علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
جلد ۴ شماره ۳ (الف)، ۱۳۸۵
صفحات ۱۷-۲۶
مقاله عادی

شناسایی اشعار شاهنامهٔ فردوسی به کمک شبکه عصبی مصنوعی

امیرشهاب شاهمیری سعید شیری قیداری رسول دژکام

دانشکدهٔ مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

شناسایی شاعر[۱] یکی از گرایش‌های دانش شناسایی نویسنده[۲] است که از مسایل مهم در پردازش زبان‌های طبیعی[۳] به‌شمار می‌رود. این نوشتار دست‌آورد پژوهشی است که هدف آن تشخیص هوشمند اشعار شاهنامهٔ فردوسی به‌کمک شبکه عصبی مصنوعی از نوع پیشخور با پس‌انتشار[۴] بوده است. برای سنجش کارایی شبکه، ویژگی‌های اشعار به‌روش درخت تصمیم[۵] نیز آزموده و نتایج آنها با خروجی شبکهٔ عصبی مقایسه شده است. در هنگام استخراج ویژگی‌های گوناگون اشعار برای ارزیابی سیستم، کوشش شده تا شیوهٔ تفکر انسان برای تشخیص اشعار شاعران شبیه‌سازی گردد و بدین منظور، ۶۴ گونه دادهٔ ورودی از ۲۲ ویژگی گوناگون از اشعار ایرانی استخراج و در سه دستهٔ فیزیکی، مفهومی و آوایی دسته‌بندی شده است. همچنین برای سنجش و ارزیابی عملی نمونه‌ها، یک پایگاه داده واژگان با فیلدهای ویژه طراحی و پیاده‌سازی گردیده که امکان مقیاس‌پذیری کمّی اشعار را فراهم آورده است.

پس از آموزش و آزمایش نمونه‌ها در هر دو روش، مشخص شد که دسته‌بندی اشعار آزمایشی در دو دستهٔ اشعار شاهنامه‌ای و غیرشاهنامه‌ای به‌کمک درخت تصمیم تا ۸۵/۵٪ درستی و به‌روش شبکهٔ عصبی مصنوعی تا ۱۰۰٪ درستی را به‌همراه داشته است. همچنین برای کاهش خودکار و تعیین ارزش هر یک از ویژگی‌های استخراج شده، یک شبکهٔ عصبی پرسپترون تک‌لایه[۶] به‌کار گرفته شد که به‌کمک آن، با حفظ دقت مطلوب، تعداد ویژگی‌های مورد نیاز به‌مقدار قابل توجهی کاهش یافت.

کلمات کلیدی: شناسایی شاعر، شناسایی نویسنده، طبقه‌بندی متن، شبکه عصبی مصنوعی، دسته‌بندی، پردازش زبان طبیعی، هوش مصنوعی، شاهنامه فردوسی، درخت تصمیم گیری.

۱- مقدمه

شناسایی نویسنده از روی نثر و سبک نوشتاری یا به‌عبارت دیگر ویژگی‌های نهفته در متون نوشته شده توسط وی، یکی از مباحث به‌نسبت جدید در زمینهٔ هوش مصنوعی و پردازش زبان طبیعی به‌شمار می‌رود. در این مبحث کوشش می‌شود تا با استخراج ویژگی‌هایی از درون متن، و پردازش و تحلیل آن به‌کمک انواع روش‌های هوش مصنوعی، نویسندهٔ متن شناسایی گردد. شناسایی شاعر نیز یکی از شاخه‌های تخصصی شناسایی نویسنده است که در آن می‌توان افزون بر ویژگی‌های مربوط به رشته حروف، ویژگی‌های دیگر مانند وزن، لحن و آهنگ جملات را نیز پردازش کرد.

ایدهٔ شناسایی نویسنده از مبحث طبقه‌بندی متن[۷] ــ که خود سرفصلی از دانش فهم زبان‌های طبیعی است[۱] ــ سرچشمه می‌گیرد که در آن کوشش می‌شود تا با تجزیه و تحلیل واژگان، دستورزبان و مفهوم یک جمله و با کمک گرفتن از دانش مربوط به واژگان، معنای آن جمله برای ماشین درک گردد[۲]. البته برای شناسایی نویسنده و شاعر به‌کارگیری تکنیک‌های مورد استفاده در پردازش زبان‌های طبیعی الزام‌آور نیست، اما از آنجا که عکس‌العمل درست ماشین به هر یک از ویژگی‌های سیگنال، آوا، واژه، دستور، معنا یا مفهوم جمله، درک زبان طبیعی به‌شمار می‌آید[۳]، شناسایی شاعر نیز در این شاخه از دانش هوش مصنوعی می‌گنجد.

در یکی از نخستین پژوهش‌های انجام شده در زمینهٔ دسته‌بندی متن در سال ۱۹۶۱ با توجه به محتوای عنوان، متون را برپایهٔ رخداد واژگان کلیدی برگزیده در آنها، به دسته‌های عناوین متعلق می‌ساخت[۴]. در سال ۱۹۹۱ "لی" و همکارانش برای طبقه‌بندی متن با کاربرد ویژه در دسته‌بندی فرم‌های تلکس در بانک‌ها از ترکیب روش دسته‌بندی شبکهٔ عصبی مصنوعی با برخی اطلاعات غیرآماری، از سیستم استخراج ویژگی بهره جست و به دقت درستی ۹۰ درصد رسید[۵]. در سال ۱۹۹۲ "ژاکوبز" برای دسته‌بندی اخبار، برخی روش‌های پردازش زبان طبیعی بر پایهٔ پایگاه دانش[۸] را به روش‌های آماری افزود و نتیجه گرفت که هر چند که روش‌های آماری در پیاده‌سازی ساده هستند و نیاز کمی به دستکاری کاربر دارند، اما افزودن پیش‌پردازشی به روش‌های پردازش زبان طبیعی، دقت و سرعت کار را

بسیار افزایش می‌دهند[۶]. "گوتنر" و همکارانش در سال ۱۹۹۳ به‌کمک شبکهٔ عصبی مصنوعی سیستمی را برای دسته‌بندی نامه‌های الکترونیکی پدید آوردند که از یک فرهنگ واژگان بهره می‌برد و قابلیت یادگیری داشت. دسته‌بندی با این روش با دقت ۷۹/۱٪ درستی، نزدیک به دقت انسان با ۷۹/۴٪ درستی بوده است[۷]. "ریلوف" در سال ۱۹۹۴ برای اثبات این که استخراج ویژگی‌ها از متن، مهم‌ترین بخش از فرایند طبقه‌بندی متن است، سه روش را ــ به‌نام‌های نشان‌های ارتباطی، که عبارات زبانی را به‌کار می‌گیرد، نشان‌های ارتباطی تکامل‌یافته[۹] که از عبارت‌های زبانی و محتوای متن استفاده می‌کند و الگوریتم دسته‌بندی متن موردگرا که تکه‌های بزرگ متن را به‌کار می‌گیرد ــ باهم آزمود و نشان داد که روش‌های دوم و سوم دقت بسیار بالایی دارند[۸]. در سال ۱۹۹۵ "وینر" و همکارانش شبکهٔ عصبی غیرخطی را برای تعیین عنوان متون به‌کار گرفت. وی در این پژوهش دو روش گزینش اصطلاح[۱۰] و شاخص‌گذاری معنای پنهان[۱۱] را آزمود که نتایج آنها اندکی بهتر از دیگر روش‌ها بوده است[۹]. در سال ۱۹۹۶ هم‌زمان و در پژوهشی همانند، "یواخیمز" نیز تعداد ۲۰۰۰۰ مقاله مختلف از ۲۰ شبکهٔ خبری را به‌کمک پایگاه دادهٔ واژگانی با ۳۸۵۰۰ واژه با همان الگوریتم ساده بیزی دسته‌بندی کرد که تا ۸۹٪ دقت داشت[۱۰]. در سال ۱۹۹۶ "هاناور" ارتباطات میان ویژگی‌هایی به‌خصوص از متون شعری را ــ مانند اطلاعات آوایی، پیش‌زمینهٔ ادبی خوانندگان و اطلاعات مربوط به فن نوشتاری اشعار ــ برای دسته‌بندی آنها به‌کار گرفت[۱۱]. "مرکل" در سال ۱۹۹۸ شبکهٔ عصبی خود سازمانده سلسله‌مراتبی[۱۲] را ــ که ترکیبی از شبکه‌های خود سازمانده است ــ برای دسته‌بندی متون به‌کار گرفت که هر شبکه یکی از ویژگی‌های متن را می‌آزمود و بدین ترتیب بر سرعت کار افزود[۱۲].

در سال ۱۹۹۹ "هورن" و همکارانش توانستند اشعار ۳ شاعر هلندی را بر پایهٔ آرایش رشته حروف[۱۳] و با روش‌های دسته‌بندی مانند kمین نزدیکترین همسایه[۱۴]، شبکهٔ عصبی مصنوعی و دسته‌بندی‌کننده ساده بیزی با ۸۰-۷۰٪ دقت برای دسته‌بندی دو شاعر و ۷۰٪ دقت برای دسته‌بندی سه شاعر از هم تشخیص دهند[۱۳]. در سال ۲۰۰۰ "استاتاماتوس" و همکارانش توانستند بر اساس محتوا و شیوهٔ نگارش، شناسایی نویسنده و گروه متون به‌زبان یونانی جدید موجود بر روی اینترنت را به‌انجام رسانند[۱۴]. در سال ۲۰۰۱ "آکیکو" برخی از فنون احتمالاتی زبانشناسی در پردازش زبان طبیعی را برای بهبود عملکرد روش‌های خودکار طبقه‌بندی متن به‌کار گرفت[۱۵]. در سال ۲۰۰۱ "پاولوف" و همکارانش ادعا کردند که می‌توانند با الگوریتم‌هایی مانند DFA[۱۵] و WTMM[۱۶] ــ که تعداد، افزایش و کاهش حروف را تحلیل می‌کنند[۱۷] ــ رشته‌های طولانی از واژگان را از نظر کمی مقیاس‌گذاری کنند[۱۶]. در سال ۲۰۰۳ "دوول" و همکارانش گزارش کردند که توانسته‌اند با ماشین بردار پشتیبان سبک[۱۸] جنسیت و زمینهٔ گفتاری نویسندگان نامه‌های الکترونیکی مورد آزمایش خود را بر پایهٔ تعداد حروف و واژگان، نمادهای سجاوندی و ساختار نامه، با نزدیک به حدود۷۰٪ دقت تشخیص دهند[۱۷].

در پژوهشی که گزارش و نتایج آن را پیش رو دارید، کوشش شده است تا اشعار حکیم ابوالقاسم فردوسی توسی، حماسه‌سرای بزرگ ایرانی، از متون نظم و نثر دیگر نویسندگان و شاعران بازشناسی شود. بدین منظور برای استخراج ویژگی‌ها، بیش از ۵۰ پارامتر ویژگی از شعر فارسی به‌طور عام و شاهنامهٔ فردوسی به‌طور خاص ــ که در سه گروه فیزیکی، مفهومی و آوایی دسته‌بندی می‌شود ــ به سیستم پیشنهاد گردیده، که البته برخی از آنها به‌کمک یک شبکهٔ عصبی پرسپترون تک‌لایه و تحلیل‌های آماری، کاهش ویژگی[۱۹] می‌یابند.

همچنین برای آزمون عملی نتایج این پژوهش در برنامه‌ای نرم‌افزاری که طراحی و پیاده‌سازی گردیده، از یک پایگاه دادهٔ ویژه به‌عنوان "فرهنگ واژگان" استفاده شده است. از آنجا که برای آزمون سیستم و یافتن و تنظیم ویژگی‌ها و پارامترهای شعری باید سیستم با شعرهای شاهنامه آزمایش می‌شد و نیز از آنجا که اختلافاتی بین ادیبان و فردوسی‌شناسان بر سر تایید اصل بودن برخی از ابیات شاهنامه وجود دارد، روش کار یکی از این استادان برای دسته‌بندی اشعار فردوسی در دو مجموعهٔ اصل و افزوده برگزیده شده است. برای نمونه‌های منفی نیز اشعاری از دیگر شاعران و نویسندگان ــ مانند حافظ، سعدی، خیامی و مولوی ــ انتخاب شده است.

در ادامهٔ این مقاله نخست در بخش ۲ به برخی از پیش‌نیازهای آزمون عملی پروژه، مانند پایگاه دادهٔ ویژهٔ واژگان فارسی و انتخاب نمونه‌های آموزشی و آزمایشی پرداخته می‌شود. سپس در بخش ۳ استخراج ویژگی‌ها و کاهش آنها و در بخش ۴ پردازش نمونه‌ها برای آموزش و آزمایش شرح داده می‌شود و سرانجام در بخش ۵ نتایج ارزیابی خواهد شد و اهداف آینده بیان خواهد گردید.

۲- پیش‌نیازهای طرح

از آنجا که تاکنون پژوهش‌های عملی چندانی در زمینهٔ شناسایی نویسنده و شاعر در زبان فارسی انجام نپذیرفته و حتی تحقیقات مربوط به پردازش متن فارسی نیز به نتایج عملی چندانی دست نیافته است، ناگزیر برای به‌انجام رسانیدن این طرح، به چندین طرح جانبی دیگر نیز پرداخته شد و همچنین دانش‌هایی مرتبط کسب گردید.

۲-۱ پایگاه داده واژگان فارسی

نخستین مشکل این پژوهش فقدان یک پایگاه داده جامع از واژگان فارسی بود؛ به‌گونه‌ای که در آن پایگاه داده، افزون بر اطلاعات معمول، آوای واژگان نیز به‌ثبت رسیده باشد و ویژگی‌های دیگر ــ مانند ریشهٔ زبانی و نقش‌های دستوری ــ هم مشخص شده باشند. این کمبود، موجب ساخت چنین پایگاه داده‌ای در گام نخست پژوهش شد که البته ویژگی‌های هر واژه در هنگام ورود آن به‌سیستم (زمان آموزش یا آزمایش) وارد می‌گردید (جدول ۱).

جدول ۱- پایگاه داده جامع زبان فارسی

واژه	نقش	آوا	کاربردها	ریشه	...
شبکه	اسم، مفرد، غایب	ekabaS	پزشکی / کامپیوتر / الکترونیک / ...	عربی	...
شما	ضمیر، جمع، حاضر	AmoS	عمومی	فارسی	...
پارامتر	اسم، مفرد، غایب	rtemArAp	علوم پایه / علوم فنی / انسانی / ...	لاتین	...

از آنجا که برای پردازش یک شعر بیشتر به آوای درست آن توجه می‌شود[۲۳] و خط امروزی فارسی در نمایش درست و کامل آوای واژگان ناتوان است، برای نمایش آوای هر واژه، خط کهن ایرانی به‌نام خط اوستایی به‌کار گرفته شد؛ زیرا این خط با داشتن ۴۴ (تا ۶۰ در نگارش‌های گوناگون) حرف چگونگی گویش هر واژه‌ای که از زبان آریایی‌زبانان گفته می‌شود را به‌خوبی نمایش می‌دهد[۲۸-۲۴]. البته در این طرح برخی از حروف با اندکی تغییر به‌کار گرفته شده است. (شرح در پیوست ۱)

۲-۲ انتخاب نمونه‌های آموزشی

انتخاب نمونه‌های آموزشی در نگاه اول ساده به نظر می‌رسد و گمان می‌رود که نمونه‌های مثبت و منفی را می‌توان از شاهنامهٔ فردوسی و دیوان اشعار دیگر شاعران برگزید. اما حقیقت این است که بر سر اصل بودن برخی از اشعار فردوسی و نیز بسیاری دیگر از شاعران، اختلاف نظرهای عمیقی میان صاحب‌نظران و شعرشناسان وجود دارد.

به‌طور عام شکی نیست که منشیانی که در طول تاریخ به‌کار رونویسی اشعار شاعران برای افزایش تعداد نسخ آنها مشغول بوده‌اند، گاه چند بیتی هم از سر خوش‌ذوقی یا مسایل دیگر بدان افزوده‌اند و گاه نیز برخی از واژگان آنها را ــ به‌عمد یا به‌سهو ــ با اختلاف نوشته‌اند و این امر موجب اختلاف عقیده و نظر بین ادیبان گشته است[۲۹].

دربارهٔ اشعار فردوسی نیز بیشتر شعرشناسان بر "افزوده" بودن برخی از اشعار آن اتفاق نظر دارند. برای نمونه سراسر بخشی را که در زمینهٔ مدح خلفای عرب بوده و نیز بخشی که به مراودات با محمود غزنوی پرداخته است، سخن فردوسی نمی‌دانند و چنین می‌گویند که این اشعار بعدها به‌دلایل سیاسی و دینی، به‌دست دیگران به شاهنامه افزوده گردیده است[۲۹و۳۰]؛ اما دربارهٔ بسیاری دیگر از ابیات شاهنامه، چنین اتفاق نظری نیست و حتی در چگونگی نگارش یا خواندن بسیاری از واژگان نیز اختلاف عقیده وجود دارد.

در این پژوهش، برای شناخت شعر اصل فردوسی به‌منظور گردآوری نمونه‌های آموزشی مثبت شبکه، روش استاد فریدون جنیدی به‌کار گرفته شده است. از آنجا که این مکتب سختگیرانه‌ترین نظر را نسبت به اشعار اصل شاهنامه دارد و نزدیک به ۴۰۰۰۰ از ۶۰۰۰۰ بیت شاهنامه را افزوده می‌داند[۲۹]، برای شناخت اشعار اصل برگزیده شده است؛ زیرا با پیروی از سبک طبقه‌بندی این مکتب با قطعیت بیشتری می‌توان گفت که شعری که این مکتب اصل می‌داند، به‌واقع اصل هم هست و شاهنامه‌شناسان دیگر هم آن را اصل می‌دانند.

برای نمونه‌های آزمایشی منفی نیز اشعار دیگر شاعران ایرانی، به‌ویژه اشعار حافظ شیرازی، خیامی نیشابوری، سعدی شیرازی و مولوی بلخی استفاده شده است. البته باید توجه داشت که هدف این پژوهش شناخت اشعار شاهنامه از دیگر اشعار زبان فارسی بوده است؛ نه شناخت اشعار اصل و افزودهٔ شاهنامه. برای اجرای عملی طرح، ۴۰۰ بیت از اشعار اصل فردوسی و ۴۰۰ بیت از اشعار دیگر شاعران به‌عنوان نمونه‌های آموزشی به سیستم وارد شده و این نمونه‌ها بارها با یکدیگر جابه‌جا و آزمایش‌ها تکرار گردیده، که البته نتایج آنها تغییرات چندانی را در دقت شناسایی شاعر نشان نداده است.

۲-۳ دیگر پیش‌نیازهای علمی

روشن است که بررسی هر گونه مسالهٔ علمی به شناخت شایسته‌ای از آن مساله نیاز و بستگی دارد. از این‌رو برای انجام این پژوهش، نگارندگان به‌ناچار توانایی‌های زیر را در حد نیاز و توان خود به‌دست آورده‌اند:

۱. آشنایی با فن شعرشناسی (عروض و قافیه)
۲. آشنایی با دانش واژه‌شناسی
۳. شناخت تاریخی دوران فردوسی و حماسه‌هایش و نیز آشنایی با تاریخ دیگر شاعران
۴. آشنایی با خط‌های باستانی ایرانی.

۳- استخراج ویژگی‌ها

بنیاد این پژوهش در تشخیص اشعار فارسی بر این اصل نهاده شده که بی‌شک اشعار فارسی پارامترهایی برای تمایز از هم و در نتیجه دسته‌بندی دارد و این نکته، همان سبک‌های گوناگون شعری است که شاعران گوناگون از آنها پیروی می‌کنند و بنا بر آنهاست که فارسی‌زبانان می‌توانند اشعار شاعران را ــ متناسب با اطلاعاتشان از ادب فارسی ــ از یکدیگر باز شناسند. از این‌رو پس از بررسی‌های اولیه، ۲۴ ویژگی آشکار و پیشنهاد شد که در سه دسته طبقه‌بندی گردد.

۳-۱ ویژگی‌های فیزیکی

این گروه از ویژگی‌ها به ماهیت فیزیکی واژگان فارسی و آرایش و ترتیب حروف می‌پردازند انواع آنها به‌همراه مقادیرشان برای دو بیت زیر بدین شرح است:

"*روز وصل دوستداران یاد باد / یاد باد آن روزگاران یاد باد*" : حافظ (۱)

"*منم گفت با فره ایزدی / همم شهریاری همم موبدی*" : فردوسی (۲)

۱- ردیف‌دار بودن بیت: بیت (۱) دارای ردیف (یاد باد) است و بیت (۲) ردیف ندارد. این مشخصه را می‌توان به‌عنوان یک ویژگی در نظر گرفت.

۲- تعداد واژگان ردیف: یعنی تعداد واژگانی که به‌صورت ردیف در پایان دو مصراع یک بیت تکرار می‌شود. برای نمونه بیت (۱) دارای ۲ واژهٔ ردیف "یاد" و "باد" است و بیت (۲) هیچ واژهٔ ردیفی ندارد.

۳- تعداد حروف نخستین واژهٔ ردیف: تعداد حروف اولین واژهٔ ردیف؛ در صورتی که وجود داشته باشد. برای نمونه در بیت (۱) نخستین واژهٔ ردیف "یاد" دارای ۳ حرف است و بیت (۲) واژهٔ ردیف ندارد و بنابرین طول آن صفر است.

۴- هم‌قافیه بودن مصراع‌ها: بسته به قالب شعری، دو مصراع برخی از ابیات هم-قافیه است و برخی نیز نیست و این ویژگی را می‌توان به‌عنوان یک پارامتر ارزیابی در نظر گرفت. اشعار شاهنامه همواره هم‌قافیه است؛ اما دیگر شاعران هر دو گونه را سروده‌اند. هر دو بیت (۱) و (۲) هم‌قافیه است.

۵- تعداد حروف قافیهٔ مصراع: در بیت (۲) این مقدار برای واژهٔ "ایزدی" برابر با ۵ و برای واژهٔ "موبدی" برابر با ۵ است.

۶- تفاضل تعداد حروف قافیه‌ها: قدر مطلق این مقدار برای بیت (۲) برابر با صفر است.

۷- تعداد حروف هم‌قافیه: این مقدار در بیت (۲) برای حروف قافیه "دی" برابر با ۲ است.

۸- تعداد واژگان مصراع‌ها: در بیت (۱) این مقدار برای مصراع نخست برابر با ۵ و برای مصرع دوم برابر با ۶ است.

۹- تفاضل تعداد واژگان مصراع‌ها: قدر مطلق این مقدار برای بیت (۱) برابر با ۱ است.

۱۰- تعداد حروف مصراع‌ها: در بیت (۱) تعداد حروف مصراع نخست برابر با ۲۱ و در مصراع دوم برابر با ۲۲ است.

۱۱- تفاضل تعداد حروف مصراع‌ها: قدر مطلق این مقدار برای بیت (۱) برابر با ۱ است.

۱۲- نوع حروف به‌کار رفته در بیت: انتظار می‌رود که کاربرد برخی از حروف ــ مانند "پ"، "چ"، "ژ" و "گ" ــ در ابیات شاهنامه که کوشش شده در آن واژگان اصیل ایرانی به‌کار رود، بیشتر از دیگر متون دیده شود و همین‌طور حروف عربی ــ مانند "ص"، "ض"، و "ظ" ــ کمتر به‌چشم آید.

جدول ۲ نسبت تعداد تکرار حروف الفبا بر کل حروف موجود در اشعار آموزشی را نشان می‌دهد.

۱۳- ترتیب حروف: اگر هر حرف فارسی را برابر با یک عدد (مانند ۰ تا ۳۲) متناظر کنیم، ترتیب و توالی این اعداد می‌تواند معیاری برای دسته‌بندی باشد. برای نمونه، در مصراع نخست بیت (۲) داریم:

۲۷-۲۸-۲۷-۲۵-۲۲-۳-۱-۰-۲۲-۱۱-۳۰-۰-۳۱-۱۲-۱۰-۳۱

همچنین می‌توان به‌جای ترتیب حروف ترتیب آوای شعر (صامت‌ها و مصوت‌ها) را به‌عنوان ویژگی در نظر گرفت. بیشتر پژوهش‌های انجام شده برای دسته‌بندی اشعار شاعران زبان‌های اروپایی بر این پایه بوده است[۱۳].

جدول ۲- نسبت تعداد حروف الفبا بر کل حروف در اشعار آموزشی (مقادیر به درصد)

ردیف	حرف	دیگران	فردوسی	ردیف	حرف	دیگران	فردوسی
۱	ا	۱۴/۰۸	۱۴/۰۳	۱۷	ص	۰/۴۲	۰/۰۳
۲	ب	۴/۵۰	۵/۵۴	۱۸	ض	۰/۰۸	۰/۱۳
۳	پ	۰/۶۲	۱/۳۱	۱۹	ط	۰/۶۳	۰/۰۰
۴	ت	۳/۹۲	۴/۴۱	۲۰	ظ	۰/۰۷	۰/۰۰
۵	ث	۰/۰۲	۰/۱۰	۲۱	ع	۰/۷۵	۰/۰۳
۶	ج	۰/۵۵	۱/۴۴	۲۲	غ	۰/۷۵	۰/۱۰
۷	چ	۰/۵۲	۰/۸۸	۲۳	ف	۱/۴۵	۱/۱۸
۸	ح	۰/۸۴	۰/۱۳	۲۴	ق	۰/۹۳	۰/۰۳
۹	خ	۱/۷۷	۱/۷۹	۲۵	ک	۲/۸۱	۳/۰۲
۱۰	د	۹/۴۰	۷/۹۸	۲۶	گ	۱/۲۴	۲/۵۴
۱۱	ذ	۰/۰۲	۰/۰۵	۲۷	ل	۲/۷۳	۰/۶۸
۱۲	ر	۸/۸۹	۹/۷۰	۲۸	م	۶/۹۸	۳/۸۵
۱۳	ز	۲/۱۸	۳/۰۰	۲۹	ن	۸/۱۵	۸/۸۹
۱۴	ژ	۰/۰۱	۰/۱۳	۳۰	و	۵/۵۵	۷/۰۸
۱۵	س	۳/۳۴	۳/۰۲	۳۱	ه	۵/۴۲	۷/۳۳
۱۶	ش	۳/۳۸	۳/۷۰	۳۲	ی	۸/۰۲	۷/۹۳

با وجود آن‌که این ویژگی در معدود پژوهش‌های انجام شده در این زمینه در زبان‌های دیگر ــ مانند انگلیسی و هلندی ــ مورد توجه بوده است، اما با توجه به این‌که محاسبات بالا و بازدهی کمی داشته و نیز در شعر پارسی ویژگی‌های به‌مراتب بهتری برای دسته‌بندی ابیات موجود است، در این پژوهش مورد استفاده قرار نگرفته است.

۳-۲ ویژگی‌های مفهومی

این دسته از ویژگی‌ها به معنی و مفهوم و ماهیت تاریخی واژه می‌پردازند و انواع آن بدین شرح است:

۱- واژگان ناایرانی: پس از ورود اسلام به ایران بسیاری از واژگان زبان‌های سامی در قالب زبان عربی به زبان فارسی راه یافت[۳۱] که امروزه آنها را واژگان با ریشهٔ عربی می‌شناسند. همچنین از سدهٔ چهارم با نیرو یافتن مغولان و ترکان در همسایگی و درون ایران، کم‌کم برخی از واژگان مغولی و ترکی نیز به فارسی راه یافت[۳۲]. در یکی-دو سدهٔ گذشته نیز با پیشرفت اروپاییان، واژگان آنها نیز ــ به‌ویژه از زبان‌های فرانسه و انگلیسی و در زمینه‌های علمی و فنی ــ به فارسی وارد شد[۳۳]. روشن است که در شاهنامهٔ فردوسی واژگان ترکی/مغولی و اروپایی نایاب است و واژگان عربی در شاهنامه نیز ــ به‌دلیل روحیهٔ میهن‌پرستی و خواست فردوسی به پالایش زبان فارسی از واژگان عربی ــ کمیاب است[۳۸- ۳۴].

از این‌رو تعداد واژگان زبان‌های دیگر، هر یک به‌طور جداگانه، به‌عنوان معیاری برای سنجش شاهنامه‌ای بودن شعر برگزیده شده است. برای نمونه در بیت (۱) واژهٔ "وصل" از زبان عربی آمده، اما بیت (۲) واژهٔ ناایرانی ندارد.

۲- واژگان ویژه: بسیاری از واژگانی که در شاهنامه به‌کار رفته است، مفهوم یا آوایی حماسی و رزمی دارد و نام‌ها یا واژگان کهن ایرانی در شاهنامه بسیار است[۳۶] و این نکته از مهم‌ترین ویژگی‌های اشعار فردوسی به‌شمار می‌آید اما در متون پس از وی کمتر مورد توجه بوده است[۲۸].

با تهیهٔ یک «واژه‌نامهٔ ویژه» کوچک برای هر شاعر یا دیوان شعری، می‌توان نزدیکی و تعلق یک نمونه به هر دسته را بهتر سنجید. برای نمونه در شاهنامهٔ فردوسی، واژگان کهن ایرانی، مانند «گوی»، «گُرد» و «دد»، نام‌های کهن ایرانی، مانند «رستم»، «گودرز» و «کاووس»، و واژگان با مفهوم ملی و حماسی، مانند «شاه»، «ایران» و «سپاه» بیشتر به‌کار می‌روند، اما در اشعار حافظ واژگان عاشقانه ــ مانند «یار»، «عشق» و «می» ــ بیشتر دیده می‌شود. جدول ۳ واژگان برگزیده بر پایهٔ تواتر در اشعار هر شاعر را نشان می‌دهد.[۲۰]

جدول ۳- واژگان ویژه (فرهنگ واژگان) اشعار شاهنامهٔ فردوسی

ردیف	واژه	تکرار	ردیف	واژه	تکرار
۱	شاه	۵۴۴۲	۱۱	باد	۱۹۹۶
۲	اندر	۴۴۲۸	۱۲	لشکر / لشگر	۱۹۷۶
۳	گرد	۳۶۶۴	۱۳	شهر	۱۹۶۲
۴	کار	۳۰۷۰	۱۴	بد	۱۹۶۰
۵	روز	۲۹۵۲	۱۵	سخن	۱۸۶۹
۶	پیش	۲۹۳۶	۱۶	دین	۱۸۵۲
۷	سپاه / سپه	۲۸۲۵	۱۷	راه	۱۸۳۴
۸	دست	۲۶۲۱	۱۸	سپه	۱۷۴۳
۹	دل	۲۳۸۲	۱۹	ایران	۱۶۷۴
۱۰	شید	۲۲۹۷	۲۰	گوی	۱۶۱۲

مقدار عددی به‌کار رفته در روش‌های دسته‌بندی، برای واژگان موجود در واژنامهٔ فردوسی با ردیف i و واژگان موجود در واژه‌نامهٔ دیگر شاعران با ردیف j از رابطهٔ زیر به‌دست آمده است:

$$\alpha = \sum_{i=1}^{n_f}(Max_f - i) - \sum_{j=1}^{n_{\sim f}}(Max_{\sim f} - j) \qquad (۱)$$

که در آن n_f حداکثر تعداد واژه در واژنامهٔ فردوسی، $n_{\sim f}$ حداکثر تعداد واژه در واژه‌نامهٔ دیگر شاعران و Max حداکثر تعداد واژه در دو واژه‌نامهٔ فردوسی و دیگران است.

۳- بررسی مفهوم شعر: این ویژگی به مفاهیم موجود در شاهنامه می‌پردازد. برای نمونه مفهوم عشق زمینی و حتی عرفانی ــ که در اشعار بسیاری از شاعران بزرگ ایرانی مانند مولوی و حافظ پرشمار است ــ کمتر در شاهنامه دیده می‌شود و در عوض عشق به میهن، از خود گذشتگی و دادگری از ویژگی‌های ابیات شاهنامه است[۳۷]. این ویژگی در این پژوهش به‌کار نرفته است، زیرا به‌بررسی‌های زبانشناسی در قالب دانش فهم زبان طبیعی نیاز داشته است.

۳-۳ ویژگی های آوایی

این دسته از ویژگی‌ها به اطلاعات نهفته در آوای شعر، مانند وزن و هجای آن

می‌پردازند:

۱- بررسی وزن شعر: در اشعار فارسی تاکنون بیش از ۳۰۰ گونه وزن و بحر شعری شناخته شده است. نوع و وزن اشعار شاهنامهٔ فردوسی مثنوی بحر متقارب مثمن محذوف است. این وزن به‌صورت "فعولن فعولن فعولن فعل" یا "تن‌تن تن‌تن تن‌تن تن‌تن" خوانده می‌شود و در هنگام تقطیع شعر از نظر عروضی (از راست به چپ) "U - - U - - U - - U -" نگاشته می‌شود، که در آن "U" نمایشگر هجای کوتاه و "-" نمایشگر هجای کشیده است و با درنظر گرفتن "c" برای حروف صامت، "v" برای مصوت کوتاه و "V" برای مصوت بلند، بنا بر قرارداد جدول ۴ به‌دست می‌آید[۲۳و۳۸].

جدول ۴- هجای کوتاه و کشیده در زبان فارسی

هجا	نماد	نمونه
cv	U	به، دو
cV	-	با، بو، بی
cvc	-	بر، خب
cVc	-U / -	جان، مور، شیر
cvcc / cVcc	-U / -	پارس، کاشت، پرت

گفتنی است که برای تقطیع یا به‌دست آوردن هجای یک شعر، ۳ روش وجود دارد: نخست روش "کلاسیک" که ادیبان تا یکی-دو سده پیش از آن پیروی می‌کرده‌اند، دوم روش "جدید" که در سدهٔ گذشته به‌کار می‌رفته[۳۸] و سوم روش "نو" که نزدیک به یکی-دو دهه از پیدایش آن می‌گذرد[۲۳]. البته هم‌اکنون روش سنتی دیگر کاربردی ندارد و نتیجهٔ تقطیع به‌روش دوم و سوم یکسان است؛ اما در این پژوهش روش "نو"، به‌دلیل سادگی و الگوریتمیک بودن، برای تقطیع به‌کار گرفته شده است.

برای نمونه بیت (۲) آوایی این‌چنین دارد:

manamgoftbAfarreyeizadi
hamamSahreyArihamammawbadi

یا:

idazieyerrafAbtfogmanam
idabwammamahirAyerhaSmamah

که آوای آن چنین نوشته می‌شود:

cvcvccvcccVcvccvcvcVcvcV
cvcvccvccvcVcVcvcvccvccvcV

و البته هجای عروضی آن نیز در هر دو مصراع همان مثنوی بحر متقارب است.

از آنجا که اشعار بیشتر شاعران دیگر بر وزن‌هایی متفاوت سروده شده است، بنابرین سنجش وزن بیت آزمایشی با بحر اصلی می‌تواند ویژگی خوبی برای دسته‌بندی باشد. در روش‌های به‌کار گرفته شده در این پژوهش، تا ۱۶ هجا برای هر مصراع، هر یک به‌طور جداگانه به‌عنوان ویژگی سنجیده شده است.

۲- بررسی هجایی: هر مصراع از شعر فردوسی همواره ۱۱ هجا دارد و این ویژگی نیز می‌تواند ابزار خوبی برای دسته‌بندی اشعار باشد[۲۳و۳۸].

۳- بررسی تفاضل هجای دو مصراع: روشن است که تفاضل تعداد هجای دو مصراع ابیات شاهنامه، مانند بیشتر دیگر وزن‌های شعر فارسی همواره صفر است و این ویژگی می‌تواند ابیات شاهنامه را از نثر یا شعر نو ــ که دو بند از آن همواره هجای برابری ندارد ــ متمایز کند.

۴- سبک شعر: برخی از صاحب‌نظران فن شعر ویژگی‌ای را بر پایهٔ سبک یا آهنگ و درون‌مایهٔ اشعار برخی از شاعران در می‌یابند که شعر ایشان را از دیگران متمایز می‌سازد. این سبک از وزن جدا است. برای نمونه بیت زیر از سعدی است:

قضا کشتی آنجا که خواهد برد / وگر ناخدا جامه بر تن درد

این شعر بر وزن شاهنامه است و همهٔ ویژگی‌هایش با اشعار فردوسی ــ به‌جز داشتن یک واژهٔ عربی ــ هم‌خوانی دارد، اما می‌گویند اگر فردوسی چنین بیتی را می‌سرود، به‌شکل زیر درمی‌آمد:

برد کشتی آنجا که خواهد خدای / وگر جامه بر تن درد ناخدای

همان‌گونه که دیده می‌شود، بیت دوم آهنگی حماسی‌تر دارد که ویژهٔ فردوسی است. این چنین خاصیتی نزد خبرگان شعر پارسی مفهوم و قابل شناسایی و ارزیابی است. (رجوع کنید به [۴۰و۴۱]) و در این پژوهش این ویژگی به‌دلیل پیچیدگی و دشواری در تبدیل به مقادیر کمی، به‌کار نیامده است.

۵- بررسی گویش: در زمانه‌ای که فردوسی شاهنامه را می‌سروده، گویش بسیاری از واژگان با گویش‌های پس از آن هنگام تفاوت داشته است. برای نمونه واژگان "سخن" و "کهن" به‌شیوهٔ "saxon" و "kahon" خوانده می‌شده است و بنابراین نمی‌تواند با عبارتی مانند "چو من" هم‌قافیه باشد.

۳-۴ کاهش ویژگی

در هنگام پردازش ویژگی‌های گوناگون مسایل الگوریتمیک، برای دسته‌بندی یک شیی، وضعیت مطلوب آن است که شیی با کمترین تعداد ویژگی‌ها دسته‌بندی گردد [۱۸و۱۹]. در دسته‌بندی اشعار نیز انتظار می‌رود که از میان ویژگی‌های پیشنهاد شده، برخی قابل حذف باشد، بدون این‌که چندان از دقت دسته‌بندی کاسته شود. برای نمونه به‌نظر می‌رسد که تعداد واژگان مصراع، خود تابعی از تعداد حروف مصراع باشد یا با به‌کار گرفتن ویژگی واژگان ناایرانی، دیگر نیازی به سنجش ویژگی نوع حروف نباشد.

از این رو برای تعیین درجهٔ اهمیت، ویژگی‌های نمونه‌ها ــ به‌جای تحلیل آماری ــ در یک شبکهٔ عصبی پرسپترون تک‌لایه با نرون‌هایی به‌تعداد ویژگی‌های داده شده در لایهٔ ورودی، ارزیابی گردید و بر پایهٔ مشاهدهٔ تجربی نتیجه گرفته شد که وزن نرون‌های متناظر با هر ویژگی، ارزش هر ویژگی را معین می‌سازد.

شبکهٔ عصبی پرسپترون را می‌توان با یک تابع مشخصهٔ خطی (معادلهٔ ۱) نمایش داد که در حقیقت معادله‌ای برای دسته‌بندی خطی الگوها در دو دستهٔ خروجی y با بردار ورودی ویژگی‌های x است و w وزن‌هایی را در بر دارد که مشخصات ابرفضا را تعریف می‌کند[۲۰].

$$y = f(\sum_{i=1}^{n}(w_i x_i + w_0)) \qquad (۲)$$

جدول ۵ ترتیب اهمیت برخی از این ویژگی‌ها که به‌کمک شبکهٔ عصبی پرسپترون تهیه شده است را در قالب سه دستهٔ کلی جای هجا، نوع حروف به‌کار رفته و دیگر ویژگی‌ها نشان می‌دهد.

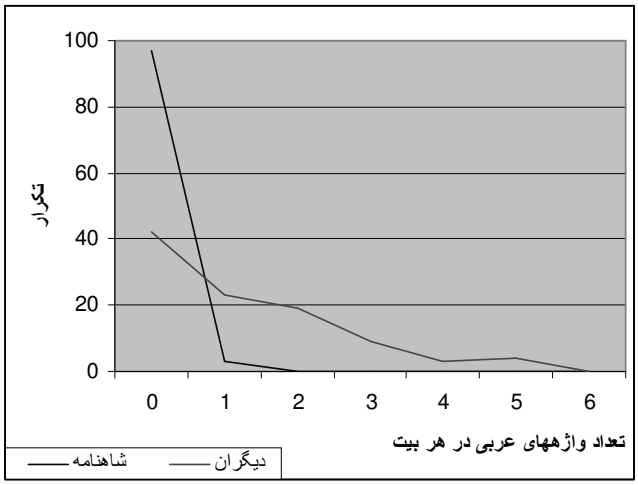
البته به‌نظر می‌رسد که این گونه پردازش برای رتبه‌بندی برخی از ویژگی‌ها ــ مانند تعداد واژگان ناایرانی و هجاهای هر مصراع ــ که به‌صورت خطی رشد می‌کنند یا منحنی آنها از درجهٔ ۱ است، بسیار مناسب باشد، اما ویژگی‌هایی ــ مانند تعداد حروف مصراع یا تعداد هجای مصراع ــ که منحنی درجه ۲ می‌سازند، به‌درستی دسته‌بندی نگردد (شکل ۱ و ۲).

از این رو برای این دست از ویژگی‌ها یک پردازش آماری نیز بر پایهٔ میانگین و انحراف معیار صورت گرفته است. این‌گونه کاهش ویژگی، توانایی کاهش خودکار ویژگی‌ها و در نتیجه، خودیادگیری را برای بهینه‌سازی خودکار سیستم بدان می‌افزاید[۲۲].

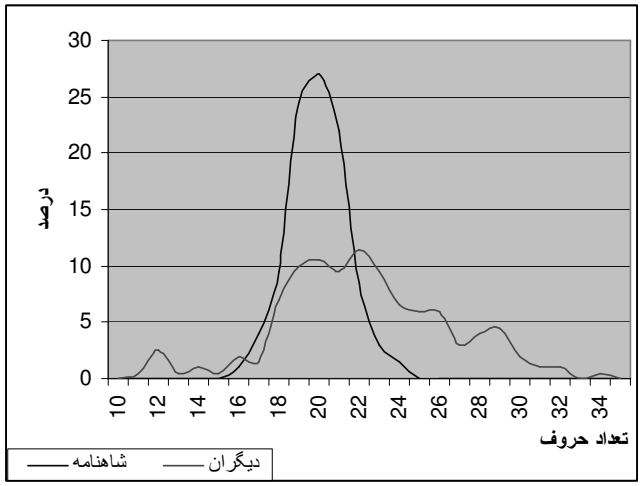
همچنین برای برخی از ویژگی‌ها ــ مانند تعداد واژگان اروپایی و ترکی ــ به‌دلیل آن‌که شبکهٔ عصبی اطلاعات کمی از نمونه‌های هر دو گروه داشته، نتوانسته است آنها را به‌درستی تمایز گذارد و از این‌رو وزنی مناسب را دریافت نکرده است (و در این آزمایش‌ها منظور نگردیده)؛ اما به‌نظر می‌رسد با افزایش تعداد نمونه‌های آموزشی منفی و مثبت، این‌گونه ویژگی‌ها نیز جایگاه خود را بازیابند.

جدول ۵- رتبه‌بندی ویژگی‌های گوناگون شعری بر پایهٔ وزن‌های شبکهٔ عصبی پرسپترون

هجا	رتبه	حروف	رتبه	ویژگی‌های دیگر	رتبه
هجای ۱	۶	پ	۶	تعداد حروف قافیهٔ مصراع	۹
هجای ۲	۱۱	ث	۴	تعداد حروف مصراع	۷
هجای ۳	۱۲	ج	۱۳	تعداد حروف واژهٔ ردیف	۱۰
هجای ۴	۳	ح	۹	تعداد حروف هم‌قافیه	۱۱
هجای ۵	۸	ذ	۷	تعداد هجای مصراع	۳
هجای ۶	۱۰	ز	۸	تعداد واژگان مصراع	۸
هجای ۷	۱۳	ژ	۳	تعداد واژگان عربی	۴
هجای ۸	۱۵	ص	۱۱	تعداد واژگان ردیف	۱۲
هجای ۹	۱۶	ط	۲	تفاضل تعداد حروف قافیه	۱۴
هجای ۱۰	۹	ظ	۱	تفاضل تعداد حروف مصراع	۱۳
هجای ۱۱	۷	ع	۱۰	تفاضل تعداد واژگان مصراع	۱۵
هجای ۱۲	۱	ق	۵	هم‌قافیه بودن بیت	۶
هجای ۱۳	۲	ک	۱۶	کلید واژه	۱
هجای ۱۴	۴	گ	۱۴	ردیف دار بودن بیت	۵
هجای ۱۵	۵	ل	۱۲	هماهنگی با الگوی وزنی	۲
هجای ۱۶	۱۴	م	۱۵		



شکل ۱- نوسانات دو ویژگی تعداد حروف هر مصراع



شکل ۲- تعداد واژه‌های عربی در هر مصراع برای اشعار شاهنامه و دیگران

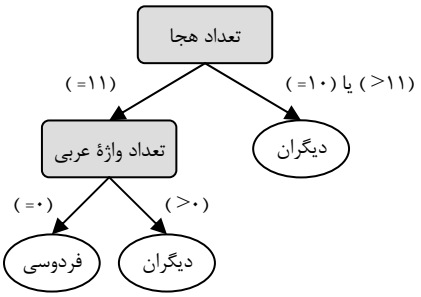
۴- آزمایش نمونه‌ها

در این پژوهش، دو روش جداگانهٔ الگوریتمیک (درخت تصمیم) و هوشمند (شبکهٔ عصبی مصنوعی) برای سنجش مقادیر ویژگی‌ها و در نتیجه شناسایی اشعار به‌کار رفته است.

۴-۱ درخت تصمیم

روش کار درخت تصمیم بدین گونه است که تصمیم برای دسته‌بندی متن یا واژه در چند مرحله و با پیمایش درخت انجام می‌پذیرد؛ بدین ترتیب که در هر گره با توجه به ویژگی‌های متن یا واژه، گره بعدی انتخاب می‌شود و گره‌های پایانی، کلاس متن را مشخص می‌کنند. در این روش اگر در هر گره دسته‌بندی کننده‌ای مجزا به‌کار رود، روش سلسله مراتبی است[۲۲].

شکل ۳ یک درخت تصمیم نوعی را بر پایهٔ دو ویژگی تعداد هجا و واژگان عربی نشان می‌دهد.



شکل۳- درخت تصمیم ساده برای دسته‌بندی اشعار فردوسی بر پایهٔ تعداد هجا و واژهٔ عربی

البته روشن است که با توجه به آن‌که همواره دسته‌ها با یکدیگر هم‌پوشانی‌هایی را دارند، نمی‌توان با چنین قطعیتی تعلق به یک دسته را تعیین کرد و در عمل نیز درخت بسیار پیچیده‌تر از این شکل بوده است. جدول ۶ دقت شناسایی ۴۰۰ نمونه از اشعار شاهنامه و دیگران را به‌کمک درخت تصمیم ساخته شده بر پایهٔ اولویت ویژگی‌های جدول ۵ نشان می‌دهد.

جدول ۶- دقت شناسایی درخت تصمیم در سنجش ویژگی‌های گوناگون شعری؛ I) ویژگی‌های فیزیکی، II) ویژگی‌های مفهومی و III) ویژگی‌های آوایی و IV تا VII) تعداد ویژگی‌های برگزیده، بر پایهٔ رتبه‌بندی جدول ۳

ردیف	ویژگی‌ها	تعداد ویژگی‌ها	دقت (درصد)	ردیف	ویژگی‌ها	تعداد ویژگی‌ها	دقت (درصد)
۱	I	۲۸	۷۵/۵۰	۶	II و III	۳۶	۸۳/۷۵
۲	II	۲	۸۲	۷	I و II و III	۶۴	۸۵/۵۰
۳	III	۳۴	۷۵/۲۵	۸	IV	۴۸	۸۲
۴	I و II	۳۰	۷۹/۲۵	۹	V	۳۲	۷۵/۲۵
۵	I و III	۵۸	۷۸/۷۵	۱۰	VII	۱۶	۶۵/۲۵

۴-۲ شبکهٔ عصبی مصنوعی

در این روش یک شبکه عصبی مصنوعی پیشخور با پس‌انتشار و دو لایهٔ پنهان با تعداد دلخواهی نرون در لایهٔ نخست (به‌تعداد ویژگی‌های مورد سنجش) و یک نرون خروجی به‌کار گرفته شده است. دلیل برگزینش این شبکه آن بوده که در

پردازش توابعی که ماهیت ایستا و غیرپویا دارند، بسیار خوب عمل می‌کند[۲۱] و البته نتایج عملی آن در این پژوهش نیز بسیار خوب بوده است. چکیدهٔ این نتایج ــ که پس از چندین بار آزمون شبکه هر بار با ۲۰۰ نمونهٔ آموزشی (از اشعار فردوسی و دیگران) و ۲۰۰ نمونهٔ آزمایشی (از اشعار فردوسی و دیگران) به‌دست آمده ــ در جدول ۷ نمایش داده شده است. تعداد اپک‌ها[۲۱] به اندازه‌ای انتخاب شده که خطای شبکه به مقداری ثابت برسد. در این جدول آزمون‌های گروه IV تا VII با تعداد دلخواهی از ویژگی‌ها به‌ترتیب اولویتی که در جدول ۵ به‌دست آمده، برگزار شده است و نتایج نشان می‌دهد که کاهش هوشمندانهٔ ویژگی‌ها تا کمتر از نصف ویژگی‌های آغازین، آسیب چندانی به دقت شبکه وارد نمی‌آورد.

۵- نتیجه‌گیری

در این مقاله با هدف تشخیص هوشمند اشعار شاهنامهٔ فردوسی، کوشش شد تا شیوهٔ تفکر انسان برای تشخیص اشعار شاعران شبیه‌سازی گردد و از این‌رو ۶۴ گونه ورودی برگرفته از ۲۲ ویژگی از اشعار پارسی، در سه گروه فیزیکی، مفهومی و آوایی استخراج و به سیستم ارایه گردید. اشعار به دو روش درخت تصمیم و شبکه عصبی مصنوعی پیشخور با پس‌انتشار تحلیل و دسته‌بندی شد. همچنین برای کاهش خودکار و تعیین ارزش هر یک از ویژگی‌های استخراج شده، یک شبکهٔ عصبی پرسپترون تک‌لایه به‌کار گرفته شد.

پس از آموزش و آزمایش نمونه‌ها در هر دو روش، مشخص شد که دسته‌بندی اشعار آزمایشی در دو دستهٔ اشعار شاهنامه‌ای و غیرشاهنامه‌ای به‌روش درخت تصمیم ۸۵/۵٪ درستی و به‌کمک شبکهٔ عصبی مصنوعی تا ۱۰۰٪ درستی را به‌همراه داشته است. همچنین مشاهده گردید که با اعمال نتایج تجربی به‌دست آمده از شبکهٔ پرسپترون، با حفظ دقت مطلوب، بار شبکهٔ عصبی کاهش یافته است.

در راه انجام این پژوهش، پیش‌نیازهای زیر به انجام رسیده است:

۱- پروسه‌هایی برای تقطیع خودکار شعر و متن فارسی طراحی و پیاده‌سازی شد.

۲- به‌منظور آوانگاری نوشتار و گفتار، خط کهن ایرانی (اوستایی) بازشناسی و استفاده گردید.

۳- پایگاه دادهٔ ویژه‌ای برای کاربردهای پردازش زبان و گفتار فارسی طراحی و پیاده‌سازی شد.

۴- نتایجی به‌دست آمد که می‌تواند به‌عنوان نظری محاسباتی، برای کمک به شعرشناسان و ادیبان در تشخیص اشعار اصل شاهنامه از افزوده به‌کار رود.

افزون بر این موارد، این تحقیق در حد بضاعت خود موجب برقراری پلی بین اندیشمندان علوم انسانی (ادبیات) و کارشناسان علوم کامپیوتر (هوش مصنوعی) گردیده است.

همچنین مهم‌ترین نتایج فنی برگرفته از این پژوهش بدین شرح است:

۱- با استخراج و به‌کارگیری درست ویژگی‌هایی از متن یا شعر می‌توان حتی بدون استفاده از روش‌های مفهوم‌گرای پردازش زبان طبیعی به شناسایی و تحلیل آنها پرداخت.

۲- نتایج کار شبکهٔ عصبی مصنوعی به‌کار گرفته شده، به‌روشنی از روش درخت تصمیم بهتر بوده است.

۳- بر پایهٔ مشاهده و تجربه، وزن گره‌های خروجی متناظر با ویژگی‌های ورودی به شبکهٔ پرسپترون تک‌لایه معیاری برای شناخت درجهٔ اهمیت هر ویژگی در کل شبکه بوده است.

۴- با گزینشی درست از ویژگی‌ها، حتی با تعدادی اندک از نمونه‌های آموزشی نیز می‌توان دقت شناسایی را تا ۱۰۰٪ افزایش داد.

اهداف کوتاه مدت زیر در دستورکار پژوهش‌های آینده جای دارد:

۱- شناخت اشعار دیگر شاعران به‌طور جداگانه

۲- شناخت اشعار چند شاعر به‌طور همزمان

۳- بهینه‌سازی نوع و تعداد ویژگی‌ها

۴- آزمایش روش‌های دسته‌بندی دیگر

۵- کار بر مسالهٔ کاهش ویژگی با شبکه‌های عصبی غیر خطی.

همچنین افزودن تناسبات هجایی در انتخاب واژگان خروجی سیستم‌های پردازش زبان طبیعی و در نتیجه سرایش شعر توسط کامپیوتر، در دستور کار اهداف بلند مدت جای گرفته است.

مراجع

[1] S. J. Russell, and P. Norvig, "Artificial Intelligence: A Modern Approach," *Prenticr-Hall International Inc*., 1994.

[2] J. Allen, "Natural Language Understanding," *The Benjamin/Cummings Publishing Co., 2nd Edition,* 1994.

جدول ۷- دقت شناسایی شبکه عصبی پیشخور با پس‌انتشار در سنجش با ویژگی‌های گوناگون شعری؛ I) ویژگی‌های فیزیکی، II) ویژگی‌های مفهومی، III) ویژگی‌های آوایی و IV تا VII) تعداد ویژگی‌های برگزیده، بر پایهٔ رتبه‌بندی جدول ۳

ردیف	ویژگی‌ها	دقت (درصد)	کمترین مقدار (نمونه‌های مثبت)	بیشترین مقدار (نمونه‌های منفی)	تعداد اپک	تعداد گره‌های لایهٔ یکم (ویژگی‌ها)	خطای شبکه
۱	I	۸۹/۵۳	۰/۷۵	-۰/۱	۱۵۰	۲۸	$۱۰^{-۳}$
۲	II	۸۲/۳۳	۰/۸۵	-۰/۱	۱۰۰	۲	$۱۰^{-۲}$
۳	III	۹۷/۲۰	۰/۲	-۰/۱	۱۵۰	۳۴	$۱۰^{-۳}$
۴	I و II	۱۰۰	۱	-۰/۱	۳۰۰	۱۳	$۱۰^{-۱۰}$
۵	I و III	۱۰۰	۱	-۰/۱	۳۰	۴۶	$۱۰^{-۸}$
۶	II و III	۱۰۰	۱	-۰/۱	۸۰	۳۵	$۱۰^{-۱۰}$
۷	I و II و III	۱۰۰	۱	-۰/۱	۲۵	۶۴	$۱۰^{-۸}$
۸	IV	۱۰۰	۱	-۰/۱	۵۰	۴۸	$۱۰^{-۸}$
۹	V	۱۰۰	۰/۹	-۰/۱	۱۰۰	۳۲	$۱۰^{-۳}$
۱۰	VII	۹۷/۲۲	۰/۴	۰	۱۵۰	۶۴	$۱۰^{-۲}$

[3] S. L. Tanimoto, "The Elements of Artificial Intelligence Using Lisp, " *Computer Science Press, 2nd Edition,* 1995.

[4] M. E. Maron, "Automatic indexing: an experimental inquiry," *Journal of the Association for Computing Machinery*, vol. 8, no. 3, pp. 404-417, 1961.

[5] W. Li, B. Lee, F. Krausz, and K. Sahin, "Text classification by a neural network," *Proceedings of the 23rd Annual Summer Computer Simulation Conference, Baltimore, USA*, pp. 313-318, 1991.

[6] P. S. Jacobs, "Joining statistics with NLP for text categorization," *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing, Association for Computational Linguistics, Morristown, US*, editor Marcia Bates and Oliviero Stock, pp. 178-185, 1992.

[7] P. Geutner, U. Bodenhausen, and A. Waibel, "Flexibility Through Incremental Learning: Neural Networks for Text Categorization," *Proceedings of WCNN-93, World Congress on Neural Networks*, pp. 24-27, 1993.

[8] E. Riloff, and W. Lehnert, "Information extraction as a basis for high-precision text classification," *ACM Transactions on Information Systems*, vol. 12, no.3, pp. 296-333, 1994.

[9] E. D.Wiener, J. O. Pedersen, and A. S. Weigend, "Neural network approach to topic spotting," *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, pp. 317-332, 1995.

[10] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," *Computer Science Technical Report CMU-CS-96-118*, *Carnegie Mellon University*, 1995.

[11] D. Hanauer, "Integration of phonetic and graphic features in poetic text categorization judgments," *Poetics*, vol. 23, no.5, pp. 363-380, 1996.

[12] D. Merkl, "Text classification with self-organizing maps: Some lessons learned," *Neurocomputing*, vol. 21, no. 1/3, pp. 61-77, 1998.

[13] J. F. Hoorn, S. L. Frank, W. Kowalczyk, and F. Van der Ham, "Neural Network Identification of Poets Using Letter Sequences," *Literary and Linguistic Computing*, vol. 14, pp. 311-338, 1999.

[14] S. Statamatos, N. Fakotakis, and G. Kokkiniakis, "Automatic text categorization in terms of genre and author," *Computational Linguistics*, vol. 26, no. 4, pp. 471-495, 2000.

[15] A. Aizawa, "Linguistic Techniques to Improve the Performance of Automatic Text Categorization," *Proceedings of NLPRS-01, 6th Natural Language Processing Pacific Rim Symposium*, pp. 307-314, 2001.

[16] A. N. Pavlov, W. Ebeling, L. Molgedey, A. R. Ziganshin, and V. S. Anishchenko, "Scaling features of texts, images and time series," *Physica A 300*, pp. 310–324, 2001.

[17] O. De Vel, M. Corney, A. Anderson, and G. Mohay, "Language and Gender Author Cohort Analysis of E-mail for Computer Forensics," *Queensland University of Technology*, 2003.

[18] K Fukunaga, "Statistical Pattern Recognition," *Academic Press Inc*, 1990.

[19] A. R. Webb, "Statistical Pattern Recognition," *John Wiley & Sons Ltd., 2nd Edition*, 2002.

[20] M. Sonka, V. Hlavac, and R. Boyle, "Image Processing, Analysis and Machine Vision," *Chapman & Hall Computing*. 1993.

[21] R. Hetch-Nielsen, "Neurocmputing," *Addison-Wesley Publishing Company,* 1989.

[22] T. M. Mitchel, "Machine Learning," *McGraw-Hill*, 1997.

[۲۳] ت. وحیدیان کامیار، ع. زرین‌کوب، ح. زرین‌کوب، "ادبیات فارسی: قافیه و عروض و نقد ادبی"، شرکت چاپ و نشر کتابهای درسی ایران، ۱۳۷۷.

[۲۴] م. ابوالقاسمی، "راهنمای زبان‌های باستانی ایران"، سمت، جلد ۱و۲، ۱۳۷۶.

[۲۵] ه. رضی، "دستورزبان اوستایی"، سازمان انتشارات فروهر، ۱۳۶۸.

[۲۶] ا. بهرامی، "فرهنگ واژه‌های اوستا"، نشر بلخ، ۱۳۶۹.

[۲۷] ذ. بهروز، "خط و فرهنگ"، سازمان انتشارات فروهر، ۱۳۶۳.

[۲۸] ج. جنیدی، "نامهٔ پهلوانی: خودآموز خط و زبان پهلوی اشکانی و ساسانی"، نشر بلخ، ۱۳۶۰.

[۲۹] فردوسی توسی، "شاهنامه فردوسی"، به‌تصحیح فریدون جنیدی، نشر بلخ، در حال انتشار.

[۳۰] فردوسی طوسی، شاهنامه فردوسی، نسخه ژول مول، انتشارات سخن، ۱۳۶۹.

[۳۱] ع. زرین‌کوب، "دو قرن سکوت"، انتشارات جاویدان، ۲۵۳۶.

[۳۲] غ. سلیم، "محمود غزنوی سرآغاز واپس‌گرایی در ایران"، نشر بلخ، ۱۳۸۳.

[۳۳] م. ماحوزی، "فارسی عمومی"، انتشارات اساطیر، ۱۳۷۶.

[۳۴] ع. ا. دهخدا، "لغت‌نامه"، انتشارات دانشگاه تهران، (لوح فشرده)، ۱۳۸۲.

[۳۵] م. معین، "فرهنگ فارسی"، انتشارات امیرکبیر، ۱۳۶۱.

[۳۶] ح. شهیدی مازندرانی، "فرهنگ شاهنامه: نام کسان و جایها"، نشر بلخ، ۱۳۷۷.

[۳۷] م. ثاقب‌فر، "شاهنامهٔ فردوسی و فلسفهٔ تاریخ ایران"، انتشارات قطره/ معین، ۱۳۷۷.

[۳۸] س. شمیسا، "آشنایی با عروض و قافیه، ویراست سوم"، انتشارات فردوس، ۱۳۶۶.

[۳۹] امیرشهاب شاهمیری، "تعیین شاعر به‌کمک روش‌های یادگیری ماشین"، پایان‌نامهٔ کارشناسی ارشد در رشتهٔ هوش ماشین و رباتیک، دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)، ۱۳۸۵.

[۴۰] محمدرضا شفیعی کدکنی، "موسیقی شعر"، موسسهٔ انتشارات آگه، ۱۳۸۴.

[۴۱] محمدرضا شفیعی کدکنی، "صور خیال"، موسسهٔ انتشارات آگه، ۱۳۸۳.

امیرشهاب شاهمیری دارای درجهٔ کارشناسی ارشد مدیریت در گرایش امور فرهنگی از دانشگاه علوم و تحقیقات تهران در سال ۱۳۸۲ و کارشناسی ارشد کامپیوتر در گرایش هوش مصنوعی و رباتیک از دانشگاه صنعتی امیرکبیر در سال ۱۳۸۵ است. وی به پژوهش در زمینه‌های پردازش زبان و متن فارسی، الگوریتم‌های یادگیری ماشین، شبکه‌های عصبی مصنوعی، ترکیب علوم فرهنگی و انسانی با هوش مصنوعی، و روش‌های

کشف و تصحیح خطا علاقه دارد و تاکنون هشت مقاله در زمینه‌های فوق در کنفرانس‌ها و نشریات داخلی به انتشار رسانیده است.

آدرس پست الکترونیکی ایشان عبارتست از:

shahmiri@ce.aut.ac.ir

سعید شیری قیداری دارای درجهٔ دکتری کامپیوتر از دانشگاه کوبه ژاپن در گرایش سیستمهای هوشمند مصنوعی در سال ۲۰۰۲ است. در حال حاضر وی عضو هیات علمی دانشکده کامپیوتر دانشگاه صنعتی امیر کبیر می باشد. تحقیقات ایشان در زمینه های هوش مصنوعی، روباتیک، مکاترونیک، یادگیری ماشین و بینائی ماشین متمرکز است. ایشان در زمینه روبوکاپ نیز فعایت نموده و عضو کمیته ملی روبوکاپ ایران هستند.

آدرس پست الکترونیکی ایشان عبارتست از:

shiry@aut.ac.ir

رسول دژکام دارای درجهٔ کارشناسی ارشد کامپیوتر در گرایش هوش مصنوعی و رباتیک از دانشگاه صنعتی امیرکبیر در سال ۱۳۸۵ است. وی به زمینه‌های پژوهشی هستی‌شناسی و شبکه‌های عصبی مصنوعی علاقه‌مند بوده، تاکنون سه مقاله در زمینه‌های فوق در کنفرانس‌ها و نشریات داخلی به انتشار رسانیده است

آدرس پست الکترونیکی ایشان عبارتست از:

dezhkam@ce.aut.ac.ir

پیوست ۱

ردیف	اوستایی	آوا	ردیف			ردیف		
۱	a	اَ	۱۲	p	پ	۲۴	S	ش
۲	A	آ	۱۳	t	ت	۲۵	f	ف
۳	e	اِ	۱۴	C	ث	۲۶	q	غ
۴	E	اِ کشیده	۱۵	j	ج	۲۷	k	ک
۵	o	اُ	۱۶	c	چ	۲۸	g	گ
۶	O	اُ کشیده	۱۷	x	خ	۲۹	l	ل
۷	u	او	۱۸	d	د	۳۰	m	م
۸	U	او کشیده	۱۹	D	ذ	۳۱	n	ن
۹	i	ای	۲۰	r	ر	۳۲	v	و
۱۰	I	ای کشیده	۲۱	z	ز	۳۳	w	و میان دو لب
۱۱	b	ب	۲۲	Z	ژ	۳۴	h	ه
			۲۳	s	س	۳۵	y	ی

[1] Poet Identification
[2] Author Identification
[3] Natural Language Processing
[4] Feed-Forward Backpropagation
[5] Decision Tree
[6] Single Perceptron
[7] Text Categorization
[8] Knowledge-based
[9] Augmented Relevancy Signatures Algorithm
[10] Term Selection
[11] Latent Semantic Indexing (LSI)
[12] Hierarchically Self-Organizing Maps (HSOM)
[13] Letter Sequence
[14] *k*-Nearest Neighbour
[15] Detrended Fluctuation Analysis
[16] Wavelet Transform Modulus Maxima

[17] Fluctuation Analysis of Sequences
[18] Light Support Vector Machine (SVM^{light})
[19] Feature Reduction
[20] برای شمارش این واژگان از نرم‌افزار «درج» نگارش ۲ و ۳ کمک گرفته شده است.
[21] Epochs