نشریه علمی‌ترویجی محاسبات نرم

سال ششم/ شمارهٔ دوم/ پاییز و زمستان ۱۳۹۶/ صفحهٔ ۶۰ـ۷۱

دریافت مقاله: ۱۳۹۲/۷/۸

پذیرش مقاله: ۱۳۹۴/۸/۸

بررسی مؤلفه‌های سبکی نویسندگان پیام‌های الکترونیکی با تکیه بر پژوهش‌های انجام‌شده

سمیرا زنگویی[۱]، حسنعلی نعمتی شمس‌آباد[۲]

[۱] کارشناس ارشد مهندسی فناوری اطلاعات، دانشگاه تهران، تهران، ایران
samirazangoei@yahoo.com
[۲] استاد دانشکده مدیریت فناوری اطلاعات دانشگاه تهران، تهران، ایران
nemati@ut.ac.ir

چکیده: شناسایی نویسنده از مسائل مهم در دسته‌بندی متون و پردازش زبان‌های طبیعی است که برای نشان دادن خصوصیات نویسنده از میان متون مختلف است. پیشرفت سریع ارتباطات اینترنتی، ابزارهای اینترنتی با هویت ناشناس مانند ایمیل و وبلاگ را به روش‌های ارتباطی محبوبی برای مرتکبین اعمال غیرقانونی تبدیل کرده و مسائل امنیتی خاصی را به وجود آورده است. در این مقاله به بررسی پژوهش‌های انجام‌شدهٔ شناسایی نویسنده در محیط‌های مجازی پرداخته شده است. پس از بررسی این پژوهش‌ها مهم‌ترین مؤلفه‌های تأثیرگذار در شناسایی نویسنده مشخص و دسته‌بندی شده‌اند. بررسی‌های این تحقیق ترکیب ۷ مؤلفه را در به دست آوردن بالاترین دقت در شناسایی نویسنده به دست آورده است.

واژه‌های کلیدی: شناسایی نویسنده، مؤلفه‌های شناسایی نویسنده، ویژگی‌های سبک نوشتاری.

۱. مقدمه

با گسترش سریع فناوری‌های اینترنت و برنامه‌های کاربردی، سوءاستفاده از پیام‌های اینترنتی برای اهداف نامناسب یا غیرقانونی، تبدیل به یک نگرانی عمده برای جامعه شده است. طبیعت ناشناس توزیع پیام‌های آنلاین، ردیابی هویت را به یک مشکل بحرانی تبدیل می‌کند. امروزه به‌کارگیری علوم مهندسی برای حل این‌گونه مسائل به‌سرعت و با موفقیت رو به افزایش است؛ بنابراین بسیاری از دانشمندان علوم کامپیوتر به این سمت گرایش پیدا کرده‌اند تا بتوانند از روش‌های خودکار هوشمند به‌منظور شناسایی نویسنده کمک بگیرند. هدف نهایی در ارائهٔ این روش‌ها، افزایش هرچه بیشتر دقت شناسایی نویسنده است. ایدهٔ شناسایی نویسنده از مبحث طبقه‌بندی متن[1] که خود سرفصلی از دانش فهم زبان‌های طبیعی است [۱] گرفته شده [۲] و در آن کوشش می‌شود تا با تجزیه و تحلیل واژگان، دستور زبان و مفهوم یک جمله و نیز با کمک گرفتن از دانش مربوط به واژگان، معنای آن جمله برای ماشین درک گردد.

این مقاله به جمع‌آوری و بررسی تحقیقات انجام‌شده در زمینهٔ شناسایی نویسنده در محیط‌های مجازی پرداخته است. زبان مورد بررسی در این تحقیق زبان انگلیسی است. در این تحقیق ابتدا مشخص می‌شود که بررسی پژوهش‌های انجام‌شده در چه زمینه‌ای از تحلیل نویسنده صورت می‌پذیرد سپس به بررسی مؤلفه‌های تأثیرگذار در شناسایی نویسنده پرداخته می‌شود. در انتها نتایج و مشاهدات به‌دست‌آمده از این تحقیق بیان شده است.

۲. پیشینهٔ تحلیل نویسنده

تحلیل نویسنده به فرایند آزمایش خصوصیات قطعه‌ای از نوشته برای نتیجه‌گیری در تعیین نویسندهٔ آن اطلاق می‌شود. تحلیل نویسنده برخاسته از تحقیقات زبان‌شناسی است که سبک‌شناسی هم گفته می‌شود. تکنیک‌های پیشرفته‌ای از جمله روش‌های یادگیری ماشین در این دامنه به کار گرفته شده‌اند. گری[2]، سالیس[3] و مکدونل[4] [۳] چهار جنبه از تحلیل نویسنده را که می‌توانند در قالب نرم‌افزار پیاده‌سازی شوند به کار گرفتند. بر اساس برخی از تعاریفِ گری و همکارانش می‌توان تحقیقات شناسایی نویسنده را در سه زمینه طبقه‌بندی کرد:

۱.۲. شناسایی نویسنده[5]

به تعیین احتمال تعلق قطعه‌ای از نوشتار به یک نویسندهٔ خاص، از طریق بررسی سایر نوشته‌ها توسط آن نویسنده می‌پردازد. همچنین در برخی از ادبیات، به‌ویژه توسط محققان زبان‌شناسی با نام خصوصیات نویسنده[6] هم کاربرد دارد. منشأ این زمینهٔ تحلیل نویسنده برخاسته از تحقیقات منطق‌دان انگلیسی آقای آگوست دمورگان[7] در قرن هجدهم است که پیشنهاد داد مشکلات نویسندگی را می‌توان با تعیین متونی که از کلمات بیشتری استفاده می‌کنند حل‌وفصل کرد. فرضیات او توسط مندنهال[8] [۴] کسی که به تألیف خصوصیات نویسندگان بیکن[9]، مارلو[10] و شکسپیر[11] پرداخت، مورد بررسی قرار گرفت. کامل‌ترین و قانع‌کننده‌ترین تحقیقی که در این زمینه انجام شد توسط موستلر و والیس[12] [۵] بود. نتایج آن‌ها به‌طور کلی توسط محققان تاریخی پذیرفته شد و نقطهٔ عطفی در این زمینهٔ تحقیقاتی شد.

۲.۲. توصیف نویسنده[13]

به جمع‌آوری خصوصیات یک نویسنده و تولید مشخصات او بر اساس نوشته‌های او می‌پردازد. برخی از این ویژگی‌ها عبارت‌اند از جنس، پس‌زمینهٔ آموزشی و فرهنگی و آشنایی با زبان. این تحقیقات نسبتاً جدید در جهتی خارج از تحقیقات شناسایی نویسنده رشد پیدا کردند. کریج[14] [۶] برای اولین بار ارتباط میان هویت و خصوصیات نویسندگی را با تجزیه و تحلیل نمایشنامه‌های نوشته‌شده توسط میدلتون توماس ارائه

1. Text Categorization
2. Gray
3. Sallis
4. MacDonell

5. Authorship identification
6. authorship attribution
7. Augustus de Morgan
8. Mendenhall
9. Bacon
10. Marlowe
11. Shakespeare
12. Mosteller and Wallace
13. Authorship characterization
14. Craig

کرد. او از کلمات رایج برجسته که به بهترین وجه می‌توانست تبعیض را نشان دهد، برای توصیف عادات نوشتاری استفاده کرد. کورنی و همکاران [۷] به کاوش اختلاف سبک نوشتن با پس زمینه‌های مختلف تحصیلاتی از نویسندگان پرداختند. کوپل و همکارانش [۸] شواهد قطعی ارائه دادند که نشان می‌دهد سبک نوشتن مردان متفاوت از زنان در استفاده از ضمایر و انواع خاصی از اصلاحات و اسامی می‌باشد. در سال ۲۰۰۹، تحقیقی توسط کورنی و کوپل [۹] صورت گرفت که در آن به بررسی میزان تأثیر جنسیت در تحلیل نویسنده پرداختند. اگرچه این معیار قابل قبولی در شناسایی نویسنده است، در تحقیقات بعدیِ نویسندگان زیاد مورد بررسی قرار نگرفت. در ادامهٔ تحقیقاتِ کورنی و کوپل در همان سال ۲۰۰۹، معیارهای دیگری از جمله سن و پیشینهٔ زبان توسط کورنی و آرگومان مورد بررسی قرار گرفت. این معیارها چنان بود که به سایر محققان در تصمیم‌گیری انتخاب بهترین معیارها کمک کرد.

۳.۲. تشخیص تشابه[1]

به مقایسهٔ قطعات متعدد نوشتار و تعیین اینکه آیا آن‌ها توسط یک نویسنده واحد تولید شده می‌پردازد. بیشتر مطالعات انجام‌شده که در این رده قرار دارند، مربوط به کشف سرقت ادبی است. دزدی ادبی شامل تکرار کامل و یا بخشی از کار بدون کسب اجازه از نویسندهٔ اصلی است. از آنجایی که تشخیص شباهت بسیار متفاوت از شناسایی نویسنده در جنبه‌های مختلف است، این موضوع فراتر از محدودهٔ این تحقیق است.

در جدول (۱) برای سه زمینهٔ تحلیل نویسنده (شناسایی نویسنده، توصیف نویسنده، تشخیص تشابه) برچسبی تعیین شده و بر اساس این سه زمینه در جدول (۲) به جمع‌آوری محققانی که در این سه زمینه کار کرده‌اند، پرداخته شده است.

جدول (۱): سه زمینهٔ تحلیل نویسنده و برچسب تعیین‌شده

مسئله	برچسب
شناسایی نویسنده	AI (Authorship identification)
توصیف نویسنده	AC (Authorship characterization)
تشخیص تشابه	SD (Similarity Detection)

1. Similarity Detection

جدول (۲): تحقیقات تحلیل نویسنده

Research	AI	AC	SD
Mosteller	×		
de Vel	×	×	
Thisted	×		
Yule	×		
Elliot	×		
Tomoji	×	×	
Binongo			×
Baayen	×	×	
Gray et al		×	
Bosch	×		
Foster	×	×	
Diederich	×		
Brainerd	×		
Farringdon	×		
McCallum		×	
Khmelev	×		

تحلیل نویسنده در سال‌های اخیر، روی پیام‌های الکترونیکی اعمال شده است. همان طور که در جدول (۲) مشاهده می‌شود بیشترین تحقیقات انجام‌شده در زمینهٔ شناسایی نویسنده است. در این تحقیق به بررسی تحقیقات انجام‌شده در شناسایی نویسنده پرداخته شده است. در ادامه ابتدا به بررسی تحقیقاتی که روی ویژگی‌های سبک‌شناسی کار کرده‌اند پرداخته می‌شود. سپس تکنیک‌ها و متغیرهای شناسایی نویسنده بررسی می‌شود. در انتها نتیجه‌گیری از این بخش صورت می‌گیرد.

۳. ویژگی‌های سبک‌شناسی برای شناسایی نویسنده

در تحقیقات اولیه، محققان کلمات مورد استفادهٔ نویسندگان مختلف را برای شناسایی نویسنده تحلیل می‌کردند. برای دستیابی به شناسایی نویسنده در کاربردهای مختلف، خصوصیات مستقل از متن مورد نیاز است. از جمله کارهای اولیه در این زمینه بررسی خصوصیاتی از جمله طول جمله [۱۰] و غنای کلمات است که یول[2] [۱۱] پیشنهاد کرد. بعدها باروز[3] [۱۲] مجموعه‌ای بیشتر از ۵۰ کلمه با فرکانس بالا را مورد بررسی قرار داد. هلمز[4] [۱۳] استفاده از کلمات کوتاه (کلمات

2. Yule
3. Burrows
4. Holmes

دو یا سه‌حرفی) و کلمات صدادار (کلماتی که با حرف صدادار شروع شوند) را تحلیل کرد. ویژگی‌های مبتنی بر کلمه و ویژگی‌های مبتنی بر حرف، تأثیر زیادی در بیشترین تحقیقات محققان در شناسایی نویسنده داشته است. در اولین کارهایی که از موستلر و والیس ارائه شد [۱۴]، بررسی کلمات تابع مشاهده می‌شد. در ادامه، محققان دیگری از جمله باین و همکارانش [۱۵]، باروز [۱۶]، هلمز و فوریس [۱۷] و توییدی [۱۸] با مشخصهٔ ممتاز کلمات تابع و قدرت آن در شناسایی نویسنده موافق بودند. از آنجایی که کاربرد کلمات تابع از روی قواعد نحو در جمله تعیین می‌شد، در این تحقیق کلمات تابع جزء ویژگی‌های نحوی در نظر گرفته می‌شود. برگرفته از تحقیقات زبان‌شناسی، تکیه‌کلام‌ها نیز از جمله ویژگی‌های نحوی مهمی هستند که در تحقیقات نویسندگان اعمال شده است. استاماتوس، فاکوتاکیس و کوکینس [۱۹] به معرفی متدهای اتوماتیک کاملی برای استخراج این نوع کلمات پرداختند و یافته‌های خود را با عنوان ویژگی‌های لغوی مطرح کردند. همچنین در تحقیقات بعدی‌شان مشاهده کردند که ویژگی‌های نحوی و لغوی تأثیر بسیار زیادی در شناسایی نویسنده دارد. همزمان با تحقیقات انواع ویژگی‌ها، ویژگی‌های ساختاری توجه زیادی را به خود جلب کرد. انسان‌ها خصوصیات متفاوتی در سازمان‌دهی گفتار دارند. این خصوصیات از جمله طول پاراگراف گواه محکمی از سبک نوشتاری شخصی است. این نکته در سندهای آنلاین برجسته‌تر به نظر می‌آید؛ زیرا اگرچه مضمون کمتری را می‌رساند، از نظر ساختاری قابل انعطاف‌تر است. دی ول و همکارانش [۲۰] از ویژگی‌های ساختاری و دیگر ویژگی‌ها برای شناسایی نویسندگان ایمیل استفاده کردند که به کارایی بالایی دست یافتند. در تحقیقات دیگری که توسط محققان صورت گرفت، ویژگی‌های خاص متن نیز مورد تأیید قرار گرفت. از کاربردهای موفق ویژگی‌های خاص متن می‌توان در تحقیق مارتیندل[1] و همکارش [۲۱] اشاره کرد. در بررسی‌هایی که انجام دادند، ویژگی‌های لغوی و ویژگی‌های خاص متن در نظر گرفته شده‌اند. در تحقیق دیگری که توسط ذینگ[2] و همکارانش [۲۲] ارائه شد، تعداد ۱۰ ویژگی خاص متن در نظر گرفته شده است و این ویژگی‌ها روی متن‌های به‌دست‌آمده از مجرمان اینترنتی، پیاده‌سازی شده‌اند. نتایج به‌دست‌آمده از تحقیقشان موفقیت‌آمیز بوده و باعث افزایش دقت شناسایی نویسنده شده است. در تحقیق دیگری که توسط عباسی و چن [۲۳] برای شناسایی نویسنده صورت گرفت، خصوصیات سبک نوشتاری را به پنج دستهٔ مختلف ویژگی‌های سبک نوشتاری، ویژگی‌های لغوی، ویژگی‌های ساختاری، ویژگی‌های خاص متن و ویژگی‌های طرز فکر خصوصی نویسنده تقسیم‌بندی کردند. این تقسیم‌بندی ویژگی‌ها بعدها مورد استفادهٔ پژوهشگران متعددی قرار گرفت. همچنین تحقیق دیگری توسط اقبال و همکارانش [۲۴] در سال ۲۰۰۸ در تقسیم‌بندی خصوصیات نویسنده صورت گرفته است.

اگرچه تاکنون یک مجموعهٔ کامل از ویژگی‌های شناسایی نویسنده که مورد قبول همگان باشد و دامنهٔ گسترده‌ای داشته باشد ارائه نشده است، کارایی شناسایی نویسنده بستگی به مجموعهٔ ویژگی‌های انتخاب‌شده و تکنیک‌های تحلیل دارد.

تکنیک‌های شناسایی نویسنده

در مطالعات اولیه، بیشترین ابزارهای تحلیلی در تحلیل نویسنده روش‌های آماری تک‌متغیری بودند. پیشگام این‌گونه مطالعات مندنهال [۲۱] بود که بر پایهٔ نمودار هیستوگرام از توزیع طول کلمه برای نویسندگان مختلف بهره برده است. از دیگر ابزارهای محبوب طبقه‌بندی توزیع ثابت کلمهٔ طبقه‌بندی بیز محققان موستلر و والیس [۱۴] بود که در طول تحقیقات طولانی مدتشان به دست آمد. فرایند استاتیک Cusum ابزار دیگری در تحلیل نویسنده بود که توسط فرینگدون اعمال شد. اساس این روش بدین گونه بود که انحراف‌معیار متغیرهای جمع‌آوری‌شده را به دست آورده و سپس گراف مقایسه را ترسیم می‌کنند. به‌رغم موفقیتی که این روش کسب کرد، هلمز بیان کرد که روش تحلیل Cusum روشی غیرقابل اعتماد است؛ زیرا ثبات نتایج برای عناوین مختلف را تضمین نمی‌کند. مشکل دیگر موجود در

1. Martindale

2. Zheng

روش‌های تک‌متغیری، این بود که فقط برای یک یا دو نوع ویژگی مناسب است. به همین دلیل روش‌های چندمتغیره به وجود آمدند. باروز [۱۲] برای اولین بار روش تحلیل اجزای اصلی را روی کلمات تابع به کار برد. بدین صورت که بعد از ترکیب متغیرها به شکل گراف پیاده‌سازی می‌شوند. فاصلهٔ بین گراف‌ها نشان‌دهندهٔ شباهت بین سبک نویسندگان است. نتایج رضایتمندی که بر این روش به دست آمد، باعث شد که در ادامه، تحقیقات بر اساس روش‌های چندمتغیره صورت بگیرد. بیبر [۲۵] آنالیز عاملی را مدل‌سازی کرد. نتایج او نشان داد که آنالیز عاملی می‌تواند اهمیت متغیرها و ویژگی‌ها را به دست آورد. بعدها تحلیل خوشه‌ای و تحلیل تفکیکی توسط لدگر و مریام [۲۶] و هلمز [۲۷] معرفی شد. نتایج به‌دست‌آمده از کارهای این محققان نشان‌دهندهٔ اثبات اعتبار روش‌های چندمتغیره شد و کارایی این روش‌ها را افزایش داد.

با ظهور کامپیوترهای قدرتمند، استفاده از روش‌های یادگیری ماشین در شناسایی نویسنده تحریک شد. توییدی [۲۸] از روش‌های شبکه‌های عصبی که پرسپترون چندلایه‌ای هم نامیده می‌شود، برای شناسایی نویسنده استفاده کرد. آن‌ها از سه لایهٔ پنهان و دو لایهٔ خروجی استفاده کردند. شبکه‌های با هستهٔ مبتنی بر تابع نیز توسط لو[1] و ماتیوز[2] [۲۹] در شناسایی نویسنده به کار گرفته شدند. دیدریچ و همکارانش روش‌های یادگیری ماشین را معرفی کردند. نتایجی که آن‌ها به دست آوردند حاصل بررسی نوشته‌های هفت نویسنده از مجموعهٔ ۲۶۵۲ روزنامهٔ نوشته‌شده توسط نویسندگان مختلف بود. این روش حدود ۶۰ تا ۸۰درصد درست تشخیص می‌داد. تحقیق جدید دیگری در شناسایی ایمیل نویسندگان بر اساس متن ایمیل صورت گرفت. دی ول و همکارانش [۲۰] از روش یادگیری ماشین برای دسته‌بندی ۱۵۰ ایمیل از سه نویسنده استفاده کردند. در این تحقیق آن‌ها به دقت ۸۰ درصد دست یافتند. آرگومان [۳۰] و همکارانش بعدها به بررسی ۵۰۰ ایمیل نوشته‌شده توسط ۲۰ نویسنده اقدام کردند. در تحقیقی که در سال ۲۰۱۱ توسط لیو [۳۱] صورت گرفت، شناسایی نویسنده با کمک روش شبکه‌های عصبی مورد بررسی قرار گرفت که در این تحقیق دقت به‌دست‌آمده ۸۰/۴۹درصد گزارش شده است.

به‌طور کلی، روش‌های یادگیری ماشین نسبت به روش‌های استاتیک، به دقت بیشتری دست می‌یابند. روش‌های یادگیری ماشین می‌توانند روی مجموعهٔ بزرگی از ویژگی‌ها با کمک مدل‌های ریاضی پیاده‌سازی شوند.

علاوه بر تکنیک‌های شناسایی نویسنده، متغیرهایی از جمله تعداد نویسندگان و تعداد پیام‌ها برای شناسایی و آموزش مدل‌های طبقه‌بندی نویسنده در شناسایی نویسنده تأثیرگذار است.

۴. متغیرهای شناسایی نویسنده

مسئلهٔ شناسایی نویسنده جزء مسائل دسته‌بندی است. سطح پیچیدگی این مسئله توسط تعدادی متغیر تعیین می‌شود. برای مثال، تعداد نویسندگان و تعداد مستندهای قابل دسترس برای مجموعهٔ آموزشی در دقت پیش‌بینی تأثیرگذار است. هورن[3]، فرانک[4] و هام[5] [۳۲] برای شناسایی شاعر با کمک شبکه‌های عصبی به دقت ۸۰ تا ۹۰درصد بین دو شاعر دست یافتند درحالی‌که وقتی انتخاب بین سه شاعر بود، دقت به ۷۰درصد کاهش یافت. آن‌ها این پیشنهاد را دادند که اعتبار طبقه‌بندی با افزایش تعداد شاعران کاهش می‌یابد. استاماتوس و همکارانش [۱۹] به آزمایش سایز مجموعه داده‌های آموزشی در کارایی شناسایی نویسنده پرداختند. آن‌ها به این نتیجه رسیدند که کارایی طبقه‌بندی با افزایش نوشته‌های نویسندگان در مجموعه دادهٔ آموزشی افزایش می‌یابد. به‌رغم اینکه این افزایش سایز مجموعه دادهٔ آموزشی باعث افزایش کارایی می‌شود، بیشتر تحقیقات انجام‌گرفته در شناسایی نویسنده، روی تعداد کمی از نویسندگان در مجموعه داده‌های آموزشی صورت گرفته است.

جدول (۳) خلاصه‌ای از تحقیقات انجام‌شده (قسمت‌های ۳، ۴ و ۵) را با بیان نتایج، نقاط قوت و ضعف آن‌ها در تحلیل نویسنده بیان کرده است:

1. Lowe
2. Matthews

3. Hoorn
4. Frank
5. Ham

جدول (۳): خلاصه‌ای از کارهای گذشته

نام محقق	سال تحقیق	موضوع تحقیق و مسئلهٔ آن	نتایج	نقاط قوت	نقاط ضعف
گری، سالیس و مکدونل	۱۹۹۷	پیاده‌سازی چهار جنبه از تحلیل نویسنده در قالب نرم‌افزار	ایجاد زمینهٔ مناسب برای تحقیقات شناسایی نویسنده	ایجاد زمینهٔ مناسب برای تحقیقات شناسایی نویسنده	جزء تحقیقات اولیهٔ شناسایی نویسنده
مندنهال	۱۸۸۷	تألیف خصوصیات نویسندگان بیکن، مارلو و شکسپیر	به دست آوردن موفقیت در شناسایی نویسنده	—	—
کریج	۱۹۹۹	تجزیه و تحلیل نمایشنامه‌های نوشته‌شده توسط میدلتون توماس و ارتباط میان هویت و خصوصیات نویسندگی	نشان دادن تبعیض به بهترین وجه	استفاده از کلمات رایج برجسته	نیاز به تکمیل داده‌ها
کورنی و کوپل	۲۰۰۹	بررسی میزان تأثیر جنسیت در تحلیل نویسنده	کسب نتیجهٔ موفقیت‌آمیز	در نظر گرفتن معیار مناسب	توسط نویسندگان بعدی مورد استفاده قرار نگرفت
استاماتوس، فاکوتاکیس و کوکینس	۲۰۰۱	معرفی متدهای اتوماتیک کاملی برای استخراج نوع کلمات	نشان دادن تأثیر زیاد ویژگی‌های لغوی و نحوی	استفاده از ویژگی‌های لغوی و نحوی	نیاز به در نظر گرفتن ویژگی‌های دیگر
دی ول	۲۰۰۱	استفاده از از ویژگی‌های ساختاری و دیگر ویژگی‌ها برای شناسایی نویسندگان ایمیل	دستیابی به کارایی بالای این ویژگی‌ها	استفاده از ویژگی‌های ساختاری	نیاز به در نظر گرفتن ویژگی‌های دیگر
مارتیندل	۱۹۹۵	بررسی ویژگی‌های لغوی و ویژگی‌های خاص متن	تأیید تأثیر ویژگی‌های لغوی و ویژگی‌های خاص متن	استفاده از ویژگی‌های لغوی و ویژگی‌های خاص متن	نیاز به در نظر گرفتن ویژگی‌های دیگر
ذینگ	۲۰۰۳	بررسی ویژگی‌های خاص متن بر روی متون مجرمان اینترنتی	افزایش دقت شناسایی نویسنده	افزایش دقت	نیاز به تکمیل ویژگی‌ها
موستلر و والیس	۱۹۶۴	استفاده از ابزارهای محبوب طبقه‌بندی توزیع ثابت کلمه طبقه‌بندی بیز	ثبت روش بیز	ایجاد زمینهٔ مناسب برای تحقیقات بعدی	غیرقابل اعتماد بودن روش
هلمز	۱۹۹۲	بررسی تحلیل خوشه‌ای و تحلیل تفکیکی	اثبات اعتبار روش‌های چندمتغیره	افزایش کارایی روش ارائه‌شده	پایین بودن دقت
دی ول	۲۰۰۱	استفاده از روش‌های یادگیری ماشین برای دسته‌بندی ایمیل	دست یافتن به دقت ۸۰	استفاده از روشی نوین	تعداد کم نویسندگان
هورن،فرانک وهام	۱۹۹۹	شناسایی شاعر با کمک شبکه‌های عصبی	کسب دقت ۸۰ تا ۹۰	دقت تقریباً بالا	تعداد کم مجموعه داده آموزشی
استاماتوس	۲۰۰۱	آزمایش سایز مجموعه داده‌های آموزشی در کارایی شناسایی نویسنده	افزایش کارایی طبقه‌بندی با افزایش نوشته‌های نویسندگان	ایجاد زمینهٔ مناسب در تحقیقات بعدی	—
لیو	۲۰۱۱	شناسایی نویسنده با کمک شبکه‌های عصبی	دست یافتن به دقت مناسب	دقت تقریباً بالا	—

همان طور که در جدول (۳) مشاهده می‌شود، در سال‌های مختلف در زمینهٔ شناسایی نویسنده، تحقیقات متعددی صورت گرفته که به نتایج متعددی هم دست یافته است. هر تحقیق به نوبهٔ خود نقاط ضعف و قوتی داشته است. با بررسی نقاط ضعف و قوت این تحقیقات، این نتیجه به دست می‌آید که در نظر گرفتن تعداد کمی از ویژگی‌ها منجر به پایین آمدن دقت می‌شود. همچنین با تحلیل نتایج جدول (۳) این نتیجه به دست می‌آید که بیشتر محققان ویژگی‌های لغوی را در تحقیقات خود به کار برده‌اند، اما با توجه به پایین بودن دقت شناسایی، محققان بعدی در صدد آن شدند که ویژگی‌های شناسایی نویسنده را افزایش دهند. همچنین مشاهده می‌شود که به‌کارگیری روش‌های یادگیری ماشین منجر به بالا رفتن دقت شناسایی نویسنده می‌شود. بنابراین بر آن شدیم تا به بررسی تأثیرگذارترین مؤلفه‌ها در شناسایی نویسنده بپردازیم.

در جدول (۴) به جمع‌آوری زمینه‌های کاری محققان در شناسایی نویسنده پرداخته شده است. به عبارتی دیگر، در این جدول نشان داده شده که هریک از محققان چه ویژگی‌هایی (ویژگی‌های لغوی، ویژگی‌های نحوی، ویژگی‌های ساختاری و ویژگی‌های خاص متن) را به کار برده و از چه تکنیک‌هایی (تکنیک‌های استاتیک یا تکنیک‌های یادگیری ماشین) استفاده کرده و چه متغیرهایی (متغیر تعداد نویسندگان و متغیر سایز مجموعه داده آموزشی) را مد نظر قرار داده‌اند.

جدول (۱): زمینه‌های کاری نویسندگان

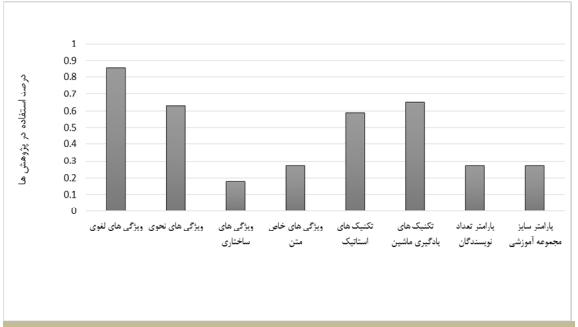
نام محقق	ویژگی‌های لغوی	ویژگی‌های نحوی	ویژگی‌های ساختاری	ویژگی‌های خاص متن	تکنیک‌های استاتیک	تکنیک‌های یادگیری ماشین	متغیر تعداد نویسندگان	متغیر سایز مجموعهٔ آموزشی
Mosteller & Wallace		■			■			
Ledger & Merriam	■				■			
Merriam & Matthews	■					■		
Kjell	■					■		
Martindale & McKenzie	■	■		■	■	■		
Mealand	■	■			■			
Holmes & Forsyth	■	■			■	■		
Farringdon	■	■			■			
Baayen et al	■	■			■			
Tweedie et al		■				■		
Craig	■				■			
Hoorn et al	■					■		
Binongo & Smith	■				■			
Diederich et al	■					■		
De Vel et al	■	■	■	■		■		
Stamatatos et al		■			■			
Khmelev & Tweedie	■					■	■	■
Corney et al	■	■	■	■		■	■	■
Baayen et al	■	■		■	■		■	■
Peng et al	■	■				■	■	■
Zheng et al	■	■	■	■	■		■	■
Argamon et al	■	■	■	■		■	■	■

با تحلیل و بررسی جدول (۴) به دسته‌بندی کاملی از مؤلفه‌های تأثیرگذار در شناسایی نویسنده پرداخته شده است. در جدول (۵) یک دسته‌بندی از این مؤلفه‌ها ارائه شده است.

جدول (۵): مؤلفه‌های تأثیرگذار در شناسایی نویسنده

مؤلفه‌های تأثیرگذار در شناسایی نویسنده	
ویژگی‌های سبک‌شناسی شناسایی نویسنده	✓ ویژگی‌های لغوی ✓ ویژگی‌های نحوی ✓ ویژگی‌های ساختاری ✓ ویژگی‌های خاص متن
تکنیک‌های شناسایی نویسنده	✓ تکنیک‌های استاتیک ✓ تکنیک‌های یادگیری ماشین
متغیرهای شناسایی نویسنده	✓ متغیر تعداد نویسندگان ✓ متغیر سایز مجموعهٔ آموزشی

همان طور که در جدول (۵) بیان شده، ۸ مؤلفه برای شناسایی نویسنده جمع‌آوری شده است. حال بررسی می‌کنیم که هریک از این عوامل در چه تعداد از پژوهش‌ها تاکنون مورد استفاده قرار گرفته و پژوهش‌هایی که از این مؤلفه‌ها استفاده کرده‌اند، به چه دقتی در شناسایی نویسنده رسیده‌اند.



نمودار (۱): درصد استفادهٔ هریک از مؤلفه‌ها در پژوهش‌های انجام‌شده تاکنون

نمودار (۱) نشان‌دهندهٔ درصد استفادهٔ هریک از مؤلفه‌ها در پژوهش‌های انجام‌شده تاکنون است.

حال به بررسی دقت به‌دست‌آمده در شناسایی نویسنده در استفاده از این مؤلفه‌ها پرداخته می‌شود. بنابراین دقت همهٔ پژوهش‌های انجام‌شده برای تک‌تک این مؤلفه‌ها جمع‌آوری کرده و میانگین دقت به دست آمده است.

همان طور که در نمودار ۲ مشاهده می‌شود، دقت شناسایی نویسنده برای هریک از مؤلفه‌های نام‌برده محاسبه شده است. با بررسی نمودار (۲) مشاهده می‌شود از بین ۴ ویژگی لغوی، نحوی، ساختاری و خاص متن بیشترین درصد متعلق به ویژگی‌های لغوی است. از آنجایی که ویژگی‌های لغوی نشان‌دهندهٔ سبک نوشتن واژگان مربوط به نویسنده است، در تحقیقات بیشتر نویسندگان نقش بسزایی داشته است و به‌کارگیری از آن در شناسایی نویسنده باعث افزایش دقت شناسایی نویسنده می‌شود. ویژگی‌های نحوی نیز بعد از ویژگی‌های لغوی بیشترین دقت را به خود اختصاص داده است. منظور از ویژگی‌های نحوی، نوع سبک نوشتاری نویسنده در سطح جمله است که تأثیر بسزایی در شناسایی نویسنده ایفا می‌کند. ویژگی‌های خاص متن اشاره به کلمات کلیدی مورد استفادهٔ نویسندگان است و ویژگی‌های ساختاری نیز نشان‌دهندهٔ عادات نویسنده در هنگام سازمان‌دهی یک قطعه از نوشتار است. این دو ویژگی به‌تنهایی نسبت به ویژگی‌های لغوی و نحوی، تأثیر کمتری در دقت شناسایی نویسنده دارند، اما با ترکیب این ویژگی‌ها با ویژگی‌های لغوی و نحوی دقت شناسایی نویسنده به طرز چشمگیری افزایش می‌یابد.

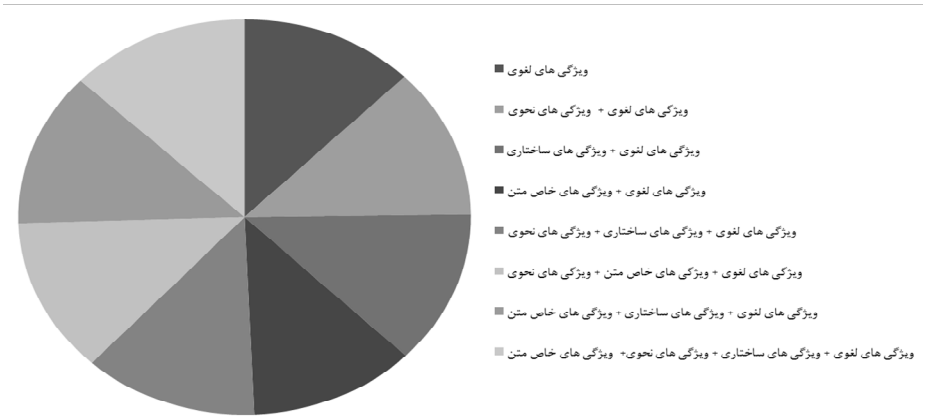
نمودار (۲): دقت به‌دست‌آمده شناسایی نویسنده در استفاده از ۸ مؤلفه

۵. بحث و نتیجه‌گیری

برای اینکه که بتوان مجموعه‌ای از تأثیرگذارترین مؤلفه‌ها در شناسایی نویسنده را ارائه کرد، ابتدا به بررسی دقت به‌دست‌آمده در ترکیب ویژگی‌ها با دو تکنیک یادگیری ماشین و تکنیک‌های ایستا پرداخته شد. به عبارتی دیگر، ابتدا تحقیقاتی که با کمک تکنیک یادگیری ماشین به شناسایی نویسنده پرداختند، در نظر گرفته و سپس دقت به‌دست‌آمده از ترکیب ویژگی‌ها استخراج شد. به همین ترتیب برای تکنیک ایستا هم دقت به‌دست‌آمده از ترکیب ویژگی‌ها استخراج شد. جدول (۶) این دقت‌ها را بیان کرده است.

جدول (۶): دقت به‌دست‌آمده از ترکیب ویژگی‌ها در دو تکنیک یادگیری ماشین و ایستا

	ویژگی	دقت
تکنیک یادگیری ماشین	ویژگی‌های لغوی	۹۱/۷۵
	ویژگی‌های لغوی + ویژگی‌های نحوی	۹۰/۳۵
	ویژگی‌های لغوی + ویژگی‌های ساختاری	۹۲/۶۳
	ویژگی‌های لغوی + ویژگی‌های خاص متن	۸۷/۶۵
	ویژگی‌های لغوی + ویژگی‌های ساختاری + ویژگی‌های نحوی	۹۱/۶۷
	ویژگی‌های لغوی + ویژگی‌های خاص متن + ویژگی‌های نحوی	۹۳/۴۵
	ویژگی‌های لغوی + ویژگی‌های ساختاری + ویژگی‌های خاص متن	۹۲/۵۹
	ویژگی‌های لغوی + ویژگی‌های ساختاری + ویژگی‌های نحوی + ویژگی‌های خاص متن	۹۴/۷۸
تکنیک ایستا	**ویژگی**	**دقت**
	ویژگی‌های لغوی	۸۶/۳۶
	ویژگی‌های لغوی + ویژگی‌های نحوی	۷۰/۱۳
	ویژگی‌های لغوی + ویژگی‌های ساختاری	۶۴/۶۳
	ویژگی‌های لغوی + ویژگی‌های خاص متن	۶۶/۶۵
	ویژگی‌های لغوی + ویژگی‌های ساختاری + ویژگی‌های نحوی	۷۱/۸۱
	ویژگی‌های لغوی + ویژگی‌های خاص متن + ویژگی‌های نحوی	۸۳/۵۵
	ویژگی‌های لغوی + ویژگی‌های ساختاری + ویژگی‌های خاص متن	۸۰/۵۶
	ویژگی‌های لغوی + ویژگی‌های ساختاری + ویژگی‌های نحوی + ویژگی‌های خاص متن	۸۶/۷۰



نمودار (۳): دقت به‌دست‌آمده از ترکیب ویژگی‌ها در تکنیک‌های یادگیری ماشین

نمودار (۳) دقت‌های به‌دست‌آمده از ترکیب ویژگی‌ها را در تکنیک‌های یادگیری ماشین و نمودار (۴) دقت‌های به‌دست‌آمده از ترکیب ویژگی‌ها را در تکنیک ایستا نشان می‌دهند. با توجه به دقت‌های به‌دست‌آمده در جدول (۶) نتیجه گرفته شد که مجموعهٔ کاملی از ویژگی‌ها دقت بیشتری را نتیجه خواهد داد. همچنین تکنیک‌های یادگیری ماشین در همهٔ موارد دقت‌های بیشتری را نتیجه داده است.

اینک تأثیر دو مؤلفهٔ «متغیر تعداد نویسندگان» و «متغیر سایز مجموعهٔ آموزشی» بررسی می‌شود که در جدول (۷) دقت‌های به‌دست‌آمده از تأثیر این دو مؤلفه قابل مشاهده است.

نمودار (۴): دقت به‌دست‌آمده از ترکیب ویژگی‌ها در تکنیک ایستا

جدول (۷): تأثیر دو مؤلفهٔ «متغیر تعداد نویسندگان» و «متغیر سایز مجموعهٔ آموزشی»

	ویژگی		دقت
تکنیک یادگیری ماشین	ویژگی‌های لغوی + ویژگی‌های ساختاری + ویژگی‌های نحوی + ویژگی‌های خاص متن	متغیر تعداد نویسندگان	۹۲/۱۸
		متغیر سایز مجموعهٔ آموزشی	۹۴/۳۲
		متغیر تعداد نویسندگان + متغیر مجموعهٔ آموزشی	۹۷/۱۵
	ویژگی		**دقت**
تکنیک ایستا	ویژگی‌های لغوی + ویژگی‌های ساختاری + ویژگی‌های نحوی + ویژگی‌های خاص متن	متغیر تعداد نویسندگان	۸۲/۹
		متغیر سایز مجموعهٔ آموزشی	۸۷/۴۳
		متغیر تعداد نویسندگان + متغیر سایز مجموعهٔ آموزشی	۸۸/۱۳

با توجه به جدول (۷) نتیجه می‌شود در نظر گرفتن ۷ مؤلفهٔ ویژگی‌های لغوی، ویژگی‌های ساختاری، ویژگی‌های نحوی، ویژگی‌های خاص متن، تکنیک یادگیری ماشین، متغیر تعداد نویسندگان و متغیر سایز مجموعهٔ آموزشی، دقت را به ۹۷/۱۵ می‌رساند.

بر اساس مطالب فوق می‌توان نتیجه گرفت که برای افزایش دقت در شناسایی نویسنده، نیاز به استخراج تأثیرگذارترین مؤلفه‌هاست که با تحلیل و بررسی تحقیقات انجام‌شده، ۷ مؤلفهٔ تأثیرگذار در شناسایی نویسنده استخراج شد که دقت شناسایی نویسنده را تا حد ۹۷/۱۵ درصد افزایش می‌دهد.

مراجع

[1] Russell S., Norvig P., and Canny J., *Artificial Intelligence: A Modern Approach*, ser. Prentice Hall series in artificial intelligence. Prentice Hall, 2003.

[2] Lewis D. D., Yang Y., Rose T. G., and Li F., *"RCV1: A new benchmark collection for text categorization research,"* J. Mach. Learn. Res., vol. 5, pp. 361–397, 2004.

[3] Gray A., Sallis P., and MackDonell S., *"Software forensics: Extending authorship analysis techniniques to computer programs,"* ser. Information Science Discussion Papers Series. University of Otago, 1997.

[4] Mendenhall T. C., *"The characteristic curves of composition,"* Science, vol. ns-9, no. 214S, pp. 237–246, 1887.

[5] Mosteller F. and Wallace D. L., *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*, ser. Springer Series in Statistics. Springer-

Verlag New York, 1984.

[6] Craig H., *"Authorial attribution and computational stylistics: if you can tell authors apart, have you learned anything about them?"* Literary and Linguistic Computing, vol. 14, no. 1, pp. 103–113, 1999.

[7] Corney M., de Vel O. Y., A. Anderson, and G. M. Mohay, *"Genderpreferential text mining of e-mail discourse,"* in 18th Annual Computer Security Applications Conference (ACSAC 2002), 9-13 December 2002, Las Vegas, NV, USA, 2002, pp. 282–289.

[8] Koppel M., Argamon S., and Shimoni A. R., *"Automatically categorizing written texts by author gender,"* LLC, vol. 17, no. 4, pp. 401–412, 2002.

[9] Koppel M., Schler J., and Argamon S., *"Computational methods in authorship attribution,"* JASIST, vol. 60, no. 1, pp. 9–26, 2009.

[10] Yule G. U., *"On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship,"* Biometrika, vol. 30, no. 3/4, pp. 363–390, 1939.

[11] Yule G. U., *The Statistical Study of Literary Vocabulary*. Cambridge,UK: Cambridge University Press, 1944.

[12] Burrows J. F., *"Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style,"* Literary and Linguistic Computing, vol. 2, no. 2, pp. 61–70, 1987.

[13] HOLMES D. I., *"The Evolution of Stylometry in Humanities Scholarship,"* Literary and Linguistic Computing, vol. 13, no. 3, pp. 111–117, 1998.

[14] Mosteller F. and Wallace D. L., *Inference and Disputed Authorship: The Federalist*, ser. The David Hume Series. Center for the Study of Language and Inf; New edition, 2008.

[15] Baayen H., van Halteren H., and Tweedie F., *"Outside the cave of shadows: using syntactic annotation to enhance authorship attribution,"* Literary and Linguistic Computing, vol. 11, no. 3, pp. 121–132, 1996.

[16] Burrows J. F., *"'an ocean where each kind. . .': Statistical analysis and some major determinants of literary style,"* Computers and the Humanities, vol. 23, no. 4, pp. 309–321, 1989.

[17] HOLMES D. I. and FORSYTH R. S., *"The Federalist Revisited: New Directions in Authorship Attribution,"* Literary and Linguistic Computing, vol. 10, no. 2, pp. 111–127, 1995.

[18] Tweedie F. J. and Baayen R. H., *"How variable may a constant be? measures of lexical richness in perspective,"* Computers and the Humanities, vol. 32, no. 5, pp. 323–352, 1998.

[19] Stamatatos E., Fakotakis N., and Kokkinakis G., *"Computer-based authorship attribution without lexical measures,"* Computers and the Humanities, vol. 35, no. 2, pp. 193–214, 2001.

[20] de Vel O., Anderson A., Corney M., and Mohay G., *"Mining e-mail content for author identification forensics,"* SIGMOD Rec., vol. 30, no. 4, pp. 55–64, 2001. [Online]. Available: http://doi.acm.org/10.1145/604264.604272

[21] Martindale C. and McKenzie D., *"On the utility of content analysis in author attribution: The federalist,"* Computers and the Humanities, vol. 29, no. 4, pp. 259–270, 1995.

[22] Zheng R., Qin Y., Huang Z., and Chen H., *"Authorship analysis in cybercrime investigation,"* in Intelligence and Security Informatics, First NSF/NIJ Symposium, ISI 2003, Tucson, AZ, USA, June 2-3, 2003, Proceedings, 2003, pp. 59–73.

[23] Abbasi A. and Chen H., *"Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace,"* ACM Trans. Inf. Syst., vol. 26, no. 2, pp. 7:1–7:29, 2008.

[24] Iqbal F., Hadjidj R., Fung B. C., and Debbabi M., *"A novel approach of mining write-prints for authorship attribution in e-mail forensics,"* Digital Investigation, vol. 5, pp. S42 – S51, 2008.

[25] Biber D., *Dimensions of Register Variation. A Cross-Linguistic Comparison*. Cambridge University Press, 1995.

[26] LEDGER G. and MERRIAM T., *"Shakespeare, Fletcher, and the Two Noble Kinsmen,"* Literary and Linguistic Computing, vol. 9, no. 3, pp. 235–248, 1994.

[27] Holmes D. I., *"A stylometric analysis of mormon scripture and related texts,"* Journal of the Royal Statistical Society. Series A (Statistics in Society), vol. 155, no. 1, pp. 91–120, 1992.

[28] Tweedie F. J., Singh S., and Holmes D. I., *"Neural network applications in stylometry: The federalist papers,"* Computers and the Humanities, vol. 30, no. 1, pp. 1–10, 1996.

[29] Lowe D. and Matthews R., *"Shakespeare vs. fletcher: A stylometric analysis by radial basis functions,"* Computers and the Humanities, vol. 29, no. 6, pp. 449–461, 1995.

[30] Argamon S., Saric M., and Stein S. S., *"Style mining of electronic messages for multiple*

authorship discrimination: first results," in Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003, 2003, pp. 475–480.

[31] Liu S., Liu Z., Sun J., and Liu L., *"Application of synergetic neural network in online writeprint identification,"* International Journal of Digital Content Technology and its Applications, vol. 5, pp. 126–135, 03 2011.

[32] Hoorn J., Frank S., Kowalczyk W., and van der Ham F., *"Neural network identification of poets using letter sequences,"* Literary and Linguistic Computing, vol. 14, no. 3, pp. 311–338, 1999.