فصلنامه علمی- ترویجی مطالعات دفاع مقدس
دوره ٤، شماره ۳، پیاپی (۱۵)، پاییز ۱۳۹۷، صص ۱٤٤ - ۱۲۱

روش‌های یادگیری ماشین در بررسی ویژگی‌های زبان شعری در اشعار شاعران دفاع مقدس[۱]

(مطالعهٔ موردی: اشعار دو شاعر دفاع مقدس؛ قیصر امین‌پور و محمدرضا عبدالملکیان)

کامیار جوانمردی[۲*]؛ منوچهر اکبری[۳]

تاریخ دریافت: ۱۳۹۷/۰۲/۲۸ تاریخ پذیرش: ۱۳۹۷/۰۵/۱۶

چکیده: *شناسایی سبک نویسنده و پردازش زبان طبیعی از اهمیت ویژه‌ای برخوردار است و پژوهش در این زمینه به دو صورت کیفی و کمی انجام می شود. ازآنجایی‌که شعر و ادبیات همواره یکی از افتخارات تاریخی زبان فارسی به‌شمار می‌آید، شناسایی سبک نویسندگان و شاعران به‌صورت گسترده، بدون اعمال نظرات انسانی و به شیوهٔ کمی، امری ضروری است. در این پژوهش کاربرد روش‌های آماری برای شناسایی سبک نویسنده مورد بررسی قرار می‌گیرد؛ به‌همین‌دلیل، ویژگی‌های واژگانی، حرفی و نحوی از متن‌های منتخب استخراج می‌شود. هدف اصلی مقاله، استخراج ویژگی‌های صوری متن و دسته‌بندی اشعار مربوط به دو شاعر حوزهٔ دفاع مقدس (قیصر امین‌پور و محمدرضا عبدالملکیان) است. به‌این‌منظور، از دو دسته‌بند K نزدیک‌ترین همسایه و بیز ساده جهت انجام عمل دسته‌بندی و انتساب داده‌ها استفاده شد. بررسی هر کدام از دسته‌بندها با استفاده از معیارهای ارزیابی، انجام شد. نتایج ارزیابی‌ها روی سه نوع ویژگی نشان داد، ویژگی‌های واژگانی بدون حذف واژه‌های دستوری در دسته‌بند بیز ساده با ۹۲ درصد معیار F از بالاترین دقت در میان انواع ویژگی‌ها برخوردار است. این نتیجه، نشان‌دهنده کارآیی قابل ملاحظهٔ این نوع ویژگی در شناسایی سبک نویسنده است.*

واژگان اصلی: *شناسایی نویسنده، پردازش متن، سبک شناسی، دسته‌بندی خودکار متن، شعر دفاع مقدس.*

۱. این مقاله از طرح پژوهشی با همین عنوان بدست‌آمده که در پژوهشگاه علوم و معارف دفاع مقدس انجام شده است.

۲*. کارشناسی ارشد دانشگاه صنعتی شریف، نویسندهٔ مسئول (javanmardi69@yahoo.com).
۳. استاد زبان و ادبیات فارسی دانشگاه تهران.

مقدمه

شناسایی نویسنده[1] و ویژگی‌های زبانی متن یکی از مسائل مهم در پردازش زبان طبیعی[2] است. با استفاده از تجزیه‌وتحلیل رایانشی[3]، سبک ادبی نویسنده، روش‌های مختلف دسته‌بندی متن[4] و یا نوشتهٔ نویسنده مشخص می‌شود. شناسایی نویسنده یکی از مسائل چالش‌برانگیز حوزهٔ ادبیات و هم‌چنین، پردازش متن است. شناسایی نویسنده، در واقع نشان‌دادن برخی از ویژگی‌های خاص نویسنده در متن و نوشته است. با این عمل می‌توان نوشته‌های نویسنده را از دیگران متمایز نمود و یا نوشته‌هایی که نویسنده‌ای ناشناس دارد، به نویسنده‌ای مشخص منسوب کرد.

شناسایی نویسنده در ادبیات و سبک‌شناسی همواره با نظرات کارشناسان و صاحب‌نظران در حوزهٔ ادبیات انجام می‌گیرد. این‌گونه تحلیل‌ها باعث ایجاد اختلاف میان صاحب‌نظران شده است. برای نمونه، می‌توان به نظرات متفاوت نسبت به برخی از اشعار خیام، حافظ و شاعران دیگر از ادبیات فارسی و همچنین، نوشته‌های شکسپیر و بیکن از ادبیات انگلیسی اشاره کرد.

ایجاد تنوع در مطالعات سبک‌شناسی و ورود به عرصهٔ پهناور ادبیات با کمک علم آمار و ارقام، وسوسه‌ای جذاب و لذت‌بخش به‌نظر می‌رسد. نگاشت احساسات شاعرانه، مضامین و مفاهیم پیچیدهٔ ادبی به ارقام، اعداد و بررسی عینی‌تر آن‌ها و ساخت پل ارتباطی میان این دو حوزه از اهداف اصلی این پژوهش است. به‌ویژه در عرصهٔ زبان فارسی که به‌نظر مطالعات اندکی در آن انجام شده است. شعر معاصر به‌دلیل قابل‌فهم‌بودن، قرابت فکری و زبانی از لحاظ واژگان و ساختار جمله به زبان معیار کنونی و شعر نو، انتخاب مناسبی به‌نظر می‌رسد و اهداف پژوهش را باوضوح بالاتری نشان می‌داد.

شناسایی نویسنده در حوزهٔ ادبیات، هوش‌مصنوعی و حقوق قضایی اهمیت ویژه‌ای دارد و در شاخهٔ بازیابی اطلاعات[5] و زبان‌شناسی رایانشی[6] قرار می‌گیرد. شناسایی سبک و سیاق نویسندهٔ متون ادبی، رسمی و حتی نامه‌های تهدیدآمیز با استفاده از ویژگی‌های کمی، بسیار کاربردی است.

1. Author identification
2. Natural language processing
3. Computational analysis
4. Text classification
5. Information retrivial
6. Computational linguistics

«هر چه‌قدر حرف نو زدن در ادبیات مشکل است، ورود کامپیوتر به این عرصه از آن هم مشکل‌تر است. با وجود پیشرفت‌های فراوان در تحقیقات کامپیوتری (به‌طوری‌که، محققان در برخی زبان‌ها همانند انگلیسی و اسپانیایی موفق به سرایش شعر به وسیله کامپیوتر شده‌اند)، هنوز اساتید ادبیات و ادیبان ما حاضر به قبول توانایی‌های این ابزار نیستند. از طرفی اساتید کامپیوتر اطلاع چندانی با مشکلات ادبیات ندارند و حاضر نیستند کار نوینی انجام دهند. باتوجه‌به اینکه تاکنون پژوهش‌های اندکی در این زمینه انجام شده، ورود به این عرصه امری ضروری به‌نظرمی رسد» (مجیری و مینایی، ۱۳۸۷ : ۲).

مسئلهٔ اصلی این پژوهش، کاربرد شیوه‌های آماری روش‌های یادگیری ماشین در بررسی سطح زبانی شعر است. شناسایی سبک دو شاعر دفاع مقدس قیصر امین‌پور و محمدرضا عبدالملکیان با استفاده از روش‌های آماری به عنوان مطالعهٔ موردی جهت نشان‌دادن این مسئله انتخاب شده است. انتخاب این دو شاعر به‌دلیل ویژگی‌های زبانی و شعری آن‌ها، برخورداری از سبک ویژه و شاخص‌بودن اشعار آنان در زمینه شعر جنگ و به‌ویژه شعر دفاع مقدس بوده است.

مهم‌ترین مرحلهٔ انجام این کار و اصلی‌ترین گام برای آغاز تحلیل آماری، سبک‌سنجی گزینش و تعریف ویژگی‌ها است. این ویژگی‌ها در شناسایی سبک نویسنده یا شاعر، شرایطی را نشان خواهد داد که قابلیت کمی‌سازی و تحلیل آماری را فراهم می‌کند. شاعرانی که در این پژوهش مورد بررسی قرار گرفته‌اند، آثارشان مشخص و در دسترس است. بخشی از این اشعار پس از جمع‌آوری، به‌عنوان اسناد آموزش[1] و بخشی دیگر برای سنجش دقت سامانهٔ نهایی این پژوهش استفاده می‌شوند. داده‌هایی که در بخش آموزش قرار ندارند، اسناد آزمون[2] یا آزمایش نامیده می‌شوند. درنهایت در این پژوهش با استفاده از ویژگی‌های واژگانی، نحوی، حرفی و بهره‌گیری از دو دسته‌بند K نزدیک‌ترین همسایه[3] و بیز ساده[4] سبک دو شاعر حوزهٔ دفاع مقدس، قیصر امین‌پور و محمدرضا عبدالمکیان از یکدیگر متمایز خواهد شد.

1. Training document
2. Test document
3. K-nearest neighbors
4. Naïve Bayes

در طراحی و پیاده‌سازی تمام بخش‌ها از جمله دسته‌بندها و برچسب‌زن نقش نحوی کلمات، از ابزارهای آماده استفاده نشده است. همچنین، در تمامی مراحل پژوهش، زبان برنامه‌نویسی پایتون[1] برای طراحی و پیاده‌سازی برنامه‌ها کاربرد داشته است. این امر باعث پیاده‌سازی سطوح مختلف پردازش زبانی مستقل می‌شود تا در پژوهش‌های آینده، قابلیت بهره‌برداری داشته و آمارهای پیش‌بینی‌نشده، امکان محاسبه و بررسی بیشتری داشته باشند.

پیشینهٔ پژوهش

نخستین پژوهش در زمینهٔ شناسایی مؤلف به سال ۱۷۱۱ میلادی بازمی‌گردد که روی متن کتاب مقدس انجام شده است. کشیش آلمانی به نام اچ. بی. ویتر[2] در سال ۱۷۱۱ میلادی به این نکته اشاره کرد، نام‌های متفاوتی از خداوند در کتب پنج‌گانه وجود دارد که نویسندگان متعددی آن‌ها را نوشته‌اند.

پژوهش‌های شناسایی نویسندهٔ متن با استفاده از روش‌های نوین آماری و رایانشی، در اواخر قرن نوزدهم آغاز شده است (Olsson, 2004: 3). «تلاش‌های اولیه جهت سنجش سبک نوشتاری نویسنده به قرن نوزدهم بازمی‌گردد. این مطالعات توسط مندهال در سال ۱۸۸۷ روی نمایش‌نامه‌های شکسپیر انجام شد. سپس، مطالعات آماری توسط زیف در سال ۱۹۳۲ و یول در سال‌های میانی قرن بیستم انجام شد. مطالعات دقیق‌تر توسط ماستلر و والاس در سال ۱۹٦٤ و روی پایگاه‌دادهٔ فدرالیست پیپر[3] انجام شد که بی‌شک، یکی از قدرتمندترین و مؤثرترین تحقیقات در تعیین هویت نویسنده بوده است. این پژوهش‌ها، آغاز مطالعهٔ غیرسنتی تشخیص هویت نویسنده محسوب می‌شود» (فرهمندپور و همکاران، ۱۳۹۱: ۲۷).

در مطالعه‌ای (Abbasi & Chen, 2005: 68) محتوای سایت‌ها و پیام‌های گروه‌های تروریستی برای شناسایی خودکار، بررسی شده است. در این تحقیق از تحلیل پیام‌های آنلاین موجود در وب‌سایت‌ها و انجمن‌های گفتگوی[4] عربی و انگلیسی و چارچوب‌های موجود برای

1. Python programming language
2. H. B. Witter
3. The Federalist Paper
4. Web forum

شناسایی نویسنده استفاده شده است. در این پژوهش ویژگی‌هایی ازجمله طول واژه، حذف و اضافه، علائم سجاوندی، ریشهٔ واژگان، ساختار واژه و محتوا بررسی شده است.

پژوهش دیگر (Baayen et al., 2002: 2) به زبان هلندی انجام گرفته است. در این پژوهش از هشت دانشجوی هلندی رشتهٔ ادبیات خواسته شده تا هر یک، سه اثر در سبک‌های تخیلی، تحلیلی و توصیفی با طول هزار واژه بنویسند. درنهایت، نُه اثر از هر نویسنده و در کل از ۷۲ یادداشت به‌عنوان پیکره پژوهش استفاده شده است. دقت این کار در صورتی که ٤٢ واژهٔ دستوری و ۸ علامت سجاوندی لحاظ شده باشد، ۸۳/۵ درصد است و اگر ۵۰ واژهٔ دستوری و ۸ علامت سجاوندی درنظر گرفته شود، دقت به ۸۸/۱ درصد خواهد رسید.

در پژوهش دیگری (Patton & Can, 2004: 468) چهار داستان از یک نویسنده بررسی شده است. این تحلیل با استفاده از ویژگی‌های پرکاربردترین واژه‌ها، شمارش هجاها، اطلاعات نوع کلمه و برچسب نحوی، طول جملات برحسب واژه‌ها، طول واژه در متن و طول واژه در واژگان، انجام گرفته است. در این بررسی، هر داستان در قالب پنج‌هزار واژه قرار گرفته و هر یک از ویژگی‌های مذکور باتوجه‌به پیکره به‌دست آمده، شمارش شده است. براساس نتایج، دو رمان اول از همدیگر متمایز و دو رمان آخر درهم تنیده شده‌اند. بهترین ویژگی‌هایی که در این پژوهش براساس جداسازی و تفکیک میان متون به‌دست آمد، پرکاربردترین واژه‌ها و طول جمله است. این دو ویژگی باعث دسته‌بندی متون به ترتیب با دقت ۸۷ و ۸۱ درصد شد.

در پژوهش‌هایی، روش‌های آماری و نحوی در ترکیب با یکدیگر استفاده شده‌اند. در پژوهشی (Mechti et al., 2015: 2) روش‌های آماری و سبکی در الگوریتم‌های یادگیری ماشین ترکیب شده است. مهدی بازرگان، از نخستین افرادی است که به ویژگی‌های آماری در متون فارسی توجه کرد. وی با استفاده از «جمله‌نگار» که نمایش هندسی از تعداد واژه‌های تشکیل‌دهندهٔ جملات و تعداد جملات است به بررسی دیباچهٔ گلستان پرداخت. نتایج این پژوهش نشان می‌دهد، شاخص‌های اصلی جمله‌نگار در بخش‌های دیگر گلستان کم‌وبیش همانند دیباچه است؛ اما، جمله‌نگار مقدمهٔ گلستان که نوشتهٔ محمدعلی فروغی است، چهرهٔ متفاوتی دارد (استاجی، ۱۳۸۷: ۱۷).

در پژوهشی شاهمیری و بروجردی (۱۳۸۵) سه ویژگی فیزیکی، آوایی و مفهومی را از اشعار خیام، حافظ، فردوسی و مولوی استخراج کرده‌اند. سپس، با استفاده از شبکهٔ عصبی مصنوعی و درخت تصمیم‌گیری دسته‌بندی و شناسایی انجام داده‌اند. در مطالعهٔ دیگری (آذین، ۱۳۹۲) اشعار نیما یوشیج، مهدی اخوان‌ثالث، احمد شاملو و سهراب سپهری با استفاده از روش‌های یادگیری ماشین، بررسی شده است. این روش‌ها در تحلیل مستقل بوده و از معیارهای انسانی تأثیر نمی‌گیرند.

چارچوب نظری

شناسایی سبک نویسنده از مسائل چالش‌برانگیز حوزهٔ ادبیات و پردازش متن است. «واژهٔ سبک، عربی و از ریشهٔ «سَبَک» به معنی «ذوب کردن و در قالب ریختن» است. سابقهٔ کاربرد اصطلاح سبک در زبان عربی به سال‌های سوم هجری‌قمری در آثار جاحظ و ابوهلال عسکری می‌رسد. در ادبیات فارسی و عربی واژه‌های طرز، روش، سیاق و نمط مترادف سبک است. در زبان فارسی اصطلاح سبک‌شناسی را نخستین بار محمدتقی بهار برای کتاب خود «سبک‌شناسی نثر» استفاده کرد. در زبان‌های لاتین «استیل[1]» به معنی سبک است که قلمی نوک تیز بوده و با آن حروف را روی صفحه‌های سخت می‌تراشیدند. بعدها این واژه به‌طور مجازی به روش خاص شکل دادن به حروف در نوشتار تعبیر شده است. هر اثر ادبی همانند هر انسانی دارای مشخصاتی باشد؛ ازاین‌رو، در پاره‌ای از صفات و خصال با آثار دیگر شریک است» (سمیعی گیلانی، ١٣٨٦: ٥٥).

بررسی و مطالعهٔ سبک، از موضوع‌های نوین در حوزهٔ زبان‌شناسی است. در سبک‌شناسی، دانش گونه‌های کاربردی زبان به عنوان شیوهٔ عمل تعریف می‌شود؛ اما، واژهٔ سبک در کاربرد جاری زبان، بر معنا و مفاهیم متعددی همانند سبک چیدمان منزل و سبک ورزشی دلالت دارد. همچنین در مواردی که هنرمند به درجه‌ای از تسلط دست می‌یابد، اذعان می‌شود که وی در هنر خود صاحب سبک است (سمیعی گیلانی، ١٣٨٤: ٨٧).

تاکنون عموم پژوهش‌ها در حوزهٔ سبک‌شناسی به شیوهٔ کیفی انجام شده است. در این روش، پژوهشگر با در نظر گرفتن دانش و درک خود از اثر مورد مطالعه و مکاتب ادبی، به

1. style

نتیجه‌گیری می‌رسد. «در این شیوه نتایج به‌دست آمده معمولاً موردی و غیرقابل تعمیم است؛ زیرا برخلاف آزمایش‌هایی که در رشته‌های تجربی انجام می‌شود، این مطالعات نمی‌تواند مجدداً توسط افراد و شرایط گوناگون آزمایش شود. ازاین‌رو دانش زبان‌شناسی که تمایل بسیاری به شیوه‌های علمی و تجربی دارد، به‌دنبال شیوه‌های کمی و قابل‌تکرار بوده است. این امر منجر به ظهور حوزه مطالعات سبک‌شناسی و تشخیص نویسنده شده است» (یاحقی و ایزانلو، ۱۳۸۵: ۱۵۳).

در این پژوهش، شناسایی نویسنده در قالب یک مسئلهٔ دسته‌بندی بررسی شده است. برای انجام بررسی‌ها، فضای مسئله به دسته‌هایی به تعداد نویسندگانی که در داده‌های آموزشی هستند، تقسیم می‌شود. روش‌های مختلفی برای دسته‌بندی اطلاعات وجود دارد که در این پژوهش به دو دسته‌بندی اشاره می‌شود.

دسته‌بند بیز ساده[1]

در زبان طبیعی، ارجاع سند D به مجموعه‌های از پیش تعریف شده $C= \{c_1,c_2,\ldots,c_{|c|}\}$ اطلاق می‌شود. به‌طورمعمول، ساختار آموزشی با نظارت[2] برای آموزش دسته‌بند استفاده می‌شود. الگوریتم آموزشی، مجموعه‌ای از N نمونهٔ برچسب خوردهٔ آموزشی $\{(d_i,c_j): i,j=1,\ldots,N\}$ فراهم می‌کند که با استفاده از آن یک تابع دسته‌بندی ایجاد می‌شود و اسناد را به دسته‌های نگاشت می‌دهد. d_i سند i و C_j دستهٔ j است که d_i به آن نگاشت می‌شود. در این رابطه D و C متغیرهای تصادفی برای ارزش‌های عددی سند و دسته هستند (آذین، ۱۳۹۲ : ۱۳). رایج‌ترین الگوریتم آموزش برای دسته‌بندی متون براساس قانون بیز (رابطهٔ شمارهٔ ۱) است.

رابطهٔ شمارهٔ ۱:

$$P(C = c|D = d) = \frac{P(C = c).P(D = d|C = c)}{P(D = d)}$$

دسته‌بند K نزدیک‌ترین همسایه

1. Naive Bayes Classifiers
2. Supervised learning framework

این دسته‌بند به‌دلیل سادگی، کاربرد بسیاری در طبقه‌بندی متون دارد. مبنای کار این الگوریتم، مقایسهٔ متن آزمایش شده با متون آموزش داده شده و در نهایت تعیین میزان شباهت میان آن‌ها است. متون آموزشی با n ویژگی به‌عنوان یک نقطه در فضای n بُعدی نمایش داده می‌شود. هنگامی که متن ناشناخته‌ای الگوریتم دریافت می‌کند، فضای الگو برای یافتن متون آموزشی که مشابه متن ناشناخته باشد، جستجو می‌کند (سرایی و شاهقلیان، ١٣٨٩).

«در این شیوه، نمونه‌ها در یک فضای n- بعدی به‌صورت نقاطی متمایز قرار می‌گیرند و دسته‌بندی با استفاده از اندازه‌گیری و مقایسهٔ فاصلهٔ اقلیدسی[1] نقاط مختلف روی بردار ویژگی‌ها انجام می‌گیرد. در مدل‌های برداری[2] که از روش K نزدیک‌ترین همسایه استفاده می‌شود، دو مقیاس اندازه‌گیری فاصله، یعنی فاصلهٔ اقلیدسی و کسینوسی[3] کاربرد دارد» (آذین، ١٣٩٢ : ١٣).

معیارهای ارزیابی

دسته‌بندی متون با استفاده از مقایسه داده‌های پیش‌بینی شده و مقادیر واقعی، ارزیابی و سنجش می‌شود. جدول شمارهٔ ١ ماتریس دسته‌بندی ساده‌ای را برای مسئلهٔ دسته‌بندی نشان می‌دهد.

جدول شمارهٔ ١: دسته‌بندی دوحالته

		مقدار خروجی سامانهٔ دستهٔ پیش‌بینی شده[4]	
		درست (مثبت)	نادرست (منفی)
مقدار واقعی[7]	درست	مثبت صحیح[6]	منفی غلط[5]
	نادرست	مثبت غلط[9]	منفی صحیح[8]

1. Euclidean distance
2. Vector models
3. Cosine similarity
4. Predicted Class
5. False Negative
6. True Positive
7. True Class
8. True Negative
9. False Positive

در جدول شمارهٔ ۱، مقادیر موجود در خانه‌های مثبت و منفی مبین دقت و پیش‌بینی صحیح سامانه است. همچنین، مقادیر موجود در خانه‌های مثبت و منفی غلط، مبین خطای دسته‌بند و پیش‌بینی اشتباه است. معیارهای دقت، بازخوانی و معیار F براساس جدول شمارهٔ ۲ به‌دست می‌آیند.

جدول شمارهٔ ۲: معیارهای ارزیابی

معیار	فرمول
دقت	$\frac{\text{مثبت صحیح}}{\text{مثبت غلط} + \text{مثبت صحیح}}$
بازخوانی	$\frac{\text{مثبت صحیح}}{\text{منفی غلط} + \text{مثبت صحیح}}$
معیار F	$\frac{2 \times \text{بازخوانی} \times \text{دقت}}{\text{بازخوانی} + \text{دقت}}$

داده‌ها و مباحث تفصیلی تحقیق

در این پژوهش، ویژگی‌های صوری در سه سطح واژگانی، نحوی و حرفی دو شاعر به صورت کمّی بررسی شده است؛ ازاین‌رو، هر شعر به‌عنوان یک سند و مجموعهٔ اسناد به‌عنوان پیکرهٔ[1] سامانه پنداشته شده است. برای محاسبهٔ این ویژگی‌ها از زبان برنامه‌نویسی پایتون و ابزار پردازش زبان طبیعی[2] استفاده شده است. ویژگی‌های کمی محاسبه شده، به‌عنوان ویژگی‌های دسته‌بندها در امر دسته‌بندی متون لحاظ شده است. همچنین، دو دسته‌بند K نزدیکترین همسایه و بیز ساده کاربرد داشته است که این دسته‌بندها عمل انتساب متون را به شاعران موردنظر انجام می‌دهند.

پیش‌ازآنکه پیکره‌ای برای انجام پردازش استفاده شود، باید ابتدا روی داده‌ها پیش‌پردازش انجام گیرد تا صورت‌های غیراستاندارد موجود در پیکره به شکل استاندارد تبدیل شوند؛ زیرا،

1. Corpus
2. Natural language toolkit

چنانچه حروف، نشانه‌های نگارشی و کلمات فارسی به شکل یکسانی نوشته نشوند، متون مورداستفاده توسط سامانه‌های رایانه‌ای قابل‌تحلیل نخواهند بود.

جدول شمارهٔ ۳: تعداد کل اشعار پیکره شاعران

شاعر	تعداد اشعار
امین‌پور	۱۲۲
عبدالملکیان	۳۷٦

پیکرهٔ کاربردی این پژوهش، متن اشعار قیصر امین‌پور و محمدرضا عبدالملکیان است که تعداد آن‌ها در جدول شمارهٔ ۳ نشان داده شد. تمام اشعار به‌صورت دستی تایپ و هر شعر در قالب یک سند و مجموع اشعار به‌عنوان کل پیکره برای استخراج ویژگی‌ها و دسته‌بندی به‌کار گرفته شد. کتاب‌های شعر قیصر امین‌پور که در این مطالعه استفاده شد، کتاب‌های از آینه‌های ناگهان دفتر اول، از آینه‌های ناگهان دفتر دوم، از تنفس صبح، بی‌بال پریدن و دستور زبان عشق است. همچنین، از آثار محمدرضا عبدالملکیان دفتر شعرهای ساده با تو حرف می‌زنم، هوای حوصله ابری‌ست، دنیا به کبوترانش پشت کرده است و پل خواب انتخاب شد. هر کدام از کتاب‌های شعر این دو شاعر شامل چندین شعر و نوشته است. در جدول شمارهٔ ۳، تعداد اشعار و در جدول شمارهٔ ٤ تعداد کل کلمات اشعار هر شاعر نمایش داده شده است.

جدول شمارهٔ ٤: تعداد کل کلمات پیکره اشعار شاعران

شاعر	تعداد کل کلمات
امین‌پور	١٤٥١٣
عبدالملکیان	١٨٢٥٥

ویژگی‌های واژگانی

بردار ویژگی‌های واژگانی که شامل تعداد کلمه در هر سند یا شعر است، به‌دست می‌آید. درواقع، هزار کلمه پراهمیت از لحاظ معیار tf-idf[1] به‌عنوان بردار ویژگی‌های واژگانی درنظر گرفته می‌شود.

الگوی وزن‌دهی بر مبنای بسامد کلمات و خواص تکرار کلمات در مجموعهٔ اسناد تعیین می‌شود. وزن‌های محلی، تابعی از تکرار هر کلمه در یک سند است. همچنین، کلمه‌ای که بیشترین تکرار را داشته باشد، در زمرهٔ کلمه‌های مهم سند است (کامیار، ۱۳۹۰). پنج کلمه از پراهمیت‌ترین کلمه‌ها برای بررسی واژگانی اشعار در جدول شمارهٔ ۵ نمایش داده شده است.

جدول شمارهٔ ۵: واژه‌های پرکاربرد پیکرهٔ اشعار بر اساس معیار tf-idf

ترتیب اهمیت	کلمه	معیار tf-idf
۱	با	۰/۰۰٤٤۳۹۹٦
۲	حالا	۰/۰۰٤۳۷۱٤٥
۳	من	۰/۰۰٤۳٤۱۲۰
٤	نه	۰/۰۰٤۲٦٥٦۲
٥	هم	۰/۰۰٤۲۲۲۷۷

یکی از روش‌های ساده برای نشان‌دادن متون، استفاده از بردار وقوع کلمات است. در بسیاری از پژوهش‌های شناسایی نویسنده، برای مطالعه و شناخت سبک وی از ویژگی‌های واژگانی استفاده می‌شود. به این شیوه در حوزهٔ بازیابی اطلاعات و دسته‌بندی متون مبتنی بر موضوع[2]، کیسهٔ لغات[3] می‌گویند. در این روش، متن به‌عنوان مجموعه‌ای از کلمات درنظرگرفته می‌شود که هر کلمه تعداد وقوع یا بسامدی دارد و مستقل از اطلاعات بافتی[4] است.

1. Term frequency-inverse document frequency
2. Topic-based text classification
3. Bag of words
4. Contextual information

بااین‌حال، تمایزهایی میان شناسایی سبک نویسنده با روش‌های بازیابی اطلاعات و دسته‌بندی متون مبتنی بر موضوع وجود دارد. کلمه‌هایی که تکرار بسیاری دارند (مانند حروف تعریف، حروف اضافه، ضمایر و حروف ربط)، از ویژگی‌های مفید در شناسایی سبک نویسنده به‌شمارمی‌آیند؛ اما، این نوع کلمات در روش‌های دسته‌بندی متون براساس موضوع، معمولاً کلمات زائد درنظر گرفته می‌شوند و در همان ابتدای کار به‌عنوان یکی از مراحل پیش‌پردازش از متن حذف می‌شوند. این نوع کلمات را ایست‌واژه[1] می‌نامند و ازآنجایی‌که فاقد اطلاعات معنایی هستند، در محاسبات لحاظ نمی‌شوند.

کلمات دستوری یا همان ایست‌واژه‌ها به‌وفور و به‌صورت ناخودآگاه، فارغ از موضوع و بافت‌های گوناگون استفاده می‌شوند؛ بنابراین، این نوع کلمات در موضوع‌های گوناگون می‌توانند عامل مهمی در شناسایی سبک نویسنده باشند (Stamatatos, 2009: 541).

پس از تعیین ویژگی‌ها و تشکیل بردار برای هریک، با استفاده از دو دسته‌بند به انتساب اسناد آزمایش به شاعران پرداخته می‌شود. برای آزمایش دسته‌بند بیز ساده، ۱۵ درصد داده‌ها به‌عنوان داده آزمون استفاده می‌شود و مابقی به‌صورت داده آموزش در دسته‌بند مورد استفاده قرار می‌گیرد. جدول شمارهٔ ٦ نمایان‌گر تعداد اسناد (اشعار)، داده‌های بخش‌های آموزش و آزمون برای هر دسته (شاعر) است.

جدول شمارهٔ ٦: تعداد اشعار انتخاب شده برای بخش آموزش و آزمون

	داده‌های آموزش	داده‌های آزمون
امین‌پور	١٠٤	١٨
عبدالملکیان	٣٣٠	٤٦
مجموع	٤٣٤	٦٤

پس از اجرای مرحلهٔ آموزش، داده‌های آزمون در دسته‌بند قرار می‌گیرند. درنهایت، دسته‌بند بیز ساده با ویژگی‌های واژگانی آموزش داده می‌شود. سپس، نتایج ارزیابی این روش با

1. Stop word

معیارهای دقت، بازخوانی و معیار F (جدول شمارهٔ ۷) و همچنین، نتایج ماتریس درهم‌ریختگی[1] (جدول شمارهٔ ۸) اندازه‌گیری می‌شود.

به‌منظور بهبود هر الگوریتمی، نیاز است تا این مدل بررسی شود و نقاطی که دچار خطا است، شناسایی شود. تحلیل خطا در دسته‌بندها از طریق ماتریس درهم‌ریختگی یا جدول وابستگی[2] انجام می‌گیرد. ماتریس درهم‌ریختگی برای دسته‌بندی با N دسته، یک ماتریس N در N است. در این ماتریکس، درایهٔ x, y، شامل تعداد دفعات داده‌ای است که در دستهٔ x قرار دارد و توسط مدل، به‌عنوان داده‌ای از دستهٔ y قرار می‌گیرد.

جدول شمارهٔ ۷: نتایج ارزیابی دسته‌بند بیز ساده بدون حذف ایست‌واژه‌ها روی ویژگی‌های واژگانی

	بازخوانی	دقت	معیار F
امین‌پور	۹٤/٤٤	۸٥/۰۰	۸۹/٤۷
عبدالملکیان	۹۳/٤۷	۹۷/۷۲	۹٥/٥٥
میانگین	۹۳/۹٦	۹۱/۳٦	۹۲/٥۱

جدول شمارهٔ ۸: نتایج ماتریس درهم‌ریختگی دسته‌بند بیز ساده بدون حذف ایست‌واژه‌ها روی ویژگی‌های واژگانی

	عبدالملکیان	امین‌پور
عبدالملکیان	٤۳	۳
امین‌پور	۲	۱٦

در مثال دسته‌بند بیز ساده، ویژگی‌های واژه‌ها، جداول و اعداد درون درایه‌های قطر اصلی ماتریس بدین‌معنا است که اسناد توسط دسته‌بند درست پیش‌بینی شده‌اند. برای نمونه، از میان ٤٦ سند عبدالملکیان برای آزمایش در دسته‌بندی، ٤۳ مورد به‌درستی تشخیص داده شده‌اند.

1. Confusion Matrix
2. Contingency Table

همچنین، سه نمونهٔ دیگر به‌عنوان اشعار امین‌پور دسته‌بندی شده‌اند. تحلیل خطا به‌این‌صورت، بخش بسیار مهمی از انواع کاربرد زبان‌شناسی رایانشی است و می‌تواند مبین مشکلاتی در داده‌های آموزش باشد. مهم‌تر از همه، تحلیل خطا در توسعهٔ انواع دانش و الگوریتم‌ها برای استفاده در حل مسائل مفید است (Jurafsky & Martin, 2006).

در گام بعدی، تمام ایست‌واژه‌ها از میان اشعار حذف شده‌اند و بار دیگر دسته‌بندی بیز ساده روی داده‌ها، آزمایش شد. برخی از ایست‌واژه‌ها در جدول شمارهٔ ۹ نشان داده شده است. سپس، نتایج ارزیابی دسته‌بند و ماتریس درهم‌ریختگی بعد از حذف ایست‌واژه‌ها در جداول شمارهٔ ۱۰ و ۱۱ درج شد.

جدول شمارهٔ ۹: برخی از انواع ایست‌واژه‌های زبان فارسی

ضمیر	حرف‌اضافه	حرف ربط	فعل
من	از	و	است
ما	برای	اما	بود
تو	به	که	شد
همین	را	پس	هستند

جدول شمارهٔ ۱۰: ارزیابی دسته‌بند بیز ساده پس از حذف ایست‌واژه‌ها روی ویژگی‌های واژگانی

	بازخوانی	دقت	معیار F
امین‌پور	۹٤/٤٤	٦٥/۳۸	۷۷/۲۷
عبدالملکیان	۸۰/٤۳	۹۷/۳٦	۸۸/۰۹
میانگین	۸۷/٤۳	۸۱/۳۷	۸۲/٦۸

جدول شمارهٔ ۱۱: ماتریس درهم‌ریختگی دسته‌بند بیز ساده پس از حذف ایست‌واژه‌ها روی ویژگی‌های واژگانی

	عبدالملکیان	امین‌پور
عبدالملکیان	۳۷	۹
امین‌پور	۱	۱۷

نتایج به‌دست آمده، پایین‌بودن دقت پیش‌بینی دسته‌بند را نشان می‌دهد. این امر مؤثربودن عدم‌حذف ایست‌واژه‌ها و واژه‌های دستوری در دسته‌بندی شناسایی سبک نویسنده را ثابت می‌کند؛ ازاین‌رو، حذف این واژه‌ها تأثیر منفی در دسته‌بندی اسناد دارد.

سپس، دسته‌بند K نزدیک‌ترین همسایه برای دسته‌بندی اسناد استفاده شده است. آزمایش‌های مختلف با مقادیر متفاوت K روی اسناد آموزش انجام گرفت. در این مرحله، جهت انتخاب K، مقادیر متفاوتی آزمایش شد که در نهایت مقدار آن هشت درنظر گرفته شد. نتایج ارزیابی این دسته‌بند با به‌کارگیری فاصلهٔ کسینوسی و اقلیدسی به‌ترتیب در جداول شمارهٔ ۱۲ و ۱۳ نشان داده شده است.

جدول شمارهٔ ۱۲: دسته‌بند K نزدیک‌ترین همسایه با استفاده از فاصلهٔ کسینوسی روی ویژگی‌های واژگانی

	بازخوانی	دقت	معیار F
امین‌پور	۹۵/۷۷	٦٦/۰۱	۷۸/۱٦
عبدالملکیان	۹۰/۳۳	۹۹/۰۹	۹٤/۵۰
میانگین	۹۳/۰۵	۸۲/۵۵	۸٦/۳۳

جدول شمارهٔ ١٣: دسته‌بند K نزدیکترین همسایه با استفاده از فاصلهٔ اقلیدسی روی ویژگی‌های واژگانی

	بازخوانی	دقت	معیار F
امین‌پور	٨٣/٣٣	٦٠/٠٠	٦٩/٧٦
عبدالملکیان	٧٨/٢٦	٩٢/٣٠	٨٤/٧٠
میانگین	٨٠/٧٩	٧٦/١٥	٧٧/٢٣

همان‌گونه که در جداول فوق مشاهده می‌شود، نتایج دسته‌بند K نزدیک‌ترین همسایه با استفاده از دو فاصلهٔ اقلیدسی و کسینوسی تفاوت نسبتاً قابل‌توجهی در دسته‌بندی نشان می‌دهند؛ بااین‌حال، نتایج با داشتن ویژگی‌های واژگانی قابل‌قبول است. همچنین، استفاده از فاصلهٔ کسینوسی، حدود ١٠ درصد بر کارآیی این دسته‌بند افزوده است.

ویژگی‌های نحوی

در این مرحله، برداری با صد ویژگی تشکیل شد که تمامی آن‌ها، دونگاشت‌های نحوی متداول در هر سند را نشان می‌دهند. در جدول شمارهٔ ١٤، پنج دونگاشت رایج به‌همراه بسامدشان نمایش داده شده است.

جدول شمارهٔ ١٤: پنج دونگاشت پرتکرار از نقش اجزای کلام در پیکره اشعار

	دونگاشت نقش اجزای کلام	وقوع در پیکره
١	N_N	٦٢١٧
٢	P_N	٣٥٠٨
٣	PUNC_N	٢٨٧٦
٤	N_PUNC	٢٧٠٩
٥	CON_N	٢٥١١

در استخراج دونگاشت‌های نحوی، دقت برچسب‌زنی از اهمیت بالایی برخوردار است. به‌طوری‌که، هرچه نتایج به‌دست‌آمده از برچسب‌زن صحت بیشتری داشته باشد، کارآیی ویژگی‌های نحوی در سامانه افزایش می‌یابد. در این پژوهش برای برچسب‌زنی از پیکرهٔ دو میلیونی بی‌جن‌خان به‌عنوان دادهٔ آموزش استفاده شده است. این برچسب‌زنی نحوی با استفاده از مدل مخفی مارکوف[1] طراحی شده است. دقت این برچسب‌زنی با معیار F ، ۸۹ درصد است. در پیکرهٔ بی‌جن‌خان علائم و برچسب‌هایی استفاده شده‌اند که در جدول شمارهٔ ۱۵ نمایش داده شده است (بی‌جن‌خان، ۱۳۸۱).

جدول شمارهٔ ۱۵: علائم و برچسب‌های نحوی نقش اجزای کلام

	برچسب انگلیسی	برچسب فارسی
۱	N	اسم
۲	N_PL	اسم جمع
۳	N_SING_NPR	اسم مفرد خاص
٤	P	حرف‌اضافه
٥	PP	گروه حرف‌اضافه

پس از استخراج ویژگی‌ها و تشکیل بردار، انتساب اسناد آزمایش به شاعران با استفاده از دسته‌بندها، مورد مطالعه قرار گرفت. در این بخش ویژگی‌های واژگانی به روش K نزدیک‌ترین همسایه، برای آزمایش ۱۵ درصد از داده‌های کلی جدا شد و با استفاده از بقیهٔ اسناد به آموزش دسته‌بند پرداخته شد. نتایج ارزیابی دسته‌بندی به روش K نزدیک‌ترین همسایه، با داشتن بردار ویژگی‌های نحوی و کاربرد دو فاصلهٔ اقلیدسی و کسینوسی در جداول شمارهٔ ١٦ و ۱۷ نمایش داده شده است.

1. Hidden Markov Model

جدول شمارهٔ ١٦: نتایج ارزیابی دسته‌بند K نزدیک‌ترین همسایه با استفاده از فاصلهٔ اقلیدسی روی ویژگی‌های نحوی

	بازخوانی	دقت	معیار F
امین‌پور	٦٣/٠٢	٧٢/٨١	٦٧/٥٦
عبدالملکیان	٩١/٠٨	٨٦/٦٦	٨٨/٨٨
میانگین	٧٧/٠٥	٧٩/٧٤	٧٨/١٩

جدول شمارهٔ ١٧: نتایج ارزیابی دسته‌بند K نزدیک‌ترین همسایه با استفاده از فاصلهٔ کسینوسی روی ویژگی‌های نحوی

	بازخوانی	دقت	معیار F
امین‌پور	١٠٠/٠٠	٧٢/٠٠	٨٣/٧٢
عبدالملکیان	٨٤/٧٨	١٠٠/٠٠	٩١/٧٦
میانگین	٩٢/٣٩	٨٦/٠٠	٨٧/٧٤

سپس، در مرحلهٔ بعد، بردار ویژگی‌های نحوی با استفاده از دسته‌بند بیز ساده انجام شد که نتایج ارزیابی این روش و ماتریس ابهام به‌ترتیب در جداول شمارهٔ ١٨ و ١٩ نمایش داده شده است. باتوجه‌به نتایج به‌دست‌آمده در این بخش، به‌نظرمی‌رسد استفاده از ویژگی‌های نحوی در حوزهٔ شناسایی نویسنده منجر به دقت مناسبی در دسته‌بندی و انتساب متن می‌شود.

جدول شمارهٔ ١٨: نتایج ارزیابی دسته‌بند بیز ساده روی ویژگی‌های نحوی

	بازخوانی	دقت	معیار F
امین‌پور	٩٤/٤٤	٧٣/٩١	٨٢/٩٢
عبدالملکیان	٨٦/٩٥	٩٧/٥٦	٩١/٩٥
میانگین	٩٠/٧	٨٥/٧٣	٧٤/٤٤

جدول شمارهٔ ۱۹: نتایج ماتریس درهم‌ریختگی دسته‌بند بیز ساده روی ویژگی‌های نحوی

	عبدالملکیان	امین‌پور
عبدالملکیان	٤٠	٦
امین‌پور	١	١٧

ویژگی‌های حرفی

برای استخراج ویژگی‌های حرفی، ابتدا پیکره به‌عنوان دنباله‌ای از رشته‌های حرفی درنظر گرفته شد و سپس، صد، دونگاشت حرفی متداول استخراج شد. در جدول شمارهٔ ۲۰، پنج دونگاشت پرکاربرد به‌همراه بسامدشان قابل‌مشاهده است.

جدول شمارهٔ ۲۰: پنج دونگاشت پرتکرار حرفی در پیکرهٔ اشعار

	دونگاشت حرفی	بسامد
۱	ه _	۳۷۰۵
۲	ی _	۳۵۸۹
۳	_ ب	۲۸۱٤
٤	ر _	۲۸۰۲
٥	ن _	۲۷٤۳

سپس، بسامد هریک از این ویژگی‌ها در هر سند محاسبه و وارد بردار ویژگی‌های حرفی شد. نتایج دسته‌بندی به روش K نزدیک‌ترین همسایه با استفاده از دو معیار فاصلهٔ اقلیدسی و کسینوسی به‌ترتیب در جداول شمارهٔ ۲۱ و ۲۲ نشان داده شده است.

جدول شمارهٔ ٢١: نتایج ارزیابی دسته‌بند K نزدیک‌ترین همسایه با استفاده از فاصلهٔ اقلیدسی روی ویژگی‌های حرفی

	بازخوانی	دقت	معیار F
امین‌پور	٩٤/٤٤	٦٠/٧١	٧٣/٩١
عبدالملکیان	٧٦/٠٨	٩٧/٢٢	٨٥/٣٦
میانگین	٨٥/٢٦	٧٨/٩٦	٧٩/٦٣

جدول شمارهٔ ٢٢: نتایج ارزیابی دسته‌بند K نزدیک‌ترین همسایه با استفاده از فاصلهٔ کسینوسی روی ویژگی‌های حرفی

	بازخوانی	دقت	معیار F
امین‌پور	٨٣/٣٣	٦٥/٢١	٧٣/١٧
عبدالملکیان	٨٢/٦٠	٩٢/٦٨	٨٧/٣٥
میانگین	٨٢/٩٧	٧٨/٩٥	٨٠/٢٦

استفاده از ویژگی‌های حرفی در مطالعهٔ سبک شاعران امکان بررسی هجاها، وزن و قافیه که از عوامل مهم در متمایز کردن سبک آن‌هاست، در اختیار محققان قرار می‌دهد. روش پایانی در این بخش، استفاده از دسته‌بند بیز ساده است. نتایج ارزیابی و ماتریس ابهام به‌ترتیب در جداول شمارهٔ ٢٣ و ٢٤ قابل‌مشاهده است.

جدول شمارهٔ ٢٣: نتایج ارزیابی دسته‌بند بیز ساده روی ویژگی‌های حرفی

	بازخوانی	دقت	معیار F
امین‌پور	٨٨/٨٨	٦٦/٦٦	٧٦/١٩
عبدالملکیان	٨٢/٦٠	٩٥/٠٠	٨٨/٣٧
میانگین	٨٥/٧٤	٨٠/٨٣	٨٢/٢٨

جدول شمارهٔ ۲٤: نتایج ماتریس درهم‌ریختگی دسته‌بند بیز ساده روی ویژگی‌های حرفی

	عبدالملکیان	امین‌پور
عبدالملکیان	۳۸	۸
امین‌پور	۲	۱٦

باتوجه‌به نتایج ویژگی‌های واژگانی و نحوی، پایین‌بودن تعداد اسناد در نتایج دسته‌بندهای K نزدیک‌ترین همسایه و بیز ساده تأثیر قابل‌توجهی دارد. باتوجه‌به ماتریس ابهام این دو دسته‌بند، تعداد کم اشعار قیصر امین‌پور در پیکره باعث ایجاد دقت پایین‌تری در سامانه شده است.

بحث و نتیجه‌گیری

در این پژوهش سعی شد تا شیوهٔ شناسایی خودکار اشعار دو شاعر دفاع مقدس یعنی قیصر امین‌پور و محمدرضا عبدالملکیان ارزیابی شود؛ به‌همین‌منظور، برخی از ویژگی‌های زبانی اشعار از طریق روش‌های آماری و الگوریتم‌های یادگیری ماشین، استخراج شد. همچنین، شناسایی ویژگی‌های زبانی شاعران با استفاده از تحلیل در سه سطح واژگانی، نحوی و حرفی انجام گرفت. ویژگی‌های واژگانی با استخراج هزار واژه پر کاربرد بر اساس معیار Tf-idf و ویژگی‌های نحوی و حرفی با استخراج صد دونگاشت پرکاربرد در پیکرهٔ اشعار به‌دست آمد.

نتایج ارزیابی دسته‌بند بیز ساده بدون حذف ایست‌واژه‌ها روی ویژگی‌های واژگانی (جدول شمارهٔ ۷) نشان می‌دهد، معیار F برای امین‌پور ۸۹ درصد است؛ بنابراین، سامانهٔ طراحی شده، از میان صد شعر قیصرامین‌پور که در بخش آموزش مشاهده نکرده و برای نخستین مرتبه با آنان روبه‌رو شده است، ۸۹ شعر را به درستی متعلق به قیصر امین‌پور می‌داند. همچنین، سامانه یازده شعر از قیصر امین‌پور را به اشتباه در دستهٔ اشعار عبدالملکیان قرار داده است.

نتایج ماتریس درهم‌ریختگی (جدول شمارهٔ ۸) نشان می‌دهد، از میان ٤٦ شعر عبدالملکیان که سامانهٔ پژوهش برای نخستین مرتبه با آنان روبه‌رو شده، ٤۳ شعر را به‌عنوان اشعار عبدالملکیان دسته‌بندی کرده است. همچنین، سامانه به اشتباه ۳ شعر را به امین‌پور نسبت داده است. به‌این‌ترتیب، از میان ۱۸ شعر امین‌پور که در بخش آزمون در نظر گرفته شده است، سامانه ۲ شعر را به اشتباه از عبدالملکیان و ۱٦ شعر دیگر را به‌درستی از امین‌پور شناسایی کرده است.

باید اذعان داشت در چنین پژوهش‌هایی، کاستی‌هایی نسبت به تحلیل کارشناسان وجود دارد؛ اما، به‌این‌معنی نیست که این روش‌ها ناکارآمد هستند. در بررسی‌های آماری، روش‌های بسیاری برای یادگیری ماشینی با استفاده از الگوریتم‌های مختلف و اهداف مشخص طراحی شده است. هر یک از این روش‌ها، مزایا و کاستی‌های منحصر به فردی دارند. شاید این‌گونه روش‌ها ساده به‌نظر می‌رسند؛ اما، از آنان در تحلیل‌های انسانی استفاده می‌شود. ترکیب چندین روش و استفاده از آن‌ها در سطوح متفاوت زبانی می‌تواند در بررسی‌های ادبی مفید باشد.

در این پژوهش، اشعار دو شاعر دفاع مقدس مورد مطالعه قرار گرفته است. این شاعران در حوزه‌ای مشخص و فضایی یکسان اشعاری سروده‌اند. مطالعهٔ این آثار با استفاده از ویژگی‌های واژگانی مهم، دونگاشت‌های نحوی و حرفی در سطح کل اشعار یا پیکرهٔ آموزش انجام گرفته است. بررسی نهایی این معیارها نشان می‌دهد، ویژگی‌های واژگانی در دسته‌بند بیز ساده از بالاترین دقت برخوردار است؛ ازاین‌رو، دسته‌بند بیز ساده با استفاده از این سطح از ویژگی‌ها قادر به شناسایی سبک شاعران است.

دسته‌بند بیز در تمام ویژگی‌ها، دقتی بالاتر از دسته‌بند K نزدیک‌ترین همسایه دارد. دسته‌بند بیز ساده علاوه بر سرعت بالا و سادگی، دقت بالایی دارد. علت این نتیجه، کم‌بودن پیکرهٔ موردمطالعه است؛ زیرا، بیز ساده در پیکره‌های کم‌حجم، دقت مناسبی دارد.

نتایج به‌دست آمده، تأثیر عدم‌حذف ایست‌واژه‌ها و واژه‌های دستوری در دسته‌بندی شناسایی سبک نویسنده را ثابت می‌کند. حذف این‌گونه واژه‌ها بر دسته‌بندی مبتنی‌بر موضوع و شناسایی سبک نویسنده، تأثیر مثبتی و بر دسته‌بندی اسناد، تأثیر منفی دارد.

دسته‌بند K نزدیک‌ترین همسایه، یکی از روش‌های یادگیری ماشین در دسته‌بندی متون است. در این پژوهش این روش برای شناسایی سبک نویسنده بررسی شد. معیاری که برای محاسبهٔ فاصلهٔ ویژگی‌ها روی بردارها در نظر گرفته می‌شود، اقلیدسی است؛ اما، در پردازش متن فاصلهٔ کسینوسی لحاظ می‌شود. در این پژوهش هر دو فاصله موردمطالعه قرار گرفت و اختلاف آنان تعیین شد. به‌کارگیری فاصلهٔ کسینوسی دقت دسته‌بند را حدود ١٠ درصد افزایش می‌دهد. این نتیجه نشان می‌دهد، فاصلهٔ کسینوسی در حالت‌ها و بخش‌های مختلف پردازش متن برای شناسایی نویسنده مفید است.

کتابنامه

منابع فارسی

آذین، زهرا (۱۳۹۲). *شناسایی خودکار شاعران شعر نو با استفاده از ویژگی‌های زبانی*، پایان‌نامهٔ کارشناسی‌ارشد زبان‌شناسی، تهران: دانشگاه صنعتی شریف، دانشکدهٔ زبان‌شناسی.

استاجی، اعظم (۱۳۸۷). تشخیص مؤلف متون ادبی و قانونی، بحثی در زبان‌شناسی قانونی، *نشریهٔ زبان و زبان‌شناسی*، دورهٔ چهارم، شمارهٔ ۲، ۳۲-۱۵.

بی‌جن‌خان، محمود (۱۳۸۱). *طرح مدل‌سازی زبان فارسی مرحله دوم*، آزمایشگاه گروه زبان شناسی دانشکدهٔ ادبیات و علوم انسانی دانشگاه تهران.

سرایی، محمدحسین و شاهقلیان، آذر (۱۳۸۹). کاوش متون فارسی بر مبنای روش طبقه‌بندی، *نشریهٔ انجمن کامپیوتر ایران*، جلد هشتم، شمارهٔ ۱، ۱۳-۸.

سمیعی گیلانی، احمد (۱۳۸٤). سبک، *نشریهٔ نامهٔ فرهنگستان*، شمارهٔ ٦، ۱۰۲-۸٦.

سمیعی گیلانی، احمد (۱۳۸٦). مبانی سبک‌شناسی شعر، *نشریهٔ ادب پژوهش*، شمارهٔ ۲، ٤۹-۷٦.

شاهمیری، امیرشهاب و مطش بروجردی، محمدرضا (۱۳۸٦). تعیین شاعر به کمک روش‌های یادگیری ماشین، *مجموعه مقالات سومین کنفرانس بین‌المللی فناوری و دانش*، مشهد: دانشگاه فردوسی مشهد.

فرهمندپور، زینب؛ نیک‌مهر، هومن؛ منصوری‌زاده، محرم و طبیب‌زاده‌قمصری، امید (۱۳۹۱). یک سیستم نوین هوشمند تشخیص هویت نویسندهٔ فارسی‌زبان براساس سبک نوشتاری، *نشریهٔ محاسبات نرم*، دورهٔ اول، شمارهٔ ۲، ۳۵-۲٦.

کامیار، حسین (۱۳۹۰). *روش جدید وزن‌دهی معنایی به کلمات در کاربردهای پردازش متن*، پایان‌نامهٔ کارشناسی ارشد مهندسی کامپیوتر، مشهد: دانشگاه فردوسی مشهد، دانشکدهٔ مهندسی.

مجیری، محمدمهدی و مینایی، بهروز (۱۳۸۷). تشخیص وزن عروضی اشعار فارسی: کاربرد جدیدی از متن کاوی، *دومین کنفرانس داده‌کاوی ایران*، تهران: دانشگاه صنعتی امیرکبیر.

یاحقی، محمدجعفر و ایزانلو، علی (۱۳۸۵). سبک سنجی، نقد و بررسی شیوهٔ آماری کیوسام در انتساب یک اثر، *نشریهٔ زبان و ادبیات فارسی دانشگاه خوارزمی*، دورهٔ چهاردهم، شمارهٔ ۵۳-۵۲، ۱۹۰-۱۵۱.

منابع انگلیسی

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist group web forum messages, *IEEE Intelligent Systems*, 20 (5), 67-75.

Baayen, H; Halteren, H. V; Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. *JADT 2002: Sixth International Conference on Textual Data Statistical Analysis*, 29-37.

Jurafsky, D., & Martin, J. H. (2006). *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*, United State: Prentice Hall.

Mechti, S; Jaoua, M; Faiz, R; Belguith, L. H., & Bsir, B. (2015). On the Empirical Evaluation of Author Identification Hybrid Method Notebook for PAN at CLEF 2015. *CLEF 2015 Evaluation Labs and Workshop*, France: Toulouse.

Olsson, J. (2004). *Forensic linguistics, an introduction to language, crim and law*. London, New York: Continuum.

Patton, J. M., & Can, F. (2004). A stylometric analysis of Yasar Kemal's Ince Memed tetralogy, *Computers and the Humanities*, 38 (4), 457-467.

Salton, G. B. (1988). Term-weighting approaches in automatic text Retrieval, *Information Processing & Management*, 24 (5), 513-523.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods, *Journal of the American Society for information Science and Technology*, 60 (3), 538-556.