

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشگاه الزهراء (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 کیوان پور
استاد مشاور
دکتر رضا عزمی

دانشجو
مریم باحجب ایمانی

اسفند ماه سال ۱۳۸۹

کلیه دستاوردهای این تحقیق متعلق به
دانشگاه الزهراء (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 بزرگواریم 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 راهنمایی های 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 بوده اند،
و از تمام کسانی که در طول دوران تحصیل با راهنمایی های ارزشمند علمی و معنوی خود، مرا یاری
نمودند سپاسگزاری نموده و از خداوند برای این عزیزان آرزوی توفیق و سربلندی می نمایم.
همچنین جا دارد از مرکز تحقیقات مخابرات ایران که از این پایان نامه تحت قرارداد شماره
۸۹۷۱/۵۰۰ حمایت کرده است، قدردانی شود.

چکیده

محبوبیت وب و حجم زیاد مستندات متنی 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. متن کاوی 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 متن کاوی 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 بر چسب 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 بر چسب 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

| | |
|----|---|
| ۱ | ۱- مقدمه |
| ۳ | ۱-۱- یادگیری ماشین |
| ۵ | ۲-۱- طرح مسئله |
| ۷ | ۳-۱- اهداف و نوآوری‌ها |
| ۸ | ۴-۱- ساختار پایان‌نامه |
| ۱۰ | ۲- پیشینه‌ی تحقیق |
| ۱۱ | ۱-۲- متن کاوی |
| ۱۴ | ۱-۱-۲- شباهت و تفاوت متن‌کاوی و داده کاوی |
| ۱۵ | ۲-۱-۲- چالش‌های متن کاوی |
| ۱۶ | ۳-۱-۲- معماری عمومی متن کاوی |
| ۱۷ | ۱-۳-۱-۲- پیش پردازش متن |
| ۱۸ | ۲-۳-۱-۲- پردازش متن |
| ۱۹ | ۳-۳-۱-۲- پس پردازش متن |
| ۲۰ | ۲-۲- یادگیری با نظارت |
| ۲۰ | ۱-۲-۲- طبقه‌بندی اسناد |
| ۲۳ | ۱-۱-۲-۲- بیزین ساده |
| ۲۴ | ۲-۱-۲-۲- Rocchio |
| ۲۵ | ۳-۱-۲-۲- درخت‌های تصمیم‌گیری |
| ۲۶ | ۴-۱-۲-۲- طبقه‌بند نزدیک‌ترین همسایه |
| ۲۷ | ۵-۱-۲-۲- Boosting |
| ۲۸ | ۷-۱-۲-۲- ماشین‌های بردار پشتیبان |
| ۳۱ | ۸-۱-۲-۲- شبکه‌های عصبی |
| ۳۱ | ۳-۲- یادگیری بدون نظارت |
| ۳۲ | ۴-۲- یادگیری نیمه‌نظارتی |
| ۳۳ | ۱-۴-۲- خودآموزی |
| ۳۳ | ۲-۴-۲- مدل‌های تولیدی |
| ۳۵ | ۱-۲-۴-۲- حداکثرسازی مورد انتظار |
| ۳۶ | ۴-۴-۲- هم‌آموزی و یادگیری چند منظری |
| ۳۸ | ۵-۴-۲- TSVM |

| | |
|----|--|
| ۳۹ | ۶-۴-۲- روش‌های مبتنی بر گراف |
| ۴۱ | Mincut - ۱-۶-۴-۲ |
| ۴۲ | ۲-۶-۴-۲- زمینه‌های گسسته‌ی مارکوف تصادفی: ماشین‌های بولتزمن |
| ۴۳ | ۳- پیش‌پردازش اسناد |
| ۴۴ | ۳-۱- مراحل پیش‌پردازش اسناد |
| ۴۵ | ۳-۱-۱- حذف تگ‌های XML |
| ۴۷ | ۳-۱-۲- Tokenization و حذف علائم نگارشی |
| ۴۸ | ۳-۱-۳- حذف Stopword ها |
| ۵۰ | ۳-۱-۴- پیدا کردن صورت استاندارد و ریشه‌یابی |
| ۵۱ | ۳-۱-۵- نمایش اسناد |
| ۵۳ | ۳-۱-۶- انتخاب مشخصه‌ها |
| ۵۶ | ۳-۱-۶-۱- رویکردهای آماری انتخاب مشخصه‌ها |
| ۵۹ | ۳-۱-۷- وزن‌دهی عبارات |
| ۶۲ | ۴- چهارچوب پیشنهادی |
| ۶۴ | ۴-۱- مرحله‌ی پیش‌پردازش |
| ۶۵ | ۴-۱-۱- الگوریتم پیشنهادی برای انتخاب مشخصه‌ها |
| ۶۶ | ۴-۱-۱-۱- گام فیلتر |
| ۶۷ | ۴-۱-۱-۲- گام wrapper |
| ۷۳ | ۴-۱-۲-۱- الگوریتم پیشنهادی ACO-GA |
| ۷۷ | ۴-۲- مرحله‌ی یادگیری (یادگیری دو سطحی) |
| ۷۸ | ۴-۲-۱- یادگیری نیمه‌نظارتی |
| ۷۸ | ۴-۲-۱-۱- یادگیری تجمیعی |
| ۸۰ | ۴-۲-۱-۲- رویکردهای ترکیب الگوریتم‌های یادگیری |
| ۸۲ | ۴-۲-۱-۳- معماری پیشنهادی جهت یادگیری نیمه‌نظارتی |
| ۸۸ | ۴-۲-۱-۳- روش وزن‌دهی پویا در الگوریتم پیشنهادی نیمه‌نظارتی |
| ۸۹ | ۴-۲-۳-۱- انتخاب الگوریتم‌های یادگیری در روش پیشنهادی نیمه‌نظارتی |
| ۹۰ | ۴-۲-۲- یادگیری با نظارت |
| ۹۳ | ۵- پیاده‌سازی و ارزیابی |
| ۹۴ | ۵-۱- پیاده‌سازی مرحله‌ی پیش‌پردازش |
| ۹۴ | ۵-۱-۱- استخراج مشخصه‌ها |

| | |
|-----|--|
| ۹۵ | tokenization -۱-۱-۱-۵ |
| ۹۶ | ۲-۱-۱-۵- ریشه‌یابی کلمات |
| ۱۰۱ | ۲-۱-۵- پیاده‌سازی روش انتخاب مشخصه‌ها |
| ۱۰۳ | ۳-۱-۵- وزن‌دهی عبارات |
| ۱۰۳ | ۲-۵- پیاده‌سازی مرحله‌ی یادگیری |
| ۱۰۳ | ۷-۲-۵- پیاده‌سازی روش پیشنهادی یادگیری نیمه‌نظارتی |
| ۱۰۵ | ۳-۵- آزمون و تحلیل نتایج |
| ۱۰۶ | ۱-۳-۵- مجموعه اسناد |
| ۱۰۸ | ۲-۳-۵- معیارهای ارزیابی |
| ۱۱۰ | ۳-۳-۵- آزمون‌ها |
| ۱۱۱ | ۱-۳-۳-۵- نتایج روش پیشنهادی انتخاب مشخصه‌ها |
| ۱۱۵ | ۲-۳-۳-۵- نتایج روش پیشنهادی یادگیری نیمه‌نظارتی |
| ۱۳۱ | ۶- نتیجه‌گیری و توسعه‌های آتی |
| ۱۳۲ | ۱-۶- نتیجه‌گیری |
| ۱۳۴ | ۲-۶- توسعه‌های آتی |
| ۱۳۵ | مراجع |

فهرست جدول‌ها

| | |
|-----|---|
| ۴۲ | جدول ۱-۲: مزایا و معایب روش‌های نیمه‌نظارتی |
| ۴۹ | جدول ۱-۳: انواع stopword ها |
| ۴۹ | جدول ۲-۳: تعدادی از واژه‌های stopword انگلیسی |
| ۵۶ | جدول ۳-۳: طبقه‌بندی رویکردهای انتخاب مشخصه‌ها |
| ۶۰ | جدول ۴-۳: نمونه‌ای وزن‌دهی دودویی |
| ۹۰ | جدول ۱-۴: مزایا، معایب و پیچیدگی زمانی الگوریتم‌های رایج طبقه‌بندی متون |
| ۱۰۷ | جدول ۱-۵: ۱۰ دسته‌ی اصلی پیکره‌ی رویترز با تقسیم‌بندی استاندارد ModeApte |
| ۱۰۹ | جدول ۲-۵: جدول احتمال وقوع برای دسته‌ی C_i |
| ۱۰۹ | جدول ۳-۵: جدول احتمال وقوع برای کل دسته‌ها |
| ۱۱۱ | جدول ۴-۵: مقادیر پارامترهای متفاوت الگوریتم ژنتیک |
| ۱۱۲ | جدول ۵-۵: مقادیر پارامترهای الگوریتم پیشنهادی ACO-GA و الگوریتم‌های مورد مقایسه |
| ۱۱۲ | جدول ۶-۵: نتایج ارزیابی معیارهای یادآوری و دقت روش‌های انتخاب مشخصه‌ی ACO, GA, CHI, IG و روش پیشنهاد شده بر روی پیکره‌ی رویترز |
| ۱۱۳ | جدول ۷-۵: نتایج ارزیابی معیارهای Macro-F1 و Micro-F1 با روش‌های انتخاب مشخصه‌ی GA, CHI, IG, ACO و روش پیشنهاد شده بر روی پیکره‌ی رویترز |
| ۱۲۶ | جدول ۸-۵: میزان F1 برای هر دسته از Reuters در درصد مشخصی از داده‌های برجسپ‌دار (L) |

فهرست شکل‌ها

- شکل ۱-۲: معماری پیشنهادی متن کاوی ۱۷
- شکل ۲-۲: طبقه‌بندی دودویی سخت ۲۱
- شکل ۳-۲: طبقه‌بندی چند کلاسه‌ی تک برچسبی ۲۱
- شکل ۴-۲: طبقه‌بندی چند کلاسه‌ی چند برچسبی ۲۲
- شکل ۵-۲: طبقه‌بندی چند کلاسه‌ی نرم ۲۲
- شکل ۶-۲: نحوه‌ی عملکرد الگوریتم Rocchio در طبقه‌بندی متون ۲۴
- شکل ۷-۲: مثالی از الگوریتم K-NN ۲۶
- شکل ۸-۲: ابرصفحه‌ای با حداکثر فاصله حاشیه به مثال‌های کلاس‌های مثبت و منفی ساخته‌شده با ماشین بردار پشتیبان ۳۰
- شکل ۹-۲: اگر مدل اشتباه باشد، درست‌نمایی بالا می‌تواند به دقت کم طبقه‌بندی منتهی شود ۳۵
- شکل ۱۰-۲: الگوریتم حداکثر سازی مورد انتظار ۳۶
- شکل ۱۱-۲: هم‌آموزی ۳۷
- شکل ۱۲-۲: TSVM ۳۹
- شکل ۱۳-۲: تابع باخت TSVM ۳۹
- شکل ۱۴-۲: یادگیری نیمه‌نظارتی مبتنی بر گراف ۴۰
- شکل ۱-۳: فاز پیش‌پردازش اسناد ۴۶
- شکل ۲-۳: نمونه‌ای از نمایش سند XML ۴۷
- شکل ۳-۳: نمایش متن شکل ۳-۳ پس از حذف تگ‌های XML که دارای برچسب کلاس weather می‌باشد ۴۷
- شکل ۴-۳: کلمات شکل ۳-۴ پس از عملیات tokenization ۴۸
- شکل ۵-۳: شکل سمت راست مثالی را نمایش می‌دهد که در آن کلمات stopword از شکل سمت چپ حذف شده است. ۵۰
- شکل ۶-۳: کلمات اصلی در شکل (سمت چپ) و ریشه‌ی آن‌ها با استفاده از الگوریتم پورتر (سمت راست) ۵۱
- شکل ۷-۳: چهارچوب رویکرد wrapper در انتخاب مشخصه‌ها ۵۴
- شکل ۸-۳: چهارچوب رویکرد فیلتر در انتخاب مشخصه‌ها ۵۴
- شکل ۹-۳: نمایش فرکانس عبارت یک سند ۶۰
- شکل ۱۰-۳: ماتریس کلمه-سند نمایشگر مجموعه اسناد ۶۱
- شکل ۱-۴: چهارچوب کلی روش پیشنهادی ۶۳
- شکل ۲-۴: چهارچوب کلی مرحله‌ی یادگیری ۶۴
- شکل ۳-۴: چهارچوب کلی مرحله‌ی پیش‌پردازش اسناد ۶۵
- شکل ۴-۴: مولفه‌های روش توکار پیشنهادی برای انتخاب مشخصه‌ها ۶۶

- شکل ۴-۵: ساختار کروموزم ۷۱
- شکل ۴-۶: مثالی از عملیات crossover دو نقطه‌ای ۷۲
- شکل ۴-۷: فلوجارت الگوریتم پیشنهادی ۷۶
- شکل ۴-۸: شمای مرحله‌ی یادگیری دو سطحی ۷۷
- شکل ۴-۹: شمای کلی یادگیری تجمیعی با روش ائتلاف طبقه‌بندها ۷۹
- شکل ۴-۱۰: معماری یک مرحله‌ای پیشنهاد شده در یادگیری نیمه‌نظارتی ۸۳
- شکل ۴-۱۱: معماری دو مرحله‌ای پیشنهاد شده در یادگیری نیمه‌نظارتی ۸۵
- شکل ۴-۱۲: معماری کلی پیشنهاد شده در یادگیری نیمه‌نظارتی ۸۷
- شکل ۴-۱۳: وزن‌دهی پویا برای یکی از الگوریتم‌های یادگیری تجمیعی ۸۸
- شکل ۵-۱: شبه کد استخراج مشخصه‌ها از اسناد ۹۵
- شکل ۵-۲: شبه کد عملیات tokenization ۹۶
- شکل ۵-۳: شبه کد قوانین گام ۱ در روش ریشه‌یابی پورتر ۹۸
- شکل ۵-۴: شبه کد قوانین گام ۲ در روش ریشه‌یابی پورتر ۹۹
- شکل ۵-۵: شبه کد قوانین گام ۳ در روش ریشه‌یابی پورتر ۹۹
- شکل ۵-۶: شبه کد قوانین گام ۴ در روش ریشه‌یابی پورتر ۱۰۰
- شکل ۵-۷: شبه کد قوانین گام ۵ در روش ریشه‌یابی پورتر ۱۰۰
- شکل ۵-۸: الگوریتم توکار پیشنهادی برای انتخاب مشخصه‌های متنی ۱۰۱
- شکل ۵-۹: الگوریتم محاسبه‌ی آماره‌ی χ^2 ۱۰۱
- شکل ۵-۱۰: الگوریتم پیشنهادی بخش wrapper که ترکیبی از ACO و GA است ۱۰۲
- شکل ۵-۱۱: شبه کد وزن‌دهی به کلمات (عبارات) با روش TF-IDF ۱۰۳
- شکل ۵-۱۲: الگوریتم پیشنهادی در یادگیری نیمه‌نظارتی ۱۰۴
- شکل ۵-۱۳: الگوریتم وزن‌دهی پویا ۱۰۵
- شکل ۵-۱۴: الگوریتم نرمال‌سازی وزن‌ها ۱۰۵
- شکل ۵-۱۵: نمایش نمونه‌ای از اسناد پیکره‌رویترز ۱۰۷
- شکل ۵-۱۶: نمودار میله‌ای نتایج حاصل از معیارهای Micro-F1 و Macro-F1 ۱۱۳
- شکل ۵-۱۷: نمودار Macro-F1 و Micro-F1 در انتخاب مشخصه‌ها بر اساس درصد مشخصه‌ها با روش‌های IG، CHI، GA، ACO و روش پیشنهاد شده بر روی پیکره‌ی رویترز ۱۱۴
- شکل ۵-۱۸: نمودار F1-Micro و F1-Macro برای مقایسه‌ی معماری‌های یک مرحله‌ای ۱۱۷
- شکل ۵-۱۹: نمودار F1-Micro و F1-Macro برای مقایسه‌ی وزن‌دهی‌های معماری یک مرحله‌ای ۱۱۸
- شکل ۵-۲۰: نمودار F1-Micro و F1-Macro برای مقایسه‌ی معماری‌های دو مرحله‌ای ۱۱۹

- ۱۲۱ شکل ۵-۲۱: نمودار F1-Micro و F1-Macro برای مقایسه‌ی وزن‌دهی‌های متفاوت معماری دو مرحله‌ای
- ۱۲۲ شکل ۵-۲۲: نمودار F1-Micro و F1-Macro برای مقایسه‌ی معماری‌های سه مرحله‌ای
- ۱۲۳ شکل ۵-۲۳: نمودار F1-Micro و F1-Macro برای مقایسه‌ی وزن‌دهی‌های متفاوت معماری سه مرحله‌ای
- ۱۲۵ شکل ۵-۲۴: نمودار F1-Micro و F1-Macro برای مقایسه‌ی معماری‌های یک، دو و سه مرحله‌ای
- ۱۲۸ شکل ۵-۲۵: ارزیابی ۱۰ دسته‌ی معروف مجموعه‌ی Reuters-21578 با معماری پیشنهادی نیمه نظارتی به وسیله‌ی معیار F1
- ۱۲۹ شکل ۵-۲۶: نمودار F1-Micro مقایسه‌ی روش پیشنهادی و روش‌های دیگر یادگیری نیمه نظارتی
- ۱۳۰ شکل ۵-۲۷: نمودار F1-Macro مقایسه‌ی روش پیشنهادی و روش‌های دیگر یادگیری نیمه نظارتی

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 متن کاوی مانند [۵، ۴، ۲] 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

^۴ Data Mining

^۵ Machine Learning

^۶ Data warehouse

^۷ Visualization

^۸ Feature Extraction

نیمه‌ساخت‌یافته در مجموعه‌ها اسناد، به صورت ساخت‌یافته می‌باشند (مانند خالص کردن داده، حذف لغات stopword، ریشه‌یابی^۱، شاخص‌گذاری^۲، پالایش متن، تولید مشخصه‌ها و کاهش فضای مشخصه‌ها). مرحله‌ی 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 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 نگاشتنی از داده‌های ورودی به خروجی مشخص شده 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

^۵ Bioinformatic

• روش‌های فرا یادگیری^۱ و تجمیعی^۲:

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.

۲-۱- 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 نیمه‌ساخت یافته 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 نیمه‌ساخت یافته، 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^۳ 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 سلايق 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^۴ 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

^۱ Meta-learning

^۲ Ensemble method

^۳ Pattern Recognition

^۴ 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 بازنارنده است. در مقابل، اسناد بدون برچسب اغلب به کمیّت زیادی قابل دسترس هستند. بنابراین، یک روش دیگر 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