مروری بر روش‌های تحلیل احساس در متون فارسی

زینب رجبی[1*]، محمدرضا ولوی[2] و مریم حورعلی[3]
[1]پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران
[2و3]مجتمع برق و کامپیوتر، دانشگاه صنعتی مالک اشتر، تهران، ایران

چکیده

با رشد چشم‌گیر رسانه‌های اجتماعی مانند توییتر و افزایش نظرات کاربران در تارنماهای تجارت الکترونیکی و تارنماهای خبری، افراد و سازمان‌ها به‌طور فزاینده‌ای از نظرات در این رسانه‌ها برای تصمیم‌گیری خود استفاده می‌کنند. تحلیل احساس یکی از روش‌های تحلیل نظرات کاربران است که در سال‌های اخیر مورد توجه قرار گرفته است. تحلیل احساس روی هر زبانی نیازمندی‌های مختص به خود را دارد و به‌کارگیری روش‌ها، ابزارها و منابع زبان انگلیسی به‌طور مستقیم در زبان فارسی با محدودیت‌هایی روبه‌رو است. متون نوشته‌شده به زبان فارسی ویژگی‌های خاصی دارند که نیازمند روش‌های خاص تحلیل احساس هستند که متفاوت از زبان انگلیسی است.

در این مقاله، پژوهش‌های تحلیل احساس که روی متون به زبان فارسی انجام شده است، مورد بررسی و مقایسه قرار می‌گیرد. ابتدا رویکردهای تحلیل احساس، وظایف و سطوح تحلیل احساس تشریح می‌شود. در ادامه تلاش می‌شود که مروری روی روش‌های به کارگرفته‌شده برای وظایف تحلیل احساس متون فارسی انجام شود و جایگاه کارهای انجام‌شده در زبان فارسی روشن شود. همچنین منابع داده‌ای ایجاد و منتشرشده برای تحلیل احساس متون فارسی معرفی شده است. در نهایت با توجه به مطالعات انجام گرفته روی آخرین پیشرفت‌های تحلیل احساس، مسائل و چالش‌هایی که در زبان فارسی به آن پرداخته نشده را برشمرده و نقشه راهی برای پژوهش‌های آینده پردازش متون فارسی ارائه می‌شود.

واژگان کلیدی: تحلیل احساس، نظرکاوی، طبقه‌بندی قطبیت، مجموعه داده های تحلیل احساس، زبان فارسی

Sentiment analysis methods in Persian text: A survey

Zeinab Rajabi[*1], MohammadReza Valavi[2] & Maryam Hourali[3]
[1]Iran Telecommunication Research Center (ITRC), Tehran, Iran
[2,3]Department of Electronic and Computer, Malek-Ashtar University of Technology, Tehran, Iran

Abstract

With the explosive growth of social media such as Twitter and Instagram, reviews on e-commerce websites, and comments on news websites, individuals and organizations are increasingly using analyzing opinions in these media for their decision-making and designing strategies. Sentiment analysis is one of the techniques used to analyze users' opinions in recent years. The Persian language has specific features and thereby requires unique methods and models to be adopted for sentiment analysis, which are different from those in English and other languages. This paper identifies the characteristics and limitations of the Persian language. Sentiment analysis in each language has specified prerequisites; hence, the direct use of methods, tools, and resources developed for the English language in Persian has its limitations.

The present study aims to investigate and compare previous sentiment analysis studies on Persian texts and describe views presented in articles published in the last decade. First, the sentiment analysis levels, approaches, and tasks are described. Then, a detailed survey of the applied sentiment analysis

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

• تاریخ ارسال مقاله: ۱۳۹۸/۹/۲۴ • تاریخ پذیرش: ۱۴۰۰/۳/۲ • تاریخ انتشار: ۱۴۰۱/۷/۷ • نوع مطالعه: پژوهشی

[DOI: 10.52547/jsdp.19.2.107]

methods used for Persian texts is presented, and previous works in this field are discussed. The advantages and disadvantages of each proposed method are demonstrated. Moreover, the publicly available sentiment analysis resources of Persian texts are studied, and the characteristics and differences of each are highlighted.

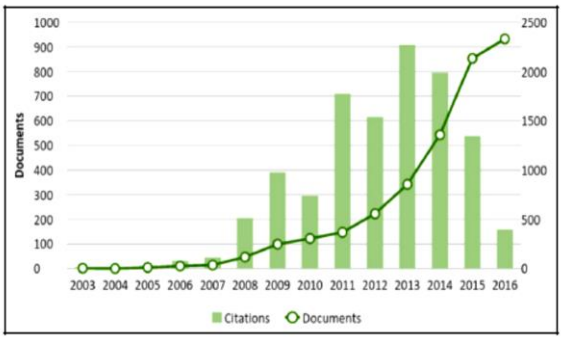
As a result, according to the recent development of the sentiment analysis field, some issues and challenges not being addressed in Persian texts are listed, and some guidelines are provided for future research on Persian texts. Future requirements of Persian text for improving the sentiment analysis system are detailed.

Keywords: Sentiment Analysis, Opinion Mining, Sentiment Classification, Sentiment Data Resource, Persian Language

۱- مقدمه

امروزه با گسترش شبکه‌های اجتماعی و سایت‌های نقدوبررسی، محتوای زیادی توسط کاربران در اینترنت تولید می‌شود. تحلیل محتواهای تولیدشده توسط کاربران کاربردهای سودمندی از قبیل بررسی رفتار مشتری، ذائقه‌سنجی کاربران، بهبود کسب‌وکار و تجارت الکترونیک را دارد [1].

تحلیل احساس یا نظرکاوی[1] حوزه مطالعاتی است که در آن نظر، احساس، ارزیابی‌ها، گرایش‌ها و تمایلات افراد نسبت به موجودیت‌هایی مانند محصولات، سرویس‌ها، سازمان‌ها، افراد، قضیه‌ها، موضوعات و ویژگی‌های هریک مورد تحلیل قرار می‌گیرد [2,3]. در این پژوهش عبارات نظرکاوی و تحلیل احساس به یک معنا به‌کارگرفته شده و همچنین در ادبیات فارسی اصطلاحات تحلیل سنجمان، اندیشه‌کاوی و عقیده‌کاوی نیز برای این حوزه استفاده شده است. تحلیل احساس با حوزه‌های پژوهشی مختلف مانند بازیابی اطلاعات، پردازش زبان طبیعی، داده‌کاوی، علوم شناختی محاسباتی و تحلیل شبکه‌های اجتماعی ارتباط نزدیک دارد.



(شکل-۱): تعداد اسناد و ارجاعات مربوط به "تحلیل احساس" طبق آمار WoS در سال‌های بین ۲۰۰۳ تا ۲۰۱۶ [4]

(Figure-1): The number of documents and citations related to 'sentiment analysis' filed according to the WoS (2003–2016).

در سال‌های اخیر حوزه پژوهشی تحلیل احساس[2] رشد قابل توجهی داشته است [4,5]. در شکل (۱) تعداد اسناد و ارجاعات مربوط به حوزه تحلیل احساس طبق آمار WoS بین سال‌های ۲۰۰۳ تا ۲۰۱۶ نشان داده شده است [4]. روند روبه رشد ارجاعات به مقالات تحلیل احساس در سال‌های اخیر نشان‌دهنده اهمیت چشم‌گیر موضوع است.

تعدادی از پژوهش‌ها مرورکلی بر روی روش‌های تحلیل احساس انجام داده‌اند [10-6] که تمرکز آنها روی روش‌های ارائه‌شده برای متون انگلیسی بوده است، علاوه‌بر آن Boudad و همکارانش [11] مروری روی تحلیل احساس در زبان عربی داشته‌اند، ولی تاکنون مروری روی پژوهش‌های تحلیل احساس در زبان فارسی انجام نشده است. زبان فارسی مورد توجه پژوهش‌گران کشورهای افغانستان، تاجیکستان و ایران است و از طرفی تحلیل متون تولیدشده به زبان فارسی در عرصه بین‌المللی نیز مورد نیاز است و بسیاری از کشورها نیز نیاز به نتایج تحلیل احساس در متون فارسی دارند. به‌عنوان نمونه در بستر تجارت الکترونیکی بسیاری از صاحبان برندها در کشورهای دیگر می‌خواهند در مورد میزان مقبولیت برند خود و رقبایشان در ایران اطلاعات به‌دست بیاورند و برنامه بازاریابی خود را گسترش بدهند. تعدادی از پژوهش‌ها از قبیل [14-12]، روی روش‌های چندزبانی تمرکز داشته‌اند ولی زبان فارسی در مجموعه آنها وجود ندارد. در همین راستا، بسیاری از پژوهش‌گران علوم رایانه به دنبال این هستند که بدانند پژوهش‌های تحلیل احساس متون فارسی در چه جایگاهی قرار دارند و چه کارهایی انجام شده و چه چالش‌هایی باقی‌مانده است و خلاءهای باقی‌مانده در کدام بخش‌ها وجود دارد و نیازمندی پژوهشی پروژه تحلیل احساس کدام‌ها هستند؛ بنابراین پژوهش انجام‌شده در این مقاله مورد توجه قرار می‌گیرد. از دیرباز تاکنون پژوهش‌ها در زمینه تأثیرگذاری و احساسات، در بسیاری از علوم، مورد توجه قرار گرفته است. علومی مانند زبان‌شناسی، روان‌شناسی، جامعه شناسی، علوم شناختی[3]، بازاریابی و علوم ارتباطات در این دسته قرار می‌گیرند [15]. بسیاری از پژوهش‌گران در این

[1] Opinion mining
[2] Sentiment analysis
[3] Cognitive science

[DOI: 10.52547/jsdp.19.2.107]

حوزه‌ها می‌خواهند از دستاوردهای حوزه تحلیل احساس استفاده نکنند و زبان مورد نظر آنها فارسی است؛ بنابراین مرور و بررسی روش‌های تحلیل احساس به آنها نیز کمک خواهد کرد. در همین راستا، در این مقاله به بررسی روش‌های استفاده‌شده در زبان فارسی و مقایسه آنها می‌پردازیم. همچنین پوشش آنها را نسبت به آخرین مسائل و رویکردهای علمی تحلیل احساس مورد بحث قرار داده و درنهایت پیشنهاداتی برای بهبود کارهای انجام‌شده در زبان فارسی با توجه به آخرین پیشرفت‌های علمی در این حوزه خواهیم داشت.

در بخش ۲ محدودیت‌های تحلیل احساس متون فارسی از سه زاویه بیان خواهد شد. در بخش ۳ سطوح مختلف تحلیل احساس، در بخش ۴ رویکردهای تحلیل احساس و در بخش ۵ وظایف تحلیل احساس مرور می‌شود. در بخش ۶ روش‌های ارائه شده برای متون فارسی مورد بررسی قرار خواهد گرفت. در بخش ۷ روش‌های ایجاد واژگان حسی مورد مقایسه قرار گرفته شده است. در بخش ۸ نیز پیکره‌های[1] استاندارد و در دسترس زبان فارسی را تشریح و در بخش ۹ به مسائل و خلاءهای تحلیل احساس در زبان فارسی پرداخته و پیشنهاداتی برای آینده خواهیم داد. درنهایت نیز در بخش ۱۰ نتیجه‌گیری خواهیم کرد.

۲- محدودیت‌های تحلیل احساس متون فارسی

۲-۱- محدودیت‌های پیش پردازشی

زبان فارسی حاوی ۳۲ حرف الفبا می‌باشد که بر عکس زبان انگلیسی از سمت راست به چپ نوشته می‌شود. زبان فارسی ساختار و گرامر مختص به خود را دارد که متفاوت از دیگر زبان‌ها است. در پردازش زبان طبیعی پیش پردازش‌گرهایی مانند لم‌یاب[2]، ریشه‌یاب[3]، جمله‌یاب[4]، واژه یاب[5]، برچسب زن ادات سخن[6]، عبارت یاب[7] وجود دارد که همگی در تحلیل احساس مورد استفاده قرار می‌گیرد. هر چه پیش‌پردازش‌ها در زبان فارسی با دقت بهتری انجام شود، پایه مناسبی برای مراحل بعدی ماست. نخست‌این‌که، وجود کلمات مرکب در این زبان استخراج کلمه را سخت کرده است، کلماتی مانند "پرداخت‌کننده"، "پشتیبانی‌کننده" دارای فاصله هستند و تشخیص آن از کلمات جدا نیاز به دقت بیشتر دارد. دوم‌این‌که، کلمات محاوره‌ای[8] و غیررسمی در زبان اندک نیستند و باید در تحلیل احساس روش‌های متناسب آن لحاظ شود."واسه"، "لپ‌تاپارو"، "فک کنم" نمونه‌هایی از کلمات محاوره‌ای است که در متن نظرات بسیار دیده می‌شود. سوم‌این‌که، وجود نیم‌فاصله در کلماتی مانند "کتاب‌های" شناسایی واژه‌ها را متفاوت کرده است و کاربران هنگام نوشتن نظرات از نیم‌فاصله استفاده نمی‌کنند؛ بنابراین برای تشخیص کلمات[9] در زبان فارسی نیازمند روش‌هایی با دقت بیشتر هستیم.

Basiri و همکارانش [16] و همچنین Asgarian و همکارانش [17] اثر پیش‌پردازش‌های مختص زبان فارسی در مسئله تحلیل احساس فارسی را مورد بررسی قرار داده‌اند و تأثیرگذاری پیش‌پردازش‌ها را تأیید کرده‌اند. از آن‌جایی که در زبان فارسی ابزارهای پیش‌پردازشی قوی مانند زبان انگلیسی در دسترس نیست، ممکن است بعضی روش‌های تحلیل احساس که در زبان انگلیسی به دقت خوبی رسیده‌اند در زبان فارسی به همان نتایج دست نیابند. همچنین ممکن است برخی روش‌ها یا ویژگی‌های به‌دست‌آمده در زبان انگلیسی برای طبقه‌بندی متون کارایی لازم را در زبان فارسی نداشته و نیازمند اصلاحات چشم‌گیری برای مؤثر واقع‌شدن باشند.

۲-۲- محدودیت وابستگی به فرهنگ

مسأله دیگر در تعیین قطبیت[10] و طبقه‌بندی نظر[11] ذهنیت افراد و طرز تفکر جامعه است. ذهنیت افراد نسبت به مثبت و منفی‌بودن واژه‌ها و کلمات متفاوت است. فرهنگ و ویژگی‌های هر جامعه‌ای در تعیین مثبت و یا منفی‌بودن نظر تأثیرگذار است. معیار بدی و خوبی در هر جامعه‌ای متفاوت از جامعه دیگر است. ممکن است در جامعه‌ای کلمه‌ای مثبت در نظرگرفته و در جامعه‌ای دیگر منفی یا خنثی تلقی شود [18]، یا جمله از نظر جامعه‌ای قضاوت مثبت نسبت به موضوعی و از نظر جامعه‌ای دیگر منفی تلقی شود. منابع واژگان حسی[12] زبان انگلیسی کمک بسیار زیادی در تعیین قطبیت می‌کنند؛ ولی نخست‌این‌که منابع به طور عمومی ایجاد شده‌اند در حالی

[1] Corpus
[2] Lemmatizer
[3] Stemmer
[4] Sentence tokenizer
[5] Tokenizer
[6] Part of speech (POS) tagger
[7] Chunker

[8] Colloquial word
[9] Tokenize
[10] Polarity detection
[11] Sentiment classification
[12] Sentiment lexicon resource

[DOI: 10.52547/jsdp.19.2.107]

که هر جامعه‌ای اصطلاحات، عبارات و کنایه‌های دارای قطبیت مخصوص به خود را دارد. دوم‌این‌که، ذهنیت‌ها و ارزش‌های یک جامعه در قطبیت یک نظر نقش تعیین‌کننده دارد. به‌عنوان نمونه در فرهنگ ایرانی "تبعیت از ولایت" مثبت تلقی می‌شود ولی در فرهنگ دیگری برای آن قطبیت لحاظ نشود؛ بنابراین نیازمند منابعی هستیم که قطبیت نسبت داده‌شده به عناصر آن، متناسب با فرهنگ جامعه ایران باشد. منابع حسی انگلیسی و ترجمه آنها به زبان فارسی ممکن است از نظر قطبیت به‌ویژه در حوزه سیاسی و اجتماعی در تضاد و از کیفیت لازم برخوردار نباشد.

۳-۲- محدودیت منابع و پیکره‌های استاندارد

در روش‌های تحلیل احساس به متون، جملات و یا واژه‌هایی که دارای برچسب قطبیت هستند، نیاز داریم. داده‌های دارای برچسب قطبیت نقش کلیدی در روش‌های یادگیری بانظارت دارند. روش‌های یادگیری شبکه‌های عصبی عمیق که به‌تازگی عملکرد خوبی داشته‌اند نیز نیاز به داده‌های برچسب‌دار با حجم زیادی دارند؛ علاوه بر این روش‌های ترکیبی که از منابع واژگان حسی در کنار روش یادگیری استفاده می‌کنند باعث افزایش دقت به‌طور چشم گیری می‌شوند؛ بنابراین از دیگر چالش‌های این حوزه می‌توان به هزینه‌بربودن تهیه داده‌های برچسب‌دار مورد نیاز در روش‌های نظارتی اشاره کرد. در زبان انگلیسی منابع مناسبی تهیه شده است، ولی در زبان فارسی هنوز نیازمند ساخت چنین منابعی هستیم. ساخت پیکره‌های استاندارد نظرات کاربران در سطح جمله، در سطح ویژگی و در سطح سند به همراه برچسب قطبیت آنها از دغدغه‌های بسیاری از پژوهش‌گران در زبان فارسی بوده است.

یکی دیگر از دغدغه‌هایی که پژوهش‌گران برای تحلیل متون فارسی دارند، ساخت منابع واژگان حسی است. برچسب‌گذاری واژگان حسی در یک زبان و همچنین برچسب‌گذاری برای واژگان دامنه خاص بسیار کار پرزحمت و زمان‌بری است؛ علاوه بر این در بسیاری از روش‌های تحلیل احساس برچسب‌گذاری واژه‌ها کافی نیست و بهتر است الگوهای زبانی[1] ، مفاهیم[2] و حتی نتایج تجزیه وابستگی[3] برای دست‌یابی به کارایی بیشتر برچسب‌گذاری شوند.

۳- سطوح مختلف تحلیل احساس

در پژوهش‌ها، تحلیل احساس در سه سطح انجام می‌شود [2]:

تحلیل احساس در سطح سند

در تحلیل احساس در سطح سند کل سند مورد طبقه‌بندی قرار می‌گیرد و قطبیت مثبت یا منفی سند تعیین می‌شود. مانند نقد و بررسی یک محصول که سامانه تعیین می‌کند که آیا کاربر با نقد و بررسی مطرح شده روی‌هم‌رفته نظر مثبت یا منفی نسبت به محصول دارد. در این سطح، فرض بر این است که در سند فقط در مورد یک موجودیت (مانند محصول) نظر داده شده است. بنابراین تحلیل احساس در این سطح برای ارزیابی‌هایی که دو موجودیت را باهم مقایسه می‌کنند ارزشی ندارد.

تحلیل احساس در سطح جمله

در این سطح، جملات مورد بررسی قرار می‌گیرد و مثبت، منفی یا خنثی‌بودن جمله تعیین می‌شود.

تحلیل احساس در سطح جنبه/ویژگی

تحلیل‌های سطح سند و سطح جمله به‌طوردقیق کشف نمی‌کند که افراد چه چیزی را دوست دارند و چه چیزی را دوست ندارند. در سطح ویژگی تحلیل‌ها را با دانه‌بندی ریز[4] انجام می‌شود. به تحلیل‌های سطح جنبه، سطح ویژگی یا مبتنی بر ویژگی[5] نیز گفته می‌شود. در بسیاری از کاربردها، هدف نظر توسط موجودیت و یا جنبه‌های مختلف موجودیت تعیین می‌شود؛ بنابراین هدف از این سطح از تحلیل، شناسایی قطبیت نسبت به موجودیت‌ها و یا جنبه‌های آن است، به‌عنوان مثال، جمله" کیفیت تماس گوشی خوب است ولی طول عمر باطری کوتاه است" دو جنبه کیفیت تماس و طول عمر باطری را ارزیابی می‌کند که قطبیت نظر روی اولی مثبت ولی قطبیت روی دومی منفی است. کیفیت تماس و طول عمر باطری هدف نظر هستند. براساس تحلیل‌های این سطح، خلاصه ساخت‌یافته از نظرات در مورد موجودیت‌ها و جنبه‌های آنها می‌تواند تولید شود، که متن بدون ساختار را به داده ساخت‌یافته تبدیل می‌کند و می‌تواند برای همه نوع تحلیل‌های کمی و کیفی مورد استفاده قرار گیرد.

۴- رویکردهای تحلیل احساس

رویکردهای تحلیل احساس به طورکلی به چهار دسته تقسیم می‌شوند [19] رویکرد مبتنی بر واژگان حسی،

[1] Linguistic pattern
[2] Concept
[3] Dependency parsing
[4] Finer-grained
[5] Feature-based

[DOI: 10.52547/jsdp.19.2.107]

رویکرد یادگیری ماشین، رویکرد مبتنی بر مفهوم و رویکرد ترکیبی. مبنای رویکردهای مبتنی بر واژه، حضور کلمات حسی و قطبیت آنها است. مهم‌ترین عامل تعیین‌کننده قطبیت در تحلیل احساس، واژگان حسی[1] است که به آنها کلمات نظر[2] نیز گفته می‌شود. واژگان حسی به کلماتی گفته می‌شود که برای بیان احساسات مثبت یا منفی مورد استفاده قرار می‌گیرد [2]. به‌عنوان مثال از کلماتی مانند خوب، عالی، وحشتناک و بد می‌توان نام برد. در همین راستا، منابع واژگان حسی، به همراه قطبیت هرکدام از آنها برای روش‌های نظرکاوی تدارک دیده شده است. قطبیت متن بر اساس قطبیت مجموع کل واژگان حسی مثبت یا منفی موجود در یک متن تعیین می‌شود. واژه‌نامه حسی شامل مجموعه‌ای از کلمات و عبارات که امتیازات منفی و مثبت به آنها اختصاص داده شده است. در مقایسه با رویکرد یادگیری ماشین، روش مبتنی بر واژه یک روش مناسب برای تحلیل احساس است، زیرا در آن منابع کمتری استفاده می‌شود و به برخی از مجموعه‌های داده حاشیه‌نویسی شده نیازی ندارد. علاوه بر این، از واژه‌نامه حسی می‌توان در رویکرد یادگیری ماشین به‌منظور ایجاد ویژگی‌های حسی نیز استفاده کرد. با این حال، یک رویکرد مبتنی بر واژه از عدم پوشش واژگان حسی در یک متن رنج می‌برد. در مقایسه با رویکرد مبتنی برواژه، رویکرد مبتنی بر یادگیری ماشین از متداول‌ترین الگوریتم‌های یادگیری ماشین برای طبقه‌بندی احساس استفاده می‌کند. رویکردهای یادگیری ماشین تحلیل احساس می‌تواند به سه شیوه یادگیری با نظارت[3]، یادگیری بدون نظارت[4] و نیمه‌نظارتی انجام شود. در بین این سه شیوه، یادگیری با نظارت به نتایج موفقیت‌آمیزتری دست یافته است. روش‌های با نظارت از حجم زیادی از داده‌ها که از قبل برچسب قطبیت خورده است استفاده می‌کنند. طبقه‌بندی نظر با استفاده از یادگیری با نظارت به‌طورمعمول به‌عنوان یک مسأله طبقه‌بندی[5] به دو ردهٔ مثبت و منفی فرموله شده است. داده‌های آموزشی و آزمایشی به‌طور معمول برای نقد و بررسی محصول مورد استفاده قرار می‌گیرد. از آنجا که این موضوع یک مسأله طبقه‌بندی متن ماست، از هر روش یادگیری با نظارت موجود می‌تواند استفاده شود، به‌عنوان مثال، طبقه‌بند بیزین ساده، SVM [20,21] و بیشینه آنتروپی. مقاله [22] برای نخستین‌بار این رویکرد را برای طبقه‌بندی نقدوبررسی‌های فیلم به دو دسته نظرات مثبت و منفی بکارگرفته است. در این مقاله طبقه‌بند بیزین ساده و SVM به‌کارگرفته شده و از unigram و کیسه کلمات به‌عنوان ویژگی طبقه‌بندی قطبیت استفاده شده است. در پژوهش‌های بعدی، بسیاری از ویژگی‌های دیگر و الگوریتم‌های یادگیری توسط تعداد زیادی از پژوهش‌گران مورد آزمایش قرار گرفت. مانند دیگر کاربردهای یادگیری ماشین با نظارت، کلید طبقه‌بندی نظر مهندسی مجموعه‌ای از ویژگی‌های مؤثر است. برخی از ویژگی‌های استفاده‌شده توسط پژوهش‌گران عبارتند از [2,3]: واژگان حسی، عبارت و نرخ رخداد آن، ادات سخن، قوانین نظرات، شیفت‌دهنده‌های احساس[6] (تغییردهنده معنا)، وابستگی نحوی.

برخی پژوهش‌ها روش بدون نظارت را پیش گرفته‌اند. به‌عنوان نمونه در پژوهش [23]، از آنجا که واژگان حسی فاکتور غالب برای طبقه‌بندی احساسات هستند، کلمات و عبارات حسی ممکن است برای طبقه‌بندی احساسات در شیوه‌ای بدون نظارت استفاده شده است. در این مقاله طبقه‌بندی را بر اساس برخی الگوهای نحوی ثابت که به احتمال زیاد برای بیان نظرات مورد استفاده قرار گیرد، انجام می‌دهد. الگوهای نحوی بر اساس برچسب‌های ادات سخن تشکیل شده است. در ادامه با کمک معیار PMI احتمال رخداد الگوی نحوی را به واژه‌نامه حسی که دارای قطبیت روشن است در داده‌های آموزشی محاسبه می‌شود.

روش‌های با نظارت عملکرد خوبی را از خود نشان داده‌اند؛ ولی مشکل آنها این است که نیاز به حجم داده برچسب‌گذاری‌شدهٔ بالایی دارند. به همین دلیل پژوهش‌گران به سراغ روش‌های نیمه‌نظارتی [24] در تحلیل احساس که نیاز به حجم کمتری از داده‌های برچسب خورده دارند رفته‌اند.

Cambria [19,25,26] دسته سومی به این دسته‌بندی اضافه کرده است که تحلیل احساس در سطح مفهوم[7] نام دارد. روش‌های تحلیل احساس در سطح مفهوم، روی تجزیه و تحلیل معنایی متن از طریق استفاده از هستان‌شناسی یا شبکه‌های معنایی تمرکز دارند، که اجازه تجمیع اطلاعات مفهومی و عاطفی در ارتباط با نظرات زبان طبیعی را می‌دهند. با تکیه بر پایگاه‌های

[1] Sentiment Lexicon
[2] Opinion words
[3] Supervised Learning
[4] Unsupervised Learning
[5] Classification problem
[6] Sentiment shifters
[7] Concept-level sentiment analysis

[DOI: 10.52547/jsdp.19.2.107]

بزرگ دانش معنایی، چنین روش‌هایی از استفاده از کلمات کلیدی به‌صورت کورکورانه و شمارش تعداد کلمات هم‌رخداد دوری گزیده، و به جای آن به ویژگی‌های ضمنی در ارتباط با مفاهیم زبان طبیعی تکیه می‌کنند. بر خلاف روش‌های صرفاً نحوی، روش‌های مبتنی بر مفهوم قادر به تشخیص قطبیت‌که به شیوه‌ای ظریف بیان می‌شود، هستند. این روش‌ها قادر به تشخیص قطبیت جملاتی که به‌صراحت احساسی منتقل نمی‌کنند، هستند و تلاش می‌کنند به‌طور ضمنی از طریق پیوند با دیگر مفاهیم به قطبیت جمله دست پیدا کنند. چارچوب CLSA(Concept-Level Sentiment Analysis) [27] چارچوبی است که برای روش‌های تحلیل احساس در سطح مفهوم استفاده می‌شود. روش‌های تحلیل احساس در سطح مفهوم در ترکیب با دسته‌بندی‌های قبلی به نتایج بهتری برای طبقه‌بندی نظر دست پیدا کرده است [28,29] رویکردهای ترکیبی، از رویکردهای یادشده به‌صورت ترکیبی استفاده می‌کنند و از مزیت‌های هر قسم بهره می‌برند.

۵- وظایف تحلیل احساس

وظایف مهم که در تحلیل احساس مورد توجه قرار می‌گیرند عبارت است از [7,27]:

- **تشخیص ذهنیت**[1]: در این وظیفه متنی که حاوی نظر است از متنی که حاوی نظر نیست جداسازی می‌شود. هدف تشخیص ذهنیت، طبقه‌بندی متن به‌صورت خودکار به جملات ذهنی یا حاوی نظر در مقابل جملات عینی و یا خنثی است. از این وظیفه تحلیل احساس برای شناسایی جملات عینی و حذف جملاتی که نظر و رأیی در آنها نیست می‌توان استفاده کرد.
- **تشخیص قطبیت یا طبقه‌بندی نظر**: تشخیص قطبیت اساسی‌ترین و مشهورترین وظیفه در تحلیل احساس است. تا حدی که در بسیاری از پژوهش‌ها عبارت *تشخیص قطبیت* و *تحلیل احساس* به‌طور جایگزین استفاده می‌شود. در وظیفه تشخیص قطبیت یا طبقه‌بندی احساس تلاش می‌شود با روش‌های با نظارت و بدون نظارت و یا مبتنی بر واژه، نظرات را در دو ردهٔ مثبت/ منفی یا سه ردهٔ مثبت/منفی/خنثی طبقه‌بندی کرد.
- **ایجاد منابع واژگان حسی:** یکی دیگر از وظایف اساسی تحلیل احساس، ایجاد منابع واژگان حسی با رویکرد مناسب است. در سال‌های اخیر، ایجاد واژگان حسی حجم بالایی از پژوهش‌های تحلیل احساس را به خود اختصاص داده است. به‌کارگیری منابع واژگان حسی در روش مبتنی بر واژگان رایج است. همچنین استفاده از واژگان حسی برای تقویت ویژگی طبقه‌بندی در رویکردهای ترکیبی، به‌طورکامل پر رنگ ماست.
- **استخراج ویژگی یا جنبه**[2]: در این وظیفه، هدف استخراج جنبه‌ها یا ویژگی‌های یک محصول، خدمات و غیره از متن است. به‌عنوان نمونه صفحه نمایش و باتری ویژگی‌های یک گوشی تلفن همراه هستند. تعداد سه پژوهش در زبان فارسی روی این وظیفه تمرکز داشته‌اند که در اینجا به بررسی آنها می‌پردازیم. گلپر رابوکی و همکارانش [30] مدلی برای استخراج جنبه بر اساس تکرار اسم در متن ارائه داده‌اند. آنها در پژوهش دیگر [31] مدل خود را بهبود داده و استخراج جنبه بر اساس تکرار و انتشار دوگانه مورد بررسی قرار داده‌اند. آنها همچنین در پژوهش سوم [32] خود روش استخراج جنبه بر اساس تجزیه وابستگی نحوی را پیشنهاد داده و به نتایج بهتری نسبت به دو روش قبل دست‌یافته‌اند.
- **تشخیص نظرات جعلی**[3]: در این وظیفه نظرات حقیقی از نظرات جعلی جداسازی می‌شود که خود مشتمل بر رویکردهای با نظارت و بدون نظارت است. د واقع وظیفه تشخیص نظرات جعلی، مسأله قراردادن نظرات در دو ردهٔ نظرات هرز و غیرهرز است [33].
- **تشخیص کنایه**[4]: سخن طعنه‌آمیز همیشه متوجه کسی یا چیزی است. هدف از طعنه فرد یا شیئی است. هدف می‌تواند خود نویسنده، مخاطب یا شخص ثالث باشد. وجود جمله طعنه‌آمیز می‌تواند معنای جمله را به‌طور کامل تغییر دهد؛ بنابراین ممکن باعث تفسیر اشتباه از جمله شود. با وجود اینکه طعنه و کنایه از جنبه روان‌شناسی و زبان‌شناسی به خوبی مورد مطالعه قرار گرفته است، ولی در علوم محاسباتی هنوز بسیار جای کار دارد. یکی از وظایف در تحلیل احساس تشخیص کنایه است.
- **مدیریت معکوس‌کننده‌ها**: در یک سامانه تحلیل احساس، نفی نقش اساسی در تشخیص قطبیت دارد و می‌تواند قطبیت را تغییر دهد. به عبارت دیگر، نفی ممکن است، قطبیت مثبت را به منفی و بالعکس تبدیل کند؛ ازاین‌رو، شناخت و مدیریت نفی در هر سامانه تحلیل احساس ضروری است. مهم‌ترین

[1] Subjectivity detection

[2] Feature extraction

[3] Fake review detection

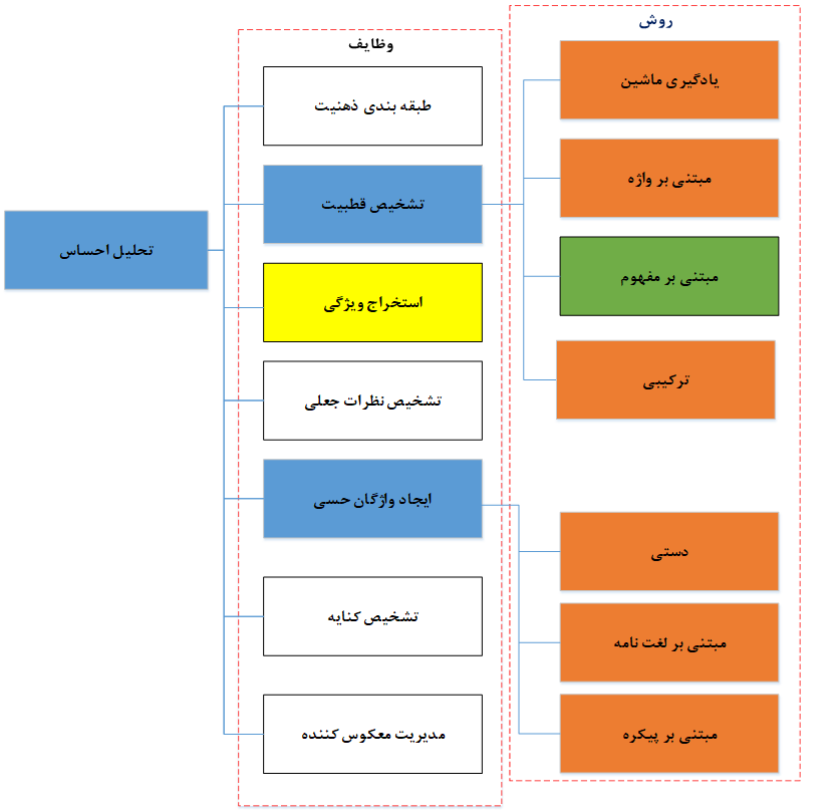
[4] Sarcasm detection

[DOI: 10.52547/jsdp.19.2.107]

معکوس‌کننده‌ها وجود کلمات و نشانه‌های نفی مانند پیشوند نفی در افعال هستند، ولی نمی‌توان فقط به کلمات و نشانه‌های نفی اکتفا کرد و نیازمند درک معنایی عمیق‌تری از متن داریم و مدیریت آن پیچیدگی‌های خاص خود را دارد [34].

۶- بررسی روش‌های ارائه‌شده برای تحلیل احساس متون فارسی

زبان‌های مختلفی در دنیا وجود دارد و هر زبانی دارای خصوصیات مختص به خود می‌باشد. به طورکلی روش‌های پردازش زبان طبیعی و به طور خاص روش‌های تحلیل احساس باید ویژگی‌های زبان را در نظر بگیرد. پژوهش [4] زبان‌هایی که بیشترین پژوهش‌های تحلیل احساس را به خود اختصاص داده یاد کرده است که زبان انگلیسی، اسپانیایی، ترکی و چینی در صدر و زبان فارسی در این فهرست قرار ندارد. زبان فارسی ساختار زبانی مختص به خود دارد و همچنین محدودیت‌هایی که درقبل اشاره شد دارد و نیازمند روش‌های جدید اصلاح‌شده است. یک روش برای یک وظیفه تحلیل احساس در زبان‌های مختلف به نتایج و دقت‌های مختلفی دست‌یافته است. وظایف تحلیل احساس در بخش قبل توضیح داده شد. در شکل (۲) وظایف تحلیل احساس به همراه روش‌های هرکدام نمایش داده شده است. *خانه‌های آبی‌رنگ*، وظایفی را نشان می‌دهد که در زبان فارسی پژوهش‌هایی را به خود اختصاص داده است که این دو وظیفه اساسی تعیین قطبیت و ایجاد واژگان حسی است. *خانه‌های سفیدرنگ*، وظایفی هستند که پس از بررسی منابع مختلف، مقاله‌ای که روی متون فارسی متمرکز شده باشد، یافت نشده است، بنابراین بهتر است در پژوهش‌های آینده تحلیل احساس متون فارسی به آنها پرداخته شود. این وظایف عبارتند از: تشخیص کنایه، طبقه‌بندی ذهنیت، تشخیص نظرات جعلی، مدیریت معکوس‌کننده. *خانه زردرنگ*، وظیفه استخراج ویژگی یا جنبه را نشان می‌دهد که تعداد انگشت شمار پژوهش در ارتباط با آن انجام شده است که در بخش قبل توضیح داده شده است. *خانه‌های نارنجی‌رنگ* ، روش‌هایی که در زبان فارسی برای دو وظیفه اساسی تشخیص قطبیت و ایجاد واژگان حسی در منابع انگلیسی وجود داشته و کم‌وبیش در زبان فارسی هم به کارگرفته شده است نشان می‌دهد. همان‌طور که در شکل دیده می‌شود، از بین روش‌ها، تشخیص قطبیت مبتنی بر مفهوم در پژوهش‌های متون فارسی مورد توجه قرار نگرفته شده که با رنگ سبز نمایش داده شده است.



(شکل-۲): دسته‌بندی وظایف تحلیل احساس و روش‌های آن.

(Figure-2): The categories of sentiment analysis tasks and their methods

[DOI: 10.52547/jsdp.19.2.107]

آبی: *وظایفی* که در تحلیل احساس متون فارسی روی آنها پژوهش انجام شده است.
سفید: *وظایفی* هستند که پس از بررسی منابع مختلف مقاله‌ای برای آنها یافت نشده است.
زرد: *وظایفی* که به تعداد انگشت شماری پژوهش در زبان فارسی روی آن انجام شده است.
نارنجی: *روش‌هایی* که در زبان فارسی برای دو وظیفه تشخیص قطبیت و ایجاد واژگان حسی به‌کارگرفته شده است.
سبز: *روش‌هایی* که در پژوهش‌های متون فارسی مورد توجه قرار نگرفته شده است.

با توجه به پژوهش‌های جمع‌آوری شده روی تحلیل احساس در متون فارسی و تمرکز آنها بر دو وظیفه تشخیص قطبیت یا طبقه‌بندی نظر و ایجاد منابع واژگان حسی، در این مقاله این دو وظیفه اصلی و اساسی بیشتر مورد توجه قرار گرفته و بقیه وظایف می‌تواند گزینه‌های بسیار مناسبی برای پژوهش‌گران زبان فارسی باشد.

نخستین پژوهش روی تحلیل احساس در زبان فارسی در سال ۲۰۱۲ انجام شده است؛ بنابراین در این مقاله، مقالات چاپ شده از سال ۲۰۱۲ تا ۲۰۱۹ مورد بررسی قرار گرفته شده است و تمرکز روی مطالعه مقالات معتبر بین‌المللی و مقالات معتبر داخلی بوده است. تعداد مقالات مرورشده در تحلیل احساس فارسی که عمدتاً روی مسأله تشخیص قطبیت و ایجاد منابع متمرکز بوده‌اند ۲۱ مقاله است که گزارش جزئیات آن در جدول (۱) آورده شده است.

(جدول-۱): مقالات گردآوری‌شده برای تحلیل احساس متون فارسی

(Table-1): Collected articles for sentiment analysis in Persian texts

محل انتشار	زبان انشار	تعداد مقالات
کنفرانس	فارسی	۴
	انگلیسی	۷
مجلات	فارسی	۳
	انگلیسی	۸

Shams و همکاران [35] اولین پژوهش تحلیل احساس متون فارسی را در سال ۲۰۱۲ انجام داده‌اند. آنها اقدام به ایجاد مجموعه اولیه واژگان حسی با نام PersianClues کرده و از ترجمه خودکار واژه‌های حسی انگلیسی با نام Subjectivity Clues در کنار لغت‌نامه انگلیسی به فارسی استفاده کرده‌اند. در این مقاله با رفع خطاهای احتمالی و تصحیح واژگان حسی، مجموعه اولیه برای مراحل بعدی آماده‌سازی شده و در ادامه روش تحلیل احساس مبتنی بر LDA[1] با نام LDASA ارائه شده است. در این روش، طبقه‌بندی قطبیت بدون نظارت با استفاده از مجموعه اولیه و مبتنی بر موضوع انجام می‌شود. قطبیت نسبت به هر موضوع مشخص می‌شود و برای تعیین موضوع با روش LDA کلمات وزن‌دهی می‌شود. ارزیابی مجموعه اولیه واژگان حسی با تعداد مختلف با شروع از ۸۰ کلمه تا ۱۶۰ کلمه با فواصل ده‌تایی انجام شده و به صحتی درحدود ۹۱ تا ۱۰۰ درصد دست یافته شده است. درنهایت برای ارزیابی کل روش، الگوریتم SVM مورد انتخاب قرار گرفته و مجموعه اولیه واژگان حسی را به‌عنوان ویژگی به آن داده شده است. در این قسمت به صحت حدود ٪۷۸ = acc به‌طور میانگین در سه دامنه هتل، گوشی تلفن همراه و دوربین دست‌یافته است.

Bagheri و همکارانش [36,37] دو پژوهش در زمینه تحلیل احساس در زبان فارسی انجام داده‌اند. در این مقاله خصوصیات زبان فارسی از قبیل پسوندهای متنوع در فعل‌ها، ادات جمع متنوع و وجود کلمات محاوره‌ای و غیررسمی در نظر گرفته شده است. در ادامه معیار اطلاعات متقابل اصلاح شده[2] به‌عنوان ویژگی مناسب در متون زبان فارسی ارائه شده است. معیار اطلاعات متقابل اصلاح شده، بهبودیافته معیار اطلاعات متقابل[3] ماست. معیار معرفی شده در این پژوهش با معیارهای دیگر اطلاعات متقابل و TFV[4] مقایسه شده و نتایج حاکی از آن است که معیار معرفی‌شده به‌عنوان ویژگی به نتایج بهتری در متون زبان فارسی رسیده است. مقدار f-measure = 0.8172، مقدار مناسبی است که نویسندگان به آن دست یافته‌اند. همچنین تعداد ۸۲۹ نظر کاربران نسبت به محصولات گوشی تلفن همراه به‌عنوان داده این پژوهش در نظر گرفته شده است.

Vaziripour و همکارانش [38] در مقاله خود روند تغییر نظرات در توئیت‌های منتشرشده در مقطع مذاکرات هسته‌ای را تحلیل می‌کنند. آنها در کار خود یک بازه زمانی در نظرگرفته و توئیت‌ها را جمع‌آوری کرده‌اند. توئیت‌ها در در دامنه سیاسی هستند. این مقاله به عنوان نخستین پژوهش در زبان فارسی که روند تغییر نظرات را در بازه زمانی مورد توجه قرار داده و تحلیل احساس را به‌صورت پویا انجام داده، ارزشمند است. در این مقاله حدود ۲۷۴۰۳۲ توئیت از ۱۷۸۴۴ کاربر یکتا جمع‌آوری و

[1] Latent Dirichlet Allocation (LDA)

[2] Modified Mutual Information (MMI)
[3] Mutual Information (MI)
[4] Term Frequency Variance

[DOI: 10.52547/jsdp.19.2.107]

حاشیه‌نویسی[1] شده است. از طبقه‌بند SVM با ویژگی شناسایی شده با خوشه‌بند[2] Brown استفاده و به دقت ۷۰٪ دست‌یافته شده است. در این مقاله برای تشخیص زیرعنوان‌ها در بازه ۴ هفته روش LDA به‌کار گرفته شده است.

Amiri و همکارانش [39] از روش مبتنی بر واژگان حسی[3] جهت تحلیل احساس متون فارسی استفاده کرده‌اند. آنها حدود ۷۱۷۹ کلمه، صفت و عبارت را از منابع برخط زبان فارسی استخراج کرده‌اند. این کلمات مشتمل بر کلمات رسمی[4]، غیر رسمی[5]، کلمات استاندارد[6]، غیر استاندارد[7] و عبارات یک کلمه یا چندکلمه‌ای که در قطبیت[8] زمینه تأثیر می‌گذارند، هستند. کلمات در اختیار افراد داوطلب با سطوح مختلف تحصیلات، سطوح مختلف گروه‌های سنی و بخش‌های مختلف جامعه ایرانی از طریق شبکه‌های اجتماعی قرار داده شده است. در مواردی که در برچسب قطبیت توافق وجود نداشته، برچسب مورد بحث قرار گرفته و به‌طور دستی تعیین شده و یا به‌عنوان خنثی برچسب‌گذاری شده است. در نهایت منبع واژگان حسی برای فرآیند تحلیل احساس استفاده و به دقت ۶۹٪ دست‌یافته شده است.

Alimardani و Aghaie [40,41] روشی ترکیبی (با نظارت و مبتنی بر واژه) برای طبقه‌بندی نظر ارائه داده‌اند. این مقاله با استفاده‌ازSentiWordNet در زبان انگلیسی و ارتباطی که از طریق WordNet در زبان فارسی پیدا می‌کند، منبع واژگان حسی دارای قطبیت ایجاد کرده است. در این مقاله از سه طبقه‌بند SVM، بیزین ساده و رگرسیون ترابری برای طبقه‌بندی با ویژگی‌های بهبود یافته به‌وسیله منبع ایجاد شده استفاده شده است. سه ویژگی Present-Absent و TF وTF-IDF[9] مد نظر قرار داده شده و هرکدام در میزان قطبیت به‌دست‌آمده از منبع واژگان ایجاد شده ضرب شده است؛ درواقع از واژه‌نامه به‌عنوان عامل وزنی در ویژگی استفاده شده است. در ادامه آزمایش‌های متعددی در حالت‌های مختلف انجام شده و بهترین حالت به دقت ٪۸۵/۵ در طبقه‌بند SVM و ویژگی TF-IDF دست‌یافته شده و علاوه بر آن، آزمایش‌های متعدد دیگر در تأثیر تعداد نمونه‌های مثبت و منفی در دقت روش، مورد بررسی قرار داده و نتایج آن در مقاله یاد شده است.

Sadidpour و همکارانش در مقاله [42] مجموعه‌ای از الگوهای قطبیت را با کمک الگوهای زبانی برای تعیین قطبیت استخراج کرده‌اند. داده‌های این مقاله از سه پایگاه خبری از مارس ۲۰۱۳ تا دسامبر ۲۰۱۵ جمع‌آوری شده است که مشتمل ۲۰۸۰۰۰ خبر سیاسی است. پیکره ایجادشده از اخبار جمع‌آوری‌شده، مشتمل بر ۱۴۰۰۰ خبر بین دو کیلو بایت تا ده کیلوبایت است. این پژوهش روی متون سیاسی به زبان فارسی انجام شده و با کمک الگوهای استخراجی به صحت بیش از ۹۰٪ دست‌یافته است. مزیت قابل توجه این پژوهش کار روی داده‌های دامنه سیاسی است.

Asgarian و همکارانش [17] در سال ۲۰۱۸ پژوهش جامعی روی زبان فارسی انجام داده‌اند که برتری زیادی نسبت به دیگر پژوهش‌ها دارد. در این مقاله اثر پیش‌پردازش‌هایی مانند حذف هرزواژه‌ها[10] ، لم‌یاب، تبدیل کلمات غیررسمی به رسمی، غلط‌یابی املایی[11] روی تحلیل احساسات مورد ارزیابی قرار گرفته است. نتایج ارزیابی‌ها حاکی از آن است که به‌کارگیری حذف هرزواژه‌ها و تبدیل کلمات غیررسمی به رسمی تأثیر مثبتی دارند؛ ولی ریشه‌یاب و غلط‌یاب املایی تأثیر چشم‌گیری نداشته و علاوه‌بر آن در این مقاله، یک وردنت فارسی با نام فردوس‌نت ایجاد و در ادامه از آن استفاده شده است. همچنین یک منبع واژگان حسی با کمک فردوس‌نت و پیکره تولیدشده نقدوبررسی کالا ایجاد شده است. برای ایجاد منبع، دو روش در این مقاله بکارگرفته شده است. روش نخست، نگاشت قطبیت SentiWordNet انگلیسی و روش دوم از الگوریتم یادگیری نیمه‌نظارتی[12] مبتنی بر مدل مخفی مارکوف[13] استفاده کرده است. نویسندگان در ادامه پژوهش خود، مهندسی ویژه‌ای برای طبقه‌بندی انجام داده‌اند. به همین منظور آنها مجموعه ویژگی‌های مناسب برای طبقه‌بندی احساس که در آخرین رویکردهای طبقه‌بندی احساس در زبان انگلیسی مورد استفاده قرار گرفته است برای زبان فارسی مورد آزمایش قرار داده‌اند. ویژگی‌هایی که مورد مقایسه قرار گرفته شده است، مواردی از قبیل واژه‌های حسی، برچسب ادات سخن، امتیاز قطبیت واژه‌های حسی، بردارهای تعبیه کلمات در Word2Vec، بردارهای تعبیه کلمات مختص

[1] Annotated
[2] Cluster
[3] Lexicon-based
[4] Formal
[5] Informal
[6] Standard
[7] Obsolete
[8] Polarity
[9] Invert Document Frequency

[10] Stop word
[11] Spell checker
[12] Semi-supervised
[13] HMM (Hidden Markov Model)

[DOI: 10.52547/jsdp.19.2.107]

تحلیل احساس، n-gram کلمات، n-gram نویسه، TF-IDF و غیره هستند. طبقه‌بندهای مختلف مانند BayesNet، LibLinear، SMO، KNN، بیشینه آنتروپی و RandomForest روی ویژگی‌های یادشده آزمایش شده‌اند. این پژوهش نقشه راه خوبی در انتخاب طبقه‌بند و انتخاب ویژگی برای کسانی که می‌خواهند به تحلیل احساس در متون فارسی بپردازند، است.

در جدول (۲) مقایسه‌ای روی پژوهش‌های صورت‌گرفته برای تعیین قطبیت متون فارسی انجام شده است. در این جدول پژوهش‌ها را از نظر رویکرد کلی تعیین قطبیت، روش تعیین قطبیت یا نوع طبقه‌بند و ویژگی مورد استفاده در طبقه‌بندی، آیا منبعی برای پژوهش خود ایجاد کرده است یا خیر، دامنه‌ای که روی آن متمرکز بوده، تعداد نظرات موجود و نام مجموعه‌داده مورد استفاده، دقت و یا بهترین دقت دست‌یافته‌شده در پژوهش، مزیت پژوهش نسبت به بقیه کارها مورد مقایسه قرار گرفته است. هر یک از ستون‌ها در ادامه به‌تفصیل توضیح داده شده است:

۱- **رویکرد کلی:** همان‌طور که در بخش ۴ یادشده رویکردهای کلی در روش‌های تحلیل احساس وجود دارد که مشتمل بر مبتنی بر واژگان حسی، یادگیری ماشین با نظارت و بدون نظارت و ترکیبی (یادگیری ماشین و مبتنی بر واژگان حسی) و همچنین روش‌های مبتنی بر مفهوم است که در این جدول برای هریک از پژوهش‌ها مورد بررسی قرار می‌گیرد. همان‌طور که در جدول (۲) مشاهده می‌شود، پژوهش‌های [38-36] از رویکرد نظارتی برای طبقه‌بندی قطبیت استفاده می‌کنند. در پژوهش [39] رویکرد مبتنی بر واژه را برای تشخیص قطبیت انتخاب و در پژوهش [40,41] رویکرد ترکیبی به‌کارگرفته شده است. همان‌طور که مشاهده می‌کنید، رویکردهای بدون نظارت و در سطح مفهوم برای طبقه‌بندی جمله و سند در زبان فارسی استفاده نشده است.

۲- **الگوریتم:** در این ستون الگوریتم مورد استفاده مانند بیزین ساده، بیشینه آنتروپی، SVM، ترابری رگرسیون و غیره مورد بررسی قرار می‌گیرد.

۳- **ویژگی مورد استفاده در طبقه‌بندی**: وظیفه تشخیص قطبیت اغلب به‌عنوان یک مسأله طبقه‌بندی مدل‌سازی می‌شود که به ویژگی‌های استخراج شده از متن به منظور تغذیه یک طبقه‌بند، وابسته است. در این ستون، ویژگی در نظرگرفته‌شده در هریک از پژوهش‌ها مورد مقایسه قرار گرفته شده است. روش‌های طبقه‌بندی از ویژگی یا ویژگی‌هایی برای انجام طبقه‌بندی استفاده می‌کنند که می‌توانند کارایی‌های متعددی داشته باشند که در این ستون مورد بررسی قرار می‌گیرند. برای زبان انگلیسی ویژگی‌های متعددی مانند کیسه کلمات، عبارت و نرخ رخداد آن، ویژگی دودویی حضور با عدم حضور کلمات، برچسب ادات سخن، قوانین نظرات، شیفت‌دهنده‌های احساس و وابستگی نحوی و یا ترکیبی از آنها مورد آزمایش قرار گرفته است. تعدادی از ویژگی‌های پرکاربرد در طبقه‌بندی قطبیت در ادامه بیشتر توضیح داده شده است:

- **عبارت و نرخ رخداد آن:** این ویژگی واژگان منفرد و واژگان چندتایی[1] به همراه تعداد فراوانی می‌باشد، که از ویژگی‌های رایج است و در طبقه‌بندی سنتی متن مبتنی بر موضوع مورد استفاده قرار می‌گیرد. در برخی موارد، موقعیت کلمه نیز ممکن است در نظر گرفته شود. همچنین طرح وزن‌دهی TF-IDF از بازیابی اطلاعات ممکن است، بسیار استفاده شود. همانند طبقه‌بندی سنتی متون، این ویژگی‌ها برای طبقه‌بندی نظر بسیار موثر نشان داده شده است.
- **ادات سخن**[2]: ادات سخن هر کلمه می‌تواند مهم باشد. کلمات ادات سخن مختلف مانند صفت، قید، فعل و اسم دارند و ممکن است، متفاوت عمل کنند. به‌عنوان مثال، نشان داده شده است که صفت‌ها تأثیرگذاری بیشتری روی قطبیت دارند. بنابراین برخی از پژوهش‌گران با صفت به‌عنوان ویژگی خاص رفتار می‌کنند. البته می‌توان تمام برچسب‌های ادات سخن و واژگان چندتایی آن را به عنوان ویژگی استفاده کرد.
- **قوانین نظرات:** به غیر از کلمات و عبارت‌های حسی، بسیاری از عبارات و یا ترکیبات زبان وجود دارند که می‌تواند برای بیان احساسات و نظرات مورد استفاده قرار بگیرند. به‌عنوان نمونه معنای عبارت، ترکیبی از عناصر آن و قوانین گرامری حاکم بر آن است که با درنظرگرفتن همه آنها قطبیت مشخص می‌شود.
- **معکوس‌کننده‌ها:** معکوس‌کننده‌ها شیفت‌دهنده‌های احساس[3] باعث تغییر جهت در قطبیت می‌شوند. قطبیت را از مثبت به منفی و یا بالعکس تغییر می‌دهند. کلمات نفی مهم‌ترین دسته شیفت‌دهنده‌های قطبیت هستند. به‌عنوان مثال، جمله "من این دوربین را دوست ندارم" منفی است. همچنین چندین نوع

[1] N-grams
[2] Part of speech (POS)
[3] Sentiment shifters

[DOI: 10.52547/jsdp.19.2.107]

دیگر از شیفت‌دهنده قطبیت وجود دارد. چنین شیفت‌دهنده‌هایی را نیز باید با احتیاط مدیریت کرد؛ چون رخداد چنین کلماتی همیشه به معنی تغییر قطبیت نیست.

- **وابستگی نحوی:** ویژگی‌های مبتنی بر وابستگی کلمات که از تجزیه وابستگی[1] یا درخت وابستگی[2] تولید می‌شوند، مورد بررسی و آزمایش پژوهش‌گران قرار گرفته است.
- **واژه‌های حسی:** در بسیاری از پژوهش‌ها از واژه‌های حسی به همراه امتیاز قطبیت آنها به‌عنوان ویژگی استفاده شده است.

Saraee و Bagheri در مقاله [43] به مسأله انتخاب ویژگی و مقایسه ویژگی‌های مناسب برای طبقه‌بند بیزین در مسأله طبقه‌بندی قطبیت برای متون فارسی پرداخته‌اند. چهار ویژگی Term Frequency، Document Frequency Variance (TFV)، اطلاعات متقابل، اطلاعات متقابل اصلاح‌شده را مد نظر داشته‌اند و نتایج حاصل طبقه‌بندی نظر را در مقاله خود یاد کرده‌اند. روی‌هم‌رفته در این مقاله معیار اطلاعات متقابل اصلاح شده، f-measure بهتری نسبت به سه ویژگی دیگر داشته است. Asgarian و همکارانش در [17] مهندسی ویژگی برای طبقه‌بندی احساس انجام داده‌اند و مجموعه ویژگی‌های مناسب که در آخرین رویکردهای طبقه‌بندی قطبیت در زبان انگلیسی مورد استفاده قرار گرفته است، برای زبان فارسی مورد آزمایش قرار داده‌اند. ویژگی‌هایی که مورد مقایسه قرار گرفته است، مواردی از قبیل واژه‌های حسی، برچسب ادات سخن، امتیاز قطبیت واژه‌های حسی، بردارهای تعبیه کلمات در word2Vec، بردارهای تعبیه اختصاصی تحلیل احساس، n-gram کلمات ،n-gram نویسه، TF-IDF غیره می‌باشند. Alimardani و Aghaie [40,41] روشی ترکیبی (با نظارت و مبتنی بر واژه) برای طبقه‌بندی نظر ارائه داده‌اند. بعد از ایجاد منبع واژگان حسی دارای قطبیت، سه ویژگی Present-Absent و TF وTF-IDF[3] را مد نظر و هرکدام را در میزان قطبیت به‌دست‌آمده از منبع واژگان ایجادشده ضرب کرده‌اند. درواقع از واژه‌نامه به‌عنوان ویژگی وزنی و تقویت ویژگی استفاده کرده‌اند.

البته به طور کلی در تحلیل احساس در هر زبانی، به‌دلیل تنکی ویژگی[4] و عدم آگاهی از زمینه[5]، فرآیند مهندسی ویژگی اغلب در جذب اطلاعات رابطه‌ای غنی و تعیین قطبیت‌هایی که حاصل از پیوند بین مفاهیم متن و مفاهیم ذهنی است، موفقیت کمتری به‌دست می‌آورند. روش‌هایی که از تعبیه کلمات استفاده می‌کنند، تلاش می‌کنند، این خلأ را از بین ببرند [44].

۴- **آیا پژوهش منبع واژگان حسی را ایجاد و در کار خود به‌کارگرفته است:** تعدادی از پژوهش‌ها با تکیه بر منابع واژگان حسی روش خود را ارائه داده‌اند و یا منابع واژگان حسی را در کنار روش‌های یادگیری ماشین استفاده نموده‌اند. از آنجا که منابع واژگان حسی در زبان فارسی اخیرا ایجاد شده و انگشت شمار هستند و در برخی کاربردها کارایی لازم را ندارند، برخی از پژوهش‌ها اقدام به ایجاد منابع با روش‌های مختلف کرده‌اند. در این ستون این مسئله مورد بررسی قرارمی‌گیرد. در بخش ۷ و جدول (۴) با جزئیات بیشتر به بررسی و مقایسه پژوهش‌هایی که نوآوری کار خود را روی ایجاد واژگان حسی قرار داده و تلاش کرده‌اند از منبع ایجادشده برای تعیین قطبیت استفاده کنند، خواهیم پرداخت.

۵- **دامنه:** تحلیل احساس روی داده‌های نظرات کاربران در دامنه‌های مختلفی می‌تواند انجام شود. در زبان انگلیسی بعضی از روش‌های تحلیل به‌طور عمومی ارائه شده‌اند و بعضی روی دامنه خاصی از جمله فروش محصولات و خدمات، سیاسی، اقتصادی، پزشکی و اجتماعی متمرکز شده‌اند. در زبان فارسی نیز پژوهش گران در بعضی دامنه‌ها کار کرده‌اند ولی همه دامنه‌ها پوشش داده نشده که در جدول (۲) مورد اشاره قرار گرفته شده است. دامنه‌هایی که از شهرت بیشتری برخوردارند، عبارتند از:

- **دامنه محصولات و خدمات:** بیشتر مورد توجه سایت‌های تجارت الکترونیکی و تحلیل نظرات کاربران نسبت به محصولات و خدمات ارائه‌شده در پایگاه‌ها را در بر می‌گیرد. صاحبان کسب‌وکارها برای بهبود کسب وکار خود تلاش می‌کنند از نظرات مشتریان نسبت به محصولات و سرویس‌ها باخبر بشوند. بیش‌تر پژوهش‌های منتشرشده در مستندات علمی روی داده‌های مرتبط با محصولات و خدمات هستند. موضوعاتی مانند فیلم، هتل، رستوران و محصولات گوشی تلفن همراه، لپ‌تاپ و نرم‌افزار بیشتر مورد توجه بوده است.

[1] Dependency parser
[2] Parse tree
[3] Invert Document Frequency
[4] Feature sparsity
[5] Context awareness

[DOI: 10.52547/jsdp.19.2.107]

- **دامنه سیاسی:** در دامنه سیاسی بیشتر نظرات کاربران در توئیتر و سایت‌های خبری نسبت به موضوعات مختلف سیاسی مورد تحلیل قرار می‌گیرد. دامنه سیاسی با توجه به اصطلاحات خاص سیاسی، نیازمندی‌های خاص دارد.
- **دامنه اجتماعی:** دامنه اجتماعی از تنوع بسیار بالایی در موضوع برخوردار است. پژوهش‌هایی که در دنیا در این دامنه قرار می‌گیرند کمتر منتشر شده و سازمان‌های خاص و یا دولت‌ها به‌طورمعمول برای شناخت اجتماعی در این دامنه فعال هستند. تحلیل احساس روی داده‌های دامنه اجتماعی با توجه به اهمیت آن در آینده‌پژوهی ظرفیت زیادی در جهت‌دهی پژوهشی پژوهش‌های آینده دارد.
- **دامنه پزشکی:** در این دامنه نظرات کاربران نسبت به مسائل پزشکی، دارو و اثرات دارو مورد تحلیل قرارمی‌گیرد.

یک روش در یک دامنه مانند محصولات الکترونیک ممکن است به دقت به‌نسبه خوبی دست پیدا کند، ولی در دامنه دیگر کارایی لازم را نداشته باشد، به همین دلیل دامنه‌ای که داده‌ها روی آن تحلیل می‌شود، اهمیت پیدا می‌کند. در بعضی از دامنه‌ها منابع بیشتری برای پشتیبانی از وظیفه اصلی که طبقه‌بندی است، نیازمند هستیم. به‌عنوان نمونه Noferesti و همکارانش [45] در دامنه پزشکی به تحلیل نظرات کاربران نسبت به دارو در زبان انگلیسی پرداخته و برای پشتیبانی از طبقه‌بندی قطبیت، با بهره‌گیری از منابعی که در وب در دامنه دارو است، پایگاه دانشی از واقعیت‌های دارای قطبیت در این دامنه ایجاد کرده است.

با توجه به اهمیت دامنه یکی از ستون‌های جدول به دامنه اختصاص داده شده است. همان‌طور که در جدول (۲) مشاهده می‌شود، تمرکز پژوهش‌های [36,37,40] روی دامنه محصولات و خدمات مانند گوشی تلفن همراه و هتل است و پژوهش‌های [38,42] روی دامنه سیاسی و پژوهش‌های [39,46,47] به‌طور عمومی هستند. در دامنه‌های اجتماعی، پزشکی و اقتصادی پژوهشی برای زبان فارسی صورت نپذیرفته است و پژوهش‌گران می‌توانند در کارهای آینده به این دامنه‌ها وارد شوند.

۶- **مجموعه‌داده:** پژوهش‌ها برای ارزیابی روش پیشنهادی از مجموعه‌داده یا پیکره نظرات استفاده می‌کنند. برخی از مجموعه داده‌هایی که به‌طور استاندارد گردآوری شده و در اینترنت منتشر شده‌اند، استفاده می‌کنند و برخی دیگر خود اقدام به جمع‌آوری داده برای پژوهش کرده‌اند؛ بنابراین گزارشی از منبع جمع‌آوری داده و تعداد داده‌ها، تعداد داده‌های مثبت و منفی و یا تعداد داده‌های مجموعه آموزش و آزمایش آورده‌اند که در این ستون یاد شده است. این موضوع از آن جهت اهمیت دارد که دقت دست‌یافته‌شده برای چه حجمی از داده‌ها است.

۷- **دقت:** در این ستون دقت دست‌یافته‌شده در پژوهش یادشده است، در برخی از پژوهش‌ها معیار صحت (Accuracy) مورد اشاره قرار گرفته و در برخی دیگر معیارهای دقت (Precision) و فراخوانی (Recall). علاوه‌بر آن روش‌های ارائه‌شده در مقالات در حالات مختلف (به‌عنوان نمونه مجموعه داده‌های مختلف) دقت‌های مختلفی را گزارش داده‌اند و روش خود را مورد ارزیابی قرار داده‌اند که در جدول (۲) بالاترین دقت گزارش‌شده، یاد شده است.

۸- **مزیت:** در این ستون مزیت و برتری پژوهش انجام‌شده، مورد توجه قرار داده می‌شود.

همان‌طور که در جدول (۲) مشاهده می‌کنید، تعداد پژوهش‌ها روی زبان فارسی در مقایسه با کارهای انجام‌شده در زبان انگلیسی اندک هستند و هرکدام به جنبه‌ای پرداخته‌اند.

۷- ایجاد واژگان حسی در متون فارسی

یکی از وظایفی که در تحلیل احساس و نظرکاوی وجود دارد، ایجاد منابع واژگان حسی مناسب است. منابع واژگان حسی به همراه میزان قطبیت آنها در هر زبانی پشتوانه روش‌های مبتنی بر واژه و روش‌های که ترکیبی هستند. روش‌های ترکیبی از منابع واژگان حسی برای تقویت ویژگی‌ها در طبقه‌بندی قطبیت استفاده می‌کنند. به همین منظور در زبان انگلیسی منابعی برای تحلیل نظرات ایجاد شده است. در زبان‌های دیگر نیز بهتر است چنین منابعی ایجاد شده و از مزیت‌های آنها بهره گرفته شود. در این بخش به بررسی و مقایسه پژوهش‌هایی که نوآوری پژوهش خود را روی ایجاد واژگان حسی در زبان فارسی گذاشته‌اند، می‌پردازیم.

Dehdarbehbahani و همکارانش [47] در پژوهشی برای شناسایی میزان قطبیت از منبع خارجی WordNet برای ساخت شبکه معنایی چند زبانه[1] استفاده می‌کنند. در

[1] A multilingual semantic network

[DOI: 10.52547/jsdp.19.2.107]

این مقاله شبکه معنایی گراف‌های زبان انگلیسی و فارسی به‌هم وصل شده و روش قدم‌زدن تصادفی برای تعیین قطبیت کلمات فارسی بکار گرفته شده است. در ارزیابی‌های روش پیشنهادی مقدار MAP[1] = 0.658 و Acc =0.914 به‌دست آمده است.

نجف‌زاده و همکارانش [46] یک روش نیمه‌نظارتی در حوزه خودآموزی برای تحلیل احساس به زبان فارسی ارائه داده‌اند. این پژوهش، نخستین کاری است که روی زبان فارسی با روش نیمه‌نظارتی انجام شده است. در این مقاله برچسب‌های حاوی احساس به کمک یک واژه‌نامه خودساخته وفقی (پویا و بدون نیاز به خبره انسانی) تعیین شده، بنابراین به طور خودکار خصیصه‌های حسی را استخراج شده است. همچنین در این مقاله از طبقه‌بند مدل مخفی ماکوف خودناظر بر روی خصیصه یادشده در کنار قوانین برای فرآیند تحلیل احساس استفاده شده است. در روش ارائه‌شده به مقدار صحت Acc= ۸۹٪ دست یافته شده است. مزیت‌هایی که روش ارائه داده شده دارد عبارتند از: ۱)خودکارسازی استخراج خصیصه‌های حسی به جای استفاده از خبره انسانی ۲)استفاده از چندنگاشت‌های عامیانه، به‌عنوان رویکرد فرااکتشافی، برای پوشش‌دادن چالش‌های بررسی‌نشده زبان فارسی ۳)استفاده از ویژگی‌های زوج مرتبی(برچسب ادات سخن، برچسب قطبیت) به‌عنوان وضعیت‌های طبقه‌بند مدل مخفی مارکوف برای افزایش دقت تحلیل احساس در زبان فارسی.

عسگریان و همکارانش [48] در سال ۱۳۹۷ با استفاده از شبکه واژگان فردوس‌نت[2] و نگاشت واژگان حسی در زبان انگلیسی (SentiWordNet)، منبعی برای واژگان حسی در زبان فارسی با نام حس‌نگار[3] ایجاد کرده‌اند. برای تولید حس‌نگار، ابتدا با استفاده از نگاشت مفاهیم (گروه‌های هم‌معنی) در شبکه واژگان پرینستون به زبان فارسی، شبکه واژگان جامع (فردوس‌نت) ساخته شده است. در ادامه، با استفاده از فردوس‌نت، میزان بار قطبیت محاسبه‌شده برای هر گروه هم‌معنی در شبکه واژگان حسی انگلیسی به گروه‌های هم‌معنی متناظر با آن در حس‌نگار نگاشت می‌شود. واژه‌های حسی زبان فارسی در چهار مقوله صفت، اسم، فعل و قید دسته‌بندی شده است. دقت precision=89% برای واژگان حسی تولیدی گزارش شده است. علاوه بر آن در این مقاله پیکره‌ای حاوی ۳۰۸۰ نظر از نظرات کاربران در سایت دیجی‌کالا با کمک ارزیاب‌ها برچسب‌گذاری شده و از این پیکره برای ارزیابی کار بهره گرفته شده است.

Dehkharghani [18] در سال ۲۰۱۹ منبع واژگان حسی مختص زبان فارسی را با نام Sentifars و با روش مبتنی بر ترجمه ایجاد کرده است. از مجموعه چهار منبع واژگان حسی موجود برای زبان انگلیسی و ترجمه آنها واژه‌های حسی مورد نیاز پژوهش استخراج شده است. در ادامه با روش دستی واژه‌های حسی برچسب‌گذاری شده تا داده‌های مقاله برای یادگیری با نظارت به روش ترابری رگرسیون آماده‌سازی شود. طبقه‌بند ترابری رگرسیون قطبیت واژگان حسی فارسی را از روی امتیاز قطبیت معادل آنها در منابع زبان انگلیسی یاد می‌گیرد. در واقع قطبیت معادل واژه‌ها در منابع زبان انگلیسی به‌عنوان ویژگی طبقه‌بندی در نظرگرفته شده است. بهترین مقدار به‌دست‌آمده برای معیار صحت[4] در تعیین قطبیت واژه‌ها Acc=95.92 است که در آن هر چهار منبع و قطبیت آنها لحاظ شده‌اند. منبع به‌طور عمومی و مستقل از دامنه تولید شده است.

Sabeti و همکارانش [49] منبع واژگان حسی لِکسی‌پِرس (LexiPers) را با روشی نیمه‌نظارتی با کمک هستان‌شناسی فارس‌نت ایجاد کرده‌اند. در این مقاله مجموعه دانه را انتخاب و با استفاده از هوش انسانی آنها را برچسب‌گذاری کرده‌اند؛ سپس با روش نیمه‌نظارتی و به‌کارگیری الگوریتم PMI مجموعه دانه گسترش داده شده است.

به طور کلی روش‌های ایجاد منابع واژگان حسی به سه دسته کلی تقسیم می‌شوند، در جدول (۳) سه روش در سطح کلان مورد اشاره قرار گرفته و مزایا و معایب هرکدام و مقالاتی که در زبان فارسی از آن روش استفاده کرده‌اند، اشاره شده است.

۱. **روش دستی:** در این روش از افراد خبره درخواست می‌کنند که واژه‌ها را برچسب‌گذاری کنند. در صورت اختلاف نظر روی برچسب‌ها رأی اکثریت مورد پذیرش قرار می‌گیرد. معیار کاپا[5] برای ارزیابی میزان اطمینان مورد استفاده قرار می‌گیرد.

مزیت روش‌های دستی در مقایسه با دو روش دیگر دقت بالا است. برچسب‌گذاری به روش‌های دستی کار سخت و بسیار زمان‌بری است و هزینه‌های زیادی دارد

[1] Mean average precision

[2] FerdosNet
[3] HesNegar

[4] Accuracy
[5] Kappa

[DOI: 10.52547/jsdp.19.2.107]

که از معایب آن به شمار می‌رود [50]. فقط حجم داده‌های مشخصی را در یک بازه زمانی می‌توان برچسب‌گذاری کرد. در پژوهش‌های فارسی [39,51,52] واژگان را به روش دستی برچسب‌گذاری کرده‌اند.

۲. **روش مبتنی بر لغت‌نامه:** دسته دوم روش‌هایی هستند که با تکیه بر لغت‌نامه‌هایی مانند WordNet و همچنین لغت‌نامه‌های ترجمه، واژه‌های حسی را برچسب‌گذاری می‌کنند.

مزیت این روش در دسترس‌بودن لغت‌نامه در هر زبانی است و دیگر اینکه از منابع واژگان حسی زبان انگلیسی که برچسب آنها درقبل تعیین شده است به‌راحتی می‌توان در کنار آنها استفاده کرد. از معایب آن این است زمینه واژه (متن اطراف واژه) در نظر گرفته نمی‌شود. همچنین لغت‌نامه‌های عمومی اطلاعاتی در خصوص یک دامنه ندارند و برای تهیه واژگان حسی در یک دامنه مناسب نیستند؛ بنابراین در روش‌های مبتنی بر لغت‌نامه زمینه و دامنه کمتر مورد توجه قرار می‌گیرد. عیب دیگری که روش‌های مبتنی بر لغت‌نامه دارند، این است که از روابط معنایی مترادف و متضاد استفاده می‌کنند، درحالی‌که لغت‌نامه‌ها نسبت به این روابط به‌روز نیستند؛ علاوه‌بر آن عیب دیگری که برای این روش می‌توان برای این روش برشمرد این است که واژه‌های محاوره‌ای و غیررسمی کمتر در لغت‌نامه‌ها حضور دارند. در کارهای انجام‌شده در زبان فارسی، پژوهش‌های [17,18,32,35,40,41,47,48] در این دسته قرار می‌گیرند.

۳. **روش مبتنی بر پیکره:** روش‌های مبتنی بر پیکره، واژه‌های حسی را از پیکره نظرات کاربران استخراج می‌کنند. روش‌های با نظارت و بدون نظارت برای این کار استفاده می‌شود. روش‌های مبتنی بر پیکره دقت کمتری نسبت به روش‌های دستی دارند، ولی از لحاظ هزینه و زمان به صرفه‌تر هستند.

مزیت این روش‌ها این است که منابع واژگان حسی در دامنه خاص به‌صورت کارآمدتر ساخته می‌شوند. به دلیل اینکه پیکره‌ها به‌طورمعمول در یک دامنه هستند، در ساخت منابع واژگان حسی در دامنه خاص با موفقیت عمل می‌کنند. معایب روش‌های مبتنی بر پیکره مشتمل بر دو مورد است. نخست این‌که واژگان به‌دست‌آمده وابستگی زیادی به داده‌های پیکره و واژه‌های رخداده در آن پیکره دارند. دوم‌این‌که آماده‌سازی پیکره‌ای که بتوانند پوشش بالایی نسبت به همه واژگان حسی داشته باشد، مشکل است. نجف‌زاده و همکاران [46] روش مبتنی بر پیکره برای ساخت واژگان حسی ارائه داده‌اند.

بعضی از پژوهش‌ها از روش‌های ترکیبی استفاده می‌کنند تا از مزیت‌های هر دو دسته استفاده کنند. برای نمونه مجموعه دانه که مجموعه محدودی از واژه‌های حسی است به‌صورت دستی برچسب‌گذاری می‌شود و بعد بقیه واژه‌ها مبتنی بر پیکره با روش با نظارت یا بدون نظارت به‌دست می‌آید. پژوهش [53] از امتیازات کاربران در سایت‌های نظرسنجی استفاده می‌کند. به‌عنوان نمونه خریداران در سایت‌های تجاری علاوه‌بر نظردهی در ارتباط با محصول، امتیاز بین یک تا پنج نیز به محصول می‌دهند که به‌عنوان معیار در نظر گرفته می‌شود. چنین روشی در تحلیل احساس در سطح سند بسیار خوب عمل می‌کند، ولی در صورت تحلیل احساس در سطح جمله، خطاهایی نیز صورت می‌پذیرد. بدین دلیل که در بیشتر موارد خریداران در نقد و بررسی‌ها جملات مثبت را در کنار جملات منفی بیان می‌کنند و مزیت‌ها را کنار معایب می‌نویسند؛ بنابراین امتیاز کلی نسبت به یک محصول نمی‌تواند قطبیت تک‌تک جملات باشد.

از منظر دیگر می‌توان روش‌های ایجاد منابع را به دستی، خودکار و نیمه‌خودکار تقسیم کرد. روش دستی، روش‌هایی هستند که برچسب‌گذاری واژه‌ها به‌طورکامل به‌صورت دستی و توسط فرد خبره صورت می‌گیرد. روش‌های خودکار روش‌هایی که بدون دخالت انسان و فقط با کمک الگوریتم‌های یادگیری ماشین به‌دست می‌آید. روش‌های نیمه‌خودکار، روش‌هایی هستند که از الگوریتم بهره می‌گیرند، ولی در بخش‌هایی از کار، از انسان برای افزایش دقت استفاده می‌کنند. در همین راستا برچسب‌گذاری نیمه‌خودکار که ترکیبی از روش‌های دستی و ماشینی است، می‌تواند گزینه بسیار مناسب‌تری باشد.

در جدول (۴) به بررسی و مقایسه پژوهش‌هایی که نوآوری کار خود را روی ایجاد واژگان حسی قرار داده‌اند و در ادامه تلاش کرده‌اند از منبع ایجادشده برای تعیین قطبیت استفاده کنند، پرداخته شده است. هریک از ستون‌های جدول (۴) عبارتند از:

روش کلی: در این ستون روش کلی ایجاد واژگان حسی که در بالا مفصل توضیح داده شده، بیان شده است.

روش جزئی ایجاد منبع واژه حسی: روش ایجاد منبع واژه حسی با جزئیات بیشتر که مختص آن پژوهش بوده در این ستون یاد شده است.

دامنه: منظور دامنه یا حوزه‌ای که واژگان حسی برای آن ایجاد شده یا در بعضی موارد داده برای ارزیابی واژگان

[DOI: 10.52547/jsdp.19.2.107]

حسی در آن دامنه است، می‌باشد. همان‌طور که در بخش۶ توضیح داده شده است، داده‌ها می‌توانند در دامنه‌های مختلف باشند؛ از جمله دامنه محصولات و خدمات، دامنه سیاسی، دامنه پزشکی و دامنه اجتماعی. لازم به توضیح است همان‌طور که در ستون مربوطه در جدول (۴) مشاهده می‌شود اکثریت پژوهش‌های انجام‌شده روی زبان فارسی [17,18,39,40,41,46,48,49,54] واژگان حسی را به‌صورت عمومی ارائه داده‌اند.

مجموعه‌داده: در این ستون مجموعه‌داده یا پیکره‌ای که برای استخراج واژگان حسی استفاده شده، یاد شده است. در برخی موارد نیز به مجموعه داده‌ای که روش ارائه‌شده روی آن ارزیابی شده اشاره شده است.

دقت واژگان حسی تولید شده: در این ستون دقت دست‌یافته‌شده در پژوهش یاد شده، در برخی از پژوهش‌ها معیار صحت (Accuracy) مورد اشاره قرار گرفته و در برخی دیگر معیارهای دقت (Precision)، فراخوانی (Recall). علاوه‌بر آن برخی از روش‌های ارائه‌شده در مقالات در حالات مختلف (به‌عنوان نمونه مجموعه‌داده‌های مختلف) دقت‌های مختلفی را گزارش و روش خود را مورد ارزیابی قرار داده‌اند که در جدول (۲) بالاترین دقت یادشده، گزارش شده است.

مزیت: در این ستون مزیت روش پیشنهادی به‌وسیلهٔ پژوهش توضیح داده شده است.

دقت بعد از اعمال منبع واژه حسی در الگوریتم طبقه‌بندی: بسیاری از پژوهش‌ها بعد از ارائه واژگان حسی، واژگان را در الگوریتم‌های یادگیری با نظارت و یا بدون نظارت استفاده کرده‌اند که دقت به‌دست‌آمده در این ستون آورده شده است.

(جدول-۲): مقایسه روش‌های تحلیل احساس در متون فارسی

(Table-2): Comparision of sentiment analysis methods in Persian texts

مرجع	رویکرد کلی	الگوریتم	ویژگی مورد استفاده در طبقه‌بندی	آیا منبعی ایجاد کرده است	دامنه	مجموعه داده	دقت	مزیت
Bagheri و همکاران [36,37]	با نظارت	بیزین ساده	اطلاعات متقابل اصلاح شده	خیر	محصولات مبایل	۸۲۹ نظر	Recall = 0.8568	ویژگی جدید برای طبقه‌بندی نظرات فارسی ارائه داده است.
Vazirpour و همکاران [38]	با نظارت	SVM	خوشه بند Brown[1]	خیر	سیاسی	۲۷۴۰۳۲ توئیت از ۱۷۸۴۴ کاربر یکتا	Acc=70%	روند تغییرات قطبیت را نسبت به زمان استخراج کرده است.
Amiri و همکاران [39]	مبتنی بر واژه	-	-	بله – به روش دستی	عمومی	منابع آنلاین	Precision= 69%	ایجاد منبع واژگان حسی به صورت دستی
Alimardani و Aghaie [40,41]	ترکیبی (مبتنی بر واژه و با نظارت)	بیزین ساده، SVM، لجستیک رگرسیون	present-absent, TF TF-IDF و هر ویژگی در میزان قطبیت واژه ضرب می‌شود.	بله- مبتنی بر لغت‌نامه	هتل	۱۸۰۵ نظر منفی ۴۶۳۰ نظر مثبت	best Acc= %۸۵,۵	آزمایشات متعدد روی نمونه‌ها با اندازه‌های مختلف انجام داده است. اثر طبقه‌بندهای مختلف با سه ویژگی مورد بررسی قرار داده است.
Sadidpour و همکاران [42]		قاعده گرا استخراج الگو	-	خیر	سیاسی	۱۴۰۰۰ خبر	Acc=90%	تمرکز روی دامنه سیاسی

[1] Brown cluster

[DOI: 10.52547/jsdp.19.2.107]

(جدول-۳): روش‌های ساخت واژگان حسی به همراه مزایا و معایب هر یک از روش‌ها و همچنین مراجعی که در تحلیل متون فارسی از آن روش استفاده کرده‌اند.

(Table-3): lexicon creation methods along with advantages and disadvantages and also references of Persian sentiment analysis studies

روش کلی	مزایا	معایب	پژوهش‌های تحلیل احساس متون فارسی که از روش استفاده کرده‌اند
ساخت دستی	دقت بالایی دارند.	پرهزینه، زمان بر است و حجم داده‌های مشخصی را در یک بازه زمانی می‌توان برچسب‌گذاری کرد.	[39,51,52,54]
مبتنی بر لغت‌نامه	در دسترس بودن لغت‌نامه در هر زبانی	زمینه واژه در نظر گرفته نمی‌شود (متن اطراف واژه) لغت‌نامه‌های عمومی اطلاعاتی در خصوص یک دامنه ندارند و برای تهیه واژگان حسی در یک دامنه مناسب نیستند. روش‌های مبتنی بر لغت‌نامه از روابط معنایی مترادف و متضاد استفاده می‌کنند، در حالیکه لغت‌نامه‌ها نسبت به این روابط به روز نیستند. واژه‌های محاوره‌ای و غیررسمی کمتر در لغت‌نامه‌ها حضور دارند.	[17,18,32,35,40,41,47,48,49]
مبتنی بر پیکره	به دلیل اینکه پیکره‌ها معمولا در یک دامنه هستند، در ساخت منابع واژگان حسی در دامنه خاص با موفقیت عمل می‌کنند.	واژگان به‌دست‌آمده وابستگی زیادی به داده‌های پیکره و واژه‌های رخداده در ان پیکره دارند. آماده‌سازی پیکره‌ای که بتوانند پوشش بالایی داشته باشد مشکل است.	[46,53](از روی امتیازات کاربران از منابع وب)

(جدول-۴): بررسی پژوهش‌هایی که تلاش کرده‌اند روشی برای ایجاد واژگان حسی برای متون فارسی ارائه دهند.

(Table-4): Review on proposed lexicon creation methods in Persian texts

مرجع	روش کلی	روش جزئی ایجاد منبع واژه حسی	دامنه	مجموعه‌داده	دقت واژگان حسی تولید شده	مزیت	دقت بعد از اعمال منبع واژه حسی در الگوریتم طبقه-بندی
shams و همکاران [35]	بدون نظارت	با استفاده منبع زبان انگیسی و ترجمه آن + بهبود روش LDA Persianclue نشانه‌های حسی مبتنی بر موضوع	هتل، مبایل و دوربین	۲۰۰ نظر مثبت و ۲۰۰ نظر منفی	Acc=٪۹۱-۱۰۰	قطبیت را با توجه به موضوع تعیین نموده است. روش بدون نظارتی استفاده نموده است.	طبقه‌بند SVM acc = ٪۷۸
Dehdarbehbahani و همکاران [47]	مبتی بر واژه نامه و نیمه-نظارتی	روش قدم زدن تصادفی از پیوند بین شبکه معنایی کلمات Wordnet و فارسنت استفاده نموده است.	عمومی		MAP = 0.658 Acc =0.914	ترکیب شبکه معنایی انگلیسی و فارسی و کمک گیری از آن برای استخراج قطبیت	-
Amiri و همکاران[39]	دستی	برچسب‌گذاری دستی ۷۱۷۹ صفت، کلمه	عمومی	منابع آنلاین		ایجاد منبع واژگان حسی به‌صورت دستی	Precision= 69%

[DOI: 10.52547/jsdp.19.2.107]

مرجع	روش کلی	روش جزئی ایجاد منبع واژه حسی	دامنه	مجموعه‌داده	دقت واژگان حسی تولید شده	مزیت	دقت بعد از اعمال منبع واژه حسی در الگوریتم طبقه‌بندی
		و عبارت را برچسب گذاری دستی کرده است.					
Alimardani و Aghaie [40,41]	مبتنی بر لغت‌نامه	ایجاد منبع واژگان حسی فارسی با کمک SentiWordNet در زبان انگلیسی	عمومی	-	-	ترکیب وزن واژگان حسی با ویژگی‌های دیگر در الگوریتم طبقه‌بندی	best Acc= %۸۵,۵
لِکسی‌پرس(LexiPers) [49]	نیمه‌نظارتی استفاده از هستان‌شناسی فارس‌نت به عنوان مبنا	مجموعه دانه به روش دستی برچسب‌گذاری شده است و با روش PMI بسط داده شده است.	عمومی		Best Acc= 0.81 Best F-= measure 0.66	در صورت گسترش فارس‌نت از الگوریتم بسط آن برای تعیین قطبیت واژه‌های جدید می‌توان استفاده کرد.	k-nearest neighbors (KNN) nearest centroid (Rocciho)
پیکرهٔ سِنتی‌پرس(SentiPers) [51,52]	با روش کاملا دستی جملات را برچسب زده است	-	کالاهای دیجیتال	-	-	با توجه به دستی بودن این منبع از دقت بالایی برخوردار است. از آنجایی که منتشر شده قابل استفاده برای پژوهش‌های دیگر است.	-
Dashtipour [54]	با روش کاملا دستی جملات را برچسب زده است	-	عمومی		-	ایجاد منبع جدید که اصطلاحات را هم لحاظ کرده و انتشار آن	SVM acc= 69.29% Naïve Bayes acc= 63.19%
نجف‌زاده و همکاران [46]	نیمه‌نظارتی	مدل مخفی مارکوف واژه نامه حسی وفقی را با روش نیمه‌نظارتی ساخته است	عمومی	۱۹۰۰ آموزش ۲۹۹ آزمون	Acc=89%	ساخت منبع واژگان حسی به روش نیمه‌نظارتی و بدون استفاده از خبره انسانی استفاده از روش نیمه‌نظارتی که نیاز به داده برچسب‌دار کمتر دارد.	-
عسگریان و همکارانش [48]		استفاده از شبکه واژگان فردوس‌نت و ترجمه منبع واژگان حسی انگلیسی	عمومی		Precision=86% F1-measure=0.8	استفاده از شبکه واژگان فردوس نت که پوشش بالاتری نسبت به بقیه شبکه واژگان دارد	-
Asgarian و همکاران [17]	نیمه‌نظارتی	مبتنی بر مدل مخفی مارکوف	عمومی		Precision=86% F1-measure=0.8	منبع واژگان حسی ایجاد کرده و در طبقه‌بندهای مختلف و در کنار ویژگی‌های مختلف استفاده کرده است	avg f-measure=88%
[18]Dehkharghani	با نظارت	واژگان مبتنی بر ترجمه منابع واژگان حسی انگلیسی استخراج کرده و طبقه‌بند لجستیک برای	عمومی		Acc=95.92	ایجاد اولین منبع واژگان حسی که به هر واژه امتیاز سه‌تایی (مثبت، منفی و عینی) داده است.	-

[DOI: 10.52547/jsdp.19.2.107]

مرجع	روش کلی	روش جزئی ایجاد منبع واژه حسی	دامنه	مجموعه‌داده	دقت واژگان حسی تولید شده	مزیت	دقت بعد از اعمال منبع واژه حسی در الگوریتم طبقه-بندی
		یادگیری قطبیت واژگان فارسی استفاده کرده					

(جدول-۵): پیکره‌های منتشرشده و در دسترس زبان فارسی

(Table-5): Available and published corpora in Persian language

نام	کلمه/جمله	تعداد	دامنه	ناشر
لکسی پرس LexiPers	کلمه	۴۲۶۱	عمومی	آزمایشگاه پردازش زبان طبیعی دانشگاه صنعتی شریف و گروه پردازش زبان طبیعی دانشگاه گیلان [49]
واژگان فارسی دارای برچسب قطبیت	بخش اول: صفت‌های برچسب‌خورده فارسی بخش دوم: صفت، فعل و اسم	۳۵۸۸ صفت، ۴۰۷۳ فعل و ۷۳۲۵ اسم	عمومی	آزمایشگاه سیستم‌های هوشمند اطلاعات دانشگاه تهران [47]
PerSent	کلمات، عبارت و اصطلاحات	۱۵۰۰ کلمه و ۷۰۰ عبارت و اصطلاح	عمومی	Dashtipour [54]
سنتی پرس SentiPers	شامل جملاتی، هم به‌صورت رسمی و هم به‌صورت نوشتاری عامیانه یا غیررسمی	۱۱۰۰	کالاهای دیجیتال	گروه پردازش زبان طبیعی دانشگاه گیلان [51,52]
پیکره قطبیت هلوکیش	سند	۶۴۲	هتل	مرادی و همکاران [53]

۸- پیکره‌های استاندارد و در دسترس در زبان فارسی

۸-۱- منابع واژگان حسی

در زبان انگلیسی منابع واژگان حسی مانند SentiWordNet[55]، General Inquier [56]، MPQA [57]، SenticNet [25] و Vader [58] برای تحلیل نظرات وجود دارد. در زبان‌های دیگر نیز بهتر است چنین منابعی ایجاد کرد و از مزیت‌های آنها بهره جست. در زبان فارسی منابع معدودی ایجاد شده است، که در ادامه به تشریح آنها می‌پردازیم. در جدول (۵) پیکره‌های منتشر شده و در دسترس زبان فارسی به همراه ویژگی‌های هرکدام آورده شده است. ستون‌های این جدول عبارتند از:

۱- **نام:** نام پیکره یا مجموعه‌داده

۲- **واژه/جمله/سند/واقعیت:** در این ستون مشخص می‌شود که پیکره حاوی کلمه یا جمله و یا واقعیت[1] با برچسب قطبیت است.

در زبان فارسی یک پیکره حاوی جملات برچسب خورده، یک پیکره حاوی اسناد برچسب خورده و سه پیکره دیگر حاوی کلمات برچسب خورده است. واقعیت‌های دارای قطبیت دوتایی‌ها و یا سه‌تایی‌هایی در یک دامنه خاص هستند که مثبت یا منفی هستند. برای نمونه (Regret, buy) دوتایی است که در حوزه نظرات کاربران نسبت به محصولات دارای قطبیت منفی است. در حال حاضر منبعی با چنین ویژگی در زبان فارسی منتشر نشده است.

۳- **تعداد:** تعداد کلمات، تعداد جملات موجود در پیکره و همچنین تعداد مثبت‌ها و منفی‌های آن در این ستون مشخص می‌شود.

۴- **دامنه:** منظور از این ستون دامنه یا دامنه‌ای است که پیکره از روی آن ساخته شده است. در بخش ۶ در ارتباط با دامنه‌های مختلف از جمله سیاسی، اجتماعی، پزشکی و اقتصادی توضیح شده است. گفتنی است در هنگام استفاده از یک پیکره مختص یک دامنه در دامنه دیگر کارایی خوبی از خود نشان نمی‌دهد. برای مثال پیکره تولید شده در دامنه محصولات لپ‌تاپ کارایی

[1] Fact

[DOI: 10.52547/jsdp.19.2.107]

لازم را در تحلیل احساس دامنه پزشکی ندارد. علاوه‌بر آن پیکره واژگان حسی به‌طورمعمول وابسته به دامنه هستند [59,60] و قطبیت واژه‌ها در یک دامنه نسبت به دامنه دیگر گاه متضاد است.

۵- **ناشر:** در این ستون مؤسسه، سازمان یا نویسنده‌ای که پیکره را منتشر کرده یاد شده است.

۸-۱-۱- منبع واژگان حسی لِکسی‌پِرس

منبع واژگان حسی لِکسی‌پِرس(LexiPers)[1] [49]، شامل زیرمجموعه‌ای از واژگان نسخه دوم فارس‌نت است که با روشی خودکار و با سه برچسب مثبت، منفی و خنثی برچسب‌گذاری شده‌اند. در این پروژه کلیه مجموعه‌های ترادف دارای نقش صفت، تعداد ۴۲۶۱ مجموعه، به‌صورت دستی و با هوش انسانی تحت عنوان مجموعه دانه[2] برچسب‌گذاری شده‌اند. این مجموعه دانه می‌تواند به‌عنوان یک استاندارد طلایی[3] و حتی یک مجموعه دانه اولیه برای توسعه و یا آزمودن سامانه‌های برچسب‌گذاری لغات و دسته‌بندی اسناد مورد استفاده قرار گیرد. تولیدکنندگان این لغت‌نامه آزمایشگاه پردازش زبان طبیعی دانشگاه صنعتی شریف و گروه پردازش زبان طبیعی دانشگاه گیلان هستند.

۸-۱-۲- واژگان فارسی دارای برچسب قطبیت

پیکره واژگان فارسی دارای برچسب قطبیت[4] [47] که در آزمایشگاه سامانه‌های هوشمند اطلاعات دانشگاه تهران تولیده شده است. از دو مجموعه‌داده تشکیل شده است:

۱- مجموعهٔ استخراج‌شده از صفت‌های برچسب‌خورده فارسی: که از روی مجموعهٔ صفت‌های زبان فارسی استخراج‌شده از فارس‌نت ساخته شده است. هر ورودی در این مجموعه می‌تواند برچسب مثبت، منفی و یا خنثی داشته باشد. برای این کار بیش از ۳۵۸۸ صفت استخراج و توسط چهار ارزیاب مستقل ارزیابی شده است.

۲- مجموعه صفت، فعل و اسم که این مجموعه از روی مجموعه صفت‌ها، فعل‌ها و اسم‌های موجود در فارس‌نت استخراج شده‌اند. به هر کلمه از این مجموعه به‌وسیلهٔ یک روش مبتنی بر یادگیری ماشین نیمه‌نظارتی یک مقدار عددی نسبت داده می‌شود. این عدد درواقع تعیین‌کننده میزان قطبیت هر کلمه است. به کلمات مثبت، عددی بزرگ‌تر از صفر و به کلمات منفی، عددی کوچکتر از صفر نسبت داده می‌شود. در این مجموعه، کلمات خنثی به صراحت تعیین نمی‌شوند و می‌توان کلمات خنثی را براساس یک حد آستانه بین کلمات مثبت و منفی تعیین کرد. این مجموعه شامل ۳۵۸۸ صفت، ۴۰۷۳ فعل و ۷۳۲۵ اسم است. گفتنی است که کلیه کلمات از روی نسخه ۱ فارس‌نت استخراج شده‌اند.

۸-۱-۳- پیکرهٔ PerSent

پیکرهٔ PerSent[5] [54,61] شامل مجموعه‌ای از کلمات و عبارات و اصطلاحات زبان فارسی با برچسب‌های قطبیت بین -۱ و +۱ است که در زمینهٔ تحلیل احساس کاربرد دارد[6]. در نسخه اول خود ۱۵۰۰ کلمه فارسی و در نسخه دوم ۷۰۰ اصطلاح و عبارت ارائه داده است. اصطلاحات فهرست‌شده در این پیکره برای تحلیل متون غیررسمی که توسط کاربران تولید شده، مناسب است. حوزه جملات این پیکره به طور عمومی و برای دامنه‌هایی مانند فیلم و محصولات قابل استفاده است. این پیکره توسط یکی از پژوهش‌گران فارسی‌زبان در خارج از ایران تهیه شده است.

۸-۲- پیکره نظرات کاربران

پیکره نظرات کاربران منابع دیگری هستند که برای روش‌های تحلیل احساس آماده‌سازی شده‌اند. این پیکره‌ها حاوی جملاتی هستند که برچسب قطبیت مثبت یا منفی دارند. در زبان فارسی دو پیکره سِنتی‌پِرس و پیکره قطبیت هلوکیش منتشر شده است که در ادامه به آنها پرداخته می‌شود.

۸-۲-۱- پیکرهٔ سِنتی‌پِرس

پیکرهٔ سِنتی‌پِرس (SentiPers) [51,52] شامل مجموعه‌ای از جملات فارسی با برچسب‌های حاوی قطبیت است که در پردازش زبان طبیعی و به‌طور مشخص در زمینهٔ تحلیل احساس یا نظرکاوی کاربرد دارد. با توجه به ویژگی‌های این پیکره، می‌توان آن را در نوع خود نخستین پیکرهٔ تحلیل احساس برای زبان فارسی با چنین مشخصاتی به‌شمار آورد. گفتنی است که دامنهٔ جملات موجود در پیکره مربوط به حوزهٔ کالاهای دیجیتال است. همچنین این پیکره شامل جملاتی، هم به‌صورت رسمی و هم به‌صورت نوشتاری عامیانه یا غیررسمی و مشتمل بر حدود ۱۱۰۰ جمله برچسب‌خورده است. این پیکره در گروه پردازش زبان طبیعی دانشگاه گیلان تولید شده است.

[1] http://www.peykaregan.ir
[2] Seed set
[3] Gold standard
[4] http://www.peykaregan.ir

[5] Persian Sentiment Lexicon(PerSent)
[6] https://www.gelbukh.com/resources/persent/

[DOI: 10.52547/jsdp.19.2.107]

۸-۲-۲- پیکره قطبیت هلوکیش

پیکرهٔ قطبیت هلوکیش [53] ازپیکره‌های نشانه‌گذاری‌شده از نظر احساسات نویسنده است. پیکره حاضر با جمع‌آوری نظرات کاربران در وبگاه هلوکیش به‌دست آمده است. کل نظرات ثبت‌شده در این وبگاه در هنگام انجام پژوهش تعداد ۳۳۱۲ نظر بوده است که از این تعداد، در ۶۴۲ مورد کاربر از طریق گزینه موجود در بخش ثبت نظر، میزان رضایت خود را نیز مشخص کرده بوده است. نظرات با توجه به درصد رضایت واردشده به دو دسته تقسیم شدند: نظراتی که درصد رضایت آنها کمتر یا مساوی ٪۳۰ بود و نظراتی که درصد رضایتشان بالاتر یا مساوی ٪۷۰ بود. با این تقسیم‌بندی ۱۰۲ نظر در دسته نظرات منفی و ۴۴۷ نظر در دسته نظرات مثبت قرار گرفتند. میانگین تعداد کلمات در هر نظر ۱۰۹ کلمه است.

۸-۳- نتایج بررسی پیکره‌های منتشرشده

همان‌طور که در جدول (۵) مشاهده می‌کنید، از پنج پیکره منتشرشده، سه تای آنها مربوط به واژگان و یکی از آنها جمله و یکی دیگر سند را برچسب‌گذاری کرده است. بنابراین نیازمند تعداد بیشتری از منابع به‌خصوص در سطح جمله و سند هستیم. علاوه‌براین منابع اکثراً به‌طور عمومی تعریف شده‌اند که در بسیاری از کاربردها نیاز به منابع با قطبیت وابسته به دامنه داریم. منابع واژگان حسی برای دامنه سیاسی، اجتماعی و دامنه پزشکی و دارو ضروری به نظر می‌رسد.

در رویکردهای مبتنی بر واژه هرچه منبع پوشش بهتری داشته و جامع‌تر باشد، به‌دقت بیشتری دست پیدا خواهیم کرد. بنابراین تعداد واژگان حسی اهمیت پیدا می‌کند و منبع با جامعیت بیشتر کارایی بالاتری دارد. در کاربردهای یادگیری ماشین هرچه داده برچسب‌گذاری‌شده بیشتر برای آموزش مدل در نظرگرفته شود، مدل به‌دست‌آمده کارایی بالاتری دارد. همان‌طور که در جدول (۵) مشاهده می‌کنید، تعداد داده‌ها در هریک از پیکره‌ها به‌خصوص پیکره سِنتی‌پِرس و پیکرهٔ قطبیت هلوکیش (حدود ۱۱۰۰ و ۶۴۲) کم هستند و نیاز به پیکره‌هایی با حجم بیشتری از داده هستیم. در کاربردهای یادگیری شبکه‌های عصبی عمیق نیز وجود حجم زیادی از داده امری ضروری است.

۹- نیازمندی‌های آینده تحلیل احساس متون فارسی

با توجه به مروری که روی کارهای انجام‌شده در تحلیل احساس متون فارسی انجام شد، و همچنین دستاوردها و آخرین رویکردهای علمی به‌دست‌آمده در زبان انگلیسی [6-10]، در این بخش به مسائل و خلاءهای تحلیل احساس در زبان فارسی پرداخته می‌شود و موضوعاتی که باید در حوزه تحلیل نظرات در زبان فارسی مورد توجه قرار بگیرد، بیان می‌شود. نکاتی که در ادامه بیان می‌شود می‌تواند نقطه شروعی برای پژوهش‌های آینده تحلیل احساس متون فارسی باشد:

۱. چالش‌های زیادی در زبان انگلیسی برای تحلیل نظرات باقی‌مانده است [62,63] در حل بعضی از این چالش‌ها در زبان انگلیسی اقدامات خوبی انجام گرفته است، سزاوار است زبان فارسی نیز همپای زبان انگلیسی به چنین دستاوردهایی برسد. تعدادی از چالش‌ها در زبان فارسی به‌ندرت به آنها پرداخته شده است؛ مانند استخراج ویژگی، تحلیل احساس جملات کنایه‌آمیز، تشخیص نظرات جعلی با توجه به متن نظر، تعیین قطبیت نظرات ضمنی و غیره. چنین موضوعاتی می‌تواند در آینده مورد پژوهش قرار بگیرد.

۲. همان‌طور که در بخش ۵ مورد اشاره قرار گرفت معکوس‌کننده‌ها باعث تغییر جهت قطبیت جمله می‌شوند. مهم‌ترین معکوس‌کننده‌ها کلمات و نشانه‌های نفی مانند پیشوند نفی در افعال جمله هستند ولی نمی‌توان فقط به حضور آنها اکتفا کرد و مدیریت آن پیچیدگی‌های خاص خود را دارد. مدیریت معکوس‌کننده‌ها یکی از مسائل چالش برانگیز تحلیل احساس است و در پژوهش‌ها روی متون انگلیسی مورد بررسی قرار گرفته است. در پژوهش‌های زبان فارسی به مسأله مدیریت معکوس‌کننده‌ها پرداخته نشده است و نیازمند است مورد توجه پژوهش‌گران قرار بگیرد.

۳. هرچه ابزارهای پیش پردازش قوی‌تری در زبان فارسی داشته باشیم، دقت روش‌های تحلیل احساس افزایش می‌یابد. یک واژه‌یاب با دقت خوب کمک بسیار جدی به پیش‌پردازش تحلیل احساس، به‌خصوص در روش‌های یادگیری شبکه‌های عصبی عمیق می‌کند؛ بنابراین روش‌ها و پیاده‌سازی‌هایی که در زبان فارسی کمک به افزایش دقت واژه‌یاب می‌کند، باید مورد پژوهش قرار بگیرد.

۴. منابع واژگان حسی مناسب برای تحلیل احساس و همچنین پیکره‌های استاندارد برچسب خورده در زبان فارسی دارای محدودیت هستند. هنوز در دامنه‌های مختلف منابع مناسب برای تحلیل احساس وجود ندارد. همان‌طور که در جدول (۵) مشاهده کرده‌اید، منابع ایجادشده بیشتر برای تحلیل احساس در دامنه عمومی

[DOI: 10.52547/jsdp.19.2.107]

و تجاری و کمتر در دامنه سیاسی، اجتماعی و اقتصادی است. ضمن این‌که منابع ایجادشده حجم پایینی دارند و کارایی کافی در بسیاری از کاربردها مانند یادگیری شبکه‌های عصبی عمیق را ندارند؛ بنابراین نیازمند منابع حسی و پیکره‌های مناسب در دامنه‌های یادشده در زبان فارسی هستیم. همچنین در زبان فارسی نیازمند روش‌هایی که کارایی لازم را برای انجام تحلیل احساس مبتنی بر دامنه[1] دارند، هستیم.

۵. ابزارها، برنامه‌های کاربردی و بسته‌های متنوعی برای تحلیل احساس در زبان انگلیسی در اینترنت توسط توسعه‌دهندگان منتشر شده است. این ابزارها بر پایه روش‌های متنوعی که وجود دارد، ساخته شده است. در زبان فارسی با کمبود چنین ابزارهایی مواجه هستیم. در دسترس‌بودن ابزارهای پیاده‌سازی‌شده در زبان فارسی کمک بسیار زیادی در جهت رشد پژوهش‌های تحلیل احساس در زبان فارسی می‌کند. پیشنهاد می‌شود که در پژوهش‌های آینده روی ایجاد چنین ابزارهایی با روش‌های علمی در زبان فارسی تمرکز کرد. ضمن اینکه تعدادی از آزمایشگاه‌های پژوهشی ابزارهایی برای پیشبرد کارهای خود ایجاد کرده‌اند ولی بنا به دلایل متعدد تمایلی به انتشار در شبکه اینترنت و در دسترس قراردادن آنها برای عموم ندارند، این امر باعث کارهای موازی و عدم پیشرفت در حوزه تحلیل احساس فارسی شده است.

۶. رویکرد تحلیل احساس مبتنی بر مفهوم [19,64-66] که مفاهیم متن و معانی ضمنی آن را در نظر می‌گیرد. بیشتر روی تجزیه و تحلیل معنایی متن از طریق استفاده از هستان‌شناسی و شبکه‌های معنایی تمرکز دارند، که اجازه تجمیع اطلاعات مفهومی و عاطفی در ارتباط با نظرات زبان طبیعی را می‌دهند. تاکنون پژوهش‌های انجام‌شده در زبان فارسی از این رویکرد استفاده نکرده‌اند؛ بنابراین استفاده از این رویکرد در تحلیل متون فارسی پیشنهاد می‌شود. ضمن این‌که در اختیار داشتن یک تجزیه‌گر مفهومی با دقت مناسب در زبان فارسی از نیازمندی‌های رویکرد مبتنی بر مفهوم است و می‌بایست در پژوهش‌های آینده به روش‌هایی متناسب با زبان فارسی پرداخته شود.

۷. نظرات به دو دسته صریح و ضمنی تقسیم شده است. نظرات صریح، نظراتی هستند که واژگان حسی به‌صراحت در آنها استفاده شده است و نظرات ضمنی، نظراتی هستند که به‌طور صریح در آنها واژه حسی استفاده نشده است، ولی حاوی نظر هستند (به‌عنوان مثال "این گوشی آنتن نمی‌دهد" حاوی واژه حسی نیست ولی واقعیتی را بیان می‌کند که منفی است) [2]. پژوهش‌ها در بخش نظرات صریح روی متون انگلیسی قابل توجه و به دقت‌های به‌نسبه خوبی رسیده‌اند ولی روش‌های تحلیل نظرات ضمنی معدود و انگشت‌شمار هستند. دسته دوم روش‌های خاص منظوره ارائه داده‌اند [67] که هنوز به جامعیت نرسیده‌اند و در ابتدای راه هستند. با توجه به اینکه بخش زیادی از نظرات کاربران در سایت‌ها نظرات ضمنی هستند، ولی از پیکره‌های استاندارد حذف می‌شوند، بهتر است در زبان فارسی هم مورد توجه قرار بگیرند. پیشنهاد می‌شود در پژوهش‌های آینده زبان فارسی تحلیل نظرات ضمنی مورد پژوهش و بررسی قرار بگیرد و همچنین پیکره‌های مورد نیاز تولید شود.

۸. از آنجایی که روش‌های بدون نظارت نیاز به پیکره برچسب خورده ندارند و در زبان فارسی هم محدودیت پیکره استاندارد داریم، نیاز هست که روش‌های بدون نظارت و همچنین نیمه‌نظارتی که نیاز به داده برچسب خورده کمتری دارند، بیشتر مورد توجه قرار گیرد.

۹. قطبیت واژگان حسی به زمینه‌ای که واژه در آن رخ می‌دهد وابسته است و قطبیت آنها ثابت نیست [68,69]، بنابراین تعداد زیادی از پژوهش‌ها در زبان انگلیسی روی این موضوع کار کرده‌اند. یکی از نقاط ضعف پژوهش‌های زبان فارسی ضعیف دیده‌شدن این مسأله است. در پژوهش‌های تحلیل احساس در زبان فارسی نیز ارزشمند است که روی تحلیل احساس وابسته به زمینه[2] تمرکز کرد.

۱۰. ایجاد پایگاه‌های دانش دارای برچسب قطبیت کمک بسیار زیادی در تحلیل احساس می‌کند. برچسب‌گذاری واقعیت‌های دارای قطبیت در تحلیل نظرات در دامنه خاص مانند دارو و ایجاد چنین منابع دانشی، دقت روش‌ها را به‌طور چشم‌گیری افزایش می‌دهد، Noferesti و همکاران در [45,70] اثر حضور چنین واقعیت‌هایی را در زبان انگلیسی مورد بررسی قرار داده‌اند. ایجاد پایگاه دانش از واقعیت‌های دارای قطبیت در زبان فارسی به همراه برچسب قطبیت آنها، پشتوانه خوبی برای تحلیل احساس متون فارسی است و می‌توان در پژوهش‌های آینده به آن پرداخت.

[1] Domain specific sentiment analysis

[2] Context-based sentiment analysis

[DOI: 10.52547/jsdp.19.2.107]

۱۱.روش‌های یادگیری شبکه‌های عصبی عمیق به دستاوردهای خوبی در زبان انگلیسی رسیده‌اند [71,72]، در پژوهش‌های آیندهٔ تحلیل متون فارسی می‌توان روی معماری‌های مناسب شبکه‌های عصبی عمیق کار کرد. Dashtipour و همکارانش [73] در مقاله‌ای روشی بدون ناظر و با معماری مبتنی بر کدگذار خودکار[1] در بستر شبکه‌های کانولوشن برای تحلیل احساس ارائه کرده‌اند. با توجه به نیازمندی‌های زبان فارسی می‌توان به‌کارگیری روش‌های یادگیری شبکه‌های عصبی عمیق در حوزه تحلیل احساس زبان فارسی را گسترش داد و از نتایج آن بهره گرفت.

۱۲.در حال حاضر تحلیل احساس روی داده‌های متنی بیشترین سهم پژوهش‌های تحلیل احساس را گرفته است. تحلیل احساس روی داده‌های تصویری، صوتی و فیلم با نام تحلیل احساس مالتی مدال[2] درهمین‌اواخر مورد توجه جامعه علمی قرار گرفته است. تحلیل احساس داده‌های صوت فارسی متفاوت از زبان انگلیسی است و نیازمند پژوهش‌های بیشتر توسط پژوهش‌گران است.

۱۰- نتیجه‌گیری

در این مقاله مروری روی کارهای انجام‌شده در تحلیل احساس و نظرکاوی متون فارسی انجام گرفت. در همین راستا، روش‌های به‌کارگرفته در مقالاتی که در منابع معتبر بین‌المللی و داخلی چاپ شده مورد مطالعه قرارگرفت؛ درنهایت نقشه راهی برای پژوهش‌های آینده تحلیل احساس در زبان فارسی ترسیم شد.

روش‌های موجود در زبان انگلیسی به‌طور کلی قابل تعمیم به زبان فارسی هستند؛ ولی باید با توجه به تفاوت‌هایی که در زبان‌ها وجود دارد، ویژگی‌های زبان فارسی و همچنین امکاناتی که در پیش‌پردازش در این زبان وجود دارد، به‌طور موشکافانه‌تری و به‌طور عملی مورد پژوهش قرار بگیرند. همچنین از آنجاکه منابع با کیفیت از قبیل واژه‌نامه قطبیت، پایگاه دانش دارای قطبیت و پیکره‌های دارای قطبیت در دامنه‌های مختلف، نقش کلیدی در تحلیل احساس در زبان فارسی دارند، باید در آینده متناسب با نیازمندی‌ها آنها را گردآوری کرد.

[1] Autoencoder
[2] Multimodal sentiment analysis

11- References ۱۱- مراجع

[1] E. Cambria, S. Poria, A. Hussain, and B. Liu, "Computational Intelligence for Affective Computing and Sentiment Analysis [Guest Editorial]", *IEEE Computational Intelligence Magazine*, vol. 14(2), pp. 16-17, 2019.

[2] B. Liu, "Sentiment analysis and opinion mining", *Synthesis lectures on human language technologies*, vol. 5(1), pp. 1-167, 2012.

[3] B. Liu, "Sentiment Analysis and Subjectivity", *Handbook of natural language processing*, vol. 2, pp. 627-666, 2010.

[4] A. Keramatfar, and H. Amirkhani, "Bibliometrics of sentiment analysis literature", *Journal of Information Science*, vol. 45(1), pp. 3-15, 2019.

[5] R. Piryani, D. Madhavi, and V.K. Singh, "Analytical mapping of opinion mining and sentiment analysis research during 2000–2015", *Information Processing & Management*, vol.53(1), pp. 122-150, 2017.

[6] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol.5(4), pp. 1093-1113, 2014.

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

[8] A. Montoyo, P. MartíNez-Barco, and A. Balahur, "Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments", *Decision Support Systems*, vol. 53(4), pp. 675-679, 2012.

[9] A.N. Jebaseeli, and E. Kirubakaran, "A survey on sentiment analysis of (product) reviews", *International Journal of Computer Applications*, vol. 47(11), 2012.

[10] D. Hussein, "A survey on sentiment analysis challenges", *Journal of King Saud University-Engineering Sciences*, vol. 30(4), pp. 330-338, 2018.

[11] N. Boudad, R. Faizi, R.O.H. Thami, and R. Chiheb, "Sentiment analysis in Arabic: A review of the literature", Ain Shams *Engineering Journal*, 2017.

[12] S.L. Lo, E. Cambria, R. Chiong, and D. Cornforth, "Multilingual sentiment analysis: from formal to informal and scarce resource languages", *Artificial Intelligence Review*,. Vol.48(4), pp. 499-527, 2017.

[13] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A.Y. Hawalah, A. Gelbukh, and Q. Zhou, "Multilingual sentiment analysis: state of the art and independent comparison of techniques", *Cognitive computation*, vol.8(4), pp. 757-771, 2016.

[DOI: 10.52547/jsdp.19.2.107]

[14] E.F. Can, A. Ezen-Can, and F. Can, "Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data", arXiv preprint arXiv:1806.04511, 2018.

[15] A. Balahur, J.M. Hermida, and A. Montoyo, "Detecting implicit expressions of sentiment in text based on commonsense knowledge", *in Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*, 2011, Association for Computational Linguistics.

[16] M.E. Basiri, A.R. Naghsh-Nilchi, and N. Ghassem-Aghaee, "A framework for sentiment analysis in persian" *Open Transactions on Information Processing*, Vol.1(3), pp. 1-14, 2014.

[17] E. Asgarian, M. Kahani, and S. Sharifi, "The impact of sentiment features on the sentiment polarity classification in Persian reviews", *Cognitive Computation*, vol. 10(1), pp. 117-135, 2018.

[18] R. Dehkharghani, "SentiFars: A Persian Polarity Lexicon for Sentiment Analysis", *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 19(2), pp. 21, 2019.

[19] E. Cambria, "An introduction to concept-level sentiment analysis," *in Mexican International Conference on Artificial Intelligence*, 2013. Springer.

[20] N. Cristianini, and J. Shawe-Taylor, "An introduction to support vector machines and other kernel-based learning methods", 2000: Cambridge university press.

[21] T. Joachims, "Making large-scale svm learning practical", I n A dvan ces in Ker n e l Meth od s. 1998, MIT Press.

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

[23] P.D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews", *in Proceedings of* the 40th annual meeting on association for computational linguistics, 2002, Association for Computational Linguistics.

[24] S. Dasgupta, and V. Ng, "Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification", in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009.

[25] S. Poria, E. Cambria, G. Winterstein, and G.-B. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis", *Knowledge-Based Systems*, Vol. 69, pp. 45-63, 2014.

[26] E. Cambria, and A. Hussain, "Sentic computing: a common-sense-based framework for concept-level sentiment analysis", Springer, Vol. 1. 2015.

[27] E. Cambria, S. Poria, F. Bisio, R. Bajpai, and I. Chaturvedi, "The CLSA model: a novel framework for concept-level sentiment analysis", *in International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, 2015.

[28] Z. Rajabi, M.R. Valavi, and M. Hourali, "A Context-Based Disambiguation Model for Sentiment Concepts Using a Bag-of-Concepts Approach", *Cognitive Computation*, pp. 1-14, 2020.

[29] K. Dashtipour, M. Gogate, J. Li, F. Jiang, B. Kong, and A. Hussain, "A hybrid Persian sentiment analysis framework: Integrating dependency grammar-based rules and deep neural networks", *Neurocomputing*, 2020. 380: pp. 1-10.

[۳۰] گلپر رابوکی، ع.، ج. رضایی نور، و س.ا. ضرغامی فر، استخراج ویژگیها در اندیشه کاوی مورد استفاده در متون فارسی, در دومین همایش ملی کامپیوتر. ۱۳۹۲.

[30] E. Golpar-Rabooki, S. Zarghamifar, and J. Rezaeenour, Feature extraction in opinion mining for Persian text, in 2nd National Conference on Computer Science, 2013.

[۳۱] گلپر رابوکی، ع.، ج. رضایی نور، و س.ا. ضرغامی فر، استخراج ویژگیها و بسط لغتنامه در اندیشه کاوی مورد استفاده در متون فارسی، در دومین همایش ملی پژوهش های کاربردی در علوم کامپیوتر و فناوری اطلاعات. ۱۳۹۳.

[31] E. Golpar-Rabooki, S. Zarghamifar, and J. Rezaeenour, "Feature extraction in opinion mining through Persian reviews", *Journal of AI and Data Mining*, vol. 3(2), pp. 169-179, 2015.

[32] E. Golpar-Rabooki, S. Zarghamifar, and J. Rezaeenour, "Feature extraction in opinion mining through Persian reviews", *Journal of AI and Data Mining*, vol. 3(2), pp. 169-179, 2015.

[۳۳] میر اجاق، ف.ا. ، م.ح. ندیمی شهرکی، ارائه یک رویکرد نظارتی برای تشخیص هرز نظرات جعلی, در کنفرانس بین‌المللی مهندسی برق و علوم کامپیوتر. ۱۳۹۴.

[DOI: 10.52547/jsdp.19.2.107]

[33] F. Mirojagh, M. Shahraki, “Fake opinion detection using a supervised approach”, International Conference on Electrical Engineering and Computer Science, 2014.

[34] S. Zirpe, and B. Joglekar, “Polarity shift detection approaches in sentiment analysis: A survey”, in 2017 International Conference on Inventive Systems and Control (ICISC), IEEE, 2017.

[35] M. Shams, A. Shakery, and H. Faili, “A non-parametric LDA-based induction method for sentiment analysis,” in The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), IEEE, 2012.

[36] A. Bagheri, and M. Saraee, “Persian sentiment analyzer: A framework based on a novel feature selection method”, *International Journal of Artificial Intelligence*, vol. 12(2), pp. 115-129, 2014.

[37] A. Bagheri, M. Saraee, and F. de Jong. “Sentiment classification in Persian: Introducing a mutual information-based method for feature selection”, in 2013 21st Iranian Conference on Electrical Engineering (ICEE), IEEE, 2013.

[38] E. Vaziripour, C. Giraud-Carrier, and D. Zappala, “Analyzing the political sentiment of Tweets in Farsi”, in Tenth International AAAI Conference on Web and Social Media, 2016.

[39] F. Amiri, S. Scerri, and M. Khodashahi. “Lexicon-based sentiment analysis for Persian Text”, in Proceedings of the International Conference Recent Advances in Natural Language Processing, 2015.

[40] S. Alimardani, and A. Aghaie, Opinion mining in Persian language using supervised algorithms, Journal of Information Systems and Telecommunication (JIST), 2015.

[۴۱] علی مردانی, س. and ع. آقایی, ارائۀ روش نظارتی برای نظرکاوی در زبان فارسی با استفاده از لغت نامه و الگوریتم .SVMمدیریت فناوری اطلاعات دانشکده مدیریت دانشگاه تهران, ۱۳۹۴. ۷(۲): p. 345-362.

[41] S. Alimardani, and A. AGHAIE, “Opinion mining in Persian language using supervised algorithms and sentiment lexicon”, Journal of *Information Technology Management*, vol.7, pp. 345-362, 2015.

[42] S. Sadidpour, , H. Shirazi, N.M. Sharef, B. Minaei-Bidgoli, and M.E. Sanjaghi, “Context-Sensitive Opinion Mining using Polarity Patterns”, *International Journal of Advanced Computer Science and Applications (IJACSA)*, pp.7, 2016.

[43] M. Saraee, and A. Bagheri, “Feature selection methods in Persian sentiment analysis”, in International Conference on Application of Natural Language to Information Systems. Springer, 2013.

[44] Li, Y, H. Guo, Q. Zhang, M. Gu, and J. Yang, “Imbalanced text sentiment classification using universal and domain-specific knowledge,” *Knowledge-Based Systems*, vol.160, pp. 1-15, 2018.

[45] S. Noferesti, and M. Shamsfard, “Using Linked Data for polarity classification of patients’ experiences”, *Journal of biomedical informatics*,vol. 57, pp. 6-19, 2015.

[۴۶] نجف زاده ، م.، س. راحتی قوچانی ، و س. قائمی یک چارچوب نیمه نظارتی مبتنی بر لغت نامه وفقی خودساخت جهت تحلیل نظرات فارسی. پردازش علائم و داده ها، ۱۳۹۷. ۳۶(۲): p. 89-101.

[46] M.Najafzadeh, S. Rahati Quchan, and R. Ghaemi, “A Semi-supervised Framework Based on Self-constructed Adaptive Lexicon for Persian Sentiment Analysis”, *Signal and Data Processing*,vol. 15(2), pp. 89-102, 2018.

[47] I. Dehdarbehbahani, A. Shakery, and H. Faili, “Semi-supervised word polarity identification in resource-lean languages”, *Neural Networks*, vol. 58, pp. 50-59, 2014.

[۴۸] عسگریان، ا. م. کاهانی، و ش. شریفی، حس نگار: شبکه واژگان حسی فارسی. فصلنامه پردازش علائم و داده ها، ۱۳۹۷. ۳۵(۱): p. 71-86.

[48] E.Asgarian, , M. Kahani, and S. Sharifi, “HesNegar: Persian Sentiment WordNet”, *Signal and Data Processing*,vol, vol.15(1), pp. 71-86, 2018.

[49] B. Sabeti, , P. Hosseini, G. Ghassem-Sani, and S.A. Mirroshandel, LexiPers: An ontology based sentiment lexicon for Persian. in GCAI. 2016.

[50] S. Deng, A.P. Sinha, and H. Zhao, “Adapting sentiment lexicons to domain-specific social media texts”, *Decision Support Systems*, vol. 94, pp. 65-76, 2017.

[۵۱] حسینی، پ.، ح. ملکی گلندوز، ع. احمدیان رمکی، م. انواری رستمکلایی، و س. میرروشندل, پیکره فارسی تحلیل احساس سنتی پرس: توسعه یک پیکره ی تحلیل احساس متنی برای زبان فارسی، در سومین همایش ملی زبان شناسی رایانشی. ۱۳۹۲.

[51] P. Hosseini, H. Maleki, A.A. Ramaki, , M. Anvari, and S.A. Mirroshandel, “A sentiment

[DOI: 10.52547/jsdp.19.2.107]

analysis corpus for Persian(SentiPers)," in 3rd National Conference on Computational Linguistics, 2013.

[52] P. Hosseini, A.A. Ramaki, H. Maleki, M. Anvari, and S.A. Mirroshandel, SentiPers: a sentiment analysis corpus for Persian. arXiv preprint arXiv:1801.07737, 2018.

[۵۳] مرادی ، م.، پ. خسروی‌زاده، و و. وزیرنژاد ساخت پیکره‌های نشانه‌گذاری‌شده با رویکرد وب به عنوان پیکره، در دومین هم‌اندیشی زبان‌شناسی رایانشی. ۱۳۹۱: تهران.

[53] M. Moradi, P. Khosravizade, and V. Bahram, "Constructing tagged corpora with a web approach as a corpus", in 2th symposium on computational Linguistics, 2012.

[54] K. Dashtipour, A. Hussain, Q. Zhou, A. Gelbukh, A.Y. Hawalah, and E. Cambria. "PerSent: a freely available Persian sentiment lexicon", in International Conference on Brain Inspired Cognitive Systems, 2016, Springer.

[55] A. Esuli, and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in Proceedings of LREC. 2006. Citeseer.

[56] P.J. Stone, D.C. Dunphy, and M.S. Smith, The general inquirer: A computer approach to content analysis, 1966.

[57] T.Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis", in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005.

[58] C.H.E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text. in Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf. 2014.

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

[60] S.Tan, and Q. Wu, "A random walk algorithm for automatic construction of domain-oriented sentiment lexicon," *Expert Systems with Applications*, vol. 38(10), pp. 12094-12100, 2011.

[61] K. Dashtipour, A. Raza, A. Gelbukh, R. Zhang, E. Cambria, and A. Hussain, "PerSent 2.0: Persian Sentiment Lexicon Enriched with Domain-Specific Words", in International Conference on Brain Inspired Cognitive Systems, Springer, 2019.

[62] A. Balahur, and G. Jacquet, Sentiment analysis meets social media–Challenges and solutions of the field in view of the current information sharing context. Information Processing & Management, 2015. 51(4): p. 428-432.

[63] D. Osimo, and F. Mureddu, "Research challenge on opinion mining and sentiment analysis", Universite de Paris-Sud, Laboratoire LIMSI-CNRS, Bâtiment, 2012. 508.

[64] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis", *IEEE Intelligent Systems*, vol. 28(2), pp. 15-21, 2013.

[۶۵] رجبی، ز.، م. ولوی، و م. حورعلی، ارائه مدلی برای غنی سازی واژگان سنجمانی بر مبنای پایگاه دانش معنایی، در کنفرانس فرماندهی و کنترل. ۱۳۹۶.

[65] Z. Rajabi, M. Valavi, and M. Hourali, "A model for enriching sentiment lexicon based on semantic knowledge base", 10st Iranian C4I Conference, 2017.

[۶۶] رجبی، ز.، م. ولوی، و م. حورعلی، ارائه یک مدل مبتنی بر زمینه برای رفع ابهام از مفاهیم حسی با کمک دانش عرفی. فصلنامه علمی-پژوهشی فرماندهی و کنترل، ۱۳۹۷. ۲(۲): p. 32-47.

[66] Z. Rajabi, M. Valavi, and M. Hourali, "A context-based model for disambiguating the sentiment concepts using the common-sense knowledge", *C4I Journal*, vol.2(2), pp. 32-47, 2018.

[67] J. Liao, S. Wang, and D. Li, "Identification of fact-implied implicit sentiment based on multi-level semantic fused representation" *Knowledge-Based Systems*, vol.165, pp. 197-207, 2019.

[68] C. Hung, "Word of mouth quality classification based on contextual sentiment lexicons", *Information Processing & Management*, vol. 53(4), pp. 751-763, 2017.

[69] H. Saif, Y. He, M. Fernandez, and H. Alani, "Contextual semantics for sentiment analysis of Twitter", *Information Processing & Management*, vol. 52(1), pp. 5-19, 2016.

[70] S. Noferesti, and M. Shamsfard, "Resource construction and evaluation for indirect opinion mining of drug reviews", *PloS one*, vol. 10(5), pp. e0124993, 2015.

[71] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews", *Data*

[DOI: 10.52547/jsdp.19.2.107]

Mining and Knowledge Discovery, 2018. Vol.8(4), p p,1253.

[72] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing", *ieee Computational intelligenCe magazine*, vol.13(3), pp. 55-75, 2018.

[73] K. Dashtipour, M. Gogate, A. Adeel, C. Ieracitano, H. Larijani, and A. Hussain, "Exploiting deep learning for persian sentiment analysis", in International Conference on Brain Inspired Cognitive Systems, Springer, 2018.

زینب رجبی مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار دانشگاه صنعتی اصفهان و مدرک کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات و مدرک دکترای خود را در رشته مهندسی کامپیوتر دانشگاه صنعتی مالک اشتر دریافت کرد. زمینه پژوهشی وی تحلیل داده‌های شبکه اجتماعی، تحلیل احساس، هستان‌شناسی، پردازش زبان طبیعی، محاسبات شناختی و غیره است. ایشان در حال حاضر پژوهش‌گر پژوهشگاه ارتباطات و فناوری اطلاعات، پژوهشکده هوش مصنوعی می‌باشد.

نشانی رایانامه ایشان عبارت است از:

z.rajabi@itrc.ac.ir

محمدرضا ولوی دوره کارشناسی و کارشناسی ارشد خود را از دانشگاه خواجه نصیر الدین طوسی و دکترای خود را از دانشگاه تربیت مدرس دریافت کرده است، زمینه علاقه‌مندی ایشان امنیت و دفاع ملی و تحلیل شبکه‌های اجتماعی است و در حال حاضر دانشیار دانشگاه صنعتی مالک اشتر است.

نشانی رایانامه ایشان عبارت است از:

valavi@mut.ac.ir

مریم حورعلی دوره کارشناسی خود را از دانشگاه تهران و کارشناسی ارشد و دکترای خود را از دانشگاه تربیت مدرس در سال ۹۰ دریافت کرده است. ایشان هم‌اکنون استادیار دانشگاه صنعتی مالک اشتر بوده و زمینه علاقه‌مندی وی پردازش زبان طبیعی، تحلیل احساس است.

نشانی رایانامه ایشان عبارت است از:

mhourali@mut.ac.ir

[DOI: 10.52547/jsdp.19.2.107]