌ها تحلیل احساسات[2] نیز نامیده می‌شود، یک زمینهٔ پژوهشی جدید است که به ارائهٔ روش‌های خودکار جهت کاوش، تحلیل، دسته‌بندی و خلاصه‌سازی نظرات می‌پردازد.

نظرات به دو دستهٔ کلی تقسیم می‌شوند (لیو، ۲۰۱۲): نظرات مستقیم[3] و نظرات غیرمستقیم[4]. در یک نظر مستقیم، یک موجودیت یا جنبه‌ای از آن به‌صورت مستقیم مورد توصیف قرار می‌گیرد. به‌عنوان مثال جملهٔ "این دارو بسیار مفید است." یک نظر مستقیم با قطبیت مثبت را نشان می‌دهد. در مقابل در یک نظر غیرمستقیم اثر یک موجودیت بر روی موجودیت‌های دیگر بیان می‌شود. به‌عنوان مثال جملهٔ "پس از مصرف این دارو دچار درد مفاصل شدم". یک نظر غیر مستقیم است که بیان‌گر اثری منفی از دارو بر روی مفاصل است.

[1] Opinion mining
[2] Sentiment analysis
[3] Direct opinions
[4] Indirect opinions

Archive of SID
www.SID.ir

یکی از وظایف اصلی در نظرکاوی تعیین قطبیت[1] یا بار سنجمانی یک نظر است. هدف این وظیفه دسته‌بندی نظرات در گروه‌های از پیش تعیین شده (اغلب در دو گروه مثبت و منفی) است. در سال‌های اخیر مسألهٔ نظرکاوی بسیار مورد توجه قرار گرفته و تلاش‌های متعددی جهت تعیین قطبیت نظرات در دامنه‌های مختلف انجام یافته است. با این وجود تمرکز این تلاش‌ها بر روی نظرات مستقیم بوده است. این درحالی است که در برخی از دامنه‌ها نظیر اقتصاد و پزشکی نظرات غیرمستقیم به دفعات رخ می‌دهند. به‌طور خاص در دامنهٔ دارو، اغلب کاربران به‌جای بیان یک نظر مستقیم دربارهٔ دارو، به توصیف اثرات مثبت و عوارضی که دارو برای آن‌ها داشته است، می‌پردازند.

بررسی‌های ما بر روی تارنمای www.druglib.com که یک تارنمای نظرسنجی محبوب در رابطه با داروهاست، 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 در این مقاله مسألهٔ تعیین قطبیت نظرات غیرمستقیم مورد توجه قرار می‌گیرد.

به‌منظور تعیین قطبیت نظرات غیرمستقیم از روش‌های یادگیری ماشین استفاده شده است. پیش از این روش‌های یادگیری ماشین در تحلیل نظرات مستقیم به‌کار گرفته شده‌اند و به‌نتایج خوبی دست یافته‌اند. مشکل اصلی این روش‌ها ساخت یک مجموعهٔ آموزش از نظرات است که **قطبیت** مثبت و منفی آن‌ها برچسب خورده باشد. ساخت دستی مجموعهٔ آموزش امری زمان‌بر و هزینه‌بر است. به‌منظور رفع این مشکل در این مقاله روشی برای ساخت نیمه‌خودکار مجموعهٔ آموزش پیشنهاد می‌شود. برای ساخت مجموعهٔ آموزش از روش‌های نظارت از راه دور[2] (مینتز و همکاران، ۲۰۰۹) الهام گرفته شده است. در این روش‌ها به‌جای به‌کارگیری یک مجموعهٔ آموزش دقیق، از یک مجموعهٔ بزرگ از داده‌های نوفه‌ای که به‌صورت خودکار به‌دست آمده‌اند، استفاده می‌شود. روش‌های نظارت از راه دور در حوزه‌های مختلف از جمله استخراج روابط (مینتز و همکاران، ۲۰۰۹) و تحلیل احساسات (لین و همکاران، ۲۰۱۲؛ گو و همکاران، ۲۰۰۹) به‌کار گرفته‌اند و نتایج حاصل از آن‌ها موفقیت‌آمیز بوده است.

نوآوری‌های این مقاله را می‌توان به‌صورت زیر خلاصه کرد:

- ارائهٔ روشی نوین برای ساخت نیمه‌خودکار یک مجموعهٔ برچسب خورده از نظرات غیرمستقیم
- معرفی، به‌کارگیری و ارزیابی ویژگی‌های لغوی، نحوی و معنایی مختلف در روش‌های یادگیری ماشین به‌منظور تعیین قطبیت نظرات غیرمستقیم

ادامهٔ این مقاله به‌صورت زیر سازمان‌دهی شده است. در ابتدا در بخش ۲ کارهای مرتبط انجام گرفته در زمینهٔ تعیین قطبیت نظرات به اختصار مرور و مشکلات آنها در تعیین قطبیت نظرات غیرمستقیم بررسی می‌شود؛ سپس در بخش‌های ۳ و ۴ جزئیات روش پیشنهادی برای ساخت نیمه‌خودکار پیکره‌ای برچسب خورده از نظرات غیرمستقیم تشریح و همچنین روش به‌کارگیری پیکره به‌منظور تعیین قطبیت نظرات جدید شرح داده می‌شود.در بخش ۵ نتایج آزمایش‌های انجام‌گرفته به‌منظور ارزیابی روش پیشنهادی را ارائه خواهیم کرد. در پایان بخش ۶ به نتیجه‌گیری می‌پردازیم.

۲- مرور کارهای پیشین

تلاش‌های متعددی در رابطه با وظایف مختلف مسألهٔ نظرکاوی در دامنه‌های مختلف انجام گرفته است. در (پنگ و همکاران، ۲۰۰۲؛ کمبریا و حسین، ۲۰۱۲؛ لیو، ۲۰۱۲) روش‌های موجود به‌خوبی معرفی و دسته‌بندی شده‌اند. در این بخش تنها به معرفی برخی از کارهای مرتبط در زمینهٔ تعیین قطبیت نظرات می‌پردازیم. کارهای انجام‌گرفته در زمینهٔ تعیین قطبیت نظرات را می‌توان به دو ردهٔ کلی تقسیم کرد: روش‌های مبتنی بر واژگان و روش‌های یادگیری ماشین. در ادامهٔ این بخش ابتدا این دو دسته را مرور می‌کنیم و سپس به معرفی کارهای انجام گرفته در زمینهٔ تحلیل نظرات غیرمستقیم می‌پردازیم.

۲-۱- روش‌های مبتنی بر واژگان

این روش‌ها برای تعیین قطبیت نظرات از یک واژگان سنجمانی[3] استفاده می‌کنند. یک واژگان سنجمانی حاوی مجموعه‌ای از لغات و عبارات با بار سنجمانی مشخص است. تلاش‌های بسیاری جهت ساخت واژگان‌های سنجمانی انجام گرفته است. تمرکز تلاش‌های اولیه بر روی ساخت واژگان‌های عمومی بوده است (اسولی و سباستیانی، ۲۰۰۶؛ نویاروسکایا و همکاران، ۲۰۱۱). با این وجود، قطبیت واژه‌ها

[1] Polarity detection
[2] Distant supervision
[3] Sentiment lexicon

Archive of SID
www.SID.ir

به‌طورکامل وابسته به دامنه است. به‌عنوان مثال واژهٔ "غیرقابل پیش‌بینی" در دامنهٔ "فیلم" می‌تواند دارای قطبیت مثبت باشد؛ درحالی‌که در رابطه با "دزدگیر ماشین" یک صفت منفی است. به‌علاوه واژگان‌های عمومی بسیاری از واژه‌های خاص دامنه را نمی‌پوشانند که این امر خود باعث کاهش دقت الگوریتم‌های نظرکاوی می‌شود.

به‌دلیل مشکلات واژگان‌های سنجمانی عمومی، تلاش‌هایی نیز جهت ساخت واژگان‌های خاص دامنه انجام گرفته است (جرجیوت و همکاران، ۲۰۱۲؛ هوآنگ و همکاران، ۲۰۱۴).

روش‌های مبتنی بر واژگان در تحلیل نظرات غیرمستقیم از دو مشکل اصلی رنج می‌برند. نخست این که اغلب واژگان‌های موجود در سطح واژه هستند. بدین معنا که در این واژگان‌ها به هر واژه یا به هر معنای[1] آن قطبیت نسبت داده می‌شود. این درحالی است که قطبیت و معنای یک جمله تنها از طریق واژه‌ها منتقل نمی‌شود؛ بلکه تعامل واژه‌ها نیز در تعیین قطبیت جمله مؤثر است. یک نظر غیرمستقیم اثر یک موجودیت بر روی موجودیت دیگر را نشان می‌دهد که این اثر اغلب در قالب فعل جمله بیان می‌شود؛ بنابراین فعل جمله تأثیر به‌سزایی در تعیین قطبیت جمله دارد. به مثال‌های زیر توجه کنید[2]:

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

روش‌های مبتنی بر واژگان از تعیین قطبیت درست این‌گونه مثال‌ها قاصرند زیرا واژه‌هایی نظیر "آکنه" و "درد" در واژگان‌های سنجمانی موجود مانند سنتی‌وردنت[3] (اسولی و سباستیانی، ۲۰۰۶) منفی هستند. این امر باعث می‌شود دو جملهٔ بالا منفی در نظر گرفته شوند. درحالی‌که قطبیت کلی هر دو جمله مثبت است. در حالت کلی در دامنهٔ پزشکی اغلب واژه‌ها مانند "بیماری"، "آلزایمر"، "تب" و "درد" منفی هستند. با این وجود، این واژه‌ها به مراتب در جملات مثبت رخ می‌دهند. بنابراین روش‌های مبتنی بر واژگان در این دامنه‌ها دقت کافی ندارند.

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

"This drug decreased my vision."

در سنتی‌وردنت واژهٔ "بینایی" بدون قطبیت است. این امر باعث می‌شود، مثال بالا بدون قطبیت در نظر گرفته شود. درحالی‌که واضح است که این جمله منفی است. به بیانی دیگر روش‌های مبتنی بر واژگان از تحلیل نظرات ضمنی (نظرات فاقد واژه سنجمانی) قاصرند. درواقع مدیریت این نوع نظرات نیاز به تحلیل معنایی عمیق‌تر جملات و عبارات دارد (بالاهور و همکاران، ۲۰۱۱؛ بالاهور و همکاران، ۲۰۱۲).

۲-۲- روش‌های یادگیری ماشین

روش‌های یادگیری ماشین به آموزش یک دسته‌بند بر روی یک پیکرهٔ برچسب‌خورده از نظرات به‌منظور دسته‌بندی نظرات بدون برچسب می‌پردازند. این روش‌ها به‌نتایج قابل‌قبولی در تعیین **قطبیت** نظرات مستقیم دست‌یافته‌اند (پنگ و همکاران، ۲۰۰۲؛ بویی و موئنز، ۲۰۰۹). ویژگی‌های مختلفی در روش‌های یادگیری ماشین به‌کار گرفته شده است. از رایج‌ترین ویژگی‌های مورد استفاده می‌توان به کیسهٔ لغات[4] (حضور یک لغت خاص و/یا تعداد رخداد آن) (پنگ و همکاران، ۲۰۰۲)، ترکیب دوتایی لغات[5] (پنگ و همکاران، ۲۰۰۲)، برچسب‌های نحوی (گو و همکاران، ۲۰۰۹)، شکلک‌ها[6] (هابرنال و همکاران، ۲۰۱۴) و مسیر بین واژه‌ها در درخت وابستگی (ناکاگاوا و همکاران، ۲۰۱۰) اشاره کرد.

برتری اصلی روش‌های یادگیری ماشین در مقایسه با روش‌های مبتنی بر واژگان این است که این روش‌ها با درنظرگرفتن اطلاعات بافتار و روابط بین واژه‌ها تا حدودی قادر به تعیین قطبیت نظرات ضمنی هستند. با این وجود روش‌های یادگیری ماشین از یک مشکل عمده رنج می‌برند و آن ساخت پیکرهٔ آموزش است.

اغلب کارهای پیشین، پیکرهٔ آموزش را به‌صورت دستی ساخته‌اند (هو و لیو، ۲۰۰۴؛ بوسکو و همکاران، ۲۰۱۳). ساخت دستی پیکرهٔ آموزش امری زمان‌بر و هزینه‌بر است که باعث می‌شود، پیکره از لحاظ کمی محدود باشد. بدین دلیل تلاش‌هایی جهت ساخت نیمه‌خودکار و خودکار پیکرهٔ آموزش انجام شده است.

در (کاجی و کیتسورگاوا، ۲۰۰۶) روشی معرفی‌شده است که با به‌کارگیری ساختار صفحات HTML و تعریف تعدادی الگوی زبانی یک پیکرهٔ برچسب‌خورده از نظرات

[1] sense

[2] مثال‌های این مقاله از تارنمای www.druglib.com انتخاب شده‌اند. از آنجا که کاربران این تارنما از کشورهای مختلف هستند که گاهاً با زبان انگلیسی آشنایی کافی ندارند، در نظرات نوشته‌شده گاهی غلط املایی و/یا گرامری دیده می‌شود.

[3] SentiWordNet

[4] bag-of-words
[5] bigram
[6] emoticons

Archive of SID
www.SID.ir

می‌سازد. در (گو و همکاران، ۲۰۰۹) از شکلک‌های مورد استفاده در توئیت‌های کاربران در توئیتر[1] برای ساخت مجموعهٔ آموزش استفاده شده است.

۲-۳- تحلیل نظرات غیرمستقیم

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 که تمرکز آن‌ها بر برچسب‌زدن نظرات غیرمستقیم به‌صورت دستی بوده است.

در (یو و همکاران، ۲۰۰۹) یک پیکره از نظرات در دامنهٔ اخبار به‌صورت دستی برچسب زده شده‌اند. در این پیکره علاوه بر نظرات مستقیم، نظرات غیرمستقیم نظیر "افزایش قیمت سهام" با برچسب خاص خود مشخص شده‌اند. در (موسات و تراسان- متو، ۲۰۱۰) روشی برای تعیین قطبیت نظرات غیرمستقیم در دامنهٔ اقتصاد معرفی شده است که درابتدا با استفاده از یک فرهنگ لغت خاص دامنهٔ اصلاحات تخصصی متن را استخراج می‌کند. شاخص‌های اقتصادی به دو گروه مثبت و منفی تقسیم شده است. همچنین تعدادی اصلاح‌کنندهٔ مثبت و منفی تعریف شده است؛ سپس با به‌کارگیری یک فرمول سادهٔ قطبیت شاخص در قطبیت اصلاح‌کنندهٔ آن ضرب شده و قطبیت کل نظر به‌دست می‌آید.

در (نوفرستی و شمس فرد، ۲۰۱۵) ابتدا مدلی برای نمایش یک نظر غیرمستقیم معرفی شده است؛ سپس بر اساس این مدل یک پایگاه دانش از نظرات غیرمستقیم ایجاد شده و از روی تاپل‌های موجود در پایگاه دانش تعدادی الگوی معنایی استخراج شده است. با به‌کارگیری تاپل‌های موجود در پایگاه دانش و نیز الگوهای استخراج‌شده از آن روشی برای تعیین قطبیت نظرات در دامنهٔ دارو پیشنهاد شده است.

در این مقاله روشی نیمه‌خودکار برای ساخت یک پیکره از نظرات غیرمستقیم ارائه می‌شود؛ سپس پیکرهٔ ساخته‌شده به‌عنوان مجموعه آموزش در روش‌های یادگیری ماشین برای تعیین قطبیت نظرات غیرمستقیم مورد استفاده قرار می‌گیرد. بدین منظور تعدادی ویژگی نحوی و معنایی جدید برای استفاده در دسته‌بندهای یادگیری ماشین معرفی می‌شود. نتایج آزمایش‌های انجام گرفته، مفیدبودن این ویژگی‌ها را در تعیین قطبیت نظرات غیرمستقیم نشان می‌دهد.

۳- روش پیشنهادی برای ساخت یک پیکره از نظرات غیرمستقیم

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

۳-۱- مجموعه داده‌ای

مجموعه داده‌ای مورد استفاده جهت ساخت پیکره از تارنمای www.druglib.com جمع‌آوری شده است. بدین منظور نظرات مطرح‌شده دربارهٔ ۸۳ دارو گردآوری شده است. این مجموعه دارای ۱۳۱۳ نظر و ۷۰۱۷ جمله است.

هر نظر دارای دو بخش است شکل (۱): متا داده‌ها و متن نظر. متا داده‌ها شامل اطلاعاتی از قبیل سن، جنسیت بیمار و نیز امتیازی که او به اثربخشی[2] و عوارض جانبی[3] دارو می‌دهد، است. متن نظر نیز شامل سه بخش است: فواید، عوارض دارویی و توضیحات. پیش‌فرض این است که کاربر در بخش فواید دربارهٔ اثرات مثبت دارو می‌نویسد. به‌عبارت دیگر انتظار می‌رود جملات این بخش دارای قطبیت مثبت باشند. به‌طور مشابه فرض می‌شود جملات نوشته‌شده در بخش عوارض دارویی دارای قطبیت منفی هستند؛ اگر چه این فرضیات به‌الزام همیشه درست نیستند. جملات مطرح‌شده در بخش توضیحات نیز محدودیتی ندارند.

Accutane review by 36 year old male patient

	Rating
Overall rating:	★★★★★★★★☆☆
Effectiveness:	Highly Effective
Side effects:	Moderate Side Effects
	Treatment Info
Condition / reason:	acne
Dosage & duration:	30 mg a day taken daily for the period of 2 months
Other conditions:	none
Other drugs taken:	none
	Reported Results
Benefits:	My skin dried up and the acne went away. This was extremely effective.
Side effects:	My face skin got a bit too dry, but manageable with lotion and drinking lots of water.
Comments:	took 30 mg daily and had my blood work monitored frequently.

(شکل-۱): مثالی از یک نظر مطرح‌شده در تارنمای www.druglib.com

[1] Twitter

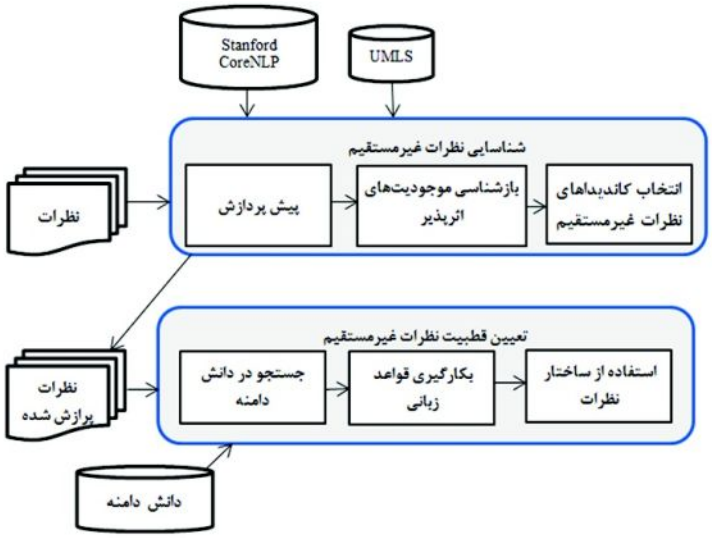
[2] effectiveness

[3] Side effects

Archive of SID

۳-۲- ساخت پیکره

به‌منظور ساخت یک پیکره از نظرات غیرمستقیم یک روش دومرحله‌ای معرفی می‌شود. در مرحلهٔ نخست جملاتی که بیان‌گر یک نظر غیرمستقیم هستند، شناسایی و در مرحلهٔ دوم **قطبیت** مثبت و منفی آنها تعیین می‌شود. نمای کلی روش پیشنهادی در شکل (۲) نشان داده شده است. در ادامه، مراحل کار با جزئیات تشریح می‌شود.



(شکل-۲): نمای روش پیشنهادی برای ساخت پیکره

۳-۲-۱- شناسایی نظرات غیرمستقیم

نظر غیرمستقیم اثر یک موجودیت را که اثرگذار نامیده می‌شود، بر موجودیتی دیگر به نام 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 ۹۰ درصد جملات استخراج‌شده با این روش به‌درستی یک نظر غیرمستقیم را نشان می‌دهند؛ اما این روش بسیاری از نظرات غیرمستقیم را استخراج نمی‌کند. علت این امر این است که اغلب نظرات غیرمستقیم فاقد موجودیت اثرگذار هستند. از آنجا که کاربران تارنماهای نظرسنجی اغلب دربارهٔ یک موجودیت مشخص (موجودیتی که نام آن در عنوان پست ذکر شده است) نظر می‌دهند، به‌ندرت نام آن موجودیت را در متن نظر خود ذکر می‌کنند؛ بلکه تنها به بیان اثر آن موجودیت می‌پردازند. به مثال‌های زیر توجه کنید:

“reduced amount of cyctic acne”
“Lips were chapped and nose was dried.”
“no migraine”

در این نظرات تنها به اثرات دارو اشاره شده است و نامی از دارو نیامده است. بنابراین فرض شده هر جمله‌ای که حاوی یک موجودیت اثرپذیر است، یک نظر غیرمستقیم را نشان می‌دهد. با این فرض به‌منظور شناسایی نظرات غیرمستقیم گام‌های زیر طی می‌شود (شکل ۲):

پیش‌پردازش: برای هر نظر در مجموعهٔ نظرات، ابتدا با توجه به ساختار نظر، تنها جملاتی که در بخش فواید یا عوارض مطرح شده‌اند، استخراج و از جملاتی که در بخش توضیحات نوشته شده‌اند، صرف‌نظر می‌شود؛ سپس با به‌کارگیری بستهٔ پردازش زبان استنفورد[1] جمله‌های یک نظر از یکدیگر جدا می‌شوند؛ سپس هر جملهٔ مرکب[2] به تعدادی واحد کوچک‌تر شکسته می‌شود. این کار با توجه به درخت وابستگی[3] جمله و ساختار حروف ربط به‌کار رفته در آن انجام می‌گیرد (دی مارنف و همکاران، ۲۰۰۶)؛ سپس جملات سؤالی، شرطی و امری حذف می‌شوند. اگرچه برخی از این نوع جملات می‌توانند حاوی نظر باشند؛ اما تکنیک تحلیل آن‌ها متفاوت از جملات خبری است. پس از آن برای هر جملهٔ باقی‌مانده با کمک ابزارهای موجود در بستهٔ پردازش زبان استنفورد قطعه‌بندی[4]، لم‌یابی[5] و برچسب‌زنی نحوی[6] انجام می‌گیرد. در پایان نیز مرجع ضمایر تشخیص داده می‌شوند (لی و

[1] http://nlp.stanford.edu/software/corenlp.shtml
[2] Compound sentence
[3] Dependency tree
[4] Tokenization
[5] lemmatization
[6] POS tagging

Archive of SID
www.SID.ir

همکاران، ۲۰۱۳) و هر ضمیر یا عبارتی که به آن ارجاع دارد، جایگزین می‌شود.

بازشناسی موجودیت‌های اثرپذیر: همان‌طورکه پیش از این گفته شد، فرض شده است هر جمله‌ای که شامل یک موجودیت اثرپذیر باشد، یک نظر غیرمستقیم است. بنابراین به‌منظور شناسایی نظرات غیرمستقیم، در ابتدا نیاز است انواع ممکن برای موجودیت‌های اثرپذیر تعیین شود. پیش‌بینی انواع ممکن برای موجودیت اثرپذیر امری دشوار است. به‌عنوان مثال اگرچه می‌دانیم "آکوتان" یک داروی پوستی است و به‌احتمال بسیاری از نظرات، اثر این دارو را بر روی پوست بیان می‌کنند، با این وجود نمی‌دانیم عوارض جانبی این دارو بر روی چه اندام‌هایی اثر می‌گذارد.

در این مقاله به‌منظور تعیین انواع ممکن برای موجودیت اثرپذیر یک روش خودکار پیشنهاد شده است. ایده این است که نوع موجودیت‌های پررخداد را به‌عنوان نامزد در نظر بگیریم. دلیل درست‌بودن این ایده این است که به‌طورمعمول موجودیت‌هایی که اغلب کاربران دربارهٔ آن نظر می‌نویسند، موجودیت‌های مهم هستند. اگرچه گاهی کاربران دربارهٔ موضوعاتی نامربوط نیز می‌نویسند؛ اما از آنجا که این مطالب برای هر کاربر متفاوت از کاربر دیگر است، موجودیت‌های نامربوط کم‌رخداد هستند. بنابراین می‌توان موجودیت‌های کم‌رخداد را حذف کرد.

برای استخراج موجودیت‌های پررخداد، دویست نظر از تارنمای www.askapatient.com جمع‌آوری و از آن به‌عنوان مجموعه توسعه[1] استفاده شده است. برای هر جمله در این مجموعه مفاهیم پزشکی برچسب زده شده‌اند. برای این کار از نرم‌افزار متامپ[2] استفاده شده است (آرونسون، ۲۰۰۱). این نرم‌افزار مفاهیم جمله را به مفاهیم تعریف‌شده در UMLS[3] نگاشت می‌کند. UMLS یک فرهنگ جامع شامل ۱/۷ میلیون مفهوم است که در ۱۳۰ گروه دسته‌بندی شده‌اند. به هریک از این گروه‌ها یک نوع معنایی[4] گفته می‌شود.

مشکل در این است که UMLS علاوه‌بر مفاهیم پزشکی مفاهیم عمومی نظیر زمان را نیز دارد. به جدول (۱) توجه کنید. این جدول مثالی را از یک نظر غیرمستقیم که توسط متامپ برچسب زده شده، نشان می‌دهد.

(جدول-۱): مثالی از موجودیت‌های تشخیص داده شده توسط متامپ

"Over the course of three nights, Avelox caused heart palpitations, confusion, and lasting muscle weakness."	
Entity	**Semantic type**
Course	Tmco
three	Qnco
nights	Tmco
Avelox	orch-phsu
caused	ftcn-cnce
heart	Bpoc
palpitations	Fndg
confusion	Fndg
muscle weakness	Fndg

همان‌طورکه در جدول (۱) دیده می‌شود، علاوه‌بر مفاهیم مدنظر (یعنی "palpitations" "confusion" و "muscle weakness") مفاهیم عمومی نظیر "course" و "nights" نیز برچسب خورده‌اند. برای حل این مشکل باید مفاهیم مهم خاص دامنه انتخاب و مفاهیم عمومی حذف شوند. مفاهیم عمومی برخلاف مفاهیم خاص دامنه در اغلب دامنه‌ها پررخداد هستند. بنابراین می‌توان مفاهیمی را که در دامنه‌های مختلف پررخداد هستند، به‌عنوان مفاهیم عمومی در نظر گرفت و آن‌ها را حذف کرد.

به‌منظور تعیین مفاهیم مهم و مرتبط دامنهٔ معیاری با نام STF-IOF[5] (تعداد رخداد نوع معنایی-معکوس تعداد رخداد نظر) تعریف شده است. این معیار اصلاح‌شده معیار CF-IOF[6] است که در (کمبریا و حسین، ۲۰۱۲) به‌منظور تعیین اهمیت یک مفهوم در یک متن خاص معرفی شده است. STF-IOF که اهمیت یک مفهوم را در یک دامنهٔ خاص اندازه می‌گیرد، به‌صورت زیر تعریف شده است:

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

که $n_{st,d}$ تعداد رخداد نوع معنایی st در مجموعه نظرات 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.

برای استفاده از رابطهٔ (۱) نیاز به یک مجموعه از نظرات است که بخشی از نظرات (به نام d) مربوط به دامنهٔ مد نظر و بخش دیگر شامل نظرات در دیگر دامنه‌ها باشد. یک مجموعه از نظرات از سه دامنهٔ دارو، رستوران[7] و

[1] development set
[2] *http://metamap.nlm.nih.gov/*
[3] *www.nlm.nih.gov/research/umls/*
[4] Semantic type
[5] Semantic Type Frequency – Inverse Opinion Frequency
[6] Concept Frequency – Inverse Opinion Frequency
[7] http://www.cs.cmu.edu/~mehrbod/RR/

Archive of SID
www.SID.ir

محصولات (هو و لیو، ۲۰۰۴) گردآوری شده است. مجموعه نظرات در دامنهٔ داروها همان مجموعه توسعه است که پیش‌تر توصیف شد. با اعمال معیار STF_IOF مفاهیم با اهمیت دامنهٔ دارو استخراج شده است. برای این منظور نوع معنایی موجودیت‌های جملات مجموعه نظرات توسط متامپ تعیین شده‌اند. اهمیت هر نوع معنایی با استفاده از معیار STF_IOF تعیین شده است. انواع معنایی که مقدار STF_IOF آن از آستانه θ بیشتر باشد به‌عنوان مفاهیم مهم دامنه انتخاب شده‌اند؛ سپس برای انواع معنایی پررخداد گروه معنایی متناظر با آن‌ها جایگزین شده است.

برای تعیین مقدار θ و α می توان از روش سعی و خطا یا از یک الگوریم تطبیق پارامتر استفاده کرد. با توجه به موفقیت الگوریتم‌های مبتنی بر اتوماتاهای یادگیر در تطبیق پارامتر (نوفرستی و رحایی، ۲۰۱۱)، برای تعیین مقدار مناسب پارامترها از اتوماتای یادگیر استفاده شده است که به‌ترتیب برای پارامترهای θ و α مقادیر 0.02 و 0.1 به‌دست آمده است.

انتخاب نامزدها: پس از تعیین انواع ممکن برای موجودیت‌های اثرپذیر هر جمله‌ای که دارای حداقل یک موجودیت اثرپذیر باشد به‌عنوان یک نامزد برای نظر غیرمستقیم در نظر گرفته شده است.

نتایج آزمایش‌های انجام‌گرفته نشان می‌دهد که روش پیشنهادی برای شناسایی نظرات غیرمستقیم دارای دقت[1] ۸۵/۰۵ درصد و فراخوانی[2] ۷۷/۸۱ درصد است.

۳-۲-۲- تعیین قطبیت نظرات غیرمستقیم

به‌منظور تعیین قطبیت 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. به‌عنوان مثال "کاهش درد" از فواید شناخته‌شدهٔ "استامینوفن" و "خشکی لب" از عوارض شناخته‌شدهٔ "آکوتان" است. در این مقاله دانش دامنه که شامل فواید و عوارض شناخته‌شدهٔ داروهای مختلف است از تارنمای *www.dailymed.com* استخراج شده است. اگر در یک نظر غیرمستقیم دربارهٔ فایده‌ای شناخته‌شده از دارو صحبت شده باشد، قطبیت آن نظر را مثبت در نظر می‌گیریم. در مقابل اگر نظر عارضه‌ای شناخته‌شده از دارو را بیان کند قطبیت آن منفی در نظر گرفته می‌شود. البته برای تعیین قطبیت، نظرات منفی‌کننده‌ها را نیز مد نظر قرار داده‌ایم. منفی‌کننده‌ها واژه‌هایی هستند که قطبیت یک جمله را تغییر می‌دهند. به‌عنوان مثال اگرچه "migraine" یک بیماری است که بر طبق دانش دامنه قطبیت منفی دارد؛ ولی "no migraine" به‌دلیل وجود منفی‌کننده "no" دارای قطبیت مثبت است. منفی‌کننده‌های یک جمله با استفاده از فهرستی از منفی‌کننده‌های رایج شامل "no"، "not" و "never" شناسایی می‌شوند. اگر این منفی‌کننده در بازهٔ یک پنجرهٔ سه‌کلمه‌ای از فعل جمله یا موجودیت اثرپذیر قرار داشته باشد، قطبیت جمله را معکوس می‌کند؛ در غیر این‌صورت از آن صرف‌نظر می‌شود. عدد سه با روش سعی و خطا تعیین شده است.

قواعد زبانی: در این مرحله قطبیت برخی از نظرات که تابه‌حال برچسب قطبیت نخورده‌اند با استفاده از قواعد زبانی تعیین می‌شود. در این مقاله دو قاعدهٔ زبانی مورد استفاده قرار گرفته است (لیو، ۲۰۱۲)

- قاعده "و": این قاعده می‌گوید جملات و عباراتی که با حروف ربط مشابه "و" به یکدیگر متصل می‌شوند، قطبیت یکسانی دارند. به‌عنوان مثال در متن "My skin dried up and the acne went away." دو جمله هر دو دارای قطبیت مثبت هستند.
- قاعده "اما": این قاعده می‌گوید جملات و عباراتی که با حروف ربط مشابه "اما" مرتبط می‌شوند، اغلب دارای قطبیت متضاد هستند. به‌عنوان مثال دو جملهٔ به‌کار رفته در متن زیر دارای قطبیت متضاد هستند.

"At first, I noticed an increase of energy, for about 4 days, but then extreme fatigue set it."

با به‌کارگیری دو قاعدهٔ بالا اگر قطبیت یک جمله مشخص باشد (در گام پیشین قطبیت آن تعیین شده باشد) می‌توان قطبیت جملهٔ دیگر را تعیین کرد.

ساختار نظرات: بسیاری از نظرات کاربران بیان‌گر اثراتی هستند که در دانش دامنه به‌وضوح و با آن الفاظ دیده نمی‌شوند. برای تعیین قطبیت این‌گونه نظرات از ساختار نظرات استفاده شده است. درابتدا فرض شد جملاتی که در بخش فواید یک نظر مطرح می‌شوند، مثبت و جملات مطرح‌شده در بخش عوارض جانبی، منفی هستند. با این وجود این فرض در برخی از موارد درست نیست. به‌عنوان مثال متن زیر در بخش فواید یک نظر نوشته شده است:

"Benefits: Imitrex just isn't right for me. I experienced flushing, sensations of tingling/prickling,

[1] precision
[2] recall

Archive of SID
www.SID.ir

weakness, drowsiness, dizziness, fast/pounding heartbeat. On top of that the packaging is just HORRIBLE. It comes 9 to a package. The packaging is a tri-fold booklet that is very rigid and has individual blisters for each tablet. It's about 6 1/2 by 3 1/2 inches, folded. I like to have migraine meds on me at all times so I ended up cutting the tablets out of the booklet so they would fit in my wallet or purse. Definitely going back to Axert!"

همان‌طورکه دیده می‌شود، اغلب جملات متن بالا قطبیت منفی دارند. به‌منظور رفع این مشکل معیاری به نام درجهٔ اطمینان تعریف شده است که درواقع بیان‌گر میزان اطمینان ما به مثبت یا منفی‌بودن یک نظر است. درجهٔ اطمینان از رابطهٔ زیر محاسبه می‌شود:

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

رابطهٔ (۲) متغیر e دوتایی (اثر، اثرپذیر) که بیان‌گر یک نظر غیرمستقیم است، نشان می‌دهد. به‌منظور عمومی‌سازی به‌جای موجودیت اثرپذیر نوع معنایی آن قرار داده می‌شود. همان‌طورکه پیش‌تر نیز گفته شد، نوع معنایی توسط نرم‌افزار متامپ به‌دست می‌آید. به‌عنوان مثال "کاهش درد" را به‌جای (reduce, pain) با (reduce, sosy) نمایش می‌دهیم که sosy طبقهٔ معنایی علائم بیماری‌ها است. دلیل جایگزینی موجودیت اثرپذیر با نوع معنایی آن، کاهش مشکل تنکی داده‌ها و افزایش پوشش الگوریتم است. مفاهیم دارای یک نوع معنایی، رفتار مشابهی در الگوریتم تعیین قطبیت نظر دارند. بدین دلیل در مواردی که مشکل تنکی داده‌ها وجود دارد و نمی‌توان یک نظر را به‌طورمستقیم تعیین قطبیت کرد، از نظرات دیگری که موجودیت اثرپذیر آن‌ها دارای نوع معنایی یکسان با موجودیت مد نظر است، استفاده می‌شود. PF و NF نیز به‌ترتیب تعداد رخداد e در بخش فواید و عوارض هستند. بر طبق رابطهٔ (۲) اگر e تنها در بخش فواید (یا عوارض) رخ داده باشد درجهٔ اطمینان آن یک می‌شود. در مقابل اگر e به تعداد یکسان در بخش فواید و عوارض رخ داده باشد، درجهٔ اطمینان آن صفر می‌شود. درواقع هرچه درجهٔ اطمینان یک اثر به یک نزدیک‌تر باشد، میزان اطمینان ما به مثبت (یا منفی) بودن آن اثر بیشتر است.

به‌منظور استفاده از معیار بالا به این صورت عمل شده است که در مجموعه جملاتی که به‌عنوان نظر غیرمستقیم شناسایی شده‌اند، اگر نظر حاوی یک فعل باشد دوتایی (ریشهٔ فعل، نوع معنایی موجودیت اثرپذیر) و اگر نظر یک عبارت بدون فعل باشد، دوتایی (اصلاح‌کننده[1]، نوع معنایی موجودیت اثرپذیر) استخراج می‌شود که البته اصلاح‌کننده می‌تواند مقدار نداشته باشد؛ سپس برای هر دوتایی درجهٔ اطمینان محاسبه شده است. تنها دوتایی‌هایی که درجهٔ اطمینان آن‌ها از یک آستانهٔ از پیش تعیین شده بیشتر باشد، انتخاب می‌شوند. در این مقاله با روش سعی‌وخطا مقدار آستانه ۰/۵ تعیین شده است. درجهٔ اطمینان را تنها برای دوتایی‌هایی که حداقل سه‌بار در مجموعهٔ نظرات رخ داده‌اند محاسبه می‌کنیم و از سایر دوتایی‌ها صرف‌نظر می‌شود. در پایان برای هر دوتایی انتخاب شده اگر تعداد رخداد آن در بخش فواید بیشتر از تعداد رخداد آن در بخش عوارض باشد قطبیت آن را مثبت و در غیر این صورت قطبیت آن را منفی در نظر می‌گیریم. در این مرحله نیز منفی‌کننده‌ها مد نظر قرار گرفته‌اند. به‌عنوان مثال اگرچه دوتایی (reduce,sosy) دارای قطبیت مثبت است، اما اگر این دوتایی در جمله‌ای که حاوی یک منفی‌کننده است، ظاهر شود قطبیت جمله منفی خواهد بود.

جهت استخراج دوتایی (ریشهٔ فعل، نوع معنایی موجودیت اثرپذیر) از یک روش فعل- محور که به‌منظور استخراج رابطه از متون پزشکی معرفی شده (شارما و همکاران، ۲۰۱۰) استفاده شده است. در این روش ابتدا فعل اصلی جمله تعیین می‌شود، سپس با استفاده از درخت وابستگی جمله اگر موجودیت اثرپذیر با فعل جمله رابطه‌ای نظیر مفعول، فاعل، اصلاح‌گر قیدی[2] و غیره داشته باشد، دوتایی مذکور استخراج می‌شود. مؤلفهٔ نخست این دوتایی ریشهٔ فعل است که از ریشه‌یاب موجود در بستهٔ پردازش زبان استنفورد به‌دست می‌آید و مؤلفهٔ دوم نوع معنایی موجودیت اثرپذیر است که توسط متامپ تعیین می‌شود.

به‌طور مشابه برای استخراج دوتایی (اصلاح‌کننده، نوع معنایی موجودیت اثرپذیر) از عبارات فاقد فعل، درابتدا موجودیت اثرپذیر مشخص می‌شود؛ سپس به کمک تعدادی قاعدهٔ زبانی که بر روی درخت وابستگی عبارت اعمال می‌شود، اصلاح‌کننده موجودیت اثرپذیر به‌دست می‌آید.

به‌عنوان مثال در عبارت "reduction of acne" موجودیت اثرپذیر "acne" است. در درخت وابستگی این عبارت، لغت "reduction" که با موجودیت اثرپذیر "acne" رابطهٔ "prep-of" دارد، به‌عنوان اصلاح‌کنندهٔ آن انتخاب می‌شود؛ سپس ریشهٔ اصلاح‌کننده با استفاده از تابع getDerivationallyRelatedForms در وردنت و نیز ریشه‌یاب استنفورد به‌دست می‌آید. در مثال بالا برای

[1] modifier

[2] Adverbial modifier

Archive of SID
www.SID.ir

اصلاح‌کنندهٔ "reduction" ریشهٔ "reduce" به‌دست می‌آید. بدین ترتیب برای عبارت مثال، دوتایی (reduce, acne) حاصل می‌شود که در مرحلهٔ بعد به‌جای موجودیت اثرپذیر نوع معنایی آن قرار داده می‌شود. جدول (۲) مثال‌هایی را از نظرات غیرمستقیم و دوتایی‌های استخراج‌شده از آن‌ها نشان می‌دهد.

(جدول-۲): مثال‌هایی از دوتایی‌های استخراج‌شده از نظرات غیرمستقیم

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 از نظراتی که قطبیت آن‌ها نامشخص باقی‌مانده است صرف‌نظر می‌کنیم.

۴- به‌کارگیری پیکره

به‌منظور ارزیابی کارآیی پیکرهٔ ساخته‌شده، از آن به‌عنوان مجموعه آموزش در روش‌های یادگیری ماشین استفاده شده است. بدین منظور برای هر جملهٔ دارای برچسب مثبت و منفی در پیکرهٔ ویژگی‌های مد نظر استخراج شده است؛ سپس مجموعه جملات به‌عنوان مجموعهٔ آموزش در یک دسته‌بند یادگیری ماشین استفاده شده است.

به‌منظور آموزش دسته‌بندهای یادگیری ماشین بر روی پیکرهٔ ساخته‌شده، پیش‌پردازش‌های مختلف بر روی آن انجام شده است. همچنین ویژگی‌های مختلفی به‌منظور دسته‌بندی به‌کارگرفته شده است که در ادامه به توصیف آنها می‌پردازیم.

۴-۱- پیش‌پردازش

پیش‌پردازش متن یکی از گام‌های رایج در بسیاری از کاربردهای پردازش زبان طبیعی است. در این مقاله نیز سه پیش‌پردازش رایج به نام‌های حذف هرزواژه‌ها، تبدیل متن به حروف کوچک و ریشه‌یابی استفاده شده است. اثر هریک از این پیش‌پردازش‌ها در دقت الگوریتم نظرکاوی در بخش ۵ مورد ارزیابی قرار گرفته است.

۴-۲- ویژگی‌ها

معرفی و انتخاب ویژگی‌های مناسب تأثیر به‌سزایی در دقت روش‌های یادگیری ماشین دارد. در زمینهٔ نظرکاوی ویژگی‌های متعددی معرفی و به‌کارگرفته شده‌اند که اغلب لغوی و نحوی هستند (مانند تک‌واژه‌ها، زوج‌واژه‌ها و برچسب نحوی). در این مقاله برخی از ویژگی‌های رایج در نظرکاوی درکنار تعدادی ویژگی جدید نظیر عبارت[1] حاوی موجودیت اثرپذیر، فعل جمله، نوع و گروه معنایی موجودیت اثرپذیر به‌کارگرفته شده‌اند. این ویژگی‌ها در جدول (۳) توصیف شده‌اند.

(جدول-۳): ویژگی‌های مورد استفاده در یادگیری ماشین

نماد	ویژگی	توصیف	اندازهٔ بردار ویژگی	
			بدون ریشه‌یابی	با ریشه‌یابی
F1	unigram	تک‌واژه‌ها	۳۵۹۵	۲۹۹۷
F2	bigram	زوج‌واژه‌ها	۱۶۴۳۸	۱۵۶۷۶
F3	POS	برچسب نحوی	۳۶	۳۶
F4	mainVerb	فعل جمله که اثر بیان‌شده بر روی موجودیت اثرپذیر را بیان می‌کند.	۱۲۰۸	۱۰۵۷
F5	affected	عبارت حاوی موجودیت اثرپذیر	۲۱۱۹	۲۰۸۰
F6	type	نوع معنایی موجودیت اثرپذیر	۱۳۰	
F7	group	گروه معنایی موجودیت اثرپذیر	۱۵	
F8	negation	آیا جمله دارای منفی‌کننده است؟	این ویژگی یک ویژگی دودویی در نظر گرفته شده است.	

همان‌طورکه در جدول (۳) مشاهده می‌شود، در این مقاله دو ویژگی معنایی به نام‌های نوع و گروه معنایی موجودیت اثرپذیر به‌کارگرفته شده است. همان‌طورکه پیش‌تر گفته شد، مفاهیم UMLS در ۱۳۰ نوع معنایی گروه‌بندی شده‌اند. این انواع خود در پانزده گروه دسته‌بندی می‌شوند که هر دسته یک گروه معنایی[2] نامیده می‌شود. به‌عنوان مثال مفهوم "بی‌خوابی" دارای نوع معنایی "علامت بیماری" و گروه معنای "اختلال" است. به‌کارگیری ویژگی‌های معنایی باعث کاهش مشکل تنکی داده‌ها و

[1] chunk
[2] Semantic group

Archive of SID
www.SID.ir

افزایش پوشش الگوریتم یادگیری می‌شود. همچنین از آنجا که در یک نظر غیرمستقیم اغلب اثر موجودیت اثرگذار بر روی موجودیت اثرپذیر توسط فعل جمله بیان می‌شود، فعل نیز به‌عنوان یک ویژگی مؤثر درنظر گرفته شده است. در بخش ۵، کارآیی هر یک از این ویژگی‌ها مورد بررسی قرار گرفته است.

۵- نتایج

در این بخش درابتدا مجموعه آزمون مورد استفاده، توصیف می‌شود؛ سپس آزمایش‌های انجام‌گرفته به‌منظور ارزیابی روش‌های پیشنهادی ارائه می‌شود. آزمایش‌های انجام‌گرفته به دو دستهٔ کلی تقسیم می‌شوند. دستهٔ نخست به ارزیابی روش پیشنهادی برای ساخت پیکره می‌پردازد و دستهٔ دوم به سنجش کارآیی روش پیشنهادی برای به‌کارگیری پیکره جهت تعیین قطبیت نظرات اختصاص دارد.

۵-۱- مجموعهٔ آزمون

از آنجا که یک مجموعهٔ آزمون عمومی که خاص مسألهٔ تعیین قطبیت نظرات غیرمستقیم در دامنهٔ دارو طراحی شده باشد وجود ندارد؛ اقدام به ساخت یک مجموعهٔ دستی از نظرات غیرمستقیم شده است. بدین منظور بخشی از نظرات مطرح‌شده دربارهٔ نوزده دارو در دو تارنمای نظرسنجی www.askeapatient.com و www.druglib.com گردآوری شده‌اند. این نوزده دارو به‌صورت تصادفی از فهرست داروهایی که بیشترین تعداد نظرات را داشته‌اند، انتخاب شده‌اند. همچنین دقت شده است که نظرات انتخاب‌شده از تارنمای www.druglib.com با مجموعه نظرات انتخاب‌شده برای ساخت پیکره، هم‌پوشانی نداشته باشند؛ سپس از دو شخص خواسته شده است که نظرات غیرمستقیم این مجموعه را استخراج کرده و به آنها برچسب قطبیت (مثبت یا منفی) اختصاص دهند. تنها جملاتی که توسط هر دو شخص به‌عنوان نظر غیرمستقیم انتخاب شده‌اند و قطبیت یکسانی نیز برای آن‌ها در نظر گرفته شده است، در مجموعهٔ آزمون نهایی قرار گرفته‌اند. بدین ترتیب مجموعه‌ای از ۲۸۵ نظرغیرمستقیم ساخته شده است که حدود شصت درصد آن‌ها مثبت و بقیه نظرات منفی هستند.

۵-۲- ارزیابی روش پیشنهادی برای ساخت پیکره

با به‌کارگیری روش پیشنهادی برای ساخت پیکره (بخش ۳)، پیکره‌ای از نظرات غیرمستقیم ساخته شد. مشخصات این پیکره در جدول (۴) ارائه شده است.

(جدول-۴): مشخصات پیکره ساخته شده

تعداد جملات	۳۸۵۶
تعداد کلمات	۳۵۰۶۴
تعداد جملات مثبت	۱۳۳۲
تعداد جملات منفی	۲۵۲۴

به‌منظور ارزیابی کیفیت پیکرهٔ ساخته شده، از مجموعه داده‌ای مورد استفاده برای ساخت پیکره (بخش ۳-۱) زیر مجموعه‌ای شامل ۸۲ نظر انتخاب شده است؛ سپس از دو شخص خواسته شده است که نظرات غیرمستقیم این مجموعه را استخراج کرده و به آن‌ها برچسب قطبیت (مثبت یا منفی) اختصاص دهند. در مواردی که این دو شخص اختلاف نظر داشتند شخص سومی قطبیت نهایی نظر را تعیین کرده است. بدین ترتیب مجموعه‌ای شامل ۳۰۷ نظر غیرمستقیم ایجاد شده است. در پایان پیکرهٔ ساخته‌شده با روش پیشنهادی با این پیکرهٔ دست‌ساز مقایسه شده است. جدول (۵) دقت روش پیشنهادی برای شناسایی نظرات غیرمستقیم و نیز روش پیشنهادی برای تعیین قطبیت نظرات را نشان می‌دهد. در ستون سوم جدول (۵) نیز دقت پیکرهٔ ساخته‌شده گزارش داده شده است.

(جدول-۵): دقت روش پیشنهادی برای ساخت پیکره

دقت روش شناسایی نظرات غیرمستقیم	دقت روش تعیین قطبیت	دقت پیکره
۸۵/۰۵	۹۰/۵۶	۷۴/۴۲

همان‌طورکه در جدول (۵) مشاهده می‌شود، دقت الگوریتم تعیین قطبیت ۹۰/۵۶ درصد است. درحالی‌که دقت یک روش پایه که تنها با توجه به ساختار نظرات به آن‌ها برچسب قطبیت اختصاص می‌دهد، ۸۳/۴۵ درصد است. درواقع، روش پایه به‌ترتیب با دقت ۷۹/۸۷ و ۸۷/۳۲ درصد، نظرات مثبت و منفی را تعیین قطبیت می‌کند. بنابراین روش پیشنهادی برای تعیین قطبیت بر روشی که به نظرات بخش

Archive of SID

فواید قطبیت مثبت و به نظرات بخش عوارض قطبیت منفی تخصیص می‌دهد، برتری دارد.

۵-۳- ارزیابی روش پیشنهادی برای به‌کارگیری پیکره

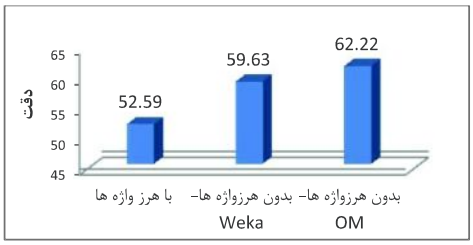
در این بخش کارآیی روش پیشنهادی برای به‌کارگیری پیکرهٔ ساخته‌شده جهت تعیین قطبیت نظرات، مورد ارزیابی قرار می‌گیرد. آزمایش‌های انجام‌گرفته بدین منظور به سه دسته تقسیم می‌شوند: ارزیابی پیش‌پردازش‌های انجام‌شده بر روی پیکره، ارزیابی اثربخشی ویژگی‌های معرفی‌شده برای تحلیل نظرات غیرمستقیم و ارزیابی الگوریتم تعیین قطبیت. در ادامه، به معرفی و شرح نتایج این آزمایش‌ها می‌پردازیم.

۵-۳-۱- ارزیابی پیش‌پردازش‌های انجام‌شده بر روی پیکره

همان‌طورکه در بخش (۴-۱) گفته شد، پیش‌پردازش‌های متعددی بر روی مجموعهٔ آموزش انجام گرفته‌اند. در ابتدا اثر هریک از این پیش‌پردازش‌ها بررسی می‌شود. در این آزمایش‌ها تنها از ویژگی تک‌واژه‌ها استفاده شده است.

در آزمایش نخست، تأثیر حذف هرزواژه‌ها بررسی شده است. بدین منظور دسته‌بند بیزین ساده (NB)[1] در نرم‌افزار وکا[2] (هال و همکاران، ۲۰۰۹) یک‌بار بر روی مجموعه آموزش با وجود هرزواژه‌ها و بار دیگر بدون هرزواژه‌ها اجرا شده است. به‌منظور حذف هرزواژه‌ها از دو فهرست هرزواژه استفاده شده است: فهرست موجود در وکا و فهرستی از هرزواژه‌های به‌کارگرفته‌شده به‌منظور نظرکاوی در دامنهٔ فیلم[3]. نتایج به‌دست‌آمده در شکل (۳-الف) ارائه شده‌اند. همان‌طورکه در شکل دیده می‌شود، حذف هرزواژه‌ها با استفاده از فهرست خاص نظرکاوی در دامنهٔ فیلم دقت را به‌میزان ۹/۶ درصد افزایش می‌دهد. دقت عبارت است از درصدی از مثال‌های مجموعهٔ آزمون که به‌درستی دسته‌بندی شده‌اند. با توجه به نتایج به‌دست‌آمده در آزمایش‌های بعدی هرزواژه‌ها حذف شده‌اند.

در آزمایش بعدی یک‌بار دسته‌بند به حروف بزرگ و کوچک حساس است و بار دیگر تنها تک‌واژه‌ها با حروف کوچک در نظر گرفته شده‌اند. همان‌طورکه در شکل (۳-ب) دیده می‌شود، حساس‌نبودن به حروف کوچک و بزرگ باعث افزایش دقت دسته‌بند درحدود ۰/۴ درصد می‌شود.



الف: بررسی اثر حذف هرزواژه‌ها

ب: بررسی حساس بودن به حروف کوچک و بزرگ

ج: بررسی اثر ریشه‌یابی

(شکل-۳): اثر پیش‌پردازش‌های انجام گرفته بر روی دقت دسته‌بندی

در آزمایش سوم اثر ریشه‌یابی توکن‌های مجموعه آموزش بررسی شده است. شکل (۳-ج) نشان می‌دهد که ریشه‌یابی دقت دسته‌بند را ۴/۵ درصد افزایش می‌دهد. در آزمایش‌های بعدی پیش‌پردازش‌های حذف هرزواژه‌ها، تبدیل توکن‌ها به حروف کوچک و ریشه‌یابی انجام شده‌اند.

۵-۳-۲- ارزیابی اثربخشی ویژگی‌های معرفی‌شده برای تحلیل نظرات غیرمستقیم

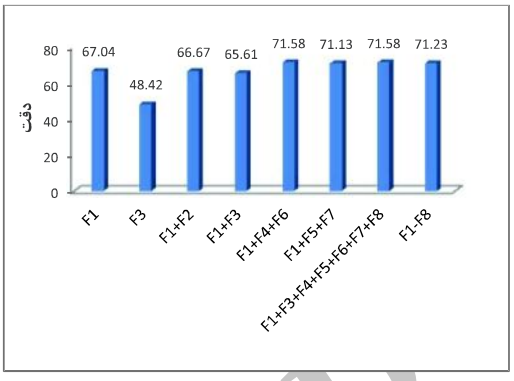
در سری بعدی آزمایش‌ها تأثیر ویژگی‌های معرفی‌شده بررسی شده است. شکل (۴) دقت دسته‌بندهای بیزین و ماشین بردار پشتیبان (SVM)[4] را با استفاده از ویژگی‌های مختلف نشان می‌دهد. برای پیاده‌سازی ماشین بردار

[1] Naïve Bayes
[2] Weka
[3] https://github.com/kuberkaul/SentimentAnalysis-MovieReviews/blob/master/sentiment%20analysis%28using%20naive%20bayes%29/src/stopwords.txt

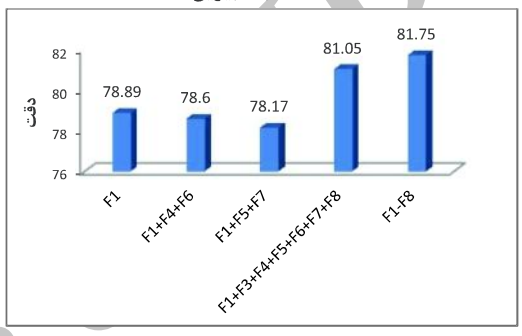
[4] Support Vector Machine

Archive of SID
www.SID.ir

پشتیبان نیز از ابزار وکا با کرنل چندجمله‌ای استفاده شده است.



الف: دسته‌بند بیزین



ب: دسته‌بند SVM

(شکل-۴): اثر به‌کارگیری ترکیب‌های مختلف از ویژگی‌های معرفی‌شده در دقت الگوریتم تعیین قطبیت

در شکل (۴-الف) دو حالت پایه در نظر گرفته شده است: زمانی‌که 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 می‌باشد، نیز استفاده شده است. علاوه‌بر آن ترکیب تک‌واژه‌ها با ویژگی‌های معنایی معرفی‌شده نیز به‌کارگرفته شده است. همان‌طورکه در شکل مشاهده می‌شود استفاده از کلیهٔ ویژگی‌های معرفی‌شده در جدول (۳) به‌جز ویژگی زوج‌واژه‌ها باعث افزایش ۴/۵ درصدی و به‌کارگیری کلیهٔ ویژگی‌های معرفی‌شده در جدول (۳) باعث افزایش ۴/۲ درصدی دقت دسته‌بند بیزین در مقایسه با حالت پایه استفاده از تک‌واژه‌ها شده است.

در شکل (۴-ب) نیز اثر ترکیب ویژگی‌های معنایی معرفی‌شده با حالت پایه تک‌واژه‌ها در دسته‌بند ماشین بردار پشتیبان بررسی شده است. همان‌طورکه مشاهده می‌شود افزودن ویژگی‌های معنایی دقت دسته‌بند را به‌طور قابل‌ملاحظه‌ای افزایش می‌دهد. در دسته‌بند ماشین بردار پشتیبان، بیشترین دقت با به‌کارگیری کلیهٔ ویژگی‌های معرفی‌شده در جدول (۳) حاصل شده است.

۵-۳-۳- ارزیابی الگوریتم تعیین قطبیت

برای تعیین قطبیت نظرات مجموعهٔ آزمون، یک الگوریتم دسته‌بندی با آموزش بر روی پیکرهٔ پردازش‌شده و با به‌کارگیری ویژگی‌های معرفی‌شده در جدول (۳) به یادگیری می‌پردازد. بدین منظور می‌توان از الگوریتم‌های دسته‌بندی و روش‌های وزن‌دهی مختلف استفاده کرد. در این راستا، در ابتدا آزمایش‌هایی جهت ارزیابی دقت دسته‌بندهای مختلف و نیز بررسی اثر وزن‌دهی به ویژگی‌ها انجام گرفته است؛ سپس دقت الگوریتم پیشنهادی برای تعیین قطبیت با دقت یک روش موجود تحلیل نظرات مستقیم مقایسه شده است.

در آزمایش نخست دقت دسته‌بندهای مختلف با یکدیگر مقایسه شده است (شکل ۵). دلیل انتخاب این دسته‌بندها استفادهٔ مکرر آن‌ها در تعیین قطبیت نظرات مستقیم و رضایت‌بخش‌بودن نتایج حاصل از آن‌ها است. این دسته‌بندها با استفاده از ترکیب همهٔ ویژگی‌های نامبرده در جدول (۳) اجرا شده‌اند. شکل (۵) نشان می‌دهد که SVM به نتایج بهتری دست یافته است. همان‌طورکه در جدول (۳) مشاهده می‌شود، اندازهٔ بردار ویژگی‌های مورد استفاده بزرگ، ولی فضای برداری متون خلوت است. در پژوهش‌های پیشین نیز نشان داده شده است که الگوریتم‌های افزایشی مانند SVM که بایاس استنتاجی دارند، عملکرد خوبی با ماتریس‌های خلوت نشان می‌دهند (کیوینن، ۱۹۹۵).

(شکل-۵): مقایسهٔ دقت دسته‌بندهای مختلف در تعیین قطبیت نظرات غیرمستقیم

در آزمایش‌های قبلی، حضور/عدم حضور تک‌واژه‌ها و زوج‌واژه‌ها به‌عنوان ویژگی‌های دودویی در نظر گرفته شده است. این دو ویژگی به‌عنوان مبنا در بسیاری از کارهای

پیشین مورد استفاده قرار گرفته‌اند. در جدول (۶) نتایج دو آزمایش دیگر ارائه شده است. در آزمایش نخست به‌جای ویژگی‌های دودویی از ویژگی‌های وزن‌دار استفاده شده است. وزن یک ویژگی با معیار TF-IDF[1] تعیین شده است. جدول (۶) نشان می‌دهد که وزن‌دهی به ویژگی‌ها دقت الگوریتم‌های یادگیری ماشین را افزایش داده است.

در آزمایش دوم برای کاهش فضای ویژگی تنها تک‌واژه‌ها و زوج‌واژه‌های پرتکرار انتخاب شده‌اند. در این آزمایش حداقل تعداد تکرار یک ویژگی پنج انتخاب شده است. عدد ۵ به‌عنوان حداقل تعداد تکرار ویژگی در بسیاری از کارهای پیشین (هابرنال و همکاران ۲۰۱۴) مورد استفاده قرار گرفته است. همان‌طورکه در جدول (۶) مشاهده می‌شود، کاهش فضای ویژگی دقت دسته‌بندها را کاهش داده است.

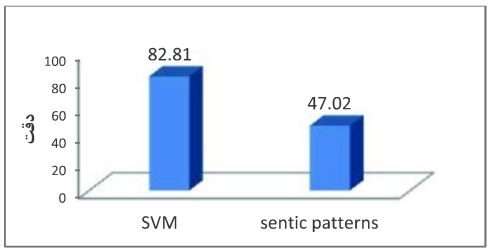
(جدول-۶): بررسی تأثیر وزن‌دهی به ویژگی‌ها و نیز به‌کارگیری ویژگی‌های پرتکرار در دقت الگوریتم تعیین قطبیت

دسته‌بند	بدون وزن‌دهی و حذف ویژگی‌های کم رخداد	TF-IDF	ویژگی‌های پرتکرار
NB	۷۱/۲۳	۷۳/۳۳	۷۰/۸۸
SVM	۸۱/۷۵	۸۲/۸۱	۸۱/۰۵

در آزمایش پایانی نتایج به‌دست‌آمده توسط دسته‌بند SVM با وزن‌دهی به ویژگی‌ها با نتایج حاصل از یکی از روش‌های برجستهٔ موجود در تحلیل نظرات مستقیم به نام الگوهای سنتیک[2] (پوریا و همکاران ۲۰۱۴) مقایسه شده است. این روش که متکی بر قواعد زبانی، محاسبات عرفی[3] و یادگیری ماشین است، از سنتیک‌نت[4] که یک پایگاه دانش از مفاهیم عرفی قطبی است، استفاده می‌کند. اگرچه در سنتیک‌نت بسیاری از مفاهیم رایج در دامنهٔ پزشکی مانند "آکنه"، "معده درد"، "افزایش وزن" و "کاهش استرس" وجود دارند که می‌توانند به‌درستی برخی از نظرات غیرمستقیم را تعیین قطبیت کنند، با این وجود این روش در تعیین قطبیت اغلب نظرات غیرمستقیم با شکست مواجه می‌شود. دلیل اصلی این امر نیز این است که الگوهای سنتیک بیشتر متکی بر مفاهیم به‌کار رفته در جمله عمل می‌کند و در اغلب موارد اثر فعل جمله را نادیده می‌گیرد.

[1] Term Frequency - Inverse Document Frequency
[2] sentic patterns
[3] common sense computing
[4] www.sentic.net

در شکل (۶) مشاهده می‌شود که روش پیشنهادی در مقایسه با الگوهای سنتیک به نتایج بهتری دست یافته است. با توجه به این‌که الگوهای سنتیک یک روش برجستهٔ موجود در زمینهٔ تحلیل نظرات مستقیم است، می‌توان نتیجه گرفت روش‌های موجود نظرکاوی در زمینهٔ تعیین قطبیت نظرات غیرمستقیم، ضعیف عمل می‌کنند. به بیانی دیگر مسألهٔ تعیین قطبیت نظرات غیرمستقیم نیازمندی‌های خاص خود را دارد و برای رسیدن به نتایج قابل قبول باید این نیازمندی‌ها مورد توجه قرار گیرند.



(شکل-۶) : مقایسهٔ روش پیشنهادی با الگوهای سنتیک

۶- نتیجه‌گیری

در این مقاله در ابتدا روشی برای ساخت نیمه‌خودکار یک پیکره از نظرات غیرمستقیم ارائه شد. با به‌کارگیری این روش یک پیکره شامل ۳۸۵۶ نظر غیرمستقیم دارای برچسب مثبت/منفی ایجاد شد؛ سپس از این پیکره به‌عنوان مجموعهٔ آموزش در روش‌های یادگیری ماشین برای دسته‌بندی نظرات یک مجموعهٔ آزمون استفاده شد. بدین منظور اثر پیش‌پردازش‌های مختلف بر روی پیکرهٔ مورد بررسی قرار گرفت. همچنین مجموعه‌ای از ویژگی‌های لغوی، نحوی و معنایی در چهار دسته‌بند بیزین ساده، ماشین بردار پشتیبان، J48 و بیشینه آنتروپی به‌کار گرفته شد.

نتایج آزمایش‌های انجام‌گرفته به‌منظور ارزیابی روش پیشنهادی نشان می‌دهد که به‌کارگیری ترکیب ویژگی‌های معرفی‌شده، دقت الگوریتم تعیین قطبیت 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 قطبیت نظرات مجموعهٔ آزمون به نتایج بسیار بهتری در مقایسه با یک روش برجستهٔ موجود در تعیین قطبیت نظرات مستقیم به نام الگوهای سنتیک می‌رسد.

Archive of SID
www.SID.ir

۷- مراجع

Aronson, A. R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Symposium, 2001.

Balahur, A., Hermida, JM., Montoyo, A., and Muñoz, R. EmotiNet: a knowledge base for emotion detection in text built on the appraisal theories. Natural Language Processing and Information Systems: Springer, 2011, pp. 27-39.

Balahur, A., Hermida, JM., and Montoyo, A. Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model. IEEE Transactions on Affective Computing, 2012, Vol. 3, pp. 88-101.

Boiy, E., and Moens, M. F. A machine learning approach to sentiment analysis in multilingual Web texts. Information retrieval, 2009, 12(5), pp. 526-558.

Bosco, C., Patti, V., and Bolioli, A. Developing corpora for sentiment analysis and opinion mining: the case of irony and senti-tut. IEEE Intelligent Systems, 2013.

Cambria, E., and Hussain, A. Sentic computing: Springer, 2012.

De Marneffe, M.-C., MacCartney, B., and Manning, C. D. Generating typed dependency parses from phrase structure parses. Proceedings of LREC, 2006.

Esuli, A., and Sebastiani, F. Sentiwordnet: A publicly available lexical resource for opinion mining. Proceedings of LREC, 2006.

Go, A., Bhayani, R., and Huang, L. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009, pp. 1-12

Goeuriot, L., Na, J.-C., Min Kyaing, W. Y., Khoo, C., Chang, Y.-K., Theng, Y.-L., and Kim, J.-J. Sentiment lexicons for health-related opinion mining. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, 2012.

Habernal, I., Ptáček, T., and Steinberger, J. Supervised sentiment analysis in Czech social media. Information Processing & Management, 2014, 50(5), pp. 693-707 .

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 2009, 11(1), pp. 10-18 .

Hu, M., and Liu, B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004.

Huang, S., Niu, Z., and Shi, C. Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. Knowledge-Based Systems, 2014, 56, pp. 191-200 .

Kaji, N., and Kitsuregawa, M. Automatic construction of polarity-tagged corpus from HTML documents. Proceedings of the COLING/ACL on Main conference poster sessions, 2006.

Kivinen, J., and Warmuth, M. K. The perceptron algorithm vs. winnow: linear vs. logarithmic mistake bounds when few input variables are relevant. Proceedings of the eighth annual conference on Computational learning theory. ACM, 1995.

Lee, H., Chang, A. ,.Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics, 2013, 39(4), pp. 885-916 .

Liu, B. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 2012, 5(1), pp. 1-167 .

Liu, K.-L., Li, W.-J., and Guo, M. Emoticon Smoothed Language Models for Twitter Sentiment Analysis. AAAI, 2012.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. Distant supervision for relation extraction without labeled data. 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, 2009, Vol. 2.

Moraes, R., Valiati, J. F., and GaviãO Neto, W. P. Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 2013, 40(2), pp. 621-633 .

Musat, C., and Trausan-Matu, S. A comparative study of the relevance of indirect and direct opinions in economic texts. Annals of DAAAM & Proceedings, 2010 .

Musat, C., and Trausan-Matu, S. The impact of valence shifters on mining implicit economic opinions. Artificial Intelligence: Methodology, Systems, and Applications, Springer, 2010, pp. 131-140.

Nakagawa, T., Inui, K., Kurohashi, S. Dependency tree-based sentiment classification using CRFs with hidden variables. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. SentiFul: A lexicon for sentiment analysis. Affective Computing, IEEE Transactions, 2011, 2(1), pp. 22-36 .

Noferesti, S., and Rajaei, M. A Hybrid algorithm based on ant colony system and learning automata for solving Steiner tree problem. International Journal of Applied Mathematics and Statistics™, 2011, 22 (S11), pp. 79-88 .

Archive of SID

Noferesti, S., and Shamsfard, M. Resource construction and evaluation for indirect opinion mining of drug reviews. PLoS One, 2015, 10(5).

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

Poria, S., Cambria, E., Winterstein, G., and Huang, G.-B. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. Knowledge-Based Systems., 2014

Sharma, A., Swaminathan, R., and Yang, H. A verb-centric approach for relationship extraction in biomedical text. Semantic Computing (ICSC), 2010.

Yu, B., Diermeier, D., and Kaufmann, S. The Wal-Mart Corpus: A multi-granularity corporate opinion corpus for opinion retrieval, classification and aggregation, working paper.

سمیرا نوفرستی در سال ۱۳۸۲ دورهٔ کارشناسی خود را رشتهٔ مهندسی کامپیوتر در دانشگاه صنعتی شریف به پایان رسانده است. در سال ۱۳۸۴ مدرک کارشناسی‌ارشد خود را در همان رشته از دانشگاه صنعتی امیرکبیر اخذ کرده است. در حال حاضر دانشجوی دکترای دانشگاه شهیدبهشتی و عضو هیئت علمی دانشگاه سیستان و بلوچستان است.

نشانی رایانامهٔ ایشان عبارت است از:

snoferesti@ece.usb.ac.ir

مهرنوش شمس فرد مدرک کارشناسی و کارشناسی ارشد ایشان در رشتهٔ مهندسی کامپیوتر از دانشگاه صنعتی شریف گرفته شده است. مدرک دکترای ایشان در گرایش هوش مصنوعی در سال ۱۳۸۱ از دانشگاه صنعتی امیرکبیر اخذ شده است. ایشان در حال حاضر عضو هیئت علمی دانشگاه شهید بهشتی و سرپرست آزمایشگاه پردازش زبان طبیعی در دانشکدهٔ مهندسی و علوم کامپیوتر هستند.

نشانی رایانامهٔ ایشان عبارت است از:

m-shams@sbu.ac.ir

Archive of SID

www.SID.ir