
بسم الله الرحمن الرحيم

مرکز آموزشهای الکترونیکی

پایان نامه کارشناسی ارشد مهندسی هوش مصنوعی

شناسایی نویسنده با ویژگی‌های آماری یک سند متنی

نگارنده: **سمانه وزیریان**

استاد راهنما

دکتر مرتضی زاهدی

استاد مشاور

دکتر حسن‌پور

شهریور ۱۳۹۵

شماره: ۳/ ۲۲۹۸
تاریخ: ۹۵/۷/۱۰
ویرایش:

باسمه تعالی

دانشگاه شاهرود
مدیریت تحصیلات تکمیلی

فرم شماره (۶)

فرم صورت جلسه دفاع از پایان نامه تحصیلی دوره کارشناسی ارشد

با تأییدات خداوند متعال و با استعانت از حضرت ولی عصر (عج) نتیجه ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم سمانه وزیریان به شماره دانشجویی ۹۲۱۴۸۵۴ رشته کامپیوترگرایش ...هوش مصنوعی تحت عنوان: شناسایی نویسنده با ویژگی‌های آماری یک سند متنی

که در تاریخ ۹۵/۶/۱۵ با حضور هیأت محترم داوران در دانشگاه شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

قبول (با درجه: عالی امتیاز ۱۹) ☒ دفاع مجدد ☐ مردود ☐

۱ـ عالی (۲۰ ـ ۱۹)
۲ـ بسیار خوب (۱۸/۹۹ ـ ۱۸)
۳ـ خوب (۱۷/۹۹ ـ ۱۶)
۴ـ قابل قبول (۱۵/۹۹ ـ ۱۴)
۵ـ بدون کمتر از ۱۴ غیر قابل قبول

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استاد راهنما	دکتر مرتضی زاهدی	استاد یار	
۲ـ استاد مشاور	دکتر حسن مهر	استاد یار	
۳ـ نماینده شورای تحصیلات تکمیلی	حسن مهر		
۴ـ استاد ممتحن	دکتر هدی مسایخی	استاد یار	
۵ ـ استاد ممتحن	دکتر علی نوریان	استاد یار	

رئیس دانشکده علوم ریاضی و کامپیوتر امضاء

از پدر و مادر عزیزم که در همه مراحل زندگی در کنار من و بعنوان حامی حضور داشتند، سپاسگزارم و

از استاد عزیزم جناب آقای مرتضی زاهدی که با نهایت صبر به اینجانب در پیشرفت و به سرانجام

رسیدن این پایان نامه کمک نمودند، نهایت تشکر را دارم.

تعهد نامه

اینجانب **سمانه وزیریان** دانشجوی دوره کارشناسی ارشد رشته مهندسی هوش مصنوعی دانشکده مرکز آموزش های الکترونیکی دانشگاه صنعتی شاهرود نویسنده پایان نامه **شناسایی نویسنده با ویژگی‌های آماری یک سند متنی** تحت راهنمائی **دکتر مرتضی زاهدی** متعهد می‌شوم.

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه های رایانه ای، نرم افزارها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی‌باشد.

چکیده

شناسایی نویسنده یکی از زیر شاخه‌های پردازش متن است و هدف آن مشخص‌کردن نویسنده یک متن با نویسنده ناشناس است که با عنوان متن دیده نشده از آن یاد می‌شود. انتخاب نویسنده برای متن دیده نشده، از میان مجموعه‌ای از نویسنده‌گان با عنوان نویسنده‌گان کاندید، انجام می‌گیرد. برای طراحی سیستم شناسایی نویسنده نیاز به ویژگی‌های مشخص کننده سبک، و انتخاب روش برای طبقه‌بندی و یا تخصیص نویسنده مناسب برای متن دیده نشده است. معمولا انتخاب ویژگی از میان داده‌های آموزشی موجود برای هر نویسنده کاندید صورت می‌گیرد.

با افزایش روند تولید متن در زبان‌های مختلف نیاز به توسعه سیستم‌های تشخیص نویسنده که مستقل از زبان نگارش متن عمل نمایند ضروری به نظر می‌رسد. در این پایان‌نامه برای حل مسئله تشخیص نویسنده از n-گرام کارکترها و n-گرام کلمات به عنوان ویژگی‌ها با مزیت استخراج آسان و مستقل از زبان، استفاده شده است. و به عنوان یک روش آماری و احتمالاتی مدل‌سازی زبانی به عنوان روش تخصیص استفاده شده است. برای بهبود نتایج ترکیب 1-گرام کلمات با 2-گرام کلمات همراه با مقدار IDF به عنوان وزن دهی احتمال n-گرام‌ها با عنوان مدل سازی زبانی تغییر یافته پیشنهاد شده است. همچنین عوامل موثر بر نتیجه ارزیابی مانند کاهش داده آموزشی و افزایش تعداد نویسنده کاندید و متعادل بودن یا نا متعادل بودن داده آموزشی در حل مسئله تشخیص نویسنده بررسی شده است.

برای ارزیابی از چهار پایگاه داده استفاده شده است. سه پایگاه داده در زبان فارسی و پایگاه داده چهارم در زبان انگلیسی است. روش مدل سازی تغییر یافته در تمام آزمایشات انجام شده بهبود را نسبت به n-گرام کاراکترها و n–گرام کلمات نشان می‌دهد. بهترین نتیجه با رسیدن به دقت ۱۰۰٪ با مدل سازی تغییر یافته در پایگاه داده فارسی بهبود خوبی را در استفاده از n-گرام‌ها نشان می‌دهد.

کلمات کلیدی: شناسایی نویسنده، تخصیص نویسنده، انگرام، مدل‌سازی زبانی، پایگاه داده WMPR-AA2016-A، پایگاه داده WMPR-AA2016-B.

فهرست مطالب

عنوان. صفحه

فهرست شکل‌ها

فهرست جدول‌ها

۱- فصل اول: معرفی و کلیات مسئله

۱-۱ مقدمه

در این پایان نامه به بررسی مسئله شناسایی نویسنده پرداخته شده است. شناسایی نویسنده که به نوعی طبقه بندی متن است در دسته پردازش متن قرار می‌گیرد که خود زیر مجموعه مجموعه بزرگتر پردازش زبان طبیعی است.

شناسایی نوسنده عمل انتساب محتمل‌ترین نویسنده از میان مجموعه نویسنده‌گان شناخته شده[۱]، به متنی است که نویسنده آن نامشخص[۲] است. در تعریف دیگر از (Juola, et al., 2006) عمل استناج محاسباتی[۳] نویسنده یک متن دیده نشده، با استفاده از اطلاعات آماری که در متن وجود دارد را تخصیص نویسنده[۴] می‌نامند. در طول این پایان‌نامه همواره منظور از شناسایی نویسنده، شناسایی با استفاده از روشهای آماری و محاسباتی است. این فرایند با نام‌های Authorship Attribution, non-traditional Authorship Attribution, Stylometric, Authorship Identification در منابع مختلف نام برده شده است

۱-۲ ضرورت حل مسئله شناسایی نویسنده

مسئله تشخیص نویسنده از زمان‌های دور اهمیت داشته و مورد بررسی پژوهشگران زمان خود بوده است اما در دهه‌های اخیر (از اواخر ۱۹۹۰) با فراگیر شدن اینترنت، استفاده از صفحات وب و استفاده از محیط‌های چت و همچنین پیشرفتی که در زمینه‌های پردازش زبان طبیعی و الگوریتم‌های یادگیری ماشین انجام شده است توجه و پیشرفت‌های بیشتری در زمینه حل مسئله تشخیص نویسنده صورت گرفته است. اما متاسفانه این تحقیقات در زمینه زبان فارسی چشم گیر نبوده است.

---

[1] known Authors

[۲] Unknown Author

[۳] Computational Inferring

[۴] Authorship Attribution

پیشرفت پردازش زبان‌های طبیعی و الگوریتم های یادگیری ماشین بر حل مسئله شناسایی نویسنده همان‌طورکه در (Stamatatos, 2009) به آن اشاره شده است می‌تواند به صورت زیر تاثیر گزار باشد:

- پردازش زبان‌های طبیعی: با گسترش ابزارهای پردازش زبان امکان استفاده از انواع جدید از ویژگی مانند ویژگی‌های نحوی برای آشکار کردن سبک نویسنده فراهم شده است.
- الگوریتم‌های یادگیری ماشین: با قدرتمندتر شدن این الگوریتم‌ها، استفاده آنها به عنوان روش تخصیص و یا طبقه بندی متن نیز با اثر مثبت همراه بوده است.

از طرف دیگر مسئله تشخیص نویسنده به خاطر کاربردهای گسترده آن در حوزه‌های مختلف، روز به روز از اهمیت بیشتری برخوردار می‌شود. از جمله کاربردهای آن می‌توان به موارد زیر اشاره کرد:

- تجزیه و تحلیل نوشته‌ها به صورت قانونی[۱]: تجزیه و تحلیل نوشته‌ها به صورت قانونی به منظور پیدا کردن نویسنده در تحقیقات قانونی و بررسی‌های مجرمانه مانند حملات سایبری صورت می‌گیرد.
- تجارت الکترونیکی[۲]: بررسی ادعاهای مالکیت نرم‌افزارهای تجاری با بررسی کد برنامه‌ها .
- شناسایی نویسنده متن‌های ادبی (Bozkurt, et al., 2007)،
- شناسایی تروریسم با استفاده از شناسایی نویسنده در پیام‌ها
- تشخیص نویسنده ایمیل و متن‌های الکترونیکی (Argamon, et al., 2003)،
- تشخیص سرقت ادبی (Alzahrani, et al., 2012).
- و ...

مسئله تشخیص نویسنده، بسته به نحوه تخصیص و هدف از حل مسئله، به چند زیر دسته تقسیم می‌شود؛ در تقسیم بندی که در (Joula, 2008) انجام شده است به سه گروه تقسیم شده است:

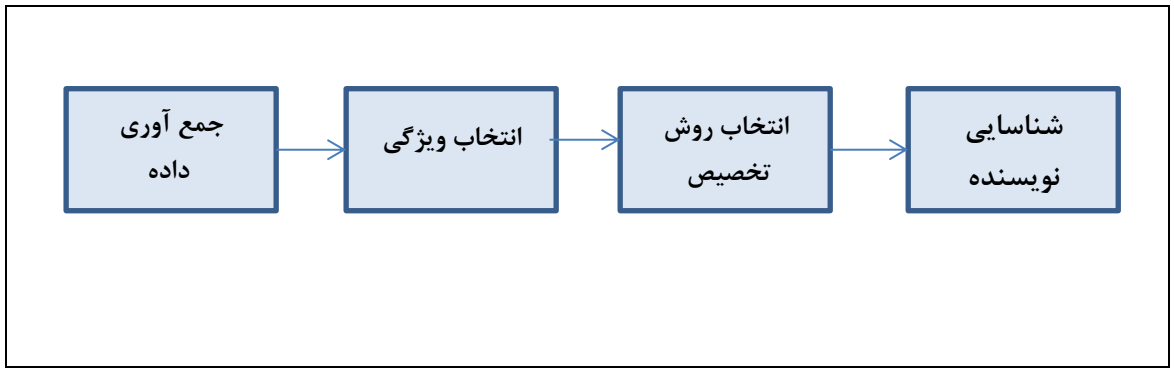
---

[۱] Forensic Analysis

[۲] Electronic Commerce

- مجموعه بسته[۱]: انتخاب نویسنده متن از روی مجموعه بسته‌ای از نویسنده‌گان کاندید که برای آنها داده آموزشی وجود دارد.
- مجموعه باز[۲]: در این حالت نویسنده متن دیده نشده می‌تواند یکی از نویسنده‌گان مجموعه بسته نویسنده‌گان کاندید با داده آموزشی باشد یا هیچکدام از آنها نباشد.
- پروفایل[۳]: در این دسته تمرکز در مشخص کردن ویژگی‌های فردی نویسنده مانند جنسیت است و یا ویژگی متن مانند آیا نوشته به زبان مادری نویسنده نگارش شده است است یا متن توسط یک نویسنده نوشته شده است یا چند نویسنده و...

به طور کلی می‌توان گفت در شناسایی نویسنده با دو مسئله روبرو هستیم، یکی ویژگی‌ها که مشخص کننده سبک نوشتاری نویسنده هستند و دیگری روش‌های کمی[۴] که ویژگی‌ها را به منظور انجام انتساب یا تخصیص به کار می‌برند. یک نمای کلی از مراحل کار در حل مسئله شناسایی شکل ۱-۱ نمایش داده شده است.



شکل۱-۱: مراحل کار در حل مسئله تشخیص نویسنده

[۱] Close- set Authorship Attribution
[۲] Open- set Authorship Attribution
[۳] Profiling
[۴] Quantitative Methods

۱-۳ بیان مسئله

همان‌طورکه عنوان شد عمل شناسایی نویسنده در مجموعه بسته، سعی بر شناسایی نویسنده و یا تخصیص نویسنده در متن مورد مناقشه از میان مجموعه نویسنده‌گان شناخته شده دارد.

در این تعریف منظور از متن مورد مناقشه یا متن دیده نشده یا متن مورد سوال یا متن با نویسنده نامشخص، متنی است که نویسنده آن مشخص نیست. منظور از نویسنده‌گان شناخته شده یا مجموعه نویسنده‌گان کاندید، مجموعه‌ای از نویسنده‌گان است که برای آنها داده آموزشی وجود دارد یا به عبارتی شناخته شده هستند و منظور از حل مسئله شناسایی نویسنده تخصیص یکی از نویسنده‌گان مجموعه کاندید به متن دیده نشده است. به این منظور به طور معمول از یک روش تخصیص و گروهی از ویژگی‌ها، اعم از لغوی، نحوی، معنایی، و...یا ترکیب ویژگی‌ها با یکدیگر استفاده می‌شود.

با افزایش روند تولید متن در زبان‌های مختلف نیاز به توسعه سیستم‌های تشخیص نویسنده که مستقل از زبان نگارش متن عمل نمایند ضروری به نظر می‌رسد. در این پایان‌نامه برای حل مسئله تشخیص نویسنده از n-گرام کارکترها و n-گرام کلمات به عنوان ویژگی‌ها با مزیت استخراج آسان و مستقل از زبان، استفاده شده است. همچنین به عنوان یک روش آماری، روش احتمالاتی مدل‌سازی زبانی به عنوان روش تخصیص استفاده شده است. برای بهبود نتایج ترکیب 1-گرام کلمات با 2-گرام کلمات و استفاده همراه با مقدار IDF به عنوان وزن دهی احتمال n-گرام‌ها با عنوان مدل سازی زبانی تغییر یافته پیشنهاد شده است.

در این مطالعه از چهار پایگاه داده، استفاده شده است که دو پایگاه داده آن توسط نگارنده جمع آوری شده است. پایگاه داده اول مجموعه رباعیات شش شاعر در زبان فارسی استفاده شده است. پایگاه داده دوم از مجموعه غزلیات هفت شاعر در زبان فارسی تهیه شده است. پایگاه داده سوم مجموعه متن‌های ادبی جمع‌آوری شده از چهل نویسنده فارسی زبان است و پایگاه داده چهارم در زبان انگلیسی و دارای پنجاه نویسنده کاندید است. همچنین در پایان عوامل موثر بر نتیجه ارزیابی مانند کاهش داده

آموزشی و افزایش تعداد نویسنده کاندید و متعادل بودن یا نـا متعـادل بـودن داده آموزشـی در حـل مسئله تشخیص نویسنده بررسی شده است.

۱-۴ ساختار پایان‌نامه

در فصل اول تعریف و مقدمه‌ای از شناسایی نویسنده و کاربردهای آن گفته شده است. در فصل دوم به دسته بندی ویژگی‌ها و متدهای تخصیص استفاده شده پرداخته شده و نمونـه پـژوهش‌هـای گذشـته مرور شده است. در فصل سوم به معرفـی چهـار پایگـاه‌داده، خصوصـیات پایگـاه داده، روش و ویژگـی انتخاب شده در این پایان‌نامه پرداخته شده است. در فصل چهارم نتایج آزمایشـات و ارزیـابی سیسـتم گزارش شده است و در فصل پنجم جمع‌بندی و نتیجه گیری انجام شده است.

۲- فصل دوم: مروری بر کارهای پیشین

۲-۱ مقدمه

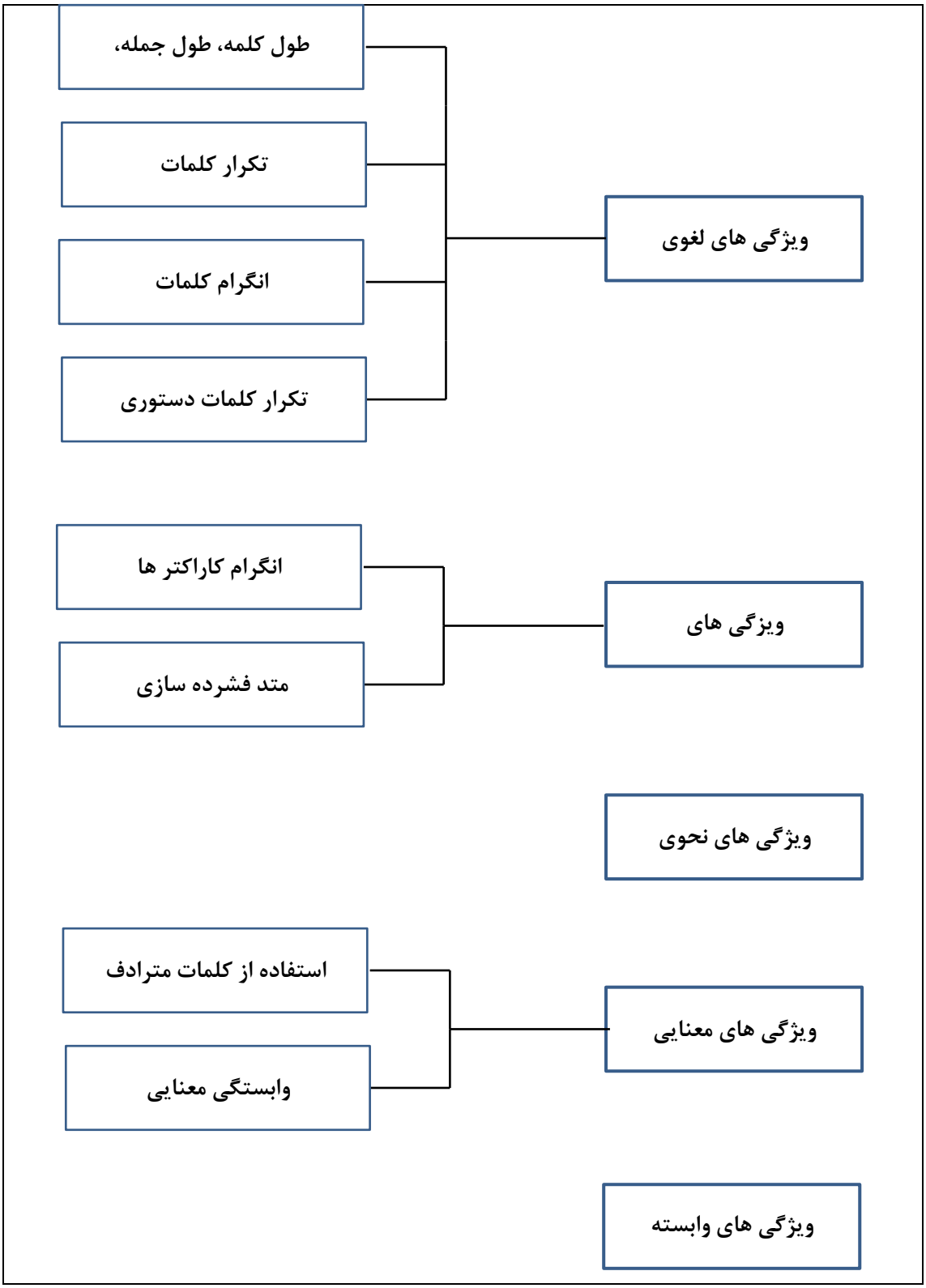
در این فصل همراه با مروری بر نمونه پژوهش‌های پیشین در حل مسئله نویسنده، دسته بندی ویژگی‌ها و روش‌های تخصیص به کار گرفته شده در این زمینه، نیز انجام شده است.

فصل ابتدا با توضیح ویژگی‌های سبکی در بخش ۲-۲ شروع می‌شود. در بخش ۲-۳ پس از تعریفی در رابطه با نحوه تشکیل داده‌های آموزشی در روش‌های مختلف به متدهای تخصیص پرداخته شده است. و در انتها در بخش ۲-۴ مروری بر کارهای انجام شده بر روی پایگاه‌داده با زبان فارسی انجام شده است.

۲-۲ ویژگی‌های سبکی

ویژگی‌های سبکی[1] ویژگی‌های معرفی کننده سبک نویسنده به صورت کمی هستند. در ادامه به طبقه بندی انواع ویژگی‌ها پرداخته شده است و همچنین به نمونه کارهای انجام شده در این راستا، اشاره شده است. بر مبنای مطالعات انجام شده در (Stamatatos, 2009) ویژگی‌ها را می‌توان به پنج دسته اصلی تقسیم بندی کرد. این دسته بندی در شکل ۲-۱ نمایش داده شده است.

[1] Stylometric Features



شکل ۲-۱: طبقه بندی ویژگی‌ها

۲-۲-۱ ویژگی‌های لغوی

در دسته بندی ویژگی‌های لغوی[1] متن به صورت رشته‌ای از توالی توکن ها در نظر گرفته شده است. هر توکن در آن می‌تواند یک کلمه یا یک عدد و یا یک علامت نگارشی باشد و حتی در تقسیم بندی بزرگتر می‌تواند به صورت جمله تقسیم بندی شود. با این دید به متن، از جمله ویژگی که در این گروه قرار می‌گیرند طول کلمه، طول جمله، غنی بودن از لحاظ لغت، (که در ساده‌ترین روش از نسبت بین تعداد لغات بدون تکرار بر کل لغات متن به دست می‌آید.) است. در اولین کارهای انجام شده در زمینه شناسایی نویسنده تنها از یک متغیر آماری[2] مانند میانگین طول کلمات یا میانگین طول جملات استفاده شده است(Holmes, (1998) .

ویژگی بعدی در این گروه، تکرار کلمات[3] و n-گرام کلمات[4] است. مزیت ویژگی‌ها در این دسته سادگی محاسبه آنها است؛ آنها برای محاسبه نیاز به ابزارهای پردازش زبان طبیعی پیشرفته‌ای ندارند و در اغلب موارد با ابزارهای ساده‌ای مثل جداکننده توکن[5] بر اساس فاصله به دست می‌آیند. این مزیت آنها را تبدیل به ویژگی مستقل از زبان می‌کند. یعنی در زبان‌های مختلف ابزار محاسبه آن یکسان است (Keselj, et al., 2003). البته در این میان استثناهایی نیز وجود دارد مثل زبان چینی‌که درآنها از فاصله برای جدا کردن کلمات استفاده نمی‌شود و برای جداکردن کلمات به ابزارهای پیچیده‌تری نیاز است. از نمونه کارها با این ویژگی می توان به کار (Stamatatos, 2006) اشاره کرد که در آن از هزار کلمه پر تکرار به عنوان ویژگی استفاده شده است و همچنین (Howedi & Mohd, 2014) که از n-گرام کلمه استفاده کرده است و یا (Luyckx & Daelemans, 2011) ۳-گرام کلمات را به کار برده است و کار (Mikros & Perifanos, 2013) که تلفیق n-گرام کاراکتر با n-گرام کلمه را به عنوان ویژگی انتخاب کرده است همچنین در کار (Rappoport & Koppel, 2013) از الگوی توالی کلمات با

[1] Lexical Features
[2] Univariant Statistic
[3] Word Frequency
[4] Word n-gram
[5] Tokenizer

شروع و پایان با کلمات پرتکرار به عنوان ویژگی تعریف شده است. به این صورت که توالی کلمات منطبق بر الگوی یاد شده از داده‌های آموزشی استخراج می‌شود و داده‌های آزمایشی براساس تطبیق با ویژگی‌های استخراج شده دسته‌بندی می‌گردد.

ویژگی دیگر که می‌توان آن را نوعی ویژگی تکرار کلمات نیز در نظر گرفت استفاده از کلمات دستوری[1] است. استفاده از کلمات دستوری از این جهت مورد توجه قرار گرفته است که به محتوای متن و موضوع بستگی ندارند و از این رو گمان می‌رود که استفاده از آنها توسط نویسنده، خارج از اراده است و در نتیجه می‌تواند نمایش دهنده سبک نوشتاری نویسنده باشند (Stamatatos, 2009). به عنوان یکی از شناخته شده‌ترین و اولین کارها که با استفاده از روش‌های آماری به حل مسئله نویسنده پرداخته است کار Mosteller و Wallace (Mosteller, et al., 1964) از این دسته از ویژگی-ها استفاده کرده است. آنها سعی داشته‌اند تا با استفاده از کلمات دستوری، دوازده عدد از مقالات فدرالی را که نویسنده آنها مشخص نبوده است را به سه نویسنده (Alexander Hamilton, James Adison وJohn Jay) تخصیص دهند. بعد از آن در کارهای اولیه بعدی از چند متغیره آماری[2] استفاده شده است از جمله (Burrows & F, 1992) که در آن از هفتاد و پنج کلمات دستوری پر تکرار به عنوان متغیر آماری انتخاب شده است و سپس با استفاده از PCA متغیرها کاهش داده شده است. از دیگرکارها در این گروه، می‌توان به کار(Boukhaled & Ganascia, 2015) اشاره کرد که درآن با مقایسه استفاده از تکرار کلمات دستوری در مقابل ویژگی‌های استخراج شده با روش قوانین پی‌درپی[3]، استفاده از کلمات دستوری را کاراتر عنوان کرده است.

[1] Function Word
[2] Multivariate Statistic
[3] Sequential Rule

۲-۲-۲ ویژگی‌های کاراکتری

در دسته ویژگی‌های کاراکتری[1] متن به صورت رشته‌ای از توالی کاراکترها در نظر گرفته می‌شود. از جمله ویژگی‌هایی که در این گروه قرار می‌گیرند نوع کاراکتر (عددی، حرفی و...)، تکرار کاراکترها، n-گرام در سطح کاراکتر[2] و متدهای فشرده سازی است. با وجود کارهای زیادی که با استفاده از کاراکتر n-گرام‌ها انجام شده است و نتایج خوبی که به دست آمده است مفید بودن آنها نشان داده شده است. مزیت استفاده از کارکتر n-گرام در این است که اختلافات جزئی در سبک نوشتاری را مشخص می‌کند و تحمل پذیری خوبی در برابر نویز در متن (متن نویزی متنی است که دارای خطاهای گرامری است و علائم نگارشی در آن درست استفاده نشده است) دارد. اما مشکلی که وجود دارد در تعین مقدار n به صورت کارا است. یک مقدار بزرگ برای n باعث در برگرفتن ویژگی‌های موضوعی علاوه بر ویژگی‌های متنی می‌شود و از طرفی انتخاب مقدار کوچک n دربرگیرنده ویژگی های متنی به طور کافی نخواهد بود. بهترین انتخاب برای مقدار n وابسته به زبان و متد استفاده شده است ((Stamatatos, 2009)، (Keselj, et al., 2003)).

از نمونه کارهای انجام شده در این زمینه، کار (Howedi & Mohd, 2014) است که درآن از کارکتر n-گرام ها ی ۱، ۲ و۳ به صورت جداگانه استفاده شده است و یا (Stamatatos, 2006) که از کنار هم قرار دادن کارکتر n-گرام‌ها با مقدار n، ۳، ۴ و ۵ با تعداد تکرارهای مشخص در یک بردار ویژگی، استفاده شده است. و همچنین کار (Escalante, 2011) که از هیستاگرام n-گرام کاراکترها استفاده کرده است.

[1] Character Features
[2] Character n-gram

۲-۲-۳ ویژگی‌های نحوی

در گروه ویژگی‌های نحوی[1] از ویژگی‌های ساختاری و نحوی متن، مانند نقش کلمات به عنوان ویژگی استفاده می‌گردد. علت قرار گرفتن این دسته از ویژگی‌های متنی در گروه ویژگی‌های سبکی، این عقیده است که نویسنده‌گان الگوهای نحوی را به صورت ناخودآگاه استفاده می‌کنند و بنابراین در تمامی نوشته‌های خود الگوهای مشابه‌ای را دنبال خواهند کرد (Stamatatos, 2009).

یکی از ساده‌ترین حالات آن استفاده از یک برچسب زن گفتار[2] است. از نمونه کاربرد آن می‌توان از کار (Diederich, et al., 2003) و (Zhao & Zobel, 2007) نام برد، در کار اول با استفاده از یک برچسب زن گفتار ساختار کلمات در سه گروه اسم، فعل، صفت و کلمات دستوری مشخص شده است و سپس بردار ویژگی متشکل از سه زیر بردار تکرار برچسب ها، تکرار ۲-گرام برچسب ها و تکرار کلمات با طول متفاوت ایجاد شده است. در کار دوم برچسب گفتار و ۲- گرام برچسب ها از جمله ویژگی‌های به کار رفته است.

حالت بعدی مربوط به استفاده از ابزارهای پردازش متن برای تشخیص جمله و عبارات اسمی و فعلی است. مانند کار (Luyckx & Daelemans, 2005) که در آن عبارات اسمی یکی از نه گروه ویژگی است که مورد استفاده قرارداده است. دامنه استفاده از این ویژگی‌ها تا استفاده از میزان خطای نحوی، وابستگی نحوی توکن‌ها و بسیاری موارد دیگر به عنوان ویژگی ادامه پیدا کرده است.

مشکلی که در رابطه با این دسته از ویژگی‌ها وجود دارد، این است که برای استخراج این گروه از ویژگی‌ها نیاز به ابزارهای پردازش متن قوی و دقیق است و به این ترتیب مسئله استخراج این ویژگی-ها از زبانی به زبان دیگر متفاوت می‌شود و به عبارت دیگر استخراج آنها را وابسته به زبان کرده است. علاوه بر آن به خاطر خطای حاصل از استفاده از ابزارهای پردازش و تجزیه کننده متن خطای استفاده از ابزارها به سیستم تحمیل می‌گردد (Stamatatos, 2009).

---

[1] Syntactic Features
[2] Part Of Speech Tagger

۲-۲-۴ ویژگی‌های معنایی

در ویژگی‌های معنایی[1] توجه روی معنا و نقش معنایی کلمات است که در سطح کلمات عبارات و یا جملات وجود دارد. برای روشن شدن نقش ویژگی معنایی، جمله "درب به علی لگد زد" را درنظر بگیرید این جمله از لحاظ گرامری و نوشتاری بدون خطا است اما چیزی که در این جمله اشتباه است معنا است. این که کلمه "درب" با وجود معنایی که دارد نمی‌تواند دراین جایگاه نحوی قرار گیرد.

می‌توان از تجزیه‌گر معنایی در مشخص کردن ویژگی‌هایی مانند لغات مترادف، متضاد، لغات با چند معنا و وابستگی معنایی[2] استفاده کرد. اما همچنان که گفته شد برای استخراج این ویژگی‌ها نیاز به ابزارهای پارس و تحلیل معنایی است. در اکثر کارها از ترکیب این ویژگی‌ها با ویژگی‌های نحوی و لغوی استفاده شده است. به عنوان نمونه در کار (Gamon, 2004) از ویژگی‌های معنایی استخراج شده از گراف وابستگی معنایی، در ترکیب با دو ویژگی نحوی و کلمات دستوری استفاده شده است.

از دیگر نمونه‌های ویژگی معنایی استفاده از تکنیک نمایه سازی معنایی نهفته[3] است. در این تکنیک ابتدا ماتریس لفظ–سند[4] تشکیل می‌گردد (در واقع هر سطر از این ماتریس بردار ویژگی هر سند می-باشد) در مرحله بعد با متد SVD[5]، فاکتورگیری روی ماتریس انجام می‌شود سپس با استفاده از مقادیر ویژه به دست آمده ماتریسی معادل ماتریس اول ساخته می‌شود و در نهایت با استفاده از متدهای اندازه‌گیری فاصله برداری، شباهت سندها در فضای جدید سنجیده می‌شود. ایده اصلی در این تکنیک بر اساس نگاشت بردار ویژگی‌ها به فضای برداری با تعداد بعد کمتر است به طوری که ویژگی‌ها در بعد جدید آشکارکننده ارتباطات معنایی مورد نظر باشند. از کارهای انجام شده در این زمینه می‌توان به کارهای (Soboroff, et al., 1997) و (Satyam, et al., 2014) اشاره کرد که از کارکتر n-گرام‌ها برای تشکیل ماتریس لفظ–سند استفاده کرده است.

---

[1] Semantic Features
[2] Semantic Dependencies
[3] Latent Semantic Indexing
[4] Term-document
[5] Singular Value Decomposition

۲-۲-۵ ویژگی‌های وابسته به کاربرد[1]

بسته به کاربرد مسئله تشخیص نویسنده در حوزه‌های مختلف مانند فروم‌ها و یا پیام‌های الکترونیکی می‌توان از ویژگی‌های مختص آن حوزه استفاده کرد. مانند کار (Abbasi & Chen, 2005) که در آن از ویژگی‌هایی از قبیل سایز و رنگ فونت استفاده شده است.

۲-۳ روش‌های کمی

پس از بررسی ویژگی‌های مختلف استفاده شده در حل مسئله تخصیص نویسنده در این بخش به دسته بندی وبررسی متدهای تخصیص در این حوزه پرداخته خواهد شد. اما قبل از آن تعریف روش-های تشکیل و به کارگیری داده‌های آموزشی در روش‌های مختلف تخصیص پرداخته شده است.

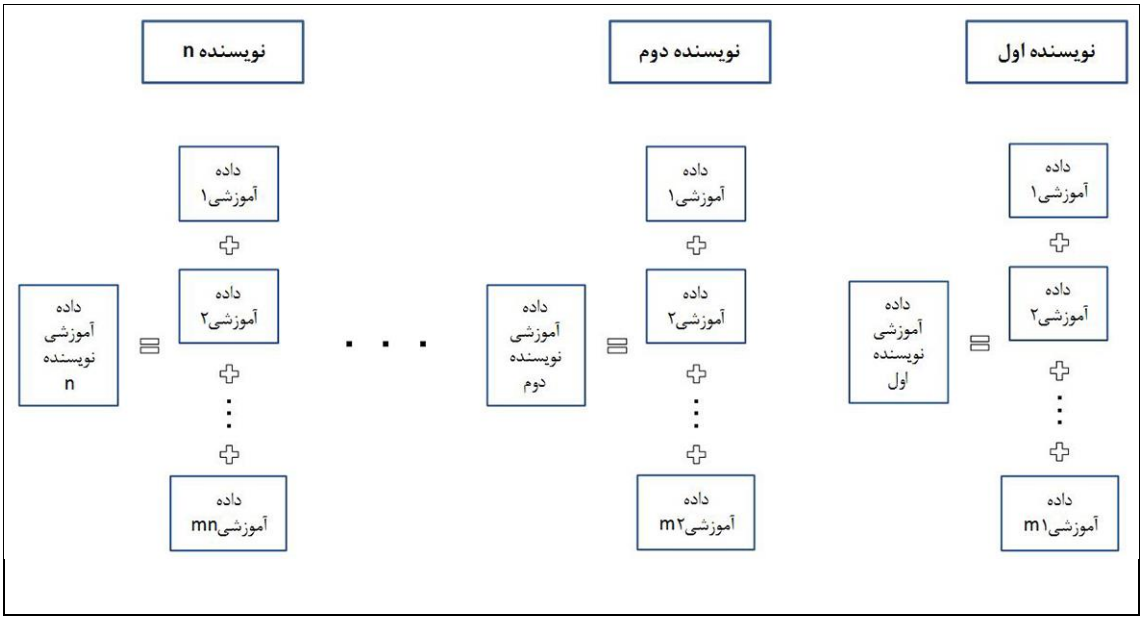
۲-۳-۱ تشکیل داده‌های آموزشی

به طور کلی در روش‌های مختلف تخصیص، داده آموزشی نویسنده‌گان کاندید به دو صورت تشکیل و استفاده می‌شود:

- مبتنی بر پروفایل[2]

تمام متن‌های آموزشی هر نویسنده در یک فایل الحاق می‌گردد به صورتی که برای هر نویسنده در مجموعه نویسنده‌گان کاندید یک فایل متنی با عنوان فایل آموزشی وجود دارد و ویژگی‌های سبکی از فایل الحاق شده استخراج خواهند شد (شکل ۲-۲).

[1] Spplication Specific
[2] Profile-Based

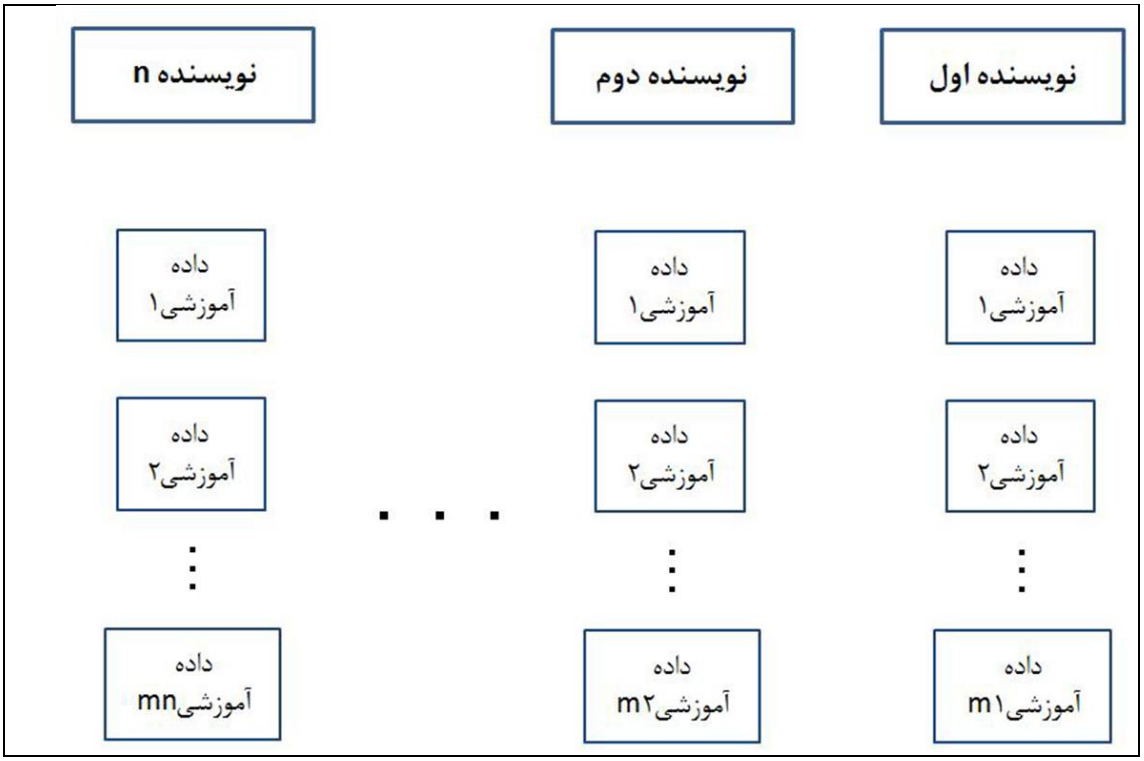


شکل ۲-۲: داده آموزشی مبتنی بر پروفایل

- مبتنی بر نمونه[1]

هر یک از فایل‌های متنی موجود برای هر نویسنده به صورت جداگانه در تخصیص شرکت می‌کنند. در نتیجه برای هر نویسنده در مجموعه کاندید می‌تواند چند فایل آموزشی وجود داشته باشد (شکل۲-۳). در ادامه به معرفی متدهای تخصیص که در حل مسئله شناسایی نویسنده به کار رفته است می‌پردازیم.

[1] Instance-Based



شکل ۲-3: داده آموزشی مبتنی بر نمونه

۲-۳-۲ روش‌های تفاضلی[1]

در این دسته از روش‌ها برای طبقه بندی کلاس‌ها (در این جا متن‌ها) با استفاده از احتمال شرطی P(c|x) به طور مستقیم نمونه x را به کلاس c نگاشت می‌کنند.(Jordan, 2002) تشکیل بردار آموزشی در این روش‌ها به هر دو صورت مبتنی بر نمونه و مبتنی بر پروفایل صورت گرفته است.

۲-۳-۲-۱ روش‌ها بر اساس فاصله

در روش‌های بر اساس فاصله[2] از یکی از معیارهای اندازه‌گیری فاصله برداری به عنوان روش تخصیص استفاده می‌گردد. از نمونه کارهایی که در این دسته قرار می‌گیرد می‌توان به کار (Stamatatos, et al., 2000) اشاره کرد که در آن با استخراج بیست و دو متغیر و تشکیل بردار ویژگی برای متن دیده نشده و اندازه‌گیری فاصله ماهالونوبی[3] از مرکز هر گروه از داده‌های آموزشی، متن دیده نشده به یکی از

[1] Discriminative Methods
[2] Distance Method
[3] Mahalonobi

نویسنده‌گان کاندید تخصیص داده می‌شود. به عنوان نمونه دیگر در کار (Chaski, 2001) با استفاده از آزمون مربع کای[1] متن‌های دیده نشده در بین چهار نویسنده کاندید دسته‌بندی شده است. در (Keselj, et al., 2003) برداری از تکرار کارکتر n-گرام‌هایی که بیشترین تکرار را دارند برای هر نویسنده کاندید با استفاده از داده‌های آموزشی تشکیل می‌شود و همین‌طور برداری مشابه برای هر یک از متن‌های دیده نشده آزمایشی تشکیل می‌شود سپس با اندازه‌گیری یک معیار فاصله بر اساس مجموع مربعات، میزان شباهت بین دو بردار سنجیده می‌شود و متن دیده نشده به نویسنده‌ای که با آن کمترین فاصله یا بیشترین شباهت را دارد تخصیص می‌یابد. مشکل این روش این‌طور بیان شده است که اگر داده آموزشی مربوط به یک نویسنده کوچک باشد باعث می‌شود که فاصله بین بردار آموزشی نویسنده با متن کوچک تر در مقایسه با سایر بردارهای آموزشی کمتر شود.

در کار (Frantzeskou, et al., 2006) مشابه کار قبل بردار ویژگی بر اسای تکرار کارکتر n-گرام‌هایی که بیشترین تکرار را دارند برای داده های آموزشی و آزمایشی تشکیل شده است اما این بار به عنوان معیار فاصله از اندازه اشتراک بردار داده آموزشی و داده آزمایشی استفاده شده است و در نهایت متن دیده نشده به نویسنده‌ای که بزرگترین معیار فاصله را با آن دارد تخصیص می‌یابد.

در (Stamatatos, 2007) ابتدا یک متن از الحاق تمام متن‌های آموزشی مربوط به تمام نویسنده‌گان کاندید ایجاد شده است و برداری از تکرار کارکتر n-گرام‌ها برای متن الحاقی تشکیل می‌شود و بردار داده نرمال خوانده می‌شود؛ سپس اختلاف بین تکرار n-گرام‌ها در متن دیده نشده با بردار نرمال محاسبه شده است و به عنوان وزن به معیار فاصله در کار (Keselj, et al., 2003) اضافه می‌گردد؛ در نهایت متن دیده نشده به نویسنده‌ای که با آن کمترین فاصله یا بیشترین شباهت را دارد تخصیص می‌یابد. در (Savoy, 2012) با در نظر گرفتن یک توزیع دو جمله‌ای روی کل کلمات برای هر کلمه وزنی را بر اساس امتیاز z[2] محاسبه شده است سپس به عنوان معیار فاصله بین متن دیده نشده و متن

[1] Chi-squared test
[2] Z score

آموزشی، میانگین اختلاف امتیاز z برای کلمات موجود درمتن دیده نشده و کلمه متنـاظر آن در داده آموزشی محاسبه می گردد. در نهایت متن دیده نشده به نویسنده‌ای که با داده آموزشی مربوط به آن کمترین فاصله را دارد تخصیص می‌یابد.

۲-۳-۲-۲ روش‌های فشرده سازی

از نمونه کارهای انجام شده در این زمینه می‌توان به کار (Marton, et al., 2005) اشاره کرد. این روش را می توان به صورت مراحل زیر توصیف کرد:

۱- داده آموزشی به صورت مبتنی بر پروفایل برای هر یک از نویسنده‌گان کاندیـد تشـکیل مـی-گردد.

۲- به وسیله یک روش فشرده‌سازی فایل آموزشی فشرده می‌گردد.

۳- در این مرحله فایل متن دیده نشده به فایل داده آموزشی هر نویسنده کاندیداضافه می‌گـردد. و متد فشرده سازی بر فایل حاصل اعمال می‌گردد.

۴- اختلاف بیتی بین دو فایل فشرده ایجاد شده در مرحله دو و مرحله سه محاسـبه مـی‌شـود و نویسنده با کمترین اختلاف، نویسنده برگزیده برای متن دیده نشده خواهد بود.

۲-۳-۲-۳ روش‌های یادگیری ماشین

از جمله رهیافت‌ها در دسته روش‌های یادگیری ماشین[1] می‌توان به موارد زیر اشاره کرد:

بردار ماشـین : ۲۰۰۱ (Vel, et al., 2001)، ۲۰۰۳(Diederich, et al., 2003) ، ۲۰۰۳ (Koppel & Schler, 2003)، ۲۰۰۴ (Gamon, 2004) ، ۲۰۰۴(BEKKERMAN & ALLAN, 2004) ، ۲۰۰۵ (Abbasi & Chen, 2005)، ۲۰۰۶(Zheng, et al., 2006) ، ۲۰۰۷(Türkoğlu, et al., 2007) ، ۲۰۰۷(Pavelec, et al., 2007) ، ۲۰۰۸(Stamatatos, 2008) ، ۲۰۱۱(Silva, et al., 2011) ، ۲۰۱۱ (Kourtis, et al., 2011)، ۲۰۱۳(Sidorov, et al., 2013) ، ۲۰۱۵(Nagaprasad, et al., 2015) .

[1] Machine Llearning Methods

درخت تصمیم[1]: ۲۰۰۴(Ramyaa, et al., 2004) ، ۲۰۰۵(Zhao & Zobel, 2005) ، ۲۰۰۵ (Uzuner & Katz, 2005)، ۲۰۰۵ (Abbasi & Chen, 2005)، ۲۰۰۶ (Zheng, et al., 2006)، ۲۰۰۹ (Koppe, et al., 2009)، ۲۰۱۳(Abdallah, et al., 2013)

شبکه عصبی[2]: ۱۹۹۴(KJELL, 1994) ، ۱۹۹۶(Tweedie, et al., 1996) ، ۲۰۰۴ (Ramyaa, et al., 2004)، ۲۰۰۶ (Zheng, et al., 2006)، ۲۰۰۸(Tearle, et al., 2008) ، ۲۰۱۲ (Jamak, et al., 2012)، ۲۰۱۲(Selman, et al., 2012) .

۲-۳-۳ روشهای تولیدی[3]

در روش‌های تولیدی از احتمال توام کلاس c و متغیر x ((P(c,x) استفاده می‌شود و احتمال شرطی P(c|x) با استفاده از قانون بیز محاسبه می‌گردد (Jordan, 2002). و داده‌های آموزشی عموما به صورت مبتنی بر پروفایل استفاده می‌شوند.

۲-۳-۳-۱ مدل احتمالاتی بیز

در این مدل از طبقه بند بیز استفاده شده است. از نمونه کارهای انجام شده در این زمینه می‌توان به کار (Peng & Schuurmans, 2003) اشاره کرد که در آن با استفاده از تکنیک‌های هموار کردن مدل-سازی زبانی و در نظر گرفتن وابستگی بین n-گرام‌ها بیز ساده را بهبود داده است. در کار (Altheneyan, 2014) از بیز روی بانک اطلاعاتی به زبان عربی استفاده شده است و از چهار مدل بیز متفاوت در ویژگی‌ها استفاده شده است. درکار (Boutwell, 2014) از کارکتر n-گرام‌های دو،سه،چهار...شش و بیز ساده استفاده شده است و در بانک اطلاعاتی جمع آوری شده از پیام‌های کوتاه به عنوان یک پایگاه داده متن کوتاه اعمال شده است. همچنین در این کار با الحاق چند متن

[1] Decision Tree
[2] Neural Network
[3] Generative Method

کوتاه در یک متن و ایجاد متن بزرگتر توانسته با فراهم کردن اطلاعات زبانی بیشتربه میزان دقت بالاتری برسد.

۲-۳-۳-۲ مدل سازی زبانی

از آنجا که مدل سازی[1] زبانی روش انتخابی روی این تحقیق است در بخش سوم به صورت کامل توضیح داده خواهد شد. در این قسمت تنها به ذکر چند نمونه از کارهای انجام شده در این زمینه می‌پردازیم. در اکثر کارها داده آموزشی به صورت پروفایل استفاده شده است و داده آموزشی هر نویسنده به صورت یک مدل زبانی در نظرگرفته شده است و احتمال هر متن با نویسنده ناشناس بر اساس مدل سازی زبانی اندازه‌گیری می‌شود و سپس با معیارهایی مانند پرپلکسیتی[2] یا KLD با یکدیگر مقایسه می‌شوند و نویسنده با بهترین مقدار به عنوان نویسنده متن ناشناس انتخاب می‌گردد. به عنوان یکی از اولین کارهای انجام شده در این زمینه کار (Peng & Schuurmans, 2003) است. در این کار از کاراکتر n-گرام‌ها به عنوان ویژگی استفاده شده است و از مدل‌سازی زبانی برای طبقه‌بندی استفاده شده است. بهترین نتیجه برای کارکترگرام‌ها با مقدار n ۲ و۳ به دست آمده است. برای نمایش دقت روش ارائه شده از چهار پایگاه داده در سه زبان انگلیسی، گریگ و چینی استفاده شده است و بهترین نتیجه را در پایگاه داده گریک[3] گزارش شده است. در کار (Zhao, et al., 2006) از مدل سازی زبانی ساده و کلمات دستوری، تگ‌های دستوری و جداکننده‌ها هر یک به صورت جداگانه و همچنین به صورت تلفیق هر سه به عنوان ویژگی استفاده شده است و با مقایسه مقدار معیار KLD[4] محاسبه شده برای متن دیده نشده با هریک از نویسنده‌گان مجموعه آموزشی، نویسنده با کمترین مقدار را به عنوان نویسنده متن ناشناس انتخاب کرده است. برای سنجش میزان دقت از سه بانک اطلاعاتی در زبان انگلیسی استفاده شده است. درکار (Azarbonyad, et al., 2015) چالش اصلی

[1] Language Modeling
[2] Perplexity
[3] Greek
[4] Kullback-Leibler Divergence

تشخیص نویسنده در متن‌های کوتاه با در نظر گرفتن تغییرات سبک نوشتاری هر نویسنده در گذر زمان است. در این مدل دوره زمانی هر نویسنده کاندید به چندین دوره با طول ثابت تقسیم شده است و سپس با استفاده از مدل سازی زبانی در سطح کاراکتر همراه با ضریبی از تابع تغییرات سبک نوشتاری در طول دوره زمانی نویسنده، احتمال تولید متن دیده نشده توسط هر نویسنده محاسبه می‌گردد و در نهایت نویسنده کاندید با بالاترین احتمال به عنوان نویسنده متن دیده نشده انتخاب می‌گردد.

در بعضی از کارها از معیارهای هموارسازی در مدل سازی زبانی برای بهبود روش‌های احتمالاتی مانند بیز استفاده شده است. در این زمینه می‌توان به کار (Peng & Schuurmans, 2003) اشاره کرد که در آن با استفاده از تکنیک‌های هموار کردن مدل‌سازی زبانی و در نظر گرفتن وابستگی بین n-گرام‌ها بیز ساده را بهبود داده است. در دسته‌ای از کارها از مقادیر وابسته به مدل‌سازی زبانی به عنوان یک ویژگی استفاده شده است مانند کار پیلا در مقاله (Pillay & Solorio, 2010) که از یک مدل دو مرحله‌ای ترکیبی از دو روش یادگیری با ناظر و بدون ناظر برای تشخیص نویسنده روی پست‌های یک فروم وب استفاده کرده است. در این مدل مقدار پرپلکسیتی مدل زبانی در سطح سه کلمه به عنوان یک ویژگی در بردار ویژگی استفاده شده است.

از انجا که استفاده از n-گرام کاراکتر و n-گرام کلمه همراه با مدل سازی زبانی راه حل به کار گرفته شده در این پژوهش است در جدول ۲-۱ تعدادی از کارهای پیشین که به نوعی از هر یک از آنها استفاده شده است لیست شده است.

۲-۱ کارهای انجام شده بر روی پایگاه‌داده‌های فارسی

در ادامه به عنوان آخرین بخش از این فصل مروری بر کارهای انجام شده بر روی پایگاه‌داده با زبان فارسی در زمینه مورد تحقیق در این پایان‌نامه یعنی شناسایی نویسنده می‌پردازیم.

کار (شاهمیری، 1385) دراین کار پایگاه داده فارسی با چهار نویسنده کاندید تشکیل شـده اسـت. ۵۰ ویژگی گوناگون استخراج شده از اشعار ایرانی در سه دسته فیزیکی، مفهومی و آوایی دسته‌بندی شده است. برای تخصیص از روش‌های درخت تصمیم و شبکه عصبی مصنوعی استفاده شـده اسـت. دقـت گزارش شده در درخت تصمیم تا ۹۴٪ و با شبکه عصبی تا ۹۵,۹٪ درستی اعلام شده، در صورتی کـه دقت انسان تنها ۳۴,۴٪ بوده است .

در کار (فرهمندپور، 1390) از ویژگی‌های واژگانی، نحوی، معنایی و وابسته به کـاربرد اسـتفاده شـده است. الگوریتم رقابت کشورهای استعماری نیز جزء ایده‌ها و روش‌های جدید عنوان شده است کـه بـه علت سرعت همگرایی بالا، به آن پرداخته شده است. همچنین ایشان الگوریتم‌هـای یـادگیری ماشـین مانند دلتا، کا نزدیک‌ترین همسایه، درخت تصمیم، شبکه‌های عصبی، ترکیـب الگـوریتم ژنتیـک و کـا نزدیک‌ترین همسایه، ترکیب الگوریتم رقابت استعماری و کا نزدیـک تـرین همسـایه و LDA را در دو پایگاه داده جمع‌آوری شده با هم مقایسه کـرده اسـت. روش دسـته بنـدی LDA بـا بهتـرین دقـت و ویژگی‌های نحوی، با بیشترین کارایی گزارش شده اسـت. دو پایگـاه داده ایجـاد شـده یکـی در یـک موضوع خاص از ۲۰ دانشجو با ۲۰۰۹ کلمه تهیه شده است و از دید نویسنده به علـت کوتـاه بـودن و محاوره‌ای بودن متون باعث نادقیق شدن پارامترهای اندازه‌گیری شده است. در پایگاه دوم از مجموعـه کتاب‌های هشت نویسنده معاصر استفاده شده است.

در کار (آذین، 1392) از پایگاه داده تشکیل شده از چهار شاعر شعر نو استفاده شده است. روش‌هـای تخصیص مورد استفاده کا نزدیک‌ترین همسایه، ماشین بردار پشتیبان وبیز سـاده اسـت. ویژگـی‌هـای مورد استفاده در سه دسته ویژگی‌های واژگانی، حرفی و نحوی و ادغام این سه گروه ویژگی بوده است و بهترین گزارش در گروه ویژگی‌های نحوی بوده است.

و در نهایت کار رمضانی و همکاران (Ramezani, et al., 2013) که به بررسی ویژگی‌هـای مختلـف در پایگاه داده R40 (در بخش ۳-۲-۳ معرفی شده است) با استفاده بردار ماشین پرداخته است.

چدول ۲-۱: لیستی از کارهای پیشین. کارهایی که در آنها از n-گرام کاراکترها یا n-گرام کلمات و یا مدل‌سازی زبانی استفاده شده است

	عنوان	ویژگی	تخصیص
۱	N-Gram Feature Selection for ۲۰۰۶ Authorship Identification	استفاده از کارکتر n-گرام ها با مفدار n ۳، ۴و ۵ و با تعداد تکرار مشخص در یک بردار ویژگی	بردار ماشین
۲	Local histograms of character n-grams for authorship attribution	کاراکتر n-گرام	استفاده از هیستاگرام
۳	Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data	۱-کارکتر n-گرام ها با مقدار n ۱، ۲ و ۳ به صورت جداگانه استفاده کرده است. ۲-کلمه n-گرام های با مقدار n۱، ۲ و ۳	بردار ماشین و بیز
۵	Searching with Style: Authorship Attribution in Classic Literature	۱- کلمات دستوری ۲-نقش‌های گرامری ۳-گرام نقش‌های گرامری ۴- ترکیب هر سه	KLD
۶	Using Function Words for Authorship Attribution: Bag of-Words vs. Sequential Rules	مقایسه استفاده از تکرار کلمات دستوری در مقابل ویژگی های استخراج شده با روش قوانین پی‌درپی	بردار ماشین
۷	Visualizing Document Authorship Using N-grams and Latent Semantic Indexing	تشکیل ماتریس لفظ–سند با استفاده از کارکتر n-گرام‌ها و SVD	فاصله کسینوسی
۷	Augmenting Naive Bayes Classifiers with Statistical Language Models	استفاده از n-گرام‌ها در سطح کاراکتر و کلمه	بهبود بیز ساده با استفاده از روشهای مدل-سازی زبانی
۹	Authorship Attribution of Short Messages Using Multimodal Features.	استفاده از n-گرام‌ها در سطح کاراکتربا مقدار n از ۲ تا ۶	استفاده از بیز ساده
۱۰	Time-aware authorship attribution for short text streams	کاراکتر n-گرام	مدل سازی زبانی
۱۱	Combining naive Bayes and n-gram language models for text classification	کاراکتر n-گرام	مدل سازی زبانی

۱۲	Authorship Attribution of Micro-Messages	۱-سه بردار n-گرام کاراکترها+n-گرام کلمات+ الگوی استخراج شده از توالی با شروع و پایان با کلمات پرتکرار و توالی از کلمات در بین آنها. ۲-دو بردار n-گرام کاراکترها + الگوی استخراج شده ۳-n-گرام کاراکترها+n-گرام کلمات ۴-n-گرام کاراکترها	بردار ماشین
۱۳	The effect of author set size and data size in authorship attribution	۱-n-گرام کلمات ۲-n-گرام کلمات دستوری ۳-n-گرام کاراکترها ۴-n-گرام ریشه کلمات ۵-n-گرام نقش دستوری کلمات	TiMBL (روشی بر اساس کا نزدیک‌ترین همسایه)
۱۴	Authorship Attribution in Greek Tweets Using Author's Multilevel N-Gram Profiles	تلفیق n-گرام کلمات با n-گرام کاراکترها	متدی با ایده جداسازی کلاس ها با حاشیه بین صفحات کلاس

۲-۲ جمع بندی

در این فصل به دسته بندی ویژگی‌ها در دسته‌های لغوی، کاراکتری، نحوی، معنایی و وابسته به کاربرد پرداخته شد و به چند نمونه از کارهای انجام شده اشاره شد. مزیت ویژگی‌ها در دسته لغوی سادگی و بی‌نیازی از ابزاراهای پردازش زبان طبیعی عنوان شد و این که این مزیت آنها را تبدیل به ویژگی مستقل از زبان می‌کند همچنین وجود کارهای زیادی که با استفاده از کاراکتر n-گرام‌ها انجام شده است و نتایج خوبی گزارش شده نشان از مفید بودن آنها دارد .-مزیت آنها در مستقل از زبان بودن، مشخص کردن اختلافات جزئی در سبک و تحمل پذیری خوب در برابر نویز شمرده شد. از طرفی مشکل تعیین مقدار n به صورت کارا در استفاده از آنها مطرح شد. ایده ویژگی‌های ساختاری و نحوی در استفاده ناخودآگاه نویسنده‌گان از الگوهای نحوی بیان شد. همچنین گفته شد در ویژگی‌های معنایی توجه روی معنا و نقش معنایی کلمات است. و در نهایت از ویژگی‌های بسته به کاربرد به

عنوان استفاده از ویژگی‌های مختص حوزه‌های مختلف مانند سایز و رنگ فونت در فروم‌ها و یا پیام های الکترونیکی یاد شد.

در بخش ۳-۲ ابتدا به معرفی انواع روش‌های تشکیل داده آموزشی در دو دسته مبتنی بر نمونه و مبتنی بر پروفایل پرداخته شده است و سپس به دسته بندی متدهای تخصیص که درپژوهش‌های پیشین به کار گرفته شده، در دسته: متدها بر اساس فاصله، روش‌های فشرده سازی و روش‌های یادگیری ماشین مانند شبکه‌های عصبی، بردار ماشین و درخت تصمیم و همچنین روش‌های احتمالاتی مانند بیز و مدل‌سازی زبانی پرداخته شده است. و در انتها در بخش ۲-۴ مروری بر کارهای پیشین در پایگاه داده با زبان فارسی صورت گرفته است.

۳- فصل سوم: مباحث نظری و روش پیشنهادی

۳-۱ مقدمه

در این فصل به توصیف و تجزیه تحلیل روش‌های استفاده شده برای حل مسئله شناسایی نویسنده در این پایان‌نامه پرداخته شده است. ابتدا با توضیح در رابطه با پایگاه داده شروع شده است سپس به چگونگی حل مسئله تخصیص نویسنده پرداخته شده است و در ادامه ویژگی های مورد استفاده و روش تخصیص توضیح داده شده است. در انتها به معرفی خصوصیاتی از پایگاه داده که در حل مسئله تخصیص نویسنده تاثیر گزار است، پرداخته شده است.

۳-۲ پایگاه داده

از مشخصات یک پایگاه داده خوب آن چنان که در (Stamatatos, 2009) آمده است استفاده از داده-هایی با نوع و موضوع یکسان است و همچنین استفاده از مجموعه نویسنده‌گانی با سطح تحصیلات ،سن، ملیت و جنسیت یکسان است. با رعایت این موارد می‌توان اطمینان حاصل کرد که این عوامل تاثیری در طبقه بندی متن ندارند و طبقه‌بندی صرفا بر اساس نویسنده‌گی صورت می‌گیرد. با این توضیحات سعی شده است تا در دو پایگاه داده WMPR-AA2016-A و WMPR-AA2016-B که توسط نگارنده جمع‌آوری شده از این اصول پیروی کنند. از طرف دیگر وجود پایگاه داده‌هایی با زبان مختلف و اندازه داده‌های آموزشی متفاوت و اندازه داده آزمایشی متفاوت شرایط را برای بررسی روش-های برررسی شده در شرایط متفاوت فراهم کرده است. بحث بیشتر در این زمینه در فصل نتایج انجام خواهد شد.

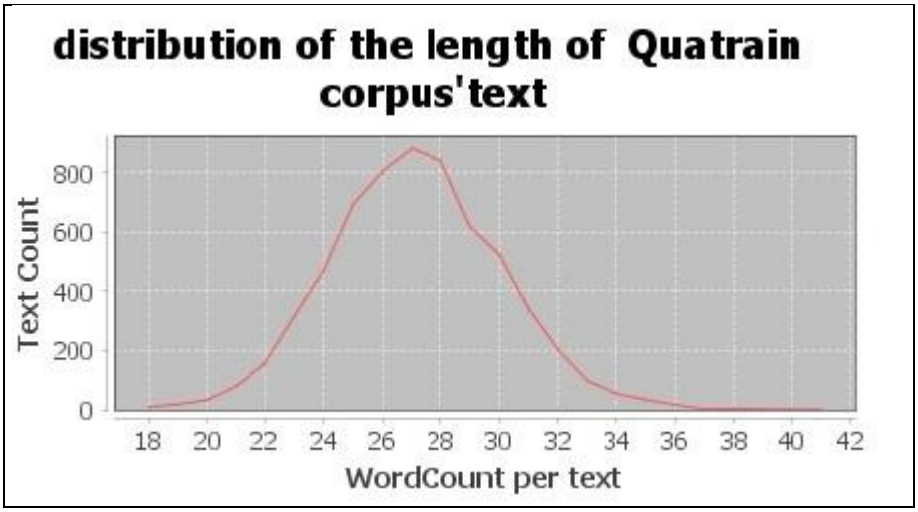
۳-۲-۱ پایگاه داده WMPR-AA2016-A

این پایگاه داده توسط نگارنده تهیه شده است و از مجموعه اشعار رباعی شش شاعر نامدار، ابوسعید ابوالخیر، عطار نیشابوری، انوری، خاقانی، مولوی، و ثنائی ایجاد شده است. در این پایگاه داده هر یک

رباعی به عنوان یک داده آزمایشی در نظر گرفته شده است از این رو با اندازه میـانگین تعـداد کلمـه، ۲۷,۴ و انحراف از میانگین ۲,۶ این پایگاه داده در دسته پایگاه داده‌هـای مـتن کوتـاه محسـوب شـده است. در شکل ۳-۱ توزیع طول مـتن در مـتن‌هـای آزمایشـی در پایگـاه داده WMPR-AA2016-A نشان داده شده است. نحوه چگونگی تخصیص داده آموزشی و داده آزمایشـی در بخـش ۴-۱ توضـیح داده شده است. جزئیات مربوط به اطلاعات پایگاه‌داده از قبیـل تعـداد کلمـه در داده‌هـای آموزشـی و داده‌های آزمایشی و کل اندازه پایگاه داده در جدول ۳-۱ نمایش داده شده است.

جدول ۳-۱: جزئیات پایگاه داده فارسی WMPR-AA2016-A

		نویسنده	(کلمات) کل	(کلمات)داده آموزشی	داده تست
پایگاه داده WMPR-AA2016-A	A0	ابوسعید ابوالخیر	19352	15507	3854
	A1	انوری	12097	9684	2413
	A2	عطار نیشابوری	62248	49808	12440
	A3	خاقانی	9712	7797	1915
	A4	مولوی	53375	42700	10657
	A5	ثنائی	11711	9389	2322



شکل ۳-۱ توزیع طول متن در پایگاه داده WMPR-AA2016-A

۳-۲-۲ پایگاه داده WMPR-AA2016-B

این پایگاه داده توسط نگارنده تهیه شده است و از مجموعه غزلیات هفت شاعر بـزرگ فارسـی زبـان، سعدی، ثنائی، عطار نیشابوری، اوحدی، ساوجی، مولوی و کرمانی تشکیل شـده اسـت. در ایـن پایگـاه

داده هر یک غزل به عنوان داده آزمایشی در نظر گرفته شده است. چگونگی تشـکیل داده آموزشـی و داده آزمایشی در بخش ۴-۲ توضیح داده شده است. جزئیات مربوط به اطلاعات پایگـاه داده از قبیـل تعداد کلمه در داده‌های آموزشی و داده‌های آزمایشی و کل اندازه پایگاه داده در جـدول ۳-۲ نمـایش داده شده است.

جدول ۳-۲: جزئیات پایگاه داده فارسی WMPR-AA2016-B

		نویسنده	کل(کلمات)	داده آموزشی (کلمات)	داده آزمایشی (کلمات)
پایگه داده WMPR-AA2016-B	B0	سعدی	88558	70847	15826
	B1	ثنائی	52512	42116	9797
	B2	عطار نیشابوری	125265	100264	23179
	B3	اوحدی	115893	92816	20563
	B4	ساوجی	49417	39555	8759
	B5	کرمانی	122841	98405	21841
	B6	مولوی	475558	200240	50083

۳-۲-۳ پایگاه داده R40

این پایگاه داده توسط رمضانی و همکارانشان تهیه شده است (Ramezani, et al., 2013) و از مجموعه نوشته‌های ۴۰ نویسنده نامی فارسی زبان مانند جمله جلال آل احمد، بزرگ علوی صادق چوبک و... در زمینه متن‌های ادبی و با سبک نگارشی نثر تهیه شده است. لذا این پایگاه داده در دسته پایگاه داده با تعداد نویسنده کاندید زیاد در نظر گرفته شده است. برای هر نویسنده پنج متن با اندازه بـین ۸۰۰ تـا ۱۰۰۰ کلمه در نظرگرفته شده است. جزئیات مربوط به اطلاعات این پایگاه داده در منبع عنوان شده، آورده نشده است لذا در این جا سعی شده تا اطلاعاتی از قبیل نام نویسـنده‌گـان و تعـداد کلمـات در

متن‌های هر نویسنده جمع آوری شود. اطلاعات جمع آوری شـده در جـدول ۳-۳ نمـایش داده شـده است.

جدول ۳-۳: جزئیات پایگاه داده فارسیR40

		نویسنده	کل (کلمات)		نویسنده	کل (کلمات)
پایگاه داده R40	R0	احمد میر علایی	۶۱۱۴	R20	ناشناس	۵۵۹۳
	R1	محمد علی کاتوزیان	۵۲۱۰	R21	غلام حسین ساعدی	۴۹۳۹
	R2	امیر حسین چهل تن	۳۶۰۵	R22	فروغ فرخزاد	۳۶۵۲
	R3	بزرگ علوی	۶۷۱۰	R23	مهرداد انتظاری	۶۱۴۱
	R4	سراسر حادثه	۱۳۶۰۵	R24	گلی ترقی	۸۲۵۳
	R5	بهنام دیانی	۱۹۶۶۴	R25	نغمه کیانی راد	۵۳۰۳
	R6	مژده کوهی	۶۴۴۶	R26	ناشناس	۵۷۴۶
	R7	ناشناس	۶۲۸۶	R27	محمد علی	۶۳۷۷
	R8	جلال ال احمد	۲۱۲۰۹	R28	ناشناس	۶۰۰۸
	R9	زهرا دلگرمی	۷۴۲۹	R29	جمال زاده	۴۱۴۳
	R10	رضا جولایی	۵۲۲۵	R30	ناشناس	۵۰۶۶
	R11	رضا قاسمی	۳۴۰۷	R31	مهرداد انتظاری	۵۳۱۶
	R12	ناشناس	۵۹۹۹	R32	میهن بهرامی	۷۶۹۶
	R13	زویا پیرزاد	۴۷۶۰	R33	ناشناس	۵۱۳۹
	R14	شهریار مدنی پور	۹۱۶۱	R34	جعفر مدرس صادقی	۳۲۱۷
	R15	صادق چوبک	۸۱۲۵	R35	فریده شجاعی	۸۲۰۴
	R16	صادق هدایت	۳۱۷۲۰	R36	الهه موذنی	۷۲۰۴
	R17	صمد بهرنگی	۷۲۹۰	R37	نیلا	۹۲۰۳
	R18	علی خدایی	۳۸۹۹	R38	پروانه شیخلو	۷۳۱۹
	R19	ناشناس	۱۰۳۷۷	R39	محنا عزیزی	۹۱۳۷

۳-۲-۴ پایگاه‌داده RCV

پایگاه داده RCV (Lewis, et al., 2004) در زبان انگلیسی تهیه شده است. و بـا پنجـاه نویسـنده در دسته پایگاه داده با تعداد نویسنده زیاد در نظر گرفته شده است. متن‌ها در این پایگاه داده به صـورت

نثر است و جنسیت نویسنده‌گان به صورت مرد و زن می‌باشد. برای هر نویسنده ۵۰ متن آموزشی و ۵۰ متن آزمایشی وجود دارد. جزئیات مربوط به اطلاعاتی از قبیل تعداد کلمه در داده‌های آموزشی، داده‌های آزمایشی و کل اندازه پایگاه داده در جدول ۳-4 نمایش داده شده است.

جدول ۳-4: جزئیات پایگاه داده انگلیسی RCV

		نویسنده	کل (کلمات)	داده آموزشی(کلمات)	داده آزمایشی (کلمات)
پایگاه داده RCV	C0	AaronPressman	46885	22499	24386
	C1	AlanCrosby	48463	22579	25884
	C2	AlexanderSmith	56958	29083	27875
	C3	BenjaminKangLim	51733	25517	26216
	C4	BernardHickey	48823	23336	25487
	C5	BradDorfman	47319	24504	22815
	C6	DarrenSchuettler	55459	26210	29249
	C7	DavidLawder	55619	29024	26595
	C8	EdnaFernandes	46412	21906	24506
	C9	EricAuchard	52967	26572	26395
	C10	FumikoFujisaki	54711	27228	27483
	C11	GrahamEarnshaw	50833	25340	25493
	C12	HeatherScoffield	52563	26961	25602
	C13	JaneMacartney	57772	29620	28152
	C14	JanLopatka	44162	22485	21677
	C15	JimGilchrist	33389	17031	16358
	C16	JoeOrtiz	47582	23568	24014
	C`17	JohnMastrini	50278	25294	24984
	C18	JonathanBirt	48259	22347	25912
	C19	JoWinterbottom	52295	26416	25879
	C20	KarlPenhaul	47637	21389	26248
	C21	KeithWeir	47146	24715	22431
	C22	KevinDrawbaugh	46122	23250	22872
	C23	KevinMorrison	53248	26612	26636
	C24	KirstinRidley	57859	26219	31640
	C25	KouroshKarimkhany	46331	22307	24024
	C26	LydiaZajc	33540	17121	16419
	C27	LynneO'Donnell	58255	28920	29335
	C28	LynnleyBrowning	58227	27521	30706
	C29	MarcelMichelson	52591	26034	26557
	C30	MarkBendeich	43106	22392	20714
	C31	MartinWolk	50056	25771	24285
	C32	MatthewBunce	51907	25439	26468
	C33	MichaelConnor	44751	22071	22680

25442	25556	50998	MureDickie	C34	
25950	28587	54537	NickLouth	C35	
23124	24580	47704	PatriciaCommins	C36	
30723	29082	59805	PeterHumphrey	C37	
27067	27199	54266	PierreTran	C38	
24850	24026	48876	RobinSidel	C39	
27252	27000	54252	RogerFillion	C40	
27344	27728	55072	SamuelPerry	C41	
28523	29007	57530	SarahDavison	C42	
26455	24463	50918	ScottHillis	C43	
24727	25228	49955	SimonCowell	C44	
25500	22892	48392	TanEeLyn	C45	
29171	28753	57924	TheresePoletti	C46	
22391	24172	46563	TimFarrand	C47	
25046	25812	50858	ToddNissen	C48	
24421	23147	47568	WilliamKazer	C49	

۳-۳ شناسایی نویسنده

همان‌طورکه در فصل اول توضیح داده شد در فرایند شناسایی نویسنده هدف تخصیص یک نویسنده از میان مجموعه نویسنده‌گان کاندید (شناخته شده)، به متنی است که نویسنده آن مشخص نیست. همچنین نیازمندی‌های این فرایند انتخاب ویژگی‌های سبکی مناسب و انتخاب روش تخصیص کارآمد به منظور انتساب نویسنده درست به متن دیده نشده معرفی شد. از این جهت در ادامه پس از توضیح مرحله پیش پردازش، ویژگی‌ها و روش تخصیص استفاده شده در این پایان‌نامه شرح داده خواهد شد.

۳-۳-۱ پیش‌پردازش

در مرحله پیش‌پردازش تمامی علامت‌های خاص، اعداد و حروف الفبای خارجی از تمامی بانک‌های اطلاعاتی حذف شده و با فضای خالی جایگزین شده‌اند. برای جداسازی جملات درپایگاه داده به ترتیب در ابتدا و انتهای هر جمله از تگ <s> و </s> استفاده شده است. تا در مرحله استخراج ویژگی‌ها ابتدا و انتهای هر جمله مشخص باشد. در پایگاه داده با سبک نوشتاری نظم هر مصراع به

عنوان یک جمله در نظر گرفته شده است و در نثر علامت نقطه و علامت سوال به عنوان مشخص کننده جمله در نظر گرفته شده است.

۳-۳-۲ انتخاب ویژگی

در این پایان نامه n-گرام کاراکتر و n-گرام کلمه به عنوان ویژگی انتخاب شده است. علت این انتخاب همان طورکه در فصل قبل توضیح داده شد این است که n-گرام کارکترها و کلمات در دسته‌ای از ویژگی‌ها قرار دارند که برای استخراج احتیاج به ابزارهای پردازش زبان ندارند. از این‌رو استخراج این ویژگی‌ها هزینه زمان اجرا و خطای ناشی از خطای ابزار پردازش زبان را ندارند. مورد پر اهمیت‌تر استقلال این ویژگی‌ها از زبان است. از این جهت که استفاده از ابزارهای پردازش زبان وابسته به زبان است و در هر زبانی ابزارهای پردازش زبان خود را نیاز دارد. به عنوان نمونه استخراج قوانین نحوی در زبان فارسی با زبان انگلیسی متفاوت است. در ادامه روش استخراج n-گرام‌ها شرح داده شده است.

۳-۳-۲-۱ n-گرام‌ها

در یک تعریف کلی می‌توان گفت، n-گرام دنباله‌ای از n توکن پشت سر هم در یک رشته از توکن‌ها *می‌باشد*. توکن‌ها بنا بر استفاده از هر جنسی اعم از کاراکتر، کلمه، ریشه‌کلمات، نقش دستوری کلمات و ... می‌توانند باشند. استخراج n-گرام‌ها را به صورت مراحل زیر می‌توان توضیح داد:

- برای استخراج n-گرام‌ها متن به صورت توالی از توکن‌ها یا به عبارت دیگر توالی از کاراکترها یا کلمات یا ریشه‌کلمات یا نقش دستوری کلمات و.... در می‌آید.
- مرحله بعد انتخاب طول n-گرام یا همان مقدار n است.
- و بعد از آن یک پنجره با طول n در نظر گرفته می‌شود.
- به اندازه n-1 کاراکتر فضای خالی برای پوشش تمام توکن‌های متن به انتهای متن اضافه می‌گردد.

- حرکت پنجره از اولین توکن متن شروع شده و جابجایی آن توکن به توکن تا آخـرین تـوکن متن ادامه می‌یابد. در هر بار حرکت پنجره آنچه در داخل پنجره قرار گرفته است، n-گرام بـا اندازه n می‌باشد. (شکل 3-2)

تعداد n-گرام‌های تولید شده مستقل از مقدار n برابر با تعداد توکن‌ها در متن بدون در نظر گرفتن n-1 توکن فاصله اضافه شده، است.

متن : یک صبح روز یکشنبه ماه تیر هوای شهر برلین تیره و خفه کننده بود.

۱- تولید توالی از توکن‌ها در شکل کاراکترها

ی،ک،–،ص،ب،ح،–،ر،و،ز،–ی،ک،ش،ن،ب،ه،–،م،ا،ه،–،ت،ی،ر،–،ه،و،ا،ی،–،ش،ه،ر،–،ب،ر،ل،ی،ن،–
،ت،ی،ر،ه،و،خ،ف،ه،ک،ن،ن،د،ه،–،ب،و،د،

۲- تولید انگرام با اندازه مشخص

حرکت پنجره با اندازه n =2

ی،ک،–،ص،ب،ح،–،ر،و،ز،–ی،ک،ش،ن،ب،ه،–،م،ا،ه،–،ت،ی،ر،–،ه،و،ا،ی،–،ش،ه،ر،–
،ب،ر،ل،ی،ن،–،ت،ی،ر،ه،و،خ،ف،ه،ک،ن،ن،د،ه،–،ب،و،د،

انگرام‌های تولید شده	اندازه
یک، ک–؛ –ص؛ صب؛ بح؛ ح–؛ –ر؛ رو؛ وز؛ ز–؛ –ی؛ یک؛ کش؛ شن؛ نب؛ به؛ ه–؛ –م؛ ما؛ اه؛ ه–؛ –ت؛ تی؛ یر؛ ر–؛ –ه؛ هو؛ و....	**۲-گرام**
یک–؛ ک–ص؛ –صب؛ صبح؛بح–؛ ح–ر–؛ –رو؛ روز؛ وز–؛ ز–ی؛ –یک؛ یکش؛ کشن؛ شنب؛ نبه؛ به–؛ –ما؛ ماه؛ اه–؛ ه–ت؛ –تی؛ تیر؛ و....	**۳-گرام**

شکل۳-۲: تولید n-گرام‌ها در یک متن. در این مثال فاصله به عنوان یک توکن در نظر گرفته شده است و با علامت "–" مشخص شده است. علامت کاما جدا کننده n-گرام‌ها و کاراکترها است.

۳-۳-۳ روش تخصیص

متد تخصیص مورد استفاده در این پایان‌نامه مدل‌سازی زبانی است. علت انتخاب این روش قرار گرفتن آن در دسته روش‌های آماری احتمالاتی است. همچنین به جهت قرار گـرفتن آن در در دسـته روش-های پویا و استفاده از احتمال پیشین باعث باز شدن راهی برای بسط احتمال پیشین و بهبود نتایج در

مدل به دست آمده درکارهای آینده می‌شود. علاوه بر آن با بررسی‌های انجام شده، این روش تاکنون روی پایگاه داده با زبان فارسی برای حل مسئله تشخیص نویسنده استفاده نشده است. در ادامه به توضیح این روش پرداخته شده است.

۳-۳-۴ مدل سازی زبانی در شناسایی نویسنده

در مدل‌سازی زبانی هدف پیدا کردن احتمال رخداد توالی از کلمات در یک مجموعه از جملات زبان است. به عبارت دیگر سنجیدن اینکه با توجه به داده‌های آموزشی که از یک زبان در دست است با چه فراوانی، توالی از کلمات می‌تواند توسط مدل زبانی داده‌های آموزشی تولید شود.

استفاده اولیه از مدل سازی زبانی در پردازش سیگنال بوده است و دامنه استفاده از آن تا حوزه‌های بایوانفورماتیک، یادگیری ماشین و بازیابی اطلاعات نیز گسترش پیدا کرده است. در ترجمه ماشینی آماری، یکی از اجزا اصلی سیستم ترجمه، استفاده از مدل‌سازی زبانی به این منظور است که ترجمه یک عبارت چقدر در زبان مورد هدف، عبارتی کاربردی و متعارفی است (Koehn, 2010) . در بازیابی اطلاعات ایده اصلی در استفاده از مدل‌سازی زبانی این است که یک سند در صورتی کاندید خوبی برای تطبیق پیدا کردن با کوئری مورد جستجو است که سند با احتمال بالایی بتواند کوئری را تولید کند. (D.Manning, et al., 2009)

پایه بیشتر تحقیقات و راه‌کارهای ارائه شده برای حل مسئله شناسایی نویسنده بر این اساس است که هر نویسنده از الگوهای نوشتاری مخصوص به خود در نوشتار استفاده می‌نماید که با عنوان "اثر انگشت مولف" ازآن یاد می‌شود و توضیح آن هم می‌تواند این باشد که هر شخصی زبان را خودش و با تجربیات و برداشت‌های شخصی خودش فرا می‌گیرد (Joula, 2008). از این مسئله می‌توان استفاده کرد و استفاده از مدل سازی زبانی در حل مسئله شناسایی نویسنده را این‌گونه توضیح داد که با جمع‌آوری مجموعه نوشته‌های هر نویسنده در مجموعه نویسنده‌گان کاندید و الحاق تمام متن‌های جمع‌آوری شده برای هر نویسنده در یک متن بزرگتر، برای هر نویسنده کاندید الگویی از زبان آن

نویسنده ایجاد شود. سپس با به‌کارگیری مدل‌سازی زبانی به دسته‌بندی متن دیده نشده بـا نویسـنده نامشخص پرداخته شود.

- محاسبه مدل سازی زبانی

همان‌طورکه گفته شد در مدل سازی زبانی هدف پیدا کردن احتمال یک کلمه بـر اسـاس n-1 کلمـه قبل از آن و یا احتمال پیدا کردن یک جمله است. ارتباط این دو احتمال به این صورت است کـه اگـر فرض شود جمله U از کلمات $u_1...u_n$ تشکیل شده است آنگاه طبق قانون زنجیره‌ای داریم :

$$P(U = u_1,\ldots,u_n) = P(u_1)\,P(u_2 \mid u_1)P(u_3|u_1,u_2)..P(u_n \mid u_1,..,u_{n-1}) \qquad ۱-۳$$

رابطه ۳-۱ ارتباط بین احتمال جملات با احتمال کلمات را نشان می‌دهد. از محاسبه احتمال کلمه به شرط وقوع کلمات قبلی به احتمال جمله رسیده است. برای محاسبه هر یک از احتمـالات شـرطی در رابطه ۳-۱ داریم :

$$P(u_n \mid u_{n-1}) = \frac{\text{Count}\left(u_1,\ldots,u_{n-1}u_n\right)}{Count\left(u_1,\ldots,u_{n-1}\right)} \qquad ۲-۳$$

در رابطه ۳-۲ صورت از شمارش تعداد عبارات حاصل از کنار هم قرار دادن کلمات u_1 تا u_n و مخرج از شمارش تعداد عبارات حاصل از کنار هم قرار دادن کلمات u_1 تا u_{n-1} در متن مدل، محاسبه مـی-گردد. مشکلی که در استفاده از رابطه ۳-۱ و قانون زنجیره‌ای وجود دارد این است کـه اگـر تمـامی ارتباط بین کلمات در نظر گرفته شود در محاسبه هر یک از احتمالات شرطی در رابطه ۳-۲ احتیاج به یک مجموعه متن بسیار بزرگ است در غیر این صورت بیشتر احتمالات صـفر خواهـد شـد و در نتیجه آن احتمال جمله U نیز صفر خواهد شد. از این رو در مدل سازی زبانی با اسـتفاده از قـانون مارکوف ارتباط بین کلمات محدود می‌شود و این محدودیت در مدل سازی زبانی سـاده بـا در نظـر نگرفتن ارتباط بین کلمات یا مستقل فرض کردن کلمات از یکدیگر منتهی می‌شـود. در نتیجـه آن رابطه ۳-۱ به صورت زیر در می‌آید (Manning & Schiitze, 2000)(Jurafsky & Martin, 2006) .

$$P(U = u_1,\ldots,u_n) = P(u_1)\,P(u_2)\ldots P(u_n) \qquad ۳-۳$$

- شناسایی نویسنده با مدل سازی زبانی ساده

اگر فرض شود مجموعه $A=\{a_1,a_2,\ldots a_m\}$، مجموعه نویسندگان کاندید هستند هدف پیدا کردن مناسب‌ترین نویسنده برای متن دیده نشده U (متن با نویسنده نامشخص) است. با توضیحاتی داده شده به راحتی می‌توان متوجه شد که مدل سازی زبانی در گروه روش‌های بر مبنای پروفایل قرار دارد (تعریف روش‌های بر مبنای پروفایل در بخش ۳-۲-۱ آورده شده است). بنابراین به منظور تشکیل پروفایل هر نویسنده ai، در مرحله اول تمامی متن‌های آموزشی مربوط به نویسنده a_i در مجموعه A، در یک متن الحاق شده و متن حاصل را d_i می‌نامیم. حاصل مرحله اول تشکیل مجموعه آموزشی $D = \{d_1, d_2 ,\ldots d_m\}$ خواهد بود. که در آن هر d_i پروفایل یا داده آموزشی یا داده مدل نویسنده ai را تشکیل می‌دهد. هدف پیدا کردن اندیس نویسنده i است به طوری که:

$$di = \text{argmax}\{P(di \mid U\} \; di \in D \qquad ۴-۳$$

در آن d_i متن آموزشی مربوط به نویسنده a_i و U متن مورد سوال (متن با نویسنده نامشخص) است. طبق قانون بیز رابطه ۳-۴ می‌تواند به صورت زیر نوشته شود:

$$di = \text{argmax}\{P(di)P(U \mid di)/P(U)\} \; di \in D \qquad ۵-۳$$

در رابطه ۳-۵ از آنجاکه P(U) یا احتمال evidence مستقل از اندیس i است و برای تمامی داده‌های آموزشی مقدار یکسانی دارد می تواند از رابطه حذف شود و $P(d_i)$ یا احتمال پیشین برای تمامی داده-های آموزشی یکسان در نظر گرفته شده در نتیجه رابطه ۳-۵ به صورت رابطه زیر در می‌آید:

$$di = \text{argmax}\{P(U \mid d_i)\} \; di \in D \qquad ۶-۳$$

در رابطه ۳-۶، $P(U|d_i)$ یا احتمال بیشترین شباهت، احتمال تولید متن مورد سوال U است که توسط مدل زبانی تشکیل شده روی داده آموزشی d_i تولید شده است. برای محاسبه احتمال $P(U|d_i)$ طبق

قانون زنجیره ای در رابطه ۳-۱ و قانون مارکوف در رابطه ۳-۲، احتمال هر جمله در متن U از حاصل-ضرب احتمال هر یک از کلمات u_1 تا u_n در مدل زبانی داده آموزشی d_i به دست می‌آید. در متن‌های بزرگ ضرب احتمال کلمات در یکدیگر ممکن است خطای میز شناور[1] را ایجاد نماید. برای جلوگیری از این خطا از لگاریتم استفاده می‌شود و در نهایت خواهیم داشت:

$$i = \text{argmax}_{d_i \in D}\{\sum \log P(u_j | LM_{d_i}) = \sum_{uj \in U} \log \frac{t_{uj,d_i}}{Ld_i}\} \qquad ۷-۳$$

که در این رابطه ui کلمات در متن مورد سوال U است. LM_{di} مدل زبانی ساخته شده روی داده آموزشی d_i است. و $\frac{t_{uj,di}}{Ld_i}$ احتمال هر کلمه u_j روی مدل زبانی ساخته شده در داده آموزشی d_i است که از تقسیم تعداد کلمات u_j در متن آموزشی d_i تقسیم بر کل کلمات متن آموزشی d_i به دست می-آید.

۳-۴ روش پیشنهادی

روش پیشنهادی با ایجاد دو تغییر در مدل سازی زبانی ساده ایجاد شده است و از این پس از آن با نام مدل‌سازی زبانی تغییر یافته یاد شده است. در ادامه به توضیح تغییرات داده شده می‌پردازیم.

۳-۴-۱ استفاده از ضریب IDF به جای حذف کلمات پرتکرار

در استفاده از کلمات به عنوان ویژگی یکی از مشکلات، رخداد کلماتی است که به طور عام و مشترک استفاده می‌شوند و با نام کلمات پرتکرار[2] شناخته شده‌اند. این کلمات به خاطر مشترک بودن و با تکرار بالا در بین بیشتر نوشته‌ها ارزش جداکنندگی زیادی به عنوان یک ویژگی ندارند و در در اکثر روش‌ها به علت نداشتن اثر منفی در روش تخصیص، حذف می‌شوند. روش حذف آنها استفاده از

[1] Floating Point Under Low
[2] Stop Words

لیستی از پرتکرارترین کلمات درمیان متون پیکره و انتخاب و حذف این کلمات به عنوان کلمات پرتکرار است. مشکلاتی که در استفاده از لیست برای حذف کلمات پرتکرار وجود دارد این است که:

- تضمینی بر اینکه آیا هیچ یک از کلمات قرار گرفته در لیست در دسته کلمات با قدرت جداکنندگی نیستند، وجود ندارد.
- آیا تمام کلمات پرتکرار در یک لیست ایستا وجود دارند؟
- استفاده از لیست ایستا روش را به زبان وابسته می‌کند.

یکی از راهکارها، که به نظر می‌رسد بتواند تا حدودی مشکلات مطرح شده را حل نماید استفاده از یک روش وزن‌دهی به کلمات است. تا با استفاده از آن، کلمات با قدرت جدا کنندگی بیشتر وزن بیشتری نسبت به کلمات پرتکرار به دست آورند و اثر بیشتری در تعیین نویسنده داشته باشند. یکی از این روش های وزن‌دهی استفاده از ضریب IDF[1] است.

$$IDF = \frac{N}{DF_u} \qquad ۸-۳$$

در رابطه ۳-۸، N تعداد متن‌های آموزشی و DF_u تعداد متن‌های آموزشی که کلمه u در آنها تکرار شده است. در مدل‌سازی زبانی تغییر یافته به جای استفاده از لیست کلمات پرتکرار، و حذف کلمات لیست از داده‌های آمورشی، هر یک از کلماتی که به عنوان یک ویژگی انتخاب می‌شوند در مقدار ضریب IDF ضرب می‌گردد. و به این ترتیب به جای حذف کلماتی که ممکن است دربرگیرنده خاصیت سبکی نویسنده باشند وزن آنها تغییر داده می‌شود. به عنوان نمونه اگر احتمال رخداد دو کلمه متفاوت u_1 وu_2 در مدل زبانی نویسنده a_1 با یکدیگر برابر باشند و فرض شود u_1 در دسته کلمات پرتکرار قرار دارد که توسط اکثر نویسنده‌گان کاندید استفاده شده است و u_2 در دسته کلماتی قرار داشته باشد که در متن‌های نویسنده‌گان کمتری شرکت دارد، در این صورت مقدار ضریب IDF برای u_2 مقدار عددی بزرگتری نسبت به ضریب IDF برای u_1 خواهد بود و نتیجه ضرب مقادیر IDF به

[1] Inversing Document Frequency

دست آمده در احتمالات برابر u_1 و u_2 مقدار عددی بزرگتری برای u_2 است. به این ترتیب به جای حذف کلمه u_1 که ممکن است دربرگیرنده خاصیت سبکی نویسنده باشد وزن آن تغییر داده می‌شود.

۳-۴-۲ ترکیب ۱-گرام و ۲ -گرام کلمات

همانطور که در قسمت قبل شرح داده شد مدل سازی در سطح دو کلمه به دنبال ارتباطات دوتایی کلمات است و مدل‌سازی در سطح سه کلمه به دنبال ارتباطات سه‌تایی بین کلمات است. از آنجا که با بالا رفتن مقدار n در n-گرام کلمات احتمال وقوع آن در یک متن کمتر می‌شود در مدل سازی در سطح سه کلمه و بالاتر نیاز به داده آموزشی با تعداد کلمه زیاد است تا بتوان برای متن دیده نشده به مقدار احتمال معنا داری دست یافت در غیر این‌صورت برای اکثر n تایی‌ها، در متن دیده نشده مقدار احتمال صفر به دست می‌آید. در مدل سازی در سطح تک کلمه، کلمات به تنهایی و بدون در نظر گرفتن ارتباط بین کلمات و در اصطلاح به صورت کیسه‌ای از کلمات[1]، به عنوان ویژگی استفاده شده است. به نظر می‌رسد استفاده از این ویژگی همیشه نمی‌تواند کافی باشد. به عنوان مثال فرض کنیم دو نویسنده کاندید با دو مجموعه آموزشی داریم و مدل‌سازی زبانی روی هر یک از این داده‌های آموزشی با نام‌های LM_1 و LM_2 پیاده‌سازی شده است. هدف تخصیص نویسنده درست به متن دیده نشده U با محتوای "خانه دوست کجاست؟ در فلق بود که پرسید سوار"، است. اگر تعداد تکرار کلمات در متن U در دو LM_1 و LM_2 یکسان فرض شود در این صورت در تخصیص متن U با دو مدل زبانی LM_1 و LM_2، هر دو امتیاز یکسانی را دریافت خواهند کرد. اما با فرض اینکه که در LM_1 علاوه بر تک کلمات موجود در متن U، ترکیبات دوتایی مانند "خانه دوست" نیز وجود دارد و این ترکیبات در LM_2 وجود ندارد و با گسترش نگاه از سطح تک کلمه به دو کلمه در این صورت در رقابت بین دو مدل‌بانی LM_1 مدل مناسب‌تری به نظر خواهد رسید. زیرا با وجود کلمات دوتایی مشترک بین متن U وLM_1، U به مدل زبانی LM_1 نزدیک‌تر است.

[1] Bag Of Words

با توجه به توضیحات داده شده به نظر می‌رسد که تلفیق دو ویژگی ۱-گرام کلمه و ۲-گرام کلمه، و استفاده آنها به عنوان ویژگی در مدل سازی زبانی می‌تواند منجر به آشکار شدن ویژگی‌های سبکی بیشتری در مسئله تشخیص نویسنده گردد. در مدل سازی زبانی تغییر یافته احتمال ۲-گرام کلمات با احتمال ۲- گرام کلمات تلفیق شده است. با این کار به مجموعه داده‌های آموزشی که علاوه بر کلمات متن دیده نشده دارای کلمات در سطح ۲-گرام از متن دیده نشده هستند امتیاز بیشتری داده شده است. علاوه براین آن‌چنان‌که در قسمت ۳-۴-۱ در استفاده از IDF به جای حذف کلمات پرتکرار توضیح داده شد برای حدف دوتایی‌های پر تکرار مانند دو تایی"من و" که در اکثر متن‌ها دیده می-شود از لگاریتم ضریب IDF در سطح دو گرام کلمه استفاده شده است. در نهایت فرمول مدل‌سازی زبانی تغییر یافته از فرمول مدل‌سازی زبانی ساده در سطح کلمه در رابطه ۳-۷ به شکل زیر تغییر پیدا خواهد کرد:

$$di = \underset{di \in D}{\operatorname{argmax}} \{ \prod_{uj \in U} IDF\ P(u_j | LM_{di})\ IDF\ P(u_j | u_{j-1}, LM_{di}) \} \qquad ۹-۳$$

که در این رابطه $P(u_j | LM_{d_i})$ احتمال تک کلمات در مدل زبانی ساخته شده روی داده آموزشی d_i مربوط به نویسنده کاندید a_i است و برای محاسبه مانند رابطه ۳-۷ از $\frac{t_{uj,di}}{Ldi}$ که احتمال هر کلمه u_j روی مدل زبانی ساخته شده در داده آموزشی di است و از تقسیم تعداد کلمات u_j در متن آموزشی di تقسیم بر کل کلمات متن آموزشی di به دست می‌آید، استفاده شده است. همچنین $P(u_j | u_{j-1}, LM_{d_i})$ احتمال ۲-گرام کلمات در مدل زبانی ساخته شده روی داده آموزشی di مربوط به نویسنده a_i است و برای محاسبه آن مطابق با رابطه ۳-۲ از تقسیم تعداد دوتایی‌های $u_{j-1}\ u_j$ در متن آموزشی d_i تقسیم بر تعداد کلمات u_j در متن آموزشی d_i به دست می‌آید. در نهایت برای تبدیل ضرب به جمع از لگاریتم استفاده شده است.

۳-۵ خصوصیات پایگاه داده

در (Stamatatos, 2009) خصوصیاتی از پایگاه داده را که در میـزان کـارایی روش‌هـای حـل مسـئله نویسنده موثر هستند به صورت زیر معرفی کرده است:

- مقدار داده آموزشی

مقدار داده آموزشی به این موضوع که چه مقدار داده آموزشی بـرای حـل مسـئله نویسـنده در نظـر گرفته شده است اشاره دارد و اینکه کارایی به دست آمده وابسته به چه مقـدار داده آموزشـی اسـت و تغییر مقدار داده آموزشی و کاهش آن چه تاثیری در نتایج دارد. این مسئله از آن جا اهمیت بیشتری پیدا می‌کند که در بعضی از مسائل وابسته به تشخیص نویسنده مانند تشخیص نویسنده در متن‌هـای مجرمانه معمولا داده آموزشی زیادی در دست نیست. البته مقدار داده آموزشی در نظـر گرفتـه شـده می‌تواند به طول داده آزمایشی نیز بستگی داشته باشد؛ زیرا زمانی که داده آزمایشی دارای متن کوتاه است برای استخراج ویژگی‌های جداکننده احتیاج به داده آموزشی بیشتری خواهـد بـود (Luyckx & Daelemans, 2011).

- تعداد نویسنده در مجموعه نویسنده‌گان کاندید

یکی از چالش‌ها در حل مسئله تخصیص نویسنده تعداد نویسنده‌گان در مجموعه نویسنده‌گان کاندیـد است. هر چه میزان تغییرات دقت در مدل با افزایش تعداد نویسنده‌گان کمتر باشد یا به عبارت دیگـر روش استفاده شده، در تعداد نویسنده زیاد کارایی خود را از دست ندهد، نشان‌دهنده پایـداری بیشـتر روش مورد استفاده است (Luyckx & Daelemans, 2011) .

از جمله کارهایی که در این زمینه بررسی انجام داده‌اند می تـوان بـه کـار (Argamon, et al., 2003) اشاره کرد که در نتایج آن گزارش شده است با افزایش تعداد نویسنده گان از ۲ به ۲۰ نویسنده میزان دقت تا ۴۰٪ کاهش پیدا کرده است. در کار (Grieve, 2007) با اعمال ویژگی‌های متفـاوت و گـزارش نتایج برای تعداد نویسنده کاندید در بازه 2 تا ۴۰ نویسنده کمترین میزان تغییرات در تقریبا بازه ۹۶ تا ۶۵ درصد و بیشترین میزان تغییرات با به کارگیری ویژگی‌هایی مانند میانگین اندازه جمـلات در بـازه

۷۰ تا ۶ درصد قرار دارد. در کار (Abbasi & Chen, 2008) با استفاده از تکنیک Writeprints که بر مبنای تبدیل karhunen-koeve می‌باشد دقت ۹۲ تا ۸۳ و ۹۶ تا ۹۱ را برای تعداد نویسنده ۲۰ تا ۱۰۰ نمایش داده است.

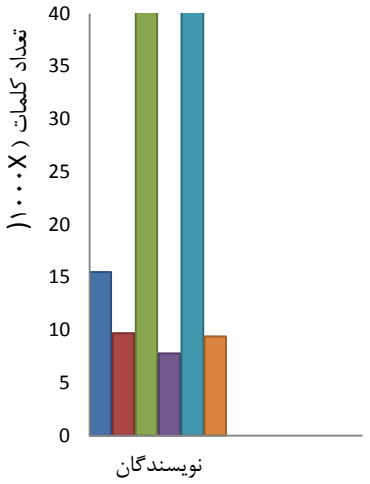
- متعادل بودن داده آموزشی

توزیع داده‌های آموزشی به صورت یکسان یا متفاوت در بین نویسنده‌گان کاندید با نام داده آموزشی متعادل و داده آموزشی نامتعادل نامیده شده است.

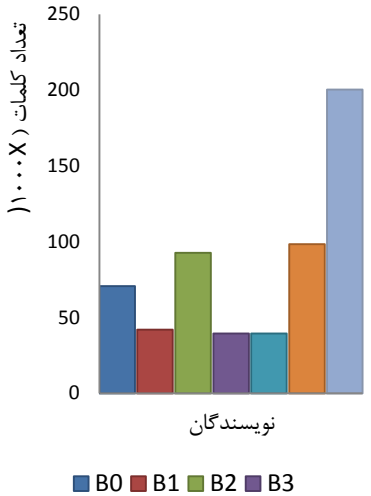
در بیشتر روش‌ها احتمال پیشین در بین تمام نویسنده‌گان کاندید یکسان فرض شده است اما در حالتی که توزیع داده‌ها در بین داده‌های آموزشی نویسنده‌گان کاندید یکسان نباشد، باعث از بین رفتن این برابری احتمال پیشین می‌شود. (Alexander Yun-chung Liu, 2004) . راه حل ارائه شده برای حل این مسئله استفاده از اضافه کردن داده آموزشی[1] و یا کم کردن داده آموزشی[2] است و راه حل‌های متفاوتی برای این منظور ارائه شده است. در ساده‌ترین حالت برای اضافه کردن داده آموزشی، داده‌های آموزشی موجود تکرار می‌شوند و برای کم کردن داده آموزشی، از داده‌های آموزشی به صورت تصادفی کم می شود ((Stamatatos, 2008))(Alexander Yun-chung Liu, 2004) , .

در ادامه توزیع داده‌های آموزشی در بین نویسندگان کاندید چهار پایگاه داده استفاده شده در این پایان‌نامه نمایش داده شده است. در بین این پایگاه داده ها، پایگاه داده WMPR-AA2016-A و WMPR-AA2016-B دو پایگاه داده نامتعادل و پایگاه داده RCV نسبتا متعادل و پایگاه داده R40 توسط رمضانی و همکارانشان متعادل معرفی شده است.

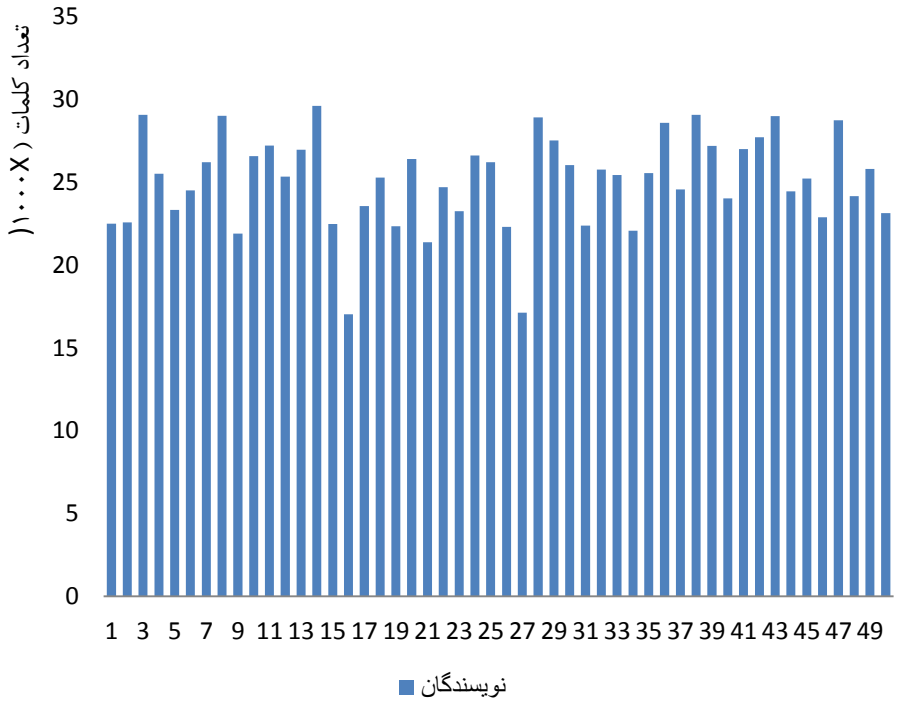
[1] Up Sampling
[2] Down Sampling



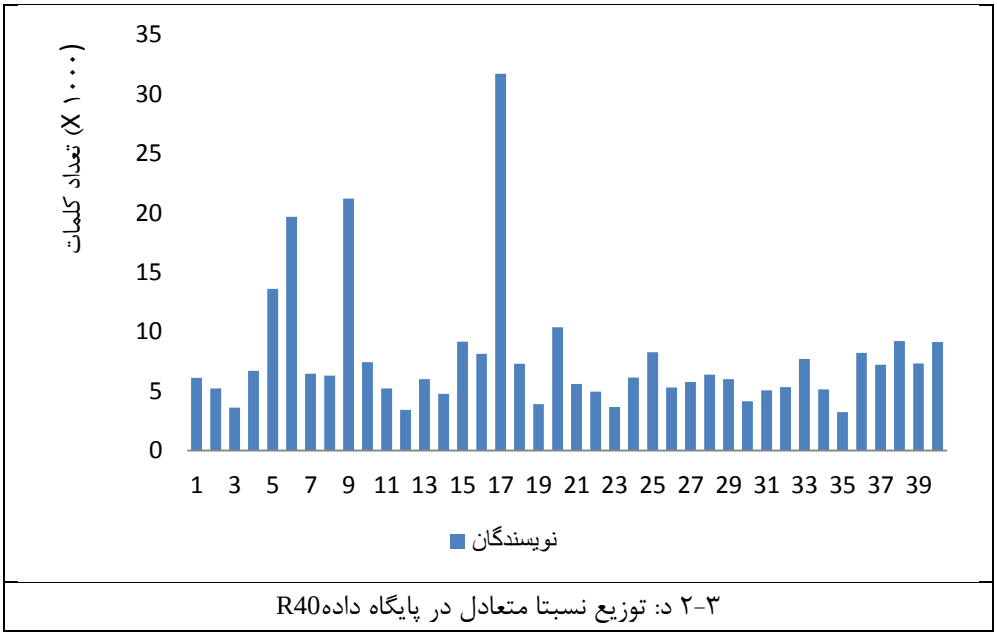
۳-۲ ب: توزیع نامتعادل در پایگاه داده WMPR-AA2016-A



۳-۲ الف: توزیع نامتعادل در پایگاه داده WMPR-AA2016-B



۳-۲ ج: توزیع نسبتا متعادل در پایگاه داده RCV



۳-۲ د: توزیع نسبتا متعادل در پایگاه داده‌R40

شکل ۳-۳: توزیع داده‌های آموزشی در بین نویسندگان کاندید.

در چهار پایگاه داده WMPR-AA2016-A، WMPR-AA2016-B، RCV و R40

۳-۶ جمع بندی

در این فصل به معرفی چهار پایگاه داده استفاده شده در این پایان نامه پرداخته شد. سه پایگاه داده در زبان فارسی و پایگاه داده چهارم در زبان انگلیسی است. پایگاه داده‌ها در سبک نگارش و اندازه داده داده آموزشی، آزمایشی و تعداد نویسنده کاندید متفاوت است. پس از آن در بخش ۳-۳ مراحل مورد نیاز برای حل مسئله شناسایی نویسنده اعم از پیش پردازش و انتخاب ویژگی و روش تخصص عنوان شده است و n-گرام کاراکترها و کلمات و مدل سازی زبانی به عنوان ویژگی و روش تخصیص انتخابی شرح داده شده است. در بخش ۳-۴ یک روش پیشنهادی با ترکیب ۱-گرام کلمات با ۲-گرام کلمات و استفاده از مقدار IDF به عنوان وزن دهی احتمال n-گرام‌ها بر اساس میزان تکرار در داده‌های آموزشی هر نویسنده، معرفی شده است.

و در نهایت در بخش ۳-۵ به معرفی خصوصیاتی از پایگاه داده که در نتایج ارزیابی در حل مسئله تخصیص نویسنده موثر هستند پرداخته شده است. و از آنها با عنوان مقدار داده آموزشی، تعداد نویسندگان کاندید و توزیع داده های آموزشی بین نویسندگان کاندید نام برده شده است.

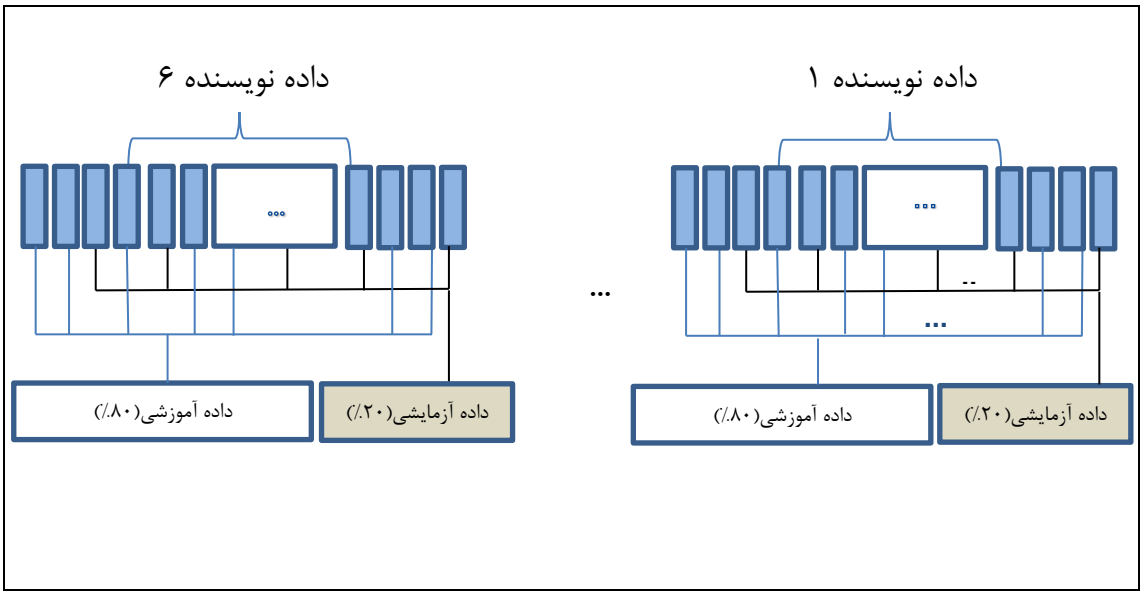
۴- فصل چهارم : نتایج

۴-۱ مقدمه

در فصل قبل به معرفی روش مدل سازی زبانی بـا اسـتفاده از n-گـرام کلمـات و n-گـرام کـاراکترهـا پرداخته شد. همچنین ترکیب 1-گرام تک کلمات با 2-گرام کلمات و استفاده از مقدار IDF به عنـوان وزن دهی احتمال n-گرامها (بر اساس میزان تکرار در دادههای آموزشی هر نویسنده) در مدل سازی زبانی با نام مدل سازی تغییر یافته، پیشنهاد شد. در این فصل به ارزیابی حل مسئله نویسنده بـا روش مدلسازی زبانی ساده با استفاده از n-گرام کلمات با مقـدار n=۱،۲و۳ و n-گـرام کاراکترهـا بـا تغییـر مقدار n از n=2 تا n=6 همچنین مدل سازی زبانی تغییر یافته در چهار پایگـاه داده معرفـی شـده در بخش ۳-۲ پرداخته شده است. سپس به بررسی اثر خصوصیات پایگاه داده که در بخـش۳-۵ توضـیح داده شد، پرداخته خواهد شد.

۴-۱ آزمایشات در پایگاه داده WMPR-AA2016-A

در آزمایشات صورت گرفته در این پایگاه داده شش نویسنده در رقابت تعیین نویسنده ناشـناس بـرای دادههای آزمایشی شرکت دارند. برای تشکیل دادههای آزمایشـی و دادههـای تسـت از مجموعـه کـل متون موجود برای هر نویسنده ۲۰٪ داده هـا بـرای داده آزمایشـی و ۸۰٪ درصـد دادههـا بـرای داده آموزشی انتخاب شده است. انتخاب دادهها برای قرارگیـری در دو مجموعـه آزمایشـی و آموزشـی بـه صورت تصادفی انجام شده است. شکل ۴-۱ نمایی کلی از این تقسیم بندی را نمایش میدهد.



شکل ۴-۱: تشکیل داده آموزشی و مجموعه داده آزمایشی در پایگاه داده WMPR-AA2016-A

۴-۱-۱ آزمایش با مدل سازی زبانی و n-گرام کلمات

نتایج حاصل از اجرای مدل‌سازی زبانی ساده با n-گرام کلمات با مقدار ۳ و n=۱،۲ در جدول ۴-۱ نمایش داده شده است. برای روشن‌تر شدن ارزیابی، علاوه بر میانگین دقت، نتایج ارزیابی به تفکیک هر نویسنده نیز محاسبه شده است.

جدول ۴-۱: نتایج ارزیابی روش مدل‌سازی زبانی با n-گرام کلمات در پایگاه داده WMPR-AA2016-A

	نویسنده	*Precision*	*Recall*	*F-measure*
مدل سازی زبانی با ۱-گرام کلمات	A0	44.0	38.73	41.19
	A1	37.19	51.72	43.26
	A2	74.69	79.78	77.15
	A3	33.68	47.05	39.26
	A4	74.91	53.26	62.26
	A5	28.92	43.20	34.65
میانگین دقت	60.27			
مدل سازی زبانی با ۲-گرام کلمات	A0	31.34	29.57	30.43
	A1	25.0	27.58	26.22
	A2	68.97	81.09	74.54
	A3	30.66	33.82	32.16
	A4	70.88	56.28	62.74
	A5	16.0	14.81	15.38
میانگین دقت	56.			

مدل سازی زبانی با ۳-گرام کلمات	A0	15.68	39.71	22.48
	A1	20.0	14.94	17.10
	A2	64.74	64.17	64.45
	A3	24.39	14.70	18.34
	A4	56.62	35.42	43.58
	A5	16.41	13.58	14.86
میانگین دقت	42.52			

۴-۱-۲ آزمایش با مدل سازی زبانی تغییر یافته

نتایج ارزیابی مدل‌سازی زبانی تغییر یافته در جدول ۳-۲ نمایش داده شده است. نتایج روی این پایگاه داده یک دقت ۵,۴۴ درصدی را نسبت به بهترین نتیجه در روش مدل سازی زبانی با n-گرام کلمات، را نشان می‌دهد. نکته قابل توجه در این است که با وجود این که به علت ناکافی بودن داده‌های آموزشی مدل سازی زبانی ساده در سطح 2-گرام کلمه موفق نبوده است اما تلفیق ۱-گرام کلمات با 2-گرام باعث بهبود نتایج شده است.

جدول ۴-۲: نتایج ارزیابی روش مدل سازی زبانی تغییر یافته در پایگاه داده WMPR-AA2016-A

نویسنده	مدل سازی زبانی تغییر یافته		
	Precision	*Recall*	*F-measure*
A0	46.49	38.73	41.40
A1	41.57	42.52	42.04
A2	73.71	88.13	80.27
A3	47.61	44.11	45.80
A4	74.57	66.33	70.21
A5	35.82	29.62	32.43
میانگین دقت	65.71%		

۴-۱-۳ آزمایش با مدل سازی و n-گرام کاراکترها

این آزمایش با استفاده از مدل سازی زبانی و n-گرام کاراکترها انجام شده است وآزمایش با افزایش مقدارn از ۱ تا ۶ تکرار شده است. میزان دقت در n=3 به بالاترین دقت در این سری از آزمایش‌ها

رسیده است و بعد از آن با افزایش مقدار n، میزان دقت کاهش پیدا کرده است. نتایج در جدول ۴-۳ نمایش داده شده است.

جدول ۴-۳: نتایج ارزیابی روش مدل سازی زبانی با n-گرام کاراکترها در پایگاه داده WMPR-AA2016-A

n	میانگین دقت در مدل سازی زبانی ساده و n-گرام کاراکترها
2	50.69
3	62.95
4	60.27
5	60.27
6	59.13

۴-۱-۴ بررسی نتایج

در میان آزمایش‌های انجام شده در پایگاه داده WMPR-AA2016-A، بهترین نتیجه در مدل سازی زبانی تغییر یافته و کمترین نتیجه با مدل سازی زبانی در ۳-گرام کلمات حاصل شده است. نتایج حاصل از آزمایش‌ها در این پایگاه داده نشان دهنده آن است که دقت مدل سازی زبانی ساده در ۱-گرام کلمات بهتر از مدل سازی زبانی ساده در سطح ۲-گرام کلمات و حتی ۳-گرام کلمات است. این می تواند به این علت باشد که با توجه به طول کوتاه داده‌های آزمایشی در این پایگاه داده داده‌های آموزشی، برای فراهم کردن اطلاعات زبانی در سطح ۲-گرام و ۳- گرام به اندازه کافی فراهم نبوده است. در مدل سازی زبانی تغییر یافته نتایج از ۱-گرام کلمات نیز بهتر شده است. علت این بهبود می تواند استفاده همزمان از دو ویژگی ۱-گرام کلمات ۲-گرام کلمات همراه با ضریب IDF باشد که به نوعی توانسته است که کاستی‌های مربوط به کمبود اطلاعات زبانی را در هریک از این ویژگی‌ها به تنهایی جبران کند. در مدل سازی در سطح کاراکتر نتایج به دست آمده با تغییر مقدار n تغییر می-کند به طوری که در$n=3$ بهترین نتیجه به دست آمده است و این نتیجه از نتایج مدل‌سازی با ۱-گرام کلمات، ۲-گرام و ۳-گرام کلمات بهتر بوده است و در رتبه بندی آزمایش‌ها در این پایگاه داده در رتبه دوم قرار گرفته است. این نتیجه می‌تواند تایید کننده قدرت بالای کاراکتر n-گرام‌ها در ثبت اختلافات

جزئی سبکی باشد. در جدول ۴-۴ نتایج حاصل از آزمایشات در پایگاه داده WMPR-AA2016-A بـه ترتیب میانگین دقت، رتبه بندی شده است.

جدول ۴-۴: رتبه بندی نتایج ازمایشات با مدل سازی زبانی در پایگاه داده WMPR-AA2016-A

رتبه	روش	دقت
۱	مدل سازی زبانی تغییر یافته	**۶۵,۷۱**
۲	مدل سازی زبانی با ۳- گرام کاراکترها	**62.95**
3	مدل سازی زبانی با ۱- گرام کلمات	60.27%
4	مدل سازی زبانی ۴و۵- گرام کاراکترها	60.27
5	مدل سازی زبانی با ۶- گرام کاراکترها	59.13
6	مدل سازی زبانی با ۲-گرام کلمات	56
7	مدل سازی زبانی با ۲- گرام کاراکترها	50.69
8	مدل سازی زبانی ساده با ۳-گرام کلمات	42.52

۴-۲ : آزمایشات در پایگاه داده WMPR-AA2016-B

در آزمایشات صورت گرفته در این پایگاه داده هفت نویسنده در رقابت تعیین نویسنده ناشـناس بـرای داده‌های آزمایشی شرکت دارند. برای تشکیل داده‌های آموزشی و داده‌های آزمایشی از مجموعـه کـل متون موجود برای هر نویسنده ۲۰٪ داده‌ها بـرای داده آزمایشـی و ۸۰٪ داده‌هـا بـرای داده آموزشـی انتخاب شده است. انتخاب داده‌ها برای قرار گیـری در دو مجموعـه آزمایشـی و آموزشـی بـه صـورت تصادفی انجام شده است. شکل ۴-۱ نمایی از این تقسیم بندی را نمایش می‌دهد.

۴-۲-۱ آزمایش با مدل سازی زبانی و n-گرام کلمات

نتایج حاصل از اجرای مدل‌سازی زبانی ساده با n-گرام کلمات با مقدار ۳ و n=۲،۱ در جدول ۴-۵ نمایش داده شده است. برای روشن‌تر شدن ارزیابی، علاوه بر میانگین دقت، نتایج ارزیابی به تفکیک هر نویسنده نیز محاسبه شده است.

جدول ۴-۵: نتایج ارزیابی روش مدل سازی زبانی با n-گرام کلمات در پایگاه داده WMPR-AA2016-B

	نویسنده	*Precision*	*Recall*	*F-measure*
مدل سازی زبانی با ۱-گرام کلمات	B0	90.08	86.50	88.25
	B1	78.57	84.61	81.48
	B2	94.01	87.30	90.53
	B3	84.88	84.39	84.63
	B4	64.58	82.66	72.51
	B5	93.22	88.70	90.90
	B6	97.55	91.93	94.65
میانگین دقت		87.95		
مدل سازی زبانی با ۲-گرام کلمات	B0	79.56	86.50	82.88
	B1	86.20	54.94	67.11
	B2	79.56	90.24	84.57
	B3	82.5	76.30	79.27
	B4	85.45	62.66	72.30
	B5	84.13	94.08	88.83
	B6	91.62	94.52	93.04
میانگین دقت		۸۵،۱۱		
مدل سازی زبانی با ۳-گرام کلمات	B0	۶۵،۶۵	۷۹،۲۶	71.48
	B1	70.73	31.86	43.93
	B2	65.65	79.26	71.82
	B3	65.85	62.42	64.09
	B4	64.86	32.0	42.85
	B5	78.21	84.94	81.44
	B6	80.67	89.04	84.65
میانگین دقت		73.32		

۴-۲-۲ آزمایش با مدل سازی زبانی تغییر یافته

مشابه پایگـاه داده WMPR-AA2016-A در آزمایشـات در پایگـاه داده WMPR-AA2016-B نیـز, روش مدل سازی زبانی تغییر یافته یک بهبود ۳,۷ درصدی را نسبت به استفاده n-گرام کلمات نشـان می‌دهد. نتایج ارزیابی در جدول ۴-۶ نمایش داده شده است.

جدول ۴-۶: نتایج ارزیابی روش مدل سازی زبانی تغییر یافته در پایگاه داده WMPR-AA2016-B

نویسنده	مدل سازی زبانی تغییر یافته		
	Precision	*Recall*	*F-measure*
B0	92.05	88.05	90.24
B1	90.90	76.92	83.33
B2	83.97	92.68	88.11
B3	85.86	91.32	88.51
B4	86.56	77.33	81.69
B5	93.75	96.77	95.23
B6	96.48	94.81	95.63
میانگین دقت	**91.56**		

۴-۲-۱ آزمایش با مدل سازی زبانی و n-گرام کاراکترها

این سری از آزمایش‌ها با استفاده از مدل سازی زبانی و n-گرام کاراکترها انجام شده است وآزمایش بـا افزایش مقدارn از ۱تا ۶ تکرار شده اسـت. میـزان دقـت در n=3 بـه بـالاترین دقـت در ایـن سـری از آزمایش‌ها رسیده است و بعد از آن با افزایش مقدار n، میزان دقت کاهش پیدا کـرده اسـت. نتـایج در جدول ۴-۷ نمایش داده شده است.

جدول ۴-۷: نتایج ارزیابی روش مدل سازی زبانی با n-گرام کاراکترها در پایگاه داده WMPR-AA2016-B

n	میانگین دقت در مدل سازی زبانی ساده و n-گرام کاراکترها
2	71.25
3	85.88
4	76.16
5	65.54
6	63.33

۴-۲-۲ بررسی نتایج

در میان آزمایش‌های انجام شده در پایگاه داده WMPR-AA2016- B بهترین نتیجه در مدل سازی زبانی تغییر یافته و کمترین در مدل سازی زبانی با ۶-گرام کاراکترها به دست آمده است. نتایج حاصل از اجرای مدل‌سازی زبانی ساده با n-گرام کلمات مشابه پایگاه داده WMPR-AA2016-A نمایانگر بهتر بودن نتایج در مدل سازی زبانی با ۱-گرام کلمات نسبت به مدل سازی زبانی با ۲-گرام و ۳-گرام کلمات است. در مدل سازی با n-گرام کاراکترها نتایج به دست آمده با تغییر مقدار n تغییر می‌کند به طوری‌که درn=6 کمترین میزان دقت و در n=3 بهترین نتیجه به دست آمده است. نکته قابل توجه حاصل شدن بهترین نتیجه در هر دو پایگاه داده WMPR-AA2016-A و WMPR-AA2016-B با n=3 در گروه آزمایشات با n-گرام کاراکترها است. این می‌تواند تا حدودی نشان دهنده وابستگی مقدار n به زبان و سبک نوشتاری باشد.

به طور کلی در تمام آزمایش‌ها در این پایگاه داده نتایج نسبت به پایگاه داده WMPR-AA2016-A رشدی بالاتر از ۲۰ درصد را داشته است. این می تواند به علت دو عامل بیشتر شدن داده‌های آموزشی و همچنین بزرگتر شدن داده‌های آزمایشی به معنای تعداد کلمات در هر متن داده آزمایشی باشد.

در جدول ۴-۸ نتایج حاصل از آزمایشات انجام شده در پایگاه داده WMPR-AA2016-B به ترتیب میانگین دقت، رتبه بندی شده است.

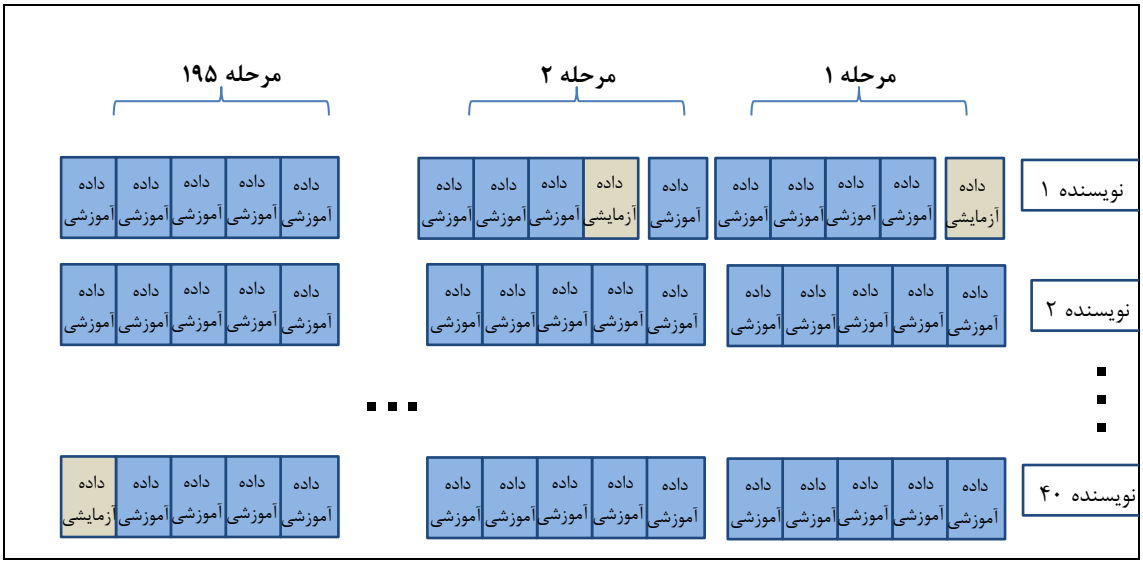
جدول ۴-۸: رتبه بندی نتایج آزمایشات با مدل سازی زبانی در پایگاه داده WMPR-AA2016-B

رتبه	روش	دقت
۱	مدل سازی زبانی تغییر یافته	**91.56**
2	مدل سازی زبانی با ۱-گرام کلمات	**87.11**
۳	مدل سازی زبانی با ۳-گرام کاراکترها	**85.88**
۴	مدل سازی زبانی با ۲-گرام کلمات	85.11
۵	مدل سازی زبانی با ۴-گرام کاراکترها	76.16
6	مدل سازی زبانی با ۳-گرام کلمات	73.32
7	مدل سازی زبانی با ۲-گرام کاراکترها	71.25
8	مدل سازی زبانی با ۵-گرام کاراکترها	65.54
9	مدل سازی زبانی با ۶-گرام کاراکترها	63.33

۴-۳ آزمایشات در پایگاه داده R40

این پایگاه داده که توضیح آن در بخش ۳-۲-۳ داده شد دارای۴۰ نویسنده است و برای هر نویسنده بین ۴ تا پنج متن با طول کلمه‌ای بین ۸۰۰ تا ۹۰۰ کلمه موجود است. و در بین سه پایگاه داده در زبان فارسی جزء پایگاه داده با تعداد نویسنده کاندید زیاد قرار می‌گیرد. به علت کم بودن متن‌های موجود برای هر نویسنده از روش اعتبار سنجی متقابل با کنار گذاشتن یک متن[1] استفاده شده است به این صورت که برای هر نویسنده یکی از متن‌های موجود به عنوان داده آزمایشی در نظر گرفته شده است و متن‌های باقی مانده برای تشکیل داده آموزشی برای هر نویسنده الحاق می‌گردد. به این ترتیب برای هر نویسنده بین چهار تا پنج مرحله وجود دارد و این مرحله به تعداد نویسنده‌گان تکرار می‌شود در مجموع در این پایگاه داده با ۴۰ نویسنده ، ۱۹۵مرحله تشکیل شده است. میانگین میزان دقت در هر مرحله به عنوان دقت روش عنوان شده است. شکل ۴-۲ تشکیل داده آموزشی و مجموعه داده آزمایشی در پایگاه داده R40 را نمایش می‌دهد.

[1] Leave-one-out Cross Validation



شکل ۴-۲: تشکیل داده آموزشی و مجموعه داده آزمایشی در پایگاه داده R40

۴-۳-۱ آزمایش با مدل سازی زبانی و n-گرام کلمات

نتایج حاصل از اجرای مدل‌سازی زبانی ساده با n-گرام کلمات با مقدار ۳ و n=۲،۱ در جدول ۴-۹ نمایش داده شده است.

جدول ۴-۹: نتایج ارزیابی روش مدل سازی زبانی با n-گرام کلمات در پایگاه داده R40

	نویسنده	*میانگین دقت*
مدل سازی زبانی با ۱-گرام کلمات	R0-R38	100
	R39	۸۰
میانگین دقت		۹۹,۴
مدل سازی زبانی با ۲-گرام کلمات	R0-R۸	100
	R9	۶۰
	R10	100
	R11	80
	R12	50
	R13-R39	100
میانگین دقت		97.43
مدل سازی زبانی با ۳-گرام کلمات	R0-R1	100
	R2	80
	R3-R4	100
	R5	80
	R6	100
	R7	80

	B6-R10	100
	R11	80
	R12	50
	R13	80
	R14-R15	100
	R16	80
	R17-R26	100
	R27	80
	R28-R29	100
	R30	75
	R31	100
	R32	80
	R33-R37	100
	R38	40
	R39	100
میانگین دقت		92.82

۴-۳-۲ آزمایش با مدل سازی زبانی تغییر یافته

نتایج حاصل از مدل سازی زبانی تغییر یافته در جدول ۴-۱۰ نمایش داده شده است.

جدول ۴-۱۰: نتایج ارزیابی روش مدل سازی زبانی تغییر یاقته در پایگاه داده R40

مدل سازی زبانی تغییر یافته	نویسنده
%۱۰۰	R0-R39
%۱۰۰	میانگین دقت

۴-۳-۳ آزمایش با مدل سازی و n-گرام کاراکترها

در این آزمایش مدل سازی زبانی با n-گرام کاراکترها انجام شده است و آزمایش با افزایش مقدار n از ۱تا 7 تکرار شده است. میزان دقت درn=2 به بالاترین دقت در این سری از آزمایش‌ها رسیده است و بعد از آن با افزایش مقدار n ، میزان دقت کاهش پیدا کرده است. نتایج در جدول 4-11 نمایش داده شده است.

جدول ۴-۱۱: نتایج ارزیابی روش مدل سازی زبانی و n-گرام کاراکترها در پایگاه داده R40

n	میانگین دقت
2	**97.43**
3	93.33
4	93.84
5	95.38
6	97.30
7	96.41

۴-۳-۴ بررسی نتایج

همانند دو پایگاه داده قبل، در این پایگاه داده نیز بهترین نتیجه در مدل سازی زبانی تغییر یافته و کمترین در مدل سازی زبانی در سطح کاراکتر با n=3 به دست آمده است.

نتایج در این پایگاه داده بهترین نتیجه در بین تمام نتایج به دست آمده در تمام پایگاه داده‌های استفاده شده را نمایش می‌دهد. که می تواند به علت طول داده آزمایشی مناسب و توزیع نسبتا متعادل داده‌های آموزشی در بین نویسنده‌گان کاندید باشد. همچنین نسبت به پژوهش پیشین (Ramezani, et al., 2013) در این پایگاه داده روش مدل سازی زبانی تغییر یافته نتیجه را بهبود داده است. در جدول ۵-۲ مقایسه بین مدل‌سازی زبانی تغییر یافته و کار (Ramezani, et al., 2013) نشان داده شده است.

نکته قابل توجه در آزمایش‌های n-گرام کاراکترها، حاصل شدن بهتربن نتیجه در n=2 است. مقدار n با وجود ثابت بودن زبان اما تفاوت در سبک نوشتاری و همچنین تعداد نویسنده کاندید تغییر کرده است. رتبه بندی نتایج آزمایشات به ترتیب میانگین دقت در جدول ۴-۱۲ نمایش داده شده است.

جدول ۴-۱۲: رتبه بندی نتایج آزمایشات با مدل سازی زبانی در پایگاه داده R40.

رتبه	نام روش	بهترین دقت
۱	**مدل سازی زبانی تغییر یافته**	**100**
2	مدل سازی زبانی با ۱-گرام کلمات	۹۹,۴
۳	مدل سازی زبانی با ۲-گرام کلمات	97.43
۴	مدل سازی زبانی با ۳-گرام کلمات	92.82
۵	مدل سازی زبانی با ۲-گرام کاراکترها	97.43
۶	مدل سازی زبانی با ۶-گرام کاراکترها	۹۷,۳۰
۷	مدل سازی زبانی با ۷-گرام کاراکترها	96.41
۸	مدل سازی زبانی با ۵ -گرام کاراکترها	95.38
۹	مدل سازی زبانی با ۴-گرام کاراکترها	93.84
9	مدل سازی زبانی با ۳-گرام کاراکترها	93.33

۴-۴ زیر مجموعه با شش نویسنده از پایگاه‌داده RCV

همان‌طورکه در بخش ۳-۵ در رابطه با خصوصیات پایگاه داده توضیح داده شد تعداد نویسنده یکی از عوامل موثر در مسئله تخصیص نویسنده است. از این جهت برای مقایسه بهتر نتایج حاصل از این پایگاه داده در زبان انگلیسی با نتایج دو پایگاه داده WMPR-AA2016-A و WMPR-AA2016-B در زبان فارسی، آزمایشی روی یک مجموعه با شش نویسنده از زیر مجموعه نویسنده‌گان پایگاه داده RCV انجام شده است. برای شرکت تمامی نویسنده‌گان در نتایج این ارزیابی، ارزیابی با ۳۰ تکرار آزمایش در زیر مجموعه‌های شش عضوی از مجموعه نویسنده‌گان ۵۰ عضوی پایگاه داده RCV انجام شده است. سعی شده است انتخاب‌ها به گونه‌ای انجام شود که تمام نویسنده‌گان در ارزیابی شرکت داشته باشند. میانگین این مجموعه سی تایی به عنوان دقت روش در نظر گرفته شده است.

۴-۴-۱ آزمایش با مدل سازی زبانی تغییر یافته و مدل سازی زبانی با n-گرام کلمات

نتایج حاصل از اجرای مدل‌سازی زبانی ساده با n-گرام کلمات با مقـدار ۳ و n=۲،۱ و همچنـین مـدل سازی زبانی تغییر یافته در جدول ۴-۱۳ نمایش داده شده است.

جدول ۴-۱۳: نتایج ارزیابی روش مدل سازی زبانی با n-گرام کلمات و مدل‌سازی زبانی تغییر یافته در زیر مجموعه ۶ عضوی از پایگاه داده RCV

نویسنده	مدل سازی زبانی با ۱-گرام کلمات	مدل سازی زبانی با ۲-گرام کلمات	مدل سازی زبانی با ۳-گرام کلمات	مدل سازی زبانی تغییر یافته
	میانگین دقت			
گروه ۱	۹۶	۹۶,۶۶	۹۶	۹۷
گروه ۲	۹۵	۹۶,۳۳	۹۵	۹۶
گروه ۳	۸۸,۶۶	۸۷,۳۳	۸۸	۸۸,۳۳
گروه۴	۹۸,۶۶	۹۹	۹۷,۳۳	۹۹
گروه۵	۹۶	۹۷,۳۳	۹۶,۳۳	۹۸
گروه۶	۹۶,۳۳	۹۵,۳۳	۹۶	۹۶,۳۳
گروه۷	۸۳	۸۵,۳۳	۸۸,۳۳	۸۶,۳۳
گروه۸	۹۰,۶۶	۹۰,۳۳	۹۲,۶۶	۹۱,۶۶
گروه۹	۹۶	۹۹,۳۳	۹۷,۶۶	۹۸,۶۶
گروه۱۰	۹۴	۹۴	۹۵,۶۶	۹۵,۳۳
گروه۱۱	۹۰,۶۶	۹۱,۶۶	۹۱	۹۱,۶۶
گروه۱۲	۹۴,۳۳	۹۳,۶۶	۹۴	۹۵,۳۳
گروه۱۳	۹۰,۳۳	۹۳	۹۴	۹۳,۳۳
گروه۱۴	۹۱	۸۹	۸۹,۶۶	۹۰
گروه۱۵	۹۵,۳۳	۹۶	۹۴,۳۳	۹۶
گروه۱۶	۹۷,۳۳	۹۷,۶۶	۹۷	۹۸,۶۶
گروه۱۷	۹۳,۳۳	۹۴,۳۳	۹۴,۶۶	۹۴,۳۳
گروه۱۸	۹۲,۶۶	۹۴	۹۱	۹۴,۶۶
گروه۱۹	۸۱,۳۳	۸۵,۶۶	۸۶,۶۶	۸۵,۳۳
20 گروه	۹۷	۹۴,۶۶	۹۶	۹۶
21گرو	۹۱,۳۳	۹۰,۶۶	۹۱,۶۶	۹۱
22گروه	۹۲	۹۲,۳۳	۹۲	۹۲,۶۶
گروه۲۳	۹۰	۸۹,٦٦	۸٦,۳۳	۹۱
گروه۲۴	۷۷,٦٦	۸۲,۳۳	۸٤	۸۱,۳۳
گروه۲۵	۹۲,٦٦	۹٤,٦٦	۹٥,٦٦	۹٤,۳۳

گروه۲۶	۹۵	۹٤	۹٦,۳۳	۹٥,۳۳
گروه۲۷	۸۸	۸۹,٦٦	۹۱	۸۹
گروه۲۸	۹۰,٦٦	۹۰	۹۰,۳۳	۹۰
گروه۲۹	۸۸,۳۳	۹۲,٦٦	۹۲,۳۳	۹۲
گروه۳۰	۹۷,٦٦	۹۷,۳۳	۹۷,۳۳	۹۸
میانگین	92.03	92.8	92.94	**93.22**

۴-۴-۲ بررسی نتایج

نکته قابل توجه در نتایج این سری از آزمایش‌ها در این است که با وجود بهتر شدن نتیجه مدل سازی زبانی در ۳-گرام کلمات نسبت به ۲-گرام و ۱-گرام کلمات، بهترین نتیجه در مدل سازی زبانی تغییر یافته به دست آمده است. البته اختلاف بهبود ایجاد شده کمتر از بهبود ایجاد شده در دو پایگاه داده WMPR-AA2016-A و WMPR-AA2016-B است. نتایج در این پایگاه داده از دو پایگاه داده WMPR-AA2016-A و WMPR-AA2016-B بهتر است. اما باید توجه داشت طول داده آزمایشی در این پایگاه داده بزرگتر است و همچنین پایگاه داده متعادل‌تری است. رتبه‌بندی نتایج آزمایشات بر اساس میانگین دقت به دست آمده در جدول ۴-۱۴ نمایش داده شده است.

جدول ۴-14: رتبه بندی نتایج آزمایشات با مدل سازی زبانی در پایگاه داده RCV

رتبه	نام روش	بهترین دقت
۱	مدل سازی زبانی تغییر یافته	**۹۳,۳۴**
2	مدل سازی زبانی با ۳-گرام کلمات	**۹۲,۹٤**
4	مدل سازی زبانی با ۲-گرام کلمات	۹۲,۸۰
5	مدل سازی زبانی با ۳۱-گرام کلمات	92.03

۴-۵ آزمایشات در پایگاه‌داده RCV

این پایگاه داده در زبان انگلیسی بوده و با ۵۰ نویسنده در دسته پایگاه داده با نویسنده زیاد محسوب می‌گردد. برای داده آزمایشی و آموزشی برای هر نویسنده ۵۰ متن وجود دارد. در روش مدل سازی زبانی همان‌طورکه عنوان شد از آنجا که یک روش بر پایه پروفایل است، برای تشکیل داده آموزش

برای هر نویسنده تمامی متن‌های آموزشی هر نویسنده الحاق می‌گردد و یک متن آموزشی بـرای هـر نویسنده تشکیل می‌شود.

۴-۵-۱ آزمایش با مدل سازی زبانی تغییر یافته و مدل سازی زبانی با n-گرام کلمات

نتایج حاصل از مدل سازی زبانی با ۱- گرام ، ۲- گرام و ۳- گرام کلمات وهمچنین مدل‌سـازی زبـانی تغییر یافته در جدول ۴-۱۵ نمایش داده شده است.

جدول ۴-۱۵: نتایج ارزیابی روش مدل سازی زبانی با n-گرام کلمات و مدل‌سازی زبانی تغییر یافته در پایگاه داده RCV

نویسنده	مدل سازی زبانی با ۱-گرام کلمات	مدل سازی زبانی با ۲-گرام کلمات	مدل سازی زبانی با ۳-گرام کلمات	مدل سازی زبانی تغییر یافته
	میانگین دقت			
C0-c49	67.48	70.	69.92	**70.48**

۴-۵-۲ آزمایش با مدل سازی زبانی و n-گرام کاراکترها

در این آزمایش مدل سازی زبانی در سطح کاراکتر انجام شده است. آزمایش با افزایش مقـدار n از تـا n=6 تکرار شده است. میزان دقت در n=3 به بالاترین دقت در این روش رسیده است و بعد از آن بـا افزایش مقدار n ، میزان دقت کاهش پیدا کرده است. نتایج در جدول ۴-۱۶ نمایش داده شده است.

جدول ۴-۱۶: نتایج ارزیابی روش مدل سازی زبانی با n-گرام کاراکترها در پایگاه داده RCV

n	ازمایش با مدل سازی و n-گرام کاراکترها
2	61.40
۳	**63.96**
4	63
5	63.24

۴-۵-۳ بررسی نتایج

در این پایگاه داده بهترین نتیجه برای مدل سازی تغییر یافته به دست آمده است. اگرچه این نتیجه با نتیجه مدل‌سازی در سطح ۲-گرام کلمه اختلاف زیادی ندارد (.48٪). به طور کلی نتایج به دست آمده در n-گرام کلمات بهتر از حاصل n-گرام کاراکترها است.در مقایسه دو پایگاه داده R40 و پایگاه داده RCVبه عنوان دو پایگاه داده با تعداد نویسنده زیاد و در دو زبان متفاوت، پایگاه داده R40 با بهترین نتیجه ۱۰۰٪، نتیجه بسیار بهتری را نسبت به بهترین نتیجه 70.48 در RCV، نمایش می‌دهد. البته عواملی مانند تعداد نویسنده‌گان بیشتر در پایگاه داده RCV نسبت به پایگاه داده R40، می‌تواند در اختلاف نتایج موثر باشد. رتبه‌بندی نتایج آزمایشات بر اساس میانگین دقت در جدول ۴-17 نمایش داده شده است. در پژوهش (Stamatatos, 2006) از این پایگاه داده استفاده شده است در جدول ۵-۳ مقایسه بین نتایج مدل‌سازی زبانی تغییر یافته با این پژوهش نمایش داده شده است. در این مقایسه مدل‌سازی زبانی تغییر یافته با دقت ۷۰,۴۸ در رتبه ششم قرار گرفته است که با دقت در رتبه اول، به اندازه ۳,۵۶ اختلاف دارد.

جدول ۴-۱۷: رتبه بندی نتایج آزمایشات در پایگاه داده RCV

رتبه	روش	بهترین دقت
۱	مدل سازی زبانی تغییر یافته	70.48
2	مدل سازی زبانی ساده با ۲-گرام کلمات	70
3	مدل سازی زبانی ساده با ۳-گرام کلمات	69.92
4	مدل سازی زبانی ساده با ۱-گرام کلمات	67.48
5	مدل سازی زبانی ساده با ۳-گرام کاراکترها	63.96
6	مدل سازی زبانی ساده با ۲-گرام کاراکترها	61.40
7	مدل سازی زبانی ساده با ۵-گرام کاراکترها	63.24
8	مدل سازی زبانی ساده با ۴-گرام کاراکترها	63

۴-۶ بررسی خصوصیات پایگاه داده

در این قسمت به بررسی خصوصیات مختلف یک پایگاه داده که در بخش ۳-۵ توضیح داده شد و تاثیری که روی نتایج می‌تواند داشته باشد می‌پردازیم. ویژگی‌های مورد بحث تعداد نویسنده‌گان، اندازه داده آموزشی و متعادل بودن یا نامتعادل بودن داده‌های آموزشی در پایگاه داده است.

۴-۶-۱ تعداد نویسنده در مجموعه نویسنده‌گان کاندید

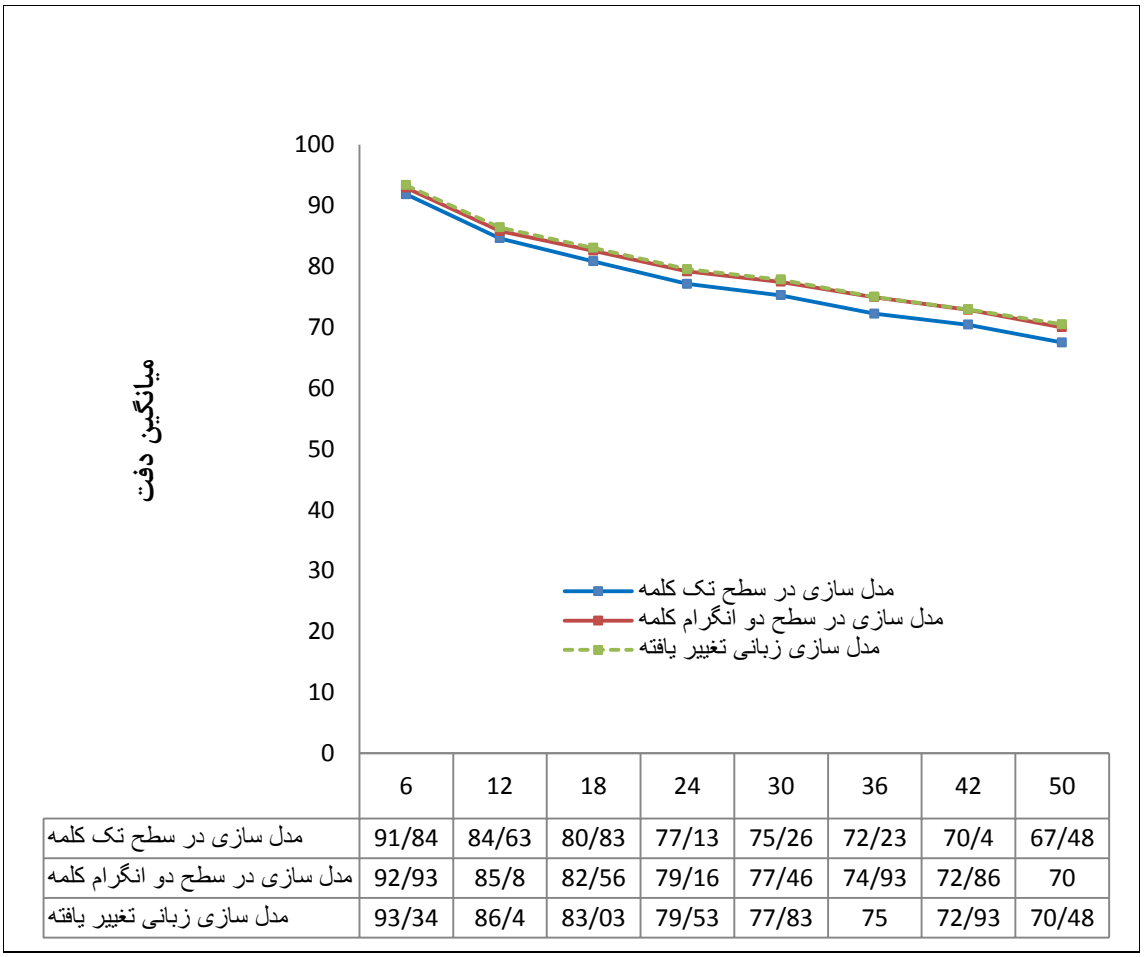
پایگاه داده RCV با ۵۰ نویسنده در دسته پایگاه داده با تعداد نویسنده زیاد قرار می‌گیرد به همین علت برای روشن شدن اثر تعداد نویسنده در میزان دقت در حل مسئله شناسایی نویسنده و بررسی این که ویژگی‌های آزمایش شده چقدر در برابر افزایش تعداد نویسنده مقاومت دارند و نتایج چگونه تغییر پیدا می‌کند؛ آزمایشی به صورتی که شرح داده خواهد شد ترتیب داده شده است.

۴-۶-۱-۱ آماده سازی زیر مجموعه‌ها

از مجموعه نویسنده‌گان پایگاه داده RCV به ترتیب زیر مجموعه هایی با تعداد نویسنده‌گان شش، دوازده، هجده، بیست و چهار، سی، سی وشش، چهل و دو و پنجاه عضوی انتخاب شده است. برای به دست آوردن میزان دقت مدل در هر یک از زیر مجموعه‌ها به صورت تصادفی سی زیر مجموعه متفاوت با تعداد نویسنده‌گان یکسان انتخاب شده است و میانگین دقت در سی زیر مجموعه به عنوان دقت زیر مجموعه عنوان شده است. به عنوان نمونه برای محاسبه میزان دقت زیر مجموعه شش تایی، سی زیر مجموعه ۶تایی از مجموعه۵۰تایی نویسنده‌گان انتخاب شده است و میانگین دقت سی زیر مجموعه به عنوان دقت در مجموعه شش تایی در نظر گرفته شده است.

۴-۶-۱-۲ بررسی نتایج:

نتیجه افرایش تعداد نویسنده گان در شکل ۳-۴ نمایش داده شده است. با افزایش تعداد نویسنده‌گان, میانگین دقت در هر مدل کاهش پیدا می‌کند. در تمامی گروه‌ها میانگین دقت در روش مـدل سـازی زبانی تغییر یافته بالاتر از دو مدل دیگر است.



	6	12	18	24	30	36	42	50
مدل سازی در سطح تک کلمه	91/84	84/63	80/83	77/13	75/26	72/23	70/4	67/48
مدل سازی در سطح دو انگرام کلمه	92/93	85/8	82/56	79/16	77/46	74/93	72/86	70
مدل سازی زبانی تغییر یافته	93/34	86/4	83/03	79/53	77/83	75	72/93	70/48

شکل ۴-۳: اثر افزایش تعداد نویسنده‌گان در میزان دقت

۴-۶-۲ کاهش داده آموزشی

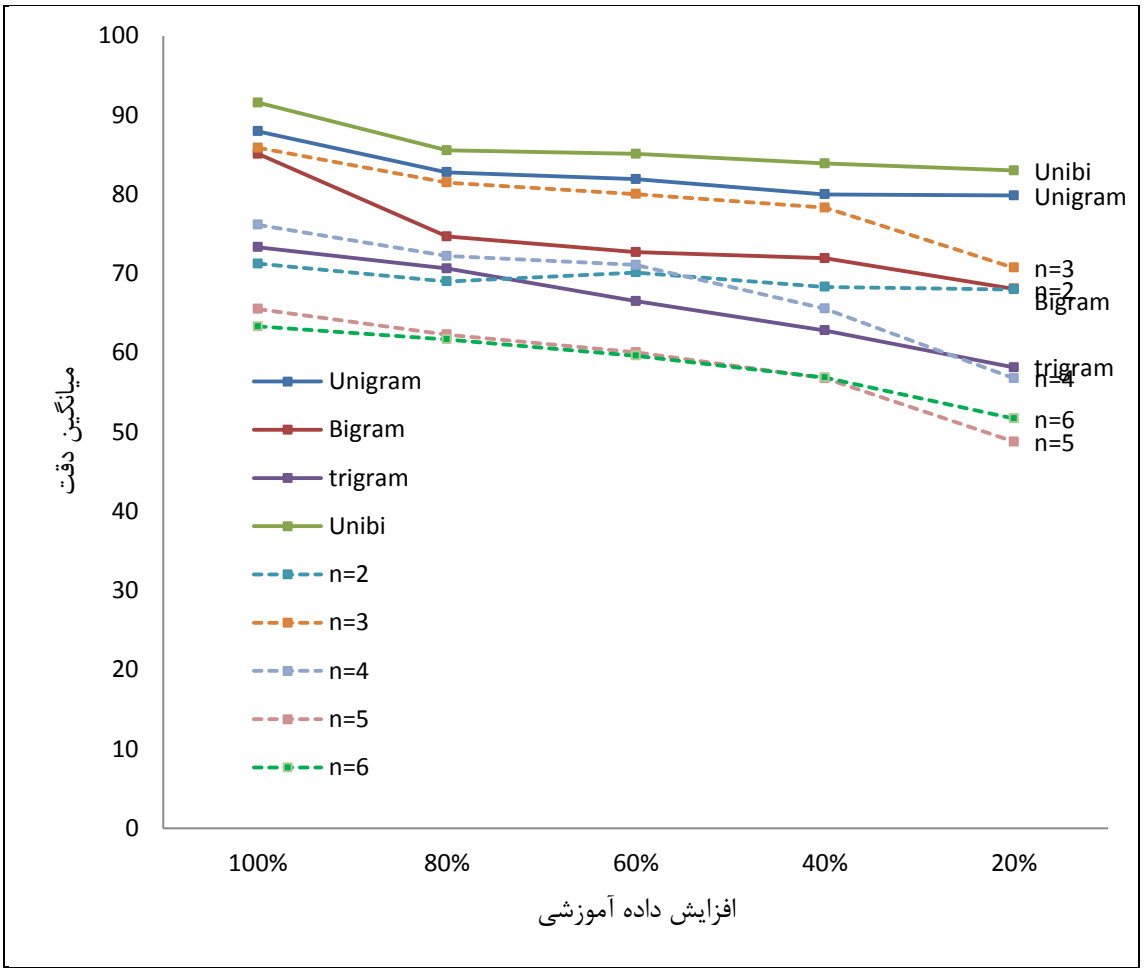
این آزمایش برای بررسی دقت مدل سازی با افزایش داده آموزشی در حل مسئله شناسـایی نویسـنده در نظر گرفته شده است. دو پایگاه داده WMPR-AA2016-A و WMPR-AA2016-B به خاطر داده آموزشی زیاد، فارسی بودن زبان و ویژگی مـتن کوتـاه در پایگـاه داده WMPR-AA2016-A انتخـاب شده‌اند.

۴-۶-۲-۱ آماده سازی داده

داده‌هـای آموزشـی در انـدازه ۲۰، ۴۰، ۶۰، ۸۰، و۱۰۰ درصـد از داده آموزشـی در دو پایگـاه داده WMPR-AA2016-A و WMPR-AA2016-B که در قسـمت ۴-۱ و۴-۲ توضـیح داده شـده اسـت، انتخاب شده‌اند. داده‌های آزمایشی داده‌های معرفـی شـده در قسـمت ۴-۱ و ۴-۲ اسـت و در تمـامی تکرار آزمایش با افزایش داده آموزشی یکسان است.

۴-۶-۲-۲ بررسی نتایج

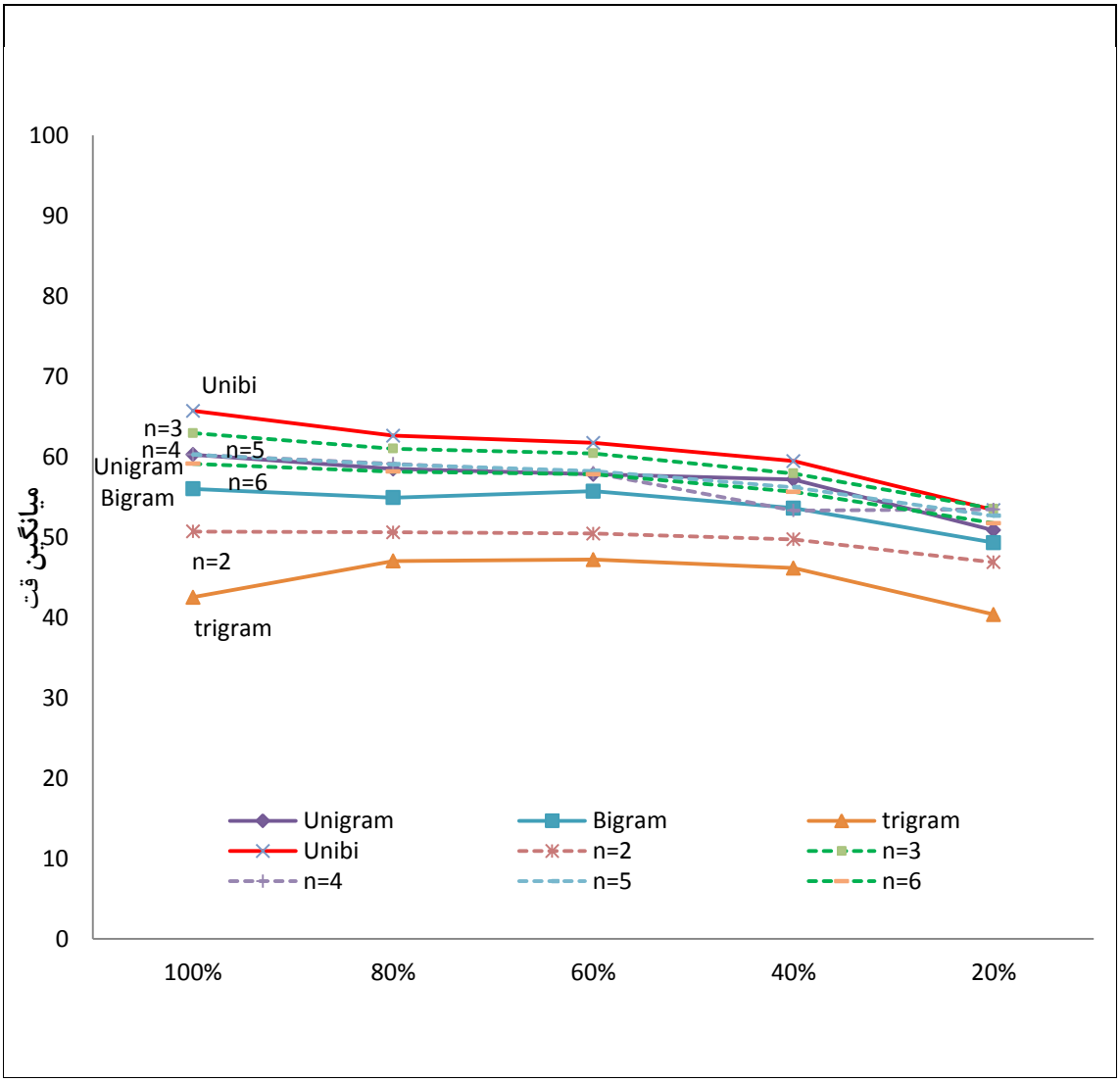
نتایج حاصل از کاهش داده آموزشـی بـا مـدل سـازی‌هـا در دو پایگـاه داده WMPR-AA2016-B و WMPR-AA2016-A به ترتیب در شکل ۴-۴ و ۴-۵، نمایش داده شده است. در هـر دو پایگـاه داده در تمامی اندازه‌های داده آموزشی، مدل سازی تغییر یافته بهتر از سایر مدل سازی‌ها عمل می‌نماید و کاهش داده آموزشی در تمام مدل سازی‌ها، با کاهش میزان دقت همراه است. مقدار دقت در زمانی‌که فقط از ۲٪ داده‌های آموزشی اسـتفاده مـی‌شـود در پایگـاه داده WMPR-AA2016-A از ۶۵,۷۱ بـه ۵۳,۳۷ کاهش پیدا کرده است و در پایگاه داده WMPR-AA2016-B از ۹۱,۵۶ به ۸۳ کـاهش پیـدا کرده است.



شکل ۴-۴: اثر کاهش داده آموزشی در میزان دقت در پایگاه داده WMPR-AA2016-B

جدول۴-۱۸: دقت در نتایج ارزیابی روش مدل سازی زبانی با کاهش داده آموزشی در پایگاه داده WMPR-AA2016-B

نسبت داده آموزشی	مدل‌سازی زبانی با ۱-گرام کلمات	مدل‌سازی زبانی با ۲-گرام کلمات	مدل‌سازی زبانی با ۳-گرام کلمات	*مدل‌سازی زبانی تغییر یافته*	مدل سازی زبانی با ۲-گرام کاراکترها	مدل سازی زبانی با ۳-گرام کاراکترها	مدل سازی زبانی با ۴-گرام کاراکترها	مدل سازی زبانی با ۵-گرام کاراکترها	مدل سازی زبانی با ۶-گرام کاراکترها
٪۲۰	79.86	68.07	58.17	**83**	68	70.74	56.79	48.79	51.72
٪۴۰	80	71.94	62.82	**83.9**	68.33	78.31	65.57	56.79	56.88
٪۶۰	81.92	72.71	66.52	**85.11**	70.13	80.03	71.08	60.06	59.63
٪۸۰	82.78	74.69	۷۰,۶۵	**85.55**	69.01	81.49	72.22	62.3	61.7
٪۱۰۰	87.95	85.11	73.32	**91.56**	71.25	85.88	76.16	65.54	63.33



شکل ۴-۵: اثر کاهش داده آموزشی در میزان دقت WMPR-AA2016-A

جدول ۴-۱۹: دقت در نتایج ارزیابی روش مدل سازی با کاهش داده آموزشی در پایگاه داده WMPR-AA2016-A

نسبت داده آموزشی	مدل‌سازی زبانی با ۱-گرام کلمات	مدل‌سازی زبانی با ۲-گرام کلمات	مدل‌سازی زبانی با ۳-گرام کلمات	*مدل‌سازی زبانی تغییر یافته*	مدل‌سازی زبانی با ۲-گرام کاراکترها	مدل سازی زبانی با ۳-گرام کاراکترها	مدل سازی زبانی با ۴-گرام کاراکترها	مدل‌سازی زبانی با ۵-گرام کاراکترها	مدل‌سازی زبانی با ۶-گرام کاراکترها
٪۲۰	50.85	49.3	40.37	**53.37**	46.87	53.53	53.45	52.64	51.74
٪۴۰	57.18	53.61	46.14	**59.46**	49.71	57.92	53.32	56.21	55.64
٪۶۰	57.83	55.72	47.19	**61.73**	50.44	60.43	58	58.24	57.83
٪۸۰	58.48	54.91	47.03	**62.63**	50.6	61	59.13	59.05	58.16
٪۱۰۰	60.27	56	42.52	**65.71**	50.69	62.95	60.27	60.27	59.13

۴-۶-۳ متعادل کردن داده آموزشی

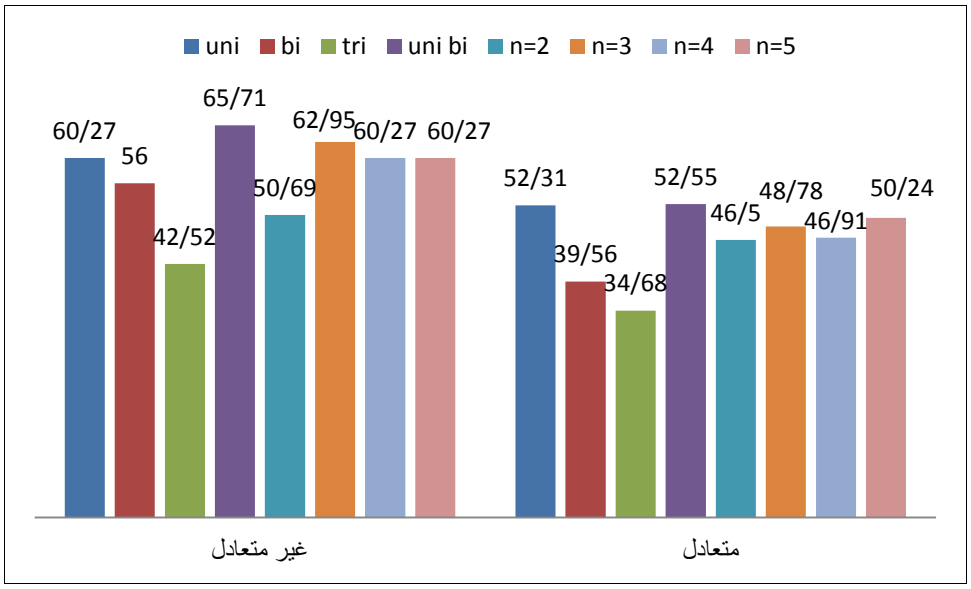
در این بخش برای بررسی اثر متعادل شدن داده آموزشی آزمایشی انجام شده است. و برای این منظور دو پایگاه داده کاملا نامتعادل WMPR-AA2016-A و WMPR-AA2016-B در نظر گرفته شده است.

۴-۶-۳-۱ آماده سازی داده آموزشی

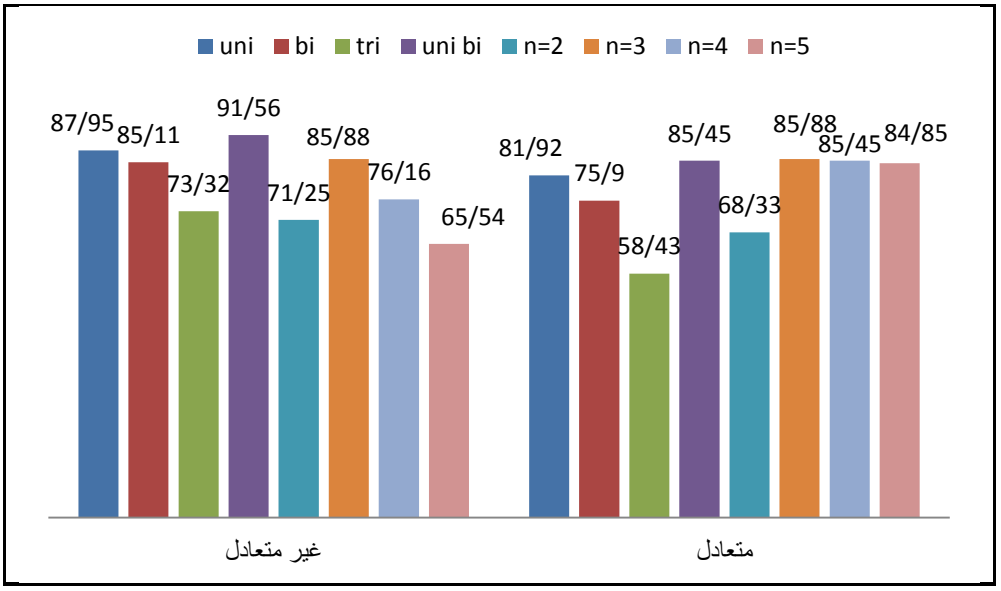
برای متعادل کردن از روش کم کردن نمونه‌ها به صورت تصادفی استفاده شده است. و به این منظور تا یکنواخت شدن توزیع داده‌های آموزشی در نویسنده‌گان کاندید از داده‌های آموزشی به صورت تصادفی داده کم شده است.

۴-۶-۳-۲ بررسی نتایج

نتایج در جدول ۴-۶ و ۴-۷ برای دو پایگاه داده WMPR-AA2016-A وWMPR-AA2016-B نمایش داده شده است. در پایگاه داده WMPR-AA2016-A پس از متعادل سازی در هیچکدام از روش‌ها مدل سازی بهبودی ایجاد نشده است. این می‌تواند از چندین دلیل مانند کارا نبودن روش متعادل سازی انتخابی یا کافی نبودن داده‌های آموزشی در اثر کم شدن داده‌های آموزشی ناشی شده باشد. در پایگاه داده WMPR-AA2016-B بعد از متعادل سازی نتایج در مدل سازی در سطح کاراکتر بهبود داشته است. به طوری که در n=3 با اختلاف کمی، بهتر از مدل تغییر یافته عمل می‌کند. اما در سطح کلمه همچنان داده‌های غیر متعادل بهتر عمل می‌کنند. به طور کلی عملکرد داده‌های متعادل شده در پایگاه داده WMPR-AA2016-B بهتر از پایگاه داده WMPR-AA2016-A عمل کرده است که می‌تواند به علت داده آموزشی بالاتر و طول داده آزمایشی بیشتر باشد.



شکل ۴-۶: تغییرات دقت در مدل سازی در دو حالت متعادل و غیر متعادل در پایگاه داده WMPR-AA2016-A.



شکل ۴-۷: تغییرات دقت در مدل سازی در دو حالت متعادل و غیر متعادل در پایگاه داده WMPR-AA2016-B.

۵- فصل پنجم: نتیجه‌گیری و پیشنهادات

۵-۱ جمع بندی

در فصل اول در بیان مسئله تشخیص نویسنده، به عنوان تخصیص نویسنده برای یک متن دیده نشده بیان شد و ضرورت حل آن با توجه به کاربردهای مختلف آن در حوزه‌های مختلف از جمله تجارت الکترونیکی، شناسایی نویسنده در متن‌های ادبی، شناسایی تروریسم با استفاده از پیام‌ها، تشخیص نویسنده در ایمیل و متن‌های الکترونیکی، تشخیص سرقت ادبی و .. و همچنین گسترش استفاده و انتشار متن با گسترش استفاده از اینترنت و صفحات وب، بیان شد.

در فصل دوم به دسته بندی ویژگی‌ها در دسته‌های لغوی، کاراکتری، نحوی، معنایی و وابسته به کاربرد پرداخته شد. و به چند نمونه از کارهای انجام شده اشاره شد. مزیت ویژگی‌ها در دسته لغوی سادگی و بی‌نیازی به ابزارهای پردازش زبان طبیعی عنوان شد و این که این مزیت آنها را تبدیل به ویژگی مستقل از زبان می‌کند. همچنین بیان شد که وجود کارهای زیادی که با استفاده از کاراکتر n-گرام‌ها انجام شده است و نتایج خوبی گزارش شده نشان از مفید بودن آنها دارد . و مزیت آن در مستقل از زبان بودن، مشخص کردن اختلافات جزئی در سبک و تحمل پذیری خوب در برابر نویز بیان شد. از طرفی مشکل تعین مقدار n به صورت کارا در استفاده از آنها مطرح شد. در رابطه با ایده استفاده از ویژگی‌های ساختاری گفته شد که نویسنده‌گان الگوهای نحوی را به صورت ناخودآگاه استفاده می‌کنند و بنابراین در تمامی نوشته‌های خود الگوهای مشابه‌ای را دنبال خواهند کرد. توجه در ویژگی‌های معنایی روی معنا و نقش معنایی کلمات عنوان شد و در نهایت ویژگی بسته به کاربرد استفاده از ویژگی‌های مختص حوزه‌های مختلف، مانند سایز و رنگ فونت در فروم‌ها و یا پیام‌های الکترونیکی تعریف شد. بعد از معرفی ویژگی‌ها، ابتدا به معرفی انواع روش‌های تشکیل داده آموزشی در دو دسته مبتنی بر نمونه و مبتنی بر پروفایل پرداخته شد و دسته بندی متدهای تخصیص درپژوهش-های پیشین در دسته‌های متدها بر اساس فاصله، روش‌های فشرده سازی و روش‌های یادگیری ماشین مانند شبکه‌های عصبی، بردار ماشین و درخت تصمیم و همچنین روش‌های احتمالاتی مانند بیز و

مدل‌سازی زبانی انجام شد و در انتها مروری بر کارهای پیشین در پایگاه داده با زبـان فارسـی صـورت گرفت.

درفصل سوم به معرفی چهار پایگاه داده استفاده شده در این پایان نامه پرداخته شد. سـه پایگـاه داده در زبان فارسی و پایگاه داده چهارم در زبان انگلیسی معرفی شد. و تفاوت آنها در سبک متفاوت نظم و نثر، اندازه داده آموزشی وآزمایشی و تعداد نویسنده کاندید بیان شد. پس از آن مراحل مورد نیاز برای حل مسئله شناسایی نویسنده پیش پردازش و انتخـاب ویژگـی و متـد تخصـص ذکـر شـد و n-گـرام کاراکترها و کلمات به عنوان ویژگی و مدل سازی زبانی برای متد تخصیص شرح داده شده‌اند. پـس از آن روشی با ترکیب ۱-گرام کلمات با ۲-گرام کلمات و اسـتفاده از مقـدار IDF بـه عنـوان وزن دهـی احتمال n-گرام‌ها (بر اساس میزان تکرار در داده‌های آموزشی هر نویسنده) در مدل سازی زبانی با نام مدل سازی تغییر یافته، پیشنهاد شد. در نهایت در بخش آخر به معرفی خصوصیاتی از پایگاه داده کـه در نتایج ارزیابی در حل مسئله تخصیص نویسنده موثر هستند پرداخته شد. و از آنها به عنـوان مقـدار داده آموزشی، تعداد نویسندگان کاندید و توزیع داده های آموزشی بین نویسنده‌گان کاندید نـام بـرده شد.

در فصل چهارم به ارزیابی حل مسئله نویسنده با روش مدل سازی زبانی ساده با استفاده از ۱-گـرام ، ۲-گرام و ۳-گرام کلمات، n-گرام‌ها کاراکترها با تغییر مقدار n و مدل سازی زبانی تغییر یافته در چهار پایگاه داده پرداخته شد. نکته قابل توجه در n-گرام کاراکترها کسب بهترین نتیجه در دو پایگاه داده با زبان و سبک نگارشی و تعداد نویسنده کاندید مشابه در مقدار n=3 است. جدا از نتایج در مدل سـازی تغییر یافته می‌توان گفت در دو پایگاه داده با نویسنده زیاد n-گرام کلمات و در دو پایگـاه داده دیگـر n-گرام کاراکترها نتایج بهتری داشته است. نتایج بهتر n-گرام کاراکترها در دو پایگاه داده با طول داده آزمایشی کوتاه‌تر، می‌تواند تاییدکننده قدرت n-گرام کاراکترها در تشخیص اختلافات جزئی در سـبک نوشتاری باشد. روش مدل سازی ترکیبی در تمام آزمایشات انجام شده در رتبه اول قرار گرفته اسـت، اگرچه اختلاف آن با روش‌های مدل سازی بر اساس n-گرام کلمه در پایگاه RCV بـا تعـداد نویسـنده

زیاد، کم است اما در دو پایگاه داده با تعداد نویسنده کم، که یکی در دسته داده‌های متن کوتاه نیز قرار می‌گیرد، با بهبود خوبی نسبت به سایر روش‌ها در رتبه اول قرار گرفته است. در پژوهش (Stamatatos, 2006) از این پایگاه داده استفاده شده است در جدول ۵-۳ مقایسه بین نتایج مدل-سازی زبانی تغییر یافته با این پژوهش نمایش داده شده است. در این مقایسه مدل سازی زبانی تغییر یافته با دقت ۷۰,۴۸ در رتبه ششم قرار گرفته است که با دقت در رتبه اول، به اندازه ۳,۵۶ اختلاف دارد. همچنین در پایگاه داده R40 نتیجه با مدل سازی تغییر یافته نسبت به بهترین نتیجه در پژوهش پیشین(Ramezani, et al., 2013) انجام شده بهبود خوبی را نشان می‌دهد. نتایج در این پژوهش با مدل‌سازی زبانی تغییر یافته در جدول ۵-۲ نمایش داده شده است. در این مقایسه در رتبه اول قرار گرفته است.

در بررسی کاهش مقدار داده آموزشی مشخص شد که کاهش داده آموزشی در تمامی پایگاه داده‌ها باعث کاهش دقت در نتایج شده است. مقدار دقت در زمانی‌که فقط از ۲۰٪ داده‌های آموزشی استفاده می‌شود در پایگاه داده WMPR-AA2016-A از ۶۵,۷۱ به ۵۳,۳۷ کاهش پیدا کرده است و در پایگاه داده WMPR-AA2016-B از ۹۱,۵۶ به ۸۳ کاهش پیدا کرده است. از کاهش ۸ درصدی در پایگاه داده WMPR-AA2016-B و همچنین ثابت ماندن نسبی رتبه بندی روش‌ها شاید بتوان این نتیجه را گرفت که روش مدل سازی زبانی با n- گرام کاراکترها و n–گرام کلمات نسبت به کاهش داده آموزشی رفتار نسبتا پایداری را داشته است. در پایگاه داده WMPR-AA2016-A پس از متعادل سازی در هیچکدام از روش‌ها مدل سازی بهبودی ایجاد نشده است. این می‌تواند از دلایلی مانند کارا نبودن روش متعادل سازی انتخابی یا کافی نبودن داده‌های آموزشی در اثر کم شدن داده‌های آموزشی ناشی شده باشد. در پایگاه داده WMPR-AA2016-B بعد از متعادل سازی، نتایج در مدل سازی با n-گرام کاراکترها بهتر شده است، به طوری که حتی در n=3 با اختلاف کم، بهتر از مدل تغییر یافته عمل می‌کند(قبل از متعادل سازی روش مدل تغییر یافته در رتبه بهترین نتیجه را داشته است) اما در سطح کلمه همچنان داده‌های غیر متعادل بهتر عمل می‌کنند. به طور کلی عملکرد داده‌های متعادل

شده در پایگاه داده WMPR-AA2016-B بهتر از پایگاه داده WMPR-AA2016-A عمل کرده است که می‌تواند به علت داده آموزشی بالاتر و طول داده آزمایشی بیشتر باشد.

جدول ۵-۱:میانگین دقت در تمامی آزمایشات انجام شده در چهار پایگاه داده

پایگاه داده WMPR-AA2016-A			پایگاه داده WMPR-AA2016-B		
رتبه	**روش**	**دقت**	**رتبه**	**روش**	**دقت**
۱	**مدل سازی زبانی تغییر یافته**	**۶۵,۷۱**	۱	**مدل سازی زبانی تغییر یافته**	**91.56**
۲	مدل سازی زبانی با ۳- گرام کاراکترها	**62.95**	2	مدل سازی زبانی با ۱-گرام کلمات	**87.11**
3	مدل سازی زبانی با ۱- گرام کلمات	60.27	۳	مدل سازی زبانی با ۳-گرام کاراکترها	**85.88**
4	مدل سازی زبانی ۴و۵- گرام کاراکترها	60.27	۴	مدل سازی زبانی با ۲-گرام کلمات	85.11
5	مدل سازی زبانی با ۶- گرام کاراکترها	59.13	۵	مدل سازی زبانی با ۴-گرام کاراکترها	76.16
6	مدل سازی زبانی با ۲-گرام کلمات	56	6	مدل سازی زبانی با ۳-گرام کلمات	73.32
7	مدل سازی زبانی با ۲- گرام کاراکترها	50.69	7	مدل سازی زبانی با ۲-گرام کاراکترها	71.25
8	مدل سازی زبانی ساده با ۳-گرام کلمات	42.52	8	مدل سازی زبانی با ۵-گرام کاراکترها	65.54
			9	مدل سازی زبانی در سطح کاراکتر با n=6	63.33
پایگاه داده R40			پایگاه داده RCV		
۱	**مدل سازی زبانی تغییر یافته**	**100**	۱	**مدل سازی زبانی تغییر یافته**	**70.48**
2	مدل سازی زبانی با ۱-گرام کلمات	۹۹,۴	2	مدل سازی زبانی ساده با ۲-گرام کلمات	70
۳	مدل سازی زبانی با ۲-گرام کلمات	97.43	3	مدل سازی زبانی ساده با ۳-گرام کلمات	69.92
۴	مدل سازی زبانی با ۳-گرام کلمات	92.82	4	مدل سازی زبانی ساده با ۱-گرام کلمات	67.48
۵	مدل سازی زبانی با ۲-گرام کاراکترها	97.43	5	مدل سازی زبانی ساده با ۳-گرام کاراکترها	63.96
۶	مدل سازی زبانی با ۶-گرام کاراکترها	۹۷,۳۰	6	مدل سازی زبانی ساده با ۲-گرام کاراکترها	61.40
۷	مدل سازی زبانی با ۷-گرام کاراکترها	96.41	7	مدل سازی زبانی ساده با ۵-گرام کاراکترها	63.24
۸	مدل سازی زبانی با ۵ -گرام کاراکترها	95.38	8	مدل سازی زبانی ساده با ۴-گرام کاراکترها	63
۹	مدل سازی زبانی با ۴-گرام کاراکترها	93.84			
پایگاه داده زیر مجموعه ۶ تایی RCV					
۱	**مدل سازی زبانی تغییر یافته**		**۹۳,۳۴**		
2	**مدل سازی زبانی در سطح سه کلمه**		**92.94**		
4	مدل سازی زبانی در سطح دو کلمه		92.94		
5	مدل سازی زبانی ساده در سطح تک کلمه		92.03		

جدول ۵-۲: مقایسه نتایج مدل سازی زبانی تغییر یافته با نتایج در (Ramezani, et al., 2013)

دقت	ویژگی	روش
۱۰۰	**مدل سازی زبانی تغییر یافته**	
.۸۴	۲- گرم کاراکتر	بردار ماشین
.۷۶	تکرار فعل‌ها	بردار ماشین

جدول ۵-۳: مقایسه نتایج مدل سازی زبانی تغییر یافته با نتایج د (Stamatatos, 2006). در (Stamatatos, 2006) با کنار هم قرار دادن کارکتر n-گرام‌ها با مقدار 5 و ۴ ,n=۳ عنوان ویژگی، استفاده شده است. آزمایش سه بار تکرار شده است و در هر بار تعدا ویژگی‌های استخراج شده افزایش پیدا کرده است.روش پیشنهاد شده در این کار با نام PM در جدول مشخص شده است.

دقت	ویژگی	روش
74.04	۳-گرام+۴ گرام+ ۵ گرام=۸۱۷۸ ویژگی	PM
۷۲,۵۶	۳-گرام+۴ گرام+ ۵ گرام=۸۱۷۸ ویژگی	IG[1]
۷۲,۴۸	۳-گرام+۴ گرام+ ۵ گرام=۸۱۷۸ ویژگی	PM
72.16	۳-گرام+۴ گرام+ ۵ گرام=۴۶۹۱ویژگی	IG
72	-گرام+۴ گرام+ ۵ گرام=۲۳۱۴ویژگی	PM
۷۰,۴۸	**مدل سازی زبانی تغییر یافته**	
۶۹,۴	۳-گرام+۴ گرام+ ۵ گرام=۲۳۱۴ویژگی	IG

[1] Informaion Gain

۵-۲ فعالیت‌ها

- ✓ گرد آوری دو مجموعه متن با شش و هفت شاعر از شاعران نامدار به زبان فارسی با نام پایگاه داده WMPR-AA2016-A و پایگاه داده WMPR-AA2016-B (فصل ۴)
- ✓ دسته بندی و بررسی در روش‌های انجام شده در حل مسئله تخصیص نویسنده (فصل ۲)
- ✓ پیشنهاد یک روش برای حل مسئله تخصیص نویسنده (فصل ۳)
- ✓ اعمال روش مدل سازی زبانی در سطح کلمه، ۲-گرام کلمه، ۳-گرام کلمه و در سطح کاراکتر با تغییر بازه متغیر n، بر روی چهار پایگاه داده با سبک نوشتاری، زبان و تعداد نویسنده متفاوت. (فصل ۶)
- ✓ بررسی تاثیر خصوصیات پایگاه داده مانند، مقدار داده آموزشی و تعداد نویسنده کاندید و توزیع داده آموزشی در بین نویسنده‌گان کاندید.

۵-۳ پیشنهادات

در این قسمت پیشنهادات برای کارهای آینده به صورت زیر ارائه می‌گردد.

- ایجاد پایگاه داده‌های استاندارد فارسی

ایجاد پایگاه داده‌های استاندارد در زبان فارسی با تنوع در تعداد نویسنده کاندید، سبک نوشتاری و داده‌های آموزشی و آزمایشی با طول متفاوت باعث ارزیابی دقیق تر و نزدیک‌تر به مسائل واقعی خواهد شد.

- تمرکز در حل مساله نویسنده در حالت مجموعه باز و شناسایی ویژگی‌های فردی

همان‌طورکه در فصل اول توضیح داده شد، شناسایی نویسنده علاوه بر حالت مجموعه بسته در دو دسته مجموعه باز و شناسایی ویژگی‌های فردی نویسنده نیز انجام می‌شود. از این رو گسترش حل مسئله شناسایی نویسنده در این دو دسته در پایگاه داده با زبان فارسی می‌تواند باعث توسعه رهیافت-های تشخیص نویسنده در زبان فارسی شود. همچنین توسعه روش‌های شناسایی نویسنده با به

کارگیری روش‌های احتمالاتی و بهبود نتایج در این زمینه با بسط احتمال پیشین به نظر می‌رسد مـی تواند باعث بهبود نتایج گردد.

- تولید نرم افزار

در فصل اول کاربردهای متفاوت شناسایی نویسنده ذکر شد. توسعه تولید نرم افزارها در این زمینه در پایگاه داده‌های فارسی می‌تواند به کاربردی کردن رهیافت‌ها در زبان فارسی کمک نماید.

منابع

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

Abbasi, A. & Chen, H., 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2), p 7.

Abdallah, E. E. et al., 2013. Simplified features for email authorship identification. *International Journal of Security and Networks*, 8(2), pp. 72-81.

Alexander Yun-chung Liu, B., 2004. *The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets.* Degree of Master of Science in Engineering ,The University of Texas at Austin.

Altheneyan, A. S. a. M. M. E. B., 2014. Naive Bayes classifiers for authorship attribution of Arabic texts. *Journal of King Saud University-Computer and Information Sciences*, 26(4), pp. 473-484.

Alzahrani, et al., 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(2), pp. 133-149.

Argamon, S., Saric, M. & Stein, a. S., 2003. Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results. *ACM SIGKDD*, Volume 19, pp. 475-480.

Azarbonyad, H., Dehghani, M., Marx, M. & Kamps, J., 2015. *Time-aware authorship attribution for short text streams.* New York, NY, USA, Acm, pp. 727-730.

BEKKERMAN, R. & ALLAN, J., 2004. *Using bigrams in text categorization.* Amherst, s.n., pp. 1-2.

Boukhaled, M. A. & Ganascia, J.-G., 2015. Using Function Words for Authorship Attribution:Bag-Of-Words vs. Sequential Rules. *Natural Language Processing and Cognitive Science: Proceedings 2014*, p. 115.

Boutwell, S. R., 2014. *Authorship attribution of short messages using multimodal features*, s.l: Master's thesis, Naval Postgraduate School.

Bozkurt, I., Bilkent Univ., A., Baghoglu, O. & Uyar, E., 2007. Authorship attribution Performance of various features and classification methods. *Computer and information sciences*, pp. 1-5.

Burrows & F, J., 1992. Not unles you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7(2), pp. 91-109.

Chaski, C. E., 2001. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, Volume 8, pp. 1-65.

D.Manning, C., Raghavan, P. & Schütze, H., 2009. Language models for information retrieval. In: *Introduction to Information Retrieval.* s.l.:Cambridge University Press, p. 237.

Diederich, J., Kindermann, J., Leopold, E. & Paass, G., 2003. Authorship Attribution with Support Vector. *Applied Intelligence ,* 19(1), pp. 109-123.

Diederich, J., Kindermann, J., Leopold, E. & Paass, G., 2003. machines, Authorship attribution with support vector. *Applied intelligence*, 19(1-2), pp. 109-123.

Escalante, H. J. T. S. M. M.-y.-G., 2011. *Local histograms of character n-grams for authorship attribution.* s.l., s.n.

Frantzeskou, G., Stamatatos, E., Gritzalis, S. & Katsikas, S., 2006. *Effective identification of source code authors using byte-level information.* ACM New York, s.n., pp. 893-896.

Gamon, M., 2004. *Linguistic correlates of style: authorship classification with deep linguistic analysis features.* Stroudsburg, s.n., p. 611.

Gamon, M., 2004. *Linguistic correlates of style: authorship classification with deep linguistic analysis features.* s.l., s.n.

Grieve, J., 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing,* 22(3).

Holmes, D. I., 1998. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing,* 13(3), pp. 111-117.

Howedi, F. & Mohd, M., 2014. Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data. *Computer Engineering and Intelligent Systems,* 5(4), pp. 48-58.

Jamak, A., Savatić, A. & Can, M., 2012. Principal component analysis for authorship attribution. *Business Systems Research,* 3(2), pp. 49-56.

Joula, P., 2008. Authorship Attribution. In: Boston: the essence of knowloge, pp. 238-239.

Juola, P., Sofko, J. & Brennan, P., 2006. A Prototype for Authorship Attribution Studies. *Literary and Linguistic Computing,* 21(2), pp. 169-178.

Jurafsky, D. & Martin, J. H., 2006. N-Grams. In: *An Introduction to Natural Language Processing,Computational Linguistics, and Speech Recognition.* s.l.:s.n.

Keselj, V., Peng, F., Cercone, N. & Thomas, C., 2003. *N-gram-based author profiles for authorship attribution.* PACLING, s.n.

KJELL, B., 1994. Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing,* 9(2), pp. 119-124.

Koehn, P., 2010. Language Modeles. In: *Statistical Machine Translation.* s.l.:Cambridge University Press, p. 181.

Koppel, M. & Schler, J., 2003. *Exploiting Stylistic Idiosyncrasies for Authorship Attribution.* s.l., s.n., p. 72.

Koppe, M., Schler, J. & Argamon, S., 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology,* 60(1), pp. 9-26.

Kourtis, Stamatatos, I. a. & Efstathios, 2011. *Author identification using semi-supervised learning.* Amsterdam, The Netherlands, s.n.

Lewis, D. D., Yang, Y., Rose, T. G. & Li, F., 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research,* Volume 5, pp. 361-97.

Luyckx, K. & Daelemans, W., 2005. Shallow Text Analysis and Machine Learning for Authorship Attribution. *LOT Occasional Series,* Volume 4, pp. 149--160.

Luyckx, K. & Daelemans, W., 2011. The effect of author set size and data size in authorship attribution. *Literary and linguistic Computing,* 26(1), pp. 35-55.

Manning, C. D. & Schiitze, H., 2000. Words. In: *Foundations of Statistical Natura Language Processing.* London, England: The MIT Press, p. 191.

Marton, Y., Wu, N. & Hellerstein, L., 2005. *On compression-based text classification.* s.l., Springer, pp. 300--314.

Mikros, G. K. & Perifanos, K., 2013. *Authorship Attribution in Greek Tweets Using Author's Multilevel N-Gram Profiles.* s.l., s.n.

Mosteller, Frederick, Wallace & David, 1964. *Inference and disputed authorship: The Federalist,* s.l.: s.n.

Nagaprasad, S. et al., 2015. Empirical Evaluations Using Character and Word N-Grams on Authorship Attribution for Telugu Text. In: *Intelligent Computing and Applications.* s.l.:Springer, pp. 613--623.

Pavelec, D., Justino, E. & Oliveira, L. S., 2007. Author Identification using Stylometric Features. *Inteligencia Artificial,* 11(36), pp. 59-66.

Peng, F. & Schuurmans, D., 2003. *Combining naive Bayes and n-gram language models for text classification.* s.l., Springer.

Pillay, S. R. & Solorio, T., 2010. *Authorship attribution of web forum posts.* Dallas, TX, IEEE.

Ramezani, R., Sheydaei, N. & Kahani, M., 2013. *Evaluating the effects of textual features on authorship attribution accuracy.* s.l., IEEE.

Ramyaa, He, C. & Rasheed, K., 2004. *Using machine learning techniques for stylometry.* s.l., s.n., pp. 897--903.

Rappoport, R. S. O. T. A. & Koppel, M., 2013. Authorship attribution of micro-messages. In: s.l.:s.n.

Satyam, A., Dawn, A. K. & Saha, S. K., 2014. *A Statistical Analysis Approach to Author Identification Using Latent Semantic Analysis.* s.l., s.n.

Savoy, J., 2012. Authorship attribution based on specific vocabulary. *ACM Transactions on Information Systems,* 30(2), p. 12.

Selman, S., Turan, K. & Kuşakçi, A. O., 2012. Distinction of The Authors of Texts Using Multilayered Feedforward Neural Networks. *SouthEast Europe Journal of Soft Computing,* 1(1), pp. 128-138.

Sidorov, G. a. V. F. a. S., Gelbukh, E. a., Alexander & Chanona-Hern, 2013. Syntactic dependency-based n-grams as classification features. In: *Advances in Computational Intelligence* . San Luis Potosi, Mexico: Springer, pp. 1-11.

Silva, R. S. et al., 2011. *twazn me!!! ;(' Automatic Authorship Analysis of Micro-Blogging Messages.* Alicante, Spain, s.n., pp. 161-168.

Soboroff, I. M., Nicholas, C. K., Kukla, J. M. & Ebert, D. S., 1997. *Visualizing document authorship using n-grams and latent semantic indexing.* New York, NY, USA, s.n.

Stamatatos, E., 2006. Authorship attribution based on feature set subspacing ensembles. *nternational Journal on Artificial Intelligence Tools,* 15(05), pp. 823-838.

Stamatatos, E., 2007. *Author Identification Using Imbalanced and Limited Training Texts.* Regensburg, s.n., pp. 237 - 241.

Stamatatos, E., 2008. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management,* 44(2), p. 790–799.

Stamatatos, E., 2008. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing \& Management,* 44(2), pp. 790--799.

Stamatatos, E., 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* , 60(3), pp. 538-556.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G., 2000. Automatic Text Categorization in Terms of Genre and Author. *Computational linguistics* , 26(4), pp. 471-495.

Stamatatos, J. H. a. E., 2006. *N-gram feature selection for authorship identification.* s.l., Springer, pp. 77--86.

Tearle, M., Taylor, K. & Demuth, H., 2008. An algorithm for automated authorship attribution using neural networks. *Literary and linguistic computing*, 23(4), pp. 425-442.

Türkoğlu, F., Diri, B. & Amasyalı, M. F., 2007. *Author Attribution of Turkish Texts by Feature Mining.* Qingdao, China, s.n., pp. 1086-1093.

Tweedie, F. J., Singh, S. & Holmes, D. I., 1996. Neural network applications in stylometry: The Federalist Papers. *Computers and the Humanities*, 30(1), pp. 1-10.

Uzuner, Ö. & Katz, B., 2005. *A comparative study of language models for book and author recognition.* Jeju Island, Korea, Springer, pp. 969-980.

Vel, O. d., Anderson, A., Corney, M. & Mohay, G., 2001. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4), pp. 55-64.

Zhao, Y. & Zobel, J., 2005. *Effective and Scalable Authorship Attribution Using Function Words.* Jeju Island, Korea, Springer, pp. 174-189.

Zhao, Y. & Zobel, J., 2007. *earching with style: Authorship attribution in classic literature.* s.l., s.n., pp. 59--68.

Zhao, Y., Zobel, J. & Vines, P., 2006. *Using relative entropy for authorship attribution.* s.l., Springer, pp. 92-105.

Zheng, R., Li, J., Chen, H. & Huang, Z., 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3), pp. 378-393.

آذین،ز1392, پایان نامه ارشد، , *شناسایی خودکار شاعران شعر نو با استفاده از ویژگی های زبانی*, مهندسی کامپیوتر، دانشگاه صنعتی شریف.

امیرشهاب ,ش.۱۳۸۵.، پایان نامه ارشد، ویژگی‌های نحوی، کارایی بیشتری را در بین سایر ویژگی‌ها دارد، مهندسی کامپیوتر و فناوری اطلاعات

فرهمندپور،زینب .1390 ,پایان پایان نامه ارشد، طراحی و پیاده سازی یک سیستم هوشمند تشخیص هویت بر اساس سبک نوشتاری فارسی، مهندسی کامپیوتر، دانشگاه بوعلی سینا همدان.

Abstract

Authorship attribution (AA) or author identification refers to the problem of determining who has written a disputed text or unseen text. In the close class authorship attribution problem, the unseen text is assigned to any one of candidate authors set, that text sample as training data are available for them. Two main requirements of authorship attribution system are features and attribute method. Features are usually selected with training data.

With increasing text in different languages, seems to be an essential need for developing authorship attribution system which is language independent. Since the procedure of extracting character n-grams and word n-grams are language-independent and require no special tools, in this thesis them have been used as features. Also language modeling has been chosen as a statistical and probabilistic attribute method. We present an approach based on language modeling called modified language modeling. It aims to offer a solution for AA problem by combinations of both bigram words weighting and unigram words weighting. Moreover, the IDF value multiplied by related word probability has been used, instead of removing stop words and balancing word probability as weights, as well.

In order to evaluate the results, four corpora have been used. Two datasets of Persian poetry, one Persian prose dataset and the fourth is English prose dataset. The accuracy of AA is calculated by language modeling with character n-grams, language modeling with word n-grams and modified language modeling on the four datasets. In all databases, modified language modeling shows improvement. The best performance is obtained for modified language by Persian prose dataset, which is 100 percent.

Keywords: Authorship Attribution, Authorship Identification, Language Modeling, WMPR-AA2016-A corpora, WMPR-AA2016-B corpora

Department of Electronic Learning

M.Sc. Thesis in Artificial Intelligence Engineering

Authorship attribution with statistical modeling on text data

By: Samane Vazirian

Supervisor:

Dr. Morteza Zahedi

Advisor:

Dr. Hamid Hasanpour

September 2016