مدیریت فناوری اطلاعات
دانشکدهٔ مدیریت دانشگاه تهران

دورهٔ ۷، شمارهٔ ۲
تابستان ۱۳۹۴
ص. ۳۶۲- ۳۴۵

ارائهٔ روش نظارتی برای نظرکاوی در زبان فارسی با استفاده از لغت‌نامه و الگوریتم SVM

سعیده علی‌مردانی[۱]، عبدالله آقایی[۲]

چکیده: به‌سبب رشد سریع شبکه‌ها و رسانه‌های اجتماعی، امکان دسترسی افراد به نظرهای دیگران افزایش یافته است. نظرها، حاوی اطلاعات ارزشمندی‌اند که با تحلیل آنها، می‌توان به گرایش‌ها و ترجیح افراد پی برد و نظرهای مثبت و منفی را نسبت به مسائل گوناگون، شناسایی کرد. نظرکاوی فرایندی است که به تحلیل عاطفه‌ها، احساس‌ها و نظرهای افراد می‌پردازد و از این طریق، اولویت افراد را شناسایی می‌کند. در این مقاله، روشی برای نظرکاوی در زبان فارسی ارائه شده است که از ترکیب لغت‌نامه و الگوریتم نظارتی ماشین بردار پشتیبان (SVM) استفاده می‌کند. برای ایجاد لغت‌نامه، از لغت‌نامهٔ SentiWordNet بهره برده شده است. در واقع این لغت‌نامه، مجموعهٔ ویژگی‌های الگوریتم SVM است. برای ارزیابی نتایج، از داده‌های دامنهٔ هتل استفاده شد. چهار فرضیه برای دستیابی به بهترین نتیجه تعریف شد که از این بین، بیشترین درستی، به فرضیهٔ حاصل‌ضرب قطبیت در تعداد تکرار کلمه‌ها اختصاص یافت.

واژه‌های کلیدی: قطبیت، لغت‌نامه، ماشین بردار پشتیبان، نظرکاوی.

۱. دانشجوی کارشناسی ارشد فناوری اطلاعات، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران
۲. استاد گروه برنامه‌ریزی و تحلیل سیستم‌ها، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران

تاریخ دریافت مقاله: ۱۳۹۳/۰۳/۲۱
تاریخ پذیرش نهایی مقاله: ۱۳۹٤/۰۲/۰۷
نویسندهٔ مسئول مقاله: سعیده علی‌مردانی
E-mail: sa1.alimardani@gmail.com

مقدمه

نظرکاوی[1] زمینه‌ای است که به مطالعهٔ نظرها[2]، احساس‌ها، ارزیابی‌ها، رفتار و عواطف افراد نسبت به موجودیت‌هایی مانند محصولات، افراد، سازمان‌ها، موضوعات، حوادث و صفات آنها می‌پردازد (لیو، ۲۰۱۲). امروزه با توجه به رشد سریع شبکه‌ها و رسانه‌های اجتماعی، دسترسی به نظرهای افراد دربارهٔ مسائل گوناگون افزایش یافته است. کاربران به‌راحتی می‌توانند از نظرها و اولویت‌های افراد آگاه شوند. استفاده و تحلیل دستی این نظرها ممکن نیست؛ زیرا حجم آنها روزبه‌روز در حال افزایش است. اینجاست که بحث نظرکاوی ضرورت می‌یابد و با تحلیل خودکار نظرها، اطلاعات ارزشمندی که در آنها نهفته است، کشف می‌شود. نظرها که در قالب داده‌های متنی جمع‌آوری می‌شوند، حاوی اطلاعات ارزشمندی‌اند که با توجه به دامنهٔ کاربرد، می‌توانند اطلاعات مفیدی را برای افراد و سازمان‌ها فراهم کنند. هدف از نظرکاوی، تعیین قطبیت[3] نظرها یا به بیانی دیگر، مثبت و منفی‌بودن هر نظر است. الگوریتم‌های متعددی در زمینهٔ نظرکاوی توسعه‌یافته که هر یک در بهبود مسئلهٔ نظرکاوی تلاش کرده است.

اگرچه در زبان فارسی به نظرکاوی کمتر توجه شده است؛ در زبان‌های دیگر از جمله انگلیسی (البِرنوز، پِلازا، گِروِس و دیاز، ۲۰۱۱؛ پَنگ و لی، ۲۰۰۲ و ترنِی، ۲۰۰۲)، چینی (وِی و پال، ۲۰۱۰؛ وَن، ۲۰۰۸ و ۲۰۰۹ و یو و ما، ۲۰۰۸) و اسپانیایی (بِروک، تُفیلُسکی، تُبادوآ، ۲۰۰۹ و گونزالِز، کَمارا، وَلدویویا و اورتِقا، ۲۰۱۳) در کانون توجه قرار دارد. بنابراین به‌دلیل پژوهش‌های اندک در این زمینه، در مقالهٔ پیش رو به نظرکاوی در زبان فارسی پرداخته‌ایم.

در این مقاله، از ترکیب الگوریتم نظارتی SVM[4] همراه با لغت‌نامه‌ای[5]، برای نظرکاوی در سطح سند استفاده شده است. برای ایجاد لغت‌نامه، از لغت‌نامهٔ SentiWordNet (باسیانلا، ایسولی، سباستیانی، ۲۰۱۰) بهره برده شده است. از مجموعهٔ لغت‌های این لغت‌نامه، به‌منزلهٔ ویژگی‌های ورودی دسته‌بندی‌کننده[6] استفاده می‌شود.

ساختار مقاله به این ترتیب است؛ در بخش دوم، کارهای انجام‌شده در زمینهٔ نظرکاوی و نظرکاوی در زبان فارسی، بررسی شده است. بخش سوم، به تشریح مراحل نظرکاوی می‌پردازد. در قسمت چهارم نظرکاوی و پیاده‌سازی آن با استفاده از الگوریتم SVM تجزیه‌وتحلیل می‌شود. بخش پایانی نیز به نتیجه‌گیری و پیشنهادهای آتی اختصاص دارد.

---

1. Opinion mining
2. Opinion
3. Opinion Orientation
4. Supervised
5. Lexicon
6. Classifier

پیشینهٔ پژوهش

تاکنون روش‌های متعددی برای نظرکاوی معرفی شده است که شامل روش‌های نظارتی و غیرنظارتی[1] می‌شود. برخی از این روش‌ها، از روابط معنایی بین کلمه‌ها و نقش دستوری کلمه‌ها استفاده کرده‌اند و برخی دیگر از لغت‌نامه‌ها برای این منظور بهره برده‌اند. در نظرکاوی بر ویژگی‌های متعددی تمرکز شده است. در ادامه به بررسی تعدادی از کارهای انجام‌گرفته در نظرکاوی و نظرکاوی در زبان فارسی پرداخته می‌شود.

یکی از روش‌های اولیه‌ای که در سطح سند در نظرکاوی انجام می‌گیرد، دسته‌بندی[2] نظرها در دو گروه مثبت و منفی بر اساس میانگین قطبیت عبارت‌های سند است. به‌طور معمول عبارت‌هایی به‌کار برده می‌شوند که صفت و قید باشند (تِرنی، ۲۰۰۲). روش‌های یادگیری ماشین، مانند بیزین ساده[3]، SVM و آنتروپی بیشینه[4]، با فرضیه‌های گوناگونی برای نظرکاوی در سطح سند بررسی شده‌اند. در این روش‌ها، محققان از دو مجموعهٔ مثبت و منفی استفاده کرده‌اند که هر یک هفت عضو دارد. استفاده از ویژگی حضورداشتن و حضورنداشتن[5] کلمه‌ها و تک‌کلمه‌ای[6] بودن، برای SVM بهترین نتیجه را درپی داشته است (پَنگ و لی، ۲۰۰۲). ژانگ و همکاران سه روش SVM، بیزین ساده و روش مبتنی بر کاراکتر[7] را با هم مقایسه کردند و دریافتند دو الگوریتم SVM و مبتنی بر کاراکتر، نتایج بهتری را نشان می‌دهد (یه، ژانگ و لا، ۲۰۰۹).

از خوشه‌بندی[8] نیز برای نظرکاوی استفاده شده است (البِرنوز و همکاران، ۲۰۱۱؛ لی و لیو ۲۰۱۲ و لی و لیو، ۲۰۱۳). روش‌های مبتنی بر خوشه‌بندی، برخلاف روش‌های نظارتی و نیمه‌نظارتی، بدون دخالت انسان، بدون نیاز به دانش زبان‌شناسی و بدون صرف زمان یادگیری، اجرا می‌شود (لی و لیو، ۲۰۱۳).

SO-CAL[9] روشی است که از لغت‌نامه‌ای شامل قطبیت لغات و شدت آنها استفاده می‌کند. در این روش، منفی‌کننده‌ها[10] و تشدیدکننده‌ها[11] نیز در نظر گرفته می‌شوند (تابوآدا، بروک،

---

1. Unsupervised
2. Classification
3. Naive Bayes
4. Maximum-entropy
5. Present-absent
6. Unigram
7. Character based
8. Clustering
9. Semantic Orientation Calculator
10. Negation
11. Intensifier

توفیلوسکی، وُل و اِستده، ۲۰۱۱). یکی از روش‌های دیگر نظرکاوی، به‌کاربردن عنوان[1] در نظرکاوی است. برای این منظور، روشی ترکیبی و غیرنظارتی در سطح سند ارائه شده است (لین و هی، ۲۰۰۹). در روش مطرح‌شدهٔ دیگری برای نظرکاوی، از SentiWordNet یا ترکیبی از SentiWordNet و SVM استفاده شده است.

محققان برای نظرکاوی از روشی بهره بردند که با استفاده از SentiWordNet اجرا می‌شود (هانگ و لین، ۲۰۱۳). در این روش کلمه‌هایی که قطبیت آنها در SentiWordNet نیست، بر اساس اینکه بیشتر در جمله‌های منفی یا مثبت حضور دارند، تعیین قطبیت می‌شوند. در این حالت درستی[2] الگوریتم نسبت به زمانی که از این کلمه‌ها استفاده نمی‌شود، بهبود می‌یابد. رائز و دیگران روشی را با ترکیب SentiWordNet و پیمایش تصادفی[3] ارائه دادند (رائِز، کَمرا، والدیویا و لُوپز، ۲۰۱۴). در این روش، الگوریتم پیمایش تصادفی، وزنی را با استفاده از SentiWordNet به مجموعه‌های مترادف می‌دهد. مارتینا و همکارش در روشی دیگر، از ویژگی Delta TFIDF[4] برای SVM در نظرکاوی استفاده کردند. این ویژگی، درواقع تفاوت TFIDF[5] هر کلمه در داده‌های آموزشی منفی و مثبت را نشان می‌دهد. به این ترتیب، کلمه‌هایی که به‌طور نامساوی در هر دو داده‌های آموزشی مثبت و منفی پراکنده شده باشند، متمایز می‌شود و کلمه‌هایی که به‌طور مساوی وجود دارند، نادیده گرفته می‌شود (مارتینا و فینین، ۲۰۰۹).

روش‌هایی نیز برای بهبود الگوریتم SVM در نظرکاوی ارائه شده است. برای مثال، بصری و همکارانش روشی با ترکیب دو الگوریتم SVM و بهینه‌سازی ازدحام ذرات[6] پیشنهاد کردند که از الگوریتم بهینه‌سازی ازدحام ذرات برای بهبود پارامترهای SVM بهره می‌برد (بَصری، حُسین، آنانتا و زِنیارجا، ۲۰۱۳). ویندهینی و همکارش (۲۰۱۴) ترکیب الگوریتم SVM و تحلیل مؤلفهٔ اصلی را برای بهبود الگوریتم SVM به‌منظور دسته‌بندی نظرها پیشنهاد دادند. آنها از تحلیل مؤلفه‌های اصلی به‌منظور کاهش ابعاد داده‌ها استفاده کردند و بدین ترتیب پیچیدگی SVM را کاهش دادند.

در زبان فارسی نیز پژوهش‌های اندکی در زمینهٔ نظرکاوی انجام گرفته است. اولین پژوهش نظرکاوی برای زبان فارسی به روش غیرنظارتی بود و با بهره‌مندی از الگوریتم LDA[7] و

1. Topic
2. Accuracy
3. Random walk
4. Delta Term Frequency–Inverse Document Frequency
5. Term Frequency–Inverse Document Frequency
6. Particle swarm optimization
7. Latent Dirichlet allocation

لغت‌نامه انجام گرفت (شمس، شاکری و فیلی، ۲۰۱۲)؛ در واقع، لغت‌نامه به جای ویژگی‌های LDA به‌کار رفته است؛ درستی استفاده از این روش حدود ۸۰ درصد گزارش شد. روش دیگر به‌صورت نظارتی و با به‌کارگیری الگوریتم SVM صورت گرفت (حاج محمدی و ابراهیم، ۲۰۱۳). در این روش، الگوریتم SVM با استفاده از ویژگی حضورداشتن و حضورنداشتن، به درستی حدود ۷۲ درصد دست یافت. پژوهش انجام‌شدهٔ دیگری در زمینهٔ نظرکاوی زبان فارسی، بر انتخاب ویژگی برای نظرکاوی تمرکز کرد (سارایی و باقری، ۲۰۱۳). در این پژوهش، ویژگی جدید اطلاعات مشترک اصلاح‌شده (MMI)[1] معرفی شده است.

لغت‌نامه

لغت‌نامه‌های زیادی برای تعیین قطبیت کلمه‌ها وجود دارد. هر یک از این لغت‌نامه‌ها، سازوکار خاصی را برای تعیین قطبیت استفاده می‌کند و روش خاصی را برای نشان‌دادن قطبیت کلمه‌ها به‌کار می‌برد. برای مثال، لغت‌نامهٔ بینگ لیو[2] شامل ۲۰۰۶ لغت مثبت و۴۷۸۳ لغت منفی است (هو و لیو، ۲۰۰۴). در این لغت‌نامه، به مقدار مثبت و منفی‌بودن کلمه‌ها اشاره‌ای نشده است و تنها فهرستی از لغات مثبت و منفی را دربرمی‌گیرد. لغت‌نامهٔ MPQA[3] نیز شامل فهرستی از لغات مثبت و منفی است. در این لغت‌نامه علاوه‌بر قطبیت هر کلمه، اطلاعاتی مانند نقش دستوری کلمه[4] نیز به چشم می‌خورد (وِیبی، ویلسون و کاردیه، ۲۰۰۵). WordNet لغت‌نامه‌ای شامل اسم، فعل، صفت و قید است. در این لغت‌نامه هر کلمه در کنار مجموعهٔ کلمه‌های هم‌خانواده‌اش قرار دارد. برای هر کلمه نقش‌های مختلف دستوری همراه با مثالی از کاربرد آن کلمه، درج شده است (میلر، ۱۹۹۵). این لغت‌نامه مثبت و منفی‌بودن کلمات را نشان نمی‌دهد. SentiWordNet توسعه‌یافتهٔ لغت‌نامهٔ WordNet است که مقدار قطبیت هر کلمه را به‌صورت عددی نشان می‌دهد. در این لغت‌نامه، مترادف‌ها و نقش دستوری هر کلمه مشخص شده است. Harvard General Inquirer لغت‌نامه‌ای شامل کلمات مثبت و منفی است (استون، اسمیت، دانفری و آگیلویه، ۱۹۶۶). علاوه‌بر دو ویژگی مثبت و منفی‌بودن، ۱۸۲ ویژگی دیگر را نیز می‌توان در این لغت‌نامه یافت.

لغت‌نامه‌ای که در این مقاله از آن بهره برده می‌شود، SentiWordNet نام دارد. این لغت‌نامه قطبیت کلمه‌ها را به‌صورت عددی نشان می‌دهد و برای زبان انگلیسی توسعه یافته است؛ لذا برای استفاده در زبان فارسی باید تغییراتی در آن اعمال شود.

---

1. Modified Mutual Information
2. Bing liu
3. Multi-Perspective Question Answering
4. Part of speech

پیچیدگی‌های زبان فارسی

برخلاف زبان‌های دیگر از جمله زبان انگلیسی، متن‌کاوی برای زبان فارسی به‌دلیل پیچیدگی بسیار با مشکلات متعددی روبه‌رو است. از آنجاکه هدف این پژوهش نظرکاوی است، داده‌های به‌کاررفته به‌صورت نظر و به شکل محاوره‌ای نوشته شده است. پس علاوه‌بر پیچیدگی‌های متون فارسی و مشکلات تحلیلی آن، مسائلی نیز به‌دلیل محاوره‌ای‌بودن زبان، به‌وجود می‌آید که پژوهش را دشوارتر می‌کند. همان‌طور که سارایی و باقری (۲۰۱۳) نیز معتقدند نظرکاوی در زبان فارسی با مشکلاتی روبه‌رو است؛ این مشکلات به‌دلیل کمبود ابزار و راه‌های مختلف، وجود پسوندهای متفاوت، فاصله‌گذاری کلمه‌ها و استفاده از کلمه‌های غیررسمی و محاوره‌ای شکل می‌گیرد.

کمبود ابزار مناسب برای زبان فارسی: در زبان فارسی به‌منظور پیش‌پردازش و تحلیل متن‌های فارسی، ابزار زیادی وجود ندارد، اما برخلاف زبان فارسی، ابزارها و روش‌های متعددی برای نظرکاوی و تحلیل متن در زبان‌های مختلف از جمله زبان انگلیسی در دسترس است که می‌توان از آنها در مراحل مختلف نظرکاوی استفاده کرد. کمبود ابزار برای پردازش متون فارسی، نظرکاوی را دشوارتر می‌کند.

کلمه‌های غیررسمی و محاوره‌ای: وقتی نوشتار به‌صورت غیررسمی و محاوره‌ای باشد، شکل کلمه‌ها تغییر می‌کند، گاهی برای یک کلمه، اشکال مختلفی به‌شکل محاوره‌ای به‌چشم می‌خورد. ممکن است کلمه‌ها به‌شکل رسمی یا محاوره‌ای نوشته شوند که افراد هر دو شکل آن را در نظرها به‌کار می‌برند. تغییر شکل کلمه‌ها در قالب محاوره‌ای با حذف، اضافه و تغییر حروف همراه است. برای مثال کلمهٔ «نمی‌توانم» در شکل محاوره‌ای به‌صورت «نمی‌تونم» با حذف «الف» بیان می‌شود، یا کلمهٔ «دارد» با تغییر حرف «د» به «ه» به‌صورت «داره» نوشته می‌شود. کلمه‌های دیگری نیز در زبان محاوره‌ای به‌کار می‌روند که به شکل کاملاً متفاوتی از شکل رسمی آن ظاهر می‌شوند، برای مثال کلمهٔ «برای» به‌صورت «واسه» نوشته می‌شود.

پیشوندها و واژه‌های غیرساده: در زبان فارسی واژه‌ها به دو دستهٔ ساده و غیرساده تقسیم می‌شوند. کلمه‌های ساده تنها از یک جزء معنادار ساخته شده‌اند و واژه‌های غیرساده بیشتر از یک جزء معنادار، دارند. واژه‌های غیرساده در سه شکل مشتق، مرکب و مشتق ـ مرکب به‌کار می‌روند. هر واژهٔ غیرساده‌ای با افزودن پسوندها و پیشوندهای مختلفی ساخته می‌شود. مشکل کلمه‌های غیرساده، فاصله‌گذاری آن است. در شکل نوشتاری کلمه‌ها، اغلب بین اجزای مختلف واژهٔ

غیرساده، فاصله‌ای گذاشته می‌شود و این فاصله تشخیص واژه را دشوار می‌کند. برای مثال کلمهٔ «خودنویس» که در زبان فارسی واژهٔ مرکبی محسوب می‌شود، در حالت نوشتاری آن ممکن است به‌صورت «خود نویس» نوشته شود، فاصلهٔ میان خود و نویس، تشخیص آن را دشوار می‌کند.

فاصله‌گذاری: در زبان فارسی برای فاصله‌گذاری از فاصله و نیم‌فاصله استفاده می‌شود. فاصله برای جداسازی کلمه‌ها از یکدیگر است و از نیم‌فاصله برای فاصلهٔ بین اجزای مختلف یک کلمه کاربرد دارد. برای مثال، در عبارت «کتاب خوب»، دو واژهٔ متفاوت وجود دارد که برای جداسازی آنها از فاصله استفاده می‌شود، اما با اینکه عبارت‌های «خودنویس»، «خاطره‌انگیز» و «کتاب‌ها» یک واژه‌اند، آنها را با نیم‌فاصله جدا می‌کنند. در حالت محاوره‌ای، اغلب این نیم‌فاصله نادیده گرفته می‌شود، در نتیجه تشخیص چنین واژه‌هایی به دقت بیشتری نیاز دارد.

روش‌شناسی پژوهش

از آنجاکه این پژوهش به بررسی و نظرخواهی هتل‌های جزیرهٔ کیش می‌پردازد، پژوهشی کاربردی شمرده می‌شود و به‌دلیل جمع‌آوری داده‌ها از چندین هتل، در گروه پژوهش‌های موردی قرار می‌گیرد. نظرکاوی این پژوهش در سطح سند انجام می‌گیرد؛ به این معنا که نظرها به‌منزلهٔ سند در نظر گرفته می‌شوند. نظرکاوی در سطح سند، به‌معنای دسته‌بندی نظرهای متنی در دو دستهٔ مثبت و منفی است (مورائِس، وَلیاتی و نِتو، ۲۰۱۳). داده‌هایی که برای نظرکاوی در این پژوهش استفاده شده است، از نظرهای جمع‌آوری‌شدهٔ تعدادی از هتل‌های کیش به‌دست آمده است. برای نظرکاوی در این پژوهش از الگوریتم SVM که الگوریتم نظارتی است، استفاده می‌شود. SVM در برابر نویز داده‌ها مقاوم است، می‌تواند با تعداد زیادی ویژگی کار کند و در کارهای مشابه مانند دسته‌بندی متن عملکرد خوبی دارد؛ از این رو ابزار مناسبی به‌شمار می‌رود. (مارتینوا و فینین، ۲۰۰۹ و جوکیمز، ۱۹۹۸).

برای استفاده از الگوریتم SVM باید داده‌ها برچسب‌دار شوند؛ از این رو داده‌هایی که برای نظرکاوی در این مقاله جمع‌آوری شده است نیز، برچسب‌گذاری شده‌اند؛ به این صورت که هر فرد بعد از درج نظر در وب‌سایت، گزینهٔ «آیا این هتل را برای اقامت توصیه می‌کنید؟» را با بلی یا خیر کامل می‌کند و از همین گزینه برای برچسب‌گذاری نظرها استفاده می‌شود. برای بررسی بیشتر و مقایسهٔ نتایج الگوریتم SVM و دستیابی به تحلیل کامل‌تر، از الگوریتم بیزین ساده نیز استفاده می‌شود. همچنین ضمن مطرح‌کردن فرضیه‌هایی، کارایی الگوریتم در هر چهار فرضیه مقایسه می‌شود.

هدف این پژوهش تعیین مثبت و منفی‌بودن نظرها در سطح سند است؛ برای این منظور باید ویژگی‌ها را در قالب مجموعه‌ای قرار داد. در این پژوهش دسته‌ای از ویژگی‌ها در مجموعهٔ $\{f۱, \ldots, fm\}$ قرار می‌گیرد. این ویژگی‌ها، همان کلمه‌های داخل لغت‌نامه است که با استفاده از لغت‌نامه SentiWordNet ایجاد می‌شود. لغت‌نامهٔ مذکور متشکل از ۳۱۲ لغت یا به بیان دیگر ۳۱۲ ویژگی است. هر سند، مجموعه‌ای از کلمه‌ها در نظر گرفته می‌شود که می‌تواند شامل هر یک از ویژگی‌ها باشد. برای استفاده از این ویژگی‌ها در الگوریتم SVM، فرض‌های زیر مطرح شده است:

فرض‌ها

۱. تعداد تکرار کلمات: تعداد تکرار کلمه‌ها یکی از ویژگی‌هایی است که از آن استفاده می‌شود؛ بدین ترتیب که هریک از کلمه‌های لغت‌نامه در بین اسناد جست‌وجو می‌شود و تعداد تکرار هر یک به‌دست می‌آید؛

۲. حضورداشتن و حضورنداشتن هر کلمه: برای این فرض، حضور هر کلمه ویژگی محسوب می‌شود؛ به این ترتیب که اگر کلمه‌ای وجود داشته باشد، برای آن مقدار عددی ۱ در نظر گرفته می‌شود و چنانچه حضور نداشته باشد، صفر می‌گیرد. در این حالت با جست‌وجوی تمام کلمه‌های لغت‌نامه در اسناد، حضور هر یک از آنها مشخص می‌شود؛

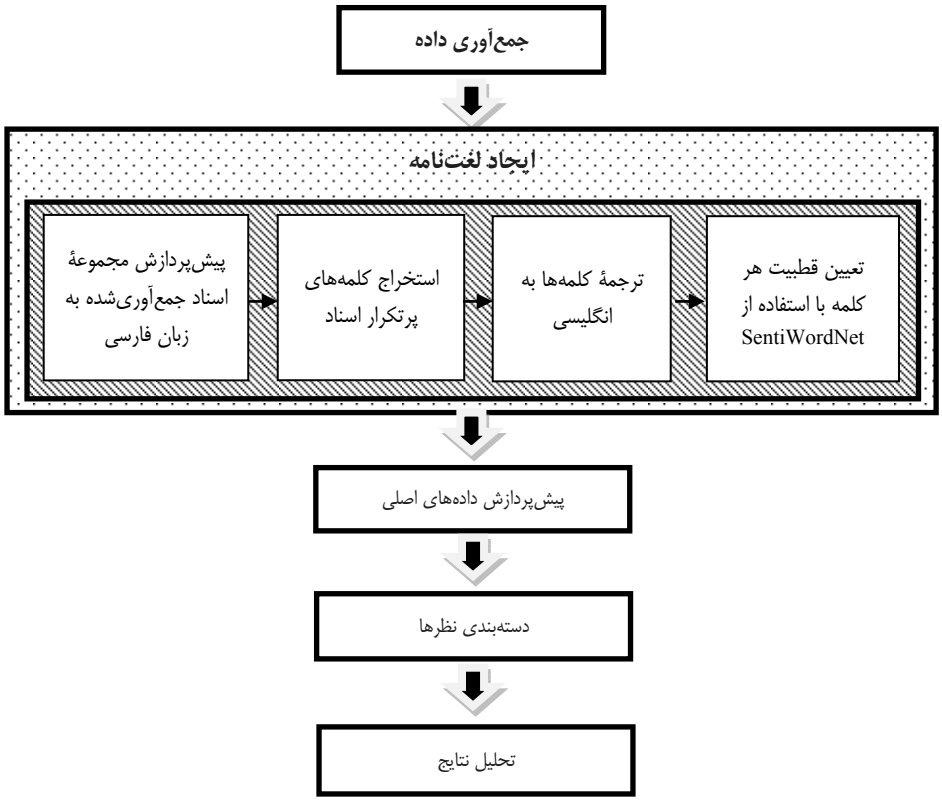
۳. حاصل‌ضرب تکرار هر کلمه در مقدار قطبیت آن: هر سند مجموعه‌ای از کلمه‌ها در نظر گرفته می‌شود؛ در نتیجه هر سند برداری از کلمه‌ها است. با جست‌وجوی هر کلمه در لغت‌نامه، قطبیت آن تعیین می‌شود و با مقدار عددی به نمایش درمی‌آید. این مقدار در تعداد تکرار هر کلمه ضرب می‌شود و وزن جدیدی را ایجاد کند. در نتیجهٔ این فرایند، برای هر کلمه وزنی ایجاد می‌شود که حاصل‌ضرب قطبیت آن کلمه در تعداد تکرارش در سند است. وزن صفر به کلمه‌هایی اختصاص دارد که در لغت‌نامه نیستند؛

۴. حاصل‌ضرب حضورداشتن و حضورنداشتن هر کلمه در مقدار قطبیت آن: ویژگی دیگر، استفاده از حضور کلمه است. بدین ترتیب، کلمه‌ای که در لغت‌نامه حضور داشته باشد، مقدار یک و کلمه‌ای که حضور نداشته باشد مقدار صفر می‌گیرد. سپس این مقدار در عدد قطبیت آن کلمه ضرب می‌شود.

به‌منظور بهره‌مندی از الگوریتم SVM، از کتابخانهٔ SMO[1] و برای الگوریتم بیزین ساده نیز از کتابخانه این الگوریتم در نرم‌افزار وکا[2] استفاده شده است.

1. Sequential Minimal Optimization
2. Weka

به‌طور کلی برای اجرای نظرکاوی مرحله‌هایی وجود دارد که گذر از هریک ضروری است. شکل ۱ این مرحله‌ها را به نمایش گذاشته است. همان‌طور که مشاهده می‌کنید، نظرکاوی از جمع‌آوری داده‌ها آغاز می‌شود و با ارزیابی نتایج، به پایان می‌رسد.



شکل ۱. مراحل اجرای پژوهش

در گام اول، داده‌ها در دو گروه جمع‌آوری می‌شود؛ گروه اول برای ایجاد لغت‌نامه به‌کار می‌رود و گروه دوم برای دسته‌بندی استفاده می‌شود. از آنجاکه ایجاد لغت‌نامه به پیش‌پردازش نیاز دارد، کلمه‌ها از داده‌های پیش‌پردازش‌شده استخراج شدند. SentiWordNet لغت‌نامه‌ای به زبان انگلیسی است، برای اینکه امکان جست‌وجو در SentiWordNet فراهم شود، واژه‌ها به انگلیسی ترجمه شدند؛ بدین ترتیب لغت‌نامه‌ای شامل واژه‌های فارسی به همراه قطبیت آنها شکل گرفت. در این مرحله برای استفاده از داده‌ها در نظرکاوی، باید داده‌های اصلی پیش‌پردازش شوند. داده‌های اصلی، همان داده‌هایی هستند که به‌کمک الگوریتم SVM

دسته‌بندی می‌شوند. برای پیش‌پردازش داده‌های اصلی، واژه‌های لغت‌نامه مبنای کار قرار گرفت. بعد از پیش‌پردازش داده‌های اصلی، به‌کمک الگوریتم SVM، واژه‌ها با در نظرگرفتن چهار دسته‌بندی شدند و اسناد در دو گروه مثبت و منفی قرار گرفتند.

یافته‌های پژوهش

در این بخش نتایج به‌کارگیری الگوریتم SVM بر داده‌های پژوهش بررسی می‌شود. همان‌طور که در بخش قبل اشاره شد پس از مطرح‌کردن چهار فرضیه، نتایج تجزیه‌وتحلیل خواهد شد.

جمع‌آوری داده‌ها

برای ارزیابی صحت روش پیشنهادی به‌منظور تعیین قطبیت نظرها، از داده‌های جمع‌آوری‌شدهٔ مربوط به هتل در تارنمای هلوکیش[1] استفاده شد. این تارنما به بررسی هتل‌های مستقر در جزیرهٔ کیش پرداخته است. داده‌های اردیبهشت ۱۳۸۹ تا فروردین ۱۳۹۳ به‌کمک نرم‌افزار خزندهٔ موزندا[2] جمع‌آوری شدند. با تنظیم این نرم‌افزار امکان جمع‌آوری خودکار نظرها فراهم می‌شود. همان‌گونه که پیش از این گفته شد، داده‌ها در دو گروه برای ایجاد لغت‌نامه و استفاده در نظرکاوی جمع‌آوری شدند. ۲۲۹۶ داده برای ایجاد لغت‌نامه و ۱۵۶۶ نظر نیز برای نظرکاوی استفاده شد. شکل ۲ نمونه‌ای از این نظرها را به نمایش گذاشته است.

شکل ۲. نمونهٔ نظر در تارنمای هلوکیش

1. www.hellokish.com
2. Mozenda web crawler

در روش‌های نظارتی داده‌ها برای دو دستهٔ مثبت و منفی برچسب‌گذاری می‌شوند، اما داده‌های جمع‌آوری‌شدهٔ این پژوهش در بطن خود چنین برچسبی را داشتند و نیازی به برچسب‌گذاری نبود.

ایجاد لغت‌نامه

در نظرکاوی برای زبان فارسی، باید به‌منظور تشخیص مثبت و منفی‌بودن کلمه‌ها، لغت‌نامه‌ای ایجاد شود. برای به‌کارگیری این لغت‌نامه در زبان فارسی می‌توان از سه راهکار بهره برد. در راهکار اول، پس از ترجمهٔ نظرهای فارسی به انگلیسی، از لغت‌نامه‌هایی به زبان انگلیسی استفاده می‌شود. در این روش لغت‌نامه تغییری نمی‌کند. راهکار دوم، ترجمهٔ لغت‌نامه به زبان فارسی است؛ بدین ترتیب که تمام واژه‌های به‌کاررفته در آن لغت‌نامه به زبان فارسی ترجمه می‌شود. راهکار سوم، ترجمهٔ تعدادی از واژه‌های پرکاربرد در دامنهٔ مد نظر به زبان انگلیسی است. با توجه به اینکه از همهٔ واژه‌ها استفاده نمی‌شود و فقط تعدادی از واژه‌های پرکاربرد در حوزهٔ داده‌های هتل مد نظر است، به ترجمهٔ تعدادی از کلمه‌های پرکاربرد به زبان انگلیسی اکتفا می‌شود. این مقاله نیز، راهکار سوم را درپیش گرفته است. در ادامه مراحل شکل‌گیری لغت‌نامه تشریح می‌شود.

مرحلهٔ اول پیش‌پردازش داده‌هاست که گام‌های متعددی را شامل می‌شود. پیش‌پردازش به‌معنای تمیزکردن و آماده‌سازی داده‌ها است. داده‌های اینترنتی معمولاً با نویزها و بخش‌های بی‌استفاده‌ای مانند برچسب‌های HTML و تبلیغات همراه‌اند. علاوه‌بر این در سطح کلمه، بسیاری از واژه‌ها تأثیری در تعیین قطب کلمه‌ها ندارند و نگه‌داشتن چنین واژه‌هایی ابعاد مسئله را افزایش می‌دهد؛ زیرا هر کلمه به‌مثابهٔ یک بعد در نظر گرفته می‌شود. در نتیجه با اعمال پیش‌پردازش، از نویزهای متن کاسته می‌شود. این فرایند شامل مراحل پاک‌سازی داده‌ها، حذف فضای خالی، گسترش کلمه‌های اختصاری، ریشه‌یابی[1]، حذف کلمات غیرمفهومی[2]، در نظرگرفتن منفی‌کننده‌ها و در انتها، انتخاب ویژگی‌هاست (هَدی، لین و شی، ۲۰۱۳).

در مقالهٔ حاضر، بیشترین پیش‌پردازش‌ها در مرحلهٔ شکل‌گیری لغت‌نامه اعمال شده است و مرحلهٔ نظرکاوی به پیش‌پردازش زیادی نیاز نداشت. روش‌های متعددی برای پیش‌پردازش متن وجود دارد که می‌توان به تناسب نیاز از همه یا تعدادی از آنها استفاده کرد. شایان ذکر است تمام مراحل پیش‌پردازش برای ایجاد لغت‌نامه به‌طور دستی انجام گرفته است.

ابتدا فقط نظرهایی که به زبان فارسی نوشته شده بودند، جدا شدند و نظرهای غیرفارسی حذف شدند. در این میان جمله‌هایی مشاهده شد که به زبان انگلیسی بود یا با حروفی به جز

1. Stemming
2. Stop word

حروف فارسی نوشته شده بود که تمام آنها پس از بررسی حذف شدند. همچنین نظرهایی به زبان فارسی وجود داشت که حاوی تعداد محدودی واژهٔ انگلیسی بود؛ این کلمه‌ها نیز پس از شناسایی، حذف شدند.

نظرهایی که فاصله و نیم‌فاصله در آن رعایت نشده بود، اصلاح شدند؛ کلمه‌هایی که به‌صورت محاوره‌ای نوشته شده بودند، به‌شکل رسمی تبدیل شدند؛ واژه‌هایی که املای درستی نداشتند، تصحیح شدند. البته شناسایی املای غلط تمام واژه‌ها امکان‌پذیر نبود و تعدادی از این کلمه‌ها مشخص و اصلاح شدند.

افعال، ضمایر و اسم‌ها ریشه‌یابی شدند. فعل‌ها می‌توانند ساده، مشتق، مرکب و عبارت فعلی باشند (بامنشین، مهدیزاده، داوطلب و پیله‌ور، ۱۳۸۹). پس از شناسایی ریشهٔ افعال، هر فعل جایگزین ریشهٔ آن شد. به‌دلیل تفاوت ریشهٔ افعال مضارع و ماضی، برای کاهش تعداد واژه‌ها، ریشهٔ افعال مضارع با ریشهٔ ماضی آن جایگزین شد. اسم‌ها نیز ممکن است به‌صورت جمع یا مفرد باشند، برای رفع این مشکل تمام پسوندهای جمع پس از شناسایی، حذف شدند و از مفرد کلمه‌ها استفاده شد. جمع‌های مکثری هم در نظرها وجود داشت که آنها نیز پس از شناسایی، به حالت مفرد تبدیل شدند. بدین ترتیب طی مراحلی که بیان شد، داده‌ها پیش‌پردازش شدند.

برای استخراج واژه‌ها از مجموعهٔ اسناد، از نرم‌افزار متن‌باز[1] وکا استفاده شد. این نرم‌افزار به‌کمک ویژگی String To Word اسناد را به مجموعه‌ای از واژه‌ها تبدیل می‌کند. وکا برای این منظور، هر سند را به‌صورت برداری از واژه‌ها به‌منزلهٔ ورودی دریافت می‌کند و مجموعه‌ای از واژه‌ها را به‌صورت خروجی تولید می‌کند. با در دست‌داشتن این مجموعه از واژه‌ها، می‌توان به آسانی واژگانی که بیشترین تکرار را دارند، استخراج کرد. در این کار پژوهشی، واژگانی که بیش از پنج بار تکرار شده بودند، انتخاب شدند و برای این کار نیز، از یکی دیگر از قابلیت‌های نرم‌افزار وکا استفاده شد؛ بدین ترتیب ۴۶۲ کلمه به‌دست آمد.

اکنون باید قطبیت کلمه‌ها مشخص شود. همان‌طور که پیش از این بیان شد، برای تعیین قطب کلمه‌ها از لغت‌نامهٔ SentiWordNet استفاده شده است. بدین ترتیب که پس از ترجمهٔ هر کلمه به انگلیسی، برای تعیین قطب آن در لغت‌نامهٔ SentiWordNet به جست‌وجو پرداخته می‌شود. هر کلمه در SentiWordNet حاوی دو قطب مثبت و منفی است. قطبیت هر کلمه به‌صورت تفاوت قطب مثبت از منفی در نظر گرفته می‌شود. به واژگانی که در SentiWordNet وجود ندارند، قطبیت صفر داده می‌شود.

1. Open Source

هر واژه ترجمه‌های متعددی دارد و با توجه به این ترجمه، قطبیت متفاوتی از آن واژه در لغت‌نامه به‌دست می‌آید. هر کلمه با توجه به نقشی که در جمله دارد و موقعیت استفاده از آن، می‌تواند معانی متفاوتی داشته باشد. در این مقاله ترجمهٔ کلمه با توجه به اینکه کدام معنی از کلمه، بیشترین تکرار را در اسناد دارد، انتخاب شده است. برای تعیین قطبیت نیز، تلاش شد با بررسی تمام جوانب، بهترین قطبیت برای کلمه مد نظر شناسایی شود. در نتیجه لغت‌نامهٔ نهایی با ۳۱۲ لغت شکل گرفت.

پیش‌پردازش داده‌های اصلی

در این بخش به مراحل انجام‌گرفته برای پیش‌پردازش داده‌های اصلی پرداخته می‌شود. مراحل پیش‌پردازش‌های داده‌های اصلی شباهتی به پیش‌پردازش‌های ایجاد لغت‌نامه ندارد. برای پیش‌پردازش داده‌های اصلی، واژگانی که در لغت‌نامه بودند، مبنای کار قرار گرفتند. ابتدا تمام اسناد جمع‌آوری‌شده با استفاده از نرم‌افزار ویراستیار تحلیل شدند. برای این منظور از برنامهٔ اصلاح نویسه‌ها و دستور خط این نرم‌افزار استفاده شد. این برنامه پس از اعمال اصلاحاتی، شامل اصلاح نویسه‌های عربی و فاصله‌گذاری، فهرستی از کلمه‌های مرکب موجود در لغت‌نامه را در اختیار محقق قرار داد. این فهرست برای اصلاح فاصله‌گذاری این کلمه‌ها در مجموعهٔ داده‌ها تهیه می‌شود. در نتیجه فقط واژگانی که از لحاظ فاصله‌گذاری، اصلاح شدند در این فهرست درج می‌شوند. بعد از اصلاح فاصله‌گذاری، واژگانی از لغت‌نامه استخراج شدند که شکل محاوره‌ای آنها متفاوت از شکل رسمی‌شان بود. این کلمه‌ها نیز مبنای اصلاح کلمه‌های محاوره‌ای در داده‌ها قرار گرفتند.

ارزیابی نتایج

در این بخش نتایج دسته‌بندی نظرها با استفاده از الگوریتم SVM و با در نظر گرفتن چهار فرضیه تحلیل شده است. برای دسته‌بندی نظرها از داده‌های دیگری استفاده شد که از همان تارنمای هلوکیش به‌دست آمد. پس از ایجاد لغت‌نامه و پیش‌پردازش داده‌های اصلی، به دسته‌بندی نظرهای جمع‌آوری‌شده پرداخته می‌شود. برای این منظور داده‌های لغت‌نامه مبنای دسته‌بندی و ویژگی‌های ورودی دسته‌بندی‌کننده قرار گرفتند.

با توجه به چهار فرضیه، چهار مجموعهٔ داده ایجاد می‌شود. در مجموعهٔ اول داده‌هایی قرار می‌گیرد که مبتنی بر فرکانس واژگان است. همهٔ کلمه‌های لغت‌نامه در بین اسناد جست‌وجو شدند و تعداد تکرار آنها در اسناد به‌دست آمد. این کار با بهره‌مندی از الگوریتم مخصوصی برای شمارش کلمه‌ها انجام گرفت. مجموعهٔ دوم داده‌ها، حضور کلمه‌ها یا نبود آنها را در اسناد نشان

می‌دهد. برای مجموعهٔ اول، حاصل‌ضرب تعداد تکرار در قطبیت محاسبه شد و در مجموعه سوم قرار گرفت و برای مجموعهٔ دوم، حاصل‌ضرب حضور و نبود کلمه در قطبیت به‌دست آمد و مجموعهٔ چهارم را شکل داد. این داده‌ها به‌منزلهٔ ورودی الگوریتم SVM، وارد نرم‌افزار وکا شدند که نتایج آن در جدول ۱ مشاهده می‌شود.

جدول ۱. درستی الگوریتم SVM و مقایسهٔ آن با الگوریتم بیزین ساده

الگوریتم	درستی SVM	درستی بیزین ساده
تعداد تکرار کلمه‌ها	۷۸/۴۴۱۴ درصد	۷۴/۰۹۹۷ درصد
حضورداشتن و حضورنداشتن کلمه‌ها	۸۱/۰۲۲۴ درصد	۸۰/۷۶۶۸ درصد
حاصل‌ضرب تعداد تکرار در قطبیت	۸۳/۵۷۸۳ درصد	۸۲/۴۲۹ درصد
حاصل‌ضرب حضورداشتن و حضورنداشتن کلمه‌ها در قطبیت	۸۱/۰۲۲۴ درصد	۸۰/۷۰۲۹ درصد

همان‌طور که در جدول مشاهده می‌کنید، الگوریتم SVM و بیزین ساده تقریباً نتایج مشابهی دارند. الگوریتم SVM با به‌کارگیری فرضیهٔ اول به نتایج بهتری نسبت به سایر فرضیه‌ها دست یافته است. در این فرضیه تعداد تکرار هر کلمه در قطبیت آن ضرب شد تا میزان تأثیر مثبت و منفی‌بودن کلمات مشخص شود. این روش در مقایسه با روش‌هایی که قبلاً برای زبان فارسی ارائه‌شده، نتایج بهتری را کسب کرده است.

نتیجه‌گیری

این مقاله به بررسی نظرکاوی در زبان فارسی پرداخت. همان‌طور که پیش از این اشاره شد، به نظرکاوی در زبان فارسی کمتر توجه شده است؛ لذا این مقاله روشی برای نظرکاوی در زبان فارسی ارائه کرد. ابتدا پژوهش‌های انجام‌گرفته در زمینهٔ نظرکاوی و نظرکاوی در زبان فارسی، مرور شدند. برای نظرکاوی از ترکیب الگوریتم SVM و لغت‌نامه استفاده شد. برای ایجاد لغت‌نامه، مجموعهٔ کلمه‌های موجود در لغت‌نامهٔ SentiWordNet به‌کار رفت. واژگان موجود در لغت‌نامه، ویژگی‌های الگوریتم SVM در نظر گرفته شدند. چهار فرضیه برای این منظور تعریف شد که از بین آنها فرضیهٔ حاصل‌ضرب قطبیت در تعداد تکرار کلمه‌ها، بیشترین درستی را به‌دست آورد. روش پیشنهادی نسبت به روش‌های دیگری که تاکنون برای زبان فارسی انجام گرفته است، نتایج بهتری را به‌دست آورد.

References

De Albornoz, J.C., Plaza, L., Gervàs, P. & Díaz, A. (2011). A joint model of feature mining and sentiment analysis for product review rating. In *33rd European Conference on IR Research*. pp. 55–66. DOI: 10.1007/978-3-642-20161-5_8.

Baccianella, S., Esuli, A. & Sebastiani, S. (2010). SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *In Proceedings of the Seventh Conference on International Language Resources and Evaluation*, 2200-2204. European Language Resources Association. Retrieved from: http://lrec.elra.info/proceedings/lrec2010/pdf/769_Paper.pdf.

Bamneshin, M., Mahdizadeh, R. & Pilehvar, A. (2011). A new stemmer for Persian verbs. In *3rd National Conference on Computer Engineering and Information Technology (CEIT2011)*. (*in Persian*)

Basari, A.S.H., Hussin, B., Ananta, G.B. & Zeniarja,J. (2013). Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. *Procedia Engineering*, 53: 453–462.

Brooke, J. (2009). Cross-Linguistic Sentiment Analysis : From English to Spanish. *International Conference RANLP*. pp. 50–54. Available in: https://www. sfu.ca/~mtaboada/docs/Brooke_et_al_RANLP_2009.pdf.

Haddi, E., Liu, X. & Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *Procedia Computer Science*, 17: 26- 32.

Hajmohammadi, M.S. & Ibrahim, R. (2013). A SVM-based method for sentiment analysis in Persian language. In Z. Zhu, ed. *international Conference on Graphic and Image Processing*. p. 876838. DOI: 10.1117/12.2010940.

Hu, M. & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*. New York, New York, USA: ACM Press, p. 168.

Hung, C. & Lin, H.-K. (2013). Using Objective Words in SentiWordNet to Improve Word-of-Mouth Sentiment Classification. *IEEE Intelligent Systems*, 28(2): 47-54.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. *10th European Conference on Machine Learning Chemnitz*. Germany: Springer Berlin Heidelberg.

Li, G. & Liu, F. (2012). Application of a clustering method on sentiment analysis. *Journal of Information Science*, 38(2): 127–139.

Li, G. & Liu, F. (2013). Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions. *Applied Intelligence*, 40(3): 441-452.

Lin, C. & He, Y. (2009). Joint Sentiment / Topic Model for Sentiment Analysis. In *the 18th ACM conference on Information and knowledge management*. pp. 375–384. DOI:10.1145/1645953.1646003.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5 (1): 1-176.

Martineau, J. & Finin, T. (2009). Delta TFIDF: An Improved Feature Space for Sentiment Analysis. *Third AAAI International Conference on Weblogs and Social Media*. Available in: http://ebiquity.umbc.edu/_file_directory_ /papers/446.pdf.

Miller, G.A. (1995). WordNet: A Lexical Database for English, *Communications of the ACM*, 38 (11): 39-41.

Molina-González, M.D., Martínez-Cámara, E. & Martín-Valdivia, M., Perea-Ortega, J. (2013). Semantic orientation for polarity classification in Spanish reviews. *Expert Systems with Applications*, 40 (18): 7250– 7257.

Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M.T. & Ureña-López, L.A. (2014). Ranked WordNet graph for Sentiment Polarity Classification in Twitter. *Computer Speech & Language*, 28(1): 93-107.

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

Pang, B. & Lee, L. (2002). Thumbs up ? Sentiment Classification using Machine Learning Techniques. In *ACL-02 conference on Empirical methods in natural language processing*. PP. 79–86. DOI:10.3115/1118693. 1118704.

Saraee, M. & Bagheri, A. (2013). Feature Selection Methods in Persian Sentiment Analysis. *Springer Berlin Heidelberg*, 7934: 303–308.

Shams, M., Shakery, A. & Faili, H. (2012). A non-parametric LDA-based induction method for sentiment analysis. *16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012)*. IEEE, pp. 216–221. 23 May, Shiraz, Fars.

Stone, P., Dunphry, D., Smith, M. & Ogilvie, D. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M.(2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307.

Turney, P.D. (2001). Thumbs up or thumbs down? *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Morristown, NJ, USA: Association for Computational Linguistics, p. 417.

Vinodhini, G., Chandrasekaran, R. M. (2014). Sentiment Mining Using SVM-Based Hybrid Classification Model. In G. S. S. Krishnan et al., eds. *Proceedings of ICC3*. New Delhi: Springer India.

Wan, X. (2008). Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pp. 553–561. Available in: http://dl.acm.org/citation.cfm?id=1613783.

Wan, X., (2009). Co-Training for Cross-Lingual Sentiment Classification. *ACL '09 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,* 1: 235–243.

Wei, B. & Pal, C. (2010). Cross lingual adaptation: an experiment on sentiment classifications. *Proceedings of the ACL 2010 Conference Short Papers*, pp. 258-262.

Wiebe, J., Wilson, T. & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39 (2-3): 165-210.

Ye, Q., Zhang, Z. & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3): 6527-6535.

Yu, L. & Ma, J. (2008). Opinion mining: A study on semantic orientation analysis for online document. In *7th World Congress on Intelligent Control and Automation*. IEEE, pp. 4548-4552. DOI: 10.1109/WCICA. 2008.4594529.