استخراج ویژگی‌ها و بسط لغت‌نامه در اندیشه‌کاوی مورد استفاده در متون فارسی

عفت گلپر رابوکی، ساقی‌السادات ضرغامی‌فر و جلال رضایی نور

چکیده: اندیشه‌کاوی به تحلیل اظهار نظرات کاربران جهت استخراج نظرات، احساسات و خواسته‌های کاربران در یک حوزه خاص می‌پردازد. دانستن نظرات افراد در یک حوزه خاص می‌تواند نقش مهمی در تصمیم‌گیری‌های کلان آن حوزه ایفا کند. به طور کلی اندیشه‌کاوی در سه سطح سند، جمله و ویژگی به استخراج نظرات کاربران می‌پردازد. اندیشه‌کاوی در سطح ویژگی به دلیل تحلیل جهت‌گیری جنبه‌های مختلف یک حوزه از دو سطح دیگر بیشتر مورد توجه قرار دارد. در این مقاله روشی به منظور استخراج ویژگی‌ها و بسط لغت‌نامه اندیشه‌کاوی ارائه شده است. این لغت‌نامه به منظور تعیین جهت‌گیری نظرات کاربران مورد استفاده قرار می‌گیرد. روش پیشنهادی شامل چهار گام اصلی است. در گام نخست لغت‌نامه اندیشه‌کاوی برای زبان فارسی ایجاد می‌شود. گام دوم مرحله پیش‌پردازش شامل تقطیع، ایجاد برچسب‌های ادات سخن و برچسب وابستگی نحوی اسناد است. گام سوم استخراج ویژگی‌ها و بسط لغت‌نامه با استفاده از روش انتشار دوگانه است و در گام چهارم ویژگی‌ها و قطبیت کلمات حاوی نظر استخراج‌شده در مرحله قبلی اصلاح شده و در نهایت قطبیت ویژگی‌ها تعیین می‌گردد. برای ارزیابی روش پیشنهادی، نتایج حاصل را با روش استخراج ویژگی بر اساس تکرار در متون فارسی که قبلاً ارائه شده است مقایسه خواهیم نمود. نتایج به دست آمده نشان می‌دهد که روش ارائه‌شده در این مقاله نسبت به روش استخراج ویژگی بر اساس تکرار در متون فارسی عملکرد بهتری دارد.

کلیدواژه: **اندیشه‌کاوی، استخراج ویژگی، بسط لغت‌نامه اندیشه‌کاوی، برچسب ادات سخن، برچسب نحوی، انتشار دوگانه.**

۱- مقدمه

با پیدایش وب ۲.۰ و ظهور شبکه‌های اجتماعی، اطلاعات بسیاری در اینترنت منتشر گردید. این داده‌های منتشرشده کاربردهای بالقوه جدیدی دارند که در هر زمان گروهی از آنها کشف می‌شوند. درصد قابل توجهی از این داده‌های انتشاریافته به صورت اسناد متنی می‌باشند که می‌توان آنها را به دو دسته تقسیم نمود: ۱) عینی[1] (واقعیت‌گرایانه) و ۲) ذهنی[2]. واقعیت‌ها، دستورات واقعی و قابل مشاهده درباره موجودیت‌های مستقل و اتفاقاتی است که در جهان می‌افتد. اما دستورات ذهنی بازتاب عواطف انسانی و یا مشاهداتی است که مردم نسبت به دنیای خارج و اتفاقات آن دارند [۱]. موتورهای جستجو قادر به بازیابی اطلاعات اسناد واقعیت‌گرایانه بر اساس کلمات کلیدی که به واقعیت‌ها اشاره دارند، هستند اما برای بازیابی و تحلیل اسناد ذهن‌گرایانه استفاده از آنها ناکارآمد به نظر می‌رسد [۲].

اندیشه‌کاوی[3] و تحلیل احساسات[4] که در دهه اخیر مورد توجه قرار گرفته است به استخراج نظرات کاربران و تشخیص قطبیت آنها، درون متون ذهن‌گرایانه می‌پردازد. طبق تعریفی که در [۵] آمده، اندیشه‌کاوی تنها تشخیص مثبت، منفی و یا خنثی‌بودن نظرات است اما در تحلیل احساسات، به هر کلمه حاوی نظر[5] وزنی بر اساس موضوع متن و قطبیت آن کلمه داده می‌شود. برای مثال در [۶] به کلمه "کثیف" به شرط موضوع هتل و قطبیت منفی، وزن ۰/۰۱ اختصاص داده شده است در حالی که همین کلمه در همین موضوع و به شرط مثبت، وزنی معادل ۰/۰۰۰۰۱ را می‌گیرد.

از جمله کاربردهای اندیشه‌کاوی می‌توان به موارد زیر اشاره نمود:

- تحلیل نظرات مشتریان برخط

با افزایش شمار پایگاه‌های اینترنتی که اقدام به جمع‌آوری نظرات بازدیدکنندگان در باب محصولی خاص یا سرویسی ویژه می‌نمایند، اهمیت اندیشه‌کاوی مشخص می‌گردد. اندیشه‌کاوی می‌تواند برای پیشنهاد خرید یا عدم خرید محصولی خاص یا استفاده از خدمات ویژه و نیز به عنوان مشاوری برای تولیدکنندگان محصولات، جهت استخراج ویژگی‌های مطلوب مشتریان و ارتقای کیفیت محصولات و خدمات استفاده گردد [۳].

- نمایش تبلیغ مناسب

با بررسی موضوعات و نظرات مطرح‌شده در یک بلاگ و یا یک انجمن می‌توان تبلیغی را که احتمال مشاهده آن بیشتر است، نمایش داد. برای مثال اگر نظرات مطرح‌شده در یک انجمن در مورد یک محصول خاص مثبت است، آن گاه احتمال مشاهده تبلیغات آن محصول توسط کاربران آن انجمن بسیار زیاد است اما در صورت منفی‌بودن نظرات، شاید بهتر باشد که محصولات رقیب را در تبلیغات نمایش داد [۴].

- بررسی افکار عمومی

برای بررسی افکار عمومی در مورد یک موضوع خاص می‌توان چندین منبع (مانند انجمن‌های خاص، تویتر و ...) در اینترنت را مورد بررسی و نظر کاربران را در مورد آن موضوع جمع‌آوری و مورد ارزیابی قرار داد.

با توجه به تعاریف بالا، تشخیص مشخصه مناسب (ویژگی‌ها یا کلمات حاوی نظرات) برای طبقه‌بندی احساسات ضروری است. کلمات حاوی نظر برای بیان احساسات مثبت و یا منفی به کار می‌روند. برای مثال کلماتی مانند "خوب"، "زیبا" و "شگفت‌انگیز" احساس مثبت را به انسان القا می‌کنند و کلماتی مانند "بد"، "زشت" و "ترس‌آور" کلماتی با

این مقاله در تاریخ ۱۴ اردیبهشت ماه ۱۳۹۳ دریافت و در تاریخ ۲ آبان ماه ۱۳۹۴ بازنگری شد.

عفت گلپر رابوکی، گروه ریاضی، دانشگاه قم، قم، (email: g_raboky@qom.ac.ir).
ساقی‌السادات ضرغامی‌فر، گروه فنی و مهندسی، دانشگاه قم، قم، (email: saghi_zarghami@yahoo.com).
جلال رضایی نور، گروه فنی و مهندسی، دانشگاه قم، قم، (email: rezaeenoor@yahoo.com).

1. Objective
2. Subjective
3. Opinion Mining
4. Sentiment Analysis
5. Opinion Word

قطبیت منفی هستند. منظور از قطبیت هر کلمه، احساس و برآوردی است که آن کلمه در ذهن متبادر می‌کند. باید به این موضوع توجه داشت که اغلب کلمات حاوی نظر صفت و قید هستند اما بعضی از اسم‌ها مانند "آشغال" و "تفاله" و فعل‌هایی مانند "متنفربودن" و "دوست‌داشتن" نیز حاوی نظر هستند.

بعد از تشخیص مشخصه مناسب، نیاز است که قطبیت آن نیز مشخص گردد. در لغت‌نامه اندیشه‌کاوی، هر کلمه حاوی نظر به همراه قطبیت آن کلمه آمده است. لغت‌نامه اندیشه‌کاوی ممکن است وزن‌دار یا بدون وزن باشد. منظور از وزن، احتمال یا عددی است که برای مثبت یا منفی‌بودن یک کلمه در نظر گرفته می‌شود. اگرچه روش‌های گوناگونی برای ایجاد لغت‌نامه اندیشه‌کاوی ارائه شده و چندین لغت‌نامه اندیشه‌کاوی نیز ساخته و در دسترس عموم قرار گرفته‌اند، اما ایجاد یک لغت‌نامه اندیشه‌کاوی که بتواند همه لغات حاوی نظرات را داشته و همه حوزه‌ها و زبان‌ها را تحت پوشش خود قرار دهد، بسیار بعید به نظر می‌رسد. یک کلمه می‌تواند در یک حوزه قطبیتی مثبت داشته باشد و در حوزه دیگر قطبیتی منفی یا خنثی. برای مثال کلمه "غیر قابل پیش‌بینی" در حوزه ابزار الکترونیکی دارای قطبیتی منفی است اما این کلمه در حوزه فیلم قطبیتی مثبت دارد. در [۶] و [۷] به منظور ایجاد لغت‌نامه اندیشه‌کاوی در زبان فارسی از ترجمه یک لغت‌نامه موجود به زبان انگلیسی و سپس برطرف نمودن اشکالات حاصل از ترجمه استفاده شده است. عدم پوشش بسیاری از کلمات حاوی نظر با استفاده از این روش از جمله مشکلاتی است که وجود دارد.

در این مقاله، روشی را ارائه می‌دهیم که علاوه بر استخراج ویژگی‌ها[1]، به بسط لغت‌نامه اندیشه‌کاوی نیز می‌پردازد. روش پیشنهادی تنها نیاز به یک لغت‌نامه پایه دارد. این لغت‌نامه شامل تعداد محدودی کلمه که تقریباً در تمام حوزه‌ها وجود داشته و قطبیتی ثابت دارند، است. روش پیشنهادی شامل چهار گام اصلی است. در گام نخست لغت‌نامه پایه اندیشه‌کاوی برای زبان فارسی ایجاد می‌شود. گام دوم مرحله پیش‌پردازش است که شامل تقطیع[2]، ایجاد برچسب‌های ادات سخن (POS)[3] و برچسب‌گذاری وابستگی نحوی[4] می‌باشد. گام سوم استخراج ویژگی‌ها با استفاده از روش انتشار دوگانه[5] است و در گام چهارم ویژگی‌ها و کلمات حاوی نظر به دست آمده در مرحله قبلی اصلاح شده و در نهایت قطبیت ویژگی‌ها تعیین می‌گردد.

۲- تحقیقات پیشین

در سال ۲۰۰۲، Pang و Lee با استفاده از الگوریتم‌های یادگیری ماشین به دسته‌بندی متون به دو دسته خنثی و دارای قطبیت پرداختند. آنها از سه الگوریتم ماشین‌های بردار پشتیبان (SVM)[6]، بیزین[7] و بیشینه بیشینه آنتروپی[8] استفاده کردند [۳].

در سال ۲۰۰۳، Riloff و همکاران روشی برای استخراج جملات ذهنی با استفاده از روش خودراه‌انداز[9] ارائه نمودند. در این شیوه ابتدا از روی یک لغت‌نامه و یک مجموعه برچسب نخورده از داده‌ها و با استفاده از دو رده‌بند، جملات در قالب دو کلاس (جملات مربوط به نظرات کاربر و سایر جملات) دسته‌بندی می‌شوند. سپس از این جملات الگوهایی استخراج شده و این الگوها در قالب یک الگوریتم تکرارشونده به رده‌بند برگردانده می‌شوند [۸].

در همان سال، Yi و همکارانش با استفاده از مدل ترکیبی ارائه‌شده در [۹]، ویژگی‌های اظهار نظرات کاربران را استخراج نمودند [۱۰]. روش آنها بر پایه استفاده از برچسب ادات سخن و برچسب ویژگی با استفاده از مجموعه آموزشی بود. آنها تنها معیار ارزیابی دقت را مد نظر قرار دادند.

در سال ۲۰۰۴، Liu و Hu ویژگی‌ها را با استفاده از تشخیص اسامی و تکرار آنها در مجموعه اسناد استخراج نمودند [۱۱]. آنها برای تشخیص اسامی از برچسب‌گذاری ادات سخن استفاده نمودند.

در سال ۲۰۰۵ روش OPINE شامل چهار مرحله شناسایی ویژگی‌ها، شناسایی نظرات مربوط به هر ویژگی، تعیین قطبیت نظرات و رتبه‌بندی نهایی معرفی شد [۱۲]. در این روش ویژگی‌ها با استفاده از محاسبه PMI کلمات شناسایی شده‌اند.

در سال ۲۰۰۷، Mei و همکارانش به استخراج ویژگی‌ها با استفاده از ایجاد الگو در یک حوزه خاص با استفاده از مدل پنهان مارکوف، مبادرت نمودند [۱۳]. بسیاری از روش‌های دیگر نیز بر مبنای ایجاد الگو به منظور استخراج ویژگی‌ها در یک حوزه خاص ارائه شده‌اند [۱۴] تا [۱۶].

در سال ۲۰۰۸، Titove و McDonald با استفاده از روش تخصیص دیریکله به استخراج ویژگی‌ها پرداخته و در نهایت رتبه‌بندی هر یک از ویژگی‌ها را با توجه به نظر کاربر در مورد آن ویژگی، مشخص کردند [۱۷]. در این تحقیق ویژگی‌ها به دو دسته دانه ریز[10] و دانه درشت[11] تقسیم شده‌اند.

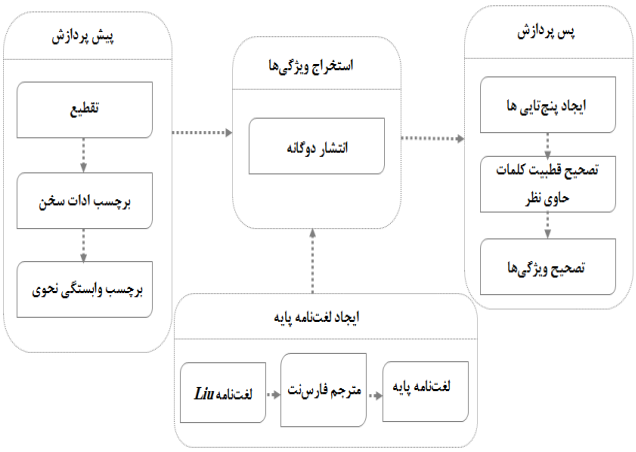
در سال ۲۰۱۱، Liu و همکارانش با استفاده از برچسب وابستگی نحوی و قواعد زبان به استخراج ویژگی‌ها و بسط لغت‌نامه پرداختند. روش پیشنهادی آنها تنها از یک لغت‌نامه پایه که حاوی تعداد محدودی از کلمات حاوی نظر بود، استفاده می‌کرد [۱۸].

در ۲۰۱۲، شمس نجف‌آبادی، روشی بدون سرپرست[12] جهت تعیین قطبیت مستندات فارسی ارائه کرده است. در روش ارائه‌شده، بعد از استخراج کلمات حاوی نظر، هر کلمه بر مبنای موضوعی که در آن قرار دارد با استفاده از دو الگوریتم PLSASA و LDASA وزن‌دهی می‌شود [۶].

در سال ۲۰۱۳ استخراج ویژگی با استفاده از روش ارائه‌شده در [۱۱] در متون فارسی در دو حوزه دانشگاه و تلفن همراه اجرا شده و ارزیابی گردید [۷]. ارزیابی‌ها نشان داد که هر چند روش استخراج ویژگی بر اساس تکرار اسامی در اسناد، دقت بالایی دارند اما نتایج معیار فراخوانی پایین‌تر از حد انتظار بود.

تحقیقات ما نشان می‌دهد که بسیاری از روش‌های پیشنهادشده به منظور استخراج ویژگی‌ها در یک حوزه خاص، نیازمند داده‌های آموزشی خاص آن حوزه هستند و همچنین برای تعیین قطبیت ویژگی‌های استخراج‌شده نیاز به لغت‌نامه اندیشه‌کاوی جامع، ضروری است. با توجه به

1. Feature Extraction
2. Tokenization
3. Part of Speech Tagging
4. Syntax Dependency Parsing
5. Double Propagation
6. Support Vector Machine
7. Naïve Bayes
8. Maximum Entropy Model
9. Bootstrapping
10. Fine-Grained
11. Coarse-Grained
12. Unsupervised



شکل ۱: شمای کلی مدل ارائه‌شده.

این که در حال حاضر، مجموعه‌های آموزشی برای این منظور در حوزه‌های مختلف در زبان فارسی وجود ندارد و همچنین به دلیل عدم وجود یک لغت‌نامه جامع در زبان فارسی، در این مقاله از روش انتشار دوگانه که در [۱۸] آمده است، به منظور استخراج ویژگی‌ها و بسط لغت‌نامه، استفاده نموده‌ایم.

۳- معرفی روش ارائه‌شده

روش ارائه‌شده در این پژوهش در سطح ویژگی است و از ۴ گام اصلی تشکیل شده است. این ۴ گام، شامل ایجاد لغت‌نامه پایه اندیشه‌کاوی، پیش‌پردازش، استخراج ویژگی‌ها و بسط لغت‌نامه با استفاده از روش انتشار دوگانه و پس‌پردازش است. در ادامه هر یک از این گام‌ها را شرح خواهیم داد.

شمای کلی مدل پیشنهادی در شکل ۱ آمده است.

۳-۱ ایجاد لغت‌نامه پایه اندیشه‌کاوی

اولین گام در اندیشه‌کاوی ایجاد لغت‌نامه است. بررسی فعالیت‌های صورت‌گرفته در سایر زبان‌ها (غیر از انگلیسی) بیانگر این مطلب است که روشی که در اکثر زبان‌ها برای ساخت لغت‌نامه استفاده می‌شود، ترجمه یکی از لغت‌نامه‌های موجود در زبان انگلیسی به زبان مقصد و سپس انجام اصلاحات روی آن است. در این مقاله نیز از همین روش برای ایجاد لغت‌نامه پایه استفاده شده است. این لغت‌نامه تنها شامل تعداد محدودی از لغات که بیانگر احساسات مثبت و یا منفی هستند می‌باشد. لغات موجود در این لغت‌نامه تقریباً در تمامی حوزه‌ها برای بیان احساسات به چشم می‌خورند. به عنوان مثال «خوب»، «زیبا»، «بد»، «زشت» و غیره. این لغت‌نامه توسط الگوریتم ارائه‌شده توسعه می‌یابد.

برای ایجاد لغت‌نامه پایه از لغت‌نامه ایجادشده در [۱۹] و [۲۰] استفاده و برای ترجمه از ابزار فارس‌نت که یک لغت‌نامه رایگان است بهره گرفته شده است [۷] و [۲۱]. در [۷] به منظور ایجاد لغت‌نامه اندیشه‌کاوی، معادل فارسی هر کلمه به همراه کلمات مترادف آن به لغت‌نامه فارسی اضافه شده است. ما تنها معادل فارسی کلمات را به دست می‌آوریم و از اضافه‌کردن مترادف‌های کلمات خودداری می‌کنیم. دلیل این امر، ایجاد یک لغت‌نامه ابتدایی است که در گام‌های بعدی با تحلیل اسناد حاوی نظر، آن را بسط خواهیم داد. به دلیل عدم وجود قطبیت در لغت‌نامه انگلیسی، قطبیت کلمات پس از ترجمه به صورت دستی تعیین می‌شوند. همچنین گروهی از کلمات که به نظر در متون فارسی پرکاربرد نیستند به صورت دستی حذف شده‌اند.

دانشگاه	N_SING
قم	N_SING
از	P
لحاظ	N_SING
اساتید	N_SING
متوسط	ADJ_SIM
و	CON
از	P
لحاظ	N_SING
امکانات	N_PL
اقتضاع	N_SING
بود	V_PA
.	DELM

شکل ۲: نمونه برچسب‌گذاری ادات سخن.

ترجمه ساده از یک لغت‌نامه اشکالاتی در بر دارد. به منظور اصلاح لغت‌نامه فارسی، کل کلمات مورد بررسی قرار گرفته و کلماتی که به اشتباه برچسب خورده بودند به صورت دستی تصحیح گردیدند. در لغت‌نامه پایه تولیدشده، تعداد کل واژه‌ها ۵۷۵ واژه است که در آن تعداد واژه‌ها با قطبیت منفی ۲۸۸ واژه، واژه‌های خنثی ۳۶ واژه و مابقی، قطبیتی مثبت دارند.

۳-۲ پیش‌پردازش

این مرحله شامل چندین گام برای ایجاد داده‌هایی که الگوریتم استخراج ویژگی در مرحله بعد به آن نیاز دارد، است. گام‌های این مرحله شامل یکسان‌سازی، تقطیع، برچسب‌گذاری ادات سخن و برچسب‌گذاری وابستگی نحوی است.

۳-۲-۱ یکسان‌سازی نگارش

برخی از حروف در زبان فارسی هستند که در استانداردهای مختلف کدگذاری به شیوه‌های متفاوتی نوشته می‌شوند. به عنوان نمونه حروف "ی" و "ک"، هر یک به چند شکل در متون فارسی یافت می‌شوند. در مرحله یکسان‌سازی این حروف به یک شکل تبدیل می‌شوند.

۳-۲-۲ یکسان‌سازی نگارش

در این گام، هر سند به کلمات تشکیل‌دهنده سند، تجزیه می‌شود. به منظور تعیین کلمات ابتدا هر اظهار نظر بر اساس علامت‌های نگارشی («.»، «؟»، «!»، «:»، «،»، «؛») شکسته شده و سپس جملات حاصل به کلمات تشکیل‌دهنده‌اش تقسیم می‌شوند.

۳-۲-۳ برچسب‌گذاری ادات سخن

برچسب‌گذاری ادات سخن به معنای به دست آوردن گونه صرفی کلمات یک متن است [۲۲]. دو مرحله اصلی در ساخت یک سامانه برچسب‌گذاری ادات سخن مبتنی بر داده وجود دارد. در مرحله نخست با استفاده از یک پیکره[1] آموزشی، الگوی برچسب‌گذاری به دست خواهد آمد. در مرحله بعدی بر اساس الگوی به دست آمده در مرحله قبل، برای هر کلمه ورودی، برچسب مناسب تولید خواهد شد. در این گام با استفاده از نرم‌افزار TNT tagger و پیکره بیژن‌خان [۲۳] و [۲۴]، برچسب‌گذاری ادات سخن بر روی اظهار نظراتی که در گام قبلی به کلمات تجزیه شده‌اند، انجام می‌گیرد. به عنوان مثال اظهار نظر "دانشگاه از لحاظ اساتید متوسط و از لحاظ امکانات افتضاح بود" بعد از اجرایی‌شدن سه گام نخست مرحله پیش‌پردازش به صورت شکل ۲ برچسب‌گذاری می‌شود.

1. Corpus

از	دانشگاه	لحاظ	اساتید	متوسط	و	از	لحاظ	امکانات	افتضاح	بود	.	
N	PREP	N	N	ADJ	CONJ	PREP	N	N	N	V	PUNC	
SBJ	MOS	POSDEP	MOZ	NPOSTMOD		AJCONJ	POSDEP	POSDEP	MOZ	MOS	ROOT	PUNC
11	11	2	3	4	5	6	7	8	11	0	11	

شکل ۳: نمونه برچسب‌گذاری وابستگی نحوی.

جدول ۱: قوانین وابستگی.

	رابطه وابستگی و واژگان	خروجی
۱	(OW dep POS(ADJ)) or (OW dep POS(ADV)POS(ADJ)) dep ∈ {CONJ}	ADJ is new Opinion word If (CONJ ∈ Contrary words) Polarity(ADJ) = - Polarity(OW) Else Polarity(ADJ) = Polarity(OW)
۲	(SBJ(POS(N)) dep OW)	SBJ(N) is new Feature
۳	(F dep MOS(POS(ADJ)))	MOS(ADJ) is new Opinion word
۴	(F dep POS(N)) or (F MOZ(POS(N))) dep ∈ {CONJ}	N is new Feature
۵	(F NPOSTMOD(POS(ADJ)) dep MOS(POS(ADJ)))	F+NPOSTMOD is new Feature

توضیحات

MOS = مُسند	F(Feature) = ویژگی	OW(opinion word) = لغت حاوی نظر
MOZ = مضاف	NPOSTMOD = صفت پسین	ADJ(Adjective) = صفت
	SJB(Subject) = فاعل	CONJ(Conjunction) = حرف ربط

۳-۲-۴ برچسب‌گذاری وابستگی نحوی

تجزیه وابستگی، رهیافتی برای تجزیه نحوی زبان طبیعی به صورت خودکار است. این رهیافت از زبان‌شناسی سنتی مبتنی بر دستور وابستگی اقتباس شده است [۲۵]. در مجموع می‌توان گفت که در تجزیه وابستگی برای هر جمله ورودی یک گراف وابستگی ساخته می‌شود و دو رهیافت عمومی برای آن وجود دارد: مبتنی بر داده و مبتنی بر دستور زبان. در رهیافت مبتنی بر داده از روش‌های یادگیری خودکار و در روش مبتنی بر دستور زبان از دستور زبان‌های صوری استفاده می‌شود. در روش یادگیری با ناظر، دو مرحله اصلی در ساخت یک سامانه تجزیه وابستگی وجود دارد. در مرحله نخست با استفاده از یک پیکره آموزشی، دستور زبان وابستگی به دست می‌آید. با به دست آمدن دستور زبان وابستگی، الگوی تجزیه به دست خواهد آمد. در مرحله بعدی بر اساس الگوی به دست آمده در مرحله قبل، برای هر جمله ورودی، گراف وابستگی تولید خواهد شد.

در این گام با استفاده از نرم‌افزار MST Parser و پیکره وابستگی زبان فارسی دادگان [۲۶] و [۲۷]، برچسب‌گذاری وابستگی نحوی بر روی اسناد انجام می‌گیرد. پس از اجرای این گام، اظهار نظر "دانشگاه از لحاظ اساتید متوسط و از لحاظ امکانات افتضاح بود" به صورت شکل ۳ برچسب‌گذاری می‌شود.

۳-۳ استخراج ویژگی‌ها و بسط لغت‌نامه (روش انتشار دوگانه)

در این گام که مهم‌ترین گام روش پیشنهادی نیز می‌باشد، ویژگی‌های (جنبه‌های) یک شیء که کاربران در مورد آنها نظرات خود را بیان کرده‌اند، استخراج می‌شوند و همچنین به بسط لغت‌نامه پایه خواهیم پرداخت. روش پیشنهادی، روشی خودراه‌انداز است که کار خود را تنها با لغت‌نامه پایه آغاز می‌کند اما در دورهای بعدی، از ویژگی‌های استخراج‌شده نیز به منظور استخراج ویژگی‌ها و بسط لغت‌نامه استفاده می‌کند.

این روش مبتنی بر قوانینی است که به طور طبیعی در ارتباطات وابستگی یک زبان وجود دارد. به عنوان مثال در جمله "این گوشی ظاهر خوبی دارد" در صورتی که بدانیم کلمه "خوب" لغت حاوی نظر است، از طریق قوانین وابستگی می‌توانیم کلمه "ظاهر" را به عنوان ویژگی استخراج کنیم. در جدول ۱ قوانین به کار رفته برای استخراج ویژگی‌ها و بسط کلمات حاوی نظر در این روش، آمده است.

قانون ۱: اگر یک کلمه، لغت حاوی نظر (OW) باشد و بعد از آن حرف ربط و صفت آمده باشد و یا بعد از حرف ربط، قید و صفت به ترتیب آمده باشد، کلمه‌ای که برچسب صفت دارد به عنوان یک کلمه حاوی نظر انتخاب می‌شود. این قانون را به صورت عکس هم در نظر می‌گیریم. یعنی اگر کلمه‌ای، لغت حاوی نظر باشد و قبل از آن حرف ربط و صفت آمده باشد، کلمه‌ای که برچسب صفت دارد یک کلمه حاوی نظر است. برای تشخیص قطبیت کلمه جدید در صورتی که حرف ربط در دسته تعویض‌گرهای ظرفیت قرار داشته باشد، قطبیت کلمه جدید، مخالف قطبیت کلمه‌ای است که به وسیله آن کلمه جدید را استخراج نموده‌ایم و در صورتی که حرف ربط در این دسته قرار نداشته باشد، دو کلمه قطبیتی یکسان دارند. برای مثال فرض کنید که کلمه "زیبا" دارای برچسب OW و قطبیت مثبت است. در این حالت با استفاده از این قانون می‌توان کلمه "جذاب" در جمله‌ای مانند "ظاهر زیبا و جذابی دارد" را به عنوان یک کلمه حاوی نظر جدید با قطبیت مثبت استخراج نمود. همچنین کلمه "شکننده" در جمله "قاب زیبا اما شکننده‌ای دارد"، یک کلمه حاوی نظر با قطبیت منفی است.

بحث دیگر در این زمینه، تعویض‌گر ظرفیت است. منظور از کلمات تعویض‌گر ظرفیت، کلماتی هستند که جمله قبل و بعد از آنها و یا کلمه قبل و بعد از آنها دارای قطبیتی متفاوت هستند. برای مثال قطبیت جمله‌ای که بعد از کلمه "اما" آمده است با قطبیت جمله‌ای که پیش از "اما" آمده است، در تضاد می‌باشد. ما لیستی از لغات تعویض‌گر ظرفیت را که در [۲۸] آمده است، ترجمه و اصلاح کرده و از آن برای تشخیص این لغات در سیستم اندیشه‌کاوی استفاده نموده‌ایم.

```
Input: Opinion word Dictionary {O}, Review Data R
Output: All possible Features {F}, The Expanded Opinion Lexicon {O-Expanded}
Function:
1. {O-Expanded} = {O}
2. {F} =Ø ,{O} =Ø, {TempF} =Ø
3. for each {O }
4.     for each parsed sentence in R
5.         label Opinion words based on Opinion words in {O}
6.     endfor
7.     remove {Oi}
8. endfor
9. for each { TempF }
10.    for each parsed sentence in R
11.        label Features based on Featurs in { TempF }
12.    endfor
13.    remove {TempFi}
14. endfor
15. for each parsed sentence in R
16.    Extract Opinion word {O'} using rule1 and add to {O} and { O-Expanded }
17. endfor
18. for each parsed sentence in R
19.    Extract Feature {F'} using rule2 and add to {F} and { TempF}
20. endfor
21. for each parsed sentence in R
22.    Extract Opinion word {O'} using rule3 and add to add to {O} and { O-
Expanded }
23. endfor
24. for each parsed sentence in R
25.    Extract Feature {F'} using rule4 and add to {F} and { TempF}
26. endfor
27. for each parsed sentence in R
28.    Extract Feature {F'} using rule5 and add to {F} and { TempF}
29. endfor
30. Repeat 3 till size({TempF}) =0,size({O})=0
```

شکل ۴: الگوریتم انتشار دوگانه.

قانون ۲: اگر یک کلمه، لغت حاوی نظر (OW) باشد و برچسب وابستگی آن مُسند باشد، با توجه به برچسب وابستگی، کلمه‌ای که در جمله برچسب فاعل (SJB) دارد را یک ویژگی جدید در نظر می‌گیریم. به عنوان مثال در جمله "دانشگاه محل تحصیل من، بسیار زیبا است" کلمه "دانشگاه" یک ویژگی جدید است.

قانون ۳: این قانون دقیقاً عکس قانون قبلی است یعنی اگر در جمله، یک کلمه، ویژگی باشد آن گاه کلمه‌ای در جمله که نقش مُسند را دارد و صفت نیز می‌باشد به عنوان کلمه حاوی نظر استخراج می‌شود.

قانون ۴: اگر یک کلمه، ویژگی (F) باشد و بعد از آن حرف ربط و اسم آمده باشد و یا دقیقاً بعد از آن اسمی با نقش مضاف آمده باشد، اسم به عنوان ویژگی انتخاب می‌شود. این قانون را به صورت عکس هم در نظر می‌گیریم یعنی اگر کلمه‌ای، ویژگی باشد و قبل از آن حرف ربط و اسم آمده باشد، اسم مورد نظر، ویژگی جدیدی است. برای مثال در جمله "کتابخانه و سلف کوچکی دارد"، در صورتی که "کتابخانه" دارای برچسب ویژگی باشد، "سلف" نیز به عنوان ویژگی جدید استخراج می‌شود.

قانون ۵: بر طبق مشاهدات اگر در جمله‌ای ویژگی و کلمه حاوی نظر مشخص شده باشد و بعد از ویژگی کلمه‌ای با نقش صفت (برچسب ادات سخن) و وابسته پسین (برچسب وابستگی) آمده باشد و این کلمه از طریق حرف ربط نیز از کلمه حاوی نظر جدا نشده باشد، ویژگی و صفت بعد از آن را می‌توان به عنوان یک عبارت اسمی و ویژگی جدید در نظر گرفت. ساختار عبارت اسمی در [۲۸] آمده است. بررسی‌ها نشان می‌دهد که بیشتر عبارت‌های اسمی که در اظهار نظرات به کار می‌رود تنها شامل اسم و صفت پسین است. به عنوان مثال در جمله "امکانات رفاهی آنجا افتضاح بود" به شرط این که امکانات ویژگی و افتضاح، کلمه حاوی نظر باشد، می‌توان "امکانات رفاهی" را یک ویژگی جدید در نظر گرفت.

الگوریتم روش انتشار دوگانه در شکل ۴ به صورت کامل نشان داده شده است. ورودی‌های این الگوریتم لغت‌نامه پایه و مجموعه اظهار نظرات کاربران است. در مرحله اول کلمات لغت‌نامه در همه اسناد جستجو شده و در صورت مشاهده در هر سند، آن کلمه برچسب جدیدی که

ویژگی	قطبیت	تاریخ	نویسنده	نوع
اسپیکر	-۱	۹۱	۹۳	S

کلمه حاوی نظر	قطبیت
ضعیف	-۱
بی‌کیفیت	-۱

شکل ۵: نمونه‌ای از پنج‌تایی ساخته‌شده.

نشان‌دهنده لغت حاوی نظر است، می‌گیرد. در مرحله بعد قوانین به ترتیب اجرا می‌شوند و کلمات استخراج‌شده در سند با توجه به نوع کلمه برچسب ویژگی و یا کلمه حاوی نظر خورده و همچنین به لیست مربوطه اضافه می‌شوند. در دور بعد، کلمات جدید استخراج‌شده (ویژگی و کلمات حاوی نظر) را در اسناد جستجو می‌کنیم. این رویه تا زمانی که دیگر هیچ کلمه جدیدی پیدا نشود، ادامه می‌یابد.

۳-۴ پس‌پردازش

در این گام، به اصلاح ویژگی‌ها و قطبیت کلمات حاوی نظر که در مرحله قبل استخراج شده‌اند می‌پردازیم.

۳-۴-۱ ایجاد پنج‌تایی‌ها و لیست کلمات حاوی نظر

در این بخش به ازای هر ویژگی در هر اظهار نظر، یک رکورد ساخته می‌شود. رکورد ساخته‌شده شامل ۵ خصوصیت ویژگی، قطبیت، تاریخ، نویسنده و نوع است. همچنین به ازای هر رکورد ساخته‌شده، لیستی از کلمات حاوی نظر که ویژگی را توصیف می‌کنند، ایجاد می‌شود. برای روشن‌شدن مطلب، رکورد ساخته‌شده برای جمله "اسپیکر روی دوربین برای پخش صدای فیلم، ضعیف و بی‌کیفیت می‌باشد" در شکل ۵ آورده شده است.

قطبیت ویژگی از مجموع قطبیت‌های کلمات حاوی نظر توصیف‌کننده ویژگی و در نظر گرفتن نقش منفی‌کننده‌ها در جمله به دست می‌آید. منظور از منفی‌کننده، کلمه یا کلماتی است که قطبیت یک جمله را وارونه می‌کند و از آنجایی که در اندیشه‌کاوی، هدف تعیین مثبت یا منفی‌بودن یک نظر است، بررسی نقش منفی‌کننده‌ها در این حوزه بسیار حایز اهمیت است. در زبان فارسی اکثر منفی‌کننده‌ها در فعل جمله ظاهر می‌شوند و فعل جزء اصلی جمله در تعیین مثبت یا منفی‌بودن جمله است. برای شناسایی منفی‌کننده‌های فعل فارسی از روشی که در [۶] آمده، پیروی می‌کنیم. در این روش، جهت تشخیص فعل از مجموعه بیژن‌خان استفاده شده است.

در این مجموعه ابتدا تمام فعل‌های منفی به صورت دستی برچسب خوردند. سپس برای گسترده‌تر شدن و پوشش بیشتر، تمام فعل‌ها با پسوندهای منفی‌کننده "ن"، "نمی" و تمام صیغه‌های "نخواه" گسترش یافتند. به طور حتم تعدادی از کلمات حاصل بی‌معنا و بدون کاربرد خواهند بود. برای مثال کلمه‌ای مانند «ناست» (ن + است) که با این روش ایجاد می‌شود کلمه صحیحی نیست ولی از آنجایی که این کلمات در مجموعه داده‌ها نمی‌آیند، این مسئله مشکلی ایجاد نمی‌کند. از فعل‌های منفی در وارونه‌کردن قطبیت و بررسی نقش منفی‌کننده‌ها استفاده می‌شود.

۳-۴-۲ تصحیح قطبیت کلمات حاوی نظر استخراج‌شده

تعدادی از کلمات حاوی نظر که با استفاده از روش انتشار دوگانه استخراج شده‌اند هیچ قطبیتی به آنها اختصاص داده نشده و دلیل این امر، استخراج این کلمات با استفاده از ویژگی‌ها می‌باشد. ویژگی‌ها به خودی

جدول ۲: مجموعه داده.

مجموعه داده	تعداد اظهار نظرات	تعداد جملات
حوزه دانشگاه	۹۰	۵۹۸
تلفن همراه	۲۵۰	۱۴۰۹

خود، هیچ قطبیتی ندارند و این کلمات حاوی نظر هستند که آنها را توصیف نموده و قطبیت آنها را تعیین می‌کنند. برای اصلاح این مشکل به این روش عمل می‌کنیم که ابتدا رکورد پنج‌تایی[1] کلمه حاوی نظر را در لیست پنج‌تایی‌ها مشخص کرده و سپس قطبیت ویژگی‌های قبل و بعد از آن رکورد را مشاهده می‌کنیم. در صورتی که قطبیت هر دو رکورد مثبت و یا منفی باشد آن را به کلمه حاوی نظر اختصاص داده و قطبیت ویژگی را در پنج‌تایی نیز تصحیح می‌کنیم. مشاهدات نشان می‌دهد که در اظهار نظرات، اگر دو جمله مثبت باشند، در اکثر موارد، جمله مابین آنها نیز مثبت است و بالعکس. در صورتی که این قانون برقرار نباشد، در مرحله بعد، قطبیت کل اظهار نظر را به دست آورده و قطبیت کلمه مورد نظر را برابر با قطبیت سند قرار می‌دهیم. سپس قطبیت ویژگی را بر اساس آن اصلاح می‌نماییم.

۳-۴-۳ تصحیح ویژگی

در [۲۹] لیستی از واژه‌های غیر مفهومی رایج در اسناد فارسی ارائه شده است. برای مثال، کلمه "لحاظ" در بسیاری از اظهار نظرات به عنوان اسم و در نتیجه به عنوان ویژگی برچسب خورده است. برای رفع این مشکل از لیست ارائه‌شده برای تصحیح ویژگی‌ها استفاده شده است. البته گروهی از این واژه‌ها نقش اسمی نداشته و در نتیجه در این قسمت کاربردی نیز ندارند [۷]. در ادامه با استفاده از [۳۰] ویژگی‌های مترادف را مشخص نموده و هر کدام از ویژگی‌ها که بیشترین تکرار را در اسناد داشتند به عنوان ویژگی اصلی در نظر گرفته و جایگزین مترادف‌های آن نمودیم. همچنین لیستی از نام دانشگاه‌ها، مدل‌ها و کارخانه‌های سازنده تلفن همراه نیز تهیه شده و از آنها برای تصحیح ویژگی‌های استخراجی استفاده کرده‌ایم زیرا طبق مشاهدات این اسامی نیز به عنوان ویژگی‌ها برچسب خورده‌اند [۷]. به این منظور، هر ویژگی که شامل یکی از این کلمات باشد را تصحیح نموده و کلمه مورد نظر را از آن ویژگی حذف می‌نماییم.

۴- تحلیل نتایج

در این قسمت به ارزیابی روش ارائه‌شده و مقایسه نتایج استخراج ویژگی‌ها با روش استخراج بر اساس تکرار [۷] خواهیم پرداخت. به این منظور روش خود را بر روی مجموعه داده‌ای که در [۷] استفاده شده اجرا می‌نماییم. این مجموعه داده در دو حوزه تلفن همراه و دانشگاه جمع‌آوری و برچسب‌گذاری شده است. جدول ۲ اطلاعات مربوط به مجموعه داده‌ها را نشان می‌دهد. اظهار نظرات کاربران در حوزه تلفن همراه از سایت http://www.digikala.com جمع‌آوری و دسته‌بندی شده‌اند. اظهار نظرات در زمینه دانشگاه نیز توسط گروهی از دانشگاهیان با استفاده از تکمیل فرم طراحی‌شده به منظور این کار به دست آمده است.

در حوزه دانشگاه در کل ۹۰ اظهار نظر انتخاب گردید که تعداد ۴۵ اظهار نظر منفی و ۴۵ اظهار نظر مثبت بودند. در حوزه تلفن همراه، ۲۵۰ نظر از سایت ذکرشده انتخاب شدند که ۱۲۵ نظر مثبت و ۱۲۵ نظر منفی بودند.

جهت ارزیابی روش‌های پیشنهادی به منظور استخراج ویژگی‌ها و لغات حاوی نظر در این پژوهش از سه معیار دقت[2]، فراخوانی[3] و معیار F[4] استفاده شده است. همچنین جهت ارزیابی قطبیت اختصاص داده شده در روش انتشار دوگانه در این پژوهش از معیار صحت[5] استفاده شده است. نتایج ارزیابی استخراج ویژگی‌ها با استفاده از الگوریتم انتشار دوگانه در مقایسه با روش ارائه‌شده در [۷] در دو حوزه دانشگاه و تلفن همراه در زیر آمده است. همچنین کلمات حاوی نظر استخراج شده و قطبیت این کلمات نیز مورد ارزیابی قرار گرفته‌اند.

۴-۱ ارزیابی ویژگی‌های استخراج‌شده

نتایج نشان می‌دهد که هر چند الگوریتم انتشار بر مبنای تکرار دقت بیشتری دارد اما فراخوانی و ارزیابی بر اساس معیار F در روش انتشار دوگانه به صورت قابل توجهی بهبود یافته است (جدول ۳).

۴-۲ ارزیابی کلمات حاوی نظر استخراج شده

نتایج ارزیابی استخراج کلمات حاوی نظر با استفاده از الگوریتم انتشار دوگانه در دو حوزه دانشگاه و تلفن همراه در زیر آمده است. به دلیل استفاده الگوریتم استخراج مبتنی بر تکرار از لغت‌نامه و عدم بسط آن در این روش، تنها روش پیشنهادی در این مقاله را مورد ارزیابی قرار می‌دهیم (جدول ۴).

۴-۳ ارزیابی قطبیت کلمات حاوی نظر

همان طور که پیشتر نیز گفته شد، گاهی قطبیت یک کلمه به حوزه‌ای که در آن به کار رفته است بستگی دارد. صحت مربوط به قطبیت کلمات حاوی نظر استخراج‌شده با استفاده از الگوریتم انتشار دوگانه مطابق جدول ۵ ارزیابی شده است.

۵- نتیجه‌گیری

در این پژوهش مدلی جدید برای استخراج ویژگی‌ها در متن فارسی ارائه شد. این مدل وابسته به حوزه خاصی نبوده و برای استخراج ویژگی‌ها و بسط کلمات حاوی نظر از وابستگی نحوی زبان استفاده می‌کند. نتایج نشان می‌دهد که این روش در مقایسه با روش استخراج بر مبنای تکرار در سطح ویژگی عملکرد بهتری دارد. همچنین با استفاده از این روش، مشکلات موجود در ایجاد یک لغت‌نامه جامع که همه حوزه‌ها را پوشش دهد، وجود ندارد زیرا این روش با استفاده از یک لغت‌نامه پایه با تعداد کلمات محدود کار خود را آغاز کرده و در ادامه به بسط این لغت‌نامه با استفاده از اظهار نظرات کاربران می‌پردازد، هر چند که روش انتشار دوگانه در تعیین قطبیت کلمات جدید حاوی نظر استخراج‌شده به خوبی عمل نمی‌کند.

به عنوان آخرین مبحث در این مقاله به معرفی کارهایی که می‌توان برای بهبود و گسترش روش ارائه‌شده در آینده انجام داد می‌پردازیم:

- شناسایی مرجع ضمیر
- عدم نادیده‌گرفتن جملاتی که به صورت ضمنی حاوی نظر هستند.
- شناسایی نماینده ویژگی[6]: بسیاری از کلمات حاوی نظر می‌توانند

1. Quintuple

2. Precision
3. Recall
4. F-Measure
5. Accuracy
6. Feature Indicator

جدول ۳: معیار دقت، فراخوانی و F برای ویژگی‌های استخراج‌شده.

مجموعه داده	دقت (Precision)		فراخوانی (Recall)		معیار F (F-Measure)	
	دانشگاه	تلفن همراه	دانشگاه	تلفن همراه	دانشگاه	تلفن همراه
استخراج مبتنی بر تکرار	۰٫۹۲	۰٫۹۴	۰٫۶۴	۰٫۷۲	۰٫۷۵	۰٫۸۱
استخراج مبتنی بر انتشار دوگانه	۰٫۸۸	۰٫۹۲	۰٫۸۳	۰٫۸۶	۰٫۸۵	۰٫۸۹

جدول ۴: معیارهای ارزیابی برای کلمات حاوی نظر استخراج‌شده.

مجموعه داده	استخراج مبتنی بر انتشار دوگانه		
	دقت	فراخوانی	معیار F
دانشگاه	۰٫۸۳	۰٫۷۹	۰٫۸۱
تلفن همراه	۰٫۸۸	۰٫۸۲	۰٫۸۵

جدول ۵: معیار صحت برای قطبیت کلمات حاوی نظر استخراج‌شده.

	صحت (Accuracy)
مجموعه داده	استخراج مبتنی بر انتشار دوگانه
دانشگاه	۰٫۷۳
تلفن همراه	۰٫۶۶

برای هر ویژگی به کار روند. برای مثال کلماتی مانند «خوب»، «بد» و غیره اما تعدادی از این کلمات نماینده ویژگی خاصی هستند. به عنوان مثال کلمه «بزرگ» نماینده ویژگی سایز در جمله «این گوشی خیلی بزرگ است» می‌باشد. می‌توان با شناسایی این کلمات در سیستم، ویژگی‌های بیشتر و دقیق‌تری را استخراج نمود.

- در نظر گرفتن موضوع در اندیشه‌کاوی
- تبدیل نگارش محاوره‌ای به نگارش رسمی
- تحلیل احساسات بیان‌شده با استفاده از الگوریتم‌های وزن‌دهی

مراجع

[1] A. Stavrianou and J. H. Chauchat, "Opinion mining issues and agreement identification in forum texts," in *Proc. 6th Int.Conf. on Computational Linguistics and Intelligent Text Processing, CICLing'05*, 51-58, Feb. 2005.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proc. Conf. on Empirical Methods in Natural Language Processingm, EMNLP'02*, pp. 79-86, Jul. 2002.

[3] M. Sepehri, "Chi-square for features selection in opinion mining in persian text," in *Proc. 2nd National Conf. on Computer/Electrical and IT Engineering, CEIC'09*, pp. 128-132, Mar. 2009.

[4] C. Nichols, Feature Selection and Weighting for Sentiment Analysis, University of Guelph, 2010.

[5] B. Liu, Sentiment Analysis and Opinion Mining, Morgan and Claypool, 2012.

[۶] م. ر. شمس نجف‌آبادی، *اندیشه‌کاوی و تحلیل نظرات در مستندات فارسی*، دانشگاه تهران، ۱۳۹۱.

[۷] س. ا. ضرغامی‌فر، "استخراج ویژگی‌ها در اندیشه‌کاوی مورد استفاده در متون فارسی،" *دومین همایش ملی کامپیوتر دانشکده فنی و حرفه‌ای سما*، صص. ۹۵-۸۹، سنندج، ۱۳۹۲.

[8] E. Riloff, J. Wiebe, and T. Wilson, "Learning subjective nouns using extraction pattern bootstrapping," in *Proc. Conf. on Natural Language Learning, CoNLL'03*, pp. 25-32, Mar. 2003.

[9] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proc. 10th Int. Conf. on Information and Knowledge Management*, pp. 403-410, Oct. 2001.

[10] J. Yi, T. Nasukawa, and R. Bunescu, "Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques," in *Proc. 3rd IEEE Int. Conf. on Data Mining*, pp. 427-434, Nov. 2003.

[11] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 168-177, Aug. 2004.

[12] A. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Proc. Conf. on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 339-346, Oct. 2005.

[13] Q. Mei, X. Ling, and M. Vondra, "Topic sentiment mixture: modeling facets and opinions in weblogs," in *Proc. Int. World Wide Web Conf. Committee*, pp. 171-180, May. 2007.

[14] Y. Liu, X. Huang, and A. An, "ARSA: a sentiment-aware model for predicting sales performance using blogs," in *Proc. 30th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 607-614, Jul. 2007.

[15] R. McDonald, K. Hannan, and T. Neylon, "Structured models for fine-to-coarse sentiment analysis," in *Proc. 45th Annual Meeting of the Association of Computational Linguistics*, pp. 432-439, Jun. 2007.

[16] Q. Su, X. Xu, and H. Guo, "Hidden sentiment association in chinese web opinion mining," in *Proc. Int. World Wide Web Conf. Committee*, pp. 959-968, Apr. 2008.

[17] I. Titov and R. McDonald, "A joint model of text and aspect ratings for sentiment summarization," *Association for Computational Linguistics*, pp. 308-316, Jun. 2008.

[18] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," *Computational Linguistics*, vol. 37, no.1, pp. 9-27, Mar. 2011.

[19] M. Hu and B. Liu, "Mining and summarizing customer reviews", in *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 168-177, Aug. 2004.

[20] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *Proc. 14th Int. World Wide Web Conf.*, pp. 341-351, May. 2005.

[21] M. Shamsfard, *et al*., "Semi-automatic development of farsnet; the persian wordnet," in *Proc. 5th Global WordNet Conf.*, pp.846-850, Aug. 2010.

[۲۲] م. ح. الهی‌منش و ب. مینایی، "برچسب‌گذاری ادات سخن متون فارسی به کمک مدل مخفی مارکوف،" *فصل‌نامه اطلاع رسانی، آموزشی و مطالعات رایانه‌ای علوم اسلامی*، شماره ۳۴، صص. ۱۰۶-۱۰۲، بهار ۱۳۹۰.

[23] F. Raja, H. Amiri, and F. Oroumchian, et al., "Evaluation of part of speech tagging on persian text," in *Proc. Second Workshop on Computational Approaches to Arabic Script-Based Languages*, Linguistic Institute Stanford University, pp. 120-127, Jul. 2007.

[24] S. Tasharofi, *et al*., "Evaluation of statistical part of speech tagging of Persian text," in *Proc. Int. Symp. on Signal Processing and its Applications*, 4 pp., Feb. 2007.

[25] Dadegan Research Group, *Persian Dependency Treebank Version 1.0, Annotation Manual and User Guide*, Supreme Council of Information and Communication Technology (SCICT), 2012.

[26] M. Rasooli, M. Kouhestani, and M. Moloodi, "Development of a persian syntactic dependency treebank," in *Proc. 2013 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT'13*, pp. 306-314, Atlanta, USA, Jun. 201.

[۲۷] گروه پژوهشی دادگان، *پیکره وابستگی نحوی زبان فارسی (نسخه ۱.۰)*، تهران، دبیرخانه شورای عالی اطلاع‌رسانی، بازیابی از http://dadegan.ir/perdt، ۱۳۹۱.

[۲۸] س. کاووسی‌نژاد، "حذف در گروه اسمی زبان،" *نامه فرهنگستان*، صص. ۱۲۷-۱۰۹، ۱۳۷۹.

[۲۹] م. سنجی و م. ر. داورپناه، "شناسایی واژه‌های غیر مفهومی (رایج) در نمایه‌سازی خودکار مدارک فارسی،" *فصل‌نامه کتابداری و اطلاع‌رسانی*، جلد ۱۲، شماره ۴، صص. ۳۶-۹، زمستان ۱۳۸۹.

[۳۰] ف. ا. خداپرستی، *فرهنگ جامع واژگان مترادف و متضاد زبان فارسی*، شیراز، دانشنامه فارس، ۱۳۷۶.

[31] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *Proc. 14th Int. World Wide Web Conf.*, pp. 342-351, Chiba, Japan, May 2005.

عفت گلپر رابوکی تحصیلات خود را در مقطع کارشناسی رشته ریاضی در سال ۱۳۷۳ در دانشگاه صنعتی امیرکبیر و کارشناسی ارشد و دکترای ریاضی را در سال‌های ۱۳۷۵ و ۱۳۹۰ در دانشگاه صنعتی شریف به پایان رساند. هم‌اکنون عضو هیأت علمی دانشکده علوم پایه دانشگاه قم می‌باشد. زمینه‌های تحقیقاتی مورد علاقه نام‌برده عبارتند از: داده کاوی، پردازش تصویر و سیگنال، جبرخطی و چندخطی.

ساقی‌السادات ضرغامی‌فر مدرک کارشناسی خود را در رشته مهندسی کامپیوتر از دانشگاه آزاد اسلامی در سال ۸۱ و مدرک کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات و ارتباطات در دانشگاه قم در سال ۹۲ دریافت نمود. زمینه‌های تحقیقاتی مورد علاقه نام‌برده عبارتند از: اندیشه کاوی، متن کاوی و داده کاوی.

جلال رضایی نور مدارک کارشناسی و دکترای خود را در رشته مهندسی صنایع از دانشگاه علم و صنعت ایران به‌ترتیب در سال‌های ۱۳۸۰ و ۱۳۹۰ و نیز مدرک کارشناسی ارشد خود را در رشته مهندسی صنایع از دانشگاه امام حسین (ع) در سال ۱۳۸۳ دریافت نمود. دکتر رضائی نور از سال ۱۳۹۰ در گروه مهندسی صنایع دانشکده فنی و مهندسی دانشگاه قم مشغول به فعالیت گردید و هم‌اینک نیز عضو هیأت علمی این دانشگاه می‌باشد. زمینه‌های تحقیقاتی مورد علاقه نام‌برده شامل مدیریت دانش، مدیریت سرمایه فکری، مهندسی مجدد فرآیندها، مدیریت عملکرد، تصمیم گیری با معیارهای چندگانه و معماری سازمانی می‌باشد.