به‌کارگیری روش‌های داده‌افزایی برای تحلیل احساسات کاربران دربارهٔ بازگشایی مدارس در دوران همه‌گیری کووید-۱۹

مرضیه میر و سمیرا نوفرستی*

گروه فناوری اطلاعات، دانشکدهٔ مهندسی برق و کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران

چکیده

ازجمله روش‌های موفق برای تحلیل احساسات، روش‌های یادگیری باناظر است که با آموزش یک طبقه‌بند بر روی یک مجموعه‌دادهٔ آموزشی از نظرات دارای برچسب احساس، یک مدل پیش‌بینی‌کننده می‌سازند که قادر است، جملات جدید را طبقه‌بندی کند. در زبان فارسی، نبود داده‌های آموزشیِ کافی و دقت کم ابزارهای پردازش زبان طبیعی، به‌کارگیری الگوریتم‌های باناظر و نیز استخراج ویژگی‌های باکیفیت را با چالش جدی روبه‌رو ساخته‌است. هدف مقالهٔ حاضر به‌کارگیری روش‌های یادگیری ماشین باناظر برای طبقه‌بندی نظرات مطرح‌شده توسط کاربران فارسی زبان در رسانه‌های اجتماعی دربارهٔ بازگشایی مدارس در دوران همه‌گیری کووید-۱۹ است. برای غلبه بر مشکل کمبود داده‌های آموزشی یک روش ترکیبی برای داده‌افزایی پیشنهاد شده‌است که اندازهٔ مجموعه آموزش را حدود ۹۷درصد افزایش می‌دهد. نتایج آزمایش‌های انجام‌گرفته نشان می‌دهد که با اعمال روش پیشنهادی برای داده‌افزایی و به‌کارگیری ویژگی‌های انتخابی در این مقاله، به‌ترتیب دقت ۸۱ و ۷۹درصد برای طبقه‌بندی نظرات با استفاده از الگوریتم‌های ماشین بردار پشتیبان و شبکهٔ عصبی پیچشی حاصل می‌شود.

واژگان کلیدی: تحلیل احساسات، نظرکاوی، یادگیری باناظر، یادگیری عمیق، داده‌افزایی، کووید-۱۹.

Using Data Augmentation Techniques for Sentiment Analysis of Users' Opinions on Reopening of Schools During the Covid-19 Epidemic

Marziye Mir and Samira Noferesti

Information Technology department, Faculty of Electrical and Computer Engineering, Sistan and Baluchestan University, Zahedan, Iran

Abstract

Sentiment analysis, also called opinion mining, is one of the sub-areas of natural language processing that aims to classify texts according to the sentiments, beliefs and attitudes expressed in them. In the most current research, texts are divided into two "positive" and "negative" categories. However, there are also other categories such as good/bad" and agree/disagree, every one of which has its applications.

The purpose of this paper is to analyze the opinions expressed by users on social media about the reopening of schools during the Covid-19 outbreak using supervised machine learning techniques, and to classify them into two "agree" and "disagree" categories. Users' opinions, in this paper, are in

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۲/۳/۱۹ • تاریخ پذیرش: ۱۴۰۲/۱۲/۶ • تاریخ انتشار: ۱۴۰۳/۸/۸ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.21.2.3]

Persian. The lack of sufficient datasets and also the low accuracy of natural language processing tools are the most important problems of text processing in Persian. Due to the mentioned limitations, the use of supervised machine learning algorithms and also the extraction of effective features for training machine learning classifiers in Persian are facing a serious challenge.
In this paper, first, a small dataset of the users' opinions about the reopening of schools was collected and manually labeled. Then, a combined method was used for data augmentation of the dataset. In the proposed method, first, Persian sentences were translated into English. Then nouns, verbs and adjectives of the English sentences were replaced with their synonyms. Next, the English sentences were translated into Persian again. The new sentence with the class label of the initial sentence was added to the training set. Thus, the size of the training set increased by 97 percent. After that, the efficiency of employing the common pre-processing steps and using common feature sets in sentiment analysis of the English texts for Persian were evaluated and the best of them were selected. Considering the low accuracy of the Persian natural language processing tools, it was tried to select those features that were less dependent on the tools. Finally, machine learning classification was used to determine agree/disagree class of the user opinions of the test sets. The results of the experiments indicated that by applying the proposed method for data augmentation and using selected features in this paper, 81 and 79 percent precision was obtained for the polarity classification of opinions using SVM and CNN algorithms, respectively.

Keywords: Sentiment Analysis, Opinion mining, Supervised learning, Deep learning, Data augmentation, Covid-19

۱- مقدمه

در سال‌های اخیر، با محبوبیت‌یافتن وب و شبکه‌های مجازی، شاهد فعالیت گستردهٔ کاربران در وبلاگ‌ها، تارنماهای نظرسنجی و شبکه‌های اجتماعی هستیم. افراد زیادی به‌صورت روزمره به بیان نظرات خود در وب می‌پردازند که موجب ایجاد منابع غنی از نوشته‌های کاربران شده‌است. درحال‌حاضر خیل عظیمی از نظرات دربارهٔ انواع مختلف موجودیت‌ها در دسترس است که تحلیل آن‌ها می‌تواند در کاربردهای مختلف از جمله شناسایی نقاط ضعف و قوت خدمات و محصولات، پیش‌بینی میزان محبوبیت افراد و اشیا در طی زمان و تحلیل دلایل اضطراب و افسردگی در جوامع مفید واقع شود.

تحلیل دستی حجم زیاد نظرات کاربران در زمان معقول کاری دشوار و گاه غیرممکن است؛ ازاین‌رو، شاخه‌ای جدید به‌نام تحلیل احساسات[1] مطرح شده‌است که به استخراج و تحلیل خودکار احساس، عقیده، باور، عواطف و نگرش نویسندهٔ متن دربارهٔ یک موجودیت می‌پردازد. تحلیل احساسات که برخی از پژوهش‌گران آن را نظرکاوی نیز نامیده‌اند، در دامنه‌های متعدد مانند پزشکی، سیاست، تجارت و بازار سهام به‌کار گرفته شده‌است و روزبه‌روز دامنهٔ کاربرد آن گسترش می‌یابد [1-3].

تحلیل احساسات دارای وظایف متعددی از قبیل تعیین قطبیت[2]، استخراج و طبقه‌بندی جنبه، شناسایی نویسندهٔ متن و خلاصه‌سازی نظرات است. اصلی‌ترین وظیفهٔ تحلیل احساسات که موردتوجه اغلب پژوهش‌ها است، تعیین قطبیت است که به طبقه‌بندی نظرات در گروه‌های ازپیش تعیین‌شده می‌پردازد. در بیش‌تر پژوهش‌های موجود، نظرات در دوگروه مثبت و منفی یا در سه‌گروه مثبت، منفی و خنثی دسته‌بندی می‌شوند؛ بااین‌وجود، یک مدل تجزیه‌وتحلیل احساسات می‌تواند برای طبقه‌بندی عواطف (مانند عصبانی، شاد و غمگین)، اضطراری‌بودن محتوا (فوری و غیرفوری)، خوب یا بدبودن اخبار و نیز موافق یا مخالف‌بودن نویسندهٔ متن با موضوع مدنظر به‌کار گرفته شود[1].

با همه‌گیرشدن بیماری کووید-۱۹ و به‌دنبال آن قرنطینه‌شدن مردم، مجازی‌شدن مدارس و دانشگاه‌ها، فاصله‌گذاری اجتماعی، پیامدهای اقتصادی و نگرانی‌های حاصل از آن، نظرات مرتبط با این بیماری در شبکه‌های اجتماعی به‌سرعت در حال افزایش است. امروزه درصد قابل‌توجهی از نوشته‌های کاربران در وب به چالش‌های مرتبط با کووید-۱۹ اختصاص دارد. با گسترش نظرات مطرح‌شده دربارهٔ کووید-۱۹ توجه پژوهش‌گران به تحلیل این نظرات معطوف شده‌است. نظرات دربارهٔ کووید-۱۹موضوعات متعددی مانند اشتغال، بهداشت و آموزش را دربرمی‌گیرد. پژوهش‌گران می‌توانند با تحلیل این قبیل نظرات به سازمان جهانی بهداشت و سازمان‌های دولتی کمک کنند تا از واکنش و احساسات مردم آگاه شوند و تصمیمات مدیریتی هوشمندانه‌تری اتخاذ کنند. به‌طورخاص، کووید-۱۹ بر روی نظام آموزشی اثرات منفی شامل توقف در یادگیری، افزایش تعداد ساعات استفاده از فضای مجازی توسط دانش‌آموزان، نقص مهارت‌های اجتماعی و خطر افزایش ترک تحصیل در مناطق محروم

[1] Sentiment analysis
[2] Polarity detection

[DOI: 10.61186/jsdp.21.2.3]

برجای گذاشته‌است. این امر موجب شده‌است تا موجی از نظرات و دل‌نوشته‌های والدین و معلمین در شبکه‌های اجتماعی راه بیافتد. جمع‌آوری نظرات مطرح‌شده دربارهٔ آموزش در دوران کووید-۱۹ و تحلیل آن‌ها می‌تواند کمک شایانی به شناسایی ریشهٔ اصلی نگرانی‌ها و مشکلات موجود در این زمینه کند.

هدف این مقاله، تحلیل نظرات کاربران فارسی‌زبان توئیتر دربارهٔ بازگشایی مدارس در ایران و طبقه‌بندی نظرات به دوگروه موافق و مخالف با استفاده از الگوریتم‌های یادگیری ماشین باناظر است. بیشتر پژوهش‌های موجود در زمینهٔ تحلیل احساسات بر روی متون انگلیسی صورت گرفته‌است و در سایر زبان‌ها به‌خصوص زبان فارسی به‌دلیل محدودیت ابزارهای پردازش متن، تعداد پژوهش‌های انجام‌گرفته بسیار کمتر است. از طرفی روش‌های مورد استفاده در تحلیل احساسات به دستور زبان، گویش، لهجه و فرهنگ نظردهنده‌ها وابسته است[4]؛ بنابراین، روش‌های ارائه‌شده برای تحلیل احساسات یک زبان خاص به‌ندرت قابل‌استفاده در زبان‌های دیگر است. تحلیل احساسات در زبان فارسی نیز نیازمندی‌هایی دارد که باعث می‌شود استفاده از روش‌های موجود برای تحلیل احساسات متون انگلیسی در زبان فارسی کارایی لازم را نداشته باشند.

روش پیشنهادی این مقاله برای تعیین قطبیت نظرات کاربران دربارهٔ بازگشایی مدارس، یادگیری باناظر است. الگوریتم‌های یادگیری ماشین باناظر نیاز به یک مجموعه آموزش از نظرات برچسب‌خورده دارند که ساخت دستی آن کاری دشوار و زمان‌بر است؛ به‌این دلیل، در این مقاله ابتدا یک مجموعه‌آموزش کوچک از نظرات کاربران به‌صورت دستی برچسب‌گذاری شده، سپس از ترکیب دو روش ترجمهٔ برگشتی و جایگزینی با واژگان مترادف برای افزایش تعداد نمونه‌های مجموعه‌آموزش استفاده شده‌است؛ همچنین کارایی پیش‌پردازش‌ها و ویژگی‌های رایج برای تحلیل احساسات متون انگلیسی، در زبان فارسی بررسی و پیش‌پردازش‌ها و ویژگی‌های کارامد انتخاب شده‌اند. نتایج ارزیابی‌های انجام‌گرفته نشان می‌دهد که دقت طبقه‌بند ماشین بردار پشتیبان با آموزش بر روی مجموعه‌دادهٔ گسترش‌یافته و براساس پیش‌پردازش‌ها و ویژگی‌های انتخابی در مقایسه با مجموعه‌دادهٔ اولیه (بدون داده‌افزایی) حدود پنج‌درصد بهبود داشته‌است. همچنین به‌کارگیری روش پیشنهادی برای داده‌افزایی منجر به افزایش دقت طبقه‌بند شبکهٔ عصبی پیچشی به میزان سه‌درصد شده‌است.

ادامهٔ مقاله به‌صورت زیر سازماندهی شده‌است: بخش ۲ به مرور کارهای پیشین در زمینهٔ تحلیل احساسات به‌ویژه برای متون فارسی و نیز به معرفی کارهای مرتبط با تحلیل نظرات مطرح‌شده دربارهٔ کووید-۱۹ می‌پردازد. در بخش ۳ جزئیات روش پیشنهادی برای تحلیل نظرات کاربران دربارهٔ بازگشایی مدارس تشریح و در بخش ۴ نتایج ارزیابی کارایی روش پیشنهادی ارائه می‌شود. بخش ۵ نتیجه‌گیری است.

۲- مرور کارهای پیشین

پژوهش‌های موجود در زمینهٔ تحلیل احساسات و به‌طورخاص تعیین قطبیت را می‌توان به سه دستهٔ کلیِ روش‌های مبتنی بر واژگان، روش‌های یادگیری ماشین و روش‌های ترکیبی طبقه‌بندی کرد.

روش‌های مبتنی بر واژگان به تعیین قطبیت واژگان موجود در متن با کمک یک واژه‌نامهٔ حسی می‌پردازند و براساس قطبیت واژگان، دربارهٔ قطبیت کلی متن تصمیم می‌گیرند[5,6]. از جمله معروف‌ترین واژه‌نامه‌های حسی در زبان انگلیسی می‌توان به SentiWordNet [7] و SenticNet [8] اشاره کرد. در زبان فارسی نیز می‌توان واژه‌نامه‌های حسی PerSent [9] و LexiPers [10] را نام برد.

روش‌های یادگیری ماشین خود به سه دسته یادگیری باناظر، یادگیری بدون ناظر و یادگیری نیمه‌نظارتی تقسیم می‌شود. در یادگیری باناظر، از یک مجموعه آموزش شامل جملات یا متون دارای برچسب قطبیت، برای آموزش یک طبقه‌بند یادگیری ماشین استفاده می‌شود. این طبقه‌بند قادر به پیش‌بینی قطبیت داده‌های مجموعه‌آزمون است[11,12].

در یادگیری بدون‌ناظر، مدل بر روی مشاهدات یادگیری انجام‌داده و الگوها و ساختارهای پنهان در مجموعه‌داده‌ها را کشف می‌کند. یکی از روش‌های رایج در این دسته خوشه‌بندی نظرات است. در این روش‌ها، داده‌ها بدون دخالت انسان به تعدادی خوشه تقسیم‌بندی می‌شوند[13]. از جمله روش‌های یادگیری بدون‌ناظر برای تحلیل احساسات در زبان فارسی، می‌توان به روش ابتکاری LDSA اشاره کرد[14].

در یادگیری نیمه‌نظارتی، مجموعهٔ کوچکی از داده‌ها به‌صورت دستی برچسب‌گذاری می‌شود و سعی در گسترش این مجموعه به‌صورت خودکار است. در [15] یک چارچوب نیمه‌نظارتی برای نظرکاوی متون فارسی ارائه

[DOI: 10.61186/jsdp.21.2.3]

شده‌است. در این روش، برچسب قطبیت واژگان متن به‌کمک یک واژه‌نامهٔ خودساخت وفقی (پویا و بدون نیاز به خبرهٔ انسانی) تعیین می‌شود؛ درواقع، خصیصه‌های حسی به‌طور خودکار استخراج می‌شوند؛ سپس از مدل مخفی مارکوف خودناظر بر روی خصیصه‌های استخراج‌شده در کنار قوانین مبتنی بر معیار شباهت برای فرایند نظرکاوی استفاده می‌شود.

مطالعات نشان می‌دهد که روش‌های یادگیری ماشین باناظر درکل بهتر از روش‌های مبتنی بر واژگان عمل می‌کنند؛ باوجوداین، روش‌های یادگیری ماشین از یک مشکل عمده رنج می‌برند و آن ساخت مجموعه‌آموزش است. روش‌هایی نیز معرفی شده‌اند که از ترکیب واژگان حسی و یادگیری ماشین استفاده می‌کنند[16, 17].

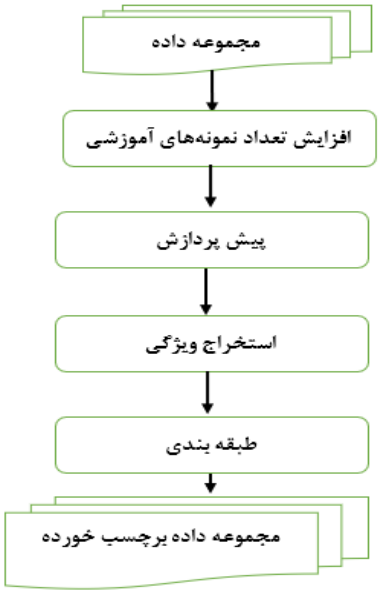
در ادامه به معرفی برخی از پژوهش‌های انجام‌گرفته در زمینهٔ تحلیل احساسات دربارهٔ کووید-۱۹ پرداخته می‌شود. در پژوهش [18] از شبکهٔ عصبی عمیق برای یافتن رابطهٔ بین احساسات مردم کشورهای همسایه دربارهٔ شیوع ویروس کرونا با تحلیل توئیت‌های کاربران استفاده شده‌است. در پژوهش [19] بر اساس هشتگ کووید-۱۹ و ویروس کرونا توئیت‌های موجود جمع‌آوری شده و با استفاده از الگوریتم بیز ساده به تحلیل احساسات آن‌ها پرداخته است.

نتایج نشان می‌دهد که ۳۶درصد توئیت‌ها مثبت، شانزده‌درصد آن‌ها منفی و سایر توئیت‌ها خنثی هستند. در پژوهش [20] نیز به تحلیل احساسات کاربران در رابطه کووید-۱۹ پرداخته شده‌است. بدین منظور از رابط برنامه‌نویسی کاربردی توئیتر برای جمع‌آوری توئیت‌های مرتبط با ویروس کرونا استفاده شده‌است و با استفاده از روش‌های یادگیری ماشین، احساسات در سه گروه مثبت، منفی و خنثی طبقه‌بندی شده‌اند. در پژوهش [21] برای طبقه‌بندی نظرات توئیتر از یک رویکرد ترکیبی (بیز ساده، ماشین بردار پشتیبان، آنتروپی بیشینه، درخت تصمیم و جنگل تصادفی) استفاده شده که دقت ۷۴ درصد حاصل شده‌است.

پژوهش‌های متعددی در زمینهٔ تحلیل احساسات دربارهٔ تأثیر کووید-۱۹ بر بازار سهام، صنعت توریسم و دیگر کاربردها انجام شده‌است[22, 23]. با این وجود، بر طبق مطالعات انجام‌گرفته، تاکنون به تحلیل احساسات نظرات کاربران درباره بازگشایی مدارس پرداخته نشده‌است. درحالی‌که، تحلیل این قبیل نظرات می‌تواند کمک شایانی به مسئولان مربوطه برای درک احساسات و نگرانی‌های اصلی والدین کند. همچنین نتایج حاصل از این تحلیل کمک می‌کند تا مسئولان بتوانند با صرف هزینه و زمان کمتر در جهت گرفتن تصمیمات بهتر در راستای سلامت عمومی جامعه، افزایش رضایت والدین و بالابردن کیفیت آموزش در مدارس گام بردارند.

۳- روش پیشنهادی

روش پیشنهادی برای تعیین قطبیت نظرات شامل چهارمرحلهٔ اصلی است که عبارت‌اند از: افزایش تعداد نمونه‌های آموزشی، پیش‌پردازش جملات، استخراج ویژگی و طبقه‌بندی. مراحل روش پیشنهادی در شکل (۱) نشان داده شده‌است. برای طبقه‌بندی نظرات با کمک الگوریتم‌های یادگیری ماشین باناظر نیاز به‌یک مجموعه از نمونه‌های آموزشی برچسب‌دار است. در ادامه ابتدا دربارهٔ ساخت مجموعه‌داده صحبت و سپس جزئیات هر مرحله از روش پیشنهادی تشریح می‌شود.



(شکل-۱): روش پیشنهادی برای تعیین قطبیت نظرات
(Figure-1): The proposed method for polarity detection

۳-۱- ساخت مجموعه‌داده

برای ساخت مجموعه‌داده، ۷۴۱ نظر (پانصد نظر از توئیتر و ۲۴۱ نظر از بخش تبادل‌نظر تارنما ninisite.com) دربارهٔ بازگشایی مدارس در دوران همه‌گیری کووید-۱۹ جمع‌آوری شده که به هر نظر به‌صورت دستی برچسب موافق/مخالف داده شده‌است. این نظرات براساس عبارات پرتکرار بازگشایی مدارس، آموزش مجازی، ویروس کووید-۱۹ و کرونا جمع‌آوری شده‌است. از این مجموعه دویست نظر به‌صورت تصادفی به‌عنوان مجموعه‌آزمون انتخاب شده‌است.

[DOI: 10.61186/jsdp.21.2.3]

۳-۲- افزایش تعداد نمونه‌های آموزشی

به‌دلیل اینکه تهیهٔ داده‌های آموزشی و برچسب‌زدن آن‌ها به‌صورت دستی کاری دشوار و زمان‌بر است، معمولاً اندازهٔ مجموعه‌داده‌های آموزشی موجود کوچک است که این امر باعث کاهش دقت الگوریتم‌های یادگیری ماشین می‌شود. برای حل مشکل مذکور، در این مقاله روشی برای داده‌افزایی معرفی شده‌است که دو رویکرد ترجمهٔ برگشتی و جایگزینی واژگان مترادف را ترکیب می‌کند.

از آنجا که در مسئلهٔ تحلیل احساسات، واژگان دارای قطبیت، صرف‌نظر از ترتیب قرارگرفتن آن‌ها در جمله، نقش مؤثری دارند، می‌توان با ایدهٔ ترجمهٔ برگشتی جملات جدید ایجاد کرد. در ترجمهٔ برگشتی ابتدا جمله به یک زبان مقصد ترجمه، سپس جملهٔ حاصل از ترجمهٔ مجدد به زبان مبدأ برگردانده می‌شود. در بسیاری از موارد، جملهٔ جدید با جملهٔ اولیه متفاوت است؛ ولی قطبیت یکسانی دارد؛ بنابراین، می‌توان از این جمله به‌عنوان یک نمونهٔ آموزشی جدید استفاده کرد.

ازطرف دیگر، به‌دلیل اینکه واژگان مترادف قطبیت یکسانی دارند، روش جایگزینی با واژگان مترادف نیز می‌تواند برای گسترش مجموعهٔ آموزش مورد استفاده قرارگیرد. بدین ترتیب، یک روش ترکیبی معرفی شده که دارای گام‌های زیر است:

۱) جملهٔ اولیه به‌کمک مترجم گوگل به زبان انگلیسی ترجمه می‌شود.

۲) فقط برای صفات، افعال و اسامی (نقش‌های نحوی مؤثر بر تحلیل احساسات)، با احتمال p، به‌کمک وردنت[1] مترادف‌ها پیدا و جایگزین کلمهٔ اصلی در جمله می‌شوند. وردنت یک واژه‌نامهٔ انگلیسی است که واژگان انگلیسی را در گروه‌هایی از واژگان مترادف دسته‌بندی می‌کند [24]. در این مقاله مقدار p برابر ۰/۵ درنظر گرفته شده‌است.

۳) جملهٔ انگلیسی دوباره به زبان فارسی برگردان می‌شود.

۴) درصورتی‌که جملهٔ برگردان‌شده متفاوت از جملهٔ اولیه باشد، با همان برچسب کلاس جملهٔ اولیه به مجموعه آموزش اضافه می‌شود.

گفتنی است که داده‌افزایی فقط بر روی داده‌های مجموعه آموزش اعمال می‌شود و داده‌های مجموعهٔ آزمون بدون تغییر در تعداد حفظ خواهند شد.

[1] WordNet

۳-۳-پیش پردازش

پیش‌پردازش یکی از مراحل رایج در پردازش متن و یادگیری ماشین است که در بیشتر موارد بر افزایش کارایی الگوریتم‌های تحلیل متن تأثیرگذار است. در این مقاله، پیش‌پردازش‌های رایج در زبان انگلیسی شامل نرمال‌سازی، تقطیع واژه‌ها[2]، حذف هرزواژه‌ها، ریشه‌یابی، برچسب‌گذاری مقولهٔ نحوی واژگان و قطعه‌بندی[3] مورد بررسی قرارگرفت. نتایج ارزیابی‌های انجام‌گرفته نشان داد که مجموعه پیش‌پردازش‌های نرمال‌سازی، تقطیع واژه‌ها و حذف هرزواژه‌ها بیشترین افزایش را در دقت طبقه‌بندی نظرات فارسی داشته‌است (به بخش ۴-۲ مراجعه شود)؛ بنابراین، درنهایت این مجموعه پیش‌پردازش‌ها مورد استفاده قرارگرفت که در ادامه شرح داده می‌شوند.

- **نرمال‌سازی:** هدف نرمال‌سازی، یکسان‌سازی متن ورودی با جایگزین‌کردن کاراکترهای معادل استاندارد است. در پردازش متون فارسی، با توجه به شباهت رسم‌الخط فارسی با رسم‌الخط عربی، مشکل استفاده از نویسه‌های عربی معادل وجود دارد؛ که از جملهٔ آن‌ها می‌توان به حروف «ک»، «ی» و همزه اشاره کرد. در نخستین قدم لازم است مشکلات مربوط به این قبیل حروف را با یکسان‌سازی آن‌ها برطرف کرد. همچنین لازم است نویسهٔ نیم‌فاصله و فاصله (مانند«کتاب‌خانه» و «کتاب خانه») در کاربردهای مختلف آن اصلاح شود و نویسه‌های اعراب، تشدید، تنوین و «ـ» که برای کشش نویسه‌های چسبان مورد استفاده قرار می‌گیرند و موارد مشابه آن حذف یا یکسان‌سازی شوند.
- **تقطیع واژه‌ها:** فرایندی که طی آن جمله به یک‌سری واژه تبدیل می‌شود، تقطیع واژه‌ها نامیده می‌شود. هدف تقطیع واژه‌ها، جداسازی واحدهای با معنی یک جمله است. برای این منظور، به‌طورمعمول از علائمی نظیر نقطه، خط تیره، فاصله و ویرگول استفاده می‌شود.
- **حذف هرزواژه‌ها:** هرزواژه‌ها که ایست‌واژه‌ها نیز نامیده می‌شوند، واژگانی هستند که در متن تکرار فراوان دارند، اما از نظر معنایی دارای اهمیت کمی هستند. از جملهٔ این واژگان در فارسی می‌توان به «اگر»، «و»، «بر» و «که» اشاره کرد. در اغلب عملیات متن‌کاوی با حذف این واژگان نتیجهٔ پردازش بهبود می‌یابد و سبب کاهش بار محاسباتی و افزایش سرعت پردازش می‌شود. برای زبان فارسی نیز فهرست‌هایی از هرزواژه‌ها ارائه شده که میانگین دارای هفتصد کلمه است. این

[2] Tokenization
[3] Chunking

[DOI: 10.61186/jsdp.21.2.3]

فهرست‌ها شامل منفی‌کننده‌هایی مانند «نه» و «هرگز» می‌شود که در تحلیل احساسات از اهمیت ویژه‌ای برخوردارند و حذف آن‌ها کارایی الگوریتم تعیین قطبیت را کاهش می‌دهد؛ بنابراین، قبل از اعمال این مرحله، واژگان منفی‌کننده از فهرست هرزواژه‌ها حذف شده‌اند.

۳-۴- استخراج ویژگی

کارایی الگوریتم‌های یادگیری ماشین به‌شدت وابسته به مجموعه ویژگی‌هایی است که برای آموزش طبقه‌بندها استفاده می‌شوند. در زبان انگلیسی ویژگی‌های متعددی معرفی شده و مورد استفاده قرارگرفته‌اند. برای استخراج اغلب ویژگی‌ها نیاز به ابزارهای پردازش زبان طبیعی است. در زبان فارسی، دقت پایین ابزارها و نیز ماهیت متفاوت خود زبان باعث می‌شود ویژگی‌های موفق در زبان انگلیسی کارایی لازم را نداشته باشند.

در این پژوهش، ویژگی‌های رایج تحلیل احساسات در زبان انگلیسی شامل تک‌واژه‌ها، ترکیب دوتایی واژگان متوالی[1]، نقش نحوی واژگان، درخت وابستگی، قطعات (عبارات) متن، فرکانس واژگان، آزمون مربع کای، منفی‌بودن فعل جمله و قطبیت کلی متن (که با استفاده از قطبیت واژگان آن محاسبه می‌شود) در تحلیل نظرات فارسی مورد ارزیابی قرار گرفته‌است. برای استخراج تک‌واژه‌ها، نقش نحوی واژگان، درخت وابستگی و قطعات متن از ابزارهای موجود کمک گرفته شده‌است. در ادامه به معرفی و نحوهٔ استخراج سایر ویژگی‌ها پرداخته می‌شود:

- **قطبیت کلی متن:** در روش‌های ترکیبی تعیین قطبیت که در سال‌های اخیر موردتوجه پژوهش‌گران بوده‌اند، از قطبیت کلی متن که براساس یک واژگان حسی به‌دست می‌آید به‌عنوان یک ویژگی یادگیری ماشین استفاده شده‌است. در این مقاله برای تعیین این ویژگی ابتدا قطبیت واژگان متن با کمک واژگان حسی PerSent تعیین شده‌است؛ سپس اگر تعداد واژگان مثبت متن بیشتر از تعداد واژگان منفی باشد، قطبیت موافق درنظر گرفته می‌شود و برعکس.
- **منفی‌بودن فعل جمله:** در تحلیل احساسات، منفی‌کننده‌ها نقش مهمی ایفا می‌کنند. برای مثال دو جملهٔ "من با بازگشایی مدارس موافق هستم" و "من با بازگشایی مدارس موافق نیستم" را درنظر بگیرید؛ هر دو جمله ساختار و واژگان مشابهی دارند، اما قطبیت اولی «موافق» و قطبیت دومی «مخالف» است؛ زیرا در دومی از فعل منفی استفاده شده‌است. دراغلب موارد، وجود منفی‌کننده‌ها باعث معکوس‌شدن قطبیت متن می‌شود؛ ازاین‌رو، شناسایی منفی‌کننده‌ها می‌تواند بر کارایی الگوریتم تعیین قطبیت اثرگذار باشد. در زبان فارسی یکی از پررخدادترین اشکال منفی‌کننده، افعال منفی هستند؛ بنابراین، در این مقاله از منفی‌بودن فعل به‌عنوان یک ویژگی دودویی برای طبقه‌بند یادگیری ماشین استفاده شده‌است. برای شناسایی افعال منفی به این صورت عمل شده‌است که در ابتدا باتوجه به نقش نحوی واژگان جمله، واژگانی که به‌عنوان فعل برچسب خورده‌اند مشخص می‌شوند؛ سپس پیشوند و پسوند کلمه حذف می‌شود و ریشه کلمه به‌دست می‌آید. اگر در ابتدای فعل اصلی حرف «ن» باشد، اما بعد از ریشه‌یابی، ریشهٔ فعل با «ن» شروع نشود، نتیجه گرفته می‌شود که «ن» پیشوند منفی‌کننده است و فعل اولیه به‌عنوان یک فعل منفی درنظر گرفته‌می‌شود.
- ترکیب دوتایی واژگان متوالی: به‌دلیل اینکه در برخی موارد یک کلمه به‌تنهایی حس جمله را به‌درستی مشخص نمی‌کند، از ترکیب دو کلمهٔ متوالی به‌عنوان ویژگی استفاده می‌شود. می‌توان ترکیب سه یا بیشتر از واژگان متوالی را نیز درنظر گرفت، اما با افزایش این تعداد، ابعاد بردار ویژگی‌ها بسیار بزرگ می‌شود که منجر به صرف زمان بیشتر در فرایند یادگیری می‌شود.
- فرکانس واژگان و آزمون مربع کای: آزمون مربع کای یک آزمون آماری است که براساس اختلاف فراوانی نمونه‌های مشاهده‌شده و فراوانی مورد انتظار آن‌ها در یک طبقه به ارزیابی میزان ارتباط متغیرهای اسمی می‌پردازد. در روش پیشنهادی برای استفاده از آزمون مربع کای تنها واژگان با فرکانس بالا یعنی واژگانی که بیشتر از دیگر لغات در متن ظاهر شده‌اند، انتخاب می‌شوند. برطبق مطالعات موجود، واژگان با فرکانس بالا تأثیر بیشتری در استخراج دانش مفید از متن دارند. با این توصیف، برای یافتن واژگان مهم، ابتدا لغات با بالاترین تکرار فهرست شده و سپس با استفاده از آزمون مربع کای تنها واژگان معنی‌دار این فهرست انتخاب می‌شوند. توجه شود که قبل از این مرحله، حذف هرزواژه‌ها انجام گرفته‌است. بنابراین، هرزواژه‌هایی که پرتکرارند در فهرست نهایی واژگان مهم ظاهر نمی‌شوند.

بر طبق بررسی‌های انجام‌شده در بخش ۳-۴ مجموعه‌ویژگی‌های درخت وابستگی، ترکیب دوتایی

[1] Bigram

[DOI: 10.61186/jsdp.21.2.3]

واژگان، فرکانس واژگان و آزمون مربع کای بیشترین دقت را برای طبقه‌بندهای یادگیری ماشین به‌همراه داشته‌اند. به‌این دلیل، در روش پیشنهادی تنها از این ویژگی‌ها استفاده شده‌است.

۳-۵- طبقه‌بندی

در این مرحله، مجموعه‌داده که مراحل پیش‌پردازش و استخراج ویژگی بر روی آن انجام شده‌است به‌عنوان مجموعه‌آموزش به یک الگوریتم یادگیری ماشین داده می‌شود. الگوریتم یادگیری ماشین با آموزش بر روی این مجموعه یک مدل پیش‌بینی‌کننده می‌سازد که قادر است به جملات جدید برچسب موافق یا مخالف اختصاص دهد. در این مقاله از دو الگوریتم ماشین بردار پشتیبان و شبکهٔ عصبی پیچشی برای تعیین قطبیت نظرات استفاده شده‌است.

۴- نتایج ارزیابی روش پیشنهادی

برای پیاده‌سازی کلیه مراحل روش پیشنهادی از زبان برنامه‌نویسی پایتون و برای ارزیابی دقت طبقه‌بندی از معیارهای رایج دقت، فراخوانی، صحت و معیار F استفاده شده‌است. در این بخش، ابتدا مجموعه‌دادهٔ مورد استفاده معرفی، سپس نتایج آزمایش‌های انجام‌گرفته برای ارزیابی کارایی روش پیشنهادی ارائه می‌شود.

۴-۱- مجموعه‌دادهٔ مورد استفاده

مجموعه‌دادهٔ مورد استفاده در این مقاله متشکل از ۷۴۱نظر مطرح‌شده توسط کاربران دربارهٔ بازگشایی مدارس در دوران همه‌گیری کووید-۱۹ است که مشخصات آن در جدول (۱) نشان داده شده‌است.

(جدول-۱): مشخصات مجموعه‌دادهٔ مورد استفاده
(Table-1): The characteristics of the used dataset

تعداد نظرات	۷۴۱
تعداد نظرات با برچسب موافق	۳۰۱
تعداد نظرات با برچسب مخالف	۴۴۰
تعداد واژگان منحصر به فرد	۲۵۶۷
تعداد واژگان	۱۰۵۷۲
متوسط طول هر جمله	۱۴.۲۷

ارزیابی تأثیر پیش‌پردازش‌های مختلف بر کارایی الگوریتم طبقه‌بندی نظرات به‌منظور ارزیابی تأثیر پیش‌پردازش‌ها و ویژگی‌های مختلف بر کارایی طبقه‌بندی از سیصد دادهٔ آموزشی استفاده شده‌است. نتایج حاصل از اعمال پیش‌پردازش‌های ذکرشده در بخش ۳-۳ در جدول (۲) نمایش داده شده‌است. همان‌طور که مشاهده می‌شود، ابتدا تنها جملهٔ اصلی به‌عنوان ورودی طبقه‌بند ماشین بردار پشتیبان مورد استفاده قرار گرفته و دقت هفتاددرصد حاصل شده‌است. دلیل انتخاب طبقه‌بند ماشین بردار پشتیبان رایج‌بودن استفاده از آن در تحلیل احساسات و موفقیت‌آمیزبودن نتایج حاصل از آن است. در اینجا به‌دلیل کوچک‌بودن مجموعه‌داده، برای ساخت مجموعه‌آزمون، از روش ارزیابی متقابل با ده‌حلقه[1] استفاده شده‌است.

(جدول-۲): تأثیر پیش‌پردازش بر کارایی طبقه‌بندی
(Tabel-2): Effect of pre-processing on the performance of the classification

حالت	ویژگی	دقت	فراخوانی	صحت	معیارF
۱	جملهٔ اصلی	۷۰	۷۰	۷۰	۷۰
۲	جملهٔ اصلی +نرمال‌سازی+ تقطیع واژه	۷۶	۷۷	۷۷	۷۷
۳	جملهٔ اصلی +نرمال‌سازی+ تقطیع واژه +ریشه‌یابی	۷۴	۷۵	۷۵	۷۴
۴	جملهٔ اصلی +نرمال‌سازی+ تقطیع واژه+ حذف هرزواژه‌ها	۷۵	۷۶	۷۶	۷۶

در حالت دوم در جدول (۲)، ابتدا جملهٔ اصلی نرمال شده و سپس روی آن تقطیع واژه انجام گرفته‌است. در این حالت دقت طبقه‌بندی شش‌درصد افزایش داشته‌است. در زبان فارسی روش‌های متعددی برای نگارش واژگان یکسان وجود دارد که باعث می‌شود این واژگان توسط ماشین متفاوت درنظر گرفته شوند. ازجملهٔ این موارد می‌توان به وجود اشتباهات رایج نگارشی در مستندات فارسی مانند "آیین نامه" به‌جای آئین‌نامه، انواع مختلف نگارش برای یک کلمه مانند "اتاق" و "اطاق"، به‌کاربردن همزه به‌صورت‌های مختلف مانند مسئول و مسوول، استفاده از "ا"و "آ" به‌جای یکدیگر مانند "فرایند" و "فرآیند"، استفاده از پیشوند فعل "می" به‌صورت چسبان و غیرچسبان مانند "می‌تواند" و "میتواند"، تنوع نحوهٔ به‌کاربردن علامت جمع "ها" مانند "آنها" و "آن‌ها" و تنوع نگارش "ی" اضافه در واژگان مختومه به "ه" همانند "مدرسه بزرگ" و "مدرسه‌ی بزرگ" اشاره کرد. بنابر

[1] 10-fold cross validation

[DOI: 10.61186/jsdp.21.2.3]

دلایل یادشده، نرمال‌سازی متن تأثیر به‌سزایی در افزایش دقت طبقه‌بندی دارد که در آزمایش‌های انجام‌گرفته مشهود است.

در حالت سوم در جدول (۲)، بر روی جملات ریشه‌یابی انجام گرفته‌است. مشاهده می‌شود که ریشه‌یابی دقت طبقه‌بندی را دو درصد کاهش داده‌است. این در حالی است که در بسیاری از پژوهش‌های مرتبط با تحلیل احساسات در متون انگلیسی، ریشه‌یابی (به‌دلیل نگاشت واژگان یکسان با اشکال مختلف به‌یک کلمهٔ واحد) تأثیر به‌سزایی در افزایش کارایی الگوریتم طبقه‌بندی داشته‌است؛ باوجوداین، زبان فارسی در ساختار و قاعده با زبان انگلیسی متفاوت است. برخی از مشکلات ذاتی مربوط به متون فارسی که در کاهش دقت ریشه‌یابی تأثیرگذار است، عبارت است از: نبود قواعد گرامری خوش‌تعریف مانند آنچه که در زبان انگلیسی وجود دارد، وجود واژگان چندجزئی مانند "برف‌پاک‌کن"، اتصال برخی از ضمائر مفعولی و نشانه‌های جمع به کلمهٔ قبلی (مانند "مرا"، "بردنش" و "کتابها")، شباهت حروف نشانه‌های جمع و ضمائر مفعولی با برخی از پسوندها و حروف اصلی واژگان فارسی (برای نمونه "ان" در "دانش‌آموزان"نشانهٔ جمع است درحالی‌که "ان" در "خندان" نشانهٔ صفت فاعلی و در "بوستان" بخشی از کلمه است)، شکل مختصر شدهٔ افعال مانند "خوشحالند" و جمع مکسر مانند "کتب" و "مدارس". به‌دلیل مشکلات مذکور، ابزارهای ریشه‌یابی در زبان فارسی از دقت لازم برخوردار نیستند و برخلاف زبان انگلیسی ریشه‌یابی واژگان منجر به کاهش دقت طبقه‌بندی شده‌است. به‌این دلیل، در ادامه از ریشه‌یابی صرف‌نظر شده‌است.

در حالت چهارم در جدول (۲) هرزواژه‌ها حذف شده‌اند. با حذف هرزواژه‌ها در مقایسه با حالت دوم که از جملهٔ نرمال‌شده استفاده می‌کند، دقت یک‌درصد کاهش یافته‌است؛ ولی برای اینکه در مراحل بعدی سرعت کار بالاتر برود و اینکه در انتخاب واژگان با فرکانس بالا هرزواژه‌ها ظاهر نشوند این واژگان از متن حذف شده‌اند. در ادامه به جمله‌ای که بر روی آن عملیات نرمال‌سازی، تقطیع واژه و حذف هرزواژه‌ها انجام شده‌است، جملهٔ پیش‌پردازش‌شده گفته می‌شود.

ارزیابی تأثیر ویژگی‌های معرفی‌شده بر کارایی الگوریتم طبقه‌بندی نظرات در جدول (۳) تأثیر به‌کارگیری مجموعه ویژگی‌های مختلف بر کارایی طبقه‌بند ماشین بردار پشتیبان برای تحلیل نظرات فارسی بررسی شده‌است. گفتنی است که این آزمایش پس از اعمال پیش‌پردازش‌های گفته شده در بخش ۳-۳ بر روی جملات مجموعه‌داده، انجام گرفته‌است و مجموعه ویژگی‌های یادشده درهر سطر جدول به‌همراه بردار واژگان حاصل از جملهٔ پیش‌پردازش‌شده است. درواقع بردار واژگان جملهٔ پیش‌پردازش‌شده به‌عنوان مجموعه ویژگی پایه درنظر گرفته شده‌است که مطابق جدول (۲) با به‌کارگیری آن معیار F و صحت ۷۶درصد برای طبقه‌بند ماشین بردار پشتیبان به‌دست آمده‌است.

(جدول-۳): تأثیر ویژگی‌های معرفی‌شده بر کارایی طبقه‌بندی
(Table-3): Effect of the introduced features on the performance of the classification

مجموعه ویژگی	صحت	دقت	فراخوانی	معیارF
{برچسب نحوی}	۷۶	۷۵	۷۶	۷۵
{قطعه‌بندی}	۷۵	۷۴	۷۵	۷۴
{درخت وابستگی}	۷۸	۷۷	۷۸	۷۷
{درخت وابستگی، قطبیت جمله با واژگان حسی PerSent}	۶۲	۶۲	۶۲	۶۱
{درخت وابستگی، دوتایی}	۷۸	۸۳	۷۸	۸۰
{درخت وابستگی، دوتایی، فرکانس واژگان و آزمون مربع کای}	۸۰	۸۰	۷۹	۷۹
{درخت وابستگی، دوتایی، فرکانس واژگان و آزمون مربع کای، منفی‌بودن فعل جمله}	۷۷	۷۶	۷۷	۷۶

همان‌طور که در جدول (۳) مشاهده می‌شود، مجموعه ویژگی‌های درخت وابستگی، ترکیب دوتایی واژگان، فرکانس واژگان و آزمون مربع کای بیشترین کارایی را برای طبقه‌بند یادگیری ماشین به‌همراه داشته‌اند که در ادامه تنها از آن‌ها استفاده شده‌است. نکتهٔ جالب‌توجه این است که برخی از ویژگی‌های رایج در زبان انگلیسی مانند منفی‌بودن فعل جمله یا قطبیت جمله براساس یک واژگان حسی که در اغلب پژوهش‌های انجام‌شده بر روی متون انگلیسی باعث بهبود دقت طبقه‌بندی شده‌اند، در اینجا کارایی طبقه‌بند ماشین بردار پشتیبان را کاهش داده‌اند. علت اصلی دقت کم ابزارهای موجود در زبان فارسی است. به‌طورخاص در رابطه با ویژگی منفی‌بودن فعل جمله، دلیل اصلی این امر دقت کم الگوریتم‌های ریشه‌یابی واژگان در زبان فارسی و پیچیدگی‌های ذاتی آن است. درواقع، خطاهای ریشه‌یابی باعث می‌شود در بسیاری از

[DOI: 10.61186/jsdp.21.2.3]

موارد تشخیص منفی‌بودن فعل جمله به‌درستی انجام نشود. دربارهٔ ویژگی قطبیت متن براساس واژگان حسی نیز، مسئلهٔ اصلی این است که در اینجا هدف تعیین مثبت یا منفی‌بودن جمله نیست، بلکه به جملات برچسب موافق یا مخالف بازگشایی مدارس داده شده‌است؛ بنابراین، جمله‌ای مانند "آموزش مجازی فاجعه‌بار است" که براساس واژگان حسی برچسب منفی می‌گیرد (به‌دلیل واژهٔ فاجعه‌بار)، در مجموعه‌داده دارای برچسب موافق بازگشایی مدارس است؛ در واقع، در اینجا موضوع جمله که دربارهٔ "آموزش مجازی"، "بازگشایی مدارس" و... است نیز حائز اهمیت است که درحال‌حاضر در مجموعه‌داده موجود نیست و می‌تواند به‌عنوان کار آتی مدنظر قرار گیرد.

۴-۲-ارزیابی روش پیشنهادی برای داده‌افزایی

به‌منظور ارزیابی کارایی روش پیشنهادی برای داده‌افزایی، ابتدا مجموعه‌دادهٔ اولیه شامل۷۴۱ نظر به تصادف به دو بخش آموزش و آزمون تقسیم شده‌است. مجموعهٔ آموزش دارای ۵۴۱ نمونه و مجموعهٔ آزمون دارای دویست نمونه است. نمونه‌های مجموعهٔ آزمون به‌گونه‌ای انتخاب شده‌اند که نیمی از جملات برچسب موافق و نیمی دیگر برچسب مخالف داشته باشند؛ سپس دو آزمایش برای ارزیابی روش پیشنهادی انجام گرفته‌است. در آزمایش نخست یک‌بار از ترجمهٔ برگشتی به تنهایی و بار دیگر از ترکیب ترجمهٔ برگشتی با جایگزینی واژگان مترادف استفاده شده‌است. در جدول (۴) کارایی طبقه‌بند ماشین بردار پشتیبان با آموزش بر روی مجموعه‌دادهٔ اولیه و مجموعه‌های داده حاصل از این دو روش مقایسه شده‌است.

(جدول-۴): مقایسهٔ روش ترجمهٔ برگشتی و روش ترکیبی
(Table-4): Comparison of back-translation and combined methods

روش	تعداد دادهٔ اضافه‌شده	دقت	فراخوانی	صحت	معیارF
بدون‌داده‌افزایی	۰	۷۶	۷۴	۷۴	۷۵
ترجمهٔ برگشتی	۵۰۳	۷۵	۷۳	۷۳	۷۴
روش ترکیبی	۵۲۳	۸۱	۷۹	۷۹	۸۰

همان‌طور که در جدول (۴) مشاهده می‌شود معیار F روش ترکیبی در مقایسه با روش پایه (بدون داده‌افزایی) و داده‌افزایی براساس ترجمهٔ جملات، به‌ترتیب پنج و شش‌درصد بهبود داشته‌است. یکی از مشکلاتی که روش ترجمهٔ برگشتی با آن مواجه است، واژگان مبهم است. واژگان مبهم، واژگانی هستند که چندین معنای متفاوت دارند و بسته به جمله معنای آن مشخص می‌شود. رفع ابهام مقوله‌ای پیچیده در پردازش زبان طبیعی است و هنوز روش دقیقی برای آن ارائه نشده‌است. به‌همین‌دلیل روش‌های ترجمهٔ ماشین گاهی در رفع ابهام واژگان دچار مشکل شده و ترجمهٔ مناسبی برای واژگان مبهم انتخاب نمی‌کنند. مشکل دیگر این است که گاهی جملهٔ انگلیسی پس از برگردان به فارسی عیناً همان جملهٔ اولیه می‌شود. این حالت برای جملات کوتاه که در مجموعه‌دادهٔ مورد استفاده پررخداد هستند بیشتر اتفاق می‌افتد. در این موارد هیچ نمونهٔ آموزشی جدیدی تولید نمی‌شود که این مشکل را می‌توان با جایگزین‌کردن واژگان با مترادفشان تا حدودی حل کرد. بنابر دلایل مذکور روش ترکیبی افزایش بیشتری در اندازهٔ مجموعه‌آموزش داشته و به کارایی بالاتری دست یافته‌است.

در آزمایش دوم، از ترکیب ترجمهٔ برگشتی و جایگزینی با واژگان مترادف استفاده شده‌است، اما در جایگزینی با واژگان مترادف یک‌بار تنها اسامی، یک‌بار تنها افعال، یک‌بار تنها صفات و بار دیگر هر سه‌گروه جایگزین شده‌اند. دلیل انتخاب این سه نقش نحوی، تأثیر بسزای آن‌ها بر قطبیت متن در مقایسه با سایر نقش‌های نحوی مانند ضمایر است. جدول (۵) نشان می‌دهد که جایگزینی هر سه‌گروه اسامی، افعال و صفات کارایی بالاتری دارد و در این حالت بیشترین مقدار برای معیار F (هشتاد درصد) حاصل شده‌است. همچنین جایگزینی هر سه نقش نحوی مذکور منجر به بیشترین نرخ افزایش اندازهٔ مجموعهٔ داده‌ای (۹۶/۶۷درصد) در مقایسه با سایر روش‌ها شده‌است. بنابراین، در روش پیشنهادی از جایگزینی افعال، صفات و اسامی استفاده شده‌است. نکتهٔ حائز اهمیت دیگر در جدول (۵) این است که بیشترین کارایی به‌ترتیب مربوط به جایگزینی افعال، صفات و بعد اسامی است، درحالی‌که بیشترین نرخ افزایش به‌ترتیب با جایگزینی افعال، اسامی و صفات حاصل شده‌است.

خلاصه، طبقه‌بند ماشین بردار پشتیبان با آموزش بر روی مجموعه‌دادهٔ گسترش‌یافته با روش پیشنهادی برای داده‌افزایی به‌دقت ۸۱درصد در طبقه‌بندی نظرات کاربران در رابطه با بازگشایی مدارس دست یافته‌است که در مقایسه با روش‌های پایه دقت بالاتری دارد. همچنین روش پیشنهادی برای داده‌افزایی قادر به گسترش مجموعه‌دادهٔ اولیه به‌میزان ۹۶/۶۷درصد است.

[DOI: 10.61186/jsdp.21.2.3]

(جدول-۵): کارایی روش پیشنهادی برای داده‌افزایی با جایگزینی واژگان مترادف برای نقش‌های نحوی مختلف

(Table-5): Efficiency of the proposed method for data augmentation with the replacement of synonym words for different syntactic roles

روش ترکیبی	تعداد داده اضافه شده	دقت	فراخوانی	صحت	معیار F
با جایگزینی اسامی	۵۱۶	۷۴	۷۳	۷۳	۷۳
با جایگزینی افعال	۵۱۸	۷۹	۷۸	۷۸	۷۸
با جایگزینی صفات	۳۲۰	۷۹	۷۷	۷۷	۷۸
با جایگزینی هر سه	۵۲۳	۸۱	۷۹	۷۹	۸۰

در آزمایش پایانی، تأثیر روش پیشنهادی برای داده‌افزایی بر بهبود عملکرد روش پیشنهادشده در مرجع [25] برای تحلیل احساسات کاربران رسانه‌های اجتماعی فارسی با استفاده از شبکهٔ عصبی پیچشی (CNN[1]) مورد بررسی قرار گرفته‌است. برای این منظور، پس از داده‌افزایی، ابتدا پیش‌پردازش‌های حذف هرزواژه‌ها، نرمال‌سازی و تقطیع واژه‌ها انجام شده‌است و سپس با کمک طبقه‌بند CNN نظرات تعیین قطبیت شده‌اند. مطابق مرجع [25]، طبقه‌بند CNN دارای لایه‌های اصلی تعبیه‌ساز[2]واژگان، کانولوشن (با ۱۲۸فیلتر با اندازهٔ هستهٔ ۳)، ادغام بیشینه‌ای[3] و به‌طورکامل متصل[4] با تابع بیشینهٔ نرم است. همچنین، ابعاد بردار واژگان ورودی برابر ۱۰۰، نرخ حذف تصادفی[5] در لایهٔ ماقبل آخر برابر ۰/۵ و الگوریتم بهینه‌سازی آدام با نرخ یادگیری ۰/۰۰۱ درنظر گرفته شده‌است.

نتایج جدول (۶)، میانگین پنج‌بار اجرای الگوریتم‌ها را نشان می‌دهد. همان‌طور که مشاهده می‌شود، روش پیشنهادی با گسترش مجموعهٔ آموزش توانسته‌است معیارهای دقت و فراخوانی طبقه‌بند CNN را به‌ترتیب به‌میزان ۳ و ۳/۴درصد بهبود دهد.

مقایسهٔ جدول‌های (۴) و (۶) نشان می‌دهد که روش پیشنهادی برای داده‌افزایی با طبقه‌بند ماشین بردار پشتیبان به‌نتایج بهتری در مقایسه با شبکهٔ عصبی پیچشی رسیده‌است که دلیل اصلی آن می‌تواند کوچک‌بودن اندازهٔ مجموعه آموزش باشد، زیرا الگوریتم‌های یادگیری عمیق برای استخراج خودکار ویژگی‌ها نیاز به حجم قابل توجهی دادهٔ آموزشی دارند.

(جدول-۶): تأثیر داده‌افزایی بر کارایی CNN

(Table-6): The effect of data augmentation on CNN performance

روش	دقت	فراخوانی	صحت
CNN [25]	۷۷	۷۴.۶	۷۴.۶
روش پیشنهادی	۸۰	۷۸	۷۸

۵- نتیجه‌گیری

در این مقاله روشی کارا برای تحلیل احساسات نظرات کاربران فارسی‌زبان توئیتر دربارهٔ بازگشایی مدارس در دوران کرونا با استفاده از یادگیری باناظر ارائه شد. در روش پیشنهادی ابتدا یک مجموعه آموزش کوچک به‌صورت دستی برچسب‌گذاری شد و سپس با یک روش ترکیبی تعداد نمونه‌های آن افزایش یافت. همچنین نشان داده شد که به‌دلیل ماهیت خاص زبان فارسی و نیز دقت کم ابزارهای پردازش متن فارسی، برخی پیش‌پردازش‌ها و ویژگی‌های رایج تحلیل احساسات در زبان انگلیسی کارایی لازم را ندارند؛ بنابراین، پیش‌پردازش‌ها و ویژگی‌های مختلفی مورد ارزیابی قرارگرفت و مجموعه پیش‌پردازش‌ها و ویژگی‌های مناسب برای زبان فارسی تعیین و در پایان براساس ویژگی‌های منتخب به طبقه‌بندی نظرات با الگوریتم ماشین بردار پشتیبان پرداخته شد. نتایج ارزیابی‌هایی انجام‌گرفته نشان‌داد که روش پیشنهادی برای داده‌افزایی منجر به گسترش مجموعه آموزش به میزان ۹۶/۶۷ درصد شده‌است. همچنین دقت طبقه‌بند ماشین بردار پشتیبان بر روی مجموعه آموزش پس از داده‌افزایی در مقایسه با مجموعه‌دادهٔ اولیهٔ پنج‌درصد بهبود داشته‌است. همچنین، برطبق ارزیابی‌های انجام‌گرفته، روش پیشنهادی برای داده‌افزایی دقت طبقه‌بند شبکهٔ عصبی پیچشی را به میزان سه‌درصد افزایش داده‌است.

به‌عنوان کار آینده قصد داریم ویژگی‌های مورد بررسی در این مقاله را در یک مدل شبکهٔ عصبی پیچشی چند کاناله (MCNN[6]) به‌کار گرفته و تأثیر این ویژگی‌ها را بر بهبود عملکرد طبقه‌بندی نظرات بررسی کنیم. به‌علاوه، به‌دلیل این که کارایی مدل‌های یادگیری عمیق بسیار وابسته به حجم داده‌های آموزشی است، در ادامهٔ پژوهش حاضر برآنیم تا با درنظرگرفتن نرخ افزایش بیشتر (تولید

[1] Convolutional Neural Network
[2] Embedding
[3] Max-Pooling
[4] Fully connected
[5] Dropout
[6] Multichannel Convolutional Neural Network

[DOI: 10.61186/jsdp.21.2.3]

چندین جمله به ازای هرجمله در مجموعه آموزش اولیه)، نمونه‌های مصنوعی بیشتری تولید کنیم و همچنین با درنظرگرفتن معماری پیچیده‌تری برای شبکهٔ عصبی پیچشی دقت طبقه‌بندی را افزایش دهیم.

6- مراجع 6-Refrences

[1] Liu, B., "Sentiment analysis and opinion mining", Synthesis lectures on human language technologies, Vol. 5, No. 1, pp. 1-167, 2012.

[2] Ahangari Ahangarkolaei, M., Sebti, A. and Yaghoubi, M., "Automatically generate sentiment lexicon for the Persian stock market", Signal and Data Processing, Vol. 20, No. 2, pp. 3-20, 2023.

[3] Noferesti, S. and Shamsfard, M., "A semantic framework based on domain knowledge for opinion mining of drug reviews", Journal of applied research and technology, Vol. 20, No. 6, pp. 652-667, 2022.

[4] Rajabi, Z., Valavi, M. and Hourali M., "Sentiment analysis methods in Persian text: A survey", Signal and Data Processing, Vol. 19, No. 2, pp. 107-132, 2019.

[5] Catelli, R., Pelosi, S., Comito, C., Pizzuti, C. and Esposito, M., "Lexicon-based sentiment analysis to detect opinions and attitude towards COVID-19 vaccines on Twitter in Italy", Computers in Biology and Medicine, Vol. 158, pp. 106876, 2023.

[6] Huang, M., Xie, H., Rao, Y., Liu, Y., Poon, L. K. and Wang, F. L., "Lexicon-based sentiment convolutional neural networks for online review analysis", IEEE Transactions on Affective Computing, 2020.

[7] Baccianella, S., Esuli, A. and Sebastiani, F., "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining", LREC. 2010.

[8] Cambria, E., Liu, Q., Decherchi, S., Xing, F. and Kwok, K., "SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis", In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 3829-3839, 2022.

[9] Dashtipour, K., Hussain, A., Zhou, Q., Gelbukh, A., Hawalah, A.Y. and Cambria, E., "PerSent: A freely available Persian sentiment lexicon", In International Conference on Brain Inspired Cognitive Systems, pp. 310-320, Springer, Cham, 2016.

[10] Sabeti, B., Hosseini, P., Ghassem-Sani, G. and Mirroshandel, S.A., "LexiPers: An ontology based sentiment lexicon for Persian", arXiv preprint arXiv:1911.05263, 2019.

[11] Revathy, G., Alghamdi, S.A., Alahmari, S.M., Yonbawi, S.R., Kumar, A. and Haq, M.A., "Sentiment analysis using machine learning: Progress in the machine intelligence for data science", Sustainable Energy Technologies and Assessments, Vo. 53, pp. 102557, 2022.

[12] Wang, Y., Chen, Q., Shen, J., Hou, B., Ahmed, M. and Li, Z., "Aspect-level sentiment analysis based on gradual machine learning", Knowledge-Based Systems, Vol. 212, pp.106509, 2021.

[13] Riaz, S., Fatima, M., Kamran, M. and Nisar, M.W., "Opinion mining on large scale data using sentiment analysis and k-means clustering", Cluster Computing, Vol. 22, No. 3, pp. 7149-7164, 2019.

[14] Shams, M., Shakery, A. and Faili, H., "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.

[15] Najafzadeh, M., Rahati Quchan, S. and Ghaemi, R., "A Semi-supervised Framework Based on Self-constructed Adaptive Lexicon for Persian Sentiment Analysis", Signal and Data Processing, Vol. 15, No. 2, pp. 89-102, 2018.

[16] Mendon, S., Dutta, P., Behl, A. and Lessmann, S., "A Hybrid approach of machine learning and lexicons to sentiment analysis: enhanced insights from twitter data of natural disasters", Information Systems Frontiers, pp.1-24, 2021.

[17] Ahangari, M. and Sebti, A., "A Hybrid Approach to Sentiment Analysis of Iranian Stock Market User's Opinions", International Journal of Engineering, Vol. 36, No. 3, pp.573-584, 2023.

[18] Imran, A.S., Daudpota, S.M., Kastrati, Z. and Batra, R., "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets", IEEE Access, Vol. 8, pp.181074-181090, 2020.

[19] Manguri, K. H., Ramadhan, R. N. and Amin, P. R. M., "Twitter sentiment analysis on worldwide COVID-19 outbreaks", Kurdistan Journal of Applied Research, pp. 54-65, 2020.

[20] Kaur, C. and Sharma, A., "Twitter Sentiment Analysis on Coronavirus using Textblob", EasyChair, 2020.

[21] Ra, M., Ab, B. and Kc, S., "COVID-19 outbreak: Tweet based analysis and visualization towards the influence of coronavirus in the World", 2020.

[22] Costola, M., Hinz, O., Nofer, M. and Pelizzon, L., "Machine learning sentiment analysis, Covid-19 news and stock market reactions", Research in International Business and Finance, pp. 101881, 2023.

[23] Leelawat, N., Jariyapongpaiboon, S., Promjun, A., Boonyarak, S., Saengtabtim, K., Laosunthara, A., Yudha, A.K. and Tang, J., "Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic

[DOI: 10.61186/jsdp.21.2.3]

using machine learning", Heliyon, Vol. 8, No. 10, pp. e10894, 2022.
[24] Miller, G. A., "WordNet: a lexical database for English", Communications of the ACM, Vol. 38, No. 11, pp. 39-41, 1995.

[۲۵] روحانیان، مرتضی، صالحی، مصطفی، درزی، علی و رنجبر، وحید، "تحلیل احساس در رسانه‌های اجتماعی فارسی با رویکرد شبکهٔ عصبی پیچشی"، مهندسی برق و مهندسی کامپیوتر ایران، شمارهٔ ۳۸، صفحات ۵۹-۶۶، ۱۳۹۹.

مرضیه میر در سال ۱۳۹۶ دورهٔ کارشناسی خود را در رشتهٔ مهندسی فناوری اطلاعات در دانشگاه ملی زابل به پایان رسانده‌است. ایشان درحال‌حاضر دانشجوی مقطع کارشناسی ارشد رشتهٔ مهندسی فناوری اطلاعات گرایش مدیریت سیستم‌های اطلاعاتی در دانشگاه سیستان و بلوچستان است.
نشانی رایانامه ایشان عبارت است از:
marziyemir95@gmail.com

سمیرا نوفرستی در سال ۱۳۸۲ دورهٔ کارشناسی خود را در رشتهٔ مهندسی کامپیوتر در دانشگاه صنعتی شریف به پایان رسانده‌است. درسال ۱۳۸۴ مدرک کارشناسی ارشد خود را در همان رشته از دانشگاه صنعتی امیرکبیر و مدرک دکترا را در سال ۱۳۹۴ از دانشگاه شهید بهشتی دریافت کرده‌است. درحال‌حاضر عضو هیئت علمی دانشگاه سیستان و بلوچستان است.
نشانی رایانامه ایشان عبارت است از:
snoferesti@ece.usb.ac.ir

[DOI: 10.61186/jsdp.21.2.3]