
بسم الله الرحمن الرحيم

دانشگاه الزهرا (س)
دانشکده فنی و مهندسی

پایان نامه
جهت اخذ درجه کارشناسی ارشد
رشته مهندسی کامپیوتر گرایش هوش مصنوعی

عنوان

ارائه یک مدل متن کاوی مبتنی بر یادگیری نیمه نظارتی

استاد راهنما
دکتر محمد رضا کیوان‌پور
استاد مشاور
دکتر رضا عزمی

دانشجو
مریم باحجب ایمانی

اسفند ماه سال ۱۳۸۹

کلیه دستاوردهای این تحقیق متعلق به

دانشگاه الزهرا(س) است.

تقدیم به مادرم که مهربانی و پدرم که صبوری را به من آموختند.

سپاس و ستایش خداوند را که به من توفیق داد تا به یاری و مدد او این پایان نامه را به پایان رسانم.

بر خود واجب می دانم که،

از استاد راهنمای بزرگوارم جناب آقای دکتر کیوان پور که با دانش خود همواره مرا در تکمیل این پژوهش راهنمایی کردند،

از اساتید مشاور ارجمندم جناب آقای دکتر عزمی که در طول انجام مراحل پایان نامه همواره با آگاهی و دقت نظر خاص راهنمایی‌های ارزشمندشان را از من دریغ ننمودند،

از جناب آقای دکتر قلی زاده که داوری این پایان نامه را بر عهده داشتند،

از خانواده عزیزم، به ویژه از برادر مهربانم که همواره یاور و پشتیبان من بوده‌اند،

و از تمام کسانی که در طول دوران تحصیل با راهنمایی‌های ارزشمند علمی و معنوی خود، مرا یاری نمودند سپاسگزاری نموده و از خداوند برای این عزیزان آرزوی توفیق و سربلندی می‌نمایم.

همچنین جا دارد از مرکز تحقیقات مخابرات ایران که از این پایان نامه تحت قرارداد شماره‌ی ۸۹۷۱/۵۰۰ حمایت کرده است، قدردانی شود.

چکیده

محبوبیت وب و حجم زیاد مستندات متنیِ الکترونیکی موجود، باعث افزایش نیاز به جستجو برای استخراج دانش نهان از مجموعه‌ی مستندات متنی شده است. بنابراین، امروزه مسئله‌ی متن کاوی در زمینه‌های متعددی از جمله پزشکی، زیست-فناوری، اقتصاد و فناوری اطلاعات مورد توجه قرار گرفته است. متن‌کاوی قادر است پردازش‌هایی مانند طبقه‌بندی، خوشه‌بندی، خلاصه‌سازی و استخراج اطلاعات متنی را پوشش دهد. طبقه‌بندی متون به شیوه‌ای مناسب با میزان خطای کم و تعمیم‌پذیری بالا یکی از موضوعات مهم در حوزه‌ی متن‌کاوی است. یکی از مهم‌ترین چالش‌ها در طبقه‌بندی متون، حجم زیاد مشخصه‌های مستخرج از اطلاعات متنی می‌باشد. یادگیری از داده‌هایی که مشخصه‌های زیادی دارند نه تنها باعث افزایش هزینه‌های محاسباتی می‌شود، بلکه کارایی یادگیری را نیز کاهش می‌دهد. بر این اساس استفاده از روش‌های مناسب انتخاب مشخصه از اهمیت ویژه‌ای در این حوزه برخوردار می‌باشد. در این راستا، در پژوهش انجام شده یک روش انتخاب مشخصه‌های توکار برای حل این چالش پیشنهاد شده است که نتایج بهتری را نسبت به روش‌های رایج می‌دهد. بهره‌گیری از روش‌های یادگیری با نظارت، که از مثال‌های آموزشی بر چسب دار استفاده می‌کنند، به عنوان یکی از رویکردهای سنتی جهت طبقه‌بندی متون مطرح است. برای انجام این نوع یادگیری با دقتی منطقی، وجود تعداد کافی از مثال‌های آموزشی برچسب دار ضروری است. بدین منظور به فردی خبره نیاز است که به هر سند برچسبی نسبت دهد؛ که این کار فرآیندی خسته‌کننده، زمانبر و پر هزینه می‌باشد. بنابراین تأمین تعداد کافی از مثال‌های آموزشی برچسب دار عملی غیر ممکن است. در مقابل، اسناد بدون برچسب اغلب در حجم زیاد قابل دسترس هستند. بنابراین، رویکرد موثر و عملی دیگر در یادگیری استفاده از اسناد برچسب دار به همراه اسناد بدون برچسب در زمان یادگیری می‌باشد، این ایده مبنای اصلی رویکرد یادگیری نیمه‌نظارتی را تشکیل می‌دهد. در این حالت، الگوریتم‌های یادگیری می‌توانند از داده‌های بدون برچسب استفاده کنند، که اغلب منتهی به تابع طبقه‌بندی دقیق‌تری می‌شود. در این پژوهش، روشی مبتنی بر یادگیری تجمیعی و رویکرد خودآموزی برای انجام یادگیری نیمه‌نظارتی پیشنهاد شده است که بر اساس آزمون‌های انجام شده موجب بهبود کارایی یادگیری نیمه‌نظارتی در زمینه‌ی طبقه‌بندی متون شده است.

کلمات کلیدی: متن کاوی؛ طبقه‌بندی اسناد؛ انتخاب مشخصه‌ها؛ یادگیری نیمه‌نظارتی

فهرست مطالب

فهرست جدول‌ها

فهرست شکل‌ها

فصل ۱

مقدمه

حجم بسیار زیادی از داده‌ها امروزه به شکل دیجیتال در دسترس هستند که اکثر آن‌ها برای ماشین‌ها غیر قابل درک هستند. امروزه نرخ جمع‌آوری و ذخیره سازی الکترونیکی این داده‌ها، تقریباً در اکثر زمینه‌ها بسیار بالاست. بنابراین، اکتشاف اطلاعات مفید از داده‌های در دسترس به صورت یک الزام در حوزه‌های کاربردی مختلف مطرح است. برای مثال، ممکن است داده‌ها حاوی اطلاعاتی در مورد بازارهای مهم رقابتی مانند رقبا و مشتریان باشد و یا حاوی اطلاعات مفیدی در مورد بهینه‌سازی فرصت‌ها برای بهبود فرآیند کسب و کار باشد. بنابراین، پژوهش‌ها در رابطه با اکتشاف خودکار این نوع دانش از بانک‌های اطلاعاتی آغاز شده است. متن کاوی یکی از مهم‌ترین رویکردهای مطرح در این زمینه است که حوزه‌های کاربردی آن گستره وسیعی را شامل می‌شود.

با توجه به مطالعات انجام شده، متن کاوی به فرآیند تحلیل متن برای استخراج یا اکتشاف اطلاعات و واقعیات معتبر، جدید و از قبل ناشناخته، پنهان، مفید و قابل درک از داده‌های غیرساخت‌یافته[۱] و نیمه‌ساخت‌یافته[۲] به صورت خودکار توسط کامپیوتر اطلاق می‌شود. متن کاوی از روش‌های بازیابی اطلاعات، استخراج اطلاعات، و پردازش زبان طبیعی[۳] (NLP) استفاده می‌کند و آن‌ها را با الگوریتم‌ها و روش‌های اکتشاف دانش پایگاه داده، داده‌کاوی[۴]، یادگیری ماشین[۵]، آمار، پایگاه داده‌ها، انبار داده[۶]، مدیریت دانش ، طراحی واسط کاربری، و مصورسازی[۷] مرتبط می‌سازد. با این تفاوت که این تحلیل بر متن اسناد تمرکز می‌کند [۳،۲،۱].

متن کاوی را می‌توان در گستره وسیعی از حوزه‌های کاربردی به کار گرفت. به عنوان نمونه‌ای از حوزه‌های کاربردی متن کاوی می‌توان به حوزه‌هایی مانند اجتماعی، سیاسی، صنعتی، فناوری اطلاعات و پزشکی اشاره کرد [۷،۵،۶،۴].

بر اساس چارچوب‌های مطرح شده در متن‌کاوی مانند [۲, ۴, ۵] می‌توان چارچوبی شامل سه مرحله پیشنهاد داد. مرحله‌ی اول شامل پیش پردازش است که مرکزیت آن بر روی استخراج مشخصه‌ها[۸] و انتخاب آن‌ها نمایشگر اسناد می‌باشد. این عملگرهای پیش‌پردازش مسئول تبدیل شکل داده‌های ذخیره شده‌ی بدون ساختار یا

---

۱ Unstructured
۲ Semi-structured
۳ Natural Language Processing
٤ Data Mining
٥ Machine Learning
٦ Data warehouse
۷ Visualization
۸ Feature Extraction

نیمه‌ساخت‌یافته در مجموعه‌ها اسناد، به صورت ساخت‌یافته می‌باشند (مانند خالص کردن داده، حذف لغات stopword، ریشه‌یابی[۱]، شاخص گذاری[۲]، پالایش متن، تولید مشخصه‌ها و کاهش فضای مشخصه‌ها). مرحله‌ی دوم دوم پردازش است که در این مرحله روش‌های زیادی از جمله بسیاری از روش‌های یادگیری ماشین و داده‌کاوی مورد استفاده قرار می‌گیرد. در واقع در این مرحله، اطلاعات پنهان و مفید استخراج می‌شوند. در مرحله‌ی آخر به ارزیابی و تعیین میزان کارایی روش‌ها و ایده‌های پیاده‌سازی شده در مرحله‌ی قبل پرداخته می‌شود؛ و در انتها الگوریتم اصلاح شده‌ی نهایی ارائه می‌شود.

یکی از مهم‌ترین مؤلفه‌های چارچوب متن‌کاوی، طبقه‌بندی متون[۳] است؛ که وظیفه‌ی آن طبقه‌بندی مجموعه‌ای مجموعه‌ای از اسناد متنی به صورت خودکار به مجموعه‌های از پیش تعریف شده است. برای مثال می‌توان به جدا کردن نامه‌های هرز[۴] از نامه‌های الکترونیکی غیر هرز و یا پیدا کردن صفحات وب با دادن موضوع مشخص اشاره کرد.

در مرحله‌ی پیش‌پردازش برای طبقه‌بندی متون، معمولاً داده‌های متنی شامل رشته‌هایی از حروف هستند که به نمایشی مناسب برای یادگیری تبدیل می‌شوند. سپس انتخاب مشخصه‌ها که شامل روشی برای کاهش فضای ویژگی می‌باشد، انجام می‌شود و در انتهای این مرحله کلمات هر متن سند می‌توانند با استفاده از روش‌هایی وزن‌دهی شود. در مرحله‌ی بعد، یعنی پردازش، از الگوریتم‌های یادگیری ماشین برای طبقه‌بندی متون استفاده می‌شود و در مرحله‌ی آخر نتایج پردازش‌های انجام شده توسط معیارهای کارایی ارزیابی می‌شود [۳, ۶].

۱-۱- یادگیری ماشین

در انسان‌ها، یادگیری از طریق فرآیندهای مشخص و تغییرات در سیستم عصبی اتفاق می‌افتد. فعالیت یادگیری معمولاً شامل مجموعه‌ای از مشاهدات در رابطه با پدیده‌های طبیعی و فرآیندی برای تبدیل آن اطلاعات به دانش می‌باشد.

سیستم‌های مصنوعی از طریق تغییراتی که آن‌ها به نمایش‌های داخلی خودشان، داده‌ها، مدل‌ها و/یا ساختارها (در سطح نرم‌افزاری و سخت‌افزاری) می‌دهند، یادگیری می‌کنند. این ماشین‌ها برای حل وظایف معینی ساخته

---

[۱] Stemming
[۲] Indexing
[۳] Text Classification
[۴] Spam

می‌شوند و هدف از یادگیری آن‌ها، بهبود کارایی‌شان در آن وظایف توسط یادگیری از محیط، آموزگاران، و تجربیات خودشان در دستیابی به اهداف مشخص می‌باشد. روش‌های یادگیری ماشین برای وظایفی که پیچیدگی مسئله از لحاظ ورودی‌ها و خروجی‌های متفاوت بسیار زیاد است که برنامه‌نویس نمی‌تواند آن‌ها را به صورت صریح کد کند، می‌تواند بسیار مفید باشد.

- **یادگیری نظارتی**[۱]:

روش‌های یادگیری ماشین معمولاً از طریق نیاز آن‌ها به نوع داده‌ای که می‌توانند استفاده کنند، طبقه بندی می‌شوند. الگوریتم‌هایی با نظارت کامل، همیشه به یک مربی[۲] برای آشکار کردن برچسب داده‌ها نیاز دارند. هدف این الگوریتم‌ها یادگیری نگاشتی از داده‌های ورودی به خروجی مشخص‌شده توسط مربی می‌باشد.

یادگیری طبقه‌بندها با استفاده از مثال‌های آموزشی برچسب دار برای پیش‌بینی برچسب مثال جدیدی که از قبل دیده نشده است.

- **یادگیری بدون نظارت**[۳]:

از طرف دیگر یادگیری بدون نظارت، فقط با داده‌های آموزشی و بدون نیاز به مربی کار می‌کنند. آن‌ها معمولاً سعی می‌کنند که الگوی شباهتی میان ورودی‌هایشان پیدا کنند و خوشه‌هایی از نمونه‌های داده‌های آموزشی را تشخیص دهند.

- **یادگیری نیمه نظارتی**[۴]:

یادگیری‌ای که با استفاده از مثال‌های آموزشی برچسب دار و بدون برچسب برای پیش‌بینی برچسب مثال جدید انجام می‌شود. در بسیاری از دامنه‌های یادگیری عملی (مانند طبقه‌بندی متن و تصویر و زیست فناوری [۵])، داده‌های بدون برچسب زیادی به همراه داده‌های برچسب دار محدودی وجود دارد، و در بیشتر موارد تولید داده‌های برچسب دار هزینه‌ی زیادی دارد.

---

۱ Supervised
۲ Teacher
۳ Unsupervised
۴ Semi-supervised
5 Bioinformatic

- **روش‌های فرا یادگیری[1] و تجمیعی[2]**:

الگوریتم‌های فرا یادگیری نوع خاصی از روش‌های یادگیری هستند که می‌توانند از همکاری دیگر یادگیرنده‌ها به عنوان توابع پایه‌شان استفاده کنند و مدلی از ترکیب آن‌ها تولید کنند. این مدل معمولاً کارایی بهتری نسبت به هر یک از یادگیرنده‌های پایه‌ی سازنده دارند. به دلیل اینکه این الگوریتم‌ها در رأس دیگر روش‌های یادگیری کار می‌کنند، به آن‌ها تحت عنوان فرا یادگیرنده‌ها و همچنین روش‌های تجمیعی اشاره می‌شود.

در این پژوهش نیز روشی از یادگیری نیمه‌نظارتی پیشنهاد و از آن استفاده شده است.

۱-۲- طرح مسئله

محبوبیت وب و حجم زیادی از مستندات متنی که به صورت الکترونیکی در دسترس هستند، باعث افزایش جستجوی دانش نهان در مجموعه‌ای از مستندات متنی شده است. بر اساس مطالعات انجام شده، حدود ۸۰-۹۰٪ تمام داده‌ها به شکل غیر ساخت یافته یا نیمه‌ساخت‌یافته مانند ایمیل، اسناد تمام متن، فایل‌های HTML، فایل‌های PDF ذخیره شده‌اند. بنابراین، امروزه مسئله‌ی متن کاوی یعنی یافتن دانش از متن غیر ساخت یافته یا نیمه‌ساخت‌یافته ، مورد توجه قرار گرفته است [۱, ۲].

طبقه‌بندی داده‌ها به شیوه‌ی مناسب با میزان خطای کم و تعمیم‌پذیری بالا همواره از چالش‌های مهم شناسایی الگو[3] به خصوص در زمینه‌ی طبقه‌بندی متن بوده است.

- با افزایش حجم اطلاعات، عملیات طبقه‌بندی دستی با معایب متعددی مانند نیاز به متخصصان آن دامنه‌ی از پیش تعریف شده، پرهزینه و زمانبر بودن، در معرض خطا و سلایق افراد قرار داشتن و ضرورت تکرار فرآیند برای هر سند جدید رو به رو می‌باشد که مهم‌ترین آن‌ها عبارتند از [۷]:

با رشد گسترده‌ی اطلاعات بر خط[4] از طریق وب، شبکه‌های داخلی سازمان‌ها، منابع خبری الکترونیکی، و دیگر منابع، مسئله‌ی طبقه‌بندی خودکار اسناد متنی به کلاس‌هایی از پیش تعیین شده به موضوع مهمی در بسیاری از وظایف مدیریت و سازماندهی اطلاعات تبدیل شده است. برای مثال می‌توان موردی را در نظر گرفت که ایمیل‌ها مرتب و دسته‌بندی می‌شوند، که سرتیتر آن‌ها عملیات متناسب با آن‌ها را مشخص می‌سازد. همچنین طبقه‌بندی

[1] Meta-learning
[2] Ensemble method
[3] Pattern Recognition
[4] Online

متون می‌تواند در زمینه‌های زیادی مانند زیست فناوری[۱] برای پیش‌بینی شامل تعاملات باکتریایی پروتئین-پروتئین و طبقه‌بندی اطلاعات متن‌ها و لینک‌های وب نیز کاربرد داشته باشد [۸].

یکی از چالش‌ها در داده‌های متنی، داشتن حجم بالای مشخصه‌ها و تنک بودن ماتریس کلمه-سند است. یادگیری از داده‌هایی که مشخصه‌های زیادی دارند نه تنها باعث افزایش هزینه‌های محاسباتی می‌شود، بلکه با افزایش تعداد مشخصه‌ها، کارایی یادگیری نیز ممکن است کمتر شود. بنابراین استفاده از روش‌های مناسب انتخاب مشخصه‌های از اهمیت ویژه‌ای برخوردار است. در همین راستا، در این پژوهش روشی ترکیبی از روش‌های انتخاب مشخصه‌های فیلتر و wrapper برای حل این چالش پیشنهاد شده است.

از طرف دیگر، مسائل طبقه‌بندی می‌تواند با استفاده از الگوریتم‌های یادگیری نظارت شده برای یادگیری طبقه‌بندها از مثال‌های آموزشی برچسب دار حل شود تا بتوانند برچسب مثال جدیدی که از قبل در سند دیده نشده را پیش‌بینی کنند. برای آموزش این طبقه‌بندها با دقتی منطقی، باید تعداد کافی از مثال‌های آموزشی برچسب دار فراهم شود. برای این منظور نیاز است که فردی خبره اسناد زیادی را بخواند و تصمیم بگیرد که برچسب کلاسی را به هر کدام از آن‌ها نسبت دهد. این کار فرآیندی خسته‌کننده، زمانبر و هزینه‌بر می‌باشد. بنابراین فراهم کردن تعداد کافی از مثال‌های آموزشی برچسب دار عملی بازدارنده است. در مقابل، اسناد بدون برچسب اغلب به کمیّت زیادی قابل دسترس هستند. بنابراین، یک روش دیگرِ یادگیری استفاده از اسناد برچسب دار به همراه اسناد بدون برچسب در زمان یادگیری است، این ایده مبنای اصلی رویکرد یادگیری نیمه‌نظارتی را تشکیل می‌دهد [۹].

چارچوب یادگیری نیمه نظارتی قابل به کارگیری در طبقه‌بندی می‌باشد. در این حالت، داده‌های آموزشی می‌توانند از داده‌های اضافی بدون برچسب بهره‌برداری کنند، که اغلب منتهی به تابع طبقه‌بندی دقیق‌تر می‌شود [۱۰].

در این پژوهش نیز، روشی مبتنی بر یادگیری تجمیعی و رویکرد خودآموزی[۲] در یادگیری نیمه‌نظارتی پیشنهاد شده است که بر اساس آزمایشات انجام شده موجب بهبود کارایی یادگیری نیمه‌نظارتی در زمینه‌ی طبقه‌بندی متون شده است.

---

[۱] Bioinformatic
[۲] Self-training