نشریه علمی- ترویجی محاسبات نرم
شماره دوم، پاییز و زمستان ۱۳۹۱/ صفحه ۲۶- ۳۵

یک سیستم نوین هوشمند تشخیص هویت نویسندهٔ فارسی‌زبان بر اساس سبک نوشتاری

زینب فرهمندپور*[۱]، هومان نیک‌مهر[۲]، محرم منصوری‌زاده[۳]، امید طبیب‌زاده قمصری[٤]

[۱] دانشجوی کارشناسی ارشد، دانشکده مهندسی، دانشگاه بوعلی سینا، همدان، ایران
zeinab.farahmandpoor@gmail.com
[۲] استادیار، گروه مهندسی کامپیوتر، دانشگاه اصفهان، اصفهان، ایران
nikmehr@eng.ui.ac.ir
[۳] استادیار، دانشکده مهندسی، گروه کامپیوتر، دانشگاه بوعلی سینا، همدان، ایران
mansoorm@basu.ac.ir
[٤] استادیار، دانشکده ادبیات، گروه زبان‌شناسی، دانشگاه بوعلی سینا، همدان، ایران
o.tabibzadeh@basu.ac.ir

چکیده: پیشرفت سریع ارتباطات اینترنتی و استفادهٔ نادرست از ناشناس بودن متون آنلاین، باعث به وجود آمدن مسائل امنیتی شده است. هویت ناشناس ابزارهای اینترنتی مانند ایمیل‌ها، وبلاگ‌ها و وب‌سایت‌ها به روش‌های ارتباط مورد علاقه‌ای برای کاربردهای جنایی تبدیل شده است. روابط سیاسی و اجتماعی جهان، باعث توجه زیادی به زبان فارسی شده است. این موضوع خود، موجب فراوانی متون فارسی در اینترنت گردیده است. در این مقاله، یک روش هوشمند writeprint معرفی شده است که به شناسایی یک نویسندهٔ فارسی‌زبان بر اساس سبک نوشتاری او کمک می‌کند. در این تحقیق، از ویژگی‌های واژگانی، نحوی، معنایی و وابسته به کاربرد، برای شناسایی هویت استفاده شده است. و کارآیی انواع مختلف این ویژگی‌ها و همچنین آثار روش‌های دسته‌بندی KNN و Delta به همراه ترکیب این روش‌ها با الگوریتم ژنتیک روی دو پایگاه دادهٔ جمع‌آوری شده بررسی شده است. در کنار سایر ویژگی‌ها برای پیاده‌سازی روش پیشنهادی یک pos tagger طراحی شده است که از ساختار واژه‌ها برای تشخیص اسم، صفت و قید استفاده می‌کند. آزمایش‌های انجام شده در این تحقیق نشان دادند که در بین روش‌های دسته‌بندی بررسی شده، ترکیب KNN و الگوریتم ژنتیک، دقت بالاتری را در تشخیص هویت مالک اثر ادبی تولید کرد.

واژه‌های کلیدی: تشخیص هویت نویسنده، سبک نوشتاری، writeprint.

۱. مقدمه

تشخیص نویسنده از سویی، یکی از قدیمی‌ترین و از سوی دیگر، یکی از به‌روزترین مسائل در سبک‌شناسی و بازیابی اطلاعات است. تشخیص نویسنده، تلاشی برای نشان دادن خصوصیات تولیدکننده یا نویسندهٔ یک تکه از اطلاعات زبانی تعریف می‌شود، به طوری که بتوان بین متون مختلف که نوشتهٔ افراد مختلف‌اند، تمایز قائل شویم. اولین تلاش برای سنجیدن سبک نوشتاری به قرن ۱۹ میلادی باز می‌گردد. این مطالعات توسط Mendenhall [۳] در سال ۱۸۸۷ بر روی نمایشنامه‌های شکسپیر انجام شد. سپس مطالعات آماری در نیمهٔ اول قرن بیستم توسط Zipf [٤] در سال ۱۹۳۲ و Yule در سال‌های ۱۹۳۸ [٥] و ۱۹٤٤ [٦] انجام گردید. مطالعات دقیق‌تر توسط Mosteller و Wallace در سال ۱۹٦٤ [۷] و بر روی پایگاه داده The Federalist Paper انجام شد که بدون شک، یکی از قدرتمندترین و مؤثرترین تحقیقات در تعیین هویت نویسنده بوده است. این کار آغاز مطالعات غیر سنتی تشخیص هویت نویسنده محسوب می‌شود.

بر اساس فرض بسیاری از محققان، انسان‌ها الگوی مشخصی برای استفاده از زبان در نوشته‌های خود دارند که همانند نوعی اثر انگشت نویسنده عمل می‌کند و writeprint نامیده می‌شود. با گسترش اینترنت و ذات بدون مرز آن و همچنین افزایش ارتباطات آنلاین، نیاز به مسائل امنیتی تشخیص نویسنده روز به روز افزایش می‌یابد. کاربردهای تشخیص هویت نویسنده، شامل دزدی ادبی و علمی (مانند مقالات دانشگاهی و کتاب‌ها)، شناسایی نویسندهٔ متون نامناسب که به صورت ناشناس یا تحت نام مستعاری ارسال می‌شوند (مانند ایمیل‌ها یا نامه‌های تهدیدآمیز) یا حل سؤالات تاریخی راجع به متون مورد مشاجره نامعلوم می‌باشد.

زبان فارسی، یک زبان رسمی در بین کشورهای ایران، افغانستان و تاجیکستان است و بیش از ده‌ها میلیون نفر از مردم به این زبان صحبت می‌کنند. به دلایل سیاسی و مذهبی این زبان مورد توجه گروه‌های مختلفی قرار گرفته است. تفاوت‌های ساختاری آن با زبان انگلیسی(از جمله نحوهٔ ساخت اصطلاحات و جملات)، باعث کم‌توجهی به پیاده‌سازی و طراحی ابزارهای پردازش زبان طبیعی متناسب با زبان فارسی و تشخیص هویت نویسنده در این زبان باشد. پژوهش‌هایی در زمینهٔ تشخیص هویت نویسنده در زبان‌های نزدیک به زبان فارسی مانند عربی [۸]، ترکی [۹] و اردو [۱۰] نیز انجام شده است. در این مقاله، برخی روش‌های معروف در تشخیص نویسنده بر متون فارسی پیاده‌سازی شده است.

در این تحقیق، از چهار مجموعه ویژگی‌ها شامل واژگانی، نحوی، معنایی و وابسته به کاربرد استفاده نمودیم و دو روش دسته‌بندی KNN و Delta را روی داده‌های خود آزمایش کردیم. در نهایت برای بالا بردن دقت، الگوریتم ژنتیک را به کار بردیم.

ادامهٔ مقاله به صورت زیر سازماندهی گردیده است. ویژگی‌های مورد استفاده برای تشخیص سبک نوشتاری انحصاری نویسنده در بخش ۲ توضیح داده شده است. بخش ۳ به بیان روش‌های تشخیص هویت نویسنده و بخش ٤ به بررسی کاربرد این روش در زبان‌های مختلف پرداخته است. بخش ٥ روند طراحی و پیاده‌سازی سیستم پیشنهادی تشخیص نویسنده را توضیح می‌دهد. مقایسهٔ روی روش‌های پیاده‌سازی با استفاده از پایگاه‌های داده‌ای مختلف در بخش ٦ بیان گردیده است. در بخش ۷ به بحث و نتیجه‌گیری و در نهایت در بخش ۸ به بیان پیشنهادهایی برای کارهای آینده خواهیم پرداخت.

۲. ویژگی‌های سبکی برای شناسایی نویسندگان

از سال ۱۹٦٤ تا ۱۹۹۰، تحقیقات بر روی تشخیص نویسنده با تلاش برای سنجیدن سبک نوشتاری که stylometry [۷] نامیده می‌شد، در صدر قرار داشت. ویژگی‌های سبکی نوشتار، خصوصیاتی هستند که از متن استخراج می‌شوند تا تشخیص نویسنده را آسان کنند. ویژگی‌های واژگانی، نحوی، معنایی و مختص کاربرد [۱۲] چهار دستهٔ مهم از خصوصیات سبکی هستند که در تعیین نویسنده به کار می‌روند.

۲-۱. ویژگی‌های واژگانی

خصوصیات واژگانی، سنتی‌ترین و قدیمی‌ترین مجموعه ویژگی‌هایی هستند که برای شناسایی نویسنده به کار می‌روند. تاریخچهٔ این گروه از ویژگی‌ها به قرن ۱۹ میلادی، زمانی که Mendenhall [۱۳] از تعداد تکرار واژه‌های طولانی برای شناسایی نویسنده استفاده کرد، برمی‌گردد. این ویژگی‌ها که متن را به صورت یک رشته از واژه‌ها یا کاراکترها در نظر می‌گیرند، شامل طول جملات، پرمایگی واژگان، توزیع طول واژه‌ها در متن، تعداد تکرار کلمات منحصربه‌فرد، تعداد تکرار انواع n-gram با طول مختلف (تکه‌هایی از واژه‌ها با طول N کاراکتر) و... می‌باشند.

۲-۱-۱. پرمایگی واژگانی

معیارهای مختلفی تاکنون برای سنجش میزان پرمایگی واژگان یک متن ارائه شده‌اند که به طور گسترده‌ای برای تشخیص هویت نویسنده به کار گرفته می‌شوند. معمول‌ترین معیار از این دسته، type-token ratio است که به صورت $\frac{V}{N}$ محاسبه می‌شود. در این عبارت، V تعداد واژگان مجزا در متن مورد نظر و N تعداد tokenها (هر token، یک کلمه، عدد یا نشان نقطه‌گذاری[1] است) در آن متن است [۱۳].

برای اندازه‌گیری پرمایگی واژگان، Stamatatos و همکاران [۱۴] و Baayen و همکارانش [۱۵] از مجموعه‌ای از ۵ تابع به عنوان ویژگی استفاده کردند. این ۵ تابع به نام‌های K، R، W، S و D که به ترتیب توسط Yule [۱۶]، Honore [۱۷]، Brunet [۱۸]، Sichel [۱۹] و Simpson [۲۰] ارائه شده‌اند. این توابع با عبارات (۱) تا (۵) تعریف می‌شوند.

$$K = \frac{10^4\left(\sum_{i=1}^{\infty} i^2 v_i - N\right)}{N^2} \quad (۱)$$

$$R = \frac{(100 \log N)}{\left(1-\left(\frac{V_1}{V}\right)\right)} \quad (۲)$$

$$W = N^{V^{-\alpha}} \quad (۳)$$

$$S = \frac{V_2}{V} \quad (۴)$$

$$D = \sum_{i=1}^{V} V_i \frac{i(i-1)}{N(N-1)} \quad (۵)$$

در این روابط، V_i تعداد کلماتی است که دقیقاً i مرتبه تکرار شده‌اند و α پارامتری ثابت با مقدار ۰/۱۷ است. برای هر متن این توابع محاسبه شده و یک بردار از پنج ویژگی تشکیل می‌شود. در این تحقیق، از تمامی پنج تابع پرمایگی واژگان استفاده شده است.

۲-۱-۲. n-gram

در [۲۱]، n-gram تکهٔ N کاراکتری از یک رشتهٔ کاراکتری طولانی‌تر معرفی شده است. این ویژگی قادر به استخراج تفاوت‌های ظریف سبک، شامل اطلاعات واژگانی می‌باشد. مزیت این ویژگی، توانایی تحمل بالای آن نسبت به نویز (خطاهای املایی) است. برای ایجاد n-gram ابتدا متن ورودی خوانده شده و به چندین token مجزا تقسیم می‌شود. سپس اعداد، کاراکترهای کنترلی (انگلیسی، فاصله و نقطه‌گذاری) به کنار گذاشته شده و همهٔ انواع n-gram ممکن استخراج می‌شوند؛ برای مثال، bi-gram چند کلمه از جملهٔ قبلی با حذف اعداد، کاراکترهای کنترلی، انگلیسی، فاصله و نقطه‌گذاری عبارت‌اند از: |سپ|، |پس|، |سا|، |اع|، |عد|، |دا|، |اد|. اگر رشته‌ای از کاراکترها به طول K داشته باشیم، قادر به تولید k عدد uni-gram، k-1 عدد bi-gram، k-2 عدد tri-gram و k-3 عدد quad-gram می‌باشیم.

در این تحقیق از ۷۲ عدد uni-gram استفاده شده و bi-gram، tri-gram و quad-gram که از یک مقدار آستانه بیشتر بودند، به عنوان ویژگی انتخاب شدند. این مقدار آستانه از روی تجربه به دست آمده و در هر مورد نسبتی از تعداد تکرار بیشترین bi-gram، tri-gram و quad-gram بوده و برای هر پایگاه داده جداگانه محاسبه می‌شود. بدین ترتیب، انواع bi-gram را که تعداد تکرار آن‌ها بیشتر از ۰/۱ پرتکرارترین bi-gram در پایگاه دادهٔ ماست، به عنوان ویژگی انتخاب کردیم. به همین ترتیب، تعداد انواع tri-gram را که تعداد تکرار آن‌ها

1. punctuation mark

بیشتر از ۰/۰۵ پرتکرارترین tri-gram در پایگاه داده است، به عنوان ویژگی قرار دادیم. انواع quad-gram را که تعداد تکرار آن‌ها بیشتر از ۰/۱پرتکرارترین quad-gram در پایگاه دادهٔ مزبور است، به عنوان ویژگی محسوب کردیم.

۲-۱-۳. سایر خصوصیات واژگانی به کار رفته

سایر خصوصیات واژگانی استفاده شده عبارت‌اند از: میانگین طول کلمات، میانگین طول جملات بر حسب کاراکتر، میانگین تعداد کلمات موجود در جملات، توزیع کلمات کوتاه[1] مثلاً با طولی کوچک‌تر یا مساوی سه کاراکتر، توزیع کاراکترهای فارسی، توزیع تعداد پاراگراف، توزیع تکرار کلمات با طول یک تا ۳۰ (کوتاه‌ترین کلمه در زبان فارسی طولی معادل یک کاراکتر دارد و از آنجا که طول بلندترین کلمه در فارسی مشخص نیست، با تخمین‌های انجام شده مقداری برابر با ۳۰ را برای آن در نظر گرفتیم)، توزیع کاراکترهای انگلیسی، توزیع کاراکترهای عددی و توزیع کاراکتر نیم‌فاصله در متن.

۲-۲. ویژگی‌های نحوی

الگوهای به کار رفته توسط نویسنده برای تشکیل نوشته را ویژگی‌های نحوی گویند. این خصوصیت از این جهت اهمیت دارد که نویسنده به طور غیر آگاهانه، به استفاده از الگوهای نحوی یکسانی تمایل دارد؛ بنابراین، استفاده از ویژگی‌های نحوی می‌تواند روش تشخیص هویت قابل اعتمادتری نسبت به به کارگیری ویژگی‌های واژگانی محسوب شود؛ هرچند که استخراج این ویژگی‌ها به ابزارهای قوی و دقیق پردازش زبان طبیعی نیاز دارند. این ویژگی‌ها شامل تعداد تکرار function word، نشانه‌های نقطه‌گذاری، تعداد تکرار قوانین بازنویسی، تعداد عبارات اسمی، عبارات فعلی، تعداد اسم‌ها، قیدها و صفت‌های به کار رفته در متن و... می‌باشند. در این بخش، مهم‌ترین ویژگی‌های نحوی شامل توزیع تکرار رخداد عبارات اسمی، قیدی و صفت در متن هستند و توزیع انواع function word در متن، توضیح داده می‌شوند.

۲-۲-۱. توزیع تکرار عبارات اسمی، قیدی و صفت

در این تحقیق، نرم‌افزاری برای تشخیص عبارات اسمی، قیدی و صفتی، طراحی و پیاده‌سازی نمودیم. طبق [۱]، اسامی، قیدها و صفت‌ها می‌توانند ساده یا مرکب باشند. اگر ساختار کلمات ساده باشد، برای تشخیص اسم، صفت و یا قید، می‌توان آن‌ها را فهرست کرده و کلمات را با آن فهرست مقایسه کرد. اما اگر مرکب باشند، روش‌های ساختاری خاصی برای تشخیص آن‌ها نیاز است. در این نرم‌افزار مرز میان واژه‌ها را با تمام فاصله، و مرز میان واژه‌های تشکیل‌دهندهٔ کلمات مرکب را با نیم‌فاصله مشخص کردیم. البته اگر واژه‌های تشکیل‌دهندهٔ کلمات مرکب، بدون نیم‌فاصله به هم متصل شوند، با نیم‌فاصله جدا شده، و اگر واژه‌های تشکیل‌دهندهٔ کلمات مرکب بدون نیم-فاصله به هم نمی‌چسبند، نیازی به درج نیم‌فاصله نیست.

۲-۲-۲. function word

function word می‌تواند شامل حروف تعریف، حروف اضافه، حروف ربط، فعل و اسم باشند. این ویژگی‌ها که با تعداد تکرار زیادی در متون ظاهر می‌شوند و بدون آنکه معنی متن را تغییر دهند، می‌توانند جایگزین انواع دیگر function word شوند. یکی از دلایلی که function word نتایج خوبی را در تشخیص سبک نوشتاری افراد دارند، این است که مستقل از موضوع هستند. آن‌ها در هر زبانی تعداد تکرار بالایی داشته و از نظر معنایی کم‌ارزش یا بی‌معنی هستند. این کلمات در دستور زبان نقش مهمی دارند و بسیار غیر محتمل است که تحت کنترل آگاهانهٔ نویسندگان قرار گیرند. در زبان فارسی نیز این کلمات تحت عنوان stop words شناخته می‌شوند. مجموعه‌ای شامل ۹۲۲ کلمه از این نوع در مقاله [۲۲] معرفی شده‌است که ما از این مجموعه در برنامهٔ خود استفاده کردیم.

۲-۲-۳. سایر ویژگی‌های نحوی به کار رفته

توزیع کاراکترهای نقطه‌گذاری در متن، ویژگی نحوی دیگر مورد استفاده در این مقاله بوده است. این نوع از اطلاعات زبانی به آسانی برای همهٔ زبان‌های طبیعی در دسترس هستند.

1. shortwords

۲-۳. ویژگی‌های معنایی

ویژگی‌های معنایی به تحلیل دقیق‌تر متن و ویژگی‌های سطح بالاتر می‌پردازند. گراف وابستگی معنایی که شامل ویژگی‌هایی دودویی معنایی و روابط اصلاح معنایی است، زمان و وجه فعل‌های به کار برده شده توسط نویسنده و شباهت‌های معنایی بین کلمات متن، نوع دیگری از این ویژگی‌ها می‌باشند. کلمات و عباراتی که جمله‌واره‌ها را به هم مربوط می‌کنند، به عنوان حروف یا افزوده‌های ربطی شناخته می‌شوند. الگوهای مختلف استفاده از حروف ربط [۲۳] منجر به تفاوت‌های قابل ملاحظه‌ای در سبک‌های متنی می‌شوند.

افزوده‌های تشریحی، گسترشی و تفضیلی انواع افزوده‌های ربطی هستند. افزوده‌های ربطی تشریحی (واقعاً، مانند، حداقل و...) مفهوم موجود در متن را با مثال و یا تأکید مجدد عمیق می‌کنند. این افزوده‌ها می‌توانند اثر خوبی در متن داشته و باعث ایجاد انسجام در سراسر متن گردند. افزوده‌های گسترشی (و، نه، هنوز و...)، اطلاعات مرتبط جدیدی را به متن اضافه می‌کنند که حتی شاید در بعضی مواقع متضاد با مفاهیم فعلی باشند. هرچند که استفاده مکرر از این افزوده، تراکم اطلاعات متن را بالا می‌برد، اگر به اندازهٔ کافی به کار نرود، ممکن است موجب سردرگمی خواننده بین مفاهیم متعدد شود. افزوده‌های تفضیلی (اینجا، همچنین و...) متن را با جزئیات یا ارتباطات منطقی توصیف می‌کنند [۲]. در این تحقیق، در هر یک از متون، نوع افزوده‌ها مشخص شده و تعداد رخداد هر نوع از آن‌ها در متون محاسبه گردیده و به عنوان ویژگی در نظر گرفته شده است. در هر حالت، با اضافه کردن مترادفات، هر گروه این افزوده‌ها را گسترده‌تر نمودیم.

۲-٤. ویژگی‌های وابسته به کاربرد

ویژگی‌های وابسته به کاربرد را می‌توان به منظور نمایش دقیق‌تر تفاوت سبک افراد در حوزهٔ خاصی به کار برد. این ویژگی‌ها همان خصوصیات ساختاری هستند که شامل استفاده از سلام و درود و خداحافظی، استفاده از فرورفتگی‌ها، طول پاراگراف و... می‌باشند.

ویژگی‌های وابسته به کاربردی که در این مقاله استفاده شده‌اند، شامل توزیع کاراکترهای tab، enter، فاصله و linefeed در متن می‌باشند. در نهایت، تعداد ویژگی‌های استخراج شده از پایگاه داده‌ها بیش از ۱۵۰۰ ویژگی بودند.

۳. روش‌های تشخیص هویت نویسنده

در هر مسئله، تشخیص نویسنده، مجموعه‌ای از نویسندگان نامزد و مجموعه‌ای از متون از نویسندگان کاندید که نویسندهٔ آن‌ها مشخص شده است، وجود دارد؛ این مجموعه از متون را مجموعهٔ آموزشی می‌نامند. مجموعه‌ای از متون که نویسندهٔ آن‌ها مشخص نیست، مجموعهٔ آزمون نامیده می‌شود. هر یک از متون مجموعهٔ تست، باید به یکی از نویسندگان مجموعهٔ کاندید نسبت داده شود.

مدل‌های احتمالی از اولین روش‌هایی بودند که برای بررسی مجموعهٔ آزمون به کار رفتند و هم‌اکنون نیز در بسیاری از مطالعات استفاده می‌شوند [۲٤ و ۲۵]. روش‌های دیگر موجود برای دسته‌بندی متون عبارت‌اند از: فشرده‌سازی، Delta [۲٦] و الگوریتم‌های یادگیری ماشین که شامل Discriminant Analysis [۱۳، ۲٦ و ۲۷]، ماشین بردار پشتیبان SVM [۲۸-۳۱]، درخت تصمیم گیری [۳۳ و ۳٤]، شبکه‌های عصبی [۳۵ و ۳٦]، الگوریتم ژنتیک [۳۷] و... می‌باشند.

٤. تشخیص نویسندگان در زبان‌های مختلف

کاربرد تشخیص هویت نویسنده در زبان‌های مختلف روز به روز اهمیت بیشتری می‌یابد. بررسی‌ها نشان می‌دهد که مطالعات گذشته بر روی زبان‌های انگلیسی، یونانی و چینی تمرکز کرده‌اند؛ برای مثال، Stamatatos و همکاران [۳۸] تشخیص هویت نویسنده را بر روی مقاله‌های یونانی پیاده‌سازی کردند. همچنین Peng و همکاران [۳۹] آزمایش‌هایی را روی متون انگلیسی، رمان‌های چینی و روزنامه‌های یونانی انجام دادند.

در اغلب مطالعات گذشته، نتایج این تحلیل در زبان انگلیسی بهتر از زبان‌های دیگر بوده است. پیاده‌سازی تشخیص

نویسنده در زبان‌های مختلف، سختی یکسانی ندارند. اکثر ویژگی‌های سبکی نوشتاری برای زبان انگلیسی طراحی شده‌اند، ولی ممکن است برای زبان‌های دیگر به همان اندازه کارا نباشند. تفاوت‌های ساختاری و زبانی ممکن است پیاده‌سازی و استخراج ویژگی‌ها را مشکل کند؛ برای مثال، نبودن مرز کلمات در زبان چینی، استخراج ویژگی‌های واژگانی مبتنی بر کلمه را بسیار سخت می‌کند.

زبان فارسی عضوی از خانوادهٔ زبان‌های هند و اروپایی شامل زبان‌های انگلیسی و آلمانی است، ولی فارسی با رسم‌الخط عربی و از راست به چپ نوشته می‌شود. در جهان بیش از ده‌ها میلیون نفر به این زبان صحبت می‌کنند و در کشورهای ایران، افغانستان و تاجیکستان به عنوان زبان رسمی به کار می‌رود. این موضوع، اهمیت استفاده از روش‌های تشخیص هویت نویسنده را در این زبان آشکار می‌کند. تاکنون تعیین هویت نویسنده در زبان فارسی به این صورت پیاده‌سازی نشده است. هدف این تحقیق، طراحی و پیاده‌سازی سیستم تشخیص هویت نویسنده در زبان فارسی است.

۵. روند طراحی و پیاده‌سازی

در این مقاله بر اساس کارهای پیشین، روش‌های مختلف دسته‌بندی را بر روی دو مجموعه پایگاه داده فارسی جمع‌آوری شده، آزمایش کرده و نتایج را گزارش نموده‌ایم.

۵-۱. پایگاه‌های داده

مجموعه‌های آموزشی و آزمون مناسب برای یک مسئلهٔ تشخیص هویت نویسنده، باید از نظر موضوع و نوع، بررسی شود تا هویت نویسنده تنها و یا مهم‌ترین عامل جداسازی متون باشد. به طور ایده‌آل، همهٔ متون مجموعهٔ آموزشی باید دقیقاً در ارتباط با یک موضوع باشند. هرچند که تعداد کمی مجموعه با چنین ویژگی‌ای وجود دارند. سن، سطح تحصیلات و ملیت از عوامل دیگری هستند که باید هنگام فراهم کردن مجموعه‌های ارزیابی ایده‌آل، مورد بررسی دقیق قرار گیرند. این بررسی روی مجموعه‌های آزمون به منظور کاهش احتمال انتخاب سبک یک گروه گسترده از افراد نسبت به انتخاب سبک نوشتاری خود فرد، مهم‌اند. همچنین همهٔ متون نویسندگان باید در یک دورهٔ زمانی مشترک نوشته شده باشند تا احتمال تغییر سبک در طول زمان از بین برود. به دلیل نبود پایگاه دادهٔ استاندارد [۹] برای کاربرد تشخیص هویت نویسنده که نیازهای این حوزه را برآورده کند، دو پایگاه داده با عناوین «دانشگاه بوعلی سینا» و «نویسندگان هم‌عصر» را جمع‌آوری کردیم. برای پایگاه نخست، در یک موضوع خاص، از ۲۰ نفر از دانشجویان سال اول کارشناسی مهندسی دانشگاه بوعلی سینا متونی با ۲۰۰۹ کلمه جمع آوری گردید. برای هر فرد، ۱۵۰۰ کلمه برای آموزش و ۵۰۹ کلمه برای آزمایش در نظر گرفته شد.

به علت طول ناکافی [۴۰] و محاوره‌ای بودن متون، که موجب دقیق نبودن بسیاری از پارامترهای اندازه‌گیری شده (توزیع function words در متون) در پایگاه دادهٔ دانشگاه بوعلی شده است، پایگاه دادهٔ دیگری تهیه کردیم. متونی از کتاب‌ها و مقالات ۸ نویسندهٔ هم‌عصر (علی اشرف صادقی، محمدعلی فروغی، مجتبی مینوی، ابوالحسن نجفی، محمد امین ریاحی، احمد سمیعی، فتح‌الله مجتبایی، حسین معصومی همدانی) با موضوعات متفاوت که همگی به تحلیل متون ادبی پرداخته بودند، به عنوان پایگاه دادهٔ دوم جمع‌آوری نمودیم. برای هر نویسنده، یک یا دو سند با ۷۷۵۰ کلمه جمع‌آوری گردید که از این تعداد، ۵۰۰۰ کلمه برای آموزش و ۲۷۵۰ کلمه برای آزمایش در نظر گرفته شد. گفتنی است که هدف از تشخیص نویسنده، بازیابی متون بر اساس نویسندهٔ آن‌ها که تعداد زیادی نویسنده با متون بلند داشته باشیم، نیست، بلکه هدف یافتن نویسندهٔ متن مورد نظر از بین تعداد کمی نویسندهٔ مورد سوءظن است.

۵-۲. روش‌های دسته‌بندی

در این تحقیق، پس از عمل پیش‌پردازش روی ویژگی‌های استخراج شده، روش‌های دسته‌بندی ارائه شدهٔ پیشین را روی هر دو پایگاه داده آزمودیم و نتایج را گزارش نمودیم.

در این آزمون‌ها از دو روش دسته‌بندی K Nearest Neighbor (KNN) و Delta [۲۶] استفاده شد. در مرحلهٔ بعدی با ترکیب این روش‌های دسته‌بندی با الگوریتم ژنتیک سعی در بهبود این روش‌ها با استفاده از انتخاب ویژگی داشتیم.

۵-۲-۱. روش KNN

روش KNN مدلی احتمالاتی است که در آن، برای نسبت دادن یک متن به نویسندهٔ با متون مشابه‌تر، به صورت زیر عمل می‌کنیم. در این روش، K تا از متون نویسندگان را که شباهت بیشتری با متن مورد نظر دارند، شناسایی کرده و نویسنده‌ای را که تعداد متون بیشتری از آن در بین K متن بود، به عنوان نویسندهٔ متن مورد نظر انتخاب شده است. معیار فاصله در این روش، فاصلهٔ اقلیدسی در نظر گرفته شده است.

۵-۲-۲. روش Delta

در سال ۲۰۰۲ JohnF. Burrows [۲۶] مقیاس جدیدی را برای تشخیص هویت نویسنده به نام Delta ارائه کرد. این معیار، معادل است با مقیاسی برای محاسبهٔ تفاوت نرمال شده بین تعداد تکرار کلمات در D و D' که با معادلهٔ (۶) محاسبه می‌گردد. $\{W_i\}$ مجموعه‌ای از nکلمهٔ مورد علاقه (معمولاً کلمه‌های با بیشترین تکرار برای محاسبهٔ Delta)، σ_i انحراف معیار استاندارد آن کلمه در مجموعهٔ مورد مقایسه و $f_i(D)$ تعداد تکرار کلمهٔ w_i (تعداد تکرار iامین کلمهٔ مورد نظر) است [۴۱]؛ بنابراین، برای محاسبهٔ Delta بین سند D و سند D' از عبارت زیر استفاده می‌کنیم:

$$\Delta_B^{(n)}(D-D') = \sum_1^n \frac{1}{\sigma_i}|f_i(D)-f_i(D')| \qquad (۶)$$

متغیر n تعداد کلماتی است که دفعات تکرار آن‌ها را در هر سند محاسبه می‌کنیم و B نمایانگر رابطهٔ اصلی Burrows می‌باشد. این رابطه به وضوح نشان می‌دهد که Delta، نویسنده‌های نامزد D' را با توجه به فاصلهٔ آن‌ها از متن آزمون D رتبه‌بندی می‌کند. بعد از اینکه تفاوت تعداد تکرار کلمات محاسبه شد، این مقدار توسط فاکتور $\frac{1}{\sigma_i}$ کوچک می‌شود؛ بنابراین Delta سبب به وجود آمدن برداری وزن‌دار در فضای ویژگی‌ها می‌شود و یک سند آزمون در دسته‌ای قرار می‌گیرد که بردار وزن‌دار آن، کمترین فاصله را با بردار وزن‌دار سند شناخته شدهٔ آموزشی داشته باشد.

۵-۲-۳. روش الگوریتم ژنتیک

الگوریتم ژنتیک، روشی آماری برای بهینه‌سازی و جستجو است که جزئی از محاسبات تکاملی است. ایدهٔ محاسبات تکاملی، اولین بار در سال ۱۹۶۰ توسط رچنبرگ [۴۲] که در زمینهٔ الگوریتم های تکاملی تحقیق می‌کرد، به وجود آمد. الگوریتم ژنتیک [۴۱] با مجموعه‌ای از جواب‌ها که از طریق کروموزوم‌ها نشان داده می‌شوند، شروع می‌شود. این مجموعه جواب‌ها جمعیت اولیه نام دارند. در این الگوریتم، جواب‌های حاصل از یک جمعیت برای تولید جمعیت بعدی استفاده می‌شوند. در این فرآیند، امید است که جمعیت جدید نسبت به جمعیت قبلی، بهتر باشد. انتخاب بعضی جواب‌ها از میان کل جواب‌ها (والدین) به منظور ایجاد جواب‌های جدید یا همان فرزندان بر اساس میزان مطلوبیت آن‌ها می‌باشد. این فرآیند تا برقراری شرط از پیش تعیین شده (مانند تعداد جمعیت‌ها یا میزان بهبود جواب) ادامه می‌یابد.

الگوریتم ژنتیک از بین ویژگی‌های استخراج شده از متون، ویژگی‌های مؤثر در تعیین هویت نویسنده را انتخاب می‌کند. تابع برازش (هزینه) الگوریتم ژنتیک، میزان دقت معیار شباهت روش‌های دسته‌بندی KNN و Delta در تشخیص هویت نویسنده است. هر کروموزوم استفاده شده در این الگوریتم ژنتیک به تعداد ویژگی‌های هر پایگاه داده، دارای ژن است، به طوری که هر ژن به یک ویژگی مربوط می‌گردد. اگر این ژن صفر باشد، آن خصوصیت در دسته‌بندی یا ارزیابی حذف می‌شود، ولی اگر مقدار ژن مربوط، یک باشد، آن ویژگی در دسته‌بندی شرکت می‌کند.

الگوریتم ژنتیک در این مقاله از عمگرهای جهش، تقاطع و نخبه‌گرایی برای تولید جمعیت نسل بعدی (فرزند) استفاده می‌کند. در این تحقیق، تعداد جمعیت اولیه بیست، تعداد نخبه

یک نفر است. برای انتخاب کروموزوم‌های مناسب جهت تولید نسل بعدی، الگوریتم از مدل چرخ رولت، عملگر تقاطع تک نقطه‌ای با نرخ تقاطع ۰/۸ و عملگر جهش با نرخ ۰/۲ استفاده می‌کند.

۶. مقایسه

مقایسهٔ دو روش دسته‌بندی بر روی ویژگی‌های استخراج شده از متون، در جدول (۱) خلاصه شده‌اند. روش Delta که به طور خاص، برای کاربرد تشخیص نویسنده ارائه شده، نتایج خوبی را به دست آورده است. در این روش، ویژگی‌های تکرار رخداد function words متون آزمایشی، به این الگوریتم داده شده و فاصلهٔ آن‌ها با تکرار رخداد این کلمات در متون آموزشی به دست می‌آید. نویسنده‌ای که متن آموزشی آن کمترین فاصله را با متن آزمایشی مورد نظر داشته باشد یعنی Delta کمتر، به عنوان نویسندهٔ آن متن مشخص می‌شود. روش دیگر استفاده از الگوریتم KNN برای دسته‌بندی متون است. این الگوریتم نیز نتایج خوبی را به دست آورده است. در این الگوریتم، متون آموزشی هر نویسنده به قطعاتی با کلمات مساوی تقسیم شده و برای آموزش به این الگوریتم داده شدند. این الگوریتم برای هر متن آزمایشی، k تا از نزدیک‌ترین متون آموزشی به آن را شناسایی کرده و نویسنده‌ای که تعداد متون آموزشی بیشتری در بین K متن داشته باشد، به عنوان نویسندهٔ آن متن آزمایشی انتخاب می‌شود.

برای انتخاب ویژگی‌های متمایزکننده به منظور بهبود دقت روش‌های دسته‌بندی، الگوریتم ژنتیک را با روش‌های دسته‌بندی به کار گرفته شده، ترکیب کردیم. الگوریتم ژنتیک به طور تصادفی، ویژگی‌هایی را انتخاب می‌کند و روش‌های دسته‌بندی روی این ویژگی‌ها انجام می‌شوند. الگوریتم ژنتیک با عملگرهای جهش و تقاطع و نخبه‌گرایی سعی در یافتن بهترین ویژگی‌ها را دارد. ویژگی‌هایی که بهترین دقت را در دسته‌بندی داشته، بهترین ویژگی خوانده می‌شوند. همان‌طور که در جدول (۱) نیز قابل مشاهده است، با اضافه کردن الگوریتم ژنتیک، دقت به طور قابل توجهی افزایش می یابد.

نتایج ذکر شده از طریق میانگین‌گیری روی k-fold cross validation روی متون آموزشی و تست به دست آمده است. همان‌طور که گفته شد، پایگاه داده بوعلی سینا به دلیل کوتاه و محاوره‌ای بودن، دقت کمی را نسبت به پایگاه داده نویسندگان هم‌عصر تولید کرده است.

۷. بحث و نتیجه‌گیری

در این مقاله، یک سیستم تشخیص هویت نویسنده با موفقیت در زبان فارسی پیاده‌سازی شده است. در این مقاله، علاوه بر پیاده‌سازی سیستم تشخیص هویت در زبان فارسی، ابزاری برای تشخیص اسم، قید و صفت طراحی و پیاده‌سازی شده است که از ساختارهای موجود در کلمات برای این منظور استفاده می‌کند. در جدول (۱) نتایج پیاده‌سازی روش‌های تشخیص هویت نویسنده روی متون فارسی نشان داده شده است؛ بنابراین، هرچه طول متون آموزشی و آزمایشی کمتر باشد، تشخیص هویت نویسنده دقت پایینی دارد. با اضافه کردن ویژگی‌های نحوی بیشتری در زمینهٔ قواعد بازنویسی جملات و قواعد معنایی، می‌توان دقت این روش را بهبود داد. همچنین می‌توان به تشخیص هویت نویسنده در محتوای ایمیل‌ها یا وبلاگ‌ها و یا متون مورد مشاجرهٔ قدیمی پرداخت. البته برای متون با طول کم، ویژگی‌های استخراج شده می‌بایست با دقت انتخاب شوند تا بتوانند گویای سبک نوشتاری نویسنده باشند.

جدول (۱): نتایج پیاده‌سازی دو روش دسته بندی و ترکیب آن‌ها با الگوریتم ژنتیک روی دو پایگاه داده

دقت	روش دسته‌بندی	پایگاه داده
٪۵۰	Delta	بوعلی سینا
٪۷۰	KNN	
٪۵۰	Delta+GA	
٪۸۰	KNN+GA	
٪۸۷	Delta	نویسندگان هم‌عصر
٪۱۰۰	KNN	
٪۸۷/۵	Delta+GA	
٪۱۰۰	KNN+GA	

۸. پیشنهادات برای کارهای آینده

روش‌های تشخیص هویت نویسنده به صورت خودکار همچنان ادامه دارد تا بتوان به دقت بالاتر و درجهٔ اطمینان بیشتری برای استفاده در مراجع قضایی دست یافت. در ادامهٔ این تحقیق می‌توان ویژگی‌های دیگری چون ویژگی‌های معنایی و ساختار جملات و یا گرامر مورد استفاده، استفاده از غلط‌های املایی، استفاده از زمان افعال و وجه افعال در جملات را که نیازمند ابزارهای پردازش زبان طبیعی خیلی قدرتمند است، به ویژگی‌های مورد استفاده در سیستم تشخیص هویت نویسنده زبان فارسی اضافه کرد. می‌توان روش‌های دیگر بهینه‌سازی همچون الگوریتم کلونی و سایر را به کار برد و نتایج را با دقت کارهای فعلی مقایسه کرد. می‌توان تشخیص نویسنده را برای پایگاه داده‌های با متون کوتاه مانند ایمیل‌ها و مطالب وبلاگ‌ها و دارای استاندارد دیگر به کار گرفت. در ادامهٔ این پایان‌نامه می‌توان سیستمی برای تشخیص چند نویسندگی در زبان فارسی طراحی کرد.

سپاس‌گزاری

از سرکار خانم دکتر نجمه نظری، استادیار گروه زبان و ادبیات فارسی دانشگاه بوعلی سینا همدان که در جمع‌آوری پایگاه داده آن دانشگاه ما را یاری کردند، تقدیر و تشکر می‌کنیم.

از دفتر تحقیقات فاتب که از این تحقیق حمایت نمودند، نیز تشکر می‌کنیم.

مراجع

[۱] طباطبایی، ع.، **ساختمان واژه و مقولهٔ دستوری: تشخیص مقولهٔ دستوری واژه‌ها بر اساس ملاک‌های صرفی**، تهران، پژوهشگاه فرهنگ هنر و ارتباطات وزارت فرهنگ و ارشاد اسلامی، ۱۳۸۸.

[۲] جعفری. آ.، **بررسی افزوده‌ها در زبان فارسی بر اساس رویکردهای نقشی و سوری**، ویژه‌نامه دستور شماره ۵، ضمیمه نامه فرهنگستان، جلد ۵، خرداد ۱۳۸٤.

[3] Mendenhall, T.C., *The characteristic curves of composition*, Science, IX, 237–249, 1887.

[4] Zipf, G.K., *Selected studies of the principle of relative frequency in language*, Cambridge, MA: Harvard University Press, 1932.

[5] Yule, G.U., *On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship*, Biometrika, 30, 363–390, 1938.

[6] Yule, G.U., *The statistical study of literary vocabulary*, Cambridge University Press, 1944.

[7] Mosteller, F., & Wallace, D.L., *Inference and disputed authorship: The Federalist*, Reading, MA: Addison-Wesley, 1964.

[8] Shaker, K&Corne, D , "Authorship attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis". in*2010 UK Workshop on Computational Intelligence, UKCI*. UK Workshop on Computational Intelligence, UKCI 2010, Colchester, United Kingdom, 2010.

[9] Türkoğlu F., Diri B., and Amasyali M. F., "Author attribution of Turkish texts by feature Mining". Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues. . Springer Berlin Heidelberg , 1086–1093, 2007.

[10] Raza A. A., Athar A. and Nadeem S., " N-Gram Based Authorship Attribution in Urdu Poetry", Proceedings of the Conference on Language & Technology , 88-93, 2009.

[11] Holmes, D.I., *The evolution of stylometry in humanities scholarship*, Literary and Linguistic Computing, 13(3), 111–117, 1998.

[12] Stamatatos, E., *A survey of modern authorship attribution methods*, Journal of the American Society for Information Science and Technology, 60(3): 538–56, 2009.

[13] de Vel, O., Anderson, A., Corney, M., &Mohay, G., *Mining e-mail content for author identification forensics*, SIGMODRecord, 30(4), 55–64, 2001.

[14] Stamatatos, E., Fakotakis, N.,&Kokkinakis, G., *Automatic text categorization in terms of genre and author*, Computational Linguistics, 26(4), 471–495, 2000.

[15] Baayen, R., van Halteren, H., &Tweedie, F., *Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution*, Literary and Linguistic Computing, 11(3), 121–131, 1996.

[16] Yule, G.U., *The statistical study of literary vocabulary*, Cambridge University Press, 1944.

[17] Honore, A., *Some simple measures of richness of vocabulary*, Association for Literary and Linguistic Computing Bulletin, 7(2), 172–177, 1979.

[18] Brunet, E., Vocabulairede JeanGiraudoux: *Structure et Evolution*. Slatkine, 1978.

Cybernetics, 2 (pp. 1204–1207). Washington, DC: IEEE, 2004.

[33] Zhao, Y., &Zobel, J., *Effective and scalable authorship attribution using function words,* In Proceedings of the 2nd Asia Information Retrieval Symposium. Berlin, Germany: Springer, 2005.

[34] Uzuner, O., &Katz, B., *A comparative study of language models for book and author recognition,* In Proceedings of the 2nd International Joint Conference on Natural Language Processing (pp. 969–980). Berlin, Germany: Springer, 2005.

[35] Merriam, T., &Matthews, R., *Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe,* Literary and Linguistic Computing, 9(1), 1–6, 1994.

[36] Matthews, R., & Merriam, T., *Neuralcomputation in stylometry: An application to the works of Shakespeare and Fletcher,* Literary and Linguistic Computing, 8(4), 203–209, 1993.

[37] Holmes, D.I., & Forsyth, R., The *Federalist revisited: New directions in authorship attribution,* Literary and Linguistic Computing, 10(2), 111–127, 1995.

[38] Stamatatos, E., Fakotakis, N., &Kokkinakis, G., *Computer-based authorship attribution without lexical measure,* Computers and the Humanities, 35(2), 193-214, 2001.

[39] Peng, F., Schuurmans, D., Keselj, V., & Wang, S., *Automated authorship attribution with character level language models.* Paper presented at the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003), 2003.

[40] Luyckx, K. and Daelemans, W., *The effect of author set size and data size in authorship attribution*, *Literary and Linguistic Computing,* vol. 26, pp. 35-55, April 1, 2011.

[41] Argamon, S., *Interpreting Burrows' Delta: Geometric and probabilistic foundation,* Literary and Linguistic Computing, 23(2), 131–147, 2008.

[42] Rechenberg.I.,*volutionstrategie:"OptimierungTechnisherSystemenachPrinzipien des Biologischen Evolution",* Fromman-HozlboogVerlag, Stuttgart, 1973.

[43] Holmes, D.I., & Forsyth, R., *The Federalist revisited: New directions in authorship attribution,* Literary and Linguistic Computing, 10(2), 111–127, 1995.

[19] Sichel, H.S., *On a distribution law for word frequencies,* Journal of the American Statistical Association, 70:542–547, 1975.

[20] Simpson, E.H., *Measurement of diversity*. Nature, 163:688, 1949.

[21] Cavnar, W. B. and Trenkle, J. M., *N-gram-based text categorization*, In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval (Las Vegas, US, 1994), pp. 161–175, 1994.

[22] Davarpanah M.R., SanjiM., AramidehM., *Farsi lexical analysis and stop word list*, Library Hi Tech, Vol. 27 Iss: 3, pp.435 – 449, 2009.

[23] Argamon, S., Whitelaw, C., Chase, P., Hota, S.R., Garg, N., &Levitan, S., *Stylistic text classification using functional lexical features*, Journal of the American Society for Information Science and Technology, 58(6), 802–822, 2007.

[24] Mosteller, F. and Wallace, D., "*Inference and Disputed Authorship: The Federalist"*: Addison-Wesley, 1964.

[25] Peng, F*., et al.*, "*Augmenting naive Bayes classifiers with statistical language models*," Information Retrieval*,* pp. 317–345, 2004

[26] Burrows, J.F.*"Delta:" A measure of stylistic difference and a guide to likely authorship*, Literary and Linguistic Computing, 17(3), 267–287, 2002.

[27] Chaski, C.E., *Who's at the keyboard?Authorship attribution in digital evidence investigations*, International Journal of Digital Evidence, 4(1), 2005.

[28] Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Carayannis, G., &Tambouratzis, D., *Discriminating the registers and styles in the Modern Greek language*—Part 2: Extending the feature vector to optimize author discrimination, Literary and Linguistic Computing, 19(2), 221–242, 2004.

[29] Sanderson, C., &Guenter, S., *Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation,* In Proceedings of the International Conference on Empirical Methods in Natural Language Engineering (pp. 482–491). Morristown,NJ: Association for Computational Linguistics, 2006.

[30] Li, J., Zheng, R., &Chen, H., *From fingerprint to writeprint.* Communications of the ACM, 49(4), 76–82, 2006.

[31] Diederich, J., Kindermann, J., Leopold, E., &Paass, G., *Authorship attribution with support vector machine*, Applied Intelligence, 19(1/2), 109–123, 2003.

[32] Teng, G., Lai, M., Ma, J., &Li,Y., *E-mail authorship mining based on SVM for computer forensics,* In Proceedings of the International Conference on Machine Learning and