تشخیص خودکار جنسیت نویسنده در متون فارسی

مهدی مرادی و محمد بحرانی
مرکز زبان‌ها و زبان‌شناسی، دانشگاه صنعتی شریف، تهران، ایران

چکیده

با گسترش روزافزون استفاده از اینترنت، شاهد تبادل هزاران گیگابایت اطلاعات متنی در فضای مجازی هستیم. فضای سایبر این امکان را به اشخاص می‌دهد که هویت حقیقی خود را مخفی کنند و با هویت ساختگی جدیدی وارد آن شوند. از این‌رو اهمیت حفظ امنیت این فضا، کنترل بر محتوای تولیدشده توسط کاربران و شناسایی مشخصات تولیدکنندگان محتوا هر روز پررنگ‌تر می‌شود. موضوع مورد بررسی در این پژوهش که مربوط به حوزهٔ شناسایی نویسنده است، شناسایی خودکار جنسیت نویسنده متن فارسی است. به‌منظور شناسایی جنسیت، با توجه به مطالعات زبان‌روان‌شناختی صورت گرفته، ۴۸ ویژگی روان‌شناختی و سبک‌شناسی تعریف شد. دو پیکرهٔ متنی جهت آموزش طبقه‌بندها تهیه و سپس برای شناسایی جنسیت، سه الگوریتم یادگیری ماشینی مختلف (ماشین‌بردار پشتیبان، بیز ساده و درخت تصمیم) طراحی شد. نتایج اعتبارسنجی متقابل ده‌تایی نشان داد که بیش‌ترین دقت مربوط به طبقه‌بند درخت تصمیم با دقت ٪۷۳/۸ است.

واژگان کلیدی: تشخیص جنسیت، شناسایی نویسنده، متن کاوی.

۱- مقدّمه

عصر ارتباطات و گسترش استفاده از نامه‌های الکترونیکی، پیام‌های فوری (IM)، اسناد، وبلاگ‌ها، مقالات خبری، صفحات خانگی، تالارهای گفتگو و شبکه‌های اجتماعی، متن را به اصلی‌ترین وسیلهٔ ارائه و انتقال اطلاعات تبدیل کرده است.

گم‌نامی[1] به‌عنوان یکی بارزترین ویژگی‌های جامعه مجازی، (ژنگ و همکاران، ۲۰۰۶) از یک‌سو امکان بیان آزادانه عقاید و نظرات و از سوی دیگر امکان جرایم اینترنتی را فراهم می‌سازد. بی‌چهرگی و بی‌نامی افراد در اینترنت نه‌تنها پیگرد قانونی جرم و مجرمان را در فضای رایانه‌ای مشکل می‌سازد، بلکه زمینهٔ گسترش بیش‌تر آن را نیز فراهم می‌کند. مجرمان اینترنتی برای ناشناس‌ماندن، از سرورهای گمنام‌کننده استفاده می‌کنند و یا این‌که با ارائه اطلاعات نادرست در مورد جنسیت، مکان، سن و ملیت خود دست به اقدامات خراب‌کارانه می‌زنند. از این رو استفاده از روش‌های جدید جهت ردیابی هویت بزه‌کاران ضروری به‌نظر می‌رسد. علاوه‌بر این تشخیص جنسیت نویسنده می‌تواند کاربردهای تجاری متعددی دیگری نیز داشته باشد. برای مثال با مشخص‌شدن جنسیت نویسنده می‌توان فهمید که زنان یا مردان بیش‌تر درباره چه موضوعات یا محصولاتی صحبت می‌کنند و از چه خدمات یا کالاهایی بیزارند. دسترسی به این اطلاعات برای تبلیغات هدفمند بسیار حائز اهمیت است(آرجون و همکاران، ۲۰۱۰). در حوزهٔ ترجمه ماشینی[2]، وقتی قرار است ترجمه بین دو زبانی که نظام ضمیرهایشان متفاوت است، انجام شود مشخص‌بودن جنسیت ضمیر، دقت ترجمه را افزایش خواهد داد (برای مثال ترجمه جمله "من تنها هستم" بسته به جنسیت نویسنده، در ایتالیایی دو جمله "Sono solo" برای مرد و "Sono sola" برای زن خواهد بود). کاربرد شناسایی جنسیت از دیدگاه علوم اجتماعی و تعیین خصوصیات جمعیت‌شناسی محتوای تولیدشده در اینترنت نیز می‌تواند کاربردی باشد. مطالعات

[1] Anonymity

[2] Machine translation

زبان‌شناختی بر روی تغییرات زبان زنان و مردان در بازه‌های زمانی یا در حوزه‌های مختلف از کاربردهای دیگر شناسایی جنسیت نویسنده خواهد بود.

مطالعات روان‌شناختی نشان می‌دهد که کلمات مورد استفاده هر فرد بیان‌گر ویژگی‌های شخصیتی و روانی آن فرد است (پنباکر و همکاران، ۱۹۹۹)، (استیرمن و همکاران، ۲۰۰۱). نتایج پژوهش‌های روان‌شناسان نشان می‌دهد که ویژگی‌های سبک‌شناسی[1] هر شخص (مانند کاربرد ضمایر، حروف ربط و یا حروف اضافه) با رفتارها و احساساتش در ارتباط است (پنباکر ،۲۰۰۳) . پژوهش‌گران از این ویژگی‌های زبان‌شناختی به‌عنوان ابزاری قانونی[2] جهت شناسایی نویسنده، جنسیت، سن و یا سطح تحصیلات او استفاده می‌کنند.

۲- مروری بر ادبیات

تشخیص جنسیت نویسندهٔ متن، در حوزه کارهایی قرار می‌گیرد که مربوط به شناسایی نویسنده[3](دول و همکاران، ۲۰۰۱) یا تصدیق نویسنده[4] است

(کوپل و همکاران، ۲۰۰۹). در فرایند شناسایی نویسنده، پژوهش‌گر به‌دنبال یافتن هویت واقعی نویسنده‌ای است که هویتش را مخفی نگاه داشته است.

نخستین پژوهش‌های کمّی مربوط به شناسایی نویسنده در سال ۱۸۷۷ توسط مندن هال هواشناس اهل اوهایو صورت گرفت، که پیشنهاد کرد که از توزیع طول کلمات[5] به‌عنوان ویژگی وابسته به نویسنده استفاده شود (مندنهال، ۱۸۸۷). وی در سال ۱۹۰۱ از این تکنیک برای شناسایی نویسنده اشعار مورد تردید شکسپیر-بیکون استفاده کرد (مندنهال، ۱۹۰۱) . از پیش‌گامان رویکرد آماری به مسئلهٔ سبک‌شناسی در عصر رایانه، موستلر و والاس بودند که در سال ۱۹۶۴ از لغات دستوری[6] و تحلیل بیزی[7] (طبقه‌بند بیز ساده) برای شناسایی هویت نویسندگان مقالات فدرالیست[8] استفاده کردند (موستلر و والاس ، ۱۹۶۴).

با گسترش رایانه، استفاده از ویژگی‌های سبک‌شناسی برای شناسایی نویسنده بیش از پیش مورد توجه پژوهش‌گران قرار گرفت، به‌طوری که تاکنون بیش از هزار ویژگی سبک‌شناسی مانند ویژگی‌های واژه‌بنیاد، نویسه‌بنیاد[9] (به‌عنوان مثال (یول، ۱۹۹۴)) و نیز ویژگی‌های مبتنی بر کلمات دستوری (به‌عنوان مثال موستلر و والاس(موستلر و والاس ، ۱۹۸۴)) و علایم نگارشی[10](باین و همکاران، ۲۰۰۲) برای این منظور تعریف شده است. هر چند پژوهش‌های بسیار زیادی بر روی زبان انگلیسی صورت گرفته و نتایج قابل قبولی نیز به‌دست آمده است (زاچاری و همکاران، ۲۰۱۲) ولی با این وجود کارهای محاسباتی صورت گرفته بر روی شناسایی نویسندهٔ زبان فارسی بسیار محدود و پراکنده بوده است (زینب فرهمند پور و همکاران, اسفند ۹۰). تاکنون روش‌های مختلفی برای شناسایی نویسنده استفاده شده است. پژوهش‌های نخستین در این حوزه بیش‌تر معطوف به استفاده از هیستوگرام توزیع طول کلمات (مندنهال، ۱۸۸۷)، طبقه‌بندهای بیزی (موستلر و والاس، ۱۹۸۴)، تحلیل مؤلفه‌های اصلی[11] (بوروز، ۱۹۸۴)، تحلیل خوشه[12] (هولمز، ۱۹۹۲) و غیره اشاره کرد. پیشرفت‌های سخت‌افزاری و افزایش قدرت پردازش رایانه‌ها، پژوهش‌های این حوزه را به سوی استفاده از تکنیک یادگیری ماشین مانند درخت تصمیم[13] (آپت و همکاران، ۱۹۹۸) شبکه‌های عصبی[14](تویدی و همکاران، ۱۹۹۶)، ماشین‌بردار پشتیبان[15] سوق داد (دیدریچ و همکاران، ۲۰۰۰).

مسئلهٔ مورد کنکاش در این پژوهش بررسی تمایز یا عدم تمایز زبان زنان و مردان است، اینکه جنسیت نمودی در سبک زبانی افراد دارد یا نه و کدام ویژگی‌های زبانی نقش پررنگ‌تری در تمایز زبان این دو جنس دارند. هر چند این تفاوت به قدری نیست که ارتباط بین زنان و مردان را مختل کند، ولی همان‌طور که نتایج پژوهش‌ها نشان می‌دهد، زنان و مردان هر کدام گونهٔ متفاوتی از زبان را به کار می‌برند. طی چند دهه گذشته، پژوهش‌گران از زوایای مختلفی، مسئلهٔ

---

[1] Linguistic style
[2] Forensic tools
[3] Authorship identification
[4] Authorship verification
[5] Word-length distribution
[6] Function words
[7] Bayesian analysis
[8] فدرالیست مجموعه‌ای از ۸۵مقاله است که درسال‌های۱۷۸۷و۱۷۸۸میلادی در دو نشریه در شهر نیویورک برای تشویق مردم این ایالت به امضای قانون اساسی پیشنهادی برای ایالات متحده درکنوانسیون۱۷۸۷نوشته شدند"،پوبلیوس" نام مستعار نویسندگان این مقالات بود. (ویکی پدیا)
[9] Character-based
[10] Punctuation
[11] Principle component analysis
[12] Cluster analysis
[13] Decision tree
[14] Neural networks
[15] Support vector machine

تفاوت بین زنان و مردان و نیز مسئلهٔ جنسیت و زبان را مورد پژوهش قرار داده‌اند (تالبوا، ۱۹۹۸). این نوع پژوهش‌ها که در آغاز بیش‌تر معطوف به بررسی تأثیر متغیر جنسیت در رفتار کلامی افراد در سطح آوایی و نیز شیوهٔ تعامل بودند (وواداک و همکاران، ۲۰۰۷). با انتشار مقاله روبین لیکاف ("جایگاه زنان در زبان") در ۱۹۷۵ بدل به نقطه عطفی در مطالعات زبان و جنسیت شد. لیکاف در این مقاله با برشمردن ویژگی‌های واژگانی، نحوی و کاربردشناختی[1]، مانند استفاده از واژه‌های خاص، پوچواژه[2]ها، ضمیمه‌های سؤالی[3]، عبارات غیر صریح[4] و قواعد از نوع تصحیح افراطی[5] سعی در شناسایی سبک نگارش زنان داشت (لیکاف، ۱۹۷۵). تالبوت در کتاب خود می‌گوید: الگوهای زبانی در گفتار زنان و مردان، نشان‌گر طبقه اجتماعی آنان است. در محیط کار، زنان طبقه پایین‌تر هنگام گفت گو با رؤسایشان از زبان مؤدبانه‌تری استفاده می‌کنند (تالبوت، ۱۹۹۸). طبق نتایج پژوهش مولاک و همکاران، در گفتگوهای دو نفره، جملات پرسشی بیش‌تر در کلام زنان و جملات امری بیش‌تر در کلام مردان دیده می‌شود (مولاک، ۱۹۹۸). باید توجه کرد که پژوهش‌گران بین مفهوم جنس[6] و جنسیت[7] تفاوت قائلند. جنس به ابعاد بیولوژیکی مردانگی و زنانگی فرد اشاره دارد؛ درحالی‌که جنسیت به صفات و ویژگی‌های اجتماعی دو جنس اطلاق می‌شود (اونگر، ۱۹۷۹). جنس هر فرد از بدو تولدش مشخص است؛ حال آن‌که جنسیت را جامعه تعیین می‌کند؛ که به ویژگی‌های شخصی و روانی شخص دلالت دارد. به عبارت دیگر هر زنی رفتار زنانه ندارد و هر مردی نیز لزوماً مردانه رفتار نمی‌کند. مسئلهٔ مورد بررسی این پژوهش تفاوت‌های زبانی وابسته به جنسیت است و به تفاوت‌های مرتبط با جنس پرداخته نخواهد شد.

۲-۱- جنسیت و زبان فارسی

هر چند کارهایی بر روی زبان‌های دیگر انجام شده، ولی در مورد شناسایی خودکار جنسیت نویسنده در زبان فارسی کار چندان مهمی صورت نگرفته است. مروری برکارهای (غیر محاسباتی) انجام‌شده بر روی زبان فارسی نشان از انجام پژوهش‌هایی بر روی گونهٔ نوشتاری و گفتاری فارسی دارد. نگار داوری اردکانی مقالات و پایان‌نامه‌های مرتبط با این موضوع را در مقاله‌ای فهرست کرده است (داوری اردکانی، ۱۳۸۷). در این بین، از جمله پژوهش‌هایی که موضوع را نه از دیدگاه روان‌شناختی و اجتماعی بلکه از نگاه زبان‌شناختی (بررسی نظام ساختاری زبان از نظر انعکاس موضوع جنسیت) مورد کنکاش قرار داده‌اند، می‌توان به بررسی جنسیت در واژگان (فارسیان، ۱۳۷۸)، نگاه مقابله‌ای به زبان فارسی و انگلیسی (چاکانی، ۲۰۰۰) و جنسیت و شمول معنایی (شجاع رضوی، ۱۳۸۶) اشاره کرد.

ما در این پژوهش با تعریف ۴۸ ویژگی مختلف، توانستیم با دقت به‌نسبه بالایی، جنسیت نویسنده را تشخیص دهیم.

۳- تبیین مسئله

مسئلهٔ تشخیص جنسیت نویسنده، نوعی مسئلهٔ طبقه‌بندی دوطبقه[8] است. که هدف در این مسئله قراردادن متن بی‌نام e در یکی از طبقه‌های {زن، مرد} است (رابطه ۱).

$$e\in \begin{cases} \text{اگر نویسنده e زن باشد کلاس } +1 \\ \text{اگر نویسنده e مرد باشد کلاس } -1 \end{cases} \quad (1)$$

برای بررسی فرض شماره ۱، باید ویژگی‌های تعریف شوند که در بیشتر متون نوشته‌شده توسط زنان یا متون نوشته‌شده توسط مردان به‌طورنسبی ثابت باشند. پس از مشخص‌شدن فضای ویژگی‌ها، هر متن e را می‌توان به‌صورت برداری متشکل از d بعد که در آن d تعداد کل ویژگی‌ها است، نشان داد. با داشتن مجموعه‌ای از متون از قبل دسته‌بندی‌شده، طبقه‌بند آموزش داده شده و سپس از آن برای دسته‌بندی متون بدون برچسب استفاده می‌شود. بیان مسئله به زبان ریاضی عبارت است از آموزش طبقه بند $y=f(x)$، از روی دادگان آموزش $D=(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$.

$X=\{x_i,i=1,2,\ldots N\}$ مجموعه از نمونه‌ها است که در آن $x_i=[x_{i1},x_{i2},\ldots,x_{id}]$ برداری متشکل از d بعد است. اگر $Y=\{y_i,i=1,2,\ldots,N\}$ نشان‌دهنده مجموعهٔ برچسب باشد، $y_i\in\{+1,-1\}$ نشان‌دهندهٔ طبقهٔ ۱- (مرد) و طبقه ۱+ (زن) و N تعداد نمونه در مجموعه داده است. در کل شناسایی جنسیت نویسنده را می‌توان به چهار مرحله تقسیم کرد شکل (۱).

۱. تهیهٔ پیکره دارای برچسب جنسیت

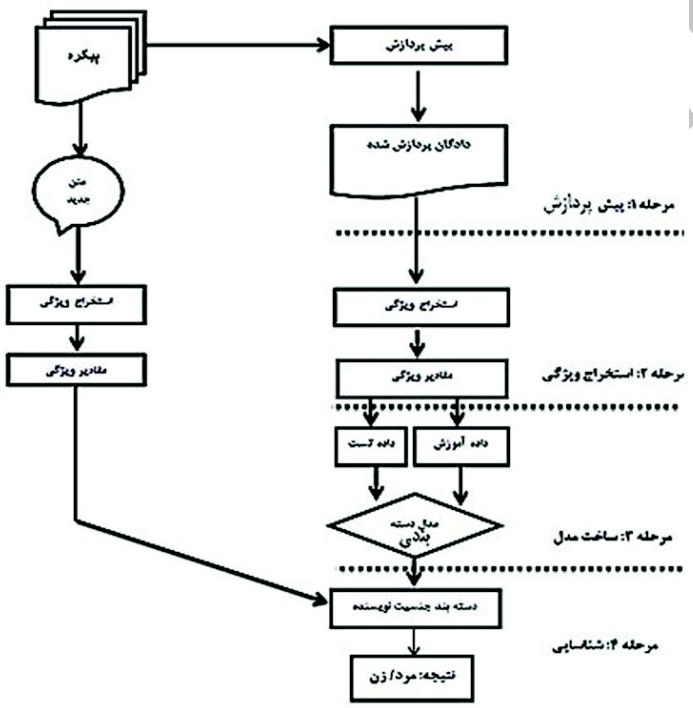
[1] Pragmatic
[2] Expletive
[3] Tag questions
[4] Hedges
[5] Hypercorrect
[6] Sex
[7] Gender
[8] Binary Classification problem

۲. شناسایی ویژگی‌های زبانی متمایز کنندهٔ جنسیت
۳. استخراج خودکار مقادیر هر ویژگی از هر متن
۴. ساخت مدل دسته‌بندی برای شناسایی جنسیت نویسنده در متون هدف

۴- تهیهٔ پیکره آموزشی (دارای برچسب جنسیت)

۴-۱ - پیکره بی‌جن‌خان (متون ادبی)

بخشی از داده‌های مورد استفاده در این پژوهش از پیکره متنی بی‌جن‌خان (بی‌جن‌خان، ۱۳۸۶) به دست آمد. از آن‌جا که متون موجود در این پیکره فاقد برچسب مشخص کننده جنسیت بودند، با توجه به اسم نویسنده داستان یا رمان، به‌صورت دستی تمامی متون موجود در بخش کتاب‌های داستانی این پیکره به دو گروه زن و مرد تفکیک شد. بخش دیگری از داده‌ها نیز از رمان‌ها، داستان‌های بلند و کوتاهی که به‌صورت پی‌دی‌اف در اینترنت موجود بودند تهیه گشت. برای حذف تأثیر عامل جنسیت مترجم بر روی متن اصلی اثر، تمامی داستان‌ها و رمان‌های غیر فارسی از مجموعهٔ مورد نظر حذف شد. پس از تهیهٔ پیکره به‌منظور حذف عامل تعداد متن بر روی استخراج ویژگی‌ها، در هر دو دسته زن و مرد تعداد مساوی ۲۵۴ داستان قرار داده شد.



(شکل- ۱): فرآیند شناسایی جنسیت نویسنده

۴-۲ - پیکره متون گفتاری

برای تهیه این پیکره، از رویکرد وب برای پیکره [1] استفاده شد. مهم‌ترین چالش، شناسایی سایتی فارسی بود که نخست محتوای متنی آن توسط خود کاربر نوشته و دوم نام نویسندهٔ هر متن نیز مشخص شده باشد. در ابتدا تصمیم بر استفاده از توییت[2] های سایت تویتر[3] یا استاتوس[4]های سایت فیس‌بوک[5] بود، ولی از آن‌جا که کاربران ایرانی به‌طورمعمول جملات دیگران را توییت کرده یا در دیوار[6]خود می‌نویسند، تصمیم گرفته شد که از نظرات کاربران در سایت هلو کیش استفاده شود (سایت هلوکیش، ۲۰۱۲). برای خزش[7] در این سایت از برنامه بیوتیفل سوپ[8] که به‌صورت کتابخانه‌ای در محیط

[1] Web for corpus
[2] Tweet
[3] Www.tweeter.com

[4] Statues= وضعیت
[5] Www.facebook.com
[6] Wall
[7] Crawling
[8] Beautiful Soup

برنامه‌نویسی پایتون وارد[1] می‌شود و قابلیت‌های بالایی در پارس صفحات HTML/XML دارد استفاده شد. برای استخراج نظرات مرتبط با نظردهندگان زن و مرد، ابتدا فهرستی تفکیک‌شده از اسامی فارسی زن و مرد تهیه شد؛ سپس نام نویسنده نظر با این فهرست اسامی تطبیق داده شده و نظرات بر حسب این فهرست تفکیک‌شده، در دو دسته زن و مرد قرار گرفت.

(جدول- ۱): پیکره داستان و رمان فارسی

	تعداد سند	میانگین تعداد کلمه در هر سند
نویسنده مرد	۲۵۴	۱۷۷۸۰
نویسنده زن	۲۵۴	۲۸۶۹۰
کل	۵۰۸	۲۳۲۳۵

(جدول- ۲): پیکره گفتاری نظرات

	تعداد سند	میانگین تعداد کلمه در هر سند
نویسنده مرد	۱۱۸۹	۸۵
نویسنده زن	۱۱۸۹	۸۸
کل	۲۳۷۸	۸۶/۵

از آن‌جا که بسیاری از متون فارسی با صفحه کلید عربی و یا غیراستاندارد نوشته شده و می‌شوند، احتمال وجود نویسه‌های غیر فارسی (برای مثال عربی یا اردو) در متون وجود دارد که درصورت عدم هنجارسازی این نویسه‌ها، نتایج پردازش‌های انجام‌شده بر روی متون قابل اطمینان نخواهند بود. قبل از انجام پژوهش، تمامی نویسه‌های غیرفارسی به معادل فارسی تبدیل شدند. جداول (۱و ۲) مشخصات این دو پیکره را نشان می‌دهد.

۵- انتخاب مجموعه ویژگی

با توجه به ادبیات پژوهش و بر پایهٔ پژوهش‌های روان‌شناختی، پنج مجموعه ویژگی مرتبط با جنسیت شامل: ۱) نویسه‌بنیاد ۲) واژه‌بنیاد ۳) نحوی ۴) ساختاربنیاد و ۵) کلمات دستوری برای هر سند محاسبه شده است. جدول شماره ۳ فهرست ۴۸ ویژگی استخراج‌شده را نشان می‌دهد.

ویژگی‌های مبتنی بر نویسه شامل شش ویژگی سبک‌شناسی هستند که به‌طور گسترده در مسائل تشخیص نویسنده به کار می‌روند و شامل ویژگی‌هایی مانند تعداد نویسه خالی، تعداد نویسه‌های ویژه (# و ٪) و غیره می‌شود (مریم، ۱۹۹۶). ویژگی‌های مبتنی بر واژه شامل یازده معیار آماری مانند غنای واژگانی، معیار یول (K) و معیار سیمپسون (D) هستند. پژوهش‌های چهار دهه گذشته نشان می‌دهد که شرایط جسمی و روانی افراد در واژه‌گزینی آن‌ها تأثیر دارد(گوتسچالک، ۱۹۹۶).تحلیل‌های انجام گرفته بر روی متون نشان داده است که تعداد کلمات مثبت (مانند عشق، خوب) در نوشته‌های افراد خوش‌قلم بیش‌تر به چشم می‌خورد؛ این افراد از کلمات دارای بار معنایی منفی (مانند زشت، کریه) در حد متعارف استفاده می‌کنند و کلمات شناختی (مانند دانستن، علت) را فراوان مورد استفاده قرار می‌دهند (پنباکر و همکاران، ۲۰۰۷). ویژگی‌های نحوی، سبک نگارشی نویسنده را در سطح جمله نشان می‌دهند. این ویژگی‌ها شامل تعداد علایم نگارشی (مانند ویرگول، نقطه، علامت تعجب و غیره) است. به‌علت تفاوت زنان و مردان در استفاده از علایم نگارشی، از این ویژگی‌ها می‌توان برای تشخیص جنسیت نویسنده استفاده کرد. به‌عنوان مثال طبق پژوهش‌ها مولاک، بسامد علامت سؤال در متون زنان نسبت به مردان بیش‌تر است (مولاک و همکاران، ۱۹۹۰).

به‌کمک ویژگی مبتنی بر ساختار، نویسنده چیدمان نوشته خود را سازماندهی می‌کند. افراد عادت‌های متفاوتی در سازماندهی متن خود دارند. ویژگی‌های مانند میانگین طول جمله، تعداد خطوط، تعداد خطوط خالی و غیره جزء ویژگی‌های ساختاری یک متن شمرده می‌شوند.

لغات دستوری، لغاتی هستند که معنای واژگانی کمی دارند، ولی کاربرد دستوری زیادی داشته و برای نشان‌دادن روابط دستوری مابین کلمات استفاده می‌شود. از این ویژگی‌ها می‌توان به ضمایر، اصوات و حروف اضافه اشاره کرد.

نشانه‌های روانی‌زبانی، مجموعه کلماتی هستند که بیان‌کنندهٔ ویژگی‌های شخصیتی و روانی شخص (مانند بسامد استفاده از صفات مثبت یا منفی، رنگ‌ها، کلمات رکیک) است.

۵-۱- استخراج خودکار و بازنمایی ویژگی‌ها

برای استخراج ویژگی‌های روانی‌زبانی جدول (۳) مانند تعداد صفات و تعداد قیدها و ضمایر، از برچسب‌زن اجزای کلام با مدل مارکوف مخفی مرتبه دو، که از روی پیکره برچسب‌خوردهٔ بی‌جن‌خان (بی‌جن‌خان، ۱۳۸۶) آموزش دیده شده بود استفاده شد. فهرست کلمات مربوط به ضمایر

[1] Import

فاعلی، پرسشی، حرف ربط، حرف ربط گروهی، صوت و حروف اضافه از واژگان زایای فارسی به دست آمد (اسلامی و همکاران، ۱۳۸۳). برای شمارش صفات مثبت و منفی در متون، به صفات موجود در واژگان زایای زبان فارسی برچسب مثبت (۱۸۴۰ صفت) و منفی (۱۵۳۶) داده شد و سپس این صفات در متن شمارش شد. برای استخراج رنگ‌ها از ویژگی سلسله‌مراتبی[1] صفحات ویکی‌پدیا فارسی استفاده شد (برای مثال در صفحه "رده رنگ‌ها" می‌توان نام رنگ‌های مختلف زبان فارسی را مشاهده کرد). برای استخراج کلمات مربوط به شک و تردید و قطعیت، فهرست واژگان مربوط به این دو حوزه در زبان انگلیسی از سایت‌های مختلف گردآوری و معادل فارسی آن‌ها نوشته و در متون شمارش شدند. فهرست کلمات رکیک با جستجو در اینترنت جمع‌آوری شد (به ضمایم مراجعه شود).

استخراج‌کنندهٔ ویژگی برای هر مستند متنی، یک بردار ۴۸ بعدی جهت بازنمایی مقادیر تولید می‌کند. از آن‌جا که اعداد مربوط به ویژگی‌ها از روش‌های مختلفی به دست آمده بود، این اعداد ممکن بود از صفر تا بیست‌هزار تغییر کنند. برای مثال مقدار مربوط به *تعداد کلمه* برای یک مستند ممکن است ۱۴۵۴۰ باشد؛ درحالی‌که *میانگین تعداد نویسه در کلمه* برای همان مستند ۳.۴ باشد. از این‌رو، برای اطمینان از این‌که در فرایند دسته‌بندی با تمامی مقادیر به‌صورت یکسان برخورد خواهد شد، تمامی مقادیر را به‌کمک روش هنجارسازی کمینه- بیشینه[2] نرمال کردیم تا در حد واسط صفر تا یک قرار گیرند:

$$Normalized\text{-}x_{ij}=\frac{x_{ij}-\min(x_j)}{\max(x_j)-\min(x_j)} \quad (۲)$$

که در این فرمول: x_{ij} ویژگی j ام در نمونه i ام و $min(x_j)$ و $max\ (x_j)$ کمترین و بیشترین مقادیر ویژگی j ام هستند.

۶- روش دسته‌بندی

برای دسته‌بندی از سه دسته‌بند بیز ساده[3]، درخت تصمیم ADtree و بردار ماشین پشتیبان[4] که به‌صورت گسترده در مسائل دسته‌بندی متون کاربرد دارند، استفاده شد. برای دسته‌بندی از بسته یادگیری ماشین وکا استفاده شد (مارک هال و همکاران، ۲۰۰۹).

[1] Hierarchical
[2] max-min normalization method
[3] Naïve Bayes
[4] Support vector machine

۶-۱- بیز ساده به‌عنوان الگوریتم دسته‌بندی

طبقه‌بندهای بیز ساده، سرعت محاسباتی بالایی داشته و با داده‌های آموزشی کم نیز دقت خوبی نشان می‌دهند. برای فرموله کردن تشخیص جنسیت نویسنده با استفاده از مدل استاندارد بردارسند به‌صورت زیر عمل می‌کنیم: هر متن را به‌صورت $D=(F_1,F_2,...,F_n)$ در نظر می‌گیریم، که در آن F_i ویژگی مورد نظر (مانند طول کلمه، تعداد صفت و غیره) است با داشتن این بردار، هدف ما تشخیص جنسیت نویسنده آن یعنی G است:

$$G \in \{male, female\} \quad (۳)$$

جنسیت G از طریق بیشینه‌کردن احتمال P(G|D) به‌دست می‌آید. طبق قانون بیز داریم :

$$P(G|D)=\frac{P(G)P(D|G)}{P(D)} \quad (۴)$$

$$P(D|G)=P(F1,F2,...,Fn|G) \quad (۵)$$

با توجه به فرض بیز ساده مبنی بر استقلال متغیرها داریم:

$$P(F_1,F_2,...,F_n|G)= P(F_1|G)P(F_2|G)...P(F_n|G), \quad (۶)$$

و فرمول به شکل زیر تغییر می‌کند:

$$gender=\underbrace{\operatorname{argmax}}_{G} P(G)\prod_{i=1}^{n} p(F_i|G) \quad (۷)$$

از آن‌جا که دادهٔ آموزشی ما برای دو جنس حجم یکسانی داشت، احتمال پیشین P(G) برای هر دو یکسان خواهد بود.

۶-۲- الگوریتم ADtree (درخت تصمیم تناوبی)

در الگوریتم درخت تصمیم‌گیری، هر گره درخت، نمایان‌گر یک ویژگی خاص است. در هر نقطه از پیمایش این درخت، تصمیم‌گیری براساس مقدار این ویژگی خاص در نمونه دسته بندی‌نشده صورت می‌گیرد. درکل، درخت تصمیم‌گیری روشی متداول است که حوزهٔ کاربرد گسترده‌ای دارد (صفویان و لاندگرب، ۱۹۹۱). درخت تصمیم ADtree، نوع وزن‌دار و تعمیم‌یافته‌ای از درخت تصمیم‌گیری است که از دو نوع گره تشکیل شده : گرهٔ تصمیم‌گیری و گرهٔ پیش‌بینی. ترکیب امتیاز گره‌های پیش‌بینی پیمایش شده و گره‌های

انتهایی، تعیین کنندهٔ دسته هر نمونه است (ماسون و همکاران، ۱۹۹۱).

(جدول- ۳): مجموعه ویژگی‌ها و توضیحات

ویژگی	توضیح ویژگی
ویژگی‌های مبتنی بر نویسه	
F_1	تعداد کل نویسه‌ها (C)
F_2	تعداد کل حروف الفبا(الف- ب)/C
F_3	تعداد کل اعداد/C
F_4	تعداد نویسه فاصله/C
F_5	تعداد نویسه تب/C
F_6	تعداد نویسه‌های ویژه/C
ویژگی‌های مبتنی بر واژه	
F_7	تعداد کل کلمات (N)
F_8	میانگین تعداد نویسه در کلمه
F_9	غنای واژگانی (کل لغات یکتا/ N)
F_{10}	کلمات طولانی (بزرگ‌تر از ۳ نویسه)/N
F_{11}	کلمات کوتاه (کوچک‌تر از ۲ نویسه) /N
F_{12}	هپکس لگومنا[1] (کلمات ۱ تکراره)/ N
F_{13}	هپکس دیسلگومنا[2] (کلمات ۲ تکراره)/N
F_{14}	معیار K یول[3]
F_{15}	معیار D سیمپسون[4]
F_{16}	معیار S سیشل[5]
F_{17}	معیار R هونور[6]
ویژگی‌های نحوی	
F_{18}	تعداد کاما (،)/C
F_{19}	تعداد نقطه (.)/C
F_{20}	تعداد دونقطه (:)/C
F_{21}	تعداد سمیکلون (؛)/C
F_{22}	تعداد علامت سوال (؟)/C
F_{23}	تعداد علامت تعجب (!)/C
F_{24}	تعداد علامت تعجب ۳تایی (!!!)/C
ویژگی‌های ساختاری[7]	
F_{25}	تعداد کل خط
F_{26}	تعداد کل جملات (S)
F_{27}	تعداد کلمه در هر جمله
F_{28}	تعدادخطوط خالی/ تعداد کل خطوط
F_{29}	میانگین طول خطوط غیر خالی

[1] Hapax legomena
[2] Hapax dislegomena
[3] Yule's K measure
[4] Simpson's D measure
[5] Sichel's S measure
[6] Honore's R measure
[7] Structural features

ادامه (جدول- ۳): مجموعه ویژگی‌ها و توضیحات

	ویژگی کلمات دستوری
F_{30}	تعداد ضمیر فاعلی / N
F_{31}	تعداد ضمیر پرسشی / N
F_{32}	حرف ربطN /
F_{33}	حرف ربط گروهی/ N
F_{34}	صوت/ N
F_{35}	حرف اضافه / N

ادامه (جدول- ۳): نشانه‌های روانی- زبانی

	ویژگی	کلمات مربوط به ویژگی
F_{36}	صفات مثبت	زیبا، معروف، سالم ...
F_{37}	صفات منفی	دشمن،آلوده،غیرقانونی...
F_{38}	رنگ‌ها	دوغی،عدسی، ماشی..
F_{39}	کلمات رکیک	دیوث، پفیوز، خرفت...
F_{40}	تعداد صفات	خوب،بد،معمولی..
F_{41}	تعداد قید	احمقانه،بیگمان، زیرکانه...
F_{42}	تعداد ضمایر	من،تو،او،ما،شما،ایشان
F_{43-46}	نسبت ضمایر	۳ ویژگی
F_{47}	شک و تردید	فرض، گمان، یحتمل
F_{48}	*قطعیت*	هیچگاه،هرگز، همیشه

۶-۳- ماشین بردار پشتیبان

این روش یکی از پرکاربردترین الگوریتم‌ها در حوزهٔ شناسایی نویسنده است(دیدریچ و همکاران، ۲۰۰۳). اساس این روش، یافتن بهترین ابرصفحه[8]دارای بیش‌ترین فاصله است، که بتواند داده‌های دو دسته را از هم تفکیک کند. این الگوریتم قادر است بدون بیش‌برازش‌شدن، تعداد زیادی ویژگی را در قالب یک مسئلهٔ دسته‌بندی خطی یا غیر خطی به کار گیرد. اگر ورودی این الگوریتم را بتوان به‌صورت خطی جدا کرد، SVM فاصلهٔ بین این دو طبقه را با جستجوی ابرصفحهٔ بهینه جدا کننده، بیشینه می‌کند. بردار وزنی بهینه ω^* را با حل مسئلهٔ بهینه‌سازی زیر می‌توان یافت:

$$\text{Minimize } j(\omega) = \frac{1}{2}||\omega||^2 \qquad (۸)$$

به‌شرطی که $y_i\ (\omega.x_i - b) \geq 1$

در مواردی که داده‌ها به‌صورت غیر خطی قابل جداسازی نیستند، SVM با استفاده از یک متغیر بی‌اثر ξ و اجازه قرار گرفتن نمونه‌های آموزشی در ناحیه‌ای مابین دو ابرصفحه‌ای که از نقطهٔ پشتیبانی دو طبقه می‌گذرد، اقدام به ساختن یک حاشیهٔ نرم[9]می‌کند (کورتکس و واپنیک،

[8] Hyperplane
[9] Soft margin

۱۹۹۵).در این حالت مسئلهٔ بهینه‌سازی به‌صورت زیر تغییر می‌کند .

$$\text{Minimize } j'(\omega) = \frac{1}{2}\left|\left|\omega^*\right|\right| + G\sum_{i=}^{N}\xi_i \qquad (۹)$$

به‌شرطی که $y_i(\omega^*.x_i - b) \le 1 - \xi, \xi \ge 0$

وقتی از SVM برای حل مسائل غیرخطی استفاده می‌شود، حقه هسته[1] کمک می‌کند تا فضای ویژگی x به یک فضای بعد بالای $\varphi(x)$ نگاشت شود و ابرصفحه دارای بیش‌ترین فاصله را در فضای جدید جستجو کند. در این مقاله برای دسته‌بندی از هستهٔ PUK[2] استفاده شده است که هم مقاوم است و هم قدرت نگاشت خوبی دارد (اوستون و همکاران، ۲۰۰۶).

۷- نتایج به‌دست‌آمده از دسته‌بندی

سه الگوریتم بیز ساده، درخت تصمیم ADtree و SVM را بر روی دو پیکره اجرا کردیم. برای ارزیابی مدل‌ها از روش ارزیابی اعتبارسنجی متقابل[3] ده‌تایی استفاده شد. دقت دسته بندی بر روی پیکرهٔ رمان و داستان به‌ترتیب ۶۶/۳٪، ۷۳/۶٪ و ۷۱/۶٪ بودند؛ درحالی‌که دقت همین دسته‌بندها بر روی پیکرهٔ متنی نظرات کاربران به‌ترتیب ۵۳٪، ۶۳٪ و ۵۸.۵٪ بود شکل (۲). به‌منظور مقایسه نتایج حاضر با نتایج تشخیص جنسیت در زبان انگلیسی، در جدول (۴) نتایج چند نمونه پژوهش در زبان انگلیسی آورده شده است.

۸- ویژگی‌های متمایزکننده

به‌منظور بررسی اهمیت هر گروه از ویژگی‌های پیشنهادشده، دسته‌بندی را پنج مرتبه با هر کدام از این ویژگی‌ها و به‌وسیله Adtree تکرار کردیم. همان‌طورکه نتایج نشان می‌دهد، تمامی ویژگی‌ها در میزان دقت طبقه‌بند دخیل بوده‌اند؛

ولی متمایزکننده‌ترین ویژگی‌ها به‌ترتیب ویژگی‌های مبتنی بر نویسه، مبتنی بر واژه و ساختاری وکم‌تأثیرترین ویژگی‌ها نیز ویژگی‌های دستوری و روانی‌زبانی بوده‌اند جدول (۵).

به‌منظور انتخاب ده ویژگی برتر و کاهش فضای ویژگی[4]، از الگوریتم بهره اطلاعات[5] استفاده شد (نجیب، ۱۹۹۹). این الگوریتم اهمیت هر ویژگی را با محاسبه بهره اطلاعاتی آن ویژگی و با توجه به طبقه مشخص می‌کند. با انجام دوبارهٔ عمل دسته‌بندی با SVM بر روی این ده ویژگی، مشاهده شد که دقت طبقه‌بند ۲ ٪ افت داشته است. جدول (۶) فهرست ده ویژگی نخست متمایزکننده را نشان می‌دهد.

(جدول- ۴): بعضی از نتایج به‌دست آمده برای زبان انگلیسی

مرجع	داده آموزشی	دسته بند	دقت (٪)
(جرمی وهمکاران، ۲۰۱۰)	وضعیت‌های[6] فیس بوک	بیز ساده	۶۷/۷
		پرسپترون	۵۹
(لای، ۲۰۱۰)	کتاب و وبلاگ	بیز ساده	۶۹
(نا چانگ و همکاران، ۲۰۱۱)	ایمیل‌های انرون	SVM	۸۲/۲۳
	گروه خبری رویتر		۷۶/۷۵

(جدول- ۵): دقت طبقه‌بندی با هر کدام از زیرویژگی‌ها

ویژگی	دقت (٪)
مبتنی بر واژه	۶۹/۱
روانی-زبانی	۵۵/۹
مبتنی بر نویسه	۶۹/۳
نحوی	۵۸/۹
ساختاری	۶۳/۸
دستوری	۵۴/۴
کل	۷۳/۸

۹- نتیجه‌گیری و بحث

این پژوهش نشان داد که به‌کمک ویژگی‌های سبک‌شناسی و روان‌شناختی جنسیت نویسنده را می‌توان تشخیص داد. نتیجهٔ طبقه‌بندی بر روی دو پیکره نشان داد که تمایز زبانی مربوط به جنسیت نویسندگان در پیکرهٔ مربوط به متون ادبی نسبت به پیکرهٔ مربوط به نظرات کاربران بیشتر است. دلیل این عدم تمایز هم ممکن است اقتضای محتوایی متون مربوط به نظرات باشد (از لحاظ حجم متن، عدم امکان بیان احساسات متفاوت، گونهٔ زبانی دو پیکره). از بین ویژگی‌های مختلف طراحی شده، ویژگی‌های واژه‌بنیاد و نشانه‌های زبانی‌روانی تمایزدهنده‌ترین عوامل در دسته‌بندی متون بودند. به‌عنوان پیشنهاد برای کارهای بعدی، جهت رسیدن به دقت‌های بالاتر می‌توان ویژگی کیسهٔ کلمات[7] و یا دیگر ویژگی‌های متمایزکننده جنسیت را نیز به مجموعه ویژگی‌ها افزود. همچنین برای رسیدن به نتایج بهتر، از داده‌های آموزشی مربوط به چت‌های فارسی استفاده کرد. برای

[1] Kernel trick
[2] Pearson VII Universal kernel
[3] Cross-validation
[4] Feature space reduction
[5] Information gain
[6] Status
[7] Bag of words

پیشرفت در حوزهٔ شناسایی نویسنده برای زبان فارسی یکی از اولویت‌ها تعریف معیارهای ارزیابی و مقایسه روش‌های مختلف با استفاده از پیکره‌های محک‌زنی[1] است. تهیه پیکره‌های مختلف زبانی با دامنه‌های (زمانی، موضوعی و غیره) مشخص در پیشرفت کار بسیار تأثیرگذار خواهد بود.

دستاوردهای این مقاله عبارت‌اند از:

- ارائهٔ نخستین دادگان دارای برچسب جنسیت برای زبان فارسی (دو پیکره نوشتاری و گفتاری).
- تهیه واژگان دسته‌بندی‌شدهٔ زبان فارسی (فهرست کلمات رکیک، رنگ‌ها، کلمات بیان‌کننده تردید و یقین، حروف ربط و غیره) و ارائه فضای ویژگی به‌کمک این واژگان.
- تهیه فهرست کاملی از اسامی مرد و زن فارسی
- درنهایت ارائه طبقه‌بندهایی جهت طبقه‌بندی متون فارسی از نظر جنسیت نویسنده

(جدول- ۶): ده ویژگی برتر محاسبه‌شده با بهره اطلاعاتی

۱	تعداد نویسه
۲	تعداد کلمات
۳	غنای واژگانی
۴	کلمات دوتکراره
۵	معیار یول
۶	کلمات کوچک‌تر از دو نویسه
۷	تعداد کل حروف الفبا
۸	کلمات یک تکراره
۹	ضمیر متصل اول شخص/مابقی ضمایر
۱۰	تعداد خطوط

ضمیمه الف:

معیار K یول:

$$\text{Yules K} = 10^4\left(-\frac{1}{N}+\sum_{i=1}^{V} V_i\left(\frac{i}{N}\right)^2\right)$$

معیار D سیمپسون:

$$\text{Simpsons D} = \sum_{i=1}^{V} V_i \frac{i}{N}\frac{i-1}{N-1}$$

معیار S سیشل:

$$\text{Sichels S} = \frac{\text{count of Hapax Dislegomena}}{V}$$

معیار R هونورس :

$$\text{Honores R} = \frac{100\log_{10}N}{1-\frac{\text{count of Hapax Legomena}}{V}}$$

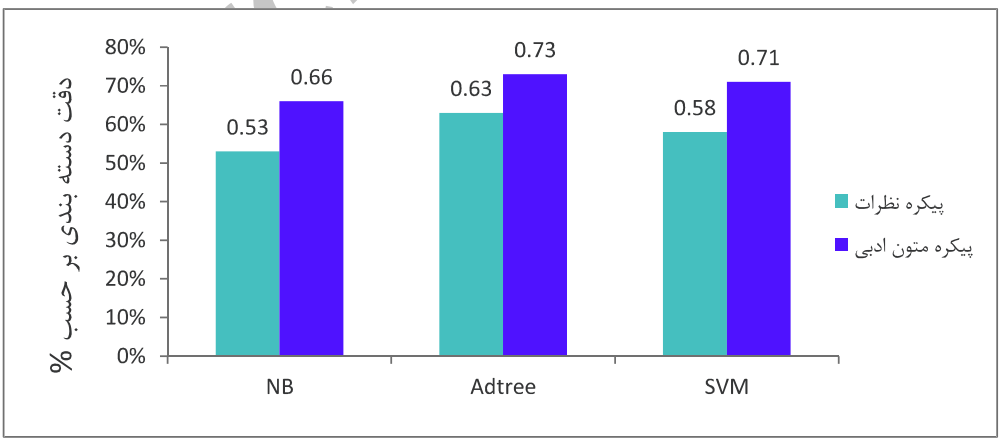
V : تعداد کلمات یکتا

V_i:تعداد کلمات یکتا که i مرتبه رخ داده است

N: تعداد کل کلمات

Hapax Dislegomena : کلماتی که فقط دوبار تکرار شده‌اند

Legomena Hapax: کلماتی که فقط یکبار تکرار شده‌اند



(شکل- ۲): الگوریتم دسته‌بندی

[1]Benchmark

ضمیمه ب (لیست کلمات):

قطعیت							
پیوسته	همیشه	همواره	مطمئنا	یقینا	مسلما	بدون شک	قطعا
صد در صد	هیچ	حاشا	هیچ وقت	هیچگاه	ابدا	هرگز	همه وقت

تردید							
گمان	حدس	یحتمل	ممکن	گویا	ممکن	احتمالا	شاید
به نظر من	محتمل	احتمالای	فرض	ظن			

حروف ربط									
که	هرچند	هم	هکذا	گرچه	چون	چه	آنچنانکه	اما	اگر
زیرا	یعنی	یا	ولو	ولیکن	ولی	و	تا	لیکن	لکن

اصوات							
براوو	بالله	به‌به	بادا	آخه	آمین	عجبا	احسنت
زکی	زهازه	یا	خوشا	وای	واسفا	وا	اوخ
جانمی	هورا	هیهات	حیف	هان	ای	اوا	ان‌شاالله
زرت	دالی	کریما	پیشت	مرده‌باد	مرسی	مرحبا	لبیک
کریما	دردا	سک سک	آفرین	کاش	دریغا	تبارک الله	آهای

پرسشی							
چی	چرا	چه	چندمین	چند	ایا	چطور	چگونه	آیا
کجایی	کجائی	کجا	کدامیک	کدامین	کدام	کی	کی	که

رنگ								
خاکستری	بنفش	برگ سنجدی	ارغوانی	آلبالویی	آبی	پوست پیازی	پسته‌ای	پرکلاغی
سرخابی	سبز	زیتونی	زرد	دوغی	دودی	دریایی	خرمایی	خاکی
فیروزه ای	عدسی	طلایی	صورتی	شبدیز	سیاه	سفید	سرمه‌ای	سرخ
قرمز	یاسی	نقره‌ای	نخودی	نارنجی	موشی	مغز	ماشی	فیلی

۱۰- مراجع

داوری اردکانی، نگار؛ عیار، عطیه. کنکاشی در پژوهش‌های زبان‌شناسی جنسیت ، مجله مطالعات راهبردی زنان، ۱۳۸۷، ۴۲-۱۶۲ .

درگاه ملی آمار . ۱۳۹۱، http: www.amar.org.ir .

زینب فرهمند پور وهمکاران. طراحی و پیاده‌سازی یک سامانه هوشمند تشخیص هویت نویسنده‌ی فارسی زبان ، ۷مین کنفراس ملی سالانه انجمن کامپیوتر ایران، اسفند ۹۰.

سعیده شجاع رضوی. جنسیت و شمول معنایی، فصل‌نامه پازند. ۱۳۸۶. ۱۰ : سال سوم.

محرم اسلامی، مسعود شریفی آتشگاه، صدیقه علیزاده لمجیری، و طاهره زندی. واژگان زایای زبان فارسی ,

مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه، تهران ،۱۳۸۳.

محمدرضا فارسیان. جنسیت در واژگان، دانشگاه تهران، ۱۳۷۸.

محمود بی‌جن‌خان. پیکره متنی زبان فارسی ، پژوهشکده پردازش هوشمند علایم, ۱۳۸۶.

Apte C, Damerau F, Weiss SM, Apte C, Damerau F, Weiss S Text mining with decision trees and decision rules, The conference on automated learning and discorery, workshop 6: learning from text and the web, 1998.

Baayen H, van Halteren H, Neijt A, Tweedie F. An experiment in authorship attribution, The 6th international conference on the statistical analysis of textual data .St. Malo, France, 2002.

BijanKhan M. The role of the corpus in writing a grammar: An introduction to a software , IranianJ-ournal of Linguistics, 2004.

Burrows J. F, Word-Patterns and Story-Shapes: the Statis-tical Analysis of Narrative Style , Literary and Lingu-istic Computing, 1987, p. 61-70.

Chakani Sarvar Language and Gender: A Contrastive View on Farsi and English DiscourseIslamic Azad University of Tabriz, 2000.

Cortes C, Vapnik V. Support-vector networks , Machine learning, 1995, p.۹۷-۲۷۳ .

De Vel, O., Anderson, A., Corney, M., and Mohay, G. Mining E-mail content for author identification fore-nsics , SIGMOD Record, .۳۰ : ٤ ۲۰۰۱, p. 55-64.

Diederich J, Kinder Mann J, Leopold E, Paass G. Authorship attribution with support vector machine , Applied Intelligence ;19, .۲۰۰۰ p.۲۳-۱۰۹ .

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. Authorship attribution with support vector machines , Applied Intelligence .۱۹ .۲۰۰۳ .p.۲۳–۱۰۹ .

Gottschalk LA Gleser GC. The measurement of psychological states through the content analysis of verbal behavior .Berkeley, University of California Press, 1969.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11, no. 1 (2009): 10-18. Hello kish .۲۰۱۲ .www.hellokish.org.

Holmes D. I ,A stylometric analysis oformon scripture and related texts , Royal Statistical Society, ۱۵۵, p.-91-120.

Internet World Usage Statistics, ۲۰۱۲ ,http: www.internetworldstats.com/stats.htm.

Jeremy K, Zach G, David K, Machine Learning and Feature Based Approaches to Gender Classification of Facebook Statuses .۲۰۱۰ .http: thekeesh.-com/cs224n/final_writ.

Koppel, M., Schler, J., and Argamon, S. Computational methods in authorship attribution , Journal of American Society Information Science Technology .٦٠ , ١ .۲۰۰۹ ..p9-26.

Lai Chao-Yue Author Gender Analysis , Applied Natural Language Processing.۲۰۱۰ .

Lakoff, R. Language and woman's place . New York : Harper and Row, 1975.

Mason Yoav Freund and Llew The Alternating Decision Tree Algorithm , the 16th International Conference on Machine Learning .۱۹۹۹ .p .124-133.

Mendenhall T. C. A mechanical solution of a literary problem , Popular Science Monthly .

Mendenhall T. C. The characteristic curves of composition , Science . ۹ , ۲۱٤ .۱۸۸۷ .p.237-248.

Merriam T. Marlowe's hand in Edward III revisited , Literary and Linguistic .۱۱ , ۱ .۱۹۹٦ . .p .

Miller Z. , Dickinson B, Wei H. Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features , International Journal of Intelligence Science .۲ .۲۰۱۲ .P.143-148.

Mosteller F. and Wallace D. L. Inference and Disputed Authorship: The Federalist, Addison-Wesley, 1964.

Mosteller F, Wallace DL. Applied Bayesian and Classical Inference: the case of the federalist papers , Springer series in statistic :Springer, 1984.

Mukherjee A. , Liu B. Improving Gender Classification of Blog Authors , Empirical Methods in Natural Language Processing, ۲۰۱۰p.۲۱۷–۲۰۷ .

Mulac A. , Wiemann, J. M., Widenmann, S. J& ,.Gibson, T. W. Male/female language differencesand effects in same-sex and mixed-sex dyads: The gender-linked language effect , Communication Monographs .۵۵ .۱۹۸۸ .p .315-335.

Mulac A. Studley LB, Blau S. The gender-linked language effect in primary and secondary students' impromptu essays , Sex Roles .۲۳ .۱۹۹۰ .p-۹ . .۱۰

Na Cheng, R. Chandramouli, K.P. Subbalakshmi Author gender identification from text , digitalin vestigation .۲۰۱۱ .p .77-88.

Nacip F.A using information gain as feature weight , The 8th Turkish Symposium on Artificial Intelligence and Nural Networks IAINN .(۹۹'Istanbul, Turkey.۱۹۹۹)

Pennebaker J. W., Mehl, M. R& ,.Niederhoffer, K. G. Psychological aspects of natural language use: Our words, our selves., Annual Review of Psychology , ۲۰۰۳

Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ. The development and psychometric properties of LIWC2007 .Austin, Texas : LIWC Inc, 2007.

Pennebaker, J. W& ,.King, L. A. Linguistic styles: Language use as an individualdifference .Personality and Social Psychology, 1999 .P. 6-14.

Safavian SR, Landgrebe D. A survey of decision tree classifier methodology , Cybernetics , ۳ .۱۹۹۱ . .۲۱p .660-674.

Stirman, S.W& ,.Pennebaker, J.W. Word use in the poetry of suicidal and non-suicidal poets. , 63, . Psychosomatic Medicine, 2001 - .۶۳ - .p .517-522.

Symantec Internet Security Threat Report, Trends for 2010 .Security Response Publications.۲۰۱۱ ,
Talbot MM Language and gender: an introduction., Wiley-Blackwell, 1998.

Tweedie FJ, Singh S, Holmes DI. Neural network applications in stylometry: the federalist papers , Computers and the Humanities.۳۰ , ۱ .۱۹۹۶ .

Unger R. K Toward a redefinition of sex and gender , American Psychologist .۳٤ .۱۹۷۹ .p .1085-1094.

Üstün, Bülent, Willem J. Melssen, and Lutgarde MC Buydens. "Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel." Chemometrics and Intelligent Laboratory Systems 81, no. 1 .2006. 29-40.

Wodak, Ruth & Benke, Gertrud "Gender as a Variation in Sociolinguistics " in The Handbook of Sociol-ingistics ,Cambridge : Blackwell, 2007.

Yule GU. The statistical study of literary vocabulary, Cambridge University Press, 1944.

Zheng R, Li J, Chen H, Huang Z . A framework for authorship identification of online messages: writing-style features and classification techniques .Journal of the American Society for Information Science and Technology, 2006.

مهدی مرادی مدرک کارشناسی ارشد زبان‌شناسی رایانشی خود را در سال ۱۳۹۲ از دانشگاه صنعتی شریف دریافت و مدرک کارشناسی خود را نیز از دانشگاه زنجان اخذ کرده است. از ایشان چند مقاله در نشریات داخلی و همایش‌های بین‌المللی و نیز یک فرهنگ تخصصی به چاپ رسیده است. وی هم‌اکنون مدرس دانشگاه آزاد واحد زنجان بوده و زمینه‌های پژوهشی مورد علاقه وی هستان‌شناسی، تشخیص نویسنده، متن‌کاوی و ساخت پیکره‌های زبانی است.
نشانی رایانامهٔ ایشان عبارت است از:
Mehdi.moradi.cl@gmail.com

محمد بحرانی درجه دکترای خود را در رشتهٔ هوش مصنوعی در سال ۱۳۸۹ از دانشگاه صنعتی شریف و درجه کارشناسی ارشد خود را از همان دانشگاه در سال ۱۳۸۲ دریافت کرده است. وی هم‌اکنون عضو هیئت علمی گروه زبان‌شناسی رایانشی در مرکز زبان‌ها و زبان‌شناسی دانشگاه صنعتی شریف است. زمینه‌های پژوهشی مورد علاقه ایشان عبارت است از: زبان‌شناسی رایانشی، پردازش زبان طبیعی، پردازش و بازشناسی گفتار و مدل‌سازی زبانی.
نشانی رایانامهٔ ایشان عبارت است از:
bahrani@sharif.edu