مقاله پژوهشی

تشخیص عددی قطبیت با کاربست شبکه‌های عمیق بازگشتی و یادگیری بانظارت در نظرکاوی بر روی مرورهای فارسی کاربران حوزه تجارت الکترونیک

سپیده جمشیدی‌نژاد، فاطمه احمدی آبکناری و پیمان بیات

چکیده: نظرکاوی، زیرشاخه‌ای از داده‌کاوی است که به حوزه پردازش زبان طبیعی وابسته بوده و با گسترش تجارت الکترونیکی، به یکی از زمینه‌های محبوب در بازیابی اطلاعات تبدیل شده است. این حوزه بر زیرمجموعه‌های مختلفی مانند تشخیص قطبیت، استخراج جنبه و تشخیص هرزنظر تمرکز دارد. اگرچه وابستگی نهانی بین این زیرمجموعه‌ها وجود دارد اما طراحی یک چارچوب جامع شامل تمامی این موارد، بسیار چالش‌برانگیز است. پژوهش‌های موجود در این حوزه اکثراً بر روی زبان انگلیسی بوده و برای تحلیل احساس، بدون توجه به زیرمجموعه‌های تأثیرگذار، فقط بر روی حالت باینری تمرکز داشته‌اند. همچنین استفاده از یادگیری ماشینی برای دسته‌بندی نظرات بسیار رایج است و در سال‌های اخیر، اغلب پژوهش‌ها از یادگیری عمیق با اهداف متفاوت استفاده کرده‌اند. از آنجا که در ادبیات پژوهشی به چارچوبی جامع با تمرکز بر زیرمجموعه‌های تأثیرگذار کمتر پرداخته شده است، از این رو در مقاله حاضر با استفاده از راهکارهای نظرکاوی و پردازش زبان طبیعی، چارچوب جامع مبتنی بر یادگیری عمیق با نام RSAD که پیشتر توسط نویسندگان این مقاله در حوزه نظرکاوی کاربران فارسی زبان توسعه داده شده بود برای تشخیص قطبیت در دو حالت باینری و غیر باینری جملات با تمرکز بر سطح جنبه بهبود داده شده که تمام زیرمجموعه‌های لازم برای تحلیل احساس را پوشش می‌دهد. مقایسه و ارزیابی RSAD با رویکردهای موجود، نشان‌دهنده استحکام آن است.

کلیدواژه: **پردازش زبان طبیعی، تحلیل احساس، تشخیص قطبیت جملات، تشخیص هرزنظر، شبکه‌های عصبی عمیق، نظرکاوی.**

۱- مقدمه

امروزه واکنش مردم نسبت به وقایع از طریق ابراز عقاید در محیط مجازی، یک پدیده طبیعی است. آنها افکار، عقاید، احساسات و هیجانات خود را نسبت به رویدادها، مسایل اجتماعی، سیاسی، خدمات و یا محصولاتی که خریداری کرده‌اند در رسانه‌های اجتماعی بیان می‌کنند. در پردازش زبان طبیعی، این واکنش‌ها از ضعیف تا قوی دسته‌بندی می‌شوند و برچسب منفی یا مثبت دارند. تحلیل احساس[۱] و نظرکاوی[۲]، کاربردهای چشم‌گیری در زمینه‌های مختلف مانند بازاریابی و مدیریت ارتباط با مشتری دارد و با استخراج آرای مشتریان از نظرات ثبت‌شده آنلاین آنها انجام می‌شود. در واقع ارائه نظرات در رسانه‌های اجتماعی به منبعی برای تصمیم‌گیری کاربران تبدیل شده است و مردم به طور فزاینده به تجربیات کاربران بالفعل درباره یک موجودیت، متکی شده‌اند. به علاوه چنین توجهی، صاحبان مشاغل را نیز ترغیب می‌کند تا از استخراج خودکار نظرات به عنوان یک نیاز استفاده کنند اما با توجه به غیر ممکن بودن بررسی تمام مرورها توسط یک کاربر، تجزیه و تحلیل خودکار چنین مجموعه‌ای، نیاز به الگوریتم‌های توسعه‌یافته مبتنی بر پردازش زبان طبیعی (NLP)[۳] و نظرکاوی دارد. همچنین تفاوت‌های زبانی، گسترش الگوریتم‌های موجود از یک زبان به زبان دیگر را به چالش کشیده و در برخی موارد غیر ممکن می‌کند. بنابراین تحلیل احساس، به علت عدم وجود یک چهارچوب جامع در تشخیص قطبیت[۴] نظرات یک حوزه پژوهشی باز است.

تحلیل احساس بر زیرمجموعه‌های مختلف مرور (پیام درج شده کاربران درباره یک محصول، خدمت و یا موجودیت در سایت‌ها و شبکه‌های اجتماعی) مانند تشخیص قطبیت، استخراج جنبه[۵] و تشخیص هرزنظر[۶] تمرکز دارد و اگرچه این زیرمجموعه‌ها به هم وابسته هستند اما طراحی یک چارچوب جامع شامل تمامی آنها، بسیار چالش‌برانگیز است. پژوهش‌های موجود در این حوزه اکثراً بر روی زبان انگلیسی بوده و برای تشخیص قطبیت، بدون توجه به زیرمجموعه‌های تأثیرگذار، فقط بر روی حالت باینری تمرکز داشته‌اند. استفاده از یادگیری ماشینی و الگوریتم‌هایی مانند درخت تصمیم[۷]، ماشین بردار پشتیبان (SVM)[۸] و بیز ساده[۹] در پژوهش‌ها، برای دسته‌بندی نظرات در تشخیص قطبیت و تشخیص هرزنظر بسیار رایج است. پژوهش‌های اخیر نیز اغلب از یادگیری عمیق[۱۰]

این مقاله در تاریخ ۲۹ آبان ماه ۱۳۹۹ دریافت و در تاریخ ۲۲ شهریور ماه ۱۴۰۰ بازنگری شد.

سپیده جمشیدی‌نژاد، گروه کامپیوتر، واحد رشت، دانشگاه آزاد اسلامی، رشت، ایران، (email: Jamshidi@phd.iaurasht.ac.ir).

فاطمه احمدی آبکناری (نویسنده مسئول)، گروه کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور، ایران، (email: Fateme.Abkenari@gilan.pnu.ac.ir).

پیمان بیات، گروه کامپیوتر، واحد رشت، دانشگاه آزاد اسلامی، رشت، ایران، (email: bayat@iaurasht.ac.ir).

1. Sentiment Analysis
2. Opinion Mining
3. Natural Language Processing
4. Polarity Detection
5. Aspect Extraction
6. Opinion Spam
7. Decision Tree
8. Support Vector Machine
9. Navi Bayes
10. Deep Learning

و الگوریتم‌های مانند شبکه عصبی کانولوشنی (CNN)[1]، انواع شبکه‌های عصبی بازگشتی (RNN)[2] و ترکیبی از آنها برای دسته‌بندی نظرات با اهداف متفاوت استفاده کرده‌اند. همچنین در مبحث مربوط به استخراج جنبه اغلب از روش‌های مبتنی بر تکرار کلمات، مدل‌سازی موضوعی و استخراج نزدیک‌ترین اسامی به کلمات حاوی نظر به عنوان جنبه، استفاده می‌شود که در زبان فارسی به دلیل تفاوت در الگوهای زبانی و وجود ساختارهای مشکل‌ساز کارایی لازم را ندارند. با توجه به این که در ادبیات پژوهشی به چارچوبی جامع با تمرکز بر زیرمجموعه‌های تأثیرگذار در تشخیص قطبیت کمتر پرداخته شده است و تمرکز اصلی اکثر آنها بر زیرمجموعه‌ها است، در مقاله حاضر با استفاده از راهکارهای نظرکاوی و پردازش زبان طبیعی، چارچوب جامع RSAD[3] که مبتنی بر یادگیری عمیق است برای تشخیص قطبیت جملات فارسی به صورت عددی در حوزه هتل‌داری داخلی بهبود داده شده است که تمام زیرمجموعه‌های لازم برای تشخیص قطبیت مانند شناسایی ذهنیت، تشخیص هرزنظر، قطبیت جملات قبلی و ... را پوشش می‌دهد. RSAD علاوه بر دسته‌بندی باینری جملات در دو دسته مثبت یا منفی، در این مقاله قطبیت جملات را در سطح جنبه و با یک عدد صحیح تعریف می‌کند. علاوه بر پیش‌پردازش مجموعه داده اولیه، پیاده‌سازی هر یک از زیرمجموعه‌ها در RSAD نیازمند وجود زیرساخت‌هایی است که در این پژوهش ایجاد شده است. به عنوان مثال، زیرساخت ایجاد شده برای روان‌شناسی کاربر و تشخیص هرزنظر، نوآوری در ساخت و استفاده از مجموعه بدیعی از ویژگی‌های داده‌ای، مشتمل بر اطلاعات متن، فراداده، خصوصیات موجودیت و ویژگی‌های احساسی است که با اعمال آن در زبان فارسی و استخراج نظرات در سه سطح سند، جمله و کلمه به تشخیص هرزنظر می‌پردازد. برای روان‌شناسی کاربر و تشخیص هرزنظر، با نوآوری در ساخت و استفاده از مجموعه بدیعی از ویژگی‌های داده‌ای مشتمل بر اطلاعات متن، فراداده، خصوصیات موجودیت و ویژگی‌های احساسی، در تشخیص هرزنظر و اعمال آن در زبان فارسی و استخراج نظرات در سطح سند، جمله و کلمه به تشخیص هرزنظر می‌پردازد و مسأله تشخیص هرزنظر را به عنوان یک مسأله دسته‌بندی دودسته‌ای با دو دسته نظرات جعلی و غیر جعلی مبتنی بر یادگیری عمیق مدل‌سازی می‌کند. همچنین لغت‌نامه احساسی در حوزه هتل‌داری برای استفاده در ماژول تشخیص قطبیت کلمات ساخته شده است. از طرفی RSAD برای تولید نتایج دقیق‌تر از تحلیل احساس سطح جنبه یا ویژگی نیز استفاده می‌کند. به این ترتیب با تعریف بخشی به نام انباشتگر[4] در کنار توجه به قوانین دستوری زبان فارسی مانند "و/ویرگول" و "تشدیدکننده" یک روش برای غلبه بر تعدادی از ترکیبات مشکل‌ساز در قالب جنبه‌های چندکلمه‌ای در زبان فارسی ارائه کرده است. روش ارائه‌شده در این بخش ساخت یک گراف وزن‌دار و جهت‌دار را بر اساس اطلاعات به دست آمده از الگوریتم الگوکاوی مکرر FP-Growth در مجموعه جملات فارسی پیشنهاد می‌کند که از محیط پایگاه داده NoSQL، گراف Neo4J و زبان جستجوی Cypher برای رسم گراف استفاده کرده است. چارچوب RSAD به عنوان یک چارچوب نرم‌افزاری دارای ویژگی‌های زیر است:

- مبتنی بر کیفیت[5]: بر اساس جریان مشخص داده‌ای و به صورت عملیاتی انجام می‌شود.
- جدایی تمرکز بر زیرمجموعه‌های مختلف[6]: بر زیرمجموعه‌های مختلف مانند استخراج جنبه، تشخیص هرزنظر و تشخیص ذهنیت تمرکز دارد.
- یکپارچگی مفهومی[7]: هدف مشخص یعنی تشخیص قطبیت را دنبال می‌کند.
- قابلیت توسعه[8]: قابل توسعه به زبان‌های دیگر است.

مقایسه و ارزیابی هر ماژول از RSAD با رویکردهای موجود، نشان‌دهنده استحکام این چارچوب در تشخیص قطبیت است.

ادامه مقاله به این صورت سازمان‌دهی شده است: بخش بعدی، مروری بر پژوهش‌های پیشین است. سپس روش پیشنهادی مقاله، ارزیابی و مقایسه نتایج آمده و در انتها نیز نتیجه‌گیری و مراجع قرار دارند.

۲- مروری بر پژوهش‌های پیشین

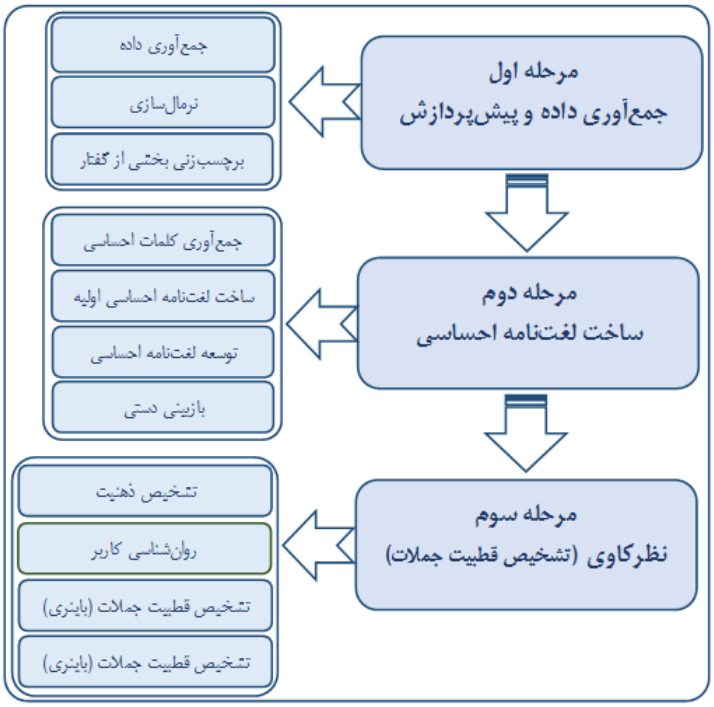
ثابتی و همکاران (۲۰۱۹) به فرایند تولید یک لغت‌نامه احساسی فارسی پرداختند. در این مقاله یک روش جدید مبتنی بر گراف برای انتخاب و گسترش بذر بر اساس هستی‌شناسی معرفی شد. سپس لغت‌نامه احساسی به یک مسأله دسته‌بندی اسناد نگاشته شد و به لغات احساسی لغت‌نامه قطبیت عددی اختصاص یافت. در ادامه نویسندگان برای دسته‌بندی از $K-$نزدیک‌ترین همسایه و نزدیک‌ترین روش‌های سنتروئید استفاده کردند. نتایج نشان می‌دهد لغت‌نامه احساسی نهایی که توسط بهترین دسته‌بند یعنی $K-$نزدیک‌ترین همسایه تولید شده است، عملکرد قابل قبولی از نظر دقت و اندازه‌گیری F در اختصاص مقادیر عددی به لغت‌نامه‌های احساسی تولیدشده دارد [۱].

تلز[9] و همکاران (۲۰۱۷) یک چارچوب چندزبانه ساده را در دسته‌بندی قطبیت برای پیاده‌سازی و استفاده آسان ارائه کردند که می‌تواند به عنوان مبانی اولیه برای مقایسه سایر سیستم‌های تحلیل احساس و همچنین نقطه شروع برای ساخت سیستم‌های جدید و پیشرفته تحلیل احساس باشد. علاوه بر تبدیلات متن، چارچوب پیشنهادی از دسته‌بندی ماشین بردار پشتیبان (با یک هسته خطی) و بهینه‌سازی hyper-parameter با استفاده از جستجوی تصادفی و $H+M$ در فضای تبدیلات متن استفاده می‌کند [۲].

ده‌خارگانی (۲۰۱۸) رویکردی نیمه‌خودکار ترکیبی برای ساخت لغت‌نامه با قطبیت کلمات پیشنهاد کرد که در زبان ترکی به عنوان یک زبان کم‌منبع آزمایش شده است. روش پیشنهادی متشکل از چندین روش از جمله احتمال وقوع کلمه و مستقل از زبان است و می‌تواند در سایر زبان‌ها با تغییرات اندک اعمال شود. از دسته‌بند لجستیک رگرسیون استفاده شده و دقت دسته‌بندی به دست آمده در استخراج و دسته‌بندی عبارات مثبت، منفی یا خنثی، اثربخشی روش پیشنهادی را تأیید می‌کند [۳].

اذانی و الفی[10] (۲۰۱۷) مدل‌های مختلف یادگیری عمیق را بر اساس شبکه‌های عصبی کانولوشنی و LSTM برای تجزیه و تحلیل احساس میکروبلاگ‌های عربی بررسی کردند. در این مقاله یک الگوی زبان عصبی با Word2vec برای بردارسازی متن ایجاد شد. سپس چندین

1. Convolution Neural Network
2. Recursive Neural Network
3. Recursive Spam Aware Deep
4. Accumulator
5. Quality_Driven
6. Separation of Concerns
7. Conceptual Integrity
8. Extensibility
9. Tellez
10. Azani and Alfy



شکل ۱: متدولوژی کلی پژوهش.

معماری یادگیری عمیق با شبکه عصبی کانولوشنی و شبکه عصبی بازگشتی LSTM طراحی و ارزیابی گردید. به طور کلی در این مقاله بیان می‌شود که جمله S از n کلمه تشکیل گردیده و با یک ماتریس $n\times k$ نشان داده شده که عنصر موجود در ردیف iام با یک بردار kبعدی از کلمه iام مطابقت دارد [۴].

دشتی‌پور و همکاران (۲۰۲۰) یک چارچوب ترکیبی جدید را برای تجزیه و تحلیل احساس در سطح مفهوم در زبان فارسی پیشنهاد کردند که قواعد زبانی و یادگیری عمیق را برای بهینه‌سازی تشخیص قطبیت ادغام می‌کند. در این مقاله قوانین حاکم بر جملات فارسی مانند متمم، وارونگی، حرف نقض، حرف اضافه و ... و روابط وابستگی سلسله‌مراتبی برای تعیین دقیق‌تر احساسات در مقایسه با رویکردهای مبتنی بر فرکانس هم‌رخداد کلمه استفاده می‌شود. رویکرد ترکیبی پیشنهادی ترکیبی از یادگیری عمیق و قوانین مبتنی بر وابستگی برای حل مسأله جملات دسته‌بندی نشده است [۵].

چاندرا و جانا[1] (۲۰۲۰) مدلی را روی توئیت توسعه دادند که از سه روش ۱) مبتنی بر قطبیت، ۲) مبتنی بر یادگیری ماشینی و یادگیری عمیق و ۳) با آموزش ویژگی‌ها برای رسیدن به دسته‌بندی استفاده می‌کند. ابتدا از روش‌های یادگیری ماشینی مانند نوی بیز، برنولی، رگرسیون لجستیک و ... برای دسته‌بندی توئیت‌ها استفاده شد. سپس با استفاده از روش مبتنی بر قطبیت، درصد مثبت یا منفی‌بودن توئیت کاربران را شناسایی کردند و سرانجام از روش‌های یادگیری عمیق مانند ترکیب شبکه عصبی کانولوشنی و شبکه عصبی بازگشتی LSTM برای دسته‌بندی استفاده گردید. روش‌های یادگیری عمیق بهترین نتیجه را در دسته‌بندی به دست آوردند [۶].

چن[2] و همکاران (۲۰۱۸) یک مدل شبکه عصبی عمیق را که ترکیب شبکه عصبی کانولوشنی و LSTM است، برای تحلیل احساس ارائه دادند. این روش می‌تواند مدت زمان آموزش مدل شبکه عصبی را از طریق LSTM کاهش دهد. مدل ارائه‌شده هم‌زمان از یک شبکه عصبی کانولوشنی در سطح جمله استفاده می‌کند تا ویژگی‌های احساسی کل جمله را استخراج کند و انتقال اطلاعات را از طریق ماتریس‌های وزنی متفاوت با هم کنترل نماید. سیستم پیشنهادی روی مجموعه داده‌های چند دامنه دو زبان SemEval۲۰۱۶ نشان داد که عملکرد بهتری نسبت به ماشین بردار پشتیبان و چند مدل شبکه عصبی مانند LSTM و شبکه عصبی کانولوشنی دارد [۷].

الماسری و همکاران (۲۰۱۷) از تجزیه و تحلیل احساس برای تحلیل موضوعات پیشگیرانه مانند بحران‌های سیاسی استفاده کردند. در این مقاله، ابزاری ارائه شد که تجزیه و تحلیل معنایی توئیت‌های عربی را با پارامترهای ترکیبی انجام می‌دهد. این پارامترها شامل زمان توئیت، پیش‌پردازش مانند ریشه‌یابی و بازخوانی، ویژگی n_grams بود که با روش مبتنی بر لغت‌نامه و روش‌های یادگیری ماشینی تحلیل شدند. آزمایش‌ها در این مقاله نشان دادند که روش یادگیری ماشین بیز ساده در پیش‌بینی قطبیت موضوع، دقیق‌ترین است [۸].

ژانگ[3] (۲۰۲۰) یک مدل دسته‌بندی احساس را بر اساس لغت‌نامه احساسی و الگوریتم شبکه عصبی کانولوشنی ارائه داد و از داده‌های پلتفرم تجارت الکترونیکی در مدت زمان معینی استفاده کرد تا صحت مدل خود را بررسی کند. مقاله شامل دو بخش است. ابتدا از روش لغت‌نامه احساسی برای حاشیه‌نویسی مجموعه داده‌های اصلی نظر استفاده کردند که توانست به طور چشم‌گیری بهره‌وری و دقت حاشیه‌نویسی را بهبود بخشد. سپس نظرات برچسب‌گذاری شده کاربران در مورد دوربین در مدل یادگیری عمیق و مدل ماشین بردار پشتیبان را وارد کردند. نتایج نشان می‌دهد که صحت مدل دسته‌بندی احساس بر اساس یادگیری عمیق نسبت به مدل ماشین بردار پشتیبان بیشتر است [۹].

1. Chandra and Jana
2. Chen
3. Zhang

۳- متدولوژی پژوهش

این پژوهش مطابق متدولوژی ارائه‌شده در شکل ۱، در سه مرحله کلی به مدل‌سازی تحلیل احساس جملات می‌پردازد. در مرحله اول بعد از جمع‌آوری مجموعه داده اولیه، عملیات پیش‌پردازش شامل نرمال‌سازی و برچسب‌زنی بخشی از گفتار[1] انجام شده است. مرحله دوم به ساخت لغت‌نامه احساسی در حوزه هتل‌داری می‌پردازد. در مرحله سوم نیز نظرکاوی و تحلیل احساس در سطح جمله و جنبه انجام خواهد شد. توضیحات هر مرحله از متدولوژی در ادامه شرح داده می‌شود.

۳-۱ جمع‌آوری مجموعه داده و پیش‌پردازش

داده‌های متنی، فرمت اصلی داده‌های رسانه‌های اجتماعی است که بازخورد، نگرش، نظرات و افکار کاربران را نسبت به محصولات، خدمات، سیاست‌ها و سایر موضوعات نشان می‌دهند [۱۰].

داده‌های متنی مقاله حاضر، نظرات فارسی کاربران در حوزه هتل‌داری در بازه زمانی پنج سال است که از دو سایت iranhotelonline.com و egardesh.com با خزنده وب[2] در بستر C# جمع‌آوری شده است. این نظرات شامل ۱۰۰۰۰ پاراگراف نظر با طول متفاوت (یک، دو، سه و ... جمله‌ای) است. بعد از جمع‌آوری نظرات، اولین چالش، پیش‌پردازش نظرات است. از آنجا که نظرات درج‌شده در وب‌سایت‌ها و شبکه‌های اجتماعی، عامیانه، گاه با غلط‌های املایی، به صورت خلاصه، با تکرار حروف و ... هستند، برای استفاده از نظرات باید مجموعه داده اولیه نرمال‌سازی شود. بنابراین نرمال‌سازی با هدف حذف نظرات خالی و نامشخص، حذف اعداد، شکلک‌ها و کلمات نامرتبط، نرمال‌سازی کلمات و نشانه‌گذاری، روی مجموعه داده متنی مقاله انجام شد. بعد از مرحله نرمال‌سازی، نقش کلمات با استفاده از برچسب‌زنی بخشی از گفتار تعیین خواهد شد. عملیات نرمال‌سازی و برچسب‌زنی بخشی از گفتار با استفاده از ابزار NLPTOOLs دانشگاه فردوسی مشهد انجام شده است [۱۱].

۳-۲ ساخت لغت‌نامه احساسی

مهم‌ترین شاخص‌های احساسات، کلمات احساسی هستند که کلمات نظر نامیده می‌شوند و ابزارهایی برای تحلیل احساس هستند. به عنوان مثال، "خوب، شگفت‌انگیز و متحیرکننده" کلمات احساسی مثبت و "بد، فقیر و وحشتناک" کلمات احساسی منفی هستند. یک لیست از چنین کلماتی، یک لغت‌نامه احساسی نامیده می‌شود (لغت‌نامه نظر). از آنجایی که در تحلیل احساس، به طور گسترده از کلمات احساسی استفاده می‌شود و با توجه به چالش‌های موجود مانند عدم وجود لغت‌نامه‌های احساسی در یک دامنه خاص در زبان فارسی و ترجمه‌بودن اکثر منابع لغوی فارسی از زبان انگلیسی، پژوهش‌های حوزه تحلیل احساس در زبان فارسی چالش‌برانگیز و وقت‌گیر است. بنابراین، این بخش با جمع‌آوری کلمات احساسی برای اولین بار اقدام به ساخت لغت‌نامه احساسی در حوزه هتل‌داری نمود. با استفاده از برچسب‌زنی بخشی از گفتار کلمات هر جمله، کلمات احساسی و حاوی نظر با نقش دستوری صفت و قید به ترتیب با برچسب ADJ و ADV استخراج و لغت‌نامه احساسی اولیه تولید شد. سپس برای یافتن مترادف‌ها و متضادهای کلمات، مجموعه اولیه با جستجو در لغت‌نامه‌های آنلاین وسیع‌تر شد. بعد از افزودن کلمات جدید به دست آمده به لیست اولیه، به شرط وجود کلمه جدید، تکرار بعدی آغاز می‌شود. زمانی که هیچ کلمه جدیدی پیدا نشد، فرایند تکرار به پایان می‌رسد. بعد از اتمام فرایند، یک مرحله بازبینی و کنترل دستی نیز برای تصدیق لیست کلمات لغت‌نامه انجام می‌شود تا خطاهای احتمالی در جمع‌آوری کلمات برطرف شود. این روش در پژوهش هو و لیو (۲۰۰۴) نیز استفاده شد [۱۲]. در انتهای این مرحله لغت‌نامه احساسی با ۳۰۶۲ کلمه شامل صفت، قید و اسم تولید شد. همچنین به همین روش برای تبدیل لغت‌نامه احساسی به لغت‌نامه احساسی حوزه هتل‌داری، مجموعه اسامی پرتکرار در حوزه هتل‌داری مانند لابی، محوطه، معماری، طراحی، تجهیزات و ... از نظرات جدا و به لغت‌نامه احساسی اضافه شدند. بنابراین در انتها لغت‌نامه احساسی در حوزه هتل‌داری با ۴۵۷۰ لغت احساسی و تخصصی شامل صفت، قید و اسم ساخته شد.

۳-۳ نظرکاوی (تشخیص قطبیت جملات)

در این مرحله با استفاده از راهکارهای نظرکاوی و پردازش زبان طبیعی، چارچوب جامع مبتنی بر یادگیری عمیق با نام RSAD برای تحلیل احساس جملات بهبود داده شده که تمام زیرمجموعه‌های لازم برای تحلیل احساس، مانند تشخیص ذهنیت، روان‌شناسی کاربر، قطبیت جملات قبلی و ... را پوشش می‌دهد. RSAD علاوه بر دسته‌بندی باینری جملات به صورت مثبت یا منفی، به طور خاص بر استخراج قطبیت عددی در سطح جنبه تأکید دارد.

چارچوب RSAD مطابق شکل ۲ برای تحلیل احساس جملات، شامل چهار زیربخش A تا D است که هر یک از آنها بر اساس یک مدل شبکه عصبی عمیق ساخته خواهد شد. همان طور که در این شکل نشان داده شده است، ماژول C وظیفه تشخیص قطبیت هر جمله را بر اساس دو ساختار ورودی مختلف از دو ماژول A و B بر عهده دارد. ماژول D بعد از ماژول C کار می‌کند تا یک مقدار عددی صحیح به خروجی ماژول C اختصاص دهد. دلیل بهبود کیفیت عددی قطبیت به دست آمده در RSAD را می‌توان تمرکز بر تمام زیرمجموعه‌های لازم برای تشخیص قطبیت و اجتماع آنها در یک چارچوب بیان کرد که در ادامه به طور مختصر به این زیرمجموعه‌ها اشاره می‌شود:

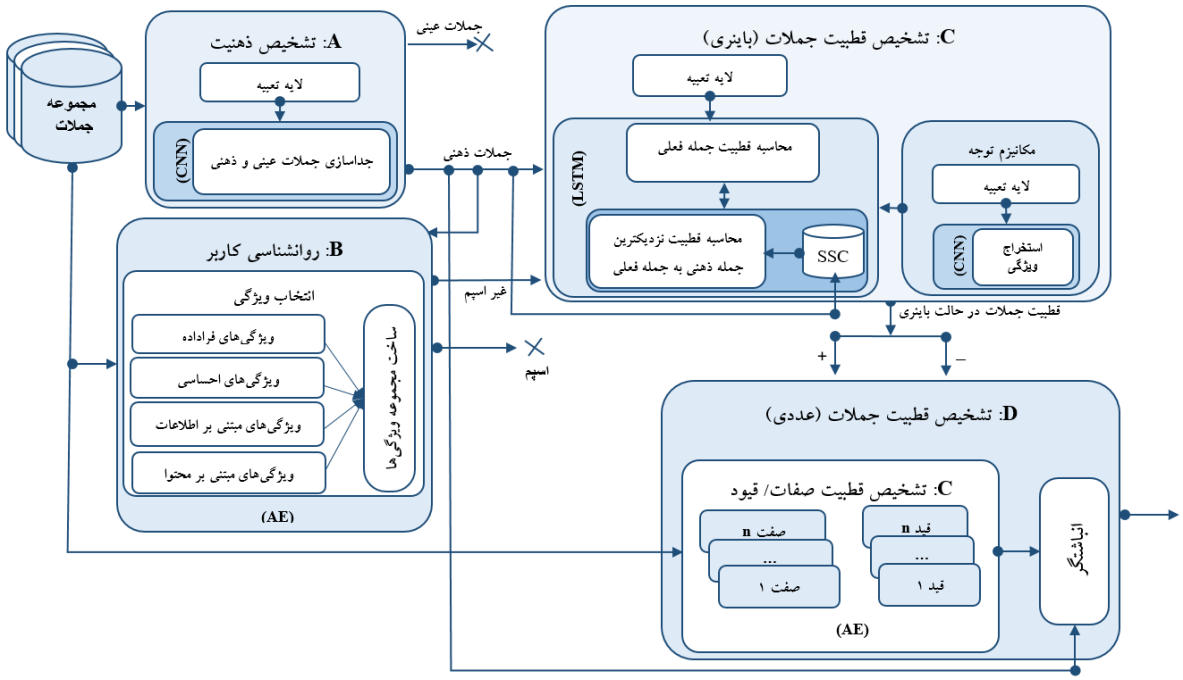
۱) تشخیص ذهنیت: وظیفه این بخش تفکیک جملات عینی (جملاتی که حاوی اطلاعات واقعی هستند) و جملات ذهنی (جملاتی که اغلب احساس، دیدگاه‌ها یا باورهای شخصی را بیان می‌کنند) بر اساس دسته‌بندی است.

۲) روان‌شناسی کاربر: در این بخش با تشخیص نظرات هرز و غیر واقعی و جداسازی آنها از نظرات واقعی کاربران، تمرکز RSAD روی نظرات واقعی گذاشته می‌شود.

۳) تشخیص قطبیت جملات (باینری): این بخش مسئول جداسازی جملات احساسی مثبت و منفی است. به طوری که برای جلوگیری از تعیین قطبیت اشتباه در نتیجه وجود ضرب‌المثل‌ها و جملات مبهم که کاربران ذکر کرده‌اند و در معنای واقعی استفاده نشده است محاسبه قطبیت جمله فعلی تحت تأثیر قطبیت نزدیک‌ترین جمله ذهنی به آن در نظر گرفته می‌شود.

۴) تشخیص قطبیت جملات (عددی): در این بخش دو قسمت مجزا برای بهبود کیفیت قطبیت نهایی جملات به صورت عددی ارائه شده است. وظیفه اصلی قسمت اول تشخیص قطبیت صفات/ قیود از طریق دسته‌بندی آنها در دو دسته مثبت یا منفی است و سپس بر اساس تعداد رخداد صفات مثبت یا منفی، یک مقدار عددی به آنها اختصاص داده می‌شود. تعداد رخداد صفات (مثبت یا منفی)، به

1. Part of Speech (POS) Tagging
2. Web Crawler



شکل ۲: چارچوب RSAD.

ازای جنبه‌های تعریف‌شده ارزیابی می‌شود. بنابراین در تشخیص قطبیت عددی جملات به نظرکاوی مبتنی بر جنبه و ملاحظات دستوری زبان فارسی پرداخته خواهد شد که وظیفه بخش دوم یعنی انباشتگر است. انباشتگر به عنوان اولین وظیفه، جنبه‌ها را استخراج می‌کند و سپس با توجه به ملاحظات دستوری زبان فارسی، به محاسبه و تخصیص قطبیت عددی به جملات می‌پردازد.

توجه به این نکته حایز اهمیت است که در RSAD فقط اجتماع زیرمجموعه‌های تأثیرگذار در تشخیص قطبیت برای بهبود کیفیت قطبیت عددی انجام نشده است، بلکه علاوه بر پیش‌پردازش مجموعه داده اولیه، در هر زیرمجموعه، ایده و نوآوری وجود دارد که شامل موارد زیر است:

۱) نوآوری در استفاده از شبکه‌های عصبی عمیق در تشخیص ذهنیت.

۲) نوآوری در ساخت و استفاده از مجموعه بدیعی از ویژگی‌های داده‌ای، مشتمل بر اطلاعات متن، فراداده، خصوصیات موجودیت و ویژگی‌های احساسی در تشخیص هرزنظر و اعمال آن در زبان فارسی و استخراج نظرات در سطح سند، جمله و کلمه به تشخیص هرزنظر می‌پردازد.

۳) نوآوری در ساخت لغت‌نامه احساسی در حوزه هتل‌داری برای استفاده در ماژول تشخیص قطبیت عددی کلمات.

۴) نوآوری در ساخت و استفاده از مجموعه بدیعی از ویژگی‌های داده‌ای برای تعیین قطبیت صفات و قیود لغت‌نامه احساسی.

۵) تولید نتایج دقیق‌تر با استفاده از تحلیل احساس سطح جنبه یا ویژگی.

از ۱۰۰۰۰ پاراگراف نظر که در مرحله اول جمع‌آوری شد، ۳۰۰۰ پاراگراف نمونه که شامل ۱۲۶۷۳ جمله با طول متفاوت است به عنوان مجموعه داده نمونه جداسازی و برای تحلیل احساس جملات، وارد چارچوب RSAD خواهد شد. در ادامه توضیحات هر ماژول به همراه روش پیاده‌سازی آن به تفصیل توضیح داده می‌شود.

۳-۳-۱ تشخیص ذهنیت

وظیفه این بخش تفکیک جملات عینی و ذهنی بر اساس دسته‌بندی است. برای اجرای چنین وظیفه‌ای از دو لایه مبتنی بر شبکه عصبی شامل لایه تعبیه کلمات و لایه کانولوشنی استفاده شده است.

برای استفاده از یک ساختار عصبی عمیق که کلمات هر جمله را درک کند باید ویژگی‌های کلمات، در لایه تعبیه استخراج شوند. Word۲vec این وظیفه را از طریق یک رویکرد بدون نظارت در ساختار شبکه عصبی دولایه، ارائه می‌دهد. مدل Word۲vec در دو نسخه مختلف شامل مدل مبتنی بر کیسه‌ای از کلمات و مدل مبتنی بر Skip-Gram موجود است. در این مقاله، از مدل Skip-Gram به دلیل کارایی آشکار آن در زمینه تجزیه و تحلیل احساس که در پژوهش حسن و همکاران (۲۰۱۷) ثابت گردید برای تولید بردار کلمات ورودی استفاده شده است [۱۳]. Word۲vec هر جمله را از ورودی دریافت می‌کند و کلمات را در یک جمله پشت سر هم در نظر گرفته و ۳۰ ستون ویژگی، از کلمات استخراج می‌کند. جزئیات ساختار این شبکه در جدول ۱ نشان داده شده و شکل ۳ دقت را در تنظیمات مختلف دورها و اندازه دسته برای این شبکه نشان می‌دهد. ورودی لایه تعبیه ۱۲۶۷۳ جمله است. این جملات، ابتدا با ابزار پیش‌پردازش پارسی‌ور در زبان فارسی برای انجام وظایف نرمال‌سازی جملات و نشانه‌گذاری پیش‌پردازش شدند [۱۴].

در بخش دوم ماژول A، یک لایه کانولوشنی پیاده‌سازی می‌شود که خروجی لایه تعبیه یعنی ماتریس تعبیه شده، ورودی آن است. از آنجا که هدف ماژول A دسته‌بندی جملات در دو دسته عینی و ذهنی است، ماتریس لایه تعبیه به همراه ستون برچسب به شبکه عصبی کانولوشنی عمیق وارد می‌شود. برای شبکه عصبی کانولوشنی، دو فیلتر با اندازه ۳ و ۵ و دو لایه Pooling تعریف شده و برای هموارسازی نتایج، تابع ReLU روی خروجی هر لایه اعمال خواهد گردید. بعد از استفاده از یک لایه Flatten، یک لایه کاملاً متصل نیز در انتها با تابع Sigmoid به کار گرفته شده است تا نتایج دسته‌بندی به دو کلاس پیاده‌سازی شده ارائه شود. جزئیات ساختار این شبکه در جدول ۱ نشان داده شده است. بهترین دقت ۹۹٫۸۸ در ۳۰ دور تکرار (شکل ۳) با توزیع ۷۰ به ۳۰ برای نمونه‌های آموزشی و آزمایشی به دست آمده است. نمونه‌های ذهنی در چارچوب تشخیص قطبیت به ماژول B می‌روند تا تحت کنترل هرزنظر پردازش شوند.

جدول ۱: جزئیات ساختار چهار شبکه عصبی استفاده‌شده در تشخیص قطبیت جملات.

Network Type	Container Module	Number of Layers	Epoch	Batch Size	Accuracy
Convulsive Neural Network (CNN)	A	Embedding Layer sentences = corpus size = 30 window = 3 min count = 1 workers = 1 Iteration = 100 CNN Layer Layer1 = Kernel size = 5, activation = ReLU Layer2 = Pool size = 2, activation = ReLU Layer3 = Kernel size = 3, activation = ReLU Layer4 = Pool size = 2, activation = ReLU Layer5 = Flatten Layer6 = Dense(1), activation = Sigmoid	۳۰	۱۲۸	۹۹٫۸۸
Auto-Encoder	B	Layer1 = Dense(160), activation = ReLU Layer2 = Dense(120), activation = ReLU Layer3 = Dense(80), activation = ReLU Layer4 = Dense(40), activation = ReLU Layer5 = Dense(20), activation = ReLU Layer6 = Dense(8), activation = ReLU Layer7 = Dense(1), activation = Sigmoid	۳۰	۶۴	۹۹٫۸۹
Long Short-Term Memory (LSTM)	C	Embedding Layer sentences = corpus size = 64 window = 3 min count = 1 workers = 1 Iteration = 100 Attention Layer Layer1 = Kernel size = 3, activation = ReLU Layer2 = Pool size = 2, activation = ReLU LSTM Layer Layer3 = LSTM(100) Layer4 = Dense(1), activation = Sigmoid	۴۰	۲۵۶	۹۹٫۹۳
Auto-Encoder	D	Layer1 = Dense(160), activation = ReLU Layer2 = Dense(120), activation = ReLU Layer3 = Dense(80), activation = ReLU Layer4 = Dense(60), activation = ReLU Layer5 = Dense(40), activation = ReLU Layer6 = Dense(20), activation = ReLU Layer7 = Dense(10), activation = ReLU Layer8 = Dense(1), activation = Sigmoid	۳۰	۵۱۲	۹۴٫۹۲

۳-۳-۲ روان‌شناسی کاربر

هدف ماژول B تمایز کاربران اسپم (کاربران با نظر غیر واقعی و هرزنظر) از غیر اسپم (کاربران با نظر واقعی) است تا از نفوذ نظرات هرز در روند تشخیص قطبیت جمله جلوگیری شود. به طور کلی مرحله تشخیص هرزنظر با نوآوری در ساخت و استفاده از مجموعه بدیعی از ویژگی‌های داده‌ای مشتمل بر اطلاعات متن، فراداده، خصوصیات موجودیت و ویژگی‌های احساسی و اعمال آن در زبان فارسی در سطح سند، جمله و کلمه به عنوان یک مسأله دسته‌بندی دودسته‌ای با دو دسته نظرات جعلی و غیر جعلی مدل‌سازی شده است. ابتدا ۱۲۶۷۳ از کاربران مختلف و با طول متفاوت متنی، انتخاب شد. مجموعه داده اولیه برای مدل‌سازی تشخیص هرزنظر با ۲۳ ستون شامل یک ستون برچسب و ۲۲ ستون ویژگی پیشنهادی ایجاد شد. به طور کلی برای تشخیص هرزنظر در مرورها و نظرات از سه یا چهار نوع داده اصلی استفاده شده که عبارت است از:

۱) محتوای مرور و نظر: این دسته شامل ویژگی‌های زبان‌شناختی و یا معناشناختی برای فریب و نیرنگ کاربران است. در این گروه می‌توان n-گرم‌ها و برچسب‌زنی بخشی از گفتار را تعریف کرد، مانند مرور تکراری یا تعداد کلمات.

۲) فراداده در مرور و نظر: از این داده‌ها می‌توان انواع متعددی از الگوهای رفتاری غیر عادی کاربران را کشف کرد، مانند شناسه کاربری نویسنده.

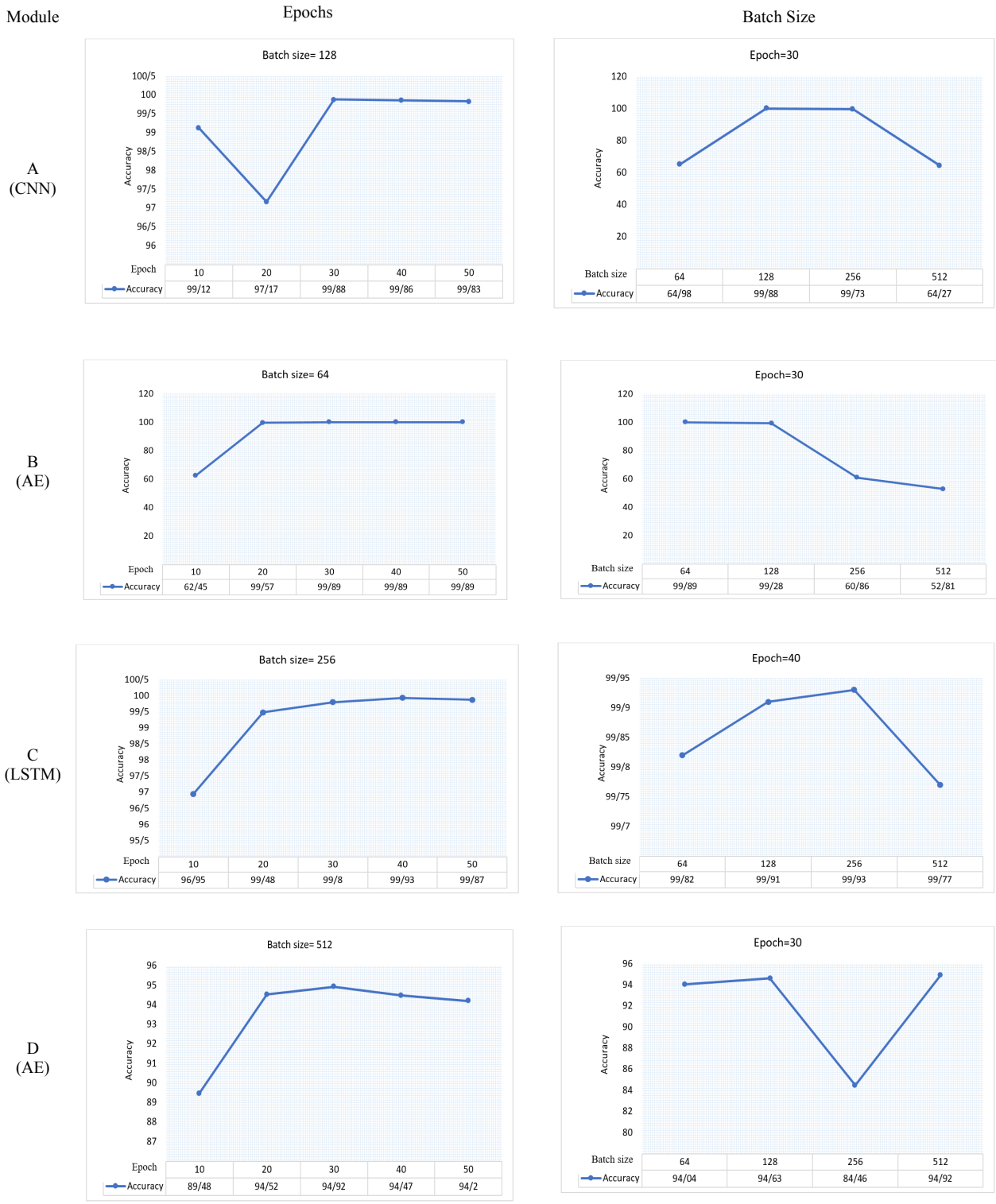
۳) اطلاعات مبتنی بر موجودیت: این داده‌ها اطلاعاتی درباره موجودیت مورد بحث هستند و به شرح آن می‌پردازند، مانند رتبه کلی هتل و قیمت اتاق‌های هتل.

۴) ویژگی‌های احساسی: این دسته از ویژگی‌ها مربوط به ذهنی یا عینی‌بودن جملات و مثبت یا منفی‌بودن کلمات احساسی استفاده‌شده در متن هستند، مانند قطبیت نظر یا تعداد کلمات مثبت و منفی [۱۵] و [۱۶].

بنابراین در این پژوهش مجموعه‌ای از ویژگی‌ها برای زبان فارسی بر اساس چهار گروه ویژگی ذکرگردیده در بالا توسعه داده شد. جدول ۲ دسته‌بندی مجموعه ۲۲ ویژگی را در چهار گروه ویژگی‌های ذکرشده نشان می‌دهد.

مجموعه ویژگی‌های توسعه‌یافته در این پژوهش برای فرایند تشخیص هرزنظر (جدول ۲) شامل ۲۲ ویژگی (بدون در نظر گرفتن ستون برچسب به عنوان جعلی یا غیر جعلی) است.

تعدادی از این ویژگی‌ها توسط پژوهشگران دیگر معرفی و استفاده شده‌اند [۱۵] تا [۱۷]. تعدادی از ویژگی‌های معرفی‌شده، نسبت به پژوهش‌های دیگر بدون تغییر هستند و فقط با مجموعه داده پژوهش حاضر تطبیق یافتند. این ویژگی‌ها (جدول ۳، گروه A) در زبان انگلیسی تعریف شده‌اند و مقادیر آنها در سطح جمله با توجه به مقالات پایه پر شده است. همچنین تغییراتی در معنی و کاربرد تعدادی از ویژگی‌ها ایجاد و سپس با مجموعه داده پژوهش حاضر تطبیق داده شدند. نمونه‌هایی از



شکل ۳: دقت چهار شبکه عصبی ماژول‌های مختلف چارچوب RSAD در تنظیمات متفاوت دور و اندازه دسته.

ویژگی‌ها، (جدول ۳، گروه B) "زمان ارسال نظر" است که به عنوان "فاصله زمانی ارسال نظر نسبت به زمان افتتاح هتل" تغییر یافت. مثال دیگر در این گروه ویژگی "اختلاف بین رتبه‌بندی محصول توسط کاربر و میانگین امتیاز محصول" است که به "تطبیق قطبیت نظر با رتبه کلی هتل" تغییر داده شد و از قطبیت نظر به جای رتبه‌بندی محصول استفاده گردید. علاوه بر این دو گروه ویژگی، برای اولین بار تعدادی از ویژگی‌های ابتکاری به مجموعه ویژگی‌ها اضافه شده است (جدول ۳، گروه C). تعدادی از ویژگی‌ها در گروه C عبارت است از:

۱) مرورهای حاوی تبلیغات: این نوع هرزنظرها امروزه به روندی برای تبلیغ سایر هتل‌ها یا خدمات مرتبط با گردشگری تبدیل شده‌اند. برخی از آنها از طریق ارائه خدمات بر اساس موقعیت هتل از نظر فکری طراحی می‌شوند.

۲) نظرات عمومی: کاربران این عقاید را در مورد کل موجودیت بیان می‌کنند نه در مورد هر جنبه از موجودیت مانند "هتل خوبی بود". با توجه به این که بیان‌کنندگان هرزنظر معمولاً مهمان هتل نبوده و هیچ تجربه‌ای در مورد هر جنبه از هتل ندارند نظرات ارسالی آنها، نظرات کلی است.

۳) علامت قطبیت نظر: در این ویژگی تعداد کل نظرات با قطبیت

جدول ۲: دسته‌بندی مجموعه ویژگی‌های تشخیص هرزنظر در چهار گروه.

محتوای مرور	فراداده	اطلاعات موجودیت	احساسی
مرور تکراری	شناسه کاربری تکراری	رضایت از قیمت هتل	علامت قطبیت نظرات
شباهت مرور	تعداد مرورهای یک نویسنده	رتبه کلی هتل	تعداد کلمات احساسی مثبت
مرور حاوی تبلیغات	درصد مرورهای نویسنده نسبت به کل مرورها	–	تعداد کلمات احساسی منفی
تعداد کلمات	زمان ارسال نظر از زمان افتتاح هتل	–	تعداد جملات ذهنی
تعداد جملات	تطبیق قطبیت نظر با رتبه کلی هتل	–	تعداد جملات عینی
نظرات عمومی	–	–	درصد کلمات احساسی مثبت به کل کلمات نظر
–	–	–	درصد کلمات احساسی منفی به کل کلمات نظر
–	–	–	درصد جملات ذهنی به کل جملات نظر
–	–	–	درصد جملات عینی به کل جملات نظر

جدول ۳: دسته‌بندی ویژگی‌های پژوهش در سه گروه A، B و C.

C	B	A
نظرات عمومی	زمان ارسال نظر از زمان افتتاح هتل	مرور تکراری
علامت قطبیت نظر	تطبیق قطبیت نظر با رتبه کلی هتل	شناسه کاربری تکراری
مرور حاوی تبلیغات	تعداد جملات ذهنی	رتبه کلی هتل
تعداد کلمات	تعداد جملات عینی	شباهت مرورها
تعداد جملات	تعداد کلمات احساسی مثبت	رضایت از قیمت هتل
–	تعداد کلمات احساسی منفی	تعداد مرورهای یک نویسنده
–	–	درصد کلمات احساسی مثبت به کل کلمات مرور
–	–	درصد کلمات احساسی منفی به کل کلمات مرور
–	–	درصد مرورهای نویسنده نسبت به کل مرورها
–	–	درصد جملات ذهنی به کل جملات نظر
–	–	درصد جملات عینی به کل جملات نظر

مثبت و تعداد کل نظرات با قطبیت منفی را محاسبه می‌کنیم و بعد از تفریق، علامت نهایی قطبیت نظر را به دست می‌آوریم. زیرا بسیاری از بیان‌کنندگان هرزنظر از بعضی احساسات مثبت در تعدادی جنبه‌های هتل استفاده می‌کنند، اما هدف اصلی آنها از بین بردن شهرت هتل با ارائه نظرات منفی در مورد بسیاری از جنبه‌های دیگر است.

۴) تعداد کلمات و تعداد جملات: در مجموعه داده‌های فارسی این پژوهش کشف شد که بیان‌کنندگان هرزنظر فارسی معمولاً نظرات جعلی خود را در قالب مرور با طول کوتاه‌تر، تعداد جملات و کلمات کمتر، ارسال می‌کنند زیرا هیچ تجربه‌ای در مورد جنبه‌های موجودیت (در اینجا هتل) برای بیان مشاهدات واقعی خود ندارند.

سپس برچسب‌زنی داده‌ها و آماده‌سازی مجموعه داده انجام می‌شود. برای انجام این وظیفه از ۲۲ ویژگی جدول ۲ استفاده شد. قبل از برچسب‌زنی داده‌ها، مقدار ویژگی‌ها را از طریق اسکن نظرات در سطح سند، جمله و کلمه با استفاده از لغت‌نامه احساسی، به ویژه برای ویژگی‌های احساسی پر کردیم. به عنوان مثال برای تشخیص جملات ذهنی یا عینی هنگامی که یک کلمه احساسی در لغت‌نامه در چنین جملاتی ظاهر می‌شود استفاده شده است. بنابراین با توجه به وجود کلمه احساسی، جمله ذهنی "یک" و جمله عینی "صفر" مقداردهی شد و سپس تعداد آنها در جریان مقداردهی استفاده خواهد شد. نظرات عمومی نیز نظرات کلی هستند که در مورد موجودیت هتل بیان می‌شوند نه جنبه‌ای از هتل که "یک" برای نظرات عمومی و "صفر" برای جملاتی است که در مورد جنبه‌ها صحبت می‌کنند. برای تعداد جملات و کلمات نیز شمارش انجام شد و مقدار نهایی ویژگی مورد نظر، تعداد این ویژگی‌ها است. فرایند برچسب‌زنی به صورت تکراری انجام می‌شود. برای پرکردن ستونی که با نام برچسب به مجموعه داده، اضافه شده است، محاسبات OR میان ۲۲ ویژگی $\langle f1, f2, \ldots, f22 \rangle$ در نظر گرفته شد. مقادیر برچسب "بله" برای هرزنظر و "خیر" برای غیر هرزنظر تعریف شده است.

برای پیاده‌سازی ماژول B از بخش رمزگذار شبکه خودرمزنگار[1] عمیق استفاده می‌شود به طوری که نیمه رمزگشای آن حذف شده است. برای کسب بهترین نتیجه، شبکه با پنج لایه مخفی کاملاً متصل با ۲۲۰، ۱۱۰، ۵۵، ۲۸ و ۱۴ گره، از طریق تابع فعال‌ساز ReLU و لایه خروجی با یک گره، با تابع فعال‌ساز سیگموئید در پایتون پیاده‌سازی شده است. داده‌های آموزشی به آزمایشی توزیع ۷۰ به ۳۰ دارند و دو پارامتر loss و Optimizer به ترتیب با مقادیر binary_crossentropy و adam تنظیم شده‌اند. بهترین نتیجه ۹۱/۹۹ در ۱۰ دور به دست آمده است. ساختار این شبکه عصبی عمیق در جدول ۱ ارائه شده و شکل ۳ دقت را در تنظیمات مختلف دورها و اندازه دسته برای این شبکه نشان می‌دهد.

۳-۳-۳ تشخیص قطبیت جملات (باینری)

ماژول C مسئول محاسبه قطبیت جملات به صورت مثبت یا منفی است. برای انجام این وظیفه، ماژول C روی جملات ذهنی خروجی ماژول A و نظرات کاربران غیر اسپم که از طریق ماژول B شناسایی شده‌اند کار می‌کند. همچنین برای جلوگیری از تعیین قطبیت اشتباه در نتیجه وجود ضرب‌المثل‌ها و جملات مبهم که کاربران ذکر کرده‌اند و در معنای واقعی استفاده نشده است، محاسبه قطبیت جمله فعلی تحت تأثیر قطبیت نزدیک‌ترین جمله ذهنی به آن در نظر گرفته شده است. برای این منظور

1. Autoencoders

جدول ۴: مجموعه طراحی‌شده و پیشنهادی ویژگی‌ها برای تشخیص قطبیت صفات/ قیود.

ردیف	ویژگی	توضیحات
۱	درجه صفت	تفاوت صفت‌هایی مانند خوب و عالی را نشان می‌دهد که شدت‌های مختلف دارند.
۲	آیا صفت در کاربرد غیر رسمی معنای دیگری دارد؟	تعدادی کلمات در زبان عامیانه با مفهومی متفاوت از مفهوم واقعی خود استفاده می‌شوند. مثلاً "مهندس" در زبان عامیانه فارسی به یک فرد بی‌احتیاط اشاره دارد.
۳	تعداد جملات با قطبیت منفی نسبت به کل جملاتی که تا کنون از این صفت استفاده کرده‌اند.	–
۴	تعداد جملات با قطبیت مثبت نسبت به کل جملاتی که تا کنون از این صفت استفاده کرده‌اند.	–
۵	تعداد جملات دارای قطبیت منفی با فعل نفی به کل جملاتی که تا کنون از این صفت استفاده کرده‌اند.	–
۶	آیا می‌توان کلمه نفی را به این صفت پیوند داد؟	برخی از صفت‌ها مانند مناسب می‌توانند پیشوند منفی مانند نامناسب بپذیرند.
۷	آیا کلمات نفی قطبیت جمله را تغییر داده است؟	در اینجا تأثیر کلمات نفی یا فعل منفی بر قطبیت جمله بررسی می‌شود.
۸	تعداد جملاتی که صفت در آنها ظاهر می‌شود.	–
۹	آیا صفت با توجه به جنبه تغییر می‌کند؟	به عنوان مثال "گرم" در جنبه "غذا" معنای مثبتی دارد اما در جنبه "دمای اتاق" معنی منفی دارد.

از یک ساختار شبکه عصبی عمیق بازگشتی LSTM استفاده می‌شود. ماژول C با سه لایه شامل لایه تعبیه، لایه توجه و لایه LSTM در ادامه شرح داده می‌شود.

لایه اول لایه تعبیه است که ورودی آن ۸۰۹۳ جمله ذهنی بدون اسپم می‌باشد. این جملات که از دو ماژول A و B به دست آمده‌اند وارد شبکه عصبی کم‌عمق Word۲vec شدند. ابتدا عملیات پیش‌پردازش شامل نرمال‌سازی جملات و نشانه‌گذاری آنها با ابزار پارسی‌ور انجام شد. برای تبدیل جملات به بردار از مدل Skip-Gram با سایز ابعاد ۶۰ برای هر کلمه در بردار تعبیه استفاده شده است. بردار کلمات هر جمله به عنوان خروجی این لایه، برای استخراج ویژگی‌های مهم به لایه توجه ارسال می‌شود.

مکانیسم توجه برای تمرکز بر بهترین ویژگی‌ها برای بهبود خلوص ورودی‌های شبکه عصبی LSTM و صرفه‌جویی در زمان پردازش استفاده شده است. این مکانیزم بر روی خروجی لایه تعبیه کار می‌کند و بعد از استخراج بهترین ویژگی‌ها، آنها را به بخش بعدی منتقل می‌نماید. مکانیسم توجه از طریق یک شبکه عصبی عمیق کانولوشنی با یک فیلتر با اندازه ۳ و یک لایه Pooling با اندازه ۲ پیاده‌سازی شده است. تابع ReLU بر روی خروجی اعمال می‌شود و این جریان به شبکه LSTM منتقل می‌گردد.

لایه LSTM بر اساس ماهیت خود به صورت بازگشتی کار می‌کند و بعد از مکانیزم توجه با ۱۰۰ نورون در پایتون پیاده‌سازی شده است. بعد از لایه LSTM، یک لایه کاملاً متصل با تابع Sigmoid وجود دارد که قطبیت جمله را به عنوان منفی یا مثبت نشان می‌دهد. طول فاصله نزدیک‌ترین جمله ذهنی قبلی به جمله فعلی ۱۵ تعیین شده است زیرا حداکثر فاصله دو جمله ذهنی در مجموعه داده در بدترین حالت ۱۵ است. در این بخش مخزن جملات ذهنی (SSC)[1] پیاده‌سازی شد که تمام جملات شناسایی‌شده ذهنی در ماژول A را نگه می‌دارد. به این ترتیب شبکه LSTM برای مشاهده حالت و پردازش سریع‌تر نیاز به بازدید از جمله با شاخص $n-۱$ به عنوان جمله قبلی از SSC را دارد تا جمله را با شاخص n محاسبه کند. SSC در اینجا، نقش دروازه فراموشی[2] در شبکه عصبی LSTM را بازی می‌کند. برای پوشش موارد استثنا، قوانین "اما" نیز در اینجا تعریف شده که در واقع دو جمله ذهنی متوالی با اتصالاتی مانند "اما"، "گرچه" و "به جز" را در نظر می‌گیرد. در این مواقع قطبیت دو جمله در تقابل با یکدیگر هستند مانند "اگرچه اتاق‌های هتل بزرگ بودند اما پارکینگ خیلی کوچک بود". در صورت وجود چنین اتصالاتی در جمله قبلی یا فعلی، با دریافت جملات ذهنی از SSC، خروجی نهایی شبکه LSTM در مقدار ۱- ضرب خواهد شد. داده‌های آموزشی به آزمایشی توزیع ۷۰ به ۳۰ دارند و دو پارامتر loss و Optimizer به ترتیب با مقادیر binary_crossentropy و adam تنظیم شده است. بهترین نتیجه ۹۹/۹۳ در ماژل C در ۴۰ دور به دست آمده است. ساختار این شبکه عصبی عمیق در جدول ۱ ارائه شده و شکل ۳ دقت را در تنظیمات مختلف دورها و اندازه دسته برای این شبکه نشان می‌دهد.

۳-۳-٤ تشخیص قطبیت جملات (عددی)

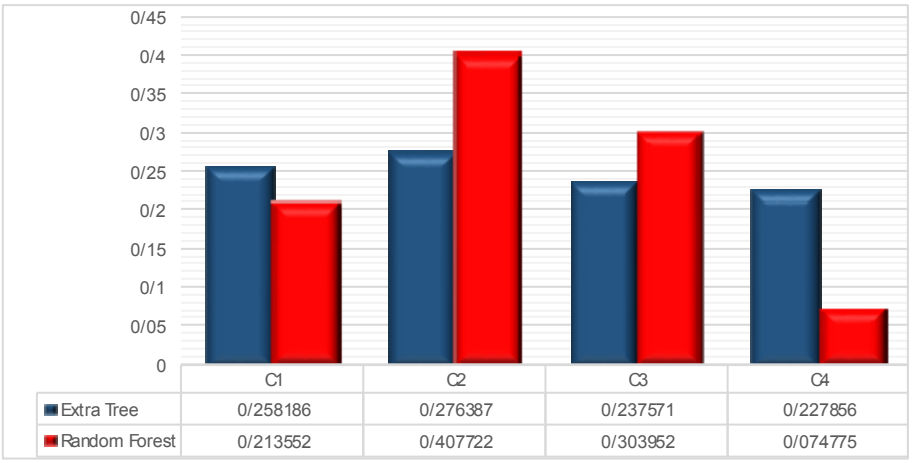
ماژول D به منظور تشخیص قطبیت جملات در حالت تخصیص عددی طراحی شده است. این وظیفه بر اساس دو بخش شامل یک ساختار شبکه عصبی عمیق و یک انباشتگر کار می‌کند. در ادامه هر بخش از ماژول D به صورت مجزا توضیح داده می‌شود.

۳-۳-٤-۱ تعیین قطبیت صفات/ قیود

بخش اول ماژول D وظیفه خود را بر اساس مجموعه طراحی‌شده و پیشنهادی از ویژگی‌های جدول ۴ انجام می‌دهد. این مجموعه ویژگی‌ها بعد از یک مرحله برچسب‌زدن به شبکه عصبی خودرمزنگار عمیق وارد می‌شوند. عملیات برچسب‌زنی برای ویژگی مانند "درجه صفت" با مقدار ۱ و ۲ برای ایجاد تفاوت بین صفاتی مانند "عالی" با "خوب" است و یا برای ویژگی "آیا صفت با توجه به جنبه تغییر می‌کند؟" در صورت مثبت‌بودن پاسخ مقدار "یک" و در صورت منفی‌بودن مقدار "صفر" داده شده است. همه ۲۴۷۴ صفت و قید موجود در لغت‌نامه احساسی بعد از برچسب‌زدن، در این شبکه با توزیع ۷۰ و ۳۰ به عنوان نمونه آموزشی و آزمایشی استفاده شدند.

از آنجا که میزان تأثیرگذاری تعدادی از ویژگی‌ها بعد از اجرای دسته‌بندی کننده Random Forest و Extra Trees در پایتون نسبت به سایر ویژگی‌ها بیشتر است (جدول ۵)، بنابراین به منظور کاهش ابعاد، بررسی واریانس جریان اصلی ویژگی‌ها و تعادل‌بخشیدن به تأثیرات همه

1. Subjective Sentence Container
2. Forget Gate



شکل ۴: اهمیت مجموعه ویژگی‌های کاهش‌یافته بعد از اجرای PCA با دو دسته‌بندی‌کننده Random Forest و Extra Trees در پایتون.

جدول ۵: اهمیت ویژگی‌های شناسایی قطبیت صفات/ قیود با Random Forest و Extra Trees.

Feature	Random Forest	Extra Trees
$f1$	۰٫۷۱۴۱۸۱۶۸	$۹٫۵۵۷۱۵۶۳۱e^{-۰۱}$
$f2$	۰٫۰	$۱٫۶۹۱۰۷۴۵۲e^{-۰۴}$
$f3$	۰٫۰۴۰۸۳۲۰۹	$۸٫۸۱۷۳۷۶۶۶e^{-۰۳}$
$f4$	۰٫۰۳۰۱۱۶۲۲	$۵٫۷۲۸۹۷۷۲۷e^{-۰۳}$
$f5$	۰٫۰۸۳۰۷۶۷۶	$۴٫۷۴۵۹۹۸۳۷e^{-۰۳}$
$f6$	۰٫۰۰۵۰۹۱۷۷	$۱٫۳۲۶۶۷۵۴۲e^{-۰۳}$
$f7$	۰٫۱۲۰۱۷۲۷۲	$۲٫۲۵۴۳۹۶۳۷e^{-۰۳}$
$f8$	۰٫۰۰۱۲۴۵۴۴	$۳٫۲۴۱۹۷۳۱۵e^{-۰۴}$
$f9$	۰٫۰۰۵۲۸۳۳۱	$۶٫۲۸۰۷۲۵۸۸e^{-۰۴}$

آنها بر روی نتیجه، الگوریتم تجزیه و تحلیل مؤلفه‌های اصلی از طریق تنظیم تعداد مؤلفه اصلی روی ۴، بر مجموعه داده پیاده‌سازی شد. شکل ۴ اهمیت مجموعه ویژگی‌های کاهش‌یافته از اجرای دسته‌بندی‌کننده Random Forest و Extra Trees در پایتون را بعد از اجرای PCA نشان می‌دهد. همان طور که مشاهده می‌شود، اهمیت ویژگی‌ها در مجموعه کاهش‌یافته بسیار متعادل‌تر است. مجموعه داده برچسب‌گذاری شده به همراه چهار ویژگی جدید C۱ تا C۴ به عنوان ورودی دوباره به ساختار شبکه عصبی خودرمزنگار عمیق وارد شدند.

برای بهترین نتیجه، شبکه با هفت لایه مخفی کاملاً متصل با ۱۶۰، ۱۲۰، ۸۰، ۶۰، ۴۰، ۲۰ و ۱۰ گره، از طریق تابع فعال‌ساز ReLU و لایه خروجی یک گره، با تابع فعال‌ساز سیگموئید که در پایتون پیاده‌سازی شده است طراحی گردید. دو پارامتر loss و Optimizer به ترتیب با مقادیر binary_crossentropy و adam تنظیم شدند و بهترین نتیجه ۹۴٫۹۲ در ۲۰ دور به دست آمد.

ساختار شبکه عصبی عمیق ماژول D در جدول ۱ ارائه شده است و شکل ۳ دقت را در تنظیمات مختلف دورها و اندازه دسته برای این شبکه نشان می‌دهد.

۳-۳-٤-۲ انباشتگر

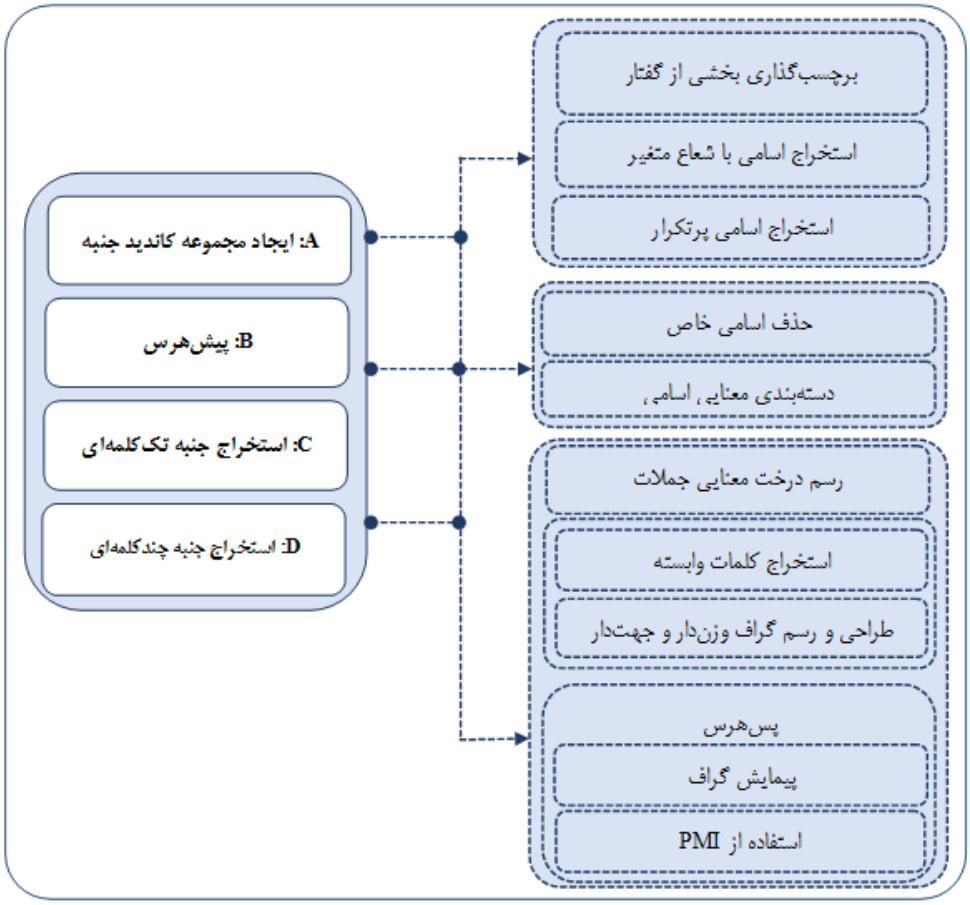
بعد از این که در بخش اول قطبیت صفات/ قیود از طریق دسته‌بندی آنها در دو دسته مثبت یا منفی با شبکه عصبی خودرمزنگار عمیق تشخیص داده شد، در بخش دوم بر اساس تعداد رخداد صفات مثبت یا منفی، به ازای تعداد جنبه‌های تعریف‌شده برای موجودیت (در اینجا هتل) یک مقدار عددی به آنها اختصاص داده می‌شود. به عنوان مثال جمله "رفتار کارکنان هتل خوب بود"، در انتهای ماژول C مثبت داده شده است اما مقدار عددی بعد از اجرای ماژول D، ۱+ خواهد بود زیرا صفت "خوب" برای توصیف جنبه "رفتار کارکنان" استفاده شده است و برای جملاتی مانند "کیفیت غذای رستوران هتل، زمان سرو غذا و تمیزی آن عالی بود" در انتهای ماژول C مثبت داده شده است اما مقدار عددی بعد از اجرای ماژول D، ۳+ خواهد بود زیرا صفت "عالی" برای توصیف سه جنبه "کیفیت غذا"، "زمان سرو" و "تمیزی" استفاده شده است. بنابراین انباشتگر در ماژول D ابتدا جنبه‌ها را استخراج می‌کند و سپس با توجه به ملاحظات دستوری زبان فارسی، به محاسبه و تخصیص قطبیت عددی به جملات در سطح جنبه می‌پردازد.

۱) استخراج جنبه

استخراج جنبه به عنوان یکی از زیرشاخه‌های تحلیل احساس، به دلیل تفاوت‌های زبانی نمی‌تواند از یک زبان به زبان دیگر گسترش یابد و بنابراین در انباشتگر برای تحلیل احساس دقیق جملات، بر استخراج جنبه صریح چندکلمه‌ای در زبان فارسی تمرکز می‌شود. برای استخراج جنبه در انباشتگر، متدولوژی ارائه‌شده در شکل ۵ استفاده شده است [۱۸]. این متدولوژی، چهار مرحله شامل ایجاد مجموعه کاندید جنبه، پیش‌هرس مجموعه کاندید، استخراج جنبه‌های تک‌کلمه‌ای و استخراج جنبه‌های چندکلمه‌ای دارد که در ادامه به توضیح هر مرحله پرداخته می‌شود.

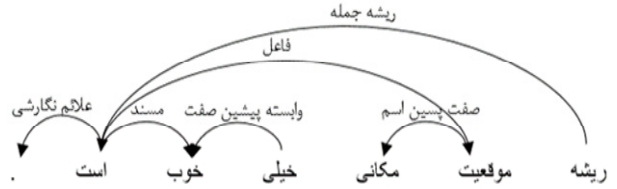
در بخش A، مطابق شکل ۵ به تشکیل مجموعه کاندید جنبه پرداخته می‌شود. در این بخش ابتدا جملات ذهنی خروجی ماژول A با ابزار NLPTools برچسب‌گذاری شدند. از آنجا که در یک جمله فارسی، اسامی با شعاع همسایگی متفاوت نسبت به کلمه احساسی می‌توانند جنبه‌های اصلی در یک موجودیت باشند، بنابراین برای استخراج جنبه‌های کاندید، همه اسامی با شعاع متغیر نسبت به کلمات احساسی حاوی نظر استخراج شدند. برای استخراج اسامی با شعاع متغیر نسبت به کلمات حاوی نظر، یک برنامه در زبان C# توسعه داده شد. از آنجا که جنبه‌ها اسامی هستند که به طور مکرر توسط کاربران در مورد یک موجودیت بیان می‌شوند، بنابراین بعد از استخراج تمامی اسامی، با استفاده از طرح وزنی TF_IDF در نرم‌افزار داده‌کاوی RapidMiner اسامی پرتکرار استخراج شدند.

مجموعه کاندید ایجادشده، شامل اسامی خاص مانند نام‌های تجاری و اسامی مترادف است، به عنوان مثال کلماتی مانند "کارکنان، کارمندان، پرسنل" و یا "وسایل، تجهیزات" که در یک گروه معنایی قرار دارند.



شکل ۵: متدولوژی استخراج جنبه در بخش انباشتگر.

شکل ۶: گراف وابستگی جمله نمونه با یک عبارت اسمی به عنوان جنبه چندکلمه‌ای با ابزار HAZM [۱۹].



شکل ۷: گراف وابستگی جمله نمونه با یک عبارت غیر اسمی به عنوان جنبه چندکلمه‌ای با ابزار HAZM [۱۹].

بنابراین در بخش "B" مطابق شکل ۵ پیش‌هرس مجموعه کاندید جنبه انجام شد که این فرایند شامل حذف نام‌های تجاری و گروه‌بندی معنایی کلمات مترادف است. برای گروه‌بندی معنایی اسامی از بخش مربوط به اسامی حوزه هتل‌داری لغت‌نامه احساسی استفاده شده است. اعضای مجموعه باقی‌مانده، مطابق بخش C از شکل ۵، مجموعه جنبه‌های تک‌کلمه‌ای را تشکیل می‌دهند. تعدادی از جنبه‌های تک‌کلمه‌ای استخراج‌شده عبارت است از: "معماری"، "استخر"، "قیمت"، "پارکینگ" و غیره. در مرحله بعد، هدف کشف جنبه‌های چندکلمه‌ای مانند "عایق صوتی"، "تمیزی اتاق" و "کیفیت رستوران" است. این کار با توجه به پیچیدگی‌های زبان فارسی بسیار چالش‌برانگیز خواهد بود و بنابراین در بخش D از شکل ۵ از سه مرحله استفاده شده است: ۱) استفاده از گراف وابستگی جملات، ۲) غلبه بر ساختارهای مشکل‌ساز و ۳) پس‌هرس که در ادامه بررسی خواهند شد.

از گراف وابستگی برای استخراج جنبه‌های چندکلمه‌ای که عبارات اسمی هستند و در مجموعه کاندیدهای اولیه وجود دارند استفاده شد. برای رسم گراف وابستگی، ابزار HAZM که مخصوص تجزیه جملات فارسی است مورد استفاده قرار گرفت [۱۹]. این ابزار، وابستگی کلمات و ساختار نحوی جمله را تعیین می‌کند. با توجه به گراف وابستگی می‌توان تشخیص داد که اگر کلمه بعد از یک اسم، نقش مضاف‌الیهی داشته باشد با ایجاد یک عبارت اسمی (ترکیب "اسم + اسم + اسم")، مانند شکل ۶ جنبه‌های چندکلمه‌ای ایجاد می‌شود. اما چالش اصلی این است که در ساختار زبان فارسی، عبارات بسیاری وجود دارند که می‌توانند جنبه چندکلمه‌ای باشند اما در قالب یک ساختار چندکلمه‌ای اسمی قرار ندارند. به عنوان مثال در جمله "موقعیت مکانی خیلی خوب است"، کلمه "موقعیت" در فارسی شامل دو کلمه "موقعیت مکانی" است و کلمه "مکانی" در زبان فارسی صفت است. گراف وابستگی یک نمونه جمله فارسی در قالب دو کلمه ذکرگردیده در شکل ۷ نشان داده شده که در آن نمی‌توان ترکیب "موقعیت مکانی" را به عنوان یک جنبه چندکلمه‌ای به دست آورد زیرا "مکانی" نقش صفت دارد. در نتیجه نمی‌توان به گراف وابستگی به عنوان ابزاری برای غلبه بر ساختارهای مشکل‌دار اعتماد کرد.

بنابراین برای غلبه بر چالش ترکیبات مشکل‌دار، ابتدا همه ترکیبات که می‌توانند به عنوان جنبه‌های چندکلمه‌ای دسته‌بندی شوند در فارسی در نظر گرفته می‌شود. این ترکیبات که در جدول ۶ لیست شده است توسط محمدی و همکاران جمع‌آوری و جمع‌بندی شده‌اند [۱۹].

جدول ۶: تعدادی از چالش‌های زبانی در جملات فارسی برای فرایند استخراج جنبه [۲۰].

ردیف	چالش	ترکیب	مثال
۱	صفت بعد از اسم اتفاق بیفتد (در ساختار فارسی) و یک اسم مرکب ایجاد کند.	اسم + صفت	رستوران سنتی هتل عالی بود.
۲	انتساب متوالی اسم در یک جمله	اسم + اسم (+ اسم)	رفتار کارکنان خوب بود.
۳	در زبان عامیانه فارسی، کلمه احساسی (قید یا صفت) بین دو اسم آمده است.	اسم + صفت + اسم	رفتار خوب کارکنان قابل تحسین بود.
۴	کلماتی که ریشه اصلی آنها صفت است اما نقش فاعل یا اسم دارند.	صفت + اسم (فرم اول)	زیبایی محوطه خوب بود.
۵	بعضی اوقات صفت با چسبیدن به اسم، صفت جدیدی را تشکیل می‌دهد.	صفت + اسم (فرم دوم)	کارکنان هتل پرحرف بودند.
۶	کاربران می‌خواهند از نسخه کوتاه‌شده استفاده کنند.	–	رستوران هتل عالی بود (کیفیت رستوران هتل عالی بود).

جدول ۷: نمونه کد CYPHER برای بازیابی جنبه‌های صریح نوع A، B و C در گراف ADG.

Compound Aspect Type	Cypher
Noun + Adjective (Type A)	MATCH p = (a) -[]- > (b) -[]- > (c) -[*1..10]- > (d: 'Verb') WHERE ANY (x IN nodes (p) WHERE (a: 'Noun') -- > (b: 'Adjective') AND EXISTS ((b: 'Adjective') -[*1..3]- > (c: 'Adjective'))) RETURN a.name, b.name;
Noun + Noun (Type B)	MATCH p = (a) -[]- > (b) -[]- > (c) -[*1..10]- > (d: 'Verb') WHERE ANY (x IN nodes(p) WHERE (a: 'Noun') -- > (b: 'Noun') -- > (c: 'Adjective')) RETURN a.name, b.name;
Noun + Adjective + Noun (Type C)	MATCH p = (a) -[]- > (b) -[]- > (c) -[*1..10]- > (d: 'Verb') WHERE ANY (x IN nodes (p) WHERE (a: 'Noun') -- > (b: 'Adjective') -- > (c: 'Noun')) RETURN a.name, c.name;

همان طور که در سه ردیف اول جدول ۶ نشان داده شده است (سطر شماره چهار یک ترکیب "اسم + اسم" و سطر شماره ۵ نشان‌دهنده یک صفت مرکب است)، سه ترکیب "اسم + صفت"، "اسم + اسم (+ اسم)" و "اسم + صفت + اسم" می‌توانند جنبه‌های چندکلمه‌ای در زبان فارسی باشند. در ادامه با تمرکز بر ترکیبات در سه قالب نام برده شده، یک گراف جهت‌دار با نام ADG پیشنهاد شده است که به عنوان گراف تشخیص جنبه معرفی می‌شود. از اطلاعات خروجی الگوریتم FP-Growth در برخی قسمت‌های ساختار ADG مانند لبه‌ها و وزن استفاده شده است. در این بخش به سه دلیل از الگوریتم FP-Growth استفاده می‌شود: ۱) سرعت قابل توجه آن نسبت به الگوریتم‌های دیگر استخراج‌کننده الگوهای مکرر (افزایش چشم‌گیر سرعت این الگوریتم به دلیل کاهش قابل توجه تعداد دفعات مراجعه به تراکنش‌های موجود در کاوش کلمات وابسته است)، ۲) استخراج کلمات وابسته و ۳) استفاده از خروجی این الگوریتم برای تشکیل پایگاه داده گرافیکی و ایجاد مدل بر اساس زبان جستجوی Cypher.

در این بخش با کشف الگوهای مکرر، مجموعه‌ای از قوانین وابستگی[1] کشف می‌شود که بر اساس آن قوانین، استناد شود که اگر کلمه x اتفاق افتاد آن گاه کلمه y نیز اتفاق خواهد افتاد. به عبارت دیگر با اجرای الگوریتم FP-Growth به دنبال این هستیم که بر اساس قوانین استخراج‌شده، بدانیم وجود چه کلماتی بر وجود مجموعه کلمات دیگر مؤثر است. خروجی مهم در این روش، مجموعه‌ای از قوانین اگر- آن گاه است که ارتباطات میان رخداد هم‌زمان مجموعه‌ای از کلمات با یکدیگر را آشکار می‌کند. مقادیر پارامترها برای support = ۱، min item set = ۲ و max item set = ۳ است، زیرا طولانی‌ترین جنبه چندکلمه‌ای که با آن روبه‌رو هستیم شامل سه کلمه خواهد بود. شکل ۸ محدود به قوانینی است که شامل ترکیبات دو اسم "کارکنان" و "رفتار" است.

بعد از استخراج قوانین مربوط به وابستگی کلمات، به رسم گراف ADG پرداخته می‌شود. در گراف ADG (V,E)، نودهای V رأس‌هایی هستند که کلماتی با برچسب‌های مختلف POS دارند که در مجموعه داده‌های مقاله ظاهر شده‌اند و بر طبق برچسب‌های دسته‌بندی شده POS، در دسته‌های مختلف نشان داده می‌شوند. به عنوان مثال، همه برچسب‌هایی که به اسامی مانند NN، NNP، NNS و NNPS اشاره می‌کنند در کلاس دیگری قرار گرفتند و همه برچسب‌هایی که به صفات با برچسب JJ، JJR و JJS اشاره می‌کنند در یک کلاس قرار دارند و غیره. برای نشان‌دادن این گروه‌های مختلف اشیا در Neo۴J از رنگ‌های مختلفی استفاده شده است. همچنین E به عنوان یال‌ها به دنباله‌ای از کلمات در یک جمله اشاره دارد. وزن یال‌ها نیز با توجه به تعداد تکرار دو گره با هم در پیکره استخراج‌شده از الگوریتم FP-Growth اختصاص داده شد. از آنجا که افعال در یک جمله فارسی در انتهای جملات ظاهر می‌شوند، بنابراین بازدید از یک گره با کلاس فعل نشان‌دهنده رسیدن به انتهای جمله در گراف ADG است. برای سادگی، مسیری با طول حداکثر ۱۰ یال به عنوان فاصله بین انتهای ترکیبات جنبه و انتهای هر جمله در کدهای Cypher در نظر گرفته شد. ساختار ADG در پایگاه داده NoSQL مبتنی بر گراف Neo۴J پیاده‌سازی شده است. بخشی از گراف ADG در سه ترکیب به گونه‌ای نشان داده شده که شکل ۹ نشان‌دهنده گراف ADG برای ترکیبات "اسم + صفت + اسم" مانند "برخورد عالی کارکنان" است.

جدول ۷ نیز شامل کدهای نمونه Cypher به عنوان زبان جستجوی Neo۴J است که برای بازیابی سه نوع جنبه ترکیبی A، B و C از ساختار گراف ADG در این بخش پیاده‌سازی شده است.

همان طور که گفته شد تمام کلمات وابسته که نتیجه پیاده‌سازی با الگوریتم FP-Growth بودند به همراه جنبه‌های کاندیدای از پیش‌هرس شده به ساختار گراف ADG اضافه شدند. سپس هر یک از ترکیبات با قوانینی که در جدول ۸ نشان داده شده است مقایسه شدند تا حضور یا حذف آنها از لیست نهایی جنبه تأیید شود. بنابراین، این قوانین به عنوان ابزار پس‌هرس برای جنبه‌های چندکلمه‌ای تدوین شده‌اند. همان طور که در جدول ۸ آمده است در صورت رعایت شرایط شرح داده شده، جنبه‌هایی که شناسایی شدند در لیست جنبه‌ها باقی می‌مانند (به عنوان مثال سطرهای شماره ۱، ۲ و ۳) و در غیر این صورت حذف می‌شوند (به عنوان مثال سطر شماره ۴). به عنوان یکی دیگر از اقدامات کنترلی و پس‌هرس،

1. Association Rules

(۰/۸۴۸: اطمینان) [کارکنان] < -- [عالی، بسیار]
(۰/۸۴۸: اطمینان) [کارکنان، برخورد] < -- [عالی، بسیار]
(۰/۸۴۸: اطمینان) [کارکنان] < -- [عالی، بسیار، برخورد]
(۰/۸۹۶: اطمینان) [کارکنان] < -- [بسیار]
(۰/۸۹۶: اطمینان) [کارکنان، برخورد] < -- [بسیار]
(۰/۸۹۶: اطمینان) [کارکنان] < -- [بسیار، برخورد]
(۰/۹۰۰: اطمینان) [کارکنان] < -- [برخورد]
(۰/۹۲۴: اطمینان) [کارکنان] < -- [مؤدبانه]
(۰/۹۲۴: اطمینان) [کارکنان، برخورد] < -- [مؤدبانه]
(۰/۹۲۴: اطمینان) [کارکنان] < -- [مؤدبانه، برخورد]
(۰/۹۳۳: اطمینان) [کارکنان] < -- [مؤدبانه، بسیار]

شکل ۸: قوانین خروجی الگوریتم FP-Growth (محدود به قوانین ترکیبات دو اسم "کارکنان" و "برخورد").

فرمول PMI برای بررسی وقوع متقابل ترکیبات مشتق‌شده استفاده گردیده و ترکیبات با امتیاز کم حذف شدند. برای حذف نمرات منفی، آستانه PMI روی صفر تنظیم شد. تعدادی از جنبه‌های مشتق‌شده تک‌کلمه‌ای و چندکلمه‌ای در حوزه هتل‌داری از انباشتگر در جدول ۹ ذکر شده است [۱۸].

۲) قوانین دستوری زبان

برای تحلیل احساس و استخراج صحیح قطبیت عددی جملات، توجه به قوانین دستوری هر زبان از اهمیت بالایی برخوردار است. بنابراین در انباشتگر بعد از استخراج جنبه، به دو قانون "و/ ویرگول" و قانون "تشدیدکننده" به عنوان ملاحظات دستوری زبان فارسی توجه شده است.

• قانون "و/ ویرگول"

بعد از استخراج جنبه، قانون "و/ ویرگول" با شرایط زیر تعریف خواهد شد:

- اگر تعداد رخداد کلمات احساسی و جنبه‌ها یکسان باشد آن گاه قطبیت عددی جمله، همان امتیاز کلمه احساسی است. به عنوان مثال در جمله "تمیزی اتاق خوب بود"، کلمه احساسی "خوب" به جنبه "تمیزی اتاق" با مقدار عددی ۱+ اختصاص می‌یابد.
- اگر برای هر جنبه، رخداد کلمات احساسی بیش از یک کلمه باشد آن گاه محاسبه عددی قطبیت با توجه به مجموع امتیاز کلمه احساسی محاسبه می‌شود. به عنوان مثال در جمله "دمای اتاق نامناسب و گرم بود"، دو کلمه احساسی "نامناسب" و "گرم" به جنبه "دمای اتاق" با مقدار عددی ۲- اختصاص می‌یابد.
- اگر جمله شامل بیش از یک جنبه با عبارت ربطی "و/ ویرگول" باشد، آن گاه ارزش عدی صفت موجود در جمله به تعداد جنبه‌های مربوط به آن ضرب خواهد شد. به عنوان مثال در "برخورد کارکنان، تمیزی اتاق‌ها و اینترنت عالی بود"، قطبیت عددی، ۳+ خواهد بود زیرا صفت "عالی" برای توصیف سه جنبه "برخورد کارکنان"، "تمیزی اتاق‌ها" و " اینترنت" استفاده شده است.

• قانون "تشدیدکننده"

در عبارات حاوی قیود تشدیدکننده مانند "بسیار، خیلی، فوق‌العاده، بی‌نهایت، بی‌اندازه" علاوه بر صفات استفاده‌شده در جمله، به فعل جمله نیز توجه می‌شود به طوری که به تناسب فعل مثبت یا منفی، قانون "تشدیدکننده" به صورت زیر تعریف خواهد شد:

- اگر فعل مثبت بود آن گاه به تناسب صفت استفاده‌شده در جمله، عدد ۱+ یا ۱- به ارزش صفت اضافه می‌شود و قطبیت نهایی جمله، مجموع ارزش صفت و ارزش قید (به تناسب صفت) است. به عنوان مثال در جمله "بسیار خوب بود"، به دلیل مثبت‌بودن فعل، عدد ۱+ به مقدار ارزش خوب یعنی ۱+ اضافه شده و مقدار نهایی قطبیت جمله ۲+ خواهد بود. در حالی که برای همین قید در عبارت "بسیار بد بود" مقدار ۲- محاسبه خواهد شد.
- اگر فعل منفی بود آن گاه قید تشدیدکننده نادیده گرفته می‌شود و قطبیت نهایی جمله، فقط به ارزش صفت وابسته است. به عنوان مثال در جمله "خیلی بد نبود" به دلیل منفی‌بودن فعل جمله، قید خیلی در نظر گرفته نمی‌شود و قطبیت نهایی جمله برابر ارزش صفت یعنی ۱- است.

در قسمت اول ماژول D، قطبیت صفات/ قیود به دست می‌آید و قسمت دوم یعنی انباشتگر، وظیفه استخراج جنبه، اعمال قانون "و/ ویرگول " و قانون "تشدیدکننده" را بر عهده دارد.

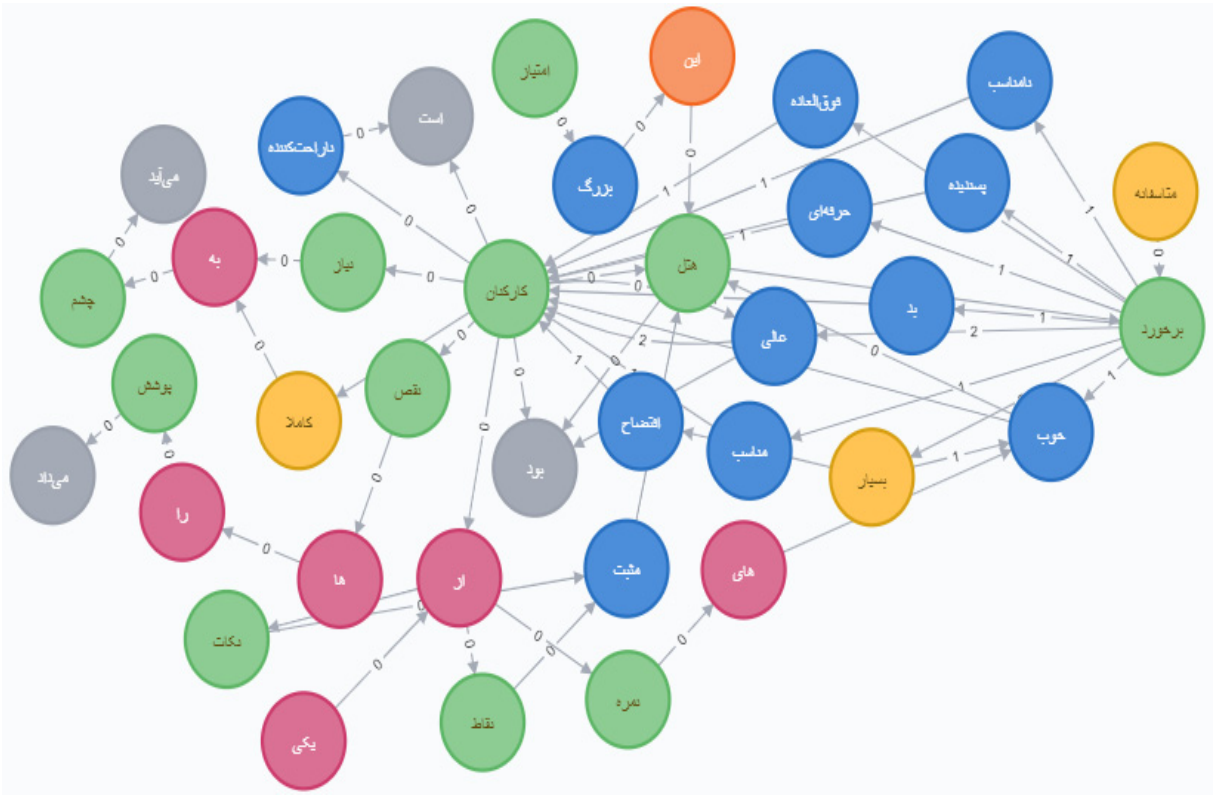
٤- مقایسه و ارزیابی

برای ارزیابی چارچوب RSAD، در بخش تشخیص قطبیت جملات در حالت باینری در ماژول C، از چهار ساختار شبکه عصبی عمیق بازگشتی استفاده شده است. این چهار روش شامل شبکه LSTM به تنهایی، LSTM دولایه[1]، LSTM دوسویه[2] و شبکه GRU[3] است که در تشخیص قطبیت جمله در حالت باینری در پژوهش‌های متعددی مانند [۲۱] تا [۲۳] استفاده شده‌اند. همان طور که در شکل ۱۰ نشان داده شد دقت عملکرد RSAD از چهار شبکه عصبی عمیق دیگر در تشخیص قطبیت جمله بهتر است. معتقدیم طبیعت بازگشتی RSDA در کنار توجه به روان‌شناسی کاربر تأثیر زیادی در نتیجه داشته است.

مرحله بعد مقایسه دقت شبکه RSAD در خروجی قسمت تشخیص قطبیت جملات در حالت عددی، در بخش صفات/ قیود در ماژول D است. در این مقایسه از روش k-نزدیک‌ترین همسایه و روش ترکیبی شامل ترکیب سه الگوریتم ماشین بردار پشتیبان، رگرسیون لجستیک و شبکه عصبی استفاده شده است. این روش‌ها در پژوهش‌های متعددی از جمله پژوهش دهخارگانی و همکاران (۲۰۱۶) بهترین نتیجه را تولید کرده است [۲۴]. مقایسه دقت شبکه RSAD در خروجی ماژول D با این روش‌ها در شکل ۱۱ آمده است.

برای ارزیابی استخراج جنبه تک و چندکلمه‌ای در انباشتگر از ماژول D نیز ابتدا از الگوریتم‌های پایه مبتنی بر تکرار، مبتنی بر برچسب‌زنی بخشی از گفتار و مبتنی بر تخصیص دریکله پنهان (LDA)[4] در مجموعه داده مقاله حاضر برای استخراج جنبه‌های تک‌کلمه‌ای استفاده شد. الگوریتم مدل‌سازی موضوعی LDA خوشه‌های کلمات را به گونه‌ای تولید می‌کند که هر خوشه برابر با یک موضوع باشد [۲۵] تا [۲۸]. رویکرد دیگر برای استخراج جنبه، روش مبتنی بر برچسب‌زنی بخشی از گفتار است که توسط هو و لیو (۲۰۰۴) و بیلر و گلدنسان (۲۰۰۸) برای استخراج جنبه‌های نادر مورد استفاده قرار گرفت، به طوری که نزدیک‌ترین اسم(های) به کلمات احساسی و حاوی نظر به عنوان جنبه‌ها استخراج شدند [۲۹] و [۳۰]. این

1. 2_Layer LSTM
2. Bi-Directional Long Short-Term Memory
3. Gated Recurrent Unit
4. Latent Direchlet Allocation



شکل ۹: گراف ADG خروجی Neo۴J برای ترکیبات "اسم + صفت + اسم" (جنبه مرکب نوع C).

جدول ۸: نمونه قوانین برای استخراج جنبه‌های صریح بر اساس کدهای CYPHER.

ترکیب	شرط	ترکیب نهایی (جنبه چندکلمه‌ای)	مثال
اسم + صفت	باقی می‌ماند/ اگر بعد از ترکیب، صفتی برای توصیف آن وجود داشته باشد، این عبارت در لیست جنبه‌ها باقی می‌ماند.	اسم مرکب (جنبه) (نوع A)	رستوران سنتی عالی بود.
اسم + اسم	باقی می‌ماند/ اگر صفتی بعد از ترکیب برای توصیف آن وجود داشته باشد، در لیست جنبه‌ها باقی می‌ماند.	اسم مرکب (جنبه) (نوع B)	تمیزی اتاق‌ها عالی بود.
اسم + صفت + اسم	باقی می‌ماند/ اگر ترکیبی از اسم + اسم در جنبه نامزد چندکلمه‌ای وجود داشته باشد، در لیست جنبه‌ها باقی می‌ماند.	جنبه (با صفت داخلی) (نوع C)	رفتار عالی کارکنان یکی از برتری‌هایش است.
صفت + اسم	حذف/ از لیست جنبه‌ها حذف و به لغت‌نامه اضافه می‌شود.	صفت جدید	فضای هتل کم‌نور بود.

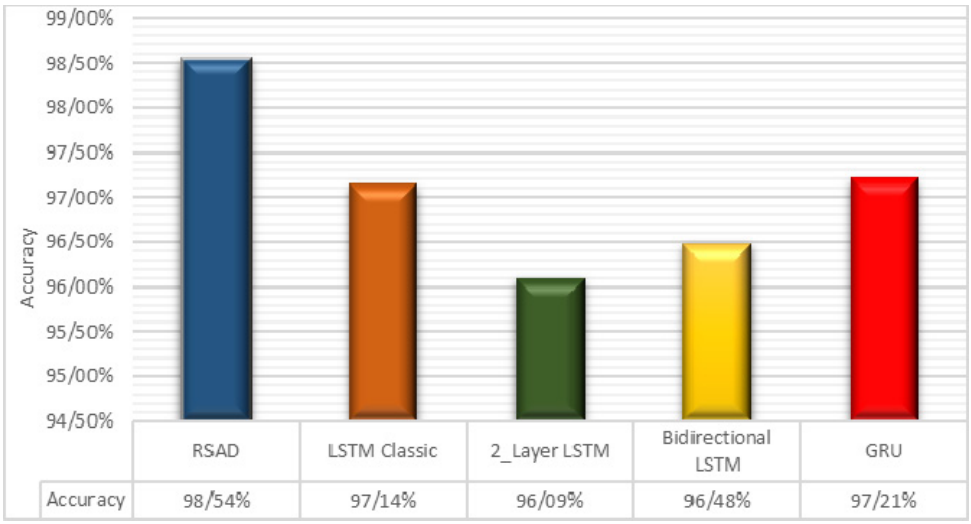
جدول ۹: تعدادی از جنبه‌های صریح استخراج‌شده از روش پیشنهادی در انباشتگر.

ردیف	جنبه
۱	قیمت
۲	اینترنت
۳	عایق صوتی
۴	دمای اتاق
۵	پارکینگ
۶	برخورد کارکنان
۷	دکوراسیون
۸	کیفیت رستوران و کافی‌شاپ
۹	تمیزی اتاق‌ها
۱۰	تمیزی محوطه
۱۱	تجهیزات
۱۲	معماری
۱۳	سرویس بهداشتی
۱۴	موقعیت مکانی و دسترسی

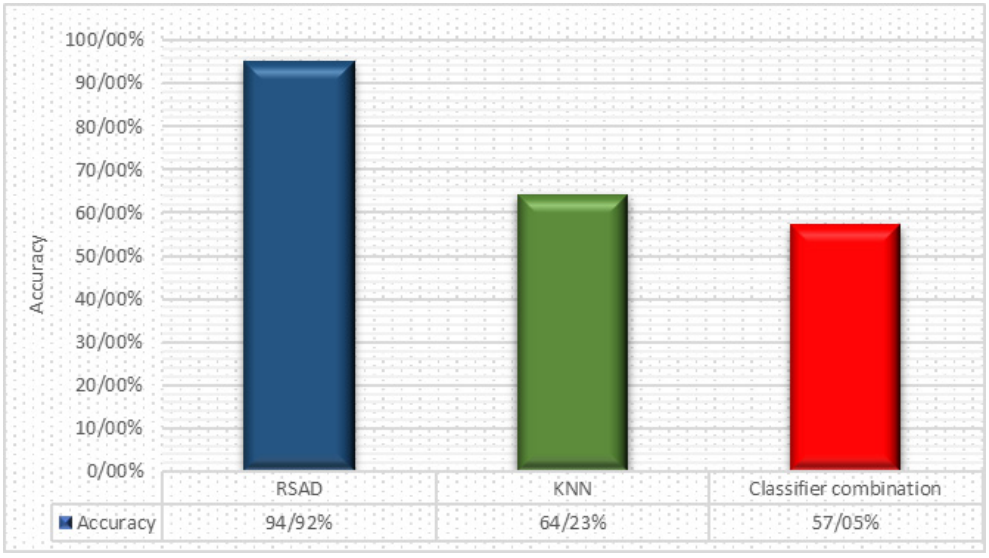
روش نیز روی مجموعه داده مقاله، پیاده‌سازی شد. الگوریتم دیگری که مورد استفاده قرار گرفت بر اساس کلمات پرتکرار با تنظیم آستانه روی ۵۰ برای تعداد کلمات پرتکرار به روشی بود که توسط کو و همکاران (۲۰۰۶) در سطح پاراگراف و سند مطابق با طرح وزنی TF_IDF استفاده شده است [۳۱]. همچنین سرانجام جنبه‌ها بر اساس قضاوت انسانی نیز استخراج شدند و در مجموع با ۵ روش استخراج جنبه، مجموعه خروجی به دست آمد.

شکل ۱۲، مقایسه روش استخراج جنبه را با سه روش مبتنی بر تکرار، مبتنی بر برچسب‌زنی بخشی از گفتار و LDA از نظر F_measure در جنبه‌های تک‌کلمه‌ای نشان داده است. قابل مشاهده است که روش مقاله حاضر به همراه رویکرد مبتنی بر تکرار با توجه به تعداد جنبه‌های استخراج‌شده و مفاهیم دقیق آنها، بهترین نتیجه را دارد. علت این تفاوت در روش مبتنی بر برچسب‌زنی بخشی از گفتار این است که در این روش فقط نزدیک‌ترین اسم به کلمه احساسی استخراج می‌شود، در حالی که همان طور که قبلاً بحث شد در ساختار جملات فارسی، اسامی نسبت به کلمات حاوی نظر شعاع‌های متغیر دارند. همچنین مدل‌سازی موضوعی LDA قادر به تمایز بین کلمات احساسی و جنبه‌ها نیست و هر دو کلاس کلمات را در یک دسته‌بندی قرار می‌دهد.

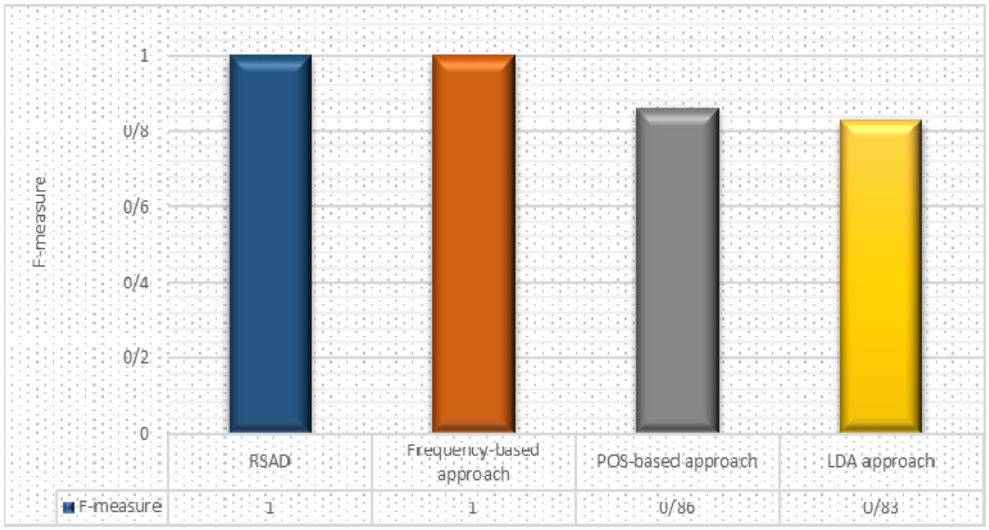
در نهایت آخرین ارزیابی روی استخراج جنبه‌های چندکلمه‌ای در بخش انباشتگر انجام گرفت. مقایسه روش پیشنهادی در استخراج جنبه



شکل ۱۰: مقایسه دقت RSAD با چهار شبکه عصبی عمیق مختلف در تشخیص قطبیت جملات (در حالت باینری).



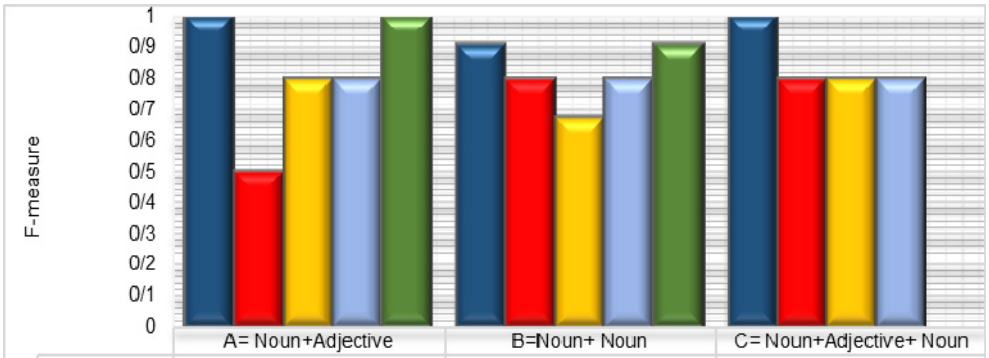
شکل ۱۱: مقایسه دقت RSAD در تحلیل احساس عددی با روش k– نزدیک‌ترین همسایه و روش ترکیبی (ماشین بردار پشتیبان، رگرسیون لجستیک و شبکه عصبی).



شکل ۱۲: مقایسه دقت RSAD در بخش انباشتگر برای استخراج جنبه‌های تک‌کلمه‌ای.

چندکلمه‌ای با ۴ روش LRT-based، SAM، ELDA و MMI-based در سه ترکیب چالش‌دار زبان فارسی در شکل ۱۳ نشان داده شده که به ترتیب در پژوهش‌های [۳۲] تا [۳۴] معرفی شده‌اند.

همان طور که مشاهده می‌شود، نتایج روش پیشنهادی از چهار روش دیگر در تشخیص جنبه‌های چندکلمه‌ای نوع C پیشی گرفته است. در تشخیص جنبه‌های چندکلمه‌ای نوع A و B برتری استخراج جنبه چندکلمه‌ای با انباشتگر برابر با روش مبتنی بر MMI است در حالی که از روش‌های مبتنی بر LRT، SAM و ELDA پیشی می‌گیرد. روش مبتنی



	A= Noun+Adjective	B=Noun+ Noun	C= Noun+Adjective+ Noun
RSAD	1	0/91	1
LRT based	0/5	0/8	0/8
SAM	0/8	0/67	0/8
ELDA	0/8	0/8	0/8
MMI_based	1	0/91	0

شکل ۱۳: مقایسه دقت RSAD در بخش انباشتگر برای استخراج جنبه‌های چندکلمه‌ای.

بر MMI فقط جنبه‌های دوکلمه‌ای را به صورت ترکیبات "اسم + صفت" و "اسم + اسم" تشخیص می‌دهد و هیچ استراتژی برای تشخیص ترکیب سوم یعنی "اسم + صفت + اسم" ندارد. روش SAM از روش LDA که قبلاً توضیح داده شده است برای استخراج جنبه‌های تک‌کلمه‌ای استفاده می‌کند و سپس با استفاده از زنجیره مارکوف، ترتیب کلمات، هم‌زمانی و تکرار کلمات را در نظر می‌گیرد. از این رو نتایج خوبی در تشخیص جنبه‌های نوع A و C دارد در حالی که نتایج آن در تشخیص جنبه‌های نوع B رضایت‌بخش نیست زیرا هم‌زمانی احساسات و جنبه‌ها را با هم در قالب "اسم + صفت" بیشتر از ترکیب "اسم + اسم" در نظر می‌گیرد. روش ELDA نتایج خوبی در شناسایی هر سه نوع دارد زیرا این روش ابتدا الگوریتم LDA را روی مجموعه داده‌ها اعمال می‌کند و سپس بر اساس سه ویژگی همراهی با کلمات جنبه، عدم تعلق به کلمات اصلی جنبه و عدم تعلق یک جنبه به جنبه‌های اصلی دیگر، هاب‌ها را استخراج می‌کند. پل‌ها را نیز به عنوان کلماتی که با سایر کلاس کلمات می‌توانند همراه شوند از لیست جنبه‌ها حذف می‌کند. ELDA بعد از یافتن هاب‌ها و استخراج قوانین از آنها، دانشی استخراج می‌کند که با آن، خروجی LDA را بهبود می‌دهند. روش ELDA و SAM از نظر نویسندگان خاص دامنه نیستند و می‌توانند به دو زبان انگلیسی و فارسی گسترش یابند و از این رو آنها بر استخراج ترکیبات مشکل‌دار در زبان فارسی متمرکز نیستند.

۵- نتیجه‌گیری

با توجه به ماهیت تأثیرگذار احساسات بیان‌شده مصرف‌کنندگان نسبت به تصمیم خرید مشتریان احتمالی، تحلیل احساس، زمینه‌ای است که برای تعیین پویای قطبیت نظرات و تشخیص اهداف واقعی مشتریان قبلی به رویکردهای هوشمندانه نیاز دارد. بنابراین در این مقاله، عملکرد چارچوب RSAD با افزودن نظرکاوی و تحلیل احساس سطح جنبه بهبود داده شد. چارچوب RSAD شامل چندین شبکه عصبی مبتنی بر یادگیری عمیق است و چالش‌های موجود در حوزه تحلیل احساس را برای رسیدن به هدف تشخیص قطبیت جمله در دو حالت باینری و عددی حل می‌نمایند. به عنوان مثال RSAD چالش‌هایی مانند تشخیص هرزنظر، تعیین قطبیت کلمات باردار حاوی نظر، استخراج جنبه و ملاحظات مربوط به قوانین زبانی و ... را پوشش داده است. مقایسه و ارزیابی RSAD با پژوهش‌های موجود در تحلیل احساس، نشان‌دهنده استحکام و قدرت چارچوب پیشنهادی است.

برای توسعه این پژوهش مواردی مانند توجه به افعال احساسی حاوی نظر در تشخیص قطبیت جملات، استفاده از روش‌های مبتنی بر یادگیری ماشینی و عمیق و ترکیب آنها با الگوریتم‌های هوش جمعی، توسعه مجموعه ویژگی‌های اولیه در تشخیص اسپم، توسعه مجموعه ویژگی‌های اولیه در تعیین قطبیت کلمات حاوی نظر، توسعه RSAD به زبان‌های دیگر و توسعه ملاحظات دستوری زبان فارسی، مد نظر است.

٦- تقدیر و تشکر

این مقاله مستخرج از رساله دکتری تخصصی نویسنده اول (سپیده جمشیدی‌نژاد) در دانشگاه آزاد اسلامی واحد رشت می‌باشد.

مراجع

[1] B. Sabeti, P. Hosseini, G. Ghassem-Sani, and S. A. Mirroshandel, *LexiPers: An Ontology Based Sentiment Lexicon for Persian*. arXiv preprint arXiv:1911.05263, 2019.

[2] E. S. Tellez, *et al.*, "A simple approach to multilingual polarity classification in Twitter," *Pattern Recognition Letters*, vol. 94, pp. 68-74, 15 Jul. 2017.

[3] R. Dehkharghani, "Building phrase polarity lexicons for sentiment analysis," *Int. J. Interact. Multim. Artif. Intell.*, vol. 5, no. 3, pp. 98-105, 2018.

[4] S. Al-Azani and E. S. M. El-Alfy, "Hybrid deep learning for sentiment polarity determination of arabic microblogs," in *Proc. Int. Conf. on Neural Information Processing,* pp. 491-500, Guangzhou, China, 14-18 Nov. 2017.

[5] 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*, vol. 380, pp. 1-10, 7 Mar. 2020.

[6] Y. Chandra and A. Jana, "Sentiment analysis using machine learning and deep learning," in *Proc. IEEE 7th Int. Conf. on Computing for Sustainable Global Development*, 4 pp., New Delhi, India, 12-14 Mar. 2020.

[7] S. Chen, C. Peng, L. Cai, and L. Guo, "A deep neural network model for target-based sentiment analysis," in *Proc. IEEE Int Joint Conf. on Neural Networks*, 7 pp., Rio de Janeiro, Brazil, 8-13Jul. 2018.

[8] M. El-Masri, N. Altrabsheh, H. Mansour, and A. Ramsay, "A web-based tool for Arabic sentiment analysis," *Procedia Computer Science*, vol. 117, pp. 38-45, 2017.

[9] M. Zhang, "E-commerce comment sentiment classification based on deep learning," in *Proc. IEEE 5th Int. Conf. on Cloud Computing and Big Data Analytics*, pp. 184-187, Chengdu, China, 10-13 Apr. 2020.

[10] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1-167, 2012.

[11] E. Asgarian, A. Saeedi, B. Stiri, and H. Ghaemi, *NLPTools* [Online]. Available: https://wtlab.um.ac.ir, 2016.

[12] M. Hu and B. Liu, "Mining opinion features in customer reviews," *AAAI*, vol. 4, no. 4, pp. 755-760, Jul. 2004.
[13] A. Hassan and A. Mahmood, "Deep learning approach for sentiment analysis of short texts," in *Proc. IEEE 3rd Int. Conf. on Control, Automation and Robotics*pp. 705-710, Nagoya, Japan, 24-26 Apr. 2017.
[14] https://github.com/ICTRC/Parsivar
[15] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proc. of the Int. Conf. on Web Search and Data Mining*, pp. 219-230, Palo Alto, CA, USA 11-12 Feb. 2008.
[16] F. H. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," in *Proc. 22nd Int. Joint Conf. on Artificial Intelligence*, pp. 2488-2493, Barcelona, Spain, 16–22 Jul. 2011.
[17] M. E. Basiri, N. Safarian, and H. K. Farsani, "A supervised framework for review spam detection in the Persian language," in *Proc. IEEE 5th Int. Conf. on Web Research*, pp. 203-207, Tehran, Iran, 24-25 Apr. 2019.
[18] S. Jamshidi-Nejad, F. Ahmadi-Abkenari, and P. Bayat, "A combination of frequent pattern mining and graph traversal approaches for aspect elicitation in customer reviews," *IEEE Access*, vol. 8, pp. 151908-151925, 2020.
[19] M. Khalash and M. Imany, "Persian Language Processing Tool," http://www.sobhe.ir/hazm, 2013.
[20] A. Mohammadi, M. R. Pajoohan, M. Montazeri, and M. Nematbakhsh, "Identifying explicit features of Persian comments," *J. of Computing and Security*, vol. 6, no. 1, pp. 1-11, Winter/ Spring 2019.
[21] G. Jain, M. Sharma, and B. Agarwal, "Spam detection in social media using convolutional and long short term memory neural network," *Annals of Mathematics and Artificial Intelligence*, vol. 85, pp. 21-44, 2019.
[22] H. Nguyen and K. Shirai, "A joint model of term extraction and polarity classification for aspect-based sentiment analysis," in *Proc. IEEE 10th In. Conf. on Knowledge and Systems Engineering*, pp. 323-328, Ho Chi Minh City, Vietnam, 1-3 Nov. 2018.
[23] C. Wu, F. Wu, S. Wu, Z. Yuan, and Y. Huang, "A hybrid unsupervised method for aspect term and opinion target extraction," *Knowledge-Based Systems*, vol. 148, pp. 66-73, 2018.
[24] R. Dehkharghani, Y. Saygin, B. Yanikoglu, and K. Oflazer, "SentiTurkNet: a Turkish polarity lexicon for sentiment analysis," *Language Resources and Evaluation*, vol. 50, no. 3, pp. 667-685, Sept. 2016.
[25] T. Hofmann, "Probabilistic latent semantic indexing," in Proc. of the 22nd Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 50-57, Berkele, CA, USA, 15-19 Aug. 1999.
[26] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The J. of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.
[27] T. Griffiths and M. Steyvers, "Prediction and semantic association," *Advances in Neural Information Processing Systems*, pp. 11-18, 2002.
[28] M. Steyvers and T. Griffiths, *Probabilistic Topic Models*: *Handbook of Latent Semantic Analysis,* pp. 439-460, Psychology Press, 2007.
[29] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. of the 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 168-177, Seattle, WA, USA, 22-25 Aug. 2004.
[30] S. Blair-Goldensohn, *et al*., "Building a sentiment summarizer for local service reviews," in *Proc. of the WWW2008 Workshop: NLP in the Information Explosion Era*, pp. 14-23, Beijing, China, 22-22 Apr. 2008.
[31] L. W. Ku, Y. T. Liang, and H. H. Chen, "Opinion extraction, summarization and tracking in news and blog corpora," in *Proc. AAAI Spring Symp.: Computational Approaches to Analyzing Weblogs*, pp. 100-107, Mar. 2006.
[32] A. Bagheri, "Integrating word status for joint detection of sentiment and aspect in reviews," *J. of Information Science*, vol. 45, no. 6, pp. 736-755, 2019.
[33] M. Shams and A. Baraani-Dastjerdi, "Enriched LDA (ELDA): combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction," *Expert Systems with Applications*, vol. 80, pp. 136-146, 1 Sept. 2017.
[34] A. Bagheri, M. Saraee, and F. de Jong, "Sentiment classification in Persian: introducing a mutual information-based method for feature selection," in *Proc. 21st Iranian Conf. on Electrical Engineering,* 6 pp., Mashhad, Iran, 14-16 May 2013.

سپیده جمشیدی نژاد دانشجوی دکتری تخصصی مهندسی کامپیوتر- سیستم‌های نرم‌افزاری در دانشگاه آزاد اسلامی واحد رشت است. نام‌برده عضو باشگاه پژوهشگران جوان و نخبگان دانشگاه آزاد اسلامی واحد رشت و مدرس دانشگاه است. زمینه‌های اصلی تحقیقات او داده‌کاوی، متن‌کاوی، پردازش زبان طبیعی، تحلیل احساس، نظرکاوی، شبکه‌های عصبی مصنوعی، یادگیری ماشینی و یادگیری عمیق است.

فاطمه احمدی آبکناری مدرک دکترای مهندسی کامپیوتر خود را از دانشگاه UTM مالزی در سال ۲۰۱۲ دریافت نمود. مدرک کارشناسی ارشد ایشان در رشته فناوری اطلاعات از دانشگاه پلی تکنیک (امیرکبیر) تهران، ایران در سال ۱۳۸۶ است. وی در حال حاضر استادیار دانشکده مهندسی کامپیوتر و فناوری اطلاعات در دانشگاه پیام نور ایران است. زمینه های اصلی تحقیقات او عبارتند از: یادگیری ماشینی، داده‌کاوی، متن‌کاوی، تحلیل احساس، نظرکاوی، شبکه های عصبی مصنوعی، یادگیری عمیق و پردازش زبان طبیعی.

پیمان بیات مدرک دکترای خود را در رشته مهندسی کامپیوتر از دانشگاه UCSI مالزی دریافت کرده است. وی در حال حاضر استادیار دانشکده مهندسی کامپیوتر دانشگاه آزاد اسلامی واحد رشت می‌باشد. زمینه‌های اصلی تحقیق او سیستم‌های توزیع شده، پردازش تصویر و داده‌کاوی است.