دانشکده مهندسی

گروه کامپیوتر

پایان‌نامه برای دریافت درجه کارشناسی ارشد در رشته مهندسی کامپیوتر گرایش هوش مصنوعی

عنوان:

طراحی و پیاده‌سازی یک سیستم هوشمند تشخیص هویت بر اساس سبک نوشتاری فارسی

استاد راهنما:

دکتر هومان نیک مهر

استاد مشاور:

دکتر محرم منصوری‌زاده

دکتر امید طبیب‌زاده

نگارش:

زینب فرهمندپور

اسفند ماه ۱۳۹۰

کلیه امتیازهای این پایان‌نامه به دانشگاه بوعلی سینا تعلق دارد.

در صورت استفاده از تمام یا بخشی از مطالب این پایان‌نامه در مجلات، کنفرانس‌ها و یا سخنرانی‌ها، باید نام دانشگاه بوعلی سینا یا استاد راهنمای پایان‌نامه و نام دانشجو با ذکر مأخذ و ضمن کسب مجوز کتبی از دفتر تحصیلات تکمیلی دانشگاه ثبت شود. در غیر این صورت مورد پیگرد قانونی قرار خواهد گرفت. درج آدرس‌های ذیل در کلیه مقالات خارجی و داخلی مستخرج از تمام یا بخشی از مطالب این پایان‌نامه در مجلات، کنفرانس‌ها و یا سخنرانی‌ها الزامی می‌باشد.

مقالات خارجی

……..……, Bu-Ali Sina University, Hamedan, Iran. ……

مقالات داخلی

.....، گروه .....، دانشکده .....، دانشگاه بوعلی سینا، همدان.

دانشگاه بوعلی سینا

مشخصات رساله/پایان نامه تحصیلی

عنوان:

طراحی و پیاده‌سازی یک سیستم هوشمند تشخیص هویت بر اساس سبک نوشتاری فارسی

نام نویسنده: زینب فرهمندپور

نام استاد/اساتید راهنما: دکتر هومان نیک‌مهر

نام استاد/اساتید مشاور: دکتر محرم منصوری، دکتر امید طبیب‌زاده

دانشکده : مهندسی	گروه آموزشی: کامپیوتر	
رشته تحصیلی: مهندسی کامپیوتر	گرایش تحصیلی: هوش مصنوعی	مقطع تحصیلی: کارشناسی ارشد
تاریخ تصویب پروپوزال: ۸۹/۶/۳۰	تاریخ دفاع: ۱۳۹۰/۱۲/۲۴	تعداد صفحات: ۱۷۱

چکیده:

تشخیص هویت نویسنده، یک مساله‌ی سبک‌شناسی است که سعی می‌کند یک متن را که نویسنده‌ی آن ناشناس است، به نویسنده‌ی واقعی آن متن نسبت دهد. این موضوع در زبان‌های مختلفی پیاده‌سازی شده و مورد بحث قرار گرفته ولی در زبان فارسی این چنین به آن پرداخته نشده بود. آنچه در این پایان‌نامه مورد بررسی قرار می‌گیرد طراحی و پیاده‌سازی یک سیستم تشخیص نویسنده بر اساس سبک نوشتاری فارسی است. در این پایان‌نامه علاوه بر طراحی و پیاده‌سازی سیستم تشخیص نویسنده متن ناشناخته، مطالعه‌ای روی مقایسه‌ی روش‌های یادگیری ماشین برای تشخیص هویت نویسنده انجام شده است. در این تحقیق ۷ روش دسته‌بندی Delta، K-nearest neighbors (KNN)، Linear Discriminant Analysis (LDA)، درخت تصمیم‌گیری، شبکه‌های عصبی، ترکیب الگوریتم ژنتیک و KNN، ترکیب الگوریتم رقابت استعماری و KNN روی ۲ پایگاه داده جمع‌آوری شده با هم مقایسه شدند.

واژه‌های کلیدی: تشخیص نویسنده، writeprint، سبک‌شناسی، دسته‌بندی

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

الرَّحْمَنُ (۱) عَلَّمَ الْقُرْآنَ (۲) خَلَقَ الْإِنسَانَ (۳) عَلَّمَهُ الْبَيَانَ (۴) سوره: الرحمن

خدای رحمان (۱) قرآن را یاد داد (۲) انسان را آفرید (۳) به او بیان آموخت (۴)

تقدیم به

پدرم که عالمانه به من آموخت تا چگونه در عرصه زندگی، ایستادگی را تجربه نمایم

و

مادرم، دریای بی کران فداکاری و عشق که وجودم برایش همه رنج بود و وجودش برایم همه مهر

سپاسگزاری

سپاس بی کران پروردگار یکتا را که هستی‌مان بخشید و به طریق علم و دانش رهنمونمان شد و به همنشینی رهروان علم و دانش مفتخرمان نمود و خوشه چینی از علم و معرفت را روزیمان ساخت.

به مصداق «من لم یشکر المخلوق لم یشکر الخالق » بسی شایسته است از استاد

فرهیخته و فرزانه جناب آقای دکتر نیک مهر

که با کرامتی چون خورشید ، سرزمین دل را روشنی بخشیدند و گلشن سرای علم و

دانش را با راهنمایی‌های کار ساز و سازنده بارور ساختند، تقدیر و تشکر نمایم.

از زحمات فراوان جناب آقای دکتر منصوری‌زاده و جناب آقای دکتر طیب‌زاده

کمال سپاس‌گذاری و قدردانی را دارم.

از همراهی و حمایت‌های همیشگی خانواده عزیزم که در تمام دوران تحصیلم مشوق و همراه همیشگی‌ام بوده و هستند، بی‌نهایت متشکرم.

فهرست مطالب

فهرست اشکال

فهرست جداول

فصل اول

پیشگفتار

۱-۱ مقدمه

شناسایی بیومتریک انواع مختلفی دارد. از جمله این روش‌ها می‌توان به اسکن عنبیه، اسکن شبکیه، شناسایی چهره، شناسایی صوت، اثر انگشت و ترکیب دست اشاره کرد. این خصوصیات در هر شخص منحصر به فرد هستند و به همین خاطر در شمار موارد قابل قیاس در تشخیص هویت افراد قرار گرفته‌اند.

هر چند استفاده از روش‌های بیومتریک، جزء روش‌های نسبتا خوب برای تشخیص هویت محسوب می‌شوند، اما به همه آنها نمی‌توان اعتماد کرد و برخی از این روش‌ها را می‌توان جعل کرد یا فریب داد.

تشخیص هویت مبتنی بر اثر انگشت و دیگر روش‌های بیومتریک، جزء روش‌های زیست-شناختی قدیمی موفق بوده است که در گذشته در کشف جرم به کار می‌رفت. الگوی یکسان و تغییرناپذیری اثر انگشت معیاری است که برای شناسایی مجرم به کار می‌رود.

۱-۲ ضرورت تحقیق

علی‌رغم مطالب گفته شده مواردی پیش می‌آید که هیچ اثر بیومتریک سنتی برای یافتن مجرم در دسترس نیست؛ مانند زمانی که تنها مدرک موجود، متنی تایپ شده است. خوشبختانه، نوع دیگری از اثر که writeprint نامیده می‌شود، وجود دارد که در نوشته‌های افراد پنهان است. همانند اثر انگشت، writeprint از ترکیبی از چندین خصوصیت همانند پرمایگی واژگان، طول کلمات، استفاده از کلمات تابعی، چینش پاراگراف‌ها، کلمات کلیدی و غیره تشکیل شده است. این خصوصیات سبک شناختی، می‌توانند شیوه نگارش نویسنده را آشکار سازند و معمولا در طول نوشتار آن فرد ثابت می-مانند در نتیجه به عنوان اساس تحلیل هویت نویسنده به کار می‌رود.

سبک‌شناسی در معنی گسترده خود، می‌تواند به تحلیل ویژگی‌های سبکی زبان گروهی خاص از یک جامعه همچون سیاستمداران یا روزنامه‌نگاران، به عنوان گونه‌ای زبانی بنشیند. گاه هدف از این مطالعات، روشنگری است؛ بدین معنا که ممکن است متنی داشته باشیم و نویسنده یا شاعر آن را نشناسیم. در این شرایط سبک‌شناسی به کمک ما می‌آید تا با تحلیل ویژگی‌های سبکی آن متن، مولف یا دست‌کم مقطع تاریخی تالیف آن متن را مشخص کنیم. البته در یک تحلیل سبکی کامل، همه سطوح زبانی اعم از صدا، صورت، ساخت و معنا، باید مورد تحلیل قرار بگیرند.

یداله ثمره با ذکر غزلی که از جانب حافظ‌پژوهی به نام «بلوخمان[1]» در سال ۱۸۸۷ ارائه شده و در هیچ یک از نسخ دیوان حافظ نیامده است، خاطرنشان کرد: بلوخمان مدعی است که این غزل متعلق به حافظ است. با تحلیل سبکی زبان این غزل و مقایسه آن با نتایج سبک‌شناسی برآمده از شعر حافظ، می‌توان میزان درستی این ادعا را تحقیق کرد.

امروزه شرایط به گونه‌ای تغییر کرده که نیاز به روش‌های دیگر برای تشخیص هویت افراد احساس می‌شود. حتی در نوعی از جرایم که سرقت علمی نامیده می‌شود و آن استفاده از ایده‌ها و نتایج دیگران است، طوری که وانمود شود که انگار کار شماست. ایده‌ها و نتایج دیگران می‌تواند در یک مقاله دیگر، در یک کتاب دیگر یا فرهنگ لغت، ترجمه از یک متن به زبان دیگر یا حتی از مقاله دوستان باشد.

از همه مهم‌تر، فراوانی متون الکترونیکی در دسترس، لزوم تحلیل نویسنده را در کاربردهای متنوع نشان می‌دهد. در زمینه‌های متفاوت مانند هوش مصنوعی (تعیین پیغام‌ها یا اعلامیه‌های تروریستی از بین تروریست‌های شناخته شده) [۱]، قوانین جزایی (مانند شناسایی نویسندگان پیغام-

[1] یداله ثمره عضو پیوسته فرهنگستان زبان و ادب فارسی و مدیر سابق پژوهشگاه گویش‌شناسی در فرهنگستان

های تهدیدآمیز، تشخیص نویسنه‌های متون خودکشی)، قوانین مدنی (قوانین کپی مورد بحث copywrite) [۲, ۳]، مناظره‌های کامپیوتری (مانند شناسایی نویسنده‌ی کد برنامه‌های نرم‌افزاری دارای سوء نیت)[۴] و تحقیقات ادبی (مانند تشخیص نویسنده‌ی ناشناس یک سری کارهای ادبی مورد مشاجره از بین نویسندگان شناخته شده)[۵ ، ۶] نیاز به تحلیل ویژگی‌های نویسنده هست. بنابراین دهه‌ی اخیر را می‌توان به عنوان یک دوره‌ی جدید در زمینه‌ی تکنولوژی تشخیص هویت نویسنده نامید. در این زمان تلاش‌هایی برای توسعه‌ی کاربردهای واقعی و عملی مرتبط با متون دنیای واقعی در صدر قرار دارد.

۱-۳ اهداف تحقیق

این پایان‌نامه علاوه بر معرفی خصوصیات سبک نویسنده فارسی زبان، و استفاده از روش‌های مختلف جداسازی نویسندگان از یکدیگر و مقایسه روش‌ها، شناسایی و کشف مجرم را آسان‌تر می‌کند. هدف، طراحی سیستمی هوشمند است که به صورت اتوماتیک و با روش‌های گوناگون به استخراج ویژگی‌های سبک نوشتاری هر فرد بپردازد و به تشخیص هویت نویسنده کمک کند.

۱-٤ روش‌های ارائه شده

ابتدا ویژگی‌های متون مورد بحث به طور جداگانه استخراج می‌شوند. در این راستا سعی در استخراج ویژگی‌های متمایز کننده متون داریم و pos tagger طراحی کردیم که قادر به شناسایی اسم، صفت و قید می‌باشد. در مرحله‌ی بعد به پیش‌پردازش ویژگی‌های استخراج شده پرداختیم و با استفاده از روش‌های انتخاب و استخراج ویژگی به انتخاب بهترین ویژگی‌ها که دقت مناسبی را روی داده‌های آموزشی و تست داشتند، پرداختیم. یکی از روش‌های انتخاب ویژگی که در پایان نامه استفاده نمودیم، ترکیب الگوریتم رقابت استعماری و روش جداسازی KNN بود که در مطالعات قبلی تشخیص نویسنده به کار گرفته نشده بود. نتیجه این الگوریتم را با ترکیب الگوریتم ژنتیک و KNN

مقایسه نمودیم و به برتری نسبی الگوریتم رقابت استعماری پی بردیم. از روش‌های دیگر دسته‌بندی که دقت آنها را روی پایگاه‌داده جمع‌آوری شده تست نمودیم می‌توان به Delta، شبکه عصبی، Linear Discriminant Analysis و درخت تصمیم‌گیری اشاره نمود. خصوصیات مناسب که برای هر نویسنده یکتا هستند، جهت جداسازی متون انتخاب می‌شوند. در انتها روش‌های ذکر شده با یکدیگر مقایسه شده و بهترین روش جداسازی برای مجموعه داده‌ای ما معرفی می‌شود.

در این پایان نامه برای ارزیابی سیستم تشخیص نویسنده در زبان فارسی، دو مجموعه داده جمع‌آوری شده و سعی شده است تا استانداردهای جمع‌آوری پایگاه‌داده در این زمینه رعایت شوند.

ویژگی روش‌های ارائه شده:

روش ارائه شده در این پایان نامه به دلیل استفاده از ویژگی‌های مختلف و کارا، دارای دقت بالا در حدود ۸۱٪ برای پایگاه‌داده اول و دقت ۱۰۰٪ برای پایگاه‌داده دوم است. لذا ویژگی‌ها و روش‌های به کار برده شده فوق می‌توانند کارا باشند. استفاده از ویژگی نحوی تشخیص اسم، صفت و قید و ترکیب الگوریتم رقابت استعماری و روش KNN نشان داد که انتخاب ویژگی تاثیر بسیار زیادی در کارایی الگوریتم یادگیری دارد. الگوریتم‌های یادگیری برای موفقیت در تشخیص نویسندگان نیازمند استفاده از مجموعه‌ی کوچکی از ویژگی‌ها می‌باشد. وجود صفات نامربوط و زائد در مرحله‌ی ساخت مدل می‌تواند منجر به کاهش دقت و محاسبات زیاد شود. در حالت کلی نمی‌توان ویژگی‌های خاصی را برای مجموعه داده‌های گوناگون مشخص نمود. از طرفی بررسی همه‌ی زیر مجموعه‌های مربوطه، نیازمند وقت زیاد بوده و در حالت کلی ناممکن است. بنابراین روش الگوریتم رقابت استعماری به دلیل سرعت بالا و همگرایی زود هنگام، روشی مناسب برای انتخاب ویژگی در زمان مناسب می‌باشد.

۱-۵ ساختار پایان نامه

در فصل اول به مقدمه‌ای از مساله تشخیص هویت نویسنده پرداخته شده است، در این فصل هدف از تشخیص هویت نویسنده، معرفی و پیگیری مشکلات، سابقه‌ی نظری و تئوری، بررسی اجمالی تاریخی بیان شده است. در فصل دوم مقدمه ای بر تشخیص هویت نویسنده گفته شده است. مروری بر کارهای پیشین تشخیص هویت نویسنده در فصل سوم توضیح داده شده است. در فصل چهارم، روش‌های انتخاب ویژگی مورد بحث قرار گرفته اند. روش‌های دسته بندی در فصل ۵ عنوان شده و در فصل ششم به پیاده‌سازی و ارزیابی سیستم تشخیص هویت نویسنده در زبان فارسی پرداخته شده است. در انتها در فصل هفتم، نتیجه‌گیری و پیشنهادهایی برای کارهای آتی بیان شده است.

فصل دوم

مقدمه‌ای بر تشخیص نویسنده

۲-۱ مقدمه

تشخیص هویت نویسنده انتساب یک متن به نویسنده‌ی واقعی آن در بین مجموعه‌ای از نویسندگان نامزد است. برای این منظور باید اطلاعات متمایز کننده نویسندگان از یکدیگر شناسایی شوند و مورد استفاده قرار گیرند. در قسمت ۲-۲ اهمیت شناسایی نویسنده توضیح داده شده است. در بخش ۲-۳ به معرفی و پیگیری مشکلات مربوط به آن پرداخته شده است. سابقه‌ی نظری و تئوری این بحث در قسمت ۲-۴ و در ۲-۵ به بررسی اجمالی و تاریخی موضوع و در ۲-۶ به تحلیل موضوع federalist paper پرداخته شده است و سر انجام در بخش ۲-۶ نتیجه‌گیری بیان شده است.

۲-۲ چرا تشخیص هویت نویسنده؟

در سال ۲۰۰۴ کتاب "why the west is losting the war on terror" منتشر شد. با توجه به تجربه‌ی گسترده نویسنده، این کتاب به تشریح وضعیت فعلی جنگ آمریکا علیه ترور می‌پرداخت و استدلال می‌کرد که سیاست ایالات متحده امریکا بسیار گمراه کننده بوده است. نویسنده‌ی این کتاب از نظر فنی ناشناس بود، اگرچه نویسنده "through our enemie's eyes" ادعا کرده بود که مقام ارشد اطلاعاتی امریکا ، نزدیک به دو دهه تجربه است. مطابق با ادعای مطرح شده در Boston Phoenix نسخه‌ی دوم جولای ۲۰۰۴، نویسنده‌ی واقعی Michael Scheuer، افسر ارشد سازمان CIA و رئیس سازمان CIA واحد اسامه بن لادن در اواخر سال‌های ۱۹۹۰ بوده است.

از سوی دیگر، بر طبق نوشته‌ی بعضی از مورخان مانند Hugh Trevor Roper نویسنده‌ی کتاب Hitler Diaries در سال ۱۹۸۳، خود هیتلر بوده است. با این وجود، یافته‌هایی به دست آمد که نشان داد متن آن کتاب روی کاغذهای جدید و با استفاده از جوهر که در سال ۱۹۴۵ در دسترس نبوده، نوشته شده است. چرا ما باید حرف‌های مورخان و روزنامه‌نگاران را بدون توجه به چگونگی آنها باور کنیم؟ چه نوع شواهدی قبل از اینکه مطمئن شویم، مورد نیاز است؟

تشخیص نویسنده یک تکه متن خاص، پرسش‌های روش‌شناختی را برای قرن‌ها به وجود آورده است. تشخیص هویت نویسنده نه تنها می‌تواند علاقه‌ی دانشمندان علوم سیاسی را برانگیزد، بلکه همانند مثال بالا، در یک مفهوم بسیار عملی‌تر می‌تواند توجه سیاستمداران، روزنامه نگاران و وکلا را نیز جلب کند.

تحولات اخیر با استفاده از تکنیک‌های جدید و کامپیوتر و پایگاه داده‌های کامپیوتری، یک استنباط عینی و اتوماتیک برای تشخیص نویسنده ایجاد کرده است.

۲-۳ معرفی و پیگیری مشکلات

تشخیص نویسنده که به طور گسترده توضیح داده شد، یکی از قدیمی‌ترین و یکی از جدیدترین مسائل در سبک‌شناسی می‌باشد. اختلاف درباره مالکیت کلمات از زمانی که وجود داشتند، مورد بحث بوده است. از زمان به وجود آمدن سندها، سوالات راجع به تصدیق هویت و خصوصیات نویسنده‌ی آنها وجود داشته است.

بنابراین می‌توان تشخیص نویسنده را به صورت تلاشی برای نشان دادن خصوصیات تولید کننده یا نویسنده‌ی یک تکه از اطلاعات زبانی تعریف کرد. این یک تعریف سنجیده است. این فیلد از علم، شامل اکثر تحقیقات انجام شده روی تشخیص گفتار است. به عبارت دیگر، اکثر تکنیک‌های تشخیص هویت نویسنده می‌توانند روی تشخیص هویت گفتار به کار گرفته شوند و در بعضی از شرایط می‌توانند بسیار مفیدتر واقع شوند.

در شرایط گسترده، ۳ مساله‌ی اصلی در تشخیص نویسنده وجود دارد. اولی داشتن یک نمونه متن است که می‌دانیم توسط یک نفر از افراد مجموعه نویسندگان نوشته شده و تعیین اینکه کدامیک