

بسم الله الرحمن الرحيم



وزارت علوم، تحقیقات و فناوری
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

طرح پژوهشی

استخراج فراداده در پارساها با رویکرد الگوریتم‌های دسته‌بندی

مجری

جلال‌الدین نصیری

همکاران

نصراله پاک‌نیت

اردیبهشت ۱۴۰۰

حقوق: پژوهشگاه علوم و فناوری اطلاعات ایران

طرح پژوهشی استخراج فراداده در پارساها با رویکرد الگوریتم‌های دسته‌بندی پیرو قرارداد شماره ۳۳/۴۶۵۰ مورخ ۹۹/۰۶/۰۴ میان پژوهشگاه علوم و فناوری اطلاعات ایران (کارفرما) و آقای دکتر جلال‌الدین نصیری اجرا شده است. گزارش حاضر یک جلد از مستندات این طرح است.

این گزارش و تمامی حقوق مادی آن بر اساس قانون حمایت حقوق مولفان و مصنفان و هنرمندان، مصوب ۱۳۴۸ و اصلاحیه‌های بعدی آن و همچنین آیین‌نامه‌های اجرایی این قانون متعلق به پژوهشگاه علوم و فناوری اطلاعات ایران و هر گونه استفاده از تمامی یا پاره‌ای از آن، شامل: نقل قول، تکثیر، انتشار، کاربرد نتایج، تکمیل و مانند آنها به صورت چاپی، الکترونیکی یا وسایل دیگر فقط با اجازه کتبی پژوهشگاه امکان‌پذیر است. نقل قول در حد هزار واژه در انتشارات علمی مانند کتاب و مقاله با درج اطلاعات کامل کتابشناختی، نیازی به مجوز پژوهشگاه ندارد. صحت مندرجات گزارش بر عهده مجری طرح پژوهشی است.



به نام خدا

ایرانداک در نیم قرن دوم کار خود همپایان جوان و کوشا؛ همراه هر پژوهش، پژوهشگر، و سیاست‌گذار

۱۳۴۷ - ۱۴۰۰

گواهی

انجام طرح پژوهشی:

- ◇ عنوان: استخراج فراداده در پارساها با رویکرد الگوریتم‌های دسته‌بندی
- ◇ مجری: دکتر جلال‌الدین نصیری
- ◇ همکار: دکتر نصرالله پاک‌نیت
- ◇ شماره قرارداد: ۳۳/۴۶۵۰
- ◇ تاریخ قرارداد: ۱۳۹۹/۶/۴
- ◇ تاریخ شروع: ۱۳۹۹/۶/۴
- ◇ تاریخ پایان: ۱۳۹۹/۱۲/۱۸
- ◇ ناظر: دکتر الهام علایی ابوذر

در پژوهشگاه علوم و فناوری اطلاعات ایران گواهی می‌شود. گزارش پایانی این طرح، بر پایه داوری در نشست ۴۰۹ (تاریخ ۱۴۰۰/۲/۱۸) شورای پژوهشی پژوهشگاه به تصویب رسیده است. لازم می‌دانم مراتب قدردانی خود را به مجری، همکار، و ناظر این طرح تقدیم دارم.

با آرزوی بهروزی
پرکیز شهریاری
معاون پژوهش و آموزش

استخراج فراداده در پارساها با رویکرد الگوریتم‌های دسته‌بندی

مدیر طرح پژوهشی: دکتر جلال‌الدین نصیری، پژوهشگاه علوم و فناوری اطلاعات ایران.

نشانی: تهران، خیابان انقلاب، چهارراه فلسطین، شماره ۱۰۹۰، طبقه ۴، اتاق ۴۲۰

صندوق پستی: ۱۳۱۵۷۷۳۳۱۴ تلفن: ۰۲۱-۶۶۴۹۴۹۸۰ (داخلی ۴۳۲)

وبگاه: <http://irandoc.ac.ir/nasiri/nasiri.html>

رایانامه: j.nasiri@irandoc.ac.ir

همکاران: نصراله پاک‌نیت

چکیده

فراداده¹ داده‌ای است که اطلاعاتی درباره‌ی داده‌ای دیگر ارائه می‌دهد. عنوان، نام دانشگاه، رشته تحصیلی، نام دانشجو، استاد راهنما، چکیده و کلمات کلیدی را می‌توان از مهم‌ترین فراداده‌های پایان‌نامه‌ها و رساله‌ها (پارساها) نام برد. از آنجاییکه فراداده‌ها در بازیابی اطلاعات نقش منحصر به فردی را ایفا می‌نمایند، استخراج صحیح و ذخیره آن‌ها در سامانه‌های بازیابی اطلاعات حائز اهمیت است. با توجه به حجم بالای متون علمی و سرعت پایین کنترل دستی فراداده در سامانه ثبت پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، همواره نمایه‌سازی پارساهای دریافتی عقب‌تر از پارساهای دریافتی بوده و در نتیجه نمایش اطلاعات در سامانه گنج ایرانداک به روز نیست. با توجه به این مساله، افزایش سرعت کنترل فراداده‌ها از موارد اولویت‌دار سامانه ثبت است.

هدف اصلی این طرح پژوهشی طراحی الگوریتمی است که بتواند فراداده‌ها را از متن کامل پارساها استخراج نماید. چنین الگوریتمی می‌تواند با بررسی سیستمی و هوشمند فراداده‌های پارساها، سرعت نمایه‌سازی پارساها در پایگاه ثبت را افزایش داده و منجر به به‌روز شدن پایگاه گنج گردد.

برای انجام این مهم پس از پیش‌پردازش‌های رایج مانند نرمال‌سازی متن، حذف ایست‌واژه‌ها و قطعه‌بندی در پردازش زبان فارسی، از هر پاراگراف ویژگی‌های ساختاری و زبانی استخراج می‌گردد. استخراج ویژگی باید به گونه‌ای طراحی گردد که فراداده‌های متفاوت قابل تمیز دادن از هم باشند. در مرحله بعد با استفاده روش‌های هوشمند دسته‌بندی چند کلاسه، روش پیشنهادی آموزش داده می‌شود. در این طرح پژوهشی برای مکاشفه بهتر و دستیابی به مقادیر بهینه دسته‌بند از الگوریتم‌های فرامکاشفه‌ای استفاده خواهد شد. ارزیابی روش پیشنهادی نشان می‌دهد در بسیاری از اقلام اطلاعاتی فراداده استخراج هوشمند و خودکار کارا بوده و امکان عملیاتی شدن ایده وجود دارد.

کلیدواژه‌ها:

دسته‌بندی چند کلاسه؛ نمایه‌سازی خودکار؛ استخراج فراداده؛ بازیابی اطلاعات

¹ Metadata

فهرست مطالب

عنوان	شماره صفحه
۱. فصل اول: کلیات.....	۱
۱-۱. مقدمه	۲
۲-۱. بیان مساله	۳
۳-۱. محدودیت‌های پژوهش	۴
۴-۱. نوآوری.....	۷
۵-۱. ساختار طرح پژوهشی	۷
۲. فصل دوم: مروری بر پیشینه پژوهش	۹
۱-۲. مقدمه.....	۱۰
۲-۲. پردازش زبان طبیعی.....	۱۰
۱-۲-۲. کاربردهای پردازش زبان طبیعی	۱۱
۲-۲-۲. پیش پردازش متن	۱۲
۳-۲. مفاهیم پایه در حوزه استخراج اطلاعات	۱۵
۱-۳-۲. هدف استخراج اطلاعات	۱۵
۲-۳-۲. وظایف سیستم استخراج اطلاعات	۱۶
۳-۳-۲. استخراج اطلاعات از اسناد علمی.....	۱۷
(۱) فراداده	۱۸
(۲) بینش‌های کلیدی.....	۱۸
۴-۲. پردازش سند	۱۹
۱-۴-۲. مراحل پردازش سند	۱۹

۲۰	۲-۴-۲. پردازش روابط منطقی در ساختار سند.....
۲۱	۵-۲. مطالعه پیشینه پژوهش در حوزه استخراج اطلاعات از نوشتارهای علمی
۲۲	۱-۵-۲. رویکردهای استخراج اطلاعات از نوشتارهای علمی
۲۳	رویکرد مبتنی بر قاعده.....
۴۲	رویکرد مبتنی بر یادگیری ماشین.....
۴۰	۲-۵-۲. پژوهش‌های انجام شده در زمینه استخراج بینش‌های کلیدی از مقالات علمی.....
۴۸	۶-۲. مروری بر الگوریتم‌های به‌کارگرفته پژوهش.....
۴۸	۱-۶-۲. ماشین بردار پشتیبان.....
۵۱	۲-۶-۲. الگوریتم خفاش.....
۵۳	۳-۶-۲. ترکیب ماشین بردار پشتیبان و الگوریتم خفاش.....
۵۵	۷-۲. جمع‌بندی.....
۵۷	۳. فصل سوم: روش پیشنهادی.....
۵۸	۱-۳. مقدمه.....
۵۸	۲-۳. روش پیشنهادی در استخراج فراداده‌ها.....
۵۹	۱-۲-۳. پیش‌پردازش.....
۶۱	۱-۲-۳. تشخیص مرز بدنه.....
۶۸	۲-۲-۳. طبقه‌بندی فراداده‌ها.....
۷۴	۱-۲-۳. پس‌پردازش.....
۷۵	۳-۳. جمع‌بندی.....
۷۷	۴. فصل چهارم: نتایج تجربی.....
۷۸	۱-۴. مقدمه.....
۷۸	۲-۴. مجموعه داده.....
۷۹	۳-۴. معیارهای ارزیابی.....
۸۱	۴-۴. ارزیابی مرحله تشخیص بدنه.....
۸۹	۱-۴-۴. ارزیابی مرحله طبقه‌بندی فراداده.....
۹۴	۲-۴-۴. ارزیابی با در نظر گرفتن جداسازی خودکار بدنه.....
۹۴	۵-۴. جمع‌بندی.....

۵. فصل پنجم: نتیجه گیری ۹۶

فهرست منابع و مآخذ ۱۰۲

فهرست جدول‌ها

عنوان	شماره صفحه
جدول ۱-۲. تفاوت رویکردهای استخراج اطلاعات از نوشتارهای علمی	۲۳
جدول ۲-۲. روش‌های استخراج فراداده مبتنی بر تعریف قواعد	۳۱
جدول ۳-۲. روش‌های استخراج فراداده با رویکرد یادگیری ماشین	۳۹
جدول ۴-۲. روش‌های استخراج بینش‌های کلیدی از مقالات	۴۷
جدول ۱-۳. ویژگی‌های استخراج شده در مرحله‌ی تشخیص مرز بدنه	۶۶
جدول ۲-۳. فهرست ویژگی‌های در نظر گرفته شده در مرحله طبقه‌بندی فراداده‌ها	۷۱
جدول ۳-۳. ادامه فهرست ویژگی‌های مستخرج در مرحله طبقه‌بندی فراداده‌ها	۷۲
جدول ۱-۴. نتایج به دست آمده در تنظیم پارامترهای SVM در مرحله تشخیص مرز به روش جستجوی شبکه‌ای	۸۲
جدول ۲-۴. مقایسه‌ی زمان اجرا و خطای آزمایش به دست آمده در الگوریتم خفاش و جستجوی شبکه‌ای	۸۵
جدول ۳-۴. مقایسه تاثیر انواع ویژگی‌ها بر عملکرد طبقه‌بند پاراگراف‌ها در مرحله تشخیص مرز بدنه	۸۶
جدول ۴-۴. نتایج به دست آمده در گام‌های اعتبارسنجی متقابل 10-fold مرحله‌ی تشخیص مرز	۸۷
جدول ۵-۴. مقایسه عملکرد طبقه‌بند در انواع پایان‌نامه‌ها/رساله‌ها در مرحله تشخیص مرز	۸۸
جدول ۶-۴. نتایج به دست آمده در تنظیم پارامترهای SVM در مرحله طبقه‌بندی فراداده به روش جستجوی شبکه‌ای	۸۹
جدول ۷-۴. مقایسه‌ی تاثیر انواع ویژگی‌ها بر عملکرد طبقه‌بند فراداده‌ها	۹۱
جدول ۸-۴. نتایج به دست آمده در گام‌های اعتبارسنجی متقابل 10-fold مرحله تشخیص مرز	۹۱
جدول ۹-۴. مقایسه عملکرد طبقه‌بند در انواع پایان‌نامه‌ها/رساله‌ها در مرحله طبقه‌بندی فراداده	۹۳

فهرست شکل‌ها

عنوان	شماره صفحه
شکل ۱-۲ فلوجارت بهینه سازی ماشین بردار پشتیبان با الگوریتم خفاش (Tharwat, Hassanien, & Elnaghi, 2017).....	۵۵
شکل ۱-۳ گام‌های پژوهش. ابتدا مدلی برای شناسایی بدنه و فراداده ساخته می‌شود. سپس از این مدل برای تشخیص مرز بدنه استفاده شده و در ادامه فراداده‌ها طبقه‌بندی می‌شوند.....	۵۹
شکل ۲-۳ نسبت تعداد پاراگراف‌های پیشگفتار به تعداد کل پاراگراف‌های طرح پژوهشی در داده‌های موجود.....	۶۰
شکل ۳-۳ چند نمونه از مرز فراداده و بدنه در پایان‌نامه. در شکل‌های (الف)، (ب) و (پ) به ترتیب عبارات‌های «فصل ۱:»، «فصل اول» و «کلیات تحقیق» شروع بدنه پایان‌نامه هستند.....	۶۳
شکل ۳-۴ نمونه‌ای از جداسازی بلوک‌ها در بخشی از پایان‌نامه. علامت ¶ نشانه‌ی انتهای پاراگراف است.....	۷۰
شکل ۵-۳ دو نمونه از شکسته شدن یک فراداده به چند پاراگراف. در شکل (الف) و (ب) به ترتیب عنوان طرح پژوهشی و استاد مشاور در دو پاراگراف نوشته شده‌اند. علامت ¶ نشانه‌ی انتهای پاراگراف است.....	۷۰
شکل ۶-۳ در شکل (الف) مقطع و رشته‌ی تحصیلی در دو پاراگراف نوشته شده، در حالی که در نمونه شکل (ب) رشته و مقطع در یک پاراگراف آمده‌اند. علامت ¶ نشانه‌ی انتهای پاراگراف است.....	۷۴
شکل ۱-۴ فراوانی انواع پایان‌نامه‌ها/رساله‌ها.....	۷۹
شکل ۲-۴ آزمایش تاثیر تعداد خفاش بر عملکرد الگوریتم : (الف) نرخ خطای آزمایش و (ب) زمان اجرا به ازای مقادیر مختلف خفاش‌ها.....	۸۳
شکل ۳-۴ آزمایش تاثیر تعداد خفاش بر عملکرد الگوریتم : (الف) نرخ خطای آزمایش به تعداد تکرارهای مختلف. (ب) زمان اجرا به ازای افزایش تعداد تکرار.....	۸۴
شکل ۴-۴ نمودار کانتوری نرخ خطا به ازای پارامترهای خطا (c) و سیگما. نقطه‌ای که با ستاره مشخص شده نقطه مینیمی است که به کمک الگوریتم خفاش به دست آمده است.....	۸۴
شکل ۵-۴ ماتریس درهم ریختگی طبقه‌بند فراداده و بدنه.....	۸۸
شکل ۶-۴ ماتریس درهم ریختگی طبقه‌بند فراداده‌ها (نرمال شده).....	۹۲
شکل ۷-۴ نمودار مقایسه (الف) صحت و (ب) بازخوانی در فراداده‌های مختلف.....	۹۴

قدردانی

اکنون که به یاری خدا استخراج فراداده در پارساها با رویکرد الگوریتم‌های دسته‌بندی را به پایان رسانده‌ام از زحمات همکاران به‌ویژه جناب آقای دکتر علیدوستی ریاست گرامی پژوهشگاه و همچنین اعضای گرامی پژوهشکده علوم اطلاعات تشکر و قدردانی می‌نمایم. همچنین از ناظر محترم طرح سرکار خانم دکتر علائی که با نظرات ارزشمند خود بهبود این طرح را تضمین نموده‌اند، تشکر می‌نمایم.

هم‌چنین سزاوار است از اعضای محترم شورای پژوهشی برای حمایت‌هایشان، سپاسگزاری نمایم.

فصل اول

کلیات

۱-۱. مقدمه

«پژوهشگاه علوم و فناوری اطلاعات ایران» (ایرانداک) که از سال ۱۳۸۸ به این نام خوانده می‌شود، بر پایه گاه‌شمار و همچنین نخستین سند در این زمینه، در یکم مهر ماه سال ۱۳۴۷ با نام «مرکز اسناد ایران» تأسیس شده است. مأموریت بنیادین ایرانداک بر پایه اساس‌نامه و برنامه استراتژیک آن؛ پژوهش، مدیریت دانش، آموزش، همکاری‌های پژوهشی و اطلاع‌رسانی، و پشتیبانی از سیاست‌گذاری علم و فناوری است.

پژوهشگاه علوم و فناوری اطلاعات با جمع‌آوری و پردازش تمام پایان‌نامه‌ها و رساله‌های وزارت علوم، تحقیقات و فناوری از جایگاه ممتاز و روبه‌رشدی برای دانشجویان، اساتید و مدیران برخوردار شده است. دستاوردهای مدیریت دانش و اطلاعات علمی و فنی کشور در سامانه‌ها و پایگاه‌های اطلاعاتی ارائه می‌شوند. در این میان، پایگاه اطلاعات گنج با صدها هزار رکورد؛ از بزرگ‌ترین، دیرین‌ترین، و پربازدیدترین پایگاه‌های اطلاعات علمی و فنی کشور است. اطلاعات پایان‌نامه‌ها و رساله‌ها و پیشنهادیه آن‌ها که در ایرانداک به ثبت می‌رسند، از ارزش بسیار زیادی برخوردار بوده و ارائه سرویس‌های ارزش افزوده از این داده‌ها به سیاست‌گذاران و تصمیم‌گیرندگان حوزه علم و فناوری و امکان تجاری‌سازی دانش در جهت تثبیت نقش مرجعیت پژوهشگاه می‌تواند مفید باشد.

با توجه به حجم بالای متون علمی و سرعت پایین کنترل اقلام اطلاعاتی در سامانه ثبت و در نتیجه عدم نمایش به روز در سامانه گنج، افزایش سرعت کنترل اقلام اطلاعاتی از موارد اولویت‌دار سامانه

ثبت است. به عنوان مثال، در حال حاضر حدود ۱۴۰ هزار متن علمی همچنان نمایه نشده وجود دارد و با توجه به سیل پایان‌نامه‌ها و رساله‌های جدید، امکان نمایه شدن سنتی آن‌ها وجود نداشته و استفاده از روش‌هایی برای افزایش سرعت نمایه‌سازی از اولویت‌های سامانه ثبت است. استخراج و تایید اقلام اطلاعاتی پایان‌نامه‌ها و رساله‌ها که به آن‌ها فراداده^۲ گفته می‌شود از وظایف نمایه‌ساز می‌باشد. در صورتی که بتوان کنترل، ورود یا تایید فراداده را بوسیله ماشین انجام داد سرعت کنترل اقلام اطلاعاتی افزایش خواهد یافت. از جمله این اقلام اطلاعاتی می‌توان به نام دانش‌آموخته، اساتید راهنما و مشاور، تاریخ و ... اشاره نمود. تجزیه و تحلیل اسناد با هدف دست یافتن به اطلاعات مهم و نهفته‌ی آن‌ها از اهمیت ویژه‌ای برخوردار است.

۲-۱. بیان مساله

استخراج خودکار فراداده‌های پایان‌نامه‌ها و رساله‌ها به معنای استخراج اطلاعات کلیدی و مهم این اسناد از آن‌ها است. نام دانش‌آموخته، نام اساتید راهنما و مشاور، دانشگاه محل تحصیل، رشته تحصیلی، چکیده و کلمات کلیدی مثال‌هایی از این فراداده‌ها هستند.

فراداده‌ها و بینش‌های کلیدی^۳ دو دسته اطلاعات نهفته در اسناد علمی هستند (Nasar, Jaffry, & Malik, 2018). فراداده‌ها ممکن است از سرآیند^۴ اسناد و یا رشته‌های متنی مراجع استخراج شوند. علاوه بر فراداده‌ها، بینش‌های کلیدی اطلاعات بالقوه‌ی دیگری در اسناد علمی هستند که از جمله آن‌ها می‌توان به نوع مسئله، حوزه‌ی مورد مطالعه، روش‌های به کار گرفته شده در حل مسئله، ارزیابی و نتایج اشاره کرد. تمرکز این طرح پژوهشی بر فراداده‌هایی است که در سرآیند پایان‌نامه‌ها و رساله‌ها

^۲ Metadata

^۳ Key insights

^۴ Header

ظاهر می‌شوند. لازم به ذکر است که دستیابی به اطلاعات سایر قسمت‌ها مانند فصل‌های اصلی و محتوای بدنه پایان‌نامه و رساله، از جمله فهرست مراجع از اهداف این پژوهش نیست.

۳-۱. محدودیت‌های پژوهش

برای استخراج اطلاعات از اسناد، می‌توان از محتوای متنی، تصویر سند و یا به صورت ترکیبی از هر دو امکان برای استخراج اطلاعات استفاده کرد. به بیان دیگر استخراج اطلاعات با سه رویکرد مبتنی بر متن، مبتنی بر تصویر و یا ترکیبی از متن و تصویر ممکن است (Toledo, Carbonell, Fornés, & Lladós, 2019). با توجه به امکان به‌کارگیری، رویکرد مبتنی بر متن، رویکردی است که در مقایسه با سایر رویکردها بیشتر مورد توجه قرار گرفته است. با این حال، در مورد این رویکرد نیز چالش‌های زیادی وجود دارد که چالش‌های زبانی، چالش‌های موجود در شناسایی مرز کلمات و چالش‌های متون علمی مهم‌ترین چالش‌های موجود در این زمینه هستند.

چالش‌های زبانی

همانطور که ذکر شد، در استخراج اطلاعات با رویکرد مبتنی بر متن تنها متن سند پردازش می‌شود. با توجه به این مهم، این روش‌ها اغلب با برخی چالش‌های زبانی روبه‌رو هستند که از مهم‌ترین آن‌ها می‌توان به موارد زیر – که در همه‌ی زبان‌ها مشترک هستند – اشاره کرد:

- به کار بردن شکل اختصاری اصطلاحات
- انواع هجی کلمات؛ مثلاً کلمات «تهران» و «طهران» دو املائی متفاوت برای یک واژه است.
- ممکن است در زبان طبیعی ابهام وجود داشته باشد؛ برای مثال در زبان فارسی برای واج /س/ چندین حرف وجود دارد (Dashtipour, Gogate, Adeel, Algarafi, Howard, & Hussain, 2017).
- منسوخ شدن برخی واژگان در اسناد قدیمی (Toledo, Carbonell, Fornés, & Lladós, 2019).

علاوه بر چالش‌های عمومی ذکر شده، چالش‌های دیگری نیز وجود دارند که در مورد زبان فارسی صادق هستند:

- کمبود مجموعه داده‌های برچسب‌دار و عمومی (Hafezi & Rezaeian, 2018)
- تفاوت ترتیب کلمات در جملات فارسی و جملات انگلیسی؛ برای نمونه فعل در زبان فارسی در انتهای جمله می‌آید. (Dashtipour, Gogate, Adeel, Algarafi, Howard, & Hussain, 2017)
- گاهی در زبان فارسی می‌توان فاعل را حذف کرد که در چنین شرایطی با توجه به فعل می‌توان فاعل را تشخیص داد.
- در فعل‌های مرکب نمی‌توان بخشی از آن را حذف کرد، زیرا معنا تغییر می‌یابد؛ مانند فعل «به کار بردن» (Rahat, Talebpour, & monemian, 2018)

چالش‌های موجود در شناسایی مرز کلمات

- به دلیل ویژگی‌های خاص زبان فارسی، کاراکتر فاصله به تنهایی نمی‌تواند نشان‌دهنده مرز واحدهای زبانی باشد. برای مثال اگر به واژه «کتاب» نشانه جمع «ها» افزوده شود، این کلمه را می‌توان به سه صورت «کتابها»، «کتاب ها» و «کتاب‌ها» نوشت؛ به طوری که این کلمات به ترتیب بدون فاصله، با فاصله و با نیم‌فاصله نوشته شده‌اند.
- نشانه‌ها و علائم عربی: به‌طور کلی در زبان فارسی، نشانه‌گذاری امری رایج نیست؛ برای مثال واژه‌های «در» و «دَر» هر دو معادل هستند. اما برخی نشانه‌های عربی مانند تنوین و همزه وارد زبان فارسی شده‌اند که موجب ایجاد تفاوت‌های املائی در نگارش برخی کلمات می‌شود: مانند «املاء» و «املا».
- جمع‌های بی‌قاعده: به جز نشانه‌های جمع فارسی، جمع برخی کلمات قاعده‌ی خاصی ندارد. چنین کلماتی که عمدتاً ریشه‌ی عربی دارند به غیر از نشانه‌های جمع، به صورت بی‌قاعده نیز

جمع بسته می‌شوند. به عنوان مثال می‌توان به کلمه «مناطق» که جمع واژه «منطقه» است، اشاره کرد.

- وجود کاراکترهای در ظاهر مشابه اما با یونی‌کد^۵ متفاوت از دیگر چالش‌های پردازش زبان فارسی است. برای مثال واج‌های /ی/ و /ک/ بیش از یک یونی‌کد دارند و این مهم سبب بروز ابهام در شناسایی حروف فارسی و عربی می‌شود (Mohseni, Ghofrani, & Faili, 2018).

چالش‌های استخراج اطلاعات از اسناد علمی

علاوه بر چالش‌های فوق که در مورد هر متنی صادق هستند، برخی چالش‌ها نیز وجود دارند که مختص متون علمی هستند:

- در صورتی که نیاز به تبدیل اسناد از قالبی به قالب دیگر باشد، ممکن است طی این فرآیند اطلاعات و فراداده‌ها ناخواسته حذف شوند. برای مثال دستیابی به محتوای متنی فایل PDF ممکن است موجب فقدان اطلاعات مهم شود. (Nasar, Jaffry, & Malik, 2018)
- تنوع در طرح‌بندی و سبک نگارش مقالات موجب می‌شود تا اطلاعات مشابه در اسناد علمی، در بخش‌های متفاوتی از صفحه قرار گیرند.
- اکثر اسناد علمی موجود در قالب PDF هستند. بنابراین دستیابی به اطلاعات ساختاری سند چون کلمات، پاراگراف‌ها، فهرست‌ها، ساختار جداول، سلسله‌مراتب بخش‌ها و ترتیب خواندن متن نیازمند پردازش ساختاری اسناد PDF است (Tkaczyk, Szostek, Fedoryszak, Dendek, & Bolikowski, 2015).

- در استخراج فراداده از رشته‌های مراجع، این مشکل وجود دارد که مراجع به شیوه‌های گوناگونی نوشته می‌شوند. برای مثال در یک شیوه ارجاعی ممکن است نام نویسندگان به

⁵ Unicode

طور کامل قید شود، در حالی که در روش دیگری نام نویسندگان کوتاه شده و خلاصه درج می‌شود یا امکان دارد که نام نویسندگان قبل از نام خانوادگی بیان شود، در مقابل در سبک دیگری باید نام خانوادگی قبل از نام قرار گیرد (Peng & McCallum, 2006).

▪ در استخراج اطلاعات از نوشتارهای علمی، تشخیص این که موارد به دست آمده متعلق به فراداده‌های موردنظر هستند یا خیر مشکل است (Meng, Hou, Yang, & Li, 2018).

۴-۱. نوآوری

در زمینه استخراج فراداده از نوشتارهای علمی، تحقیقات بسیاری بر اسناد عملی به زبان انگلیسی صورت گرفته که در اکثر آن‌ها تمرکز بر مقالات انتشار یافته در کنفرانس‌ها/ مجلات و نه پایان‌نامه بوده است. در مقابل در نوشتارهای علمی زبان فارسی پژوهش‌های اندکی انجام شده است.

الگوی پایان‌نامه‌ها و رساله‌ها در دانشگاه‌ها و مؤسسات آموزشی مختلف، متفاوت است. با این حال، در تمامی پایان‌نامه‌ها و رساله‌ها ذکر اطلاعاتی برای معرفی و شرح کلی تحقیق ضروری است که در این پژوهش به مساله استخراج خودکار برخی از این اقلام اطلاعاتی پرداخته می‌شود. به طور خلاصه، نوآوری‌های این پژوهش را می‌توان به صورت زیر بیان کرد:

- استخراج فراداده از اسناد پایان‌نامه/رساله‌ی فارسی مبتنی بر رویکرد یادگیری ماشین
- تعریف ویژگی‌های متناسب با قالب و ساختار کلی پایان‌نامه و خصوصیات زبان فارسی

۵-۱. ساختار طرح پژوهشی

این پژوهش در شش فصل ارائه شده که در فصل اول کلیات تحقیق، چالش‌ها و نوآوری بیان شد. در فصل دوم به مفاهیم پایه‌ی استخراج اطلاعات و الگوریتم‌های مورد استفاده در پژوهش، از جمله الگوریتم ماشین بردار پشتیبان و الگوریتم بهینه‌سازی خفاش، اشاره خواهد شد. فصل سوم به مرور بیشینه پژوهش و تحقیقات خارجی و داخلی انجام شده در این زمینه می‌پردازد و با گردآوری

تحقیقات مربوطه، انواع اطلاعات قابل استخراج از اسناد علمی و همچنین روش‌های به کار رفته در حل مسائل را بررسی خواهد کرد. در فصل چهارم، روش پیشنهادی برای استخراج فراداده از پایان‌نامه‌ها و رساله‌ها مطرح شده و گام‌های پژوهش به ترتیب تشریح می‌گردند. ارزیابی و آزمایش‌های انجام شده در فصل پنجم ارائه شده و هر مرحله به صورت جداگانه در این فصل مورد ارزشیابی قرار می‌گیرد. در نهایت، فصل ششم به نتیجه‌گیری این پژوهش اختصاص داشته و در آن موضوعاتی برای پژوهش‌های آتی پیشنهاد می‌گردد.

فصل دوم

مروری بر پیشینه پژوهش

۱-۲. مقدمه

در این فصل به مفاهیم پایه و پیشینه تحقیقات صورت گرفته در حوزه استخراج اطلاعات و فراداده‌ها از نوشتارهای علمی پرداخته می‌شود. این فصل از چند بخش تشکیل شده که در بخش اول، مفاهیم پایه در سه حوزه مرور می‌شود: ابتدا، به مفاهیم مرتبط با پردازش زبان طبیعی اشاره می‌گردد. در ادامه، به تعاریف پایه در حوزه استخراج اطلاعات از اسناد پرداخته می‌شود. در نهایت، برخی مفاهیم مرتبط با پردازش اسناد توضیح داده می‌شوند. در بخش دوم به پیشینه تحقیقات صورت گرفته در حوزه استخراج اطلاعات و فراداده‌ها از نوشتارهای علمی پرداخته می‌شود. همچنین الگوریتم‌های مورد استفاده در طبقه‌بندی و بهینه‌سازی و رویکردهای استخراج اطلاعات از مقالات علمی و تفاوت‌های آن‌ها نیز در این بخش مرور می‌گردد. سپس، به روش‌های پیشنهادی در این زمینه اشاره می‌شود. در پایان نیز دسته‌بند ماشین بردار پشتیبان و الگوریتم بهینه‌سازی خفاش مرور می‌شوند.

۲-۲. پردازش زبان طبیعی

زبان طبیعی به زبانی که انسان‌ها در ارتباطات خود به کار می‌برند اطلاق می‌شود. در مقابل، زبان‌های برنامه‌نویسی و نمادهای ریاضی نمونه‌ای از زبان‌های مصنوعی هستند که برخلاف زبان‌های طبیعی، در طول زمان تغییر چندانی نمی‌کنند. پویایی و تغییرپذیری از ویژگی‌های بارز زبان‌های طبیعی است و

سبب می‌شود تا نتوان از قواعد تعریف شده و مشخ در پردازش زبان طبیعی بهره برد (Bird, Klein, & Loper, 2009).

۲-۱-۲. کاربردهای پردازش زبان طبیعی

با گسترش پردازش زبان طبیعی، سیستم‌های بسیاری به وجود آمدند که هر کدام در جایگاه خود کاربرد خاصی داشته و در بهره‌برداری از حجم انبوه متون مفید بوده‌اند. در ادامه به چند مورد از بیشترین کاربردهای پردازش زبان طبیعی اشاره خواهد شد. (Otter, Medina, & Kalita, 2020)

- بازیابی اطلاعات: سیستم‌های بازیابی اطلاعات با هدف فراهم کردن اطلاعات مفید و موردنیاز کاربر، در قالب و زمان درست، به وجود آمده‌اند. یکی از مهم‌ترین مشکلات این سامانه‌ها، نیاز به رتبه‌بندی اسناد بر اساس پرس‌وجوی کاربر است.
- استخراج اطلاعات: عبارت است از فرآیند شناسایی مفاهیم از قبل تعریف شده در یک حوزه مشخص و بدون در نظر گرفتن اطلاعات غیرمرتبط. در این فرآیند علاوه بر موجودیت‌ها و خصیصه‌های آن‌ها، ارتباط بین هر کدام از موجودیت‌ها نیز به دست می‌آید. (Piskorski & Yangarber, 2013) از آن‌جا که تمرکز این تحقیق بر استخراج اطلاعات و فراداده‌هاست، در بخش‌های بعدی این فصل بیشتر در این باب توضیح داده می‌شود.
- طبقه‌بندی متن: از دیگر کاربردهای پردازش زبان طبیعی، دسته‌بندی متون در کلاس‌های از پیش تعریف شده است.
- تولید متن: در برخی از سیستم‌های پردازش زبان، نیاز است تا متونی نزدیک به زبان طبیعی انسان تولید شود. به عنوان نمونه در سامانه‌های خلاصه‌سازی متون و ترجمه از زبانی به زبان دیگر، خروجی نیز باید متن به زبان طبیعی باشد.

▪ خلاصه‌سازی متن: در این فرآیند، عناصر مهم و مفید از متن به دست می‌آید تا در نهایت بتوان خلاصه‌ای از متن ورودی را به نمایش گذاشت. روش‌های استخراجی^۷ و انتزاعی^۸ دو رویکرد اصلی در خلاصه‌سازی خودکار متون هستند. در روش اول تمرکز بر جملات است و پردازش‌هایی برای ساده‌سازی، مرتب‌سازی مجدد و الحاق آن‌ها صورت می‌گیرد. شیوه خلاصه‌سازی انتزاعی با تمرکز بر مفاهیم بیان شده در محتوای اسناد عمل کرده و خلاصه‌ای از متن را تولید می‌کند.

▪ سیستم پرسش و پاسخ^۹: مشابه فرآیند خلاصه‌سازی و استخراج اطلاعات، سیستم پرسش و پاسخ کلمات، عبارات و جملات به هم مرتبط را از اسناد گردآوری می‌کند و این اطلاعات را به عنوان پاسخ به سوال برمی‌گرداند.

▪ ترجمه متن: یکی از مهم‌ترین کاربردهای پردازش زبان طبیعی، ترجمه متن از زبانی به زبان دیگر است. در چنین سیستم‌هایی، با استفاده از روش‌های ریاضی و الگوریتم‌های موجود در این زمینه، می‌توان سندی را از یک زبان به زبان دیگر ترجمه کرد.

▪ تحلیل احساس: این دسته از سامانه‌های پردازش زبان طبیعی به تحلیل نظرات و احساسات کاربران درباره موضوعات موردنظر مانند کالاهای فروشگاه‌های اینترنتی می‌پردازند. تکنیک‌های تحلیل احساس عمدتاً بر اساس روش‌های آماری و یادگیری ماشین عمل می‌کنند (Liu & Zhang, 2012).

۲-۲-۲. پیش‌پردازش متن

یکی از ملزومات پردازش متن، پیش‌پردازش آن است تا داده‌های اضافی و ناخواسته موجود در متن حذف شده و آماده مراحل بعدی شود (Dudhabaware & Madankar, 2014). در ادامه به انواع

⁷ Extractive

⁸ Abstractive

⁹ Question Answering (QA)

پردازش‌هایی که به طور معمول بر متن‌های خام اعمال می‌شود، اشاره خواهد شد (Mohseni, Ghofrani, & Faili, 2018):

- یکسان‌سازی^{۱۰}: به منظور یکپارچه‌سازی کاراکترهای متن، در بسیاری از موارد نیاز است تا فرآیند یکسان‌سازی بر متن صورت گیرد. عوامل زیادی موجب عدم انسجام در داده‌های متنی – به خصوص در زبان فارسی – می‌شوند. از جمله این که برخی از حروف مشترک در الفبای زبان فارسی و عربی دارای کدهای متفاوت در جدول یونی کد هستند؛ مانند حروف /ی/ و /ک/ که دارای چند کد مختلف به ازای هر حرف هستند. در نتیجه با نرمال کردن متن تمامی کدهای عربی به فارسی تبدیل می‌شود. به علاوه، در برخی مواقع، کاراکترهای همزه و تنوین که ریشه عربی داشته و از عربی وارد زبان فارسی شدند، موجب بروز ابهام در متون فارسی می‌شوند. برای نمونه «املا» و «املاء» دو نگارش متفاوت برای یک کلمه هستند. از طرفی در زبان فارسی – برخلاف زبان عربی – اعراب‌گذاری مرسوم نیست. بلکه فقط در مواقع خاص و جلوگیری از ابهام در خوانش، علامت‌گذاری صورت می‌گیرد. تمامی نکات بیان‌شده را می‌توان با اعمال پردازش یکسان‌سازی برطرف نمود و در نتیجه حروف، علائم و ارقام عربی به فارسی تبدیل خواهند شد.
- جداسازی نشانه‌ها^{۱۱}: در برخی از کاربردهای پردازش زبان طبیعی نیاز است تا متن ورودی تبدیل به دنباله‌ای از نشانه‌ها یا به عبارتی «توکن‌ها» شود. به طور معمول مرز توکن‌ها براساس فاصله، علائم نگارشی و دیگر ویژگی‌های املائی (همچون وجود پسوند یا پیشوند) جداسازی شده و بدین ترتیب نشانه‌ها به دست می‌آیند.

¹⁰ Normalization

¹¹ Tokenization

- جداسازی جملات^{۱۲}: در برخی کاربردها ضروری است تا جملات موجود در متن جدا شده و کلمات موجود در هر جمله نیز تفکیک یابند. بدین ترتیب خروجی این فرآیند جملاتی است که عمدتاً با علامت نگارشی نقطه یا دیگر علائم خاتمه می‌یابند.
- تشخیص ایست‌واژه‌ها^{۱۳}: کلماتی که معنای خاصی در متن ندارند و حذف آن‌ها موجب ابهام در مفهوم کلی متن نمی‌شود، ایست‌واژه نامیده می‌شوند. برای نمونه حروف اضافه معمولاً در فهرست ایست‌واژه‌ها قرار می‌گیرند. چنین کلماتی در تمامی زبان‌ها وجود دارند و تنها ممکن است از حوزه‌ای به حوزه دیگر متفاوت باشند. برای تشخیص ایست‌واژه‌ها کافی است بررسی شود که آیا کلمه‌ی مورد نظر در فهرست ایست‌واژه‌های جمع‌آوری شده وجود دارد یا خیر.
- برچسب‌گذاری اجزای واژگانی کلام^{۱۴}: در پاره‌ای از کاربردهای پردازش زبان طبیعی مانند ترجمه متن، تبدیل متن به صوت، بررسی هجی کلمات، سیستم‌های تشخیص صوت و دیگر سامانه‌ها نیاز است تا مقوله کلمات در جمله به دست آید.
- ریشه‌یابی^{۱۵}: در بسیاری از متون، کلماتی یافت می‌شوند که ریشه آن‌ها یکی است و تنها تفاوتشان در وجود وندها (پیشوند و پسوند) است. برای نمونه ریشه کلمات «کتاب» و «کتاب‌ها» واژه کتاب است. الگوریتم‌های ریشه‌یابی برای دستیابی به ریشه کلمات تعریف شده‌اند. این الگوریتم‌ها معمولاً بدون اطلاعات قبلی درباره کلمات عمل می‌کنند و همین عامل سبب می‌شود تا لزوماً خروجی به دست آمده از تابع ریشه‌یابی، واژه معناداری نباشد. این مهم تفاوت ریشه‌یابی با روش‌های بن‌واژه‌سازی است (Jabbar, Iqbal, Khan, & Hussain, 2016).

¹² Sentencization

¹³ Stop words

¹⁴ POS tagging

¹⁵ Stemming

- بن‌واژه‌سازی^{۱۶}: این تکنیک تقریباً مشابه ریشه‌یابی است با این تفاوت که کلمات موجود در متن با اطلاعات قبلی پردازش می‌شود. این اطلاعات در پایگاه داده‌ای ذخیره می‌شوند و در نتیجه خروجی نهایی آن کلمه‌ای معنادار خواهد بود. به علاوه روش‌های بن‌واژه‌سازی برای به‌دست آوردن بن‌فعل‌ها نیز مناسب هستند.
- قطعه‌بند عبارات اسمی^{۱۷}: این توابع برای تشخیص عبارات و گروه‌های اسمی در جملات به کار می‌روند. پیش از اجرای این پردازش، باید نشانه‌های متن جداسازی شده و نقش هر واژه برچسب‌گذاری شده باشد تا گروه‌های اسمی به درستی شناسایی گردند.

۳-۲. مفاهیم پایه در حوزه استخراج اطلاعات

همان‌طور که پیش‌تر اشاره شد، استخراج اطلاعات یکی از کاربردهای پردازش زبان طبیعی است که طی آن موجودیت‌های از پیش تعریف شده‌ای از متن استخراج می‌شود. بدیهی است که این اطلاعات در هر حوزه متفاوت بوده و اهمیت و کاربردهای گوناگونی خواهند داشت.

۱-۳-۲. هدف استخراج اطلاعات

هدف از استخراج اطلاعات، دستیابی به اطلاعات ساخت‌یافته^{۱۸} از اسناد بدون ساختار^{۱۹} یا نیمه ساخت‌یافته^{۲۰} است (Adrian, Leone, Manna, & Marte, 2017). به طور کلی داده‌ها به سه نوع ساخت‌یافته، نیمه‌ساخت‌یافته و ساخت‌نیافته تقسیم می‌شوند.

¹⁶ Lemmatization

¹⁷ Noun phrase (NP) chunker

¹⁸ Structured information

¹⁹ Unstructured documents

²⁰ Semi-structured documents

- داده ساخت یافته به داده‌ای گفته می‌شود که از چندین فیلد^{۲۱} تشکیل شده باشد (Dejean, 2015). به عنوان مثال می‌توان به جداول پایگاه داده به عنوان نمونه‌ای از این دست داده‌ها اشاره کرد.
- داده نیمه‌ساخت یافته به داده‌های گفته می‌شود که نه به شکل داده‌های خام و نه به شکل داده‌های پایگاه داده هستند (Abiteboul, 1996). صفحات وب و یا فرمت‌های html, json, xml و ... نمونه‌هایی از این نوع داده هستند.
- داده‌های خامی مانند تصویر، ویدیو، صوت، متن و دیگر داده‌هایی که ساختار مشخصی ندارند، داده‌های ساختاریافته نامیده می‌شوند.
- همچنین در پردازش سند، اسناد نیز به سه دسته ساخت یافته، نیمه‌ساخت یافته و ساختاریافته تقسیم می‌شوند. دو ویژگی زیر می‌تواند از خصوصیت‌های یک سند باشد:
 - اسنادی که دارای مجموعه داده‌های از پیش تعریف شده هستند.
 - اسنادی که طرح‌بندی^{۲۲} از پیش تعریف شده‌ای داشته و اجزای مشخصی دارند.
- اگر سندی ویژگی ۱ و ۲ را با هم داشته باشد سند ساخت یافته و اگر یکی از ویژگی‌های ۱ و ۲ را داشته باشد سند نیمه‌ساخت یافته نامیده می‌شود؛ در غیر این صورت بدون ساختار خواهد بود (Cristani, Bertolaso, Scannapieco, & Tomazzoli, 2018).
- ۲-۳-۲. وظایف سیستم استخراج اطلاعات
- به طور کلی در یک سیستم استخراج اطلاعات چهار وظیفه تعریف می‌شود (Piskorski & Yangarber, 2013):

²¹ Feild

²² Layout

- شناسایی موجودیت‌های اسمی^{۲۳}: به معنای تعیین موجودیت‌ها و تخصیص آن‌ها به کلاس‌های از پیش تعریف شده است. در این فرآیند اسامی سازمان‌ها، اشخاص، مکان‌ها، عبارات زمانی، عبارات عددی و واحد پول و سایر موجودیت‌ها به دست می‌آید.
 - تحلیل اسامی هم‌ارجاع^{۲۴}: تشخیص کلماتی که در واقع به یک موجودیت اشاره دارند و به شکل‌های متفاوتی ذکر شده‌اند. برای نمونه ممکن است یک عبارت اختصاری برای واژه‌ای به کار رود. همچنین ضمائر، اشاره به موجودیت‌هایی در متن دارند که طی این فرآیند مرجع ضمیر شناخته می‌شود. گاهی نیز ممکن است مرجع برخی عبارات اسمی موجودیت خاصی باشند. برای مثال در جمله «مایکروسافت درآمد خود را فاش کرد. همچنین این شرکت از برنامه‌های آینده پرده‌برداری کرد.» عبارت «این شرکت» اشاره به مایکروسافت دارد. در حالت دیگری نیز ممکن است در برخی زبان‌ها، فاعل به صورت صریح در جمله ظاهر نشود.
 - استخراج روابط^{۲۵}: به معنای شناسایی روابط از پیش تعریف شده بین موجودیت‌هاست. برای نمونه ارتباط بین اشخاص و سازمان‌ها یا مکان‌های به دست آمده شناسایی می‌شود.
 - استخراج رویداد^{۲۶}: شناسایی رویدادها و استخراج اطلاعات جزئی از متن، خروجی این فرآیند است. در واقع اطلاعاتی درباره این که چه کسی، در کجا، چه وقت، چرا، چگونه و با چه روشی، چه کاری را انجام داده است.
- ۳-۳-۲. استخراج اطلاعات از اسناد علمی

اسناد علمی و دانشگاهی حاوی اطلاعات ارزشمندی هستند و استخراج اطلاعات از این اسناد یکی از حوزه‌های فعال است که محققان بسیاری بر این موضوع تمرکز دارند.

²³ Named Entity Recognition (NER)

²⁴ Co-reference resolution

²⁵ Relation extraction (RE)

²⁶ Event extraction (EE)

انواع داده‌هایی که می‌توان از اسناد علمی به دست آورد در دو دسته کلی فراداده‌ها و بینش‌های کلیدی^{۲۷} قرار می‌گیرند (Nasar, Jaffry, & Malik, 2018).

(۱) فراداده

در یک تعریف کلی می‌توان فراداده را به عنوان داده‌ای در نظر گرفت که اطلاعاتی درباره یک داده دیگر می‌دهد. فراداده‌ها را می‌توان به سه دسته تقسیم‌بندی نمود: توصیفی (مانند عنوان و نویسنده)، ساختاری (مانند تعداد صفحات) و اجرایی (مانند مالک) (Meng, Hou, Yang, & Li, 2018).

فراداده در کتابداری تعریف دیگری دارد. برای نمونه در اسناد علمی به اطلاعاتی مانند عنوان سند، نویسندگان، محل انتشار، تاریخ انتشار، چکیده، کلمات کلیدی و... فراداده می‌گویند. استخراج فراداده از نوشتارهای علمی با دو رویکرد کلی امکان‌پذیر است. یکی از این رویکردها تحلیل فراداده‌های سرآیند^{۲۸} مقالات است. در این رویکرد فراداده‌های توصیفی که در بخش‌های آغازین سند علمی بیان می‌شوند، استخراج می‌شوند. تجزیه رشته‌های مرجع نیز رویکرد دیگر برای دستیابی به این اطلاعات است (Nasar, Jaffry, & Malik, 2018).

(۲) بینش‌های کلیدی

با توجه به توصیف ارائه شده در (Nasar, Jaffry, & Malik, 2018) علاوه بر فراداده‌ها، اطلاعات بالقوه دیگری نیز در اسناد علمی هستند که دستیابی به آن‌ها در بهبود عملکرد سامانه‌های بازیابی اطلاعات تأثیرگذار است. این اطلاعات به شرح ذیل است:

- تعریف مسئله
- حوزه مطالعه پژوهش
- روش‌ها، الگوریتم‌ها و پردازش‌هایی که در مسئله تحقیق به کار گرفته می‌شود.
- مجموعه داده‌های به کار گرفته شده در اسناد علمی

²⁷ Key insights

²⁸ Header

- ابزارهای به کار گرفته در اجرای آزمایش‌ها
- معیارهای ارزیابی
- نتایج به دست آمده
- محدودیت‌های حوزه مطالعه
- پیشنهادهایی برای پژوهش‌های آینده

۴-۲. پردازش سند

پردازش سند^{۲۹} فرآیندی است که طی آن متن و اطلاعاتی که برای انسان قابل فهم هستند، از سند ورودی به دست می‌آید. سیستم پردازش سند می‌تواند اسناد را در قالب‌هایی چون PDF، XML، RTF و یا DOCX به عنوان ورودی دریافت کند. در صورتی که "تصویر سند" ورودی این سیستم باشد، به این فرآیند تحلیل تصویر سند^{۳۰} گویند (Cristani, Bertolaso, Scannapieco, & Tomazzoli, 2018). درک سند^{۳۱} نیز یکی دیگر از اصطلاحات پردازش سند است و به حوزه‌ای اطلاق می‌شود که مرتبط با تحلیل منطقی و معنایی اسناد است تا اطلاعات قابل فهم انسان استخراج و به شکل قابل فهم ماشین کدگذاری شوند. (Dengel & Shafait, 2014)

۱-۴-۲. مراحل پردازش سند

در صورتی که تصویر سند در اختیار باشد، ضروری است تا ابتدا تصویر مورد پردازش قرار گیرد، در ادامه نواحی متنی تشخیص داده شده و در پایان محتوای متنی استخراج شود.

مراحل پردازش سند شامل پیش‌پردازش، تحلیل طرح‌بندی فیزیکی (هندسی)^{۳۲}، تشخیص حروف – در این مرحله سیستم‌های نویسه‌خوان نوری^{۳۳} به کار می‌روند – و استخراج اطلاعات^{۳۴} است. در

²⁹ Document processing

³⁰ Document image analysis(DIA)

³¹ Document understanding

³² Physical (Geometric) layout analysis

³³ Optical character recognition

³⁴ Information Extraction (IE)

مرحله‌ی پیش‌پردازش، کارهایی چون حذف نویز^{۳۵}، تشخیص و اصلاح انحراف^{۳۶} تصویر سند و جداسازی پس زمینه^{۳۷} انجام می‌شود تا تصویر برای مراحل بعدی آماده گردد. سپس تحلیل طرح‌بندی فیزیکی سند صورت می‌گیرد؛ بدین معنا که تصویر به صورت سلسله‌مراتبی تجزیه شده و نواحی متنی و غیرمتنی جدا می‌شود. نواحی متنی به قسمت‌هایی که بیشترین شباهت را با هم دارند، تقسیم شده و هر ناحیه نیز به زیر ناحیه‌های دیگر با بیشترین شباهت تقسیم می‌شود و این تقسیم‌بندی می‌تواند تا سطح پیکسل، اجزای به هم پیوسته^{۳۸}، کلمه یا هر سطح مورد نظر دیگر ادامه یابد. همچنین معیار شباهت هر ناحیه باید از قبل تعریف شده باشد (Directions & Advances, 2007). پس از این مراحل، کلمات موجود در تصویر شناسایی شده و خروجی آن محتوای متن سند است.

تا این مرحله تنها محتوای سند به دست آمده و برای دسترس‌پذیری به اطلاعات سند، تبدیل تصویر به متن قابل فهم ماشین کافی نیست و ضروری است تا اطلاعات با هدف دسترسی و جستجو در محتوای سند استخراج شود (Toledo, Carbonell, Fornés, & Lladós, 2019). بدیهی است در صورتی که اسناد در قالب غیرتصویری در اختیار باشند و محتوای متنی آن‌ها به راحتی در دسترس قرار گیرند، نیازی به انجام برخی مراحل ذکر شده نبوده و تنها پردازش متن و تحلیل روابط منطقی بین اجزای سند کفایت می‌کند.

۲-۴-۲. پردازش روابط منطقی در ساختار سند

همان‌گونه که انسان با مشاهده اسناد مشابه (از نظر طرح‌بندی، ساختار و یا محتوا) اطلاعات موردنظر خود را - با توجه به ویژگی‌های مشترک آن‌ها - به دست می‌آورد، می‌توان نتیجه گرفت که مکان قرارگیری برخی بخش‌های متن از یک سند به سند دیگر تکرار می‌شود. به بیان دیگر هر بخش از متن برچسب مشخصی دارد؛ به طوری که برچسب موردنظر اشاره به نقش خاصی در ساختار متن دارد.

³⁵ Noise removal

³⁶ Skew detection and correction

³⁷ Background separation

³⁸ Connected component

گاهی ساختار متن، حاوی معنا و نکات مفیدی درباره مکان اطلاعات موردنظر است. برای نمونه اگر طرح‌بندی بخشی از سند دو ستونه باشد، این بخش ممکن است مربوط به جدولی در متن باشد (Adrian, Leone, Manna, & Marte, 2017) برای به‌دست آوردن روابط بین اجزای سند، ساختار سند^{۳۹} بررسی می‌شود.

در تحلیل ساختار منطقی^{۴۰} نواحی متنی مورد تجزیه و تحلیل قرار گرفته و نقش هر قسمت تعیین شده و برچسب موردنظر به آن ناحیه زده می‌شود. مصادیقی از برچسب‌های ساختار منطقی عبارتند از: عنوان، زیرعنوان، چکیده، سرتیتر^{۴۱}، پاورقی، شماره صفحه و عناوین شکل‌ها و جدول‌ها (Haralick, 1994). تحلیل ساختار منطقی سند یکی از روش‌هایی است که به یافتن موجودیت‌ها و ارتباط آن‌ها کمک می‌کند.

۲-۵. مطالعه پیشینه پژوهش در حوزه‌ی استخراج اطلاعات از نوشتارهای علمی

همان‌طور که بیان شد، از یک دیدگاه عمومی می‌توان دو دسته اطلاعات از اسناد علمی استخراج کرد. فراداده‌ها یکی از انواع اطلاعات ارزشمندی هستند که مرور مختصری بر محتوای اسناد علمی دارند. برای نمونه می‌توان به عنوان سند، نویسندگان، محل انتشار، تاریخ انتشار، کلمات کلیدی و ... اشاره کرد. مقالاتی که بر استخراج فراداده از اسناد علمی تمرکز دارند عمدتاً به دو روش اقدام به این کار می‌کنند: تحلیل صفحات آغازین مقالات و تجزیه‌ی رشته‌های مرجع^{۴۲}. به غیر از فراداده، نوع دیگری از اطلاعات بالقوه در اسناد علمی و دانشگاهی وجود دارد. اطلاعاتی چون حوزه تحقیقاتی، تعریف مسئله، روش‌ها، پاسخ‌دهی به سوالات تحقیق، مجموعه داده، ابزارها، محدودیت‌ها، معیارهای ارزیابی، نتایج به دست آمده و پیشنهاد تحقیقات آینده.

³⁹ Document structure

⁴⁰ Logical layout analysis

⁴¹ Header

⁴² References

بیشتر پژوهش‌های انجام شده زمینه استخراج اطلاعات از اسناد و مقالات علمی، بر استخراج فراداده‌ها و اطلاعات ذکر شده از مقالات- به خصوص مقالات لاتین- تمرکز کرده‌اند. در ادامه درباره رویکردهای استخراج فراداده از مقالات توضیح داده می‌شود.

۲-۵-۱. رویکردهای استخراج اطلاعات از نوشتارهای علمی

پژوهش‌هایی که بر استخراج اطلاعات از مقالات علمی تمرکز داشته‌اند، عمدتاً با دو رویکرد مبتنی بر قواعد^{۴۳} و مبتنی بر یادگیری ماشین^{۴۴} سعی در حل مسئله دارند. در رویکرد مبتنی بر قواعد، اجزای سند مقاله با قواعد از پیش تعریف‌شده‌ای (مانند عبارات منظم^{۴۵}) تطبیق می‌یابند. در رویکرد یادگیری ماشین، تکنیک‌های پردازش زبان طبیعی با الگوریتم‌های یادگیری ماشین چون مدل مخفی مارکوف، میدان تصادفی شرطی، ماشین بردار پشتیبان و یا روش‌های مبتنی بر یادگیری عمیق ترکیب می‌شوند (Nasar, Jaffry, & Malik, 2018).

مزایا و معایب رویکردهای استخراج اطلاعات از مقالات علمی

روش‌های مبتنی بر قواعد، عملکرد بهتری نسبت به رویکردهای یادگیری ماشین دارند؛ زیرا تولید قواعد به صورت دستی انجام می‌شود و عامل انسانی در تعریف این قواعد دخالت دارد. اما از سوی دیگر امکان سازگاری و وفق پیدا کردن با ساختار و طرح‌بندی‌های متفاوت انواع اسناد را ندارند. علاوه بر آن، تعریف قواعد کاری زمان‌بر و پرهزینه است. واضح است که هر چه قواعد پیچیده‌تری تعریف شوند، سیستم موردنظر می‌تواند الگوهای پیچیده‌تری را تشخیص و تطبیق دهد. در مقابل، این امر موجب طولانی‌تر شدن پردازش خواهد شد (Nasar, Jaffry, & Malik, 2018). همچنین لازم است تا تولید قواعد توسط افراد خبره و آگاه به حوزه موردنظر انجام شود. به علاوه، اگر ویژگی‌هایی

⁴³ Rule based

⁴⁴ Machine learning based

⁴⁵ Regular expression

که برای تعریف قواعد به کار می‌رود افزایش یابد، قواعد به خوبی عمل نمی‌کنند. (Day, et al.,

2007) تفاوت‌های ذکر شده در مورد دو رویکرد موردنظر در **جدول ۱-۲** خلاصه شده‌اند.

جدول ۱-۲. تفاوت رویکردهای استخراج اطلاعات از نوشتارهای علمی

مبتنی بر قواعد	مبتنی بر یادگیری ماشین
تعریف قواعد توسط عامل انسانی	یادگیری خودکار بدون دخالت عامل انسانی
عدم سازگاری با انواع اسناد	سازگاری با اسناد دیده نشده
لزوم تولید قواعد توسط افراد خبره	لزوم وجود مجموعه داده‌ی کافی برای یادگیری
عملکرد نامناسب در مواجهه با افزایش ویژگی	عملکرد نسبتاً خوب در مواجهه با ابعاد بالای ویژگی

رویکرد مبتنی بر قاعده

در روش پیشنهادی در (Klink, Dengel, & Kieninger, 2000)، قواعدی بر اساس ویژگی‌های متنی و هندسی اسناد تعریف شده تا بخش‌های مربوط به فراداده‌های موردنظر شناسایی شود. در ابتدا تصویر سند اسکن شده و محتوای آن -شامل متن، ویژگی‌های طرح‌بندی و فونت- به کمک موتور نویسه‌خوان نوری به دست می‌آید و سپس ساختار فیزیکی و منطقی سند تحلیل می‌شود. همان‌طور که در فصل اول توضیح داده شد، سلسله‌مراتبی فیزیکی سند شامل بلوک‌های متنی است که هر بلوک از چندین خط تشکیل می‌شود و در نهایت هر خط شامل چندین کلمه است. در تحلیل منطقی سند نیز عناصری نظیر عناوین، جداول، شماره‌ی صفحه و ... شناسایی می‌شود. پس از تحلیل و شناسایی ساختار سند یا به اصلاح درک سند، به عناصر به‌دست آمده از مراحل قبلی، با استفاده از قواعد تعریف شده برچسبی اختصاص می‌یابد. برچسب‌ها شامل عنوان و بدنه بخش چکیده، وابستگی‌ها^{۴۶}، بیوگرافی، عنوان جداول و تصاویر، حرف درشت اولین کلمه پاراگراف^{۴۷}، هایلایت، عنوان و بدنه بخش کلمات کلیدی، عضویت، شماره صفحه، شبه کد، اطلاعات نشر، خدمات خواننده، خلاصه و

⁴⁶ Affiliation

⁴⁷ Drop cap

بدنه متن اصلی هستند. تطبیق قواعد نیز به صورت فازی انجام می‌شود. از مزایای روش پیشنهاد شده در این پژوهش این است که کاربر امکان ویرایش خطاهای خروجی را دارد و در نتیجه قواعد می‌توانند متناسب با موارد ویرایش یافته تغییر کنند تا امکان بروز خطای مجدد کاهش یابد.

تبیین قواعد با تکیه بر پایگاه دانش^{۴۸} مبنای روش ارائه شده در (Giuffrida, Shek, & Yang, 2000) است. این اطلاعات عموماً بر اساس ویژگی‌های بصری و فضایی نوشتار است. در ادامه چند نمونه اطلاعاتی که از قبل درباره محتوای مقالات وجود دارد ذکر می‌شود.

- عنوان با فونت درشت در بالای اولین صفحه قرار دارد.
- اسامی نویسندگان معمولاً بعد از عنوان و به ترتیب نوشته می‌شود.
- اطلاعات مربوط به محل خدمت و وابستگی‌های نویسندگان پس از لیست نویسندگان نوشته می‌شود.
- اگر تنها یک فراداده وابستگی یافت شود، این فراداده مربوط به تمامی نویسندگان است.
- نوع فونت نام در نوشتن نویسندگان و وابستگی‌های آن‌ها مشابه است.
- عناوین با فونتی بزرگ‌تر از زیرعنوان‌ها درج می‌شوند.

به این ترتیب برای فراداده قواعدی مبتنی بر دانش اولیه تعریف شده و به روش فازی تطبیق و برچسب‌زنی به قسمت‌های مختلف انجام می‌شود. فراداده‌هایی که در این روش از مقالات استخراج می‌شوند عبارتند از: عنوان، نویسندگان، وابستگی‌ها و فهرست مطالب. در پژوهش مذکور تنها از مقالاتی استفاده شده که ساختار مشخصی دارند و از طرح‌های پژوهشی، پایان‌نامه‌ها، رساله‌ها یا اسنادی که ساختار متفاوتی دارند اجتناب شده است.

در (Mao, Kim, & Thoma, 2004) تلاش شده برای هر نوع سند، به صورت خودکار ویژگی‌هایی تولید شود تا استخراج اطلاعات با قواعد مطابق با ویژگی‌های ساختاری همان سند انجام گیرد. این

⁴⁸ Knowledge based

رویکرد با توجه به این واقعیت اتخاذ شده که ساختار مقالات از مجله‌ای به مجله دیگر متفاوت است. همچنین ممکن است قالب نوشتاری مجلات در طول سال‌های متمادی تغییر یابد. بنابراین به دلیل تنوع ساختاری مقالات نمی‌توان به قواعدی که وابستگی بسیاری به طرح‌بندی اسناد دارند، متکی بود. در روش پیشنهاد شده تلاش شده تا قواعد با تحلیل ویژگی‌های ساختاری و هندسی مانند اندازه فونت، نوع فونت و مختصات عناصر متنی با محاسبه توزیع آن‌ها تولید شوند. این ویژگی‌ها از خروجی نویسه‌خوان نوری به دست می‌آیند و در ادامه توزیع این ویژگی‌ها در مقاله ورودی محاسبه می‌شود. سپس برای مقالات هر مجله، به ترتیب شماره چاپ آن‌ها، مدلی به دست می‌آید؛ با این فرض که مقالات این مجله همگی ویژگی‌های ساختاری مشابهی دارند. این فرض تا زمانی برقرار است که سردبیر مجله قصد تغییر سبک نوشتاری مقالات را نداشته باشد و معمولاً چنین رویدادی کمتر رخ می‌دهد.

در (Huang, Jin, Yuan, & Han, 2006) روشی پیشنهاد شده که با استفاده از آن، دستیابی به فراداده‌ها با تطبیق قالب سند انجام‌پذیر است. تطابق با قالب اسناد بر اساس ویژگی‌های بخش‌ها^{۴۹}، پاراگراف‌ها، فونت و ویژگی‌های کلی سند است. برای مثال تعداد ستون‌ها و مکان قرارگیری شماره صفحه از ویژگی‌های هر بخش است. هر پاراگراف ممکن است در راست، چپ یا مرکز تراز شده باشد و با تورفتگی نیز همراه باشد. ویژگی‌های فونت‌ها می‌توانند اندازه، سبک و قالب، رنگ، درشت‌نمایی^{۵۰}، خط زیرین و دیگر اطلاعات مشابه باشند. طول و عرض صفحه و حاشیه‌ها نمونه‌ای از ویژگی‌های کلی سند است. در این روش با تطبیق قالب سند ورودی با ویژگی‌های ذکر شده، اطلاعات موردنظر به دست می‌آید.

استخراج فراداده از رشته‌های متنی مراجع هدف پژوهش صورت گرفته در (Day, et al., 2007) است. در این پژوهش، قواعد بر اساس قالب روش‌های مرسوم ارجاع مانند ACM، IEEE، APA،

⁴⁹ Sections

⁵⁰ Bold

MISQ، JMIS و ISR بوده و مبتنی بر ویژگی‌های لغوی و متنی ایجاد می‌شوند. بدین ترتیب فراداده‌های نویسنده، عنوان، نام مجله، سال، تعداد سال‌های چاپ مجله^{۵۱}، شماره چاپ در سال^{۵۲} و صفحات از رشته مراجع به دست می‌آیند. در ادامه اطلاعاتی که بر اساس آن‌ها قواعد مربوط به هر فراداده تعریف شده است شرح داده می‌شود:

- نویسنده: نام و نام‌خانوادگی نویسندگان به روش‌های گوناگونی نوشته می‌شود. به عنوان مثال، ممکن است بین نام و نام‌خانوادگی کاراکتر فاصله یا کاما قرار گیرد یا ممکن است نام‌خانوادگی قبل از نام نویسنده نوشته شود. همچنین در مراجع انگلیسی امکان دارد که تنها حرف اول و یا همه حروف به شکل حرف بزرگ نوشته شوند. نکته دیگر این که در برخی سبک‌های ارجاع‌دهی، نام کوچک نویسندگان کامل و در برخی دیگر تنها حرف اول نوشته می‌شود. به طور کلی خلاصه‌نویسی و جداسازی نام نویسندگان از مشکلاتی است که در تحلیل مراجع دیده می‌شود و تمامی این حالات باید در تعریف قواعد در نظر گرفته شوند.
- عنوان: در روش پیشنهادی پژوهش مذکور، هدف دستیابی به عناوین مقالات نشریات است. در این حالت فرض شده که عناوین با حروف بزرگ به سبک عنوان یا به سبک جملات عادی نوشتار ظاهر می‌شود.
- نام مجله: این فراداده با کمک پایگاه داده‌ای که شامل نام مجلات پرتکرار است استخراج می‌شود.
- سال: اعداد چهار رقمی یا دو رقمی معمولاً می‌توانند کاندیدی برای این فراداده باشند. همچنین وجود کلمات فصول و ماه‌ها نیز در دستیابی به این فراداده کمک می‌کند.
- شماره والیوم: با مشاهده کلمه volume یا شکل اختصاری Vol. شماره‌ی والیوم به دست می‌آید.

⁵¹ Volume number

⁵² Issue number

- شماره چاپ در سال: وجود واژه issue یا شکل اختصاری No. در دستیابی به این اطلاعات کمک‌کننده است.

- شماره صفحات: بنابر نوع نمایش صفحات به صورت بازه‌ای یا به سبک‌های دیگر، قواعدی تعریف می‌شود تا این فراداده نیز استخراج شود.

روش ارائه شده در (Flynn, Zhou, Maly, Zeil, & Zubair, 2007) برای دستیابی به اطلاعات موردنظر بر اساس قواعد است و قالب اسناد ورودی PDF هستند. ابتدا فایل به قالب XML تبدیل شده و محتوای متنی سند ورودی به دست می‌آید. سپس ساختار سند تحلیل می‌شود و در ادامه با استفاده از قواعدی که متناسب با ویژگی‌های ساختاری آن تعریف شده، فراداده‌ها استخراج می‌شوند. با توجه به داده‌ها و نوع فراداده‌های موردنظر محققان، پردازش اولین صفحه و پنج صفحه آخر در هر فایل ورودی کفایت می‌کند. بنابراین در اولین مرحله پردازش صفحات موردنظر از فایل ورودی در قالب PDF جدا می‌شود و این فایل کوتاه شده و ناقص برای ادامه پردازش‌ها آماده می‌شود. در این قسمت محتوای متنی با استفاده از نویسه‌خوان نوری به دست آمده و قالب سند از PDF به قالب XML تبدیل می‌گردد. در بین داده‌های مورد استفاده در این پژوهش، اسنادی وجود دارند که حاوی نوعی فرم استاندارد به نام صفحه سند گزارش^{۵۳} هستند. این فرم برای نمایش ساخت یافته اطلاعات است که در این تحقیق بیش از ۲۰ فراداده را می‌توان از این فرم به دست آورد. بنابراین در مرحله دوم پردازش، بررسی می‌شود که آیا سند ورودی سیستم شامل چنین فرمی است یا خیر. با توجه به این که ساختار فرم موردنظر کاملاً مشخص است، با تعریف قواعدی می‌توان ساختار سند را بررسی کرد. در صورتی که فرم یافت شود، سند به بخش «پردازش فرم»^{۵۴} هدایت شده و فراداده‌های موردنظر از آن به دست می‌آید. در صورت عدم مشاهده این ساختار، سند به بخش «پردازش غیر فرم»^{۵۵} وارد می‌شود و در این

⁵³ Report document page (RDP)

⁵⁴ Form processing

⁵⁵ Non-form processing

بخش نیز با قواعدی از پیش تعریف شده و بر اساس ساختار و قالب سند، فراداده‌ها استخراج می‌شوند. در ادامه، خروجی هر دو فرآیند وارد بخش پیش‌پردازش شده تا عملیات پاک‌سازی و یکسان‌سازی صورت گیرد. در آخرین مرحله صحت فراداده‌های به دست آمده بررسی می‌شود و در صورتی که صحت داده‌ها تایید نشود، لازم است توسط عامل انسانی مورد بازبینی قرار بگیرد.

در (Groza, Handschuh, & Hulpus, 2009) روشی ارائه شده که در فاز نخست، فراداده‌های صفحه اول مقالات را به دست می‌آید. این فراداده‌ها شامل عنوان، نویسندگان و چکیده هستند. سپس در فاز دوم قسمت‌های مختلف مقاله و همچنین فهرست مراجع استخراج می‌شود. قالب اسناد ورودی PDF است و پردازش با تکیه بر اطلاعات ساختاری مانند مختصات خطوط و فونت انجام می‌شود. فاز استخراج فراداده از صفحه‌ی اول، از سه مرحله تشکیل شده که به ترتیب عنوان، نویسندگان و چکیده به دست می‌آید. روش دستیابی به این فراداده‌ها بر اساس اختلاف مقادیر طول و عرض مختصات واحدهای متنی است. در فاز دوم و پردازش محتوای کامل مقاله، با بهره‌گیری از اطلاعاتی چون هیستوگرام اندازه فونت‌های ظاهر شده در متن، طول کلمات و مختصات نواحی متنی در هر صفحه سعی در دستیابی به بخش‌ها و زیر بخش‌های مقاله، از جمله فهرست مراجع دارند.

استخراج فراداده از مراجع، متکی بر پایگاه دانشی که صورت خودکار و بدون دخالت عامل انسانی تولید می‌شود، ایده‌ای است که در (Cortez E. , da Silva, Goncalves, Mesquita, & de Moura, 2007) مطرح و بررسی شده است. در واقع تفاوت این روش با سایر روش‌هایی که از پایگاه دانش بهره می‌گیرند، این است که در این روش، به جای آن که دانش توسط عامل انسانی تعریف شود، به صورت خودکار و با توجه به فراداده‌های موجود در پایگاه داده جمع‌آوری شده، تولید می‌شود. به این ترتیب که ابتدا رشته متنی به چند فیلد^{۵۶} یا به اصطلاح بلوک تقسیم‌بندی می‌شوند و در ادامه، نوع بلوک‌ها با توجه به پایگاه دانش مشخص می‌شود. بدین صورت، برای هر کلمه می‌توان

⁵⁶ Field

احتمال رخداد در هر یک از فیلدهای فراداده را محاسبه کرد. برای نمونه واژه Halloween به احتمال بالایی در «عنوان» دیده می‌شود. اما کلمه Potter می‌تواند هم در فیلد «عنوان» و هم «نام نویسنده» ظاهر شود و از این نظر ابهام دارد. در این حالت می‌توان با توجه به بلوک‌های مجاور درباره نوع بلوک موردنظر تصمیم‌گیری نمود. بدین شیوه فراداده‌های مراجع از رشته متنی ورودی به دست می‌آید. این کار در (Cortez E. , da Silva, Gonçalves, Mesquita, & de Moura, 2009) نیز بسط داده شده است. طبق نتیجه‌گیری انتهای پژوهش، واضح است باکیفیت بودن داده‌هایی که برای تولید دانش مورد استفاده قرار می‌گیرند از اهمیت بالایی برخوردار است.

استخراج فراداده از پایان‌نامه‌ها با استفاده از قواعد و عبارات منظم^{۵۷} تمرکز اصلی پژوهش (Ojokoh, Adewale, & Samuel oluwale, 2009) است. قالب اسناد پردازش شده Word، PDF و متن هستند. فراداده‌های موردنظر شامل عنوان، فهرست مطالب، چکیده، قدردانی، دیباچه، مقدمه، نتیجه‌گیری و مراجع هستند. در این روش ابتدا طرح پژوهشی دریافت شده به قالب مناسب تبدیل می‌شود. در ادامه فراداده‌ها، بر پایه قواعد تعریف شده، به دست می‌آیند و پس از نمایش، در پایگاه داده ذخیره می‌شوند. در این روش، عبارات منظم با کلمات کلیدی چون «فهرست مطالب»، «چکیده»، «دیباچه»، «فصل اول»، «مقدمه»، «نتیجه‌گیری»، «مراجع»، «فهرست کتب»^{۵۸}، «استناد»^{۵۹} و «پیوست» تعریف شده‌اند و عملیات تطبیق با استفاده از آن‌ها صورت می‌گیرد. از طرفی به دلیل این که در عنوان از کلمه کلیدی خاصی استفاده نمی‌شود و نمی‌توان بر اساس واژه‌های از پیش تعریف شده‌ای عنوان اسناد را به دست آورد، برای استخراج این فراداده به شیوه‌ای متفاوت با سایر فراداده‌ها عمل شده است.

^{۵۷} Regular expressions

^{۵۸} Bibliography

^{۵۹} Citation

در پژوهش (Guo & Jin, 2011) سیستمی عرضه شده که در آن فراداده‌های سرآیند و مراجع مقالات وب استخراج می‌شود. به این صورت که ابتدا یک خزنده^{۶۰} به جستجوی مقالات علمی در وب می‌پردازد. سپس فایل PDF مقالات دانلود شده و به فایل XML تبدیل می‌گردند. در ادامه با توجه به برخی قواعد، علمی یا غیرعلمی بودن فایل به دست آمده بررسی می‌شود. برای نمونه اگر تعداد خطوط سند از ده خط کمتر باشد، این سند مقاله نیست. اگر کلماتی چون IEEE، ACM یا PROCEEDING در اولین صفحه سند یافت شود، این سند می‌تواند یک مقاله باشد. یافتن کلمات introduction و abstract در صفحات اول و یا کلمه references در صفحات پایانی نیز گواهی بر این است که احتمالاً خرنده‌ی وب یک فایل مقاله را به دست آورده است. در پژوهش موردنظر، پس از تایید مقاله بودن فایل، با قواعد تعریف شده و با کمک پایگاه دانش، فراداده‌های سرآیند و مراجع به دست می‌آیند.

ایده اصلی پژوهش (Chen, Yang, Kao, & Ho, 2008) بازنمایی رشته متنی منابع به صورت دنباله پروتئینی و تجزیه الگوی این دنباله است. بدین ترتیب که در مرحله‌ی اول رشته متنی تبدیل به دنباله‌ای از نمادهای اسیدآمین می‌شود و سپس در مرحله‌ی دوم با توجه به الگوهایی که از قبل در یک پایگاه داده جمع‌آوری شده‌اند، شبیه‌ترین الگو به دنباله‌ی رشته متنی یافت شده و فراداده‌های موردنظر به دست می‌آیند. در این پژوهش، با توجه به اطلاعاتی که معمولاً در رشته‌های مراجع مشاهده می‌شوند، ۲۳ نماد اسیدهای آمینه مختلف تعریف شده است. برای نمونه نویسنده، عنوان، مجله و علامت نگارشی نقطه «.» به ترتیب با حروف A، T، L و D نمایش داده می‌شوند. در سیستم ارائه شده، ابتدا توکن‌های رشته متنی ورودی جداسازی می‌شوند. در ادامه، هر توکن به نماد اسید آمینه موردنظر تبدیل می‌گردد و در نتیجه دنباله‌ای پروتئینی حاصل می‌شود. در صورتی که سبک ارجاع‌دهی مشخص باشد و تمامی توکن‌ها به درستی برچسب زده شوند، از این دنباله در فاز بعدی برای استخراج

⁶⁰ Crawler

فراداده‌ها استفاده می‌شود. اگر در فرآیند تبدیل توکن‌ها به دنباله پروتئین، نوع توکن شناخته نشود، نماد X در دنباله قرار می‌گیرد. در این حالت در فاز تجزیه دنباله، نزدیک‌ترین و شبیه‌ترین الگو به این دنباله یافت شده و سپس فراداده‌های موردنظر به دست می‌آیند. در (Ferrés, Saggion, Ronzano, & Bravo, 2018) ابزاری ارائه شده که می‌تواند فایل PDF را به xml تبدیل کند. علاوه بر این، ابزار ارائه شده قادر است محتوای متنی و قسمت‌های مختلف هر مقاله مانند فراداده‌ها، بخش‌ها، زیربخش‌ها و مراجع را شناسایی کند. پژوهش‌هایی که در آن‌ها از رویکرد مبتنی بر قاعده استفاده شده، در جدول ۲-۲ خلاصه شده‌اند.

جدول ۲-۲. روش‌های استخراج فراداده مبتنی بر تعریف قواعد

نام روش	ایده‌ی اصلی	قالب	نوع فراداده
(Klink, Dengel, & Kieninger, 2000)	تعریف قواعد مبتنی بر ویژگی‌های متنی و ساختاری، قابل تطبیق با حوزه‌های مختلف	تصویر	سرآیند
(Giuffrida, Shek, & Yang, 2000)	تولید قواعد براساس دانش قبلی (پایگاه دانش) درباره مکان و نوع فراداده‌ها در اسناد	PostScript	سرآیند
(Mao, Kim, & Thoma, 2004)	تولید خودکار قواعد، با مدل‌سازی ویژگی‌های ساختاری قالب مقالات هر مجله	تصویر	سرآیند
(Huang, Jin, Yuan, & Han, 2006)	استخراج فراداده با تطبیق قالب سند و اطلاعات مرتبط با طرح‌بندی	PDF	سرآیند
(Day, et al., 2007)	تعریف قواعد مبتنی بر متن و براساس روش‌های مرسوم ارجاع‌دهی در مقالات	HTML	مراجع
(Flynn, Zhou, Maly, Zeil, & Zubair, 2007)	استخراج فراداده از اسناد متنوع، با تأکید بر اسناد فرم و غیر فرم	PDF	سرآیند
(Groza, Handschuh, & Hulpus, 2009)	شناسایی فراداده‌ها و بخش‌های مختلف مقاله با بهره‌گیری از اطلاعات ساختاری و سبک نوشتار	PDF	سرآیند
(Cortez E. , da Silva, Goncalves, Mesquita, & de Moura, 2007)	استخراج فراداده از مراجع متکی بر دانش، پایگاه دانش تولید شده به صورت خودکار و بدون دخالت عامل انسانی	متن	مراجع

سرآیند	PDF و Word متن	استخراج فراداده از پایان‌نامه‌های انگلیسی، مبتنی بر قواعد و عبارات منظم	(Ojokoh, Adewale, & Samuel oluwale, 2009)
سرآیند و مراجع	PDF	شناسایی بخش‌های مختلف مقاله و تجزیه‌ی رشته‌ی مراجع مبتنی بر قواعد و پایگاه دانش	(Guo & Jin, 2011)
مراجع	متن	بازنمایی رشته‌ی متنی منابع به صورت دنباله‌ی پروتئینی و تجزیه‌ی الگوی آن	(Chen, Yang, Kao, & Ho, 2008)
سرآیند	PDF	تبدیل PDF به XML و استخراج محتوای آن با تعریف قواعد	(Ferrés, Saggion, Ronzano, & Bravo, 2018)

رویکرد مبتنی بر یادگیری ماشین

در (Peng & McCallum, 2006) سعی شده فراداده‌های صفحات آغازین استخراج شود و همچنین رشته‌های مراجع مقاله تحلیل شوند. در این پژوهش، داده‌های ورودی متنی هستند. به منظور استخراج اطلاعات از سرآیند مقالات، مجموعه داده‌های کلمات آغازین طرح پژوهشیتا اولین بخش مقاله (معمولاً مقدمه) و یا کل کلمات صفحه اول در نظر گرفته شده و برای استخراج فراداده از منابع، رشته متنی منابع مورد استفاده قرار گرفته است. در این روش، ویژگی‌های لغوی، طرح‌بندی و هجی کلمات محاسبه شده و به روش میدان تصادفی شرطی^{۶۱} فراداده‌ها استخراج می‌شوند.

سیستم متن‌باز ارائه شده در (Councill, Giles, & Kan, 2008) با هدف تجزیه و تحلیل رشته‌های مراجع طراحی شده است. در این روش، پس از اعمال تکنیک‌های پردازش زبان طبیعی و محاسبه ویژگی‌های وابسته به متن برای هر واحد متنی، دنباله توکن‌های رشته‌ی مراجع به فراداده‌های موردنظر دسته‌بندی می‌شوند. لازم به ذکر است که طبقه‌بندی با بهره‌گیری از مدل میدان تصادفی شرطی انجام شده است. در (Marinai, 2009) پژوهشی در راستای دستیابی به نام نویسندگان و عنوان مقالات انجام شده که با تحلیل ساختاری فایل‌های PDF و با استفاده از پرسپترون چندلایه، سعی در حل مسئله دارد.

⁶¹ Conditional random field

در این پژوهش، ابتدا مقاله با قالب PDF به عنوان ورودی دریافت شده و پس از تحلیل طرح‌بندی و استخراج محتوا، نواحی متنی شناسایی می‌شوند. در ادامه ویژگی‌های هر ناحیه استخراج شده و دسته‌بند پرسپترون چندلایه نواحی را در کلاس‌های نویسنده، عنوان و سایر، طبقه‌بندی می‌کند. در نهایت، با کمک پایگاه داده‌ای که حاوی مراجع مجلات علوم کامپیوتر است، صحت اطلاعات به دست آمده بررسی می‌شود.

در (Kern, Jack, Hristakeva, & Granitzer, 2012)، نویسندگان برای دستیابی به فراداده‌ها از مقالات PDF یک روش دیگر ارائه کرده‌اند. در این روش، ابتدا بلوک‌های متنی به دست آمده از هر سند در کلاس فراداده‌های موردنظر طبقه‌بندی می‌شوند (بلوک‌های متنی که فراداده نیستند در کلاس «سایر» دسته‌بندی می‌شوند). در گام بعدی، طبقه‌بندی در سطح «توکن»‌ها انجام می‌شود. هر بلوک متنی از چندین «توکن» تشکیل شده که در این قدم هر کدام از کلمات به چندین کلاس طبقه‌بندی می‌شوند. برای مثال بلوک متنی که در گام اول برچسب «نویسنده» گرفته، در گام دوم هر کلمه آن در یکی از کلاس‌های نام، نام خانوادگی، ایمیل، وابستگی و... دسته‌بندی می‌شود. سپس در مرحله بعد فراداده‌های مورد نظر استخراج می‌شود. در این مقاله روش طبقه‌بندی بر اساس ماکسیمم آنتروپی است. در روش پیشنهادی ارائه شده در (Luong, Nguyen, & Kan, 2010) با تحلیل ساختار منطقی سند^{۶۲}، به هر بخش از مقاله بر اساس نقش آن در سند، برچسبی از فراداده‌ها تخصیص می‌یابد. همچنین نویسندگان در (Cuong, Kumar, Kan, & Lee, 2015) نیز با رویکردی مشابه سعی در حل این مسئله کرده و فراداده‌های سرآیند و مراجع از متن مقالات را استخراج کرده‌اند. در این پژوهش، ارتباط‌های مرتبه بالاتر میدان تصادفی شرطی^{۶۳} به عنوان ویژگی در نظر گرفته شده است.

استخراج نام نویسندگان و وابستگی‌ها هدف پژوهش صورت گرفته در (Do, Chandrasekaran, Cho, & Kan, 2013) است. در این پژوهش، قبل از فاز شناسایی فراداده‌ها، فایل PDF به xml تبدیل

^{۶۲} Logical structure analysis

^{۶۳} Higher order semi-CRFs (HO-SCRFs)

می‌شود تا ویژگی‌های مکانی و هندسی محتوا حفظ شود. سپس به هر خط از متن، بر اساس نقش آن در متن مقاله و با استفاده از روش ارائه شده در (Luong, Nguyen, & Kan, 2010) یک برچسب تخصیص می‌یابد. در ادامه خطوطی که یکی از دو برچسب نویسنده و یا وابستگی را داشته باشند وارد سیستم اصلی می‌شوند. این سیستم از دو بخش تشکیل شده: ابتدا نام نویسنده (نویسندگان) و وابستگی‌های آن‌ها که شامل اطلاعاتی چون ایمیل، شماره تلفن، محل خدمت و... هستند شناسایی می‌شود. در گام بعد، وابستگی‌های هر نویسنده با نام او مطابقت می‌یابد. در این پژوهش، بخش شناسایی به روش یادگیری ماشین با نظارت و روش میدان تصادفی شرطی، بخش تطبیق و طبقه‌بندی ارتباطات به روش ماشین بردار پشتیبان انجام شده است.

در (Souza, Moreira, & Heuser, 2014) با توجه به این که فراداده‌های یک مقاله در صفحه اول آن ظاهر می‌شوند تنها صفحه اول مقالات و طی دو فاز تحلیل می‌شوند: ابتدا بخش‌های مختلف صفحه اول - که عمدتاً حاوی فراداده‌های مهم و ارزشمند است - شناسایی می‌شوند. به این منظور، پنج کلاس تعریف شده است:

- سرآیند: معمولاً در بالای صفحه ظاهر شده و شامل اطلاعاتی درباره کنفرانس یا نشریه‌ای است که مقاله در آن‌جا انتشار یافته است. مکان برگزاری، سال چاپ و شماره صفحات نمونه‌هایی از این فراداده‌ها هستند.
- عنوان: موضوع اصلی مقاله یا به عبارتی عنوان در این بخش نوشته می‌شود. به طور معمول عنوان مقاله ممکن است در چند خط و با بزرگ‌ترین فونت در بالای صفحه قرار بگیرد.
- اطلاعات نویسنده: این بخش دربردارنده برخی اطلاعات درباره نویسنده (نویسندگان) است. فراداده‌هایی چون نام، نام خانوادگی، ایمیل، اطلاعات محل خدمت و آدرس از این کلاس قابل استخراج هستند.

▪ بدنه: در این بخش فراداده‌ای که بیانگر اطلاعات توصیفی و مهم مقاله باشد یافت نمی‌شود.

این بخش، شامل بخش‌ها و زیربخش‌های محتوای اصلی مقاله است. از این رو کلاس بدنه

مورد پردازش مجدد جهت استخراج فراداده قرار نمی‌گیرد.

▪ پاورقی: در این بخش ممکن است اطلاعاتی همچون حق کپی‌رایت، نام کنفرانس، صفحات،

سال انتشار، وابستگی نویسندگان و آدرس نوشته شود. پاورقی در گوشه پایین سمت چپ

صفحه با فونت ریز قرار می‌گیرد.

در این فاز، طبقه‌بندی با روش میدان تصادفی شرطی صورت می‌گیرد و ویژگی‌های مبتنی بر قالب

نوشته مانند هم‌ترازی^{۶۴}، درشت‌نمایی^{۶۵}، خط زیرین^{۶۶}، حروف کج^{۶۷} و اندازه فونت در

تصمیم‌گیری‌های مدل میدان تصادفی شرطی مورد استفاده قرار می‌گیرند. در ادامه و در مرحله دوم

مدل دیگری – که آن هم با استفاده از روش میدان تصادفی شرطی به دست آمده – برای استخراج

فراداده‌های بخش‌های سرآیند، اطلاعات نویسنده و پاورقی پردازش می‌شود. به دلیل این که کلاس

«عنوان» تنها شامل یک فراداده یعنی عنوان مقاله است، نیازی به پردازش مجدد در فاز دوم ندارد. در

این مرحله، از سرآیند مقاله فراداده‌های نام کنفرانس، سال، تاریخ، مکان و ناشر به دست می‌آید.

دستیابی به نام نویسنده(ها)، محل خدمت و ایمیل با پردازش کلاس اطلاعات نویسنده‌گان امکان‌پذیر

است. گاهی ممکن است که برخی فراداده‌ها به جای بالای صفحه در پاورقی نوشته شوند؛ بنابراین با

پردازش این بخش می‌توان به اطلاعاتی چون نام همایش، سال، تاریخ، مکان، ناشر، شماره استاندارد

بین‌المللی کتاب^{۶۸} و رایانامه دست یافت. در این روش، در فاز دوم از ویژگی‌هایی چون تعداد لغات و

وجود برخی کلمات کلیدی رزرو شده استفاده شده است.

⁶⁴ Alignment

⁶⁵ Bold

⁶⁶ Underline

⁶⁷ Italic

⁶⁸ ISBN

سیستم ارائه شده در (Tkaczyk, Szostek, Fedoryszak, Dendek, & Bolikowski, 2015) سعی در استخراج فراداده از صفحات آغازین و مراجع مقالات چاپ شده در نشریات دارد. این نرم‌افزار به صورت متن‌باز در دسترس بوده^{۶۹} و قادر است مقالات با قالب PDF را پردازش کند. روش پیشنهاد شده برای مقالاتی است که از ابتدا در قالب PDF بوده‌اند و نیازی نیست تا محتوای متنی آن‌ها با کمک نویسه‌خوان‌نوری به دست آید. فرآیند این سیستم بدین صورت است که نخست محتوای متنی و ویژگی‌های تصویری فایل PDF به دست آمده و در ادامه، خروجی به مراحل استخراج ساختار پایه^{۷۰}، استخراج فراداده^{۷۱} و استخراج فهرست مقالات ارجاعی^{۷۲} هدایت می‌شود. در مرحله استخراج ساختار پایه، ساختار سلسله‌مراتبی سند تولید می‌شود و پس از آن نواحی مقاله به چهار کلاس فراداده، بدنه، مراجع و سایر، طبقه‌بندی می‌شود. سپس نواحی با برچسب «فراداده» وارد مرحله استخراج فراداده شده و خود به انواع فراداده‌های از قبل تعریف شده طبقه‌بندی می‌گردند. در هر دو مرحله استخراج ساختار پایه و فراداده، طبقه‌بندی با روش ماشین بردار پشتیبان (Cortes & Vapnik, 1995) انجام شده است. در بخش استخراج فهرست مقالات ارجاعی، رشته‌های مراجع تجزیه شده و فراداده‌های منابع به دست می‌آید. در این بخش، به منظور تقسیم‌بندی نواحی متنی مراجع، از روش خوشه‌بندی k-means استفاده شده است و در نهایت فراداده‌های هر رشته متنی، با کمک میدان تصادفی شرطی برچسب‌زنی می‌شوند.

تطبیق گراف ساختار^{۷۳} سند، رویکرد استفاده شده توسط پژوهشگران در (Kooli & Belaïd, 2016) برای استخراج موجودیت‌ها است. در این روش موجودیت‌های مقالات علمی، اسناد سازمانی و فاکتور خرید بررسی شده است. به این ترتیب که پس از استخراج متن از تصویر ورودی، با استفاده از عبارات منظم و لغت‌نامه به موجودیت‌ها برچسب زده می‌شود. این برچسب‌زنی صرفاً بر اساس متن

⁶⁹ <http://cermine.ceon.pl/index.html>

⁷⁰ Basic structure extraction

⁷¹ Metadata extraction

⁷² Bibliography extraction

⁷³ Structure graph matching

انجام می‌شود و ویژگی‌های تصویری و ساختاری در مراحل بعدی مورد استفاده قرار می‌گیرد. در ادامه ساختارهای محلی هر بخش از تصویر سند، با یک گراف مدل می‌شود و با مدل‌هایی که سیستم از قبل یادگرفته (یادگیری به روش بدون نظارت و خوشه‌بندی حول رهبر^{۷۴}) نزدیک‌ترین مدل یافت شده و از آن برای اصلاح خطاهای احتمالی و به‌روزرسانی مدل یادگیری نیز استفاده می‌شود. در آخرین مرحله موجودیت‌ها شناسایی می‌شوند.

در (Liu, Gao, An, Jiang, & Tang, 2017) تلاش شده با به کارگیری شبکه‌های عمیق و ترکیب ویژگی‌های متنی و تصویری، فراداده‌ها از بخش‌های ابتدایی فایل PDF مقالات به دست آیند. در این روش با استفاده از شبکه عمیق کانولوشن^{۷۵} ویژگی‌های متنی و تصویری به دست آمده و خروجی به شبکه حافظه طولانی کوتاه مدت^{۷۶} فرستاده می‌شود. مطالعه دیگری که بر استخراج فراداده‌های رشته‌های متنی مراجع تمرکز داشته مقاله (An, Gao, Jiang, Liu, & Tang, 2017) است. در این روش از ترکیب مدل شبکه حافظه طولانی کوتاه مدت و میدان تصادفی شرطی برای دستیابی به فراداده‌ها استفاده شده است. همچنین مدلی برای برچسب زدن به دنباله قطعات متنی ارائه شده که برای مسئله تجزیه متن مراجع مناسب است. محققان در (Qiu & Zhou, 2019) بر اسناد از نوع کتاب‌های دیجیتال تاکید داشته و هدف آن‌ها دستیابی به فراداده‌های کتاب‌های دیجیتال است. این فراداده‌ها شامل عنوان کتاب، نویسنده، ناشر، تاریخ انتشار، شماره استاندارد بین‌المللی کتاب^{۷۷} (شابک) و ... هستند. طبقه‌بندی فراداده‌ها به روش ماشین بردار پشتیبان و مبتنی بر درخت دودویی انجام می‌شود. به این ترتیب که تمامی دسته‌ها به دو کلاس طبقه‌بندی می‌شوند و در ادامه هر کدام خود به زیرکلاس‌های دودویی طبقه‌بندی می‌گردند. این کار تا زمانی که تمامی گره‌های درخت دودویی تنها شامل یک کلاس باشند، ادامه می‌یابد.

⁷⁴ Leader clustering

⁷⁵ Convolutional neural network (CNN)

⁷⁶ Long short-term memory (LSTM)

⁷⁷ ISBN

روش پیشنهادی در (Boukhers, Ambhore, & Staab, 2019) بر شناسایی بخش مراجع، قطعه‌بندی و شناسایی فراداده‌های این بخش تمرکز دارد. در یک فرآیند آفلاین، با بهره‌گیری از روش جنگل تصادفی^{۷۸}، یک مدل یادگیری برای طبقه‌بندی خطوط به دست می‌آید. این دسته‌بند خطوط را در چهار کلاس «غیر مرجع»، «خط اول مرجع»، «خط میانه‌ی مراجع» و «آخرین خط مرجع» طبقه‌بندی می‌کند. همچنین با استفاده از میدان تصادفی شرطی، مدل یادگیری دیگری برای قطعه‌بندی و شناسایی بخش‌های مختلف رشته‌های متنی مرجع تولید می‌شود. همانطور که اشاره شد هر دو مدل یادگیری به صورت آفلاین تولید می‌شوند و از آن‌ها برای شناسایی و استخراج فراداده از خطوط مراجع مقاله ورودی استفاده می‌شود. در (Rizvi, Dengel, & Ahmed, 2019) روشی دیگر با بهره‌گیری از شبکه‌های عمیق برای شناسایی بخش مراجع مقالات ارائه شده که مبتنی بر ویژگی‌های ساختاری عمل می‌کند. در این روش، ابتدا طرح‌بندی مقاله تحلیل شده و سند مقاله از نظر تک ستونه یا چند ستونه بودن بررسی می‌شود. در ادامه، بخش مراجع تشخیص داده می‌شود. همچنین در این مقاله مجموعه داده‌ای جمع‌آوری شده که در آن درباره نحوه تهیه داده‌ها توضیح داده شده است.

چارچوب پیشنهاد شده در (Ahmed & Afzal, 2020) برای استخراج فراداده‌ها از قالب PDF است. در این روش پس از محاسبه ویژگی‌ها، الگوریتم‌های مختلفی برای یادگیری به کار گرفته شده و مقایسه‌ای بین آن‌ها صورت گرفته است. از جمله این الگوریتم‌ها می‌توان به نایو بیز^{۷۹}، ماشین بردار پشتیبان، k نزدیک‌ترین همسایه، درخت تصمیم و چند طبقه‌بند دیگر اشاره کرد. همچنین ترکیب طبقه‌بندها نیز مورد توجه قرار گرفته است. لازم به ذکر است که طبق نتایج به دست آمده در این مقاله، ماشین بردار پشتیبان بهترین نتیجه را نسبت به سایر روش‌ها داشته است.

به عنوان نمونه‌ای از پژوهش‌های داخلی صورت گرفته در استخراج فراداده از مقالات فارسی می‌توان به پژوهش صورت گرفته توسط (تن سازان & مهدوی، ۱۳۹۶) اشاره کرد. چارچوب ارائه شده در

⁷⁸ Random forest

⁷⁹ Naïve Bayes

این مقاله برای دستیابی به اطلاعات سرآیند و مراجع مقالات طراحی شده است. در این روش، در ابتدا، بخش‌های سرآیند مقاله که در ابتدای مقاله نوشته می‌شوند، از متن اصلی مقاله جدا می‌شود. همچنین قسمت مراجع که در انتهای هر مقاله‌ای قرار می‌گیرد، شناسایی می‌شود. سپس فراداده‌های سرآیند شامل عنوان، نویسندگان، وابستگی‌ها، چکیده، کلمات کلیدی و اطلاعات چاپ مقاله به دست می‌آید. فراداده‌های مراجع نیز از مراجع انگلیسی و فارسی به صورت جداگانه استخراج می‌شود. در این روش، برچسب‌زدن به فراداده‌ها با میدان تصادفی شرطی انجام شده است. مقایسه روش‌هایی که از رویکرد مبتنی بر یادگیری ماشین برای استخراج فراداده از اسناد علمی استفاده می‌کنند، در **جدول ۳-۲** به صورت خلاصه بیان شده است.

جدول ۳-۲. روش‌های استخراج فراداده با رویکرد یادگیری ماشین

نام روش	ویژگی‌ها	قالب	نوع فراداده	الگوریتم
(Peng & McCallum, 2006)	لغوی، همجی کلمات و طرح‌بندی	متن	سرآیند، مراجع	<i>CRF</i>
(Council, Giles, & Kan, 2008)	لغوی	متن	مراجع	<i>CRF</i>
(Marinai, 2009)	هندسی، لغوی و همسایگی	<i>PDF</i>	سرآیند	<i>MLP</i>
(Kern, Jack, Hristakeva, & Granitzer, 2012)	مدل زبانی، لغوی، طرح‌بندی و هیوریستیک	<i>PDF</i>	سرآیند	<i>MaxEnt</i>
(Luong, Nguyen, & Kan, 2010)	ساختاری و کلمات	<i>PDF</i>	سرآیند	<i>CRF</i>
(Cuong, Kumar, Kan, & Lee, 2015)	مراتب بالاتر میدان تصادفی شرطی	متن	سرآیند، مراجع	<i>CRF</i>
(Souza, Moreira, & Heuser, 2014)	قالب نوشتار، مختصات، لغات و کلمات	<i>PDF</i>	سرآیند	<i>CRF</i>
(Tkaczyk, Szostek, Fedoryszak, Dendek, & Bolikowski, 2015)	هندسی، لغوی، توالی، قالب نوشتار و هیوریستیک	<i>PDF</i>	سرآیند، مراجع	<i>SVM, CRF, K-means, Rules</i>
(Kooli & Belaïd, 2016)	متن و گراف ساختار	تصویر	سرآیند	خوشه‌بندی حول رهبر، الگوریتم پیشنهادی
(Liu, Gao, An, Jiang, & Tang, 2017)	متن و تصویر	<i>PDF</i>	سرآیند	<i>CNN, LSTM</i>

<i>Bi-LSTM CRF</i>	مراجع	متن	متن	(An, Gao, Jiang, Liu, & Tang, 2017)
<i>SVM</i>	سرآیند	متن، <i>PDF</i> ، <i>HTML</i>	متنی و ساختاری	(Qiu & Zhou, 2019)
درخت تصادفی، <i>CRF</i>	مراجع	<i>PDF</i>	قالب، لغوی، معنایی و شکل	(Boukhers, Ambhore, & Staab, 2019)
<i>R-CNN</i>	مراجع	تصویر	ویژگی‌های متن و تصویر به صورت خودکار در فرآیند یادگیری به دست می‌آیند	(Rizvi, Dengel, & Ahmed, 2019)
ترکیب طبقه‌بندها	سرآیند	<i>PDF</i>	فونت و سبک آن، فاصله‌ی اجزای همسایه، مکان، لغوی	(Ahmed & Afzal, 2020)
<i>CRF</i>	سرآیند، مراجع	متن	ویژگی‌های متنی	(تن سازان & مهدوی، ۱۳۹۶)

۲-۵-۲. پژوهش‌های انجام شده در زمینه‌ی استخراج بینش‌های کلیدی از مقالات علمی

تحقیقات صورت گرفته در دستیابی به بینش‌های کلیدی و اطلاعات بالقوه نوشتارهای دانشگاهی و علمی عمدتاً به تحلیل متن مقالات از جمله بخش چکیده و نتیجه‌گیری می‌پردازند و متون را در دو سطح جمله^{۸۰} و عبارت^{۸۱} پردازش می‌کنند (Nasar, Jaffry, & Malik, 2018). روش‌های استخراج بینش از مقالات علمی نیز به دو دسته مبتنی بر قاعده و مبتنی بر یادگیری تقسیم می‌شوند.

رویکرد مبتنی بر قاعده

در (Gupta & Manning, 2011)، روشی برای استخراج اطلاعاتی مانند نوآوری، روش‌ها، ابزارهای به کار گرفته شده در حل مسئله و حوزه مسئله ارائه شده است. در این روش، جملات بخش چکیده تحلیل شده، درخت تجزیه‌ای برای نمایش ارتباط بین کلمات هر جمله به دست می‌آید و اطلاعات موردنظر با انطباق با الگوهایی از قبل تعریف شده استخراج می‌شوند. در صورتی که الگوی جدیدی در

⁸⁰ Sentence level

⁸¹ Phrase level

جملات مشاهده شود، این الگو به الگوهای پیشین افزوده خواهد شد. در پژوهش (Hanyurwimfura, Liao, Njogu, & Ndatinya, 2012) شیوه‌ای برای خلاصه‌سازی مقالات پیشنهاد شده که با تحلیل بخش‌های چکیده و نتیجه‌گیری مقالات، اطلاعات مختصری درباره محتوای مقاله به دست می‌آورد. این اطلاعات به خواننده مقاله در انتخاب مقاله متناسب با اهداف خود کمک می‌کند. در این روش قواعدی تعریف شده تا نتایج، ارزیابی و ایده اصلی مقاله استخراج شود. برای مثال کلماتی چون «نتایج»، «آزمایش‌ها» یا «ارزیابی» می‌تواند در دستیابی به یافته‌های مقاله مؤثر باشد. به عنوان نمونه‌ای دیگر، جملاتی که با عبارتی مانند «این مقاله» یا «رویکرد ما» شروع می‌شوند به احتمال زیاد جملاتی هستند که درباره‌ی ایده اصلی مقاله بحث می‌کنند.

روش پیشنهادی در (Houngbo & Mercer, 2012) سعی در استخراج تکنیک‌ها و روش‌ها از مقالات تحقیقاتی در حوزه زیست‌شناسی دارد. بدین‌منظور جملاتی که احتمال دارد درباره روش به کار رفته در مقاله بحث کرده باشند، استخراج می‌شوند. برای دستیابی به این جملات، دو رویکرد مورد توجه قرار گرفته شده: رویکرد مبتنی بر قاعده که بر اساس عبارت منظمی که از قبل تعریف شده، جملاتی که در آن‌ها کلماتی چون «روش»، «تحلیل»، «الگوریتم»، «رویکرد» و «مدل» یافت شود بررسی شده و در صورتی که با الگوی تعریف شده مطابقت پیدا کرد، به عنوان جمله کاندید انتخاب می‌شود. رویکرد دوم برای حالاتی است که کلمات کلیدی موردنظر به طور صریح در جملات یافت نشود؛ در این صورت با رویکرد یادگیری ماشین و الگوریتم میدان تصادفی شرطی، جملاتی که درباره روش تحقیق مقاله توضیح داده‌اند، استخراج می‌شوند. ویژگی‌هایی که در رویکرد دوم مورد استفاده قرار گرفته‌اند، عبارتند از: نقش کلمات، شکل کلمات (وجود حروف بزرگ یا کوچک در زبان انگلیسی)، مکان کلمه (ابتدا، انتها یا میانه جمله)، وجود پسوند و پیشوند و عبارت دو کلمه‌ای.

رویکرد مبتنی بر یادگیری ماشین

در پژوهش (Teufel & Moens, 2002) شیوه‌ای برای خلاصه‌سازی مقالات علمی پیشنهاد شده که جملات را استفاده از الگوریتم نایو بی‌زین^{۸۲} به چهار کلاس هدف، پیشینه، زمینه و مغایرت دسته‌بندی می‌کند. به این ترتیب که هدف و مقصود اصلی مقاله و تاریخچه موضوع مورد بحث استخراج می‌شود. همچنین اطلاعاتی درباره این که مقاله مورد نظر هم‌راستا با کدام پژوهش‌ها است (کلاس زمینه) و یا با کدام یک از مقالات پیشین در تقابل است، به دست می‌آید. این روش در (Teufel, Siddharthan, & Batchelor, 2009) توسعه داده شده است و هر بخش نیز خود به کلاس‌های دیگری طبقه‌بندی می‌شود. دستیابی به هدف پژوهش، روش تحقیق، نتیجه و یافته‌ها از متن چکیده مقالات علوم پزشکی هدف پژوهش (Ruch, et al., 2007) است. در این روش نیز الگوریتم نایو بی‌زین برای طبقه‌بندی جملات به کار گرفته شده است.

یکی دیگر از روش‌هایی که متن مقالات را در سطح جمله تحلیل می‌کند، روش ارائه شده در (Lin, Karakos, Demner-Fushman, & Khudanpur, 2006) است. در این پژوهش جملات موجود در چکیده مقاله‌های حوزه پزشکی با روش مدل مخفی مارکوف به بخش‌های مقدمه، روش، یافته‌ها و نتیجه‌گیری تقسیم می‌شوند. همچنین در این روش، تحلیل تشخیصی خطی^{۸۳} با مدل مخفی مارکوف ترکیب شده که این تکنیک در بهتر شدن نتایج تأثیر بسزایی داشته است.

تحقیق دیگری با تمرکز بر تحلیل و طبقه‌بندی بخش چکیده مقالات در (Hirohata, Okazaki, Ananiadou, & Ishizuka, 2008) ارائه شده است. با علم به این که چکیده معمولاً با جملاتی درباره زمینه و هدف موضوع مورد نظر شروع شده، در ادامه درباره روش و یافته‌های به دست آمده توضیح مختصری داده شده و در نهایت با نتیجه‌گیری خاتمه می‌یابد، تلاش شده تا جملات چکیده به چهار کلاس هدف، روش، یافته‌ها و نتیجه دسته‌بندی شوند. به دلیل ذات این مسئله که جملات مرتبط با هم، پشت سر هم قرار می‌گیرند، برای طبقه‌بندی الگوریتم میدان تصادفی شرطی استفاده شده است.

⁸² Naïve Bayesian

⁸³ Linear Discriminant Analysis (LDA)

در (Lin, Ng, Pradhan, Shah, Pietrobon, & Kan, 2010) شیوه‌ای پیشنهاد شده که با استفاده از مدل میدان تصادفی شرطی، فراداده‌ها و بینش‌های کلیدی از مقالات حوزه پزشکی استخراج می‌شود. در این روش علاوه بر فراداده‌های «نویسندگان» و اطلاعاتی چون «رایانامه» و «محل خدمت»، اطلاعات دیگری که در مقالات پزشکی اهمیت دارند نیز به دست می‌آیند. پارامترهایی چون محدوده سنی افرادی که مورد مطالعه قرار گرفته‌اند، نام روش یا نرم‌افزاری که برای تحلیل داده‌ها استفاده شده، شیوه جمع‌آوری داده‌ها، نام پایگاه داده، نوع داده‌های مورد مطالعه، ناحیه جغرافیایی، دارویی که در روش پژوهش دخیل بوده، متغیرهایی که در طول مشاهدات جمع‌آوری شده‌اند، تعداد مشاهدات و در نهایت مدت زمان مطالعه مواردی هستند که در این روش از نوشتارهای علمی حوزه پزشکی به دست می‌آیند. همان‌طور که اشاره شد در این روش برچسب‌زنی با مدل میدان تصادفی شرطی با در نظر گرفتن ویژگی‌های لغوی و ساختاری انجام شده‌است.

مطالعه دیگری در تحلیل عناوین مقالات دو زبانه انگلیسی-ژاپنی انجام شده‌است (Kondo, Nanba, Takezawa, & Okumura, 2011) که در آن عنوان مقاله در سطح عبارت^{۸۴} پردازش می‌شود تا هدف پژوهش و تکنیک به کاررفته استخراج شود. برای این کار، عباراتی که به طور معمول در عنوان مقالات ظاهر می‌شوند، جمع‌آوری شده‌اند. برای مثال «مطالعه‌ای بر»، «بر اساس»، «با استفاده از» و ... عباراتی هستند که ممکن است در عنوان نوشتارهای علمی دیده شوند. با کمک این عبارات می‌توان حوزه و روش به کار گرفته شده در مقاله موردنظر را به دست آورد. طبقه‌بندی کلمات متوالی با روش میدان تصادفی شرطی انجام شده و در نهایت هدف اصلی، تکنیک و حوزه تحقیق به دست می‌آید. پژوهش دیگری در (Kovačević, Konjović, Milosavljević, & Nenadic, 2012) صورت گرفته است که متن مقالات را تحلیل کرده و اطلاعات موردنظر را استخراج می‌کند. این سیستم از دو بخش تشکیل شده است؛ ابتدا جملات بر اساس روش ارائه شده در (Teufel & Moens, 2002) – که

⁸⁴ Phrase level

قبلاً توضیح داده شد- طبقه‌بندی می‌شوند. سپس در مرحله بعد، جملاتی که درباره روش‌شناسی هستند خود به عبارت‌هایی تقسیم شده و عبارات با مدل میدان تصادفی شرطی، به کلاس‌های روش، وظیفه^{۸۵}، منبع^{۸۶} و پیاده‌سازی دسته‌بندی می‌شوند.

روش ارائه شده در (Guo, Korhonen, Liakata, Karolinska, Sun, & Stenius, 2010) تمرکز بر تحلیل جملات چکیده مقالات زیست-پزشکی داشته و در سه شیما^{۸۷}ی مختلف انجام شده است:

▪ نام بخش^{۸۸}: این شیما در (Hirohata, Okazaki, Ananiadou, & Ishizuka, 2008) ارائه

شده و همانطور که قبلاً به این پژوهش اشاره شد، تمرکز آن بر دسته‌بندی جملات در چهار کلاس هدف، روش، یافته‌ها و نتیجه چکیده است.

▪ نواحی استدلالی^{۸۹}: این اصطلاح اولین بار در پژوهش (Teufel & Moens, Summarizing

Scientific Articles: Experiments with Relevance and Rhetorical Status, 2002)

مطرح شد. درباره روش پیشنهادی مقاله مذکور نیز همانطور که پیش‌تر توضیح داده شد، با طبقه‌بندی جملات متن مقاله به چهار دسته هدف، پیشینه، زمینه و مغایرت، در راستای خلاصه‌سازی مقالات تلاش شده است.

▪ مفاهیم علمی مهم^{۹۰}: اصطلاح به کاررفته در (Liakata, Teufel, Siddharthan, &

Batchelor, 2010) که تحلیل مقالات مبتنی بر مفاهیم علمی قابل فهم انسان بوده و بر این

اساس مفاهیمی چون انگیزه تحقیق، هدف، موجودیتی که موضوع و بحث اصلی مقاله است (شی^{۹۱})، روش، آزمایش، مشاهده، یافته‌ها و نتیجه‌گیری به دست می‌آید.

⁸⁵ Task

⁸⁶ Resource

⁸⁷ Scheme

⁸⁸ Section names

⁸⁹ Argumentative zones (AZ)

⁹⁰ Core Scientific Concepts (coreSC)

⁹¹ Object

در هر سه شمای نام برده، الگوریتم‌های ماشین بردار پشتیبان و نایو بی‌زین برای طبقه‌بندی به کار گرفته شده است و در هر سه آزمایش، ماشین بردار پشتیبان عملکرد بهتری داشته است.

در (Fisas, Saggion, & Ronzano, 2015)، پژوهشگران علاوه بر ارائه مجموعه داده‌ای متشکل از متون علمی، آزمایشی برای طبقه‌بندی جملات این مجموعه متون انجام دادند. به این ترتیب بعد از پیش‌پردازش متن و استخراج ویژگی‌ها، جملات به دسته‌های رویکرد، پیشینه، چالش‌ها، پژوهش‌های آینده و نتایج دسته‌بندی می‌شوند. در طبقه‌بندی جملات، دو دسته‌بند ماشین بردار پشتیبان و لجستیک رگرسیون مورد استفاده قرار گرفته و عملکرد آن‌ها با یکدیگر مقایسه شده است.

در (Wu, Chang, Liou, & Chang, 2006) هدف تحلیل جملات چکیده نوشتارهای علمی است. به‌طوری‌که پیشینه، هدف، شیوه و روش به‌کاررفته، یافته‌ها و نتایج به‌دست‌آمده از چکیده استخراج می‌شود. برای این کار، چکیده مقالات از وب جمع‌آوری شده و بخش کوچکی از داده‌ها به صورت دستی برچسب‌گذاری می‌شوند. از این داده‌های برچسب‌دار برای یادگیری داده‌های باقی‌مانده استفاده می‌شود و فرآیند یادگیری خودکار با مدل مخفی مارکوف انجام شده است. در پژوهش دیگری مطالعه‌ای بر استخراج حوزه تحقیق از متن مقالات صورت گرفته است (Lakhanpal, Gupta, & Agrawal, 2015). این کار با ترکیب روش‌های پردازش زبان طبیعی و یادگیری ماشین انجام شده و برای طبقه‌بندی عبارات از الگوریتم نایو بی‌زین استفاده شده است.

چارچوب پیشنهاد شده در (Ronzano & Saggion, 2015) برای تحلیل و دستیابی به اطلاعات متون علمی در قالب‌های متن و PDF طراحی شده است و این قابلیت را دارد تا جملات مقاله را تحلیل کند و اطلاعاتی از بدنه اصلی مقاله و بخش مراجع به دست آورد. در این سیستم، تجزیه مراجع با استفاده از ابزارهای آنلاین انجام می‌گیرد و فراداده‌های مربوطه استخراج می‌شود. همچنین جملات متن اصلی با هدف دستیابی به اطلاعاتی چون رویکرد، چالش‌ها، پیشینه، نتایج و پیشنهاد پژوهش‌های آتی پردازش می‌شود. به این ترتیب که پس از جدا کردن جملات و اعمال تکنیک‌های پردازش زبان طبیعی، جملات با استفاده از الگوریتم ماشین بردار پشتیبان به کلاس‌های مذکور دسته‌بندی می‌شوند.

در انتهای پردازش سند مقاله، مدل به‌دست‌آمده از محتوای مقاله با گراف‌های موسوم به گراف دانش^{۹۲} بازنمایی می‌شود.

روش پیشنهادی در (Meng, Hou, Yang, & Li, 2018) برای استخراج اطلاعاتی چون مسئله، روش، مجموعه داده و سایر (غیرفراداده) ارائه شده و با ترکیب روش‌های پردازش زبان طبیعی و یادگیری ماشین سعی در دستیابی به اطلاعات یاد شده دارد. چارچوب پیشنهاد شده در سه گام تعریف شده است: در اولین قدم تمامی کاندیدهای کلمات کلیدی مشخص می‌شوند. سپس در مرحله بعد هر کدام از کلمات کلیدی با بردار ویژگی نمایش داده می‌شوند تا در گام سوم طبقه‌بندی شوند. الگوریتم‌های یادگیری به کارگرفته در طبقه‌بندی شامل نایو بیزین، درخت تصادفی، نزدیک‌ترین همسایه، طبقه‌بندی بوستینگ تطبیقی^{۹۳} (مبتنی بر یادگیری جمعی^{۹۴})، درخت تصمیم و رگرسیون خطی خطی هستند که در بین آن‌ها روش درخت تصادفی بهترین نتایج را داشته است.

استخراج شبه کد و اطلاعات مربوط به آن‌ها هدف اصلی پژوهش (Safder, Hassan, Visvizi, Noraset, Nawaz, & Tuarob, 2020) است. به این ترتیب که پس از تبدیل *PDF* ورودی به قالب مناسب و دستیابی به محتوای متنی و ویژگی‌های ساختاری، تمامی عناصر غیرمتنی حذف می‌شود. آنگاه، در دو مرحله، ابتدا شبه کد و سپس فراداده‌های آن- یعنی جملاتی که درباره شبه کد و الگوریتم آن بحث می‌کنند- استخراج می‌شوند. به منظور شناسایی شبه کد در متن مقالات، برای خط بردار ویژگی محاسبه شده و طبقه‌بندی با چند الگوریتم پیاده‌سازی شده که بهترین عملکرد در درخت تصادفی مشاهده شده است. همچنین برای استخراج فراداده‌های شبه کد، از رویکرد شبکه‌های عمیق استفاده گردیده است. بدین شیوه که جملات با شبکه کانولوشن به بردار تبدیل شده (تعبیه‌سازی جملات^{۹۵}) و سپس بردارها به شبکه حافظه طولانی کوتاه مدت هدایت می‌شوند.

^{۹۲} Knowledge graph

^{۹۳} Adaboost

^{۹۴} Ensemble learning

^{۹۵} Sentence level embedding

خلاصه روش‌های استخراج بینش‌های کلیدی در **جدول ۲-۴** به همراه رویکردها، الگوریتم‌ها و نوع اطلاعات به دست آمده قابل مشاهده است.

جدول ۲-۴. روش‌های استخراج بینش‌های کلیدی از مقالات

نام روش	رویکرد	الگوریتم یادگیری	فراداده‌ها
(Teufel & Moens, Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status, 2002)	یادگیری ماشین	نایو بیزین	هدف، پیشینه، زمینه و مغایرت با سایر زمینه‌ها
(Ruch, et al., 2007)	یادگیری ماشین	نایو بیزین	هدف پژوهش، روش تحقیق، نتیجه و یافته‌ها
(Lin, Karakos, Demner-Fushman, & Khudanpur, 2006)	یادگیری ماشین	مدل مخفی مارکوف	مقدمه، روش، یافته‌ها و نتیجه‌گیری
(Hirohata, Okazaki, Ananiadou, & Ishizuka, 2008)	یادگیری ماشین	میدان تصادفی شرطی	هدف، روش، یافته‌ها و نتیجه
(Lin, Ng, Pradhan, Shah, Pietrobon, & Kan, 2010)	یادگیری ماشین	میدان تصادفی شرطی	نویسندگان، وابستگی‌ها و اطلاعات مقالات پزشکی
(Kondo, Nanba, Takezawa, & Okumura, 2011)	یادگیری ماشین	میدان تصادفی شرطی	هدف و تکنیک
(Gupta & Manning, 2011)	مبتنی بر قواعد	-	نوآوری، روش‌ها، ابزارها و حوزه‌ی مسئله
(Kovačević, Konjović, Milosavljević, & Nenadic, 2012)	یادگیری ماشین	میدان تصادفی شرطی	روش، وظیفه، منابع و پیاده‌سازی
(Hanyurwimfura, Liao, Njogu, & Ndatinya, 2012)	مبتنی بر قواعد	-	نتایج، ارزیابی و ایده‌ی اصلی مقاله
(Guo, Korhonen, Liakata, Karolinska, Sun, & Stenius, 2010)	یادگیری ماشین	نایو بیزین و ماشین بردار پشتیبان	هدف، روش، پیشینه، زمینه، مغایرت با سایر حوزه‌ها، انگیزه تحقیق، آزمایش، مشاهده، یافته‌ها و نتیجه‌گیری
(Houngbo & Mercer, 2012)	مبتنی بر قواعد	-	تکنیک‌ها و روش‌ها
(Fisas, Saggion, & Ronzano, 2015)	یادگیری ماشین	ماشین بردار پشتیبان و لجستیک رگرسیون	رویکرد، پیشینه، چالش‌ها، پژوهش‌های آینده و نتایج
(Wu, Chang, Liou, & Chang, 2006)	یادگیری ماشین	مدل مخفی مارکوف	پیشینه، هدف، شیوه و روش، یافته‌ها و نتایج

حوزه‌ی تحقیق	نایو بیزین	یادگیری ماشین	(Lakhanpal, Gupta, & Agrawal, 2015)
رویکرد، چالش‌ها، پیشینه، نتایج و پیشنهاد پژوهش‌های آتی	ماشین بردار پشتیبان	یادگیری ماشین	(Ronzano & Saggion, 2015)
مسئله، روش، مجموعه‌داده	درخت تصادفی	یادگیری ماشین	(Meng, Hou, Yang, & Li, 2018)
شبه‌کد و فراداده‌های آن	درخت تصادفی، شبکه‌های عمیق	یادگیری ماشین	(Safder, Hassan, Visvizi, Noraset, Nawaz, & Tuarob, 2020)

۶-۲. مروری بر الگوریتم‌های به‌کارگرفته پژوهش

این بخش به مرور الگوریتم‌های مورد استفاده در این تحقیق اختصاص دارد. بدین منظور، در ابتدا مفاهیم اولیه طبقه‌بند ماشین بردار پشتیبان بیان شده و سپس به الگوریتم خفاش اشاره خواهد شد. از این الگوریتم فرامکاشفه‌ای برای بهینه‌سازی پارامترهای ماشین بردار پشتیبان استفاده شده است. در خاتمه نیز روشی برای ترکیب این دو الگوریتم بیان می‌شود.

۶-۲-۱. ماشین بردار پشتیبان

دسته‌بند ماشین بردار پشتیبان اولین بار در سال ۱۹۹۵ پیشنهاد گردید (Cortes & Vapnik, 1995). این الگوریتم ذاتاً برای مسائل دو کلاسه و داده‌های جداپذیر خطی مناسب است؛ گرچه روش‌هایی برای تعمیم آن در مسائل طبقه‌بندی چند کلاسه و جداسازی داده‌های غیرخطی وجود دارد. هدف طبقه‌بند ماشین بردار پشتیبان، یافتن بهترین ابرصفحه جداکننده است، به‌طوری‌که بیشترین فاصله را از داده‌های هر دو کلاس داشته باشد؛ همچنین کمترین خطا را در طبقه‌بندی داده‌های کلاس‌ها داشته باشد. برای بیان ریاضی مسئله، فرض می‌شود که n نمونه با مجموعه $X = \{(x_i, y_i)\}_{i=1}^n$ نشان داده می‌شود. به‌طوری‌که $x_i \in \mathbb{R}^m$ نمونه‌ای در فضای m بعدی است و $y_i \in \{-1, 1\}$ برچسب کلاس نمونه x_i است. ابرصفحه موردنظر در حالت خطی به صورت زیر است:

$$w^T X + b = 0 \quad (۱-۲)$$

در این معادله، بردار وزن w و پارامتر بایاس b هدف دسته‌بند ماشین بردار پشتیبان یافتن مقادیر این دو پارامتر است، به صورتی که ابرصفحه از نمونه‌های هر دو کلاس بیشترین فاصله را داشته باشد. به ترتیب دو ابرصفحه دیگر به صورت زیر تعریف می‌شود:

$$\begin{aligned} H_1: w^T x_i + b &= +1 & \text{for } y_i &= +1 \\ H_2: w^T x_i + b &= -1 & \text{for } y_i &= -1 \end{aligned} \quad (۲-۲)$$

در واقع نمونه‌های کلاس مثبت در ناحیه $w^T x_i + b \geq +1$ و نمونه‌های منفی در $w^T x_i + b \leq -1$ قرار می‌گیرند. این دو نامعادله را می‌توان ترکیب کرد و به صورت زیر نوشت:

$$y_i(w^T x_i + b) - 1 \geq 0 \quad \forall i = 1, 2, \dots, n \quad (۳-۲)$$

فاصله‌ی ابرصفحه‌های H_1 و H_2 از ابرصفحه جداکننده را به اصطلاح حاشیه^{۹۶} می‌نامند. بدین ترتیب با بیشینه کردن حاشیه می‌توان به ابرصفحه بهینه رسید. می‌دانیم کوتاه‌ترین فاصله هر نقطه دلخواه تا یک ابرصفحه از فرمول زیر به دست می‌آید:

$$d = \frac{|w^T x_i + b|}{\|w\|} \quad (۴-۲)$$

بنابراین بیشینه حاشیه به صورت زیر قابل محاسبه است:

$$\min_{x_i: y_i=+1} \frac{|w^T x_i + b|}{\|w\|} + \min_{x_i: y_i=-1} \frac{|w^T x_i + b|}{\|w\|} = \frac{1}{\|w\|} + \frac{1}{\|w\|} = \frac{2}{\|w\|} \quad (۵-۲)$$

به طوری که $\|w\|$ نُرم اقلیدسی بردار وزن‌هاست. برای بیشینه کردن حاشیه $(\frac{2}{\|w\|})$ کافی است تا

$\|w\|$ کمینه گردد:

$$\min \frac{1}{2} \|w\|^2 \quad (۶-۲)$$

^{۹۶} Margin

$$s. t. \quad y_i(w^T x_i + b) - 1 \geq 0 \quad \forall i = 1, 2, \dots, n$$

در مسائل طبقه‌بندی، به منظور تولید^{۹۷} مدل بهتر می‌توان تا حدی خطای طبقه‌بندی را تحمل کرد. از این رو مفهوم حاشیه نرم در ماشین بردار پشتیبان به وجود آمد. بدین منظور پارامترهایی برای کنترل خطا (C) و پارامتر لغزش (ξ_i) تعریف می‌شوند و معادله (۶-۲) به صورت زیر تغییر می‌یابد:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \quad (۷-۲)$$

$$s. t. \quad y_i(w^T x_i + b) - 1 + \xi_i \geq 0 \quad \forall i = 1, 2, \dots, n$$

برای حل مسائل بهینه‌سازی در معادلاتی به این شکل که همراه با یک یا چند قید هستند، یکی از روش‌ها، حل به روش ضرایب لاگرانژ است. بدین ترتیب معادله‌ی لاگرانژ به صورت زیر تشکیل می‌شود:

$$\begin{aligned} \min L_p = & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \\ & - \sum_i \alpha_i (y_i(w^T x_i + b) - 1 + \xi_i) \\ & \alpha_i \geq 0, i = 1, 2, \dots, n \end{aligned} \quad (۸-۲)$$

که در آن α_i ها ضرایب لاگرانژ هستند. با حل این معادله و مشتق‌گیری نسبت به پارامترهای تابع لاگرانژ، می‌توان مسئله را حل کرد.

فرمول‌های به دست آمده تا این بخش، با این فرض بود که داده‌های دو کلاس به صورت خطی جداپذیر باشند. در صورتی که داده‌ها جداپذیر خطی نباشند، می‌توان با توابعی به نام توابع هسته^{۹۸} ($K(x_i, x_j)$) داده‌ها در فضای m بُعدی را به فضایی با ابعاد l ($l > m$) انتقال داد، تا جایی که در فضای با ابعاد بالاتر جداپذیر خطی باشند. این نگاشت بردار غیرخطی $\varphi(x)$ صورت می‌گیرد:

^{۹۷} Generation

^{۹۸} Kernel function

$$\varphi(x) = (\varphi(x_1), \dots, \varphi(x_l))^T$$

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \quad (9-2)$$

بدین ترتیب معادله بهینه‌سازی به صورت زیر تغییر می‌یابد:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \quad (10-2)$$

$$s. t. \quad y_i(w^T \varphi(x_i) + b) - 1 + \xi_i \geq 0 \quad \forall i = 1, 2, \dots, n$$

معروف‌ترین توابع هسته عبارتند از: هسته‌ی خطی^{۹۹}، تابع پایه شعاعی^{۱۰۰}، هسته چندجمله‌ای^{۱۰۱} و هسته

سیگموئید^{۱۰۲}. ضابطه هر کدام از توابع هسته به صورت زیر است:

$$\text{Linear kernel: } K(x_i, x_j) = \langle x_i, x_j \rangle$$

$$\text{RBF kernel: } K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

$$\text{Polynomial kernel of degree } d: K(x_i, x_j) = (\langle x_i, x_j \rangle + c)^d$$

$$\text{sigmoid kernel: } K(x_i, x_j) = \tanh(\langle x_i, x_j \rangle + r) \quad (11-2)$$

همان‌طور که پیش‌تر اشاره شد، ماشین بردار پشتیبان برای مسائل دو کلاسه مناسب است؛ اما

روش‌هایی برای تعمیم این دسته‌بند نیز وجود دارد. تکنیک‌های «یکی در مقابل همه» و «یکی در مقابل

یکی» دو استراتژی متداول در حل مسائل چند کلاسه است.

۲-۶-۲. الگوریتم خفاش

ایده الگوریتم خفاش^{۱۰۳} برگرفته از شیوه‌ی جهت‌یابی خفاش با پژواک صدا بوده و برای یافتن راه‌حل

بهینه از روش این جانوران در یافتن شکار و جهت حرکت خود الهام گرفته است (Yang, 2010).

همانند سایر الگوریتم‌های فرامکاشفه‌ای، الگوریتم خفاش نیز رسیدن به پاسخ بهینه را تضمین نمی‌کند

و علاوه بر آن برای حل مسائل در فضای پیوسته مناسب است.

⁹⁹ Linear kernel

¹⁰⁰ Radial based functions (RBF)

¹⁰¹ Polynomial kernel

¹⁰² Sigmoid kernel

¹⁰³ Bat Algorithm (BA)

در این الگوریتم برای سادگی، فرضیات زیر در نظر گرفته می‌شود:

- تمامی خفاش‌ها با ارسال صدا فاصله را درک کرده و از این طریق می‌توانند طعمه و موانع را تشخیص دهند.
- در این الگوریتم فرض می‌شود که خفاش‌ها در مکان x_i و با سرعت v_i به طور تصادفی حرکت می‌کنند. فرکانس صدا از ابتدا برابر f_{min} است و با طول موج و بلندی صدا به ترتیب λ و A_0 است. خفاش‌ها این توانایی را دارند تا طول موج (و یا فرکانس) پالس صدا و نرخ پالس انتشار $r \in [0,1]$ را بر اساس این که چه میزان به هدف خود نزدیک هستند به طور خودکار تنظیم کنند.
- بلندی صدا در طول زمان تغییر می‌کند. در این الگوریتم فرض می‌شود که بلندی صدا از مقدار ثابت و مثبت A_0 تا مقدار $A_{min} > 0$ تغییر می‌کند.
- برای سادگی فرض می‌شود که از ردیابی صدا به منظور محاسبه تاخیر و نقشه‌برداری سه بعدی استفاده نخواهد شد.

فرکانس f در بازه‌ی $[f_{min}, f_{max}]$ محدود شده است که متناسب با طول موج در بازه $[\lambda_{min}, \lambda_{max}]$ است.

می‌دانیم حاصل ضرب λf برابر سرعت صوت بوده که مقداری ثابت است ($v = 340 \text{ m/s}$). بنابراین در پیاده‌سازی الگوریتم فقط مقدار فرکانس تغییر می‌یابد و طول موج بدون تغییر باقی می‌ماند.

همان‌گونه که بیان شد در این شبیه‌سازی، خفاش‌های مصنوعی هرکدام موقعیت و سرعتی در یک فضای d بعدی دارند و در گام t موقعیت x_i^t و سرعت v_i^t آن‌ها به ترتیب طبق فرمول‌های (۲-۱۲) و (۲-۱۳) به‌روزرسانی می‌شود.

$$x_i^t = x_i^{t-1} + v_i^t \quad (2-12)$$

$$v_i^t = v_i^{t-1} + (x_i^{t-1} - x_*)f_i \quad (2-13)$$

$$f_i = f_{min} + (f_{max} - f_{min})\beta \quad (2-14)$$

در فرمول (۱۶-۲) پارامتر $\beta \in [0,1]$ عددی تصادفی از توزیع یکنواخت است و بدین ترتیب مقدار فرکانس در هر مرحله به صورت تصادفی در محدوده مشخص تولید می‌شود. x_* بهترین جوابی است که تا به حال یافت شده و در هر تکرار از مقایسه مکان تمامی خفاش‌ها به دست می‌آید.

در بهره‌وری^{۱۰۴} و به منظور جستجوی محلی، پس از یافتن بهترین راه‌حل، مکان جدیدی به روش مسیر تصادفی برای خفاش یافت می‌شود (فرمول (۱۵-۲))

$$x_{new} = x_{old} + \varepsilon A^t \quad (15-2)$$

متغیر $\varepsilon \in [-1,1]$ عددی تصادفی است و $A^t = \langle A_i^t \rangle$ میانگین بلندی صدای تمام خفاش‌ها در زمان t است.

از سوی دیگر هنگامی که خفاش به شکار خود نزدیک می‌شود، بلندی صدای خود را کاهش داده و نرخ پالس را افزایش می‌دهد (فرمول (۱۶-۲)).

$$A_i^{t+1} = \alpha A_i^t, \quad r_i^{t+1} = r_i^0 (1 - \exp(-\gamma t)) \quad (16-2)$$

در این فرمول α و γ دو پارامتر ثابت هستند و α فاکتور خنک‌سازی است. از طرفی اگر $t \rightarrow \infty$ برای هر $0 < \alpha < 1$ و $\gamma > 0$ داریم:

$$A_i^t \rightarrow 0, \quad r_i^t \rightarrow r_i^0 \quad (17-2)$$

در پیاده‌سازی این الگوریتم برای سادگی حالت $\alpha = \gamma = 0.9$ در نظر گرفته شده است.
۳-۶-۲. ترکیب ماشین بردار پشتیبان و الگوریتم خفاش

در (Tharwat, Hassanien, & Elnaghi, 2017) مدلی برای ترکیب الگوریتم ماشین بردار پشتیبان و الگوریتم خفاش ارائه شده است. در واقع برای یافتن پارامترهای این دسته‌بند از الگوریتم خفاش بهره گرفته شده است. همان‌طور که در بخش‌های قبل اشاره شد، در ماشین بردار پشتیبان لازم است تا دو پارامتر پنالیتی (C) و پارامتر تابع هسته مقداردهی شوند. این بهینه‌سازی در دو آزمایش بر روی ۹

مجموعه داده دانشگاه UCI^{۱۰۵} انجام شده است. در اولین آزمایش، پارامتر پنهانی و پارامتر هسته چندجمله‌ای^{۱۰۶} d بهینه شده و در دومین آزمایش پارامتر خطا c و پارامتر هسته تابع پایه شعاعی^{۱۰۷} σ به دست آمده است. روند تعیین پارامترها به این ترتیب است که پس از نرمال کردن داده‌ها، به روش اعتبارسنجی متقابل k -fold و $k=10$ داده‌ها به دو دسته داده‌های آموزشی و داده‌های آزمایش تقسیم می‌گردند. سپس در هر گام الگوریتم خفاش اجرا شده و بهترین راه‌حل به دست آمده از این الگوریتم در نظر گرفته می‌شود و واضح است که این مقادیر بهینه سراسری نیستند. در ادامه پارامترهای به دست آمده برای آموزش داده‌ها با ماشین بردار پشتیبان اعمال می‌شود. پس از محاسبه تابع برازش^{۱۰۸} (فرمول (۱۸-۲)) بهترین وضعیت (X^*) خفاش‌ها به دست می‌آید. در این جا تابع برازش همان نرخ خطاست و N_e تعداد نمونه‌هایی که برچسب اشتباه خورده‌اند و N تعداد کل داده‌های آزمایش است. این الگوریتم تا برآورده شدن معیار خاتمه ادامه پیدا می‌کند.

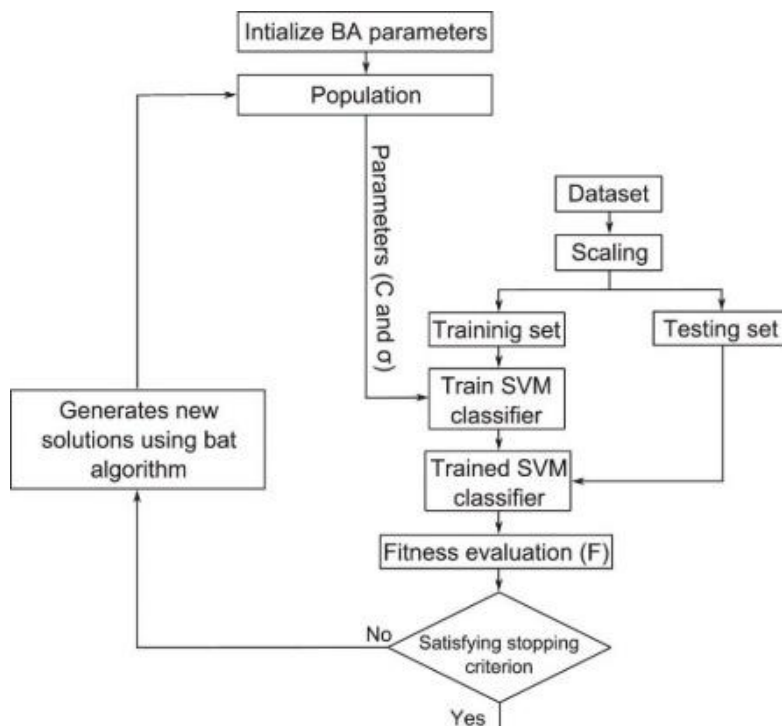
$$\text{Minimize: } F = \frac{N_e}{N} \quad (18-2)$$

¹⁰⁵ <https://archive.ics.uci.edu/ml/datasets.php>

¹⁰⁶ Polynomial kernel

¹⁰⁷ Radial Basis Function (RBF)

¹⁰⁸ Fitness function



شکل ۱-۲ فلوچارت بهینه سازی ماشین بردار پشتیبان با الگوریتم خفاش (Tharwat, Hassanien, & Elnaghi, 2017)

۷-۲. جمع‌بندی

در این فصل پیشینه استخراج اطلاعات از اسناد علمی و دانشگاهی از منظر نوع اطلاعات مورد نظر ارائه گردید. با توجه به این که دو نوع اطلاعات می‌تواند مورد هدف محققان این حوزه قرارگیرد، بخشی از این فصل به پژوهش‌های صورت گرفته در استخراج فراداده‌ها از سرآیند، مراجع یا هر دو اختصاص یافت و در بخش دیگر به مقالات ارائه شده در تشخیص سایر اطلاعات نهفته در اسناد علمی اشاره شد. در هر دو حوزه، روش‌های مبتنی بر قاعده و روش‌های مبتنی بر یادگیری ماشین مورد بررسی قرارگرفت و همان‌طور که انتظار می‌رود، در سال‌های اخیر روش‌های مبتنی بر قواعد از پیش تعریف شده کمتر مورد توجه قرارگرفته و استفاده از روش‌های یادگیری ماشین بر روش‌های سنتی پیشی گرفته است. اکثریت پژوهش‌های صورت گرفته بر مقالات همایش و مجلات، به‌خصوص زبان

انگلیسی، تمرکز داشته‌اند و تحقیقات اندکی بر روی پایان‌نامه‌ها و رساله‌ها و همچنین زبان فارسی صورت گرفته است.

فصل سوم

روش پیشنهادی

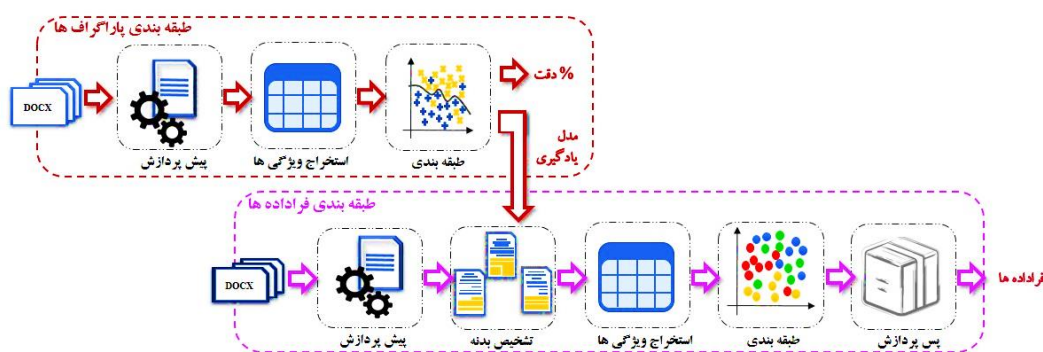
۳-۱. مقدمه

در این فصل روش پیشنهادی برای استخراج فراداده از پایان‌نامه‌ها و رساله‌های فارسی به تفصیل شرح داده می‌شود. روال کار به دو بخش تقسیم می‌شود: ابتدا صفحات آغازین پایان‌نامه یا رساله از بدنه و متن اصلی جدا می‌شود و در ادامه صفحات آغازین برای استخراج فراداده‌های موردنظر پردازش می‌گردد. هر دو مرحله با رویکرد یادگیری ماشین انجام شده و طبقه‌بند ماشین بردار پشتیبان برای طبقه‌بندی پاراگراف‌ها انتخاب شده است.

۳-۲. روش پیشنهادی در استخراج فراداده‌ها

این بخش به توضیح روش پیشنهاد شده برای استخراج فراداده از پایان‌نامه‌ها و رساله‌های فارسی می‌پردازد. در این راستا، ابتدا به پیش‌پردازش و نحوه آماده‌سازی داده‌ها برای فرآیندهای آتی اشاره می‌گردد. سپس به اهداف و جزئیات دو مرحله «تشخیص مرز بدنه» و «طبقه‌بندی فراداده‌ها» پرداخته می‌شود. گام‌های پژوهش در شکل ۱.۳ قابل مشاهده است. لازم به ذکر است که از آن‌جا که در قالب *DOCX* جداکننده صفحات ظاهر نشده و محل قرارگیری اجزای سند در هنگام اجرای نرم‌افزارهای

مربوطه محاسبه می‌شود، در این پژوهش پارگراف‌های اسناد طرح پژوهشی مورد پردازش قرار گرفته‌اند.



شکل ۳-۱ گام‌های پژوهش. ابتدا مدلی برای شناسایی بدنه و فراداده ساخته می‌شود. سپس از این مدل برای تشخیص مرز بدنه استفاده شده و در ادامه فراداده‌ها طبقه‌بندی می‌شوند.

در این پژوهش از روش ارائه شده در مقاله (Tkaczyk, Szostek, Fedoryszak, Dendek, & Bolikowski, 2015) اقتباس شده است. در این سیستم فراداده‌های سرآیند و مراجع از مقالات به دست می‌آید. بدین منظور، ابتدا نواحی متنی اسناد PDF شناسایی شده و به چهار کلاس فراداده، بدنه، مراجع و سایر طبقه‌بندی می‌شود. در ادامه، نواحی با برجسب «فراداده» وارد مرحله استخراج فراداده شده و به انواع فراداده‌های از قبل تعریف شده طبقه‌بندی می‌گردند. هر دو طبقه‌بندی با استفاده از دسته‌بند ماشین بردار پشتیبان صورت گرفته‌اند. همچنین کلاس «مراجع» نیز برای تجزیه رشته‌های متنی و استخراج فراداده‌های آن مورد پردازش قرار می‌گیرد. در این طرح پژوهشی، از نحوه عملکرد این سیستم برای دستیابی به فراداده‌های پایان‌نامه‌ها و رساله‌ها الهام گرفته شده که از جمله آن می‌توان به انتخاب ویژگی‌های محاسبه شده در فرآیند یادگیری ماشین اشاره نمود. در ادامه روش پژوهش و گام‌های آن شرح داده می‌شود.

۱-۲-۳. پیش‌پردازش

آماده‌سازی داده‌ها برای انجام فرآیند یادگیری امری ضروری است. از این رو لازم است تا بخش‌های اضافی پایان‌نامه‌ها و رساله‌ها دور ریخته شود و پیش‌پردازش‌های موردنیاز برای یکسان‌سازی متن صورت گیرد.

از آنجایی که هدف این تحقیق استخراج فراداده‌ها از پایان‌نامه‌ها و رساله‌ها است و فراداده‌های هر پایان‌نامه/رساله در صفحات آغازین ظاهر می‌شوند، بنابراین پردازش کل سند فاقد ضرورت است. برای نمونه فصل‌های اصلی سند و محتوای آن‌ها، فهرست مراجع، پیوست‌ها و فراداده‌های انگلیسی که معمولاً در انتهای سند نوشته می‌شوند مورد توجه این تحقیق نیستند. به همین علت، قبل از انجام فرآیند اصلی، چهل درصد کل پاراگراف‌های هر پایان‌نامه/رساله نگه داشته شده و شصت درصد باقی مانده پردازش نمی‌شود. به بیان دیگر فرض شده که شروع بدنه اصلی در بین چهل درصد ابتدایی پاراگراف‌های پایان‌نامه/رساله قرار داشته که در فاز تشخیص مرز بدنه، مرز شروع متن اصلی شناسایی

خواهد شد. (Error! Reference source not found.)

شکل ۳-۲ نسبت تعداد پاراگراف‌های پیشگفتار به تعداد کل پاراگراف‌های طرح پژوهشی در داده‌های موجود

به دلیل چالش‌های موجود در پردازش متون زبان فارسی، از جمله وجود برخی کاراکترهای یکسان در زبان فارسی و عربی با یونی‌کدهای متفاوت (مانند حروف /ی/ و /ک/)، وجود فاصله و نیم‌فاصله و همچنین استفاده از کاراکترها و علائم عربی، لازم است تا متن پایان‌نامه‌ها/رساله‌ها پیش از استخراج

ویژگی، با ابزارهای یکسان‌سازی مورد پردازش قرار گیرد. (درباره چالش‌ها در فصل اول توضیحات کامل داده شده است.) با نرمال‌سازی متن، تمامی کاراکترهای عربی و فارسی که از نظر ظاهری شبیه هستند، اما یونی‌کد متفاوتی دارند، با یونی‌کد کاراکتر فارسی جایگزین می‌شوند. علاوه بر آن ارقام انگلیسی و عربی تبدیل به ارقام فارسی معادل می‌شوند و کاراکترها و علائم خاص با نسخه استاندارد آن‌ها جایگزین می‌گردند. برای پیش‌پردازش، در این پژوهش از ابزار فارسی‌یار^۱ به صورت برخط استفاده شده است.

۳-۲-۱. تشخیص مرز بدنه

طبق راهنمای نگارش در وب‌سایت دفتر برنامه‌ریزی و سیاست‌گذاری پژوهشی^۲، پایان‌نامه‌ها/رساله‌های فارسی به طور کلی از الگوی زیر پیروی می‌کنند:

۱. صفحه‌ی بسم الله الرحمن الرحيم

۲. صفحه‌ی عنوان شامل:

- نام دانشگاه، پژوهشگاه یا مؤسسه آموزش عالی
- نام دانشکده، پژوهشکده، آموزشکده، مجتمع
- نوع: پایان‌نامه / رساله
- رشته و گرایش
- عنوان پایان‌نامه
- نام و نام خانوادگی دانش‌آموخته

^۱ <https://text-mining.ir/>

^۲ فایل-الگوی-تمپلیت-راهنمای-نگارش-پایان-نامه-و-رساله/2012/download/rppc.msrt.ir/fa/

- نام و نام خانوادگی استاد/استادان راهنما
 - نام و نام خانوادگی استاد/استادان مشاور
 - ماه و سال خورشیدی دفاع از پایان‌نامه
۳. برگ اصالت و مالکیت اثر
۴. برگ تأیید هیئت داوران/ صورت جلسه دفاع (به زبان فارسی)
۵. صفحه تقدیم
۶. صفحه سپاسگزاری
۷. چکیده
۸. کلیدواژه‌ها
۹. فهرست نوشتار، کوتاه‌نوشت‌ها، جدول‌ها، تصویرها، نمودارها و پیوست‌ها
۱۰. محتوای پایان‌نامه/رساله در چند فصل
۱۱. فهرست منابع
۱۲. پیوست‌ها
۱۳. واژه‌نامه توصیفی
۱۴. واژه‌نامه فارسی به انگلیسی
۱۵. واژه‌نامه انگلیسی به فارسی
۱۶. فهرست مقالات برگرفته از پایان‌نامه/رساله
۱۷. برگ تأیید هیئت داوران/ صورت جلسه دفاع (به زبان انگلیسی)
۱۸. چکیده و کلیدواژه‌ها به زبان انگلیسی
۱۹. صفحه عنوان به زبان انگلیسی

استخراج فراداده در پارساها با رویکرد الگوریتم‌های دسته‌بندی □ ۶۳

اگرچه قالب نگارش پایان‌نامه‌ها و رساله‌ها برای دانشگاه‌های مختلف متفاوت است، اما تمامی قالب‌ها از الگوی فوق تبعیت می‌کنند و تفاوت آن‌ها در ترتیب برخی بخش‌ها و ویژگی‌هایی چون نوع فونت و اندازه‌های صفحات است. همچنین نحوه نگارش محتوای اصلی و نام فصول برای رشته‌های گوناگون متفاوت است. لازم به ذکر است که برخی از بخش‌های نام برده مانند صفحات بسم‌الله الرحمن الرحیم، تقدیم، سپاسگزاری و پیوست‌ها عموماً اختیاری هستند و می‌توانند حذف شوند. در مرحله تشخیص مرز بدنه، پاراگراف شروع بدنه اصلی مشخص شده و پاراگراف‌های پس از آن کنار گذاشته می‌شوند. بنابراین طبق فهرست بالا، تنها بخش‌های مربوط به شماره‌های ۱ تا ۹ وارد مرحله طبقه‌بندی فراداده‌ها شده و سایر بخش‌ها در نظر گرفته نمی‌شوند. در **Error! Reference source not found.** نمونه‌هایی از شروع بدنه اصلی در پایان‌نامه‌ها/رساله‌ها مشاهده می‌شود. در ادامه جزئیات این مرحله تشریح می‌شود.

شکل ۳-۶: مقایسه زمان بارگذاری دایره‌وار سازه‌های مختلف	۲۰
شکل ۳-۶: مقایسه اندازه مخرن در ذخیره سازه‌های مختلف	۷۱
شکل ۳-۶: مقایسه زمان ریزش در روش‌های مختلف	۷۲
فصل ۱: وب معدنیابی	
۱-۱ وب معدنیابی	
در سال ۱۹۹۷ میلادی تیم مبراژ ^۱ مفهوم وب معدنی ^۲ را معرفی کرد. وب معدنی توسعه ای از وب جاری است که می‌توان آن را تقاضای در راستای تریزین موجود به اطلاعات موجود در وب جاری به حساب آورد [۱]. امروزه وب معدنی چالاکه خود را در جوامع تحقیقاتی باز کرده و نرم افزارهای معدنی را در آن به کار می‌برد.	

(الف)

(ب)

(پ)

شکل ۳-۳ چند نمونه از مرز فراداده و بدنه در پایان‌نامه. در شکل‌های (الف)، (ب) و (پ) به ترتیب عبارت‌های «فصل ۱:»، «فصل اول» و «کلمات تحقیق» شروع بدنه‌ی پایان‌نامه هستند.

استخراج ویژگی

همان‌طور که پیشتر بدان اشاره شد، در این طرح پژوهشی روش پیشنهادی در (Tkaczyk, Szostek, Fedoryszak, Dendek, & Bolikowski, 2015) مورد توجه قرار گرفته است. یکی از مراحل سیستم پیشنهاد شده در این مقاله، استخراج ساختار پایه است که در آن ساختار سلسله‌مراتبی سند تولید شده و سپس نواحی مقاله به چهار کلاس فراداده، بدنه، مراجع و سایر طبقه‌بندی می‌شود. در این طرح پژوهشی نیز، در فاز تشخیص مرز بدنه، پاراگراف‌ها به دو کلاس فراداده و بدنه دسته‌بندی می‌شوند و در مرحله استخراج ویژگی‌ها، برخی ویژگی‌های انتخاب شده در مرحله استخراج ساختار پایه در سیستم مذکور از پایان‌نامه‌ها و رساله‌ها به دست می‌آیند. به دلیل این که در این سیستم مقالات در قالب PDF و به زبان انگلیسی مورد پردازش قرار می‌گیرند، واضح است که تمامی این ویژگی‌ها برای پایان‌نامه‌ها و رساله‌های فارسی و در قالب DOCX مناسب نیستند. بنابراین، علاوه بر حذف برخی ویژگی‌های پیشنهادی این سیستم، تعدادی از ویژگی‌ها، متناسب با پایان‌نامه‌ها و رساله‌های فارسی تعریف شده‌اند که در ادامه به آن اشاره خواهد شد. به عنوان نمونه محاسبه تعداد کلماتی که با حروف بزرگ شروع می‌شوند در زبان فارسی معنا ندارد. همچنین در نظر گرفتن ویژگی‌هایی که وابسته به مختصات و مکان اجزای اسناد هستند در فایل‌های DOCX امکان‌پذیر نیست. به‌طور کلی پنج نوع ویژگی در سیستم مورد نظر تعریف شده است:

- هندسی: این ویژگی‌ها مبتنی بر خصیصه‌های هندسی و مکانی اجزای سند هستند. طول و عرض ناحیه^۱، نسبت طول به عرض، مختصات طولی و عرضی، فاصله تا نزدیک‌ترین ناحیه،

¹ Zone

فضای خالی بالا و پایین ناحیه، میانگین طول خطوط و مکان ناحیه در صفحه (بالا، پایین، چپ و راست) نمونه‌هایی از ویژگی‌های هندسی هستند.

- لغوی: کلمات کلیدی که ممکن است در بخش‌های مختلف مقاله ظاهر شوند؛ مانند وابستگی‌ها، قدردانی، چکیده، تاریخ، کلمات کلیدی، مراجع و نوع مقاله. در این ویژگی بررسی می‌شود که کلمات کلیدی خاص در نواحی متنی ظاهر شده‌اند یا خیر.
- توالی: این ویژگی بر اساس ارتباطات بین ناحیه‌هایی است که پشت سرهم قرار گرفته‌اند. برجسب ناحیه قبلی و برجسب دو ناحیه قبلی (در همان صفحه) یا ویژگی‌هایی که به نوعی مرتبط با دو ناحیه متوالی هستند در این دسته قرار می‌گیرند.
- قالب: شامل ویژگی‌هایی چون اندازه فونت، فضای خالی داخل ناحیه (پیسکل‌ها) و میانگین خطوط که مرتبط با سبک و قالب نوشته هستند.
- هیوریستیک: این دسته از ویژگی‌ها مربوط به ذات و طبیعت نوشته‌ها هستند. به عنوان مثال تعداد و درصد خطوط، کلمات، کلمات دارای حروف بزرگ، کاراکترها، حروف، حروف بزرگ، ارقام، فاصله، علائم نگارشی (نقطه، ویرگول، قلاب و ...) و یا بررسی این که متن با عدد شروع شده است، نمونه‌هایی از ویژگی‌های هیوریستیک هستند.

در این پژوهش تلاش شده تا ویژگی‌های مستخرج - با حذف و تعریف برخی ویژگی‌های سیستم مذکور- با ساختار پایان‌نامه‌ها و رساله‌های فارسی مطابقت یابد. به همین منظور در گام شناسایی مرز بدنه و فراداده، ویژگی‌های سیستم یادشده در فاز استخراج ساختار پایه بررسی شده که بر اساس آن، ۱۵ ویژگی برای هر پاراگراف تعریف شده که در جدول ۳-۱ قابل مشاهده هستند. تمامی ویژگی‌ها از نوع کمی هستند و در صورتی که شرطی بررسی شود، می‌تواند یکی از مقادیر صفر و یا یک را اختیار کند.

جدول ۱-۳. ویژگی‌های استخراج شده در مرحله‌ی تشخیص مرز بدنه

نوع ویژگی	توضیحات
هیورستیک	تعداد تکرار کارکتر قلاب '[' یا ']'
	تعداد کاراکتر قلاب '[' یا ']' نسبت به کل کاراکترها
	تعداد تکرار اعداد فارسی و انگلیسی
	تعداد ارقام فارسی و انگلیسی نسبت به کل کاراکترها
	تعداد تکرار کاراکتر نقطه (.)
	تعداد کاراکتر نقطه نسبت به کل کاراکترها
	تعداد خطوط به صورت تقریبی
	تعداد تکرار کاراکتر پرانتز '(' یا ')'
	تعداد تکرار کاراکتر پرانتز '(' یا ')' نسبت به کل کاراکترها
	تعداد علائم نگارشی نسبت به کل کاراکترها
	تعداد کلمات موجود در پاراگراف
	آیا در ابتدای پاراگراف عدد مشاهده می‌شود؟
	شماره پاراگراف
توالی	
لغوی	تعداد عبارات «فصل اول»، «مقدمه»، «پیشگفتار»، «دیاچه» یا «کلیات تحقیق»
هندسی	نسبت طول به عرض پاراگراف

همان‌طور که ملاحظه می‌شود، بیشتر ویژگی‌های محاسبه شده از نوع هیورستیکی هستند. ویژگی لغوی تعداد عبارات رایج در ابتدای بدنه پایان‌نامه/رساله را محاسبه می‌کند. عباراتی چون «فصل اول»، «فصل ۱»، «مقدمه» و «کلیات تحقیق» نمونه‌هایی از کلمات کلیدی هستند که در شروع بدنه پایان‌نامه/رساله مشاهده می‌شوند. طبق این فهرست، یکی از ویژگی‌های هیورستیک تعداد خطوط پاراگراف است. پیش‌تر ذکر شد که در اسناد DOCX اطلاعاتی چون جداکننده صفحات و خطوط ظاهر نمی‌شوند و مکان هر یک از اجزای سند هنگام اجرای نرم‌افزارهای مربوطه محاسبه می‌گردد. بنابراین تعداد خطوط هر پاراگراف به طور دقیق قابل محاسبه نیست. در مقابل می‌توان با اطلاعات دیگری چون اندازه فونت و فاصله پاراگراف از لبه‌های کاغذ (حاشیه)، تعداد خطوط را به صورت تقریبی محاسبه کرد.

تشخیص مرز

پس از استخراج ویژگی‌های بیان شده در بخش قبل، پاراگراف‌ها به یکی از کلاس‌های فراداده و بدنه طبقه‌بندی می‌شوند. این طبقه‌بندی با استفاده از الگوریتم ماشین بردار پشتیبان انجام می‌گیرد. پیش از شروع فرآیند یادگیری ضروری است تا بردارهای ویژگی نرمال شوند. از مزایای نرمال کردن داده‌ها این است که از غلبه اعداد بزرگ‌تر بر مقادیر کوچک جلوگیری می‌شود. در واقع مقدار تمامی ویژگی‌ها به بازه مشخص $[0, 1]$ نگاشت می‌شود و تأثیر یکسانی نسبت به هم دارند. بنابراین دقت الگوریتم طبقه‌بندی بهبود می‌یابد. از طرف دیگر تمامی محاسبات در بازه کوچک‌تر و مشخصی انجام می‌شود و در نتیجه محاسبات ساده‌تر خواهند بود (Zhao, Fu, Ji, Tang, & Zhou, 2011). در این پژوهش بردارهای ویژگی بر اساس فرمول نرمال‌سازی می‌شوند.

$$v' = \frac{v - \min}{\max - \min} \quad (1-3)$$

برای طبقه‌بندی پاراگراف‌ها، دسته‌بند ماشین بردار پشتیبان انتخاب شده و عملکرد آن همراه با چهار تابع هسته‌ای معروف بررسی شده است (فرمول‌های (۲-۳)). نتایج ارزیابی در انتخاب تابع هسته مناسب و همچنین بهینه‌سازی پارامترها در فصل بعدی تشریح خواهد شد.

$$\text{Linear kernel: } K(x_i, x_j) = \langle x_i, x_j \rangle$$

$$\text{RBF kernel: } K(x_i, x_j) = \exp\left(\frac{-||x_i - x_j||^2}{2\sigma^2}\right)$$

$$\text{Polynomial kernel of degree } d: K(x_i, x_j) = (\langle x_i, x_j \rangle + c)^d$$

$$\text{sigmoid kernel: } K(x_i, x_j) = \tanh(\langle x_i, x_j \rangle + r) \quad (2-3)$$

برای ارزیابی طبقه‌بند، جداسازی داده‌های آموزشی و آزمایش به روش اعتبارسنجی متقابل k -fold صورت گرفته (Arlot & Celisse, 2010) و مقدار k برابر با ۱۰ در نظر گرفته شده است. در روش اعتبارسنجی متقابل داده‌ها به k بخش تقسیم می‌شود و $k-1$ قسمت به عنوان داده‌های آموزشی و یک قسمت به عنوان داده‌های آزمایش برای ارزیابی طبقه‌بند مورد استفاده قرار می‌گیرد. این کار k بار تکرار می‌شود تا هر قسمت یک بار به عنوان داده آزمایش مورد بررسی قرار گیرد. این تقسیم‌بندی می‌تواند به صورت تصادفی نیز انجام شود.

پس از طبقه‌بندی، برجسب پاراگراف‌ها از ابتدای پایان‌نامه/رساله بررسی شده تا مشخص گردد از کدام پاراگراف به بعد اکثریت پاراگراف‌ها برجسب بدنه دارند. در این پژوهش فرض بر این است که در صورتی که پاراگرافی یافت شود که حداقل چهار پاراگراف بعد از آن در دسته بدنه قرار گرفته باشند، این پاراگراف ابتدای بدنه در نظر گرفته می‌شوند. سپس از اولین پاراگراف تا قبل پاراگراف موردنظر وارد مرحله بعد یعنی طبقه‌بندی فراداده‌ها می‌شود.

۲-۲-۳. طبقه‌بندی فراداده‌ها

پس از شناسایی مرز بدنه و کنار گذاشتن متن اصلی پایان‌نامه، می‌توان فراداده‌های موردنظر را از بخش‌های آغازین (پیش‌گفتار) به دست آورد. در ادامه روش پیشنهادی در استخراج فراداده از پایان‌نامه‌ها و رساله‌ها شرح داده می‌شود.

استخراج ویژگی

در فاز طبقه‌بندی فراداده‌ها، تلاش شده تا ویژگی‌های مرحله استخراج فراداده -در سیستم ارائه شده در این مقاله- با نوع داده‌های این پژوهش تطبیق یابند. به همین منظور ۵۷ ویژگی در این گام محاسبه می‌شود که شامل ۱۷ ویژگی هیوریستیک، ۷ ویژگی توالی، ۱۴ ویژگی لغوی، ۷ ویژگی هندسی و ۳

ویژگی مبتنی بر قالب هستند. علاوه بر آن، ۹ ویژگی متعلق به دسته دیگری به نام ویژگی‌های محلی^۱ هستند که در این پژوهش پیشنهاد شده‌اند. هدف از تعریف این دسته از ویژگی‌ها، در نظر گرفتن یک همسایگی از پاراگراف‌هاست. همان‌طور که قبلاً اشاره شد، در فایل DOCX جداکننده صفحات مشخص نیست. از سوی دیگر، پاراگراف‌های هر صفحه به نوعی در یک همسایگی قرار می‌گیرند و بررسی پاراگراف‌های همسایه به شناسایی بهتر پاراگراف‌های فراداده کمک می‌کند. بدین منظور ابتدا پاراگراف‌ها را در مجموعه‌ای از بلوک‌ها دسته‌بندی می‌کنیم.

جداسازی بلوک‌ها بر اساس پاراگراف‌های پوچ انجام می‌شود. در واقع می‌دانیم که هر پاراگراف از چند جمله تشکیل شده و هر جمله متشکل از تعدادی کلمه است. در نهایت هر کلمه نیز از تعدادی حرف ساخته شده است. در این میان ممکن است تعدادی پاراگراف پوچ نیز وجود داشته باشد که به دلیل کارکترهای خاصی چون فاصله^۲، خط جدید^۳ و یا تب^۴ به وجود آمده‌اند. بدین ترتیب جداسازی بلاک‌ها طبق روش زیر انجام می‌شود:

"فرض کنید سند D موجود است و مجموعه پاراگراف‌های آن به صورت $P = \{p_1, p_2, p_3, \dots, p_n\}$ باشد و برخی از پاراگراف‌های آن پوچ باشند. هر بلوک b_i ($1 \leq i \leq m$) در مجموعه بلوک‌های $B = \{b_1, b_2, b_3, \dots, b_m\}$ از تعدادی پاراگراف غیرپوچ که بین دو پاراگراف پوچ قرار گرفته‌اند تشکیل شده است."

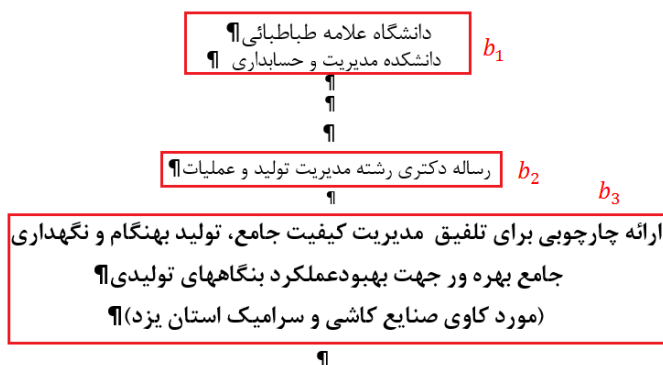
¹ Local

² Space

³ New line

⁴ Tab

استخراج فراداده در پارساها با رویکرد الگوریتم‌های دسته‌بندی □ ۷۰



شکل ۳-۴ نمونه‌ای از جداسازی بلوک‌ها در بخشی از پایان‌نامه. علامت ¶ نشانه‌ی انتهای پاراگراف است.

Error! Reference source not found. مثالی از جداسازی بلوک‌ها در بخشی از پاراگراف‌های یک رساله است. پس از به‌دست آمدن بلوک‌ها، می‌توان ویژگی‌های محلی را محاسبه کرد. برای نمونه بررسی این که پاراگراف موردنظر بزرگ‌ترین اندازه فونت را نسبت به سایر پاراگراف‌های بلوک دارد یا خیر، در این دسته از ویژگی‌ها محاسبه می‌شود. مزیت دیگر ویژگی‌های محلی، در تشخیص پاراگراف‌های فراداده‌ای است که به چند پاراگراف شکسته شده‌اند. برای مثال ممکن است عنوان پایان‌نامه‌ای در چند پاراگراف نوشته شده باشد. یا در نمونه‌ای دیگر نام اساتید راهنما در چند پاراگراف ظاهر شده باشد (**Error! Reference source not found.**).

عنوان بخش نظری و عملی:	استاد مشاور: ¶
طراحی بانک اطلاعاتی پایان‌نامه‌های رشته ارتباط تصویری ¶ و تصویرسازی در مقطع کارشناسی ارشد ایران (۱۳۹۴-۱۳۹۰) ¶	دکتر ابوالفضل تاج زاده نمین ¶
(الف)	(ب)

شکل ۳-۵ دو نمونه از شکسته شدن یک فراداده به چند پاراگراف. در شکل (الف) و (ب) به ترتیب عنوان طرح پژوهشی و استاد مشاور در دو پاراگراف نوشته شده‌اند. علامت ¶ نشانه‌ی انتهای پاراگراف است.

جدول ۳-۲ و جدول ۳-۳ به طور خلاصه ویژگی‌های استخراج شده در این بخش را نمایش می‌دهد. این ویژگی‌ها از نوع عددی و کمی هستند و در ویژگی‌هایی که شرط خاصی بررسی می‌شود، یکی از اعداد صفر یا یک را به عنوان مقدار می‌پذیرند. لازم به ذکر است که طبقه‌بندی پاراگراف‌ها از اولین

پاراگراف پایان‌نامه/رساله شروع می‌شود، بنابراین در ویژگی‌های محلی، پاراگراف‌های برجسته خورده - که به طور طبیعی پاراگراف‌های قبل از پاراگراف موردنظر هستند - در مواقع لزوم در نظر گرفته می‌شوند.

جدول ۲-۳. فهرست ویژگی‌های در نظر گرفته شده در مرحله طبقه‌بندی فراداده‌ها

نوع ویژگی	توضیحات
هیورستیک	تعداد تکرار علامت دو نقطه «:»
	نسبت تعداد تکرار علامت دو نقطه به کل کاراکترها
	تعداد تکرار اعداد فارسی و انگلیسی
	تعداد ارقام فارسی و انگلیسی نسبت به کل کاراکترها
	تعداد تکرار کاراکتر نقطه (.)
	تعداد کاراکتر نقطه نسبت به کل کاراکترها
	تعداد خطوط به صورت تقریبی
	تعداد تکرار کاراکتر ویرگول «،»
	تعداد تکرار کاراکتر ویرگول نسبت به کل کاراکترها
	تعداد علائم نگارشی نسبت به کل کاراکترها
	تعداد کلمات موجود در پاراگراف
	آیا پاراگراف تنها از اعداد تشکیل شده‌است؟
	آیا پاراگراف فقط شامل یک کلمه است؟
	نسبت تعداد کاراکتر فاصله به تعداد کل کاراکترها
	آیا پاراگراف با عدد شروع می‌شود؟
	میانگین طول کلمات در پاراگراف
	میانگین طول کلمات در پاراگراف
توالی	شماره پاراگراف
	برجسته پاراگراف قبلی
	برجسته دو پاراگراف قبل
	آیا نام فونت پاراگراف فعلی با فونت پاراگراف قبلی متفاوت است؟
	آیا اندازه فونت پاراگراف فعلی با اندازه فونت پاراگراف قبلی متفاوت است؟
	آیا پاراگراف فعلی و پاراگراف قبل از آن هر دو پررنگ هستند؟

استخراج فراداده در پارساها با رویکرد الگوریتم‌های دسته‌بندی □ ۷۲

آیا پاراگراف فعلی و پاراگراف قبل از آن هر دو /تالیك نوشته شده‌اند؟	
نسبت طول به عرض پاراگراف	هندسی
آیا پاراگراف در مرکز تراز شده است؟	
آیا قبل از پاراگراف فضای خالی (پاراگراف پوچ) وجود دارد؟	
آیا قبل از پاراگراف، تصویری وجود دارد؟	
فاصله‌ی پاراگراف از لبه راست کاغذ (حاشیه راست)	
فاصله‌ی پاراگراف از لبه چپ کاغذ (حاشیه چپ)	
میانگین پهنای کلمات در پاراگراف	

جدول ۳-۳. ادامه فهرست ویژگی‌های مستخرج در مرحله طبقه‌بندی فراداده‌ها

نوع ویژگی	توضیحات
لغوی (در این ویژگی تعداد کلمات کلیدی در بخش‌های مختلف پیشگفتار پایان‌نامه/رساله محاسبه می‌شود.)	«چکیده»
	«قدردانی»، «سیاس»، «تقدیم»، «تشکر»، «تقدیر»، «پدر»، «مادر» و کلمات مشابه
	«دانشجو»، «نگارنده»، «نگارش»، «تدوین»، «تهیه»، «مؤلف»، «پژوهشگر»
	«خرداد»، «بهمن»، «مهر»، «پاییز»، «زمستان» و کلمات مربوط به ماه‌ها و فصول سال
	«کارشناسی»، «کارشناسی ارشد»، «دکتر»، «دکتری»
	«دانشکده»، «پژوهشکده»، «آموزشکده»، «مجمع آموزش عالی»
	«رشته» و «گرایش»
	«کلمات کلیدی»، «واژگان کلیدی»، «واژه‌های کلیدی»
	«استاد راهنما»، «استاد مشاور»، «دکتر»، «پروفسور»، «مهندس»، «حجت الاسلام»
	«موضوع»، «عنوان»
	«فهرست»، «صفحه»
	«پایان‌نامه»، «رساله» یا «پروپوزال»
	«دانشگاه»، «پژوهشگاه» یا «مؤسسه آموزش عالی»
	سایر کلمات: «صورت جلسه‌ی دفاع»، «تاییدیه‌ی دفاع»، «تعهدنامه»، «سوگند نامه» و...
	اندازه‌ی فونت
قالب	آیا پاراگراف پررنگ است؟
	آیا پاراگراف /تالیك است؟

شماره بلوکی که پاراگراف موردنظر در آن قرار دارد.	محلی
آیا این پاراگراف بیشترین اندازه‌ی فونت را در بلوک دارد؟	
نسبت تعداد پاراگراف‌های فراداده به کل پاراگراف‌ها	
نسبت تعداد پاراگراف‌های با برجسب «سایر» به کل پاراگراف‌ها	
شماره پاراگراف در بلوک	
تعداد پاراگراف‌های با برجسب «نویسنده»	
تعداد پاراگراف‌های با برجسب «دانشجو»	
تعداد پاراگراف‌های با برجسب «عنوان»	
نسبت تعداد کلمات پاراگراف به تعداد کل کلمات بلوک	

طبقه‌بندی

در این گام، پاراگراف‌های صفحات آغازین – خروجی مرحله تشخیص مرز بدنه – مورد پردازش قرار گرفته و به فراداده‌های از پیش تعریف شده طبقه‌بندی می‌شوند. به‌طور کلی در بخش پیشگفتار هر پایان‌نامه/رساله، اطلاعاتی چون نام دانشگاه، دانشکده، رشته، مقطع تحصیلی، عنوان پایان‌نامه، نام دانش‌آموخته، استاد (اساتید) راهنما، استاد (اساتید) مشاور، چکیده و کلمات کلیدی به عنوان فراداده شناخته می‌شوند. کلاس استاد راهنما و مشاور به دلیل شباهت ویژگی‌هایشان سبب افت کارایی دسته‌بند می‌گردند. به همین علت این دو کلاس با هم ادغام شده و در پس‌پردازش نام اساتید راهنما و مشاور تفکیک خواهند شد.

از طرفی در الگوی نگارش برخی پایان‌نامه‌ها و رساله‌ها، رشته و مقطع تحصیلی در یک پاراگراف نوشته می‌شوند. این نکته سبب چالش در تشخیص دو فراداده مذکور خواهد شد. یکی از راه‌های برطرف نمودن این مشکل، تعریف کلاس دیگری است که در این پژوهش کلاس رشته – مقطع نام گرفته است. بنابراین ۱۰ کلاس فراداده تعریف شده‌است که با احتساب دیگر بخش‌ها همچون

استخراج فراداده در پارساها با رویکرد الگوریتم‌های دسته‌بندی □ ۷۴

تقدیم‌نامه، سپاس‌گزاری، فهرست‌ها، تعهدنامه و ... ، کلاس دیگری به نام سایر اضافه شده و در مجموع ۱۱ کلاس وجود دارد.

پایان نامه کارشناسی ارشد در رشته مدیریت صنعتی □

(ب)

پایان نامه برای دریافت درجه ی کارشناسی ارشد □
در رشته مدیریت دولتی □

(الف)

شکل ۳-۶ در شکل (الف) مقطع و رشته ی تحصیلی در دو پاراگراف نوشته شده، در حالی که در نمونه ی شکل (ب) رشته و مقطع در یک پاراگراف آمده‌اند. علامت □ نشانه ی انتهای پاراگراف است.

همانند بخش قبل نرمال‌سازی بردار ویژگی انجام شده، طبقه‌بندی با دسته‌بند ماشین بردار پشتیبان صورت گرفته و چهار تابع هسته پرکاربرد یعنی تابع خطی، چندجمله‌ای، تابع پایه شعاعی و تابع سیگموید بررسی شده که در فصل بعدی درباره انتخاب تابع هسته مناسب و پارامترهای آن‌ها توضیح داده خواهد شد. همچنین ارزیابی به روش اعتبارسنجی متقابل k -fold و $k=10$ صورت گرفته است.

۳-۲-۱. پس‌پردازش

پس از طبقه‌بندی پاراگراف‌های پیشگفتار به هر یک از فراداده‌های موردنظر، نیاز است تا پردازشی صورت گیرد تا فراداده‌ها به طور دقیق به دست آیند. به عنوان نمونه نام اساتید و دانشجو از پاراگراف‌های با برجسب‌های مذکور استخراج می‌شود. پس‌پردازش انجام شده شامل مراحل زیر است:

- تمامی پاراگراف‌ها با برجسب یکسان به یکدیگر الصاق می‌شوند.

- پاراگراف‌های با برچسب *دانشگاه*، *دانشکده*، *چکیده* و *کلمات کلیدی* بدون پردازش دیگری باقی می‌مانند.
- برای دستیابی به رشته دانشجو، می‌توان با تعریف عبارات منظم از پاراگراف‌های رشته و یا رشته - مقطع بهره گرفت.
- مقطع تحصیلی با جستجوی عبارات «کارشناسی ارشد»، «دکتری» یا «دکتر» در پاراگراف‌های مقطع یا رشته - مقطع قابل تشخیص است.
- عنوان پایان‌نامه/رساله با حذف کلماتی چون «عنوان»، «موضوع» و کاراکترهایی چون دو نقطه (:): به دست می‌آید.
- نام و نام خانوادگی دانشجو با حذف کاراکتر دو نقطه (در صورت وجود) و کلمات متداول در معرفی دانشجو قابل استخراج است.
- در دستیابی به نام اساتید، نیاز است تا اساتید راهنما و مشاور از یکدیگر تمیز داده شوند. همچنین ممکن است تعداد استادهاى راهنما (یا مشاور) بیش از یک استاد باشد. با بررسی کلمات «راهنما» یا «مشاور» می‌توان نام استاد موردنظر را شناسایی نمود. همچنین با جستجوی کلمات «استادها» و یا «اساتید» مشخص می‌شود که دو استاد راهنما (مشاور) در پایان‌نامه/رساله مشارکت کرده‌اند یا خیر. در نهایت با حذف کاراکتر دو نقطه و عبارات «استاد راهنما»، «استاد مشاور» و عبارات مشابه، می‌توان نام کامل اساتید را به دست آورد.

۳-۳. جمع‌بندی

در این فصل روش پژوهش به صورت مرحله به مرحله تشریح شد. ابتدا در گام تشخیص بدنه، پس از طبقه‌بندی پاراگراف‌ها به یکی از دو کلاس فراداده و بدنه، پاراگراف شروع فصل اول هر پایان‌نامه/رساله شناسایی می‌شود و سپس پاراگراف‌های بدنه کنار گذاشته شده و پاراگراف‌های

صفحات آغازین به گام بعد، یعنی طبقه‌بندی فراداده‌ها، وارد می‌شوند. در این فاز پاراگراف‌ها به فراداده‌های از قبل تعریف شده‌ای به ۱۱ کلاس دسته‌بندی می‌شوند. در هر دو مرحله از طبقه‌بند ماشین بردار پشتیبان بهره گرفته شده است. در انتها با انجام چند پردازش، فراداده‌ها به طور کامل استخراج می‌شوند.

فصل چهارم

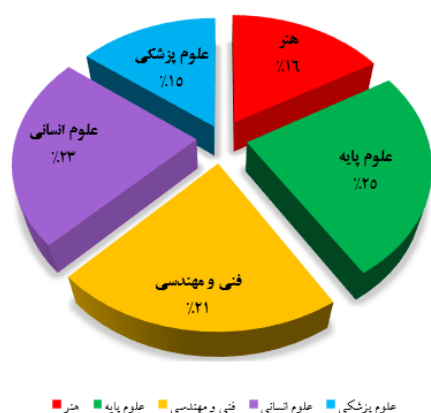
نتایج تجربی

۴-۱. مقدمه

در این فصل به ارزیابی روش پیشنهادی و تحلیل نتایج به دست آمده پرداخته می‌شود. در ابتدای فصل درباره نحوه جمع‌آوری داده‌ها و توزیع آن‌ها توضیح بحث شده و در ادامه به معیارهای ارزیابی یادگیری ماشین با نظارت اشاره می‌شود. سپس آزمایش‌های انجام شده برای ارزیابی مراحل تشخیص بدنه و طبقه‌بندی فراداده‌ها تشریح می‌گردد.

۴-۲. مجموعه داده

در زمینه این پژوهش مجموعه داده عمومی در زبان فارسی در دسترس نیست. مجموعه داده‌ی مورد استفاده به صورت تصادفی از پایان‌نامه‌ها و رساله‌های ثبت شده در سامانه ثبت پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) جمع‌آوری شده است. بدین منظور ۲۵۰ پایان‌نامه/رساله در فاز تشخیص بدنه مورد استفاده قرار گرفته است که مجموعاً شامل ۷۰,۸۶۲ پاراگراف هستند. علاوه بر آن ۳۱۵ پایان‌نامه/رساله با مجموعاً ۲۹,۶۵۲ پاراگراف در مرحله طبقه‌بندی فراداده‌ها به کار گرفته شده است. این اسناد به صورت موضوعی به پنج دسته علوم پایه، علوم پزشکی، فنی و مهندسی، هنر و علوم انسانی تقسیم می‌گردند. Error! Reference source not found. جزییات فراوانی انواع پایان‌نامه‌ها را نمایش می‌دهد.



شکل ۴-۱ فراوانی انواع پایان‌نامه‌ها/رساله‌ها

۳-۴. معیارهای ارزیابی

در ارزیابی عملکرد طبقه‌بندی، معیارهای دقت^۱، صحت^۲، بازخوانی^۳ و $F1-socre$ مورد استفاده قرار گرفته است که در ادامه درباره نحوه محاسبه آن‌ها توضیح داده می‌شود (Hossin & Sulaiman, 2015).

در مسائل طبقه‌بندی دو کلاسه، اگر تعلق یک نمونه جدید به کلاسی را مثبت و عدم تعلق به همان کلاس را منفی در نظر بگیریم، در این صورت به ازای هر نمونه یکی از چهار حالت پیش رو در انتظار خواهد بود:

- نمونه در کلاس مثبت دسته‌بندی شده و برچسب واقعی آن نیز مثبت است. (مثبت صحیح^۴)
- نمونه در کلاس منفی طبقه‌بندی شده و برچسب واقعی آن نیز منفی است. (منفی صحیح^۵)

¹ Accuracy

² Precision

³ Recall

⁴ True positive

⁵ True negative

- نمونه در کلاس مثبت دسته‌بندی شده درحالی‌که برچسب واقعی آن منفی است. (مثبت کاذب)^۱

- نمونه در کلاس منفی طبقه‌بندی شده درحالی‌که برچسب واقعی آن مثبت است. (منفی کاذب)^۲

اگر t_p, t_n, f_p و f_n به ترتیب نشان‌دهنده مثبت صحیح، منفی صحیح، مثبت کاذب و منفی کاذب در نظر بگیریم، آن‌گاه دقت الگوریتم طبقه‌بندی به صورت زیر محاسبه می‌گردد:

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \quad (1-4)$$

در واقع معیار دقت برابر ایت با میزان تشخیص‌های صحیح نسبت به کل تشخیص‌های طبقه‌بند است. به همین ترتیب میزان صحت دسته‌بند نیز با رابطه زیر به دست می‌آید:

$$Precision = \frac{t_p}{t_p + f_p} \quad (2-4)$$

معیار صحت به خوبی نشان می‌دهد که چه میزان از نمونه‌هایی که توسط طبقه‌بند در کلاس مثبت جای گرفته‌اند، به درستی طبقه‌بندی شده‌اند. از طرف دیگر، معیار بازخوانی بیان می‌کند که چه تعداد از نمونه‌هایی که برچسب واقعی آن‌ها کلاس مثبت بوده به درستی طبقه‌بندی شده‌اند (فرمول (۳-۴)).

$$Recall = \frac{t_p}{t_p + f_n} \quad (3-4)$$

معیار F1-score صحت و بازخوانی را با یکدیگر ترکیب کرده و به نوعی میانگین این دو مقدار است:

$$F1 - score = \frac{2 \times (Precision \cdot Recall)}{Precision + Recall} \quad (4-4)$$

¹ False positive

² False negative

۴-۴. ارزیابی مرحله تشخیص بدنه

در این بخش عملکرد مرحله شناسایی بدنه بررسی می‌شود و آزمایش‌های انجام گرفته تشریح می‌گردند. همچنین درباره بهینه‌سازی پارامترهای طبقه‌بند ماشین بردار پشتیبان توضیحاتی ارائه می‌شود.

همان‌طور که در فصل قبل نیز بیان شد، در این فاز پاراگراف‌های پایان‌نامه/رساله به دو کلاس فراداده و بدنه طبقه‌بندی می‌شوند تا در ادامه آن، شروع بدنه اصلی شناسایی شود و صفحات آغازین برای استخراج فراداده‌ها مورد پردازش قرار گیرند. بدین ترتیب ضروری است تا پارامترهای الگوریتم دسته‌بندی با مقادیری مناسب مقداردهی شوند. در ماشین بردار پشتیبان C و γ به ترتیب پارامترهای کنترل خطا و تابع هسته بوده که مقداردهی متفاوت آن‌ها بر مدل آموزشی به دست آمده از دسته‌بند تأثیر دارد. یکی از روش‌های متداول در یافتن مقدار پارامترها جستجوی شبکه‌ای^۱ است که در جدول ۴-۱ نتایج حاصل از این آزمایش را به طور خلاصه نشان می‌دهد. توابع خطی، چندجمله‌ای درجه چهارم، تابع پایه شعاعی و سیگموید چهار تابع هسته مورداستفاده در این آزمایش هستند. پارامترهای C و γ می‌توانند یکی از مقادیر مجموعه $\{2^i \mid i \in [-7, 7]\}$ باشند. نتایج به دست آمده مقادیری از دو پارامتر یاد شده هستند که به ازای آن‌ها بیشترین میزان $F1$ -score برای تابع هسته مورد نظر به دست آمده است. در طول این آزمایش بهترین تابع هسته برای ماشین بردار پشتیبان نیز مشخص می‌شود. همان‌طور که ملاحظه می‌شود، تابع پایه شعاعی با اختلاف زیادی نسبت به سایر توابع به $F1$ -score بالاتری دست یافته است. بنابراین به نظر می‌رسد که این تابع هسته برای مسئله طبقه‌بندی در این پژوهش مناسب‌تر است.

¹ Grid search

جدول ۴-۱. نتایج به دست آمده در تنظیم پارامترهای SVM در مرحله تشخیص مرز به روش جستجوی شبکه‌ای

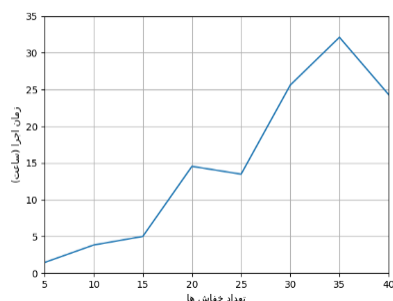
سیگموئید	RBF	چندجمله‌ای درجه ۴	خطی	تابع هسته
۶, -۴	۷, ۷	۶, ۳	۶, ۲	$\log_2 C, \log_2 \gamma$
۶۳/۳	۹۵/۷	۸۹/۱	۸۹/۹	میانگین F1-score (%)

به منظور دستیابی به پارامتر بهتر در ماشین بردار پشتیبان به همراه تابع پایه شعاعی، بهینه‌سازی دیگری با الگوریتم خفاش صورت گرفته است که در ادامه توضیح داده می‌شود.

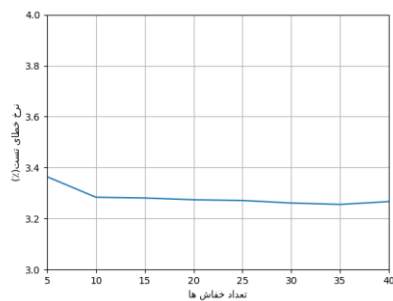
پژوهش‌های بسیاری در بررسی الگوریتم بهینه‌سازی خفاش در مقایسه با سایر الگوریتم‌های تکاملی انجام شده است. برای نمونه نتایج به دست آمده در (Naderi & Khamehchi, 2017) و (Tharwat, Hassanien, & Elnaghi, 2017) نشان می‌دهد که الگوریتم خفاش در مقایسه با الگوریتم‌های مشابه مانند الگوریتم ژنتیک، پرندگان و ... سریع‌تر به همگرایی می‌رسد و به مراتب نتایج بهتری را به همراه خواهد داشت. از طرف دیگر، می‌توان گفت که الگوریتم خفاش تعمیم‌یافته الگوریتم پرندگان بوده است و یا به بیان دیگر الگوریتم پرندگان نوع خاصی از الگوریتم خفاش است؛ چرا که تمامی پارامترها در طول اجرای الگوریتم پرندگان ثابت هستند. درحالی‌که در الگوریتم خفاش، دو پارامتر بلندی صدا و نرخ پالس صدا می‌تواند در طول زمان تغییر کند و این امر منجر به تعادل بین اکتشاف و جستجو می‌شود (Tharwat, Hassanien, & Elnaghi, 2017). از این‌رو الگوریتم خفاش برای بهینه‌سازی در این مرحله انتخاب شده است.

به منظور یافتن تعداد خفاش‌ها، آزمایشی انجام شده که در آن الگوریتم هر بار با مقادیر مختلف خفاش‌ها اجرا می‌شود و تا برقراری شرط خاتمه ادامه می‌یابد. شرط خاتمه در این آزمایش در دو حالت می‌تواند رخ دهد: الگوریتم تا تکرار مشخص و ثابتی اجرا شود یا بهترین پاسخ پس از پنج تکرار بدون تغییر باقی بماند (Tharwat, Hassanien, & Elnaghi, 2017). اجرای الگوریتم با ۵ خفاش شروع شده و هر بار با افزودن ۵ خفاش مجدد اجرا می‌شود تا به ۴۰ خفاش برسد. در هر اجرا،

کمترین خطای به دست آمده و زمان اجرای الگوریتم ثبت شده است که نتایج در **Error! Reference source not found.** (الف) و (ب) قابل مشاهده هستند. همان‌طور که در شکل مشخص است از تعداد ۱۰ خفاش به بعد میزان خطای به دست آمده تغییر چندانی نداشته و به نظر می‌رسد تعداد ۱۰ خفاش برای بهینه‌سازی با روش الگوریتم خفاش در این مسئله مناسب است. بدیهی است که در اکثر مواقع با افزایش تعداد خفاش‌ها زمان اجرای الگوریتم نیز افزایش می‌یابد.



(ب)

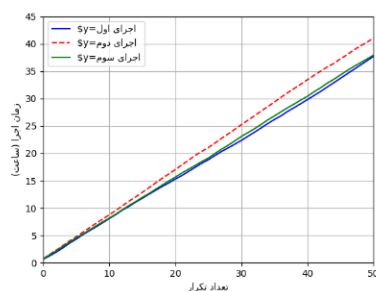


(الف)

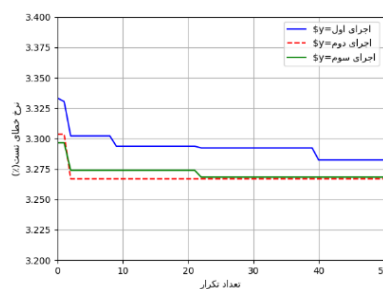
شکل ۴-۲ آزمایش تاثیر تعداد خفاش بر عملکرد الگوریتم : (الف) نرخ خطای آزمایش و (ب) زمان اجرا به ازای مقادیر مختلف خفاش‌ها

پس از تعیین تعداد خفاش‌ها، برای تعیین تعداد تکرار الگوریتم نیز آزمایش دیگری انجام شده است. بدین منظور الگوریتم با بیشینه تکرار ۵۰ اجرا شده و در پایان هر تکرار کمترین خطا و زمان اجرا ثبت گردیده است. این آزمایش سه مرتبه تکرار شده و نتایج به دست آمده در **Error! Reference source not found.** قابل مشاهده است. واضح است با افزایش تکرار الگوریتم، مقدار خطای به دست آمده کاهش یافته و در مقابل زمان اجرا افزایش می‌یابد. در **Error! Reference source not found.** فضای جستجوی الگوریتم ترسیم شده که علامت ستاره نشان دهنده مختصات مکان به دست آمده با الگوریتم خفاش است که کمترین خطا را در بین مکان‌های جستجو شده داشته است. بدیهی

است که مکان بهینه در این مسائل دست‌یافتنی نبوده و هیچ یک از الگوریتم‌های تکاملی این مهم را تضمین نمی‌کنند.

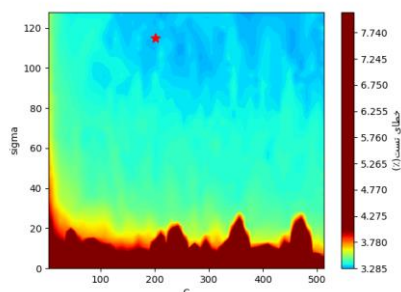


(ب)



(الف)

شکل ۴-۳ تأثیر تعداد خفاش بر عملکرد الگوریتم: (الف) نرخ خطای آزمایش به تعداد تکرارهای مختلف. (ب) زمان اجرا به ازای افزایش تعداد تکرار



شکل ۴-۴ نمودار کانتوری نرخ خطا به ازای پارامترهای خطا (c) و سیگما. نقطه‌ای که با ستاره مشخص شده نقطه مینیمی است که به کمک الگوریتم خفاش به دست آمده است.

پیش‌تر در آزمایشی مشخص شد که ماشین بردار پشتیبان در مسئله این پژوهش با تابع پایه شعاعی عملکرد بهتری نسبت به سایر توابع هسته پر استفاده دارد. در آزمایش دیگری پارامترهای ماشین بردار پشتیبان با استفاده از این تابع هسته با الگوریتم خفاش نیز مقاداردهی گردید. در جدول ۴-۲ مقایسه‌ای

بین کمترین خطای به دست آمده به روش جستجوی شبکه‌ای و الگوریتم تکاملی خفاش صورت گرفته‌است. نتایج نشان می‌دهد الگوریتم خفاش در زمان سریع‌تری به نرخ خطای کمتری دست یافته است.

جدول ۴-۲. مقایسه‌ی زمان اجرا و خطای آزمایش به دست آمده در الگوریتم خفاش و جستجوی شبکه‌ای

BA- SVM		Grid search-SVM	
درصد خطای تست	زمان اجرا (ساعت)	درصد خطای تست	زمان اجرا (ساعت)
۳/۲۶۸	۲/۳۴	۳/۳۱۴	۵/۲۲

جهت بررسی تأثیر ویژگی‌ها و مقایسه عملکرد طبقه‌بند، آزمایشی با ترکیب انواع ویژگی‌ها در مرحله تشخیص مرز انجام شده‌است. نتایج جدول ۴-۳ نشان می‌دهد که ویژگی‌های هیوریستیک به تنهایی نمی‌توانند میان پاراگراف‌های فراداده و بدنه تمایز ایجاد کند. همان‌طور که در فصل قبل اشاره شد، ویژگی‌های هیوریستیک مرتبط با خصیصه‌های ذاتی متن هستند. تعداد کلمات، خطوط، علائم نگارشی و اعداد نمونه‌ای از این ویژگی‌ها هستند که در این فاز بیشترین تعداد ویژگی‌ها را به خود اختصاص داده‌اند. با بررسی عبارات متداول در شروع متن اصلی بدنه (ویژگی لغوی) همانند «فصل اول»، «مقدمه»، «پیشگفتار»، «دیباچه» و یا «کلیات تحقیق» طبقه‌بند عملکرد بهتری داشته است و برای نمونه میزان $F1\text{-score}$ آن از ۷۹/۵ درصد به ۸۳/۴ افزایش یافته است. لازم به ذکر است که یافته‌های ثبت شده از میانگین دو کلاس به دست آمده‌اند. از طرف دیگر ترتیب پاراگراف‌ها (ویژگی توالی) تأثیر بسزایی در نتایج داشته و درصدها به بالای ۹۰ درصد افزایش پیدا کردند.

به نظر می‌رسد بهبود نتایج با در نظر گرفتن این ویژگی به این سبب بوده که بدیهی است پاراگراف‌های اول هر پایان‌نامه/رساله با احتمال بیشتری متعلق به کلاس فراداده‌ها هستند و هرچه به پاراگراف‌های جلوتر می‌رسیم احتمال تعلق به کلاس بدنه افزایش می‌یابد. بنابراین ترتیب پاراگراف‌ها می‌تواند ویژگی مناسبی در تشخیص فراداده‌ها از بدنه باشد. ترکیب ویژگی‌های هیوریستیک با ویژگی لغوی و توالی نیز به نتایج نسبتاً مناسبی منجر شده است. این در حالی است که ویژگی نسبت

طول به عرض پاراگراف (ویژگی هندسی) موجب کاهش درصدها شده و کمکی به بهبود نتایج نمی‌کند. یافته‌های به دست آمده در این آزمایش نشان می‌دهد که ترکیب این ویژگی با سایر ویژگی‌ها نیز سبب تنزل درصدها شده است. بنابراین به نظر می‌رسد که در فاز تشخیص بدنه، ویژگی هندسی مناسب نیست و می‌تواند کنار گذاشته شود.

جدول ۴-۳. مقایسه تاثیر انواع ویژگی‌ها بر عملکرد طبقه‌بند پاراگراف‌ها در مرحله تشخیص مرز بدنه

نوع ویژگی / معیارهای ارزیابی (میانگین)	دقت (%)	صحت (%)	بازخوانی (%)	F1-score (%)
هیوریتیک	۷۶/۰	۸۷/۴	۷۶/۰	۷۹/۵
هیوریتیک + هندسی	۷۶/۱	۸۷/۴	۷۶/۱	۷۹/۵
هیوریتیک + لغوی	۸۷/۳	۸۳/۶	۸۳/۲	۸۳/۴
هیوریتیک + توالی	۹۴/۴	۹۵/۵	۹۵/۲	۹۵/۳
هیوریتیک + هندسی + لغوی	۸۳/۴	۸۳/۷	۸۳/۴	۸۳/۵
هیوریتیک + توالی + لغوی	۹۶/۷	۹۵/۸	۹۵/۵	۹۵/۷
هیوریتیک + هندسی + توالی	۹۵/۲	۹۵/۵	۹۵/۲	۹۵/۳
هیوریتیک + هندسی + توالی + لغوی	۹۶/۶	۹۵/۸	۹۵/۴	۹۵/۶

برای ارزیابی طبقه‌بندی در مرحله‌ی تشخیص بدنه، از روش اعتبارسنجی k-fold استفاده شده (k=10)

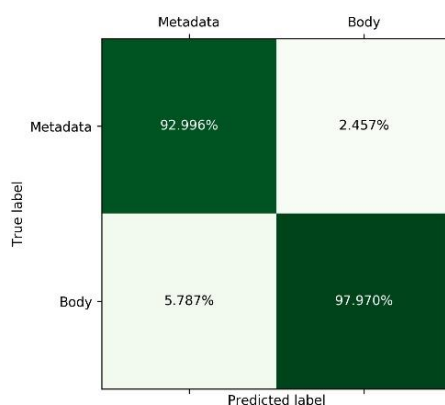
و در هر گام درصد صحت، بازخوانی و F1-score ثبت شده که در

خلاصه نتایج ارائه شده است. نتایج به دست آمده در گام‌های ارزیابی نسبتاً مشابه بوده‌اند و یافته‌های این جدول، میانگین دو کلاس هستند. لازم به ذکر است که داده‌ها در هر گام به صورت تصادفی انتخاب شده‌اند تا تنوع نمونه‌های آموزشی و آزمایش افزایش یابد.

گام‌های ارزیابی	صحت (%)	بازخوانی (%)	F1-score (%)
۱	۹۶/۰	۹۵/۸	۹۵/۹
۲	۹۵/۷	۹۵/۴	۹۵/۵
۳	۹۶/۰	۹۵/۴	۹۵/۷
۴	۹۶/۰	۹۵/۶	۹۵/۸
۵	۹۵/۷	۹۵/۳	۹۵/۵
۶	۹۵/۵	۹۵/۷	۹۵/۶
۷	۹۵/۴	۹۵/۱	۹۵/۲
۸	۹۵/۹	۹۵/۸	۹۵/۸
۹	۹۵/۶	۹۵/۱	۹۵/۴
۱۰	۹۶/۰	۹۵/۶	۹۵/۸
میانگین	۹۵/۸	۹۵/۵	۹۵/۶

جدول ۴-۴. نتایج به دست آمده در گام‌های اعتبارسنجی متقابل 10-fold مرحله‌ی تشخیص مرز

Error! Reference source not found. ماتریس درهم‌ریختگی طبقه‌بند پاراگراف‌ها در گام تشخیص بدنه را نشان می‌دهد. عدم تعادل بین نمونه‌های دو کلاس بر عملکرد طبقه‌بند بی‌تأثیر نبوده‌است. به بیان دیگر، تعداد نمونه‌های کلاس بدنه از نمونه‌های کلاس فراداده بیشتر بوده و همین امر در نتایج به دست آمده مؤثر است. مشاهده می‌شود که بیشترین خطا در تخصیص برچسب فراداده به پاراگراف‌های بدنه است (مثبت کاذب) و پس از آن خطای منفی کاذب احتمال رخداد دارد. خطای مثبت کاذب ممکن است در پاراگراف‌های کوتاه بدنه رخ دهد. برای مثال عناوین بخش‌های مختلف هر فصل می‌تواند مستعد این نوع خطا باشد. از طرف دیگر خطای منفی کاذب یعنی تخصیص برچسب بدنه به پاراگراف‌های فراداده می‌تواند در پاراگراف‌های چکیده رخ دهد. همچنین بخش‌های دیگری همچون متون آیین‌نامه‌ها و یا تقدیم‌نامه (در صورت طولانی بودن) ممکن است اشتباه طبقه‌بندی شوند.



شکل ۵-۴. ماتریس درهم ریختگی طبقه‌بند فراداده و بدنه

آزمون دیگر برای بررسی عملکرد طبقه‌بند در انواع پایان‌نامه‌ها انجام شده است (جدول ۵-۴). یافته‌های این آزمایش نشان می‌دهد که درصدهای پایان‌نامه‌های علوم انسانی بهتر از سایر پایان‌نامه‌ها است. در واقع تفاوت در ساختار پایان‌نامه‌ها و نوشتارها در رشته‌های مختلف، علت این تفاوت‌هاست. از طرفی پایان‌نامه‌های علوم پایه و انسانی به ترتیب بیشترین فراوانی را نسبت به دیگر رشته‌ها داشته‌اند که این مهم نیز در نتایج به دست آمده مؤثر بوده است.

جدول ۵-۴. مقایسه عملکرد طبقه‌بند در انواع پایان‌نامه‌ها/رساله‌ها در مرحله تشخیص مرز

نوع پایان‌نامه/رساله	میانگین دقت.٪	میانگین صحت.٪	میانگین بازخوانی.٪
علوم پایه	۹۷/۰	۹۶/۶	۹۶/۰
علوم انسانی	۹۷/۵	۹۷/۰	۹۶/۵
فنی و مهندسی	۹۶/۵	۹۵/۰	۹۴/۰
هنر	۹۷/۴	۹۶/۹	۹۶/۴
علوم پزشکی	۹۶/۱	۹۵/۸	۹۵/۱

۴-۴-۱. ارزیابی مرحله طبقه‌بندی فراداده

در این بخش آزمایش‌های انجام شده برای ارزیابی روش پیشنهادی در طبقه‌بندی فراداده‌ها تشریح می‌شود.

همانند فاز قبلی، برای تعیین پارامترهای ماشین بردار پشتیبان و انتخاب تابع هسته‌ی مناسب از روش جستجوی شبکه‌ای استفاده شده است. بدین منظور پارامتر کنترل خطا (C) و پارامتر هسته (γ) در مجموعه $\{2^i \mid i \in [-7, 7]\}$ جستجو شده و اعدادی که به ازای آن، درصد F -score طبقه‌بند با استفاده از تابع هسته مورد استفاده بیشتر بوده ثبت شده است. طبق یافته‌های به دست آمده به نظر می‌رسد در این فاز نیز تابع پایه شعاعی از سایر توابع هسته بهتر است. اگرچه درصد حاصل از حالت خطی ماشین بردار پشتیبان نیز به آن نزدیک است (جدول ۴-۶).

جدول ۴-۶. نتایج به دست آمده در تنظیم پارامترهای SVM در مرحله‌ی طبقه‌بندی فراداده به روش جستجوی شبکه‌ای

سیگموئید	RBF	چند جمله‌ای درجه ۴	خطی	تابع هسته
۶, -۶	۶, -۲	۱, -۱	۶, -۱	$\log_2 C, \log_2 \gamma$
۸۹/۵	۹۱/۳۶	۹۰/۵	۹۱/۳۴	میانگین F1-score (%)

جهت بررسی تأثیر انواع ویژگی‌ها بر عملکرد طبقه‌بند، در هر مرحله ویژگی‌های مختلفی آزمایش شده تا مشخص شود هر دسته از ویژگی‌ها چه میزان بر بهبود عملکرد دسته‌بند مؤثرند. همچنین ویژگی محلی-پیشنهادی این پژوهش- نیز با سایر ویژگی‌ها مقایسه شده است.

واضح است که در این مسئله، کلمات و عبارات به کار رفته در پاراگراف‌های فراداده‌ها در یافتن نوع فراداده بسیار کمک‌کننده هستند. برای نمونه عبارت «چکیده» در یک پاراگراف می‌تواند بیانگر این باشد که به احتمال بالایی پس از این عبارت فراداده چکیده قابل استخراج است. اما این ویژگی به تنهایی کافی نیست و افزودن ویژگی‌های هیوریستیک تأثیر بسزایی در بهبود درصدها داشته است. لازم به ذکر است که ویژگی‌های هیوریستیک نیز به تنهایی در تشخیص فراداده‌ها مناسب نیستند و نتایج به دست آمده از این نوع ویژگی به تنهایی درصدهای بسیار پایینی دارند. ویژگی‌های وابسته به سبک و قالب نوشتار مانند اندازه فونت و یا پررنگ بودن متن نیز در تشخیص فراداده‌ها از هم و تمیز آن‌ها از سایر بخش‌های پیش‌گفتار پایان‌نامه/رساله مناسب است. برای مثال عنوان پایان‌نامه/رساله در اندازه بزرگ و معمولاً پررنگ نوشته می‌شود. افزودن ویژگی‌های هندسی در این فاز موجب بهتر شدن عملکرد طبقه‌بند فراداده‌ها شده است و افزودن آن به ویژگی‌های لغوی و هیوریستیک درصد $F1-score$ را از ۸۱/۲ به ۸۶/۲ رسانده است.

یافته‌های این آزمایش نشان می‌دهد در نظر گرفتن روابط پاراگراف‌های متوالی مانند برچسب و قالب نوشتاری پاراگراف قبلی نیز در افزایش مقادیر معیارهای ارزیابی مؤثر بوده و میزان $F1-score$ به بالای نود درصد ارتقا یافته است. درصدهای حاصل از ترکیب کل ویژگی‌ها نشان می‌دهد ادغام آن‌ها با

یکدیگر نیز مفید بوده است. در نهایت ویژگی‌های پیشنهادی که به نام ویژگی‌های محلی از آن‌ها یاد شد، با تمام ویژگی‌ها ادغام شده و نتایج به دست آمده حاکی از بهتر شدن یافته‌هاست. ویژگی‌های محلی در تشخیص فراداده‌هایی که در چند پاراگراف ظاهر می‌شوند، مؤثر هستند. به عنوان نمونه نام اساتید راهنما، نام دانشجو، عنوان پایان‌نامه/رساله و پاراگراف‌های چکیده از فراداده‌هایی هستند که ممکن است در چند پاراگراف پشت سر هم قرار گیرند. نتایج به دست آمده از این آزمایش در جدول ۷-۴ خلاصه شده است.

جدول ۷-۴. مقایسه‌ی تأثیر انواع ویژگی‌ها بر عملکرد طبقه‌بند فراداده‌ها

نوع ارزیابی (میانگین)	ویژگی / معیارهای	دقت (%)	صحت (%)	بازخوانی (%)	F1-score (%)
لغوی		۸۹/۸	۷۱/۲	۶۵/۸	۶۵/۶
لغوی + هیوریستیک		۹۳/۸	۹۰/۰	۷۸/۳	۸۱/۲
لغوی + هیوریستیک + قالب		۹۴/۰	۸۹/۹	۷۸/۱	۸۱/۲
لغوی + هیوریستیک + هندسی		۹۵/۴	۹۰/۳	۸۳/۹	۸۶/۲
لغوی + هیوریستیک + توالی		۹۷/۹	۹۲/۰	۹۰/۰	۹۰/۷
کل ویژگی‌ها		۹۸/۰	۹۱/۸	۸۹/۹	۹۰/۵
کل ویژگی‌ها + ویژگی‌های محلی		۹۸/۱	۹۲/۳	۹۰/۸	۹۱/۳

در ارزیابی متقابل 10-fold میانگین درصدهای صحت، بازخوانی و $F1-score$ نشان می‌دهد که پراکندگی بین یافته‌های به دست آمده به نسبت کم است و افزایش یا کاهش قابل توجهی در میزان مقادیر گام‌ها مشاهده نمی‌شود. تنها در گام پنجم است که میزان $F1-score$ به زیر نود درصد رسیده است.

جدول ۸-۴. نتایج به دست آمده در گام‌های اعتبارسنجی متقابل 10-fold مرحله تشخیص مرز

استخراج فراداده در پارساها با رویکرد الگوریتم‌های دسته‌بندی □ ۹۲

گام‌های ارزیابی	صحت (%)	بازخوانی (%)	F1-score (%)
۱	۹۲/۱	۹۱/۴	۹۱/۶
۲	۹۲/۵	۸۹/۲	۹۰/۷
۳	۹۱/۹	۹۲/۷	۹۲/۲
۴	۹۱/۸	۹۰/۱	۹۰/۸
۵	۹۱/۱	۸۸/۸	۸۹/۸
۶	۹۴/۰	۹۱/۵	۹۲/۷
۷	۹۱/۸	۹۱/۱	۹۱/۴
۸	۹۳/۹	۹۱/۵	۹۲/۶
۹	۹۴/۷	۹۳/۷	۹۴/۰
۱۰	۹۱/۷	۸۹/۱	۹۰/۴
میانگین	۹۲/۶	۹۰/۹	۹۱/۶

ماتریس درهم‌ریختگی برای کلاس‌های فراداده‌ها در **Error! Reference source not found.** نشان داده شده است. همان‌طور که ملاحظه می‌شود، بیشترین خطاها در کلاس «سایر» رخ می‌دهد. بیشترین خطای مثبت کاذب در کلاس «سایر» نسبت به فراداده «دانشجو» است. به این معنا که نام دانشجو نسبت به فراداده‌های دیگر، بیشتر در معرض خطای طبقه‌بندی در کلاس «سایر» است. پس از آن به ترتیب کلاس «دانشگاه»، «عنوان» و «استاد راهنما» بیشترین میزان خطای مثبت کاذب را نسبت به کلاس مذکور دارند؛ اگرچه احتمال اشتباه در تمامی فراداده‌ها مشاهده می‌شود. به نظر می‌رسد بیشتر بودن نمونه‌های این کلاس نسبت به کلاس فراداده‌های پایان‌نامه، در ایجاد خطای طبقه‌بندی مؤثر بوده است.



شکل ۴-۶ ماتریس درهم‌ریختگی طبقه‌بندی فراداده‌ها (نرمال شده)

بیشترین خطای منفی کاذب نسبت به کلاس «سایر» در فراداده‌های «عنوان»، «دانشگاه» و «دانشجو» است. خطاهای بالا در این فراداده‌ها می‌تواند به این دلایل زیر باشد: در برخی پایان‌نامه‌ها، قبل از عنوان پایان‌نامه/رساله و یا دانشجو، عبارات توضیحی و متداول این بخش نوشته نمی‌شود، از طرفی تعداد نمونه‌های کلاس فراداده «دانشگاه» نیز به نسبت کم است و این مهم در نتایج به دست آمده تأثیرگذار است.

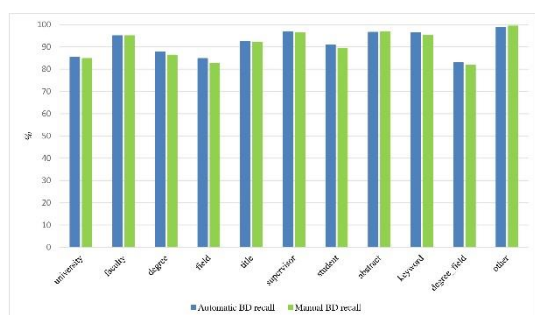
در کلاس‌های «مقطع تحصیلی»، «رشته‌ی تحصیلی» و «مقطع-رشته» نیز احتمال اشتباه در طبقه‌بندی به دلیل هم‌پوشانی با یکدیگر وجود دارد و بیشترین خطا بین کلاس‌های «مقطع» و «مقطع-رشته» رخ می‌دهد. نام دانشجو و اساتید نیز دو کلاس دیگری هستند که امکان خطا بین این دو کلاس نسبتاً بالاست. بین کلاس‌های «چکیده» و «کلمات کلیدی» نیز به دلیل متوالی بودن این دو فراداده در پایان‌نامه‌ها/رساله‌ها گاهی احتمال خطا وجود دارد.

همانند فاز قبل، آزمایشی برای بررسی عملکرد طبقه‌بند بر انواع رشته‌ها صورت گرفته است. این بار بهترین نتایج در پایان‌نامه‌های رشته‌ی مهندسی به دست آمده است. از طرفی کمترین درصد‌ها متعلق به رشته‌های پزشکی است که به نظر می‌رسد یکی از دلایل آن کمترین میزان تعداد پایان‌نامه‌ها در رشته مذکور است (جدول ۴-۹).

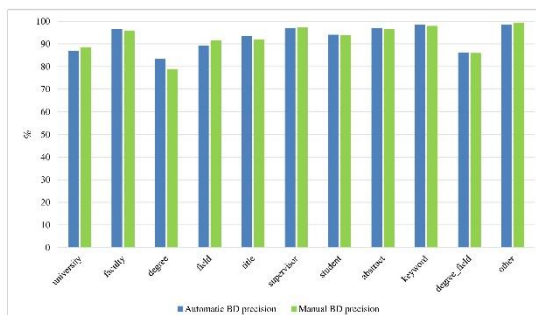
نوع پایان‌نامه	میانگین دقت.٪	میانگین صحت.٪	میانگین بازخوانی.٪
علوم پایه	۹۷/۵	۸۹/۵	۸۵/۹
علوم انسانی	۹۷/۹	۹۰/۸	۸۷/۱
فنی و مهندسی	۹۸/۱	۹۳/۲	۹۰/۹
هنر	۹۷/۴	۹۰/۶	۸۷/۹
علوم پزشکی	۷۸/۳	۷۰/۴	۶۶/۸

۲-۴-۴. ارزیابی با در نظر گرفتن جداسازی خودکار بدنه

در این زیربخش ارزیابی کل سیستم با خودکارسازی تشخیص مرز بدنه انجام شده است. در **Error! Reference source not found.** (الف) و (ب) به ترتیب مقادیر صحت و بازخوانی هر فراداده در دو حالت مقایسه شده است. حالت اول زمانی است که شناسایی مرز شروع بدنه و تعیین پاراگراف‌های فراداده به صورت دستی انجام شده و حالت دوم با خودکارسازی شناسایی مرز بدنه به روش پیشنهاد شده است. نتایج نشان می‌دهد که تفاوت زیادی میان یافته‌ها وجود نداشته و این فرآیند تقریباً به خوبی می‌تواند بدون دخالت انسان و خودکار اجرا شود.



(ب)



(الف)

شکل ۴-۷ نمودار مقایسه (الف) صحت و (ب) بازخوانی در فراداده‌های مختلف

در این فصل آزمایش‌های انجام شده برای ارزیابی روش پیشنهادی تشریح شد. بدین منظور در ارزیابی مرحله تشخیص بدنه، آزمون‌هایی برای یافتن پارامترهای مناسب در دسته‌بند ماشین بردار پشتیبان انجام شد و در ادامه میزان تأثیر انواع ویژگی‌ها بر عملکرد طبقه‌بند بررسی گردید. همچنین مقادیر به دست آمده در معیارهای ارزیابی دسته‌بند ارائه شد و در نهایت درصدهای به دست آمده در انواع پایان‌نامه‌ها مطرح شد. در مرحله طبقه‌بندی فراداده‌ها همچون مرحله قبل، آزمایش‌هایی جهت تنظیم پارامترهای ماشین بردار پشتیبان و تأثیر ویژگی‌ها بر عملکرد طبقه‌بند انجام شد و یافته‌های به دست آمده از آموزش و آزمایش مدل یادگیری بر کل داده‌ها و انواع پایان‌نامه‌ها/رساله‌ها ارائه شد. در آخر نیز عملکرد طبقه‌بندی فراداده‌ها با در نظر گرفتن تشخیص خودکار مرز بدنه با حالت دستی مقایسه گردید.

فصل پنجم

نتیجه گیری

در فصل قبل، در مورد آزمایش‌های انجام شده برای ارزیابی روش پیشنهادی بحث شد. در این فصل به نتیجه‌گیری یافته‌های این پژوهش و پیشنهاد پژوهش‌های آینده پرداخته می‌شود.

در زمینه استخراج فراداده از اسناد علمی، پژوهش‌های بسیاری بر مقالات - و نه پایان‌نامه/رساله- صورت گرفته که اغلب این پژوهش‌ها در زبان انگلیسی هستند. در زبان فارسی نیز تحقیقاتی بر روی مقالات انجام شده است. در این پژوهش، روشی برای استخراج فراداده‌ها ارائه شده و تمرکز بر اسناد از نوع پایان‌نامه/رساله در زبان فارسی بوده است. روش پیشنهادی به دو بخش تشخیص خودکار مرز بدنه و طبقه‌بندی فراداده‌ها تقسیم می‌گردد.

در مرحله نخست، هدف شناسایی پاراگراف‌های بخش پیش‌گفتار پایان‌نامه/رساله است. زیرا فراداده‌های موردنظر از قبیل نام دانشگاه، دانشکده، رشته و مقطع تحصیلی، عنوان، نام دانش‌آموخته، استاد (اساتید) راهنما و مشاور، چکیده و کلمات کلیدی در پیشگفتار هر پایان‌نامه/رساله بیان می‌شوند. بنابراین در مرحله تشخیص مرز بدنه سعی بر این است که شروع فصل اول پایان‌نامه/رساله شناسایی شود. در این مرحله پاراگراف‌ها با استفاده از الگوریتم طبقه‌بندی ماشین بردار پشتیبان به دو کلاس بدنه و فراداده طبقه‌بندی می‌شوند و پس از تشخیص اولین پاراگراف فصل اول پایان‌نامه/رساله، پاراگراف‌های قبل از آن برای استخراج فراداده‌ها به مرحله بعدی هدایت می‌شوند. در فاز بعد - همان‌طور که از نام آن پیداست- هدف طبقه‌بندی پاراگراف‌ها به فراداده‌های ذکر شده است. این طبقه‌بندی نیز با استفاده از ماشین بردار پشتیبان انجام شده است. در نهایت نیز با اعمال پس‌پردازش،

فراداده‌های نهایی به دست می‌آیند. در هر مرحله ویژگی‌های متناسب با اسناد پایان‌نامه/رساله فارسی و قالب DOCX تعریف شده است.

در ارزیابی فاز تشخیص بدنه، آزمایش‌هایی برای تنظیم پارامترهای طبقه‌بند ماشین بردار پشتیبان با مقادیر مناسب انجام شده است. اولین آزمایش با هدف یافتن بهترین تابع هسته برای دسته‌بند مذکور نشان می‌دهد که هسته تابع پایه شعاعی نسبت به سایر توابع هسته به درصدهای بهتری دست یافته است. همچنین طی این آزمایش مقادیر مناسب برای مقدار پارامتر کنترل و پارامتر تابع هسته با روش جستجوی شبکه‌ای به دست آمده است. به منظور تنظیم دقیق‌تر پارامترهای کنترل و پارامتر هسته تابع پایه شعاعی در این مرحله، آزمایش دیگری با استفاده از الگوریتم تکاملی خفاش صورت گرفته است. همان‌طور که انتظار می‌رود نتایج به دست آمده در این آزمایش نشان می‌دهد که الگوریتم خفاش می‌تواند سریع‌تر از روش جستجوی شبکه‌ای پارامترهای بهینه (بهینه محلی) را یافته و خطای طبقه‌بندی را کاهش دهد.

نتایج به دست آمده از تأثیر انواع ویژگی‌ها بر دسته‌بندی پاراگراف‌ها در مرحله تشخیص مرز گویای این است که ویژگی‌های مبتنی بر ذات و طبیعت متن - که در این پژوهش با نام ویژگی‌های هیوریستیک از آن‌ها یاد شد - به تنهایی کفایت نمی‌کند و افزودن ویژگی‌های مبتنی بر توالی پاراگراف‌ها و ویژگی‌های لغوی در بهبود عملکرد طبقه‌بند مؤثر است. در مقابل ویژگی هندسی تعریف شده نه تنها در بهبود درصدهای به دست آمده تأثیری ندارد بلکه ترکیب آن با سایر ویژگی‌ها سبب کاهش یافته‌ها می‌شود. بنابراین به نظر می‌رسد که در مرحله تشخیص مرز می‌توان ویژگی هندسی را کنار گذاشت.

در ارزیابی طبقه‌بندی پاراگراف‌ها در فاز اول، به منظور پیشگیری از بیش‌پردازش از روش اعتبارسنجی متقابل 10-fold استفاده شده است. یافته‌های این آزمون نشان می‌دهد که الگوریتم به طور

میانگین با میزان $F1-socre$ برابر با $۹۵/۶$ درصد قادر به دسته‌بندی پاراگراف‌ها است. ماتریس درهم‌ریختگی حاصل از هر ده گام نیز به دست آمده است. در این مسئله داده‌ها نامتعادل هستند و نمونه‌های کلاس بدنه بیشتر از نمونه‌های کلاس فراداده هستند. نسبت به کلاس فراداده، خطای مثبت کاذب و سپس خطای منفی کاذب به ترتیب بیشترین میزان اشتباهات در طبقه‌بندی هستند. واضح است که خطای مثبت کاذب به تخصیص برچسب فراداده به پاراگراف‌های بدنه نسبت داده می‌شود و در مقابل منظور از خطای منفی کاذب، طبقه‌بندی پاراگراف‌های فراداده در کلاس بدنه است. احتمال رخداد خطای مثبت کاذب در پاراگراف‌های کوتاه بدنه (برای نمونه عناوین بخش‌ها و زیر بخش‌ها) وجود دارد. از طرف دیگر خطای منفی کاذب می‌تواند در پاراگراف‌های طولانی پیش‌گفتار همچون پاراگراف‌های چکیده یا متون آیین‌نامه رخ دهد.

بدیهی است که پایان‌نامه‌ها/رساله‌های هر دانشگاه ساختار متفاوتی دارد. از طرفی پایان‌نامه‌ها و رساله‌های هر رشته نیز سبک و قالب متفاوتی دارند. برای نمونه در پایان‌نامه‌ها/رساله‌های رشته‌های پزشکی ممکن است چکیده از چند بخش مجزا تشکیل شده باشد یا در پایان‌نامه‌ها/رساله‌های هنر ممکن عنوان نظری با عنوان عملی پایان‌نامه/رساله متفاوت باشد. نتایج به دست آمده در ارزیابی انواع موضوعی پایان‌نامه‌ها/رساله‌ها حاکی از آن است که عملکرد طبقه‌بندی در پایان‌نامه‌ها/رساله‌های علوم انسانی بهتر از سایر رشته‌ها بوده است.

در ارزیابی مرحله طبقه‌بندی فراداده‌ها نیز همانند مرحله قبل، آزمایشی برای تعیین پارامترهای الگوریتم ماشین بردار پشتیبان و تابع هسته مناسب انجام شده است. بدین ترتیب در این فاز نیز تابع پایه شعاعی به عنوان تابع هسته الگوریتم ماشین بردار پشتیبان انتخاب شد و پارامترهای تابع هسته و کنترل خطا نیز با استفاده از روش جستجوی شبکه‌ای تنظیم شدند.

بررسی ترکیب ویژگی‌ها بر نحوه عملکرد دسته‌بند فراداده‌ها نشان می‌دهد که ویژگی‌های لغوی به تنهایی برای استخراج فراداده‌ها مناسب نیستند؛ بلکه ترکیب آن‌ها با ویژگی‌های هیورستیک، قالب، توالی و هندسی در بهبود نتایج تأثیر مثبتی دارند. در این پژوهش ویژگی‌هایی تحت عنوان ویژگی‌های محلی تعریف شدند تا محدودیت پردازش صفحات در فایل‌های DOCX تا حدودی برطرف شود و فراداده‌هایی که بنا به هر دلیلی ممکن است در چند پاراگراف قرار گیرند، بهتر شناسایی شوند. یافته‌های این آزمایش نشان داده است که ادغام ویژگی‌های پیشنهاد شده به ویژگی‌های قبلی نیز موجب افزایش درصدهای به دست آمده شده‌اند.

با انجام اعتبارسنجی متقابل به روش *10-fold* در ارزیابی طبقه‌بند فراداده‌ها، مشخص شد که *F1-score* به طور میانگین برابر با $91/6$ درصد است که نشان از عملکرد نسبتاً مناسب طبقه‌بند دارد. آنچه از ماتریس درهم‌ریختگی ۱۱ کلاس می‌توان نتیجه گرفت این است که در این مسئله نیز عدم تعادل بین داده‌های هر کلاس بر نتایج تأثیر گذاشته و بیشترین خطاها در کلاس «سایر» - که بیشترین نمونه‌ها را دارد - رخ می‌دهد. در برخی پایان‌نامه‌ها ممکن است رشته و مقطع تحصیلی در یک خط نوشته شود. بنابراین به جز کلاس‌های «مقطع تحصیلی»، «رشته تحصیلی» کلاس دیگری به نام «مقطع-رشته» تعریف شده که به دلیل هم‌پوشانی سه کلاس مذکور با یکدیگر احتمال بروز خطا وجود دارد. در این بین بیشترین خطا در کلاس‌های «مقطع» و «مقطع-رشته» رخ می‌دهد. همچنین امکان خطا بین دو کلاس نام دانشجویان و اساتید به نسبت زیاد است. بین دو کلاس «چکیده» و «کلمات کلیدی» نیز به دلیل متوالی بودن در پایان‌نامه‌ها گاهی احتمال خطا وجود دارد.

همانند فاز قبل، آزمایشی برای بررسی عملکرد طبقه‌بند فراداده بر انواع پایان‌نامه‌ها و رسالات در رشته‌های مختلف صورت گرفته است. این بار بهترین نتایج در پایان‌نامه‌ها/رساله‌های رشته‌ی مهندسی

به دست آمده است و کمترین درصدها متعلق به رشته‌های پزشکی است. به نظر می‌رسد یکی از دلایل پایین بودن درصدهای پایان‌نامه‌ها/رساله‌های پزشکی کم بودن تعداد آن‌ها است.

در ارزیابی فاز طبقه‌بندی فراداده‌ها، پاراگراف‌های بدنه به صورت دستی جدا شده و کنار گذاشته شدند. برای بررسی تأثیر خودکارسازی تشخیص مرز بدنه بر عملکرد طبقه‌بند فراداده، بار دیگر مقادیر صحت و بازخوانی مدل یادگیری دسته‌بند فراداده‌ها محاسبه شده‌اند. مقایسه این مقادیر با درصدهای قبلی نشان می‌دهد که تفاوت چندانی بین حالت دستی و اتوماتیک وجود ندارد.

محققان در پژوهش‌های آتی می‌توانند در استخراج فراداده‌ها از پایان‌نامه‌ها/رساله‌های فارسی در قالب PDF تمرکز کنند که این موضوع چالش‌های بیشتری نسبت به پژوهش انجام شده دارد. یافتن ارتباط بین فراداده‌ها (مانند اساتید و وابستگی‌های آن‌ها) می‌تواند یکی دیگر از موضوعات پیشنهادی باشد. همچنین با شناسایی خودکار مرز شروع بدنه اصلی پایان‌نامه، می‌توان با کنار گذاشتن پاراگراف‌های پیشگفتار، تنها بدنه اصلی پایان‌نامه/رساله از قبیل متن اصلی، تصاویر، جداول، منابع و سایر بخش‌ها را تحلیل کرد؛ بدون آن که بخش‌های پیشگفتار مورد پردازش قرار گیرد.

فهرست منابع و مآخذ

۱. تن سازان، ا. و مهدوی، م. (۱۳۹۶). استخراج فراداده‌های متنی از مقاله‌های علمی به زبان فارسی با مدل

آماری CRF. پژوهش‌های نظری و کاربردی در علم اطلاعات و دانش‌شناسی، ۳۲۱-۳۰۴.

2. Abiteboul, S. (1996). Querying semi-structured data. International Conference on Database Theory.
3. Adak, C., Chaudhuri, B. B., & Blumenstein, M. (2016). Named Entity Recognition from Unstructured Handwritten Document Images. 12th IAPR Workshop on Document Analysis Systems (DAS). Santorini, Greece: IEEE.
4. Adrian, W., Leone, N., Manna, M., & Marte, C. (2017). Document Layout Analysis for Semantic Information Extraction. In AI*IA 2017 Advances in Artificial Intelligence (pp. 269-281). Cham: Springer.
5. Ahmed, M., & Afzal, M. (2020). FLAG-PDFe: Features oriented metadata extraction framework for scientific publications. IEEE Access, 8, 99458-99469. doi:10.1109/ACCESS.2020.2997907
6. An, D., Gao, L., Jiang, Z., Liu, R., & Tang, Z. (2017). Citation Metadata Extraction via Deep Neural Network-Based Segment Sequence Labeling. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 1967-1970). New York, NY, USA: Association for Computing Machinery.
7. Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statist. Surv.
8. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.
9. Boukhers, Z., Ambhore, S., & Staab, S. (2019). An End-to-end Approach for Extracting and Segmenting High-Variance References from PDF Documents. ACM/IEEE Joint Conference on Digital Libraries (JCDL).
10. Chen, C.-C., Yang, K.-H., Kao, H.-Y., & Ho, J.-M. (2008). BibPro: A Citation Parser Based on Sequence Alignment Techniques. In 22nd International Conference on Advanced Information Networking and Applications - Workshops (aina workshops 2008) (pp. 1175-1180). doi:10.1109/WAINA.2008.125
11. Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning volume, 20, 273-297.
12. Cortez, E., da Silva, A. S., Gonçalves, M. A., Mesquita, F., & de Moura, E. S. (2009). A flexible approach for extracting metadata from bibliographic citations. Journal of the American Society for Information Science and Technology, 60, 1144-1158. doi:10.1002/asi.21049
13. Cortez, E., da Silva, A., Gonçalves, M., Mesquita, F., & de Moura, E. (2007). FLUX-CIM: Flexible Unsupervised Extraction of Citation Metadata. In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 215-224). Vancouver, BC, Canada: Association for Computing Machinery.
14. Councill, I., Giles, C. L., & Kan, M.-Y. (2008). ParsCit: an Open-source CRF Reference String Parsing Package. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA).
15. Cristani, M., Bertolaso, A., Scannapieco, S., & Tomazzoli, c. (2018). Future paradigms of automated processing of business documents. International Journal of Information Management, 67-75.

16. Cuong, N., Kumar, M., Kan, M.-Y., & Lee, W. (2015). Scholarly Document Information Extraction using Extensible Features for Efficient Higher Order Semi-CRFs. JCDL '15: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries.
17. Dashtipour, K., Gogate, M., Adeel, A., Algarafi, A., Howard, N., & Hussain, A. (2017). Persian Named Entity Recognition. IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC). Oxford, UK.
18. Day, M.-Y., Tsai, R. T.-H., Sung, C.-L., Hsieh, C.-C., Lee, C.-W., Wu, S.-H., . . . Hsu, W.-L. (2007). Reference metadata extraction using a hierarchical knowledge representation framework. *Decision Support Systems*, 43, 152 - 167.
19. Dejean, H. (2015). Extracting structured data from unstructured document with incomplete resources. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*. Tunis, Tunisia.
20. Dengel, A., & Shafait, F. (2014). Analysis of the Logical Layout of Documents. In *Handbook of Document Image Processing and Recognition* (pp. 177-222). London: springer.
21. Directions, M., & Advances, R. (2007). *Digital Document Processing*. London: Springer-Verlag.
22. Do, H., Chandrasekaran, M., Cho, P., & Kan, M.-Y. (2013). Extracting and Matching Authors and Affiliations in Scholarly Documents. JCDL '13: Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries.
23. Dudhabaware, R. S., & Madankar, M. S. (2014). Review on Natural Language Processing Tasks for Text Documents. 2014 IEEE International Conference on Computational Intelligence and Computing Research. IEEE.
24. Ferrés, D., Saggion, H., Ronzano, F., & Bravo, À. (2018). PDFdigest: an Adaptable Layout-Aware PDF-to-XML Textual Content Extractor for Scientific Articles.
25. Fisas, B., Saggion, H., & Ronzano, F. (2015). On the Discursive Structure of Computer Graphics Research Papers. doi:10.3115/v1/W15-1605
26. Flynn, P., Zhou, L., Maly, K., Zeil, S., & Zubair, M. (2007). Automated Template-Based Metadata Extraction Architecture. In *Proceedings of the 10th International Conference on Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers* (pp. 327–336). Berlin, Heidelberg: Springer-Verlag.
27. Giuffrida, G., Shek, E., & Yang, J. (2000). Knowledge-Based Metadata Extraction from PostScript Files. doi:10.1145/336597.336639
28. Groza, T., Handschuh, S., & Hulpus, I. (2009). A DOCUMENT ENGINEERING APPROACH TO AUTOMATIC EXTRACTION OF SHALLOW METADATA FROM SCIENTIFIC PUBLICATIONS.
29. Guo, Y., Korhonen, A., Liakata, M., Karolinska, I. S., Sun, L., & Stenius, U. (2010). Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing* (pp. 99–107). Uppsala, Sweden: Association for Computational Linguistics.
30. Guo, Z., & Jin, H. (2011). Reference metadata extraction from scientific papers. In *2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies* (pp. 45-49). doi:10.1109/PDCAT.2011.72
31. Gupta, S., & Manning, C. (2011). Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing.
32. Hafezi, L., & Rezaeian, M. (2018). Neural Architecture for Persian Named Entity. 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS). Tehran, Iran: IEEE.

33. Hanyurwimfura, D., Liao, B., Njogu, H., & Ndatinya, E. (2012). An automated Cue Word based Text Extraction. *Journal of Convergence Information Technology*, 421-429. doi:10.4156/jcit.vol7.issue10.50
34. Haralick. (1994). Document image understanding: geometric and logical layout. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Seattle, WA, USA, USA.
35. Hirohata, K., Okazaki, N., Ananiadou, S., & Ishizuka, M. (2008). Identifying Sections in Scientific Abstracts using Conditional Random Fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-{I}*.
36. Hossin, M., & Sulaiman, M. (2015). A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*.
37. Houngho, H., & Mercer, R. E. (2012). Method Mention Extraction from Scientific Research Papers. In *Proceedings of {COLING} 2012*. The COLING 2012 Organizing Committee.
38. Huang, Z., Jin, H., Yuan, P., & Han, Z. (2006). Header Metadata Extraction from Semi-structured Documents Using Template Matching. doi:10.1007/11915072_84
39. Jabbar, A., Iqbal, S., Khan, M. U., & Hussain, S. (2016). A survey on Urdu and Urdu like language stemmers and stemming techniques. *Artificial Intelligence Review*.
40. Kern, R., Jack, K., Hristakeva, M., & Granitzer, M. (2012). TeamBeam — Meta-Data Extraction from Scientific Literature. *D-Lib Magazine*.
41. Klink, S., Dengel, A., & Kieninger, T. (2000). Document Structure Analysis Based on Layout and Textual Features. *Proceedings of international workshop on document analysis systems*.
42. Kondo, T., Nanba, H., Takezawa, T., & Okumura, M. (2011). Technical Trend Analysis by Analyzing Research Papers' Titles. In *Human Language Technology. Challenges for Computer Science and Linguistics* (pp. 512-521). Berlin, Heidelberg: Springer Berlin Heidelberg.
43. Kooli, N., & Belaïd, A. (2016). Inexact graph matching for entity recognition in OCRed documents. *23rd International Conference on Pattern Recognition (ICPR)*. Cancun, Mexico: IEEE.
44. Kovačević, A., Konjović, Z., Milosavljević, B., & Nenadic, G. (2012). Mining methodologies from NLP publications: A case study in automatic terminology recognition. *Computer Speech & Language*, 105 - 126.
45. Lakhanpal, S., Gupta, A., & Agrawal, R. (2015). Towards extracting domains from research publications. *CEUR Workshop Proceedings*, 117-120.
46. Liakata, M., Teufel, S., Siddharthan, A., & Batchelor, C. (2010). Corpora for the Conceptualisation and Zoning of Scientific Papers. *Proceedings of LREC*.
47. Lin, J., Karakos, D., Demner-Fushman, D., & Khudanpur, S. (2006). Generative Content Models for Structural Analysis of Medical Abstracts. In *Proceedings of the HLT-NAACL Bio NLP Workshop on Linking Natural Language and Biology* (pp. 65-72). New York, New York: Association for Computational Linguistics.
48. Lin, S., Ng, J.-P., Pradhan, S., Shah, J., Pietrobon, R., & Kan, M.-Y. (2010). Extracting Formulaic and Free Text Clinical Research Articles Metadata Using Conditional Random Fields. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents* (pp. 90-95). Los Angeles, California: Association for Computational Linguistics.
49. Liu, B., & Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. In *Mining Text Data*. Springer.
50. Liu, R., Gao, L., An, D., Jiang, Z., & Tang, Z. (2017). Automatic Document Metadata Extraction Based on Deep Networks. *Natural Language Processing and Chinese Computing. NLPCC 2017*.

51. Liu, X., Gao, F., Zhang, Q., & Zhao, H. (2019). Graph Convolution for Multimodal Information Extraction from Visually Rich Documents. arXiv:1903.11279.
52. Luong, M., Nguyen, T., & Kan, M.-Y. (2010). Logical Structure Recovery in Scholarly Articles with Rich Document Features. *International Journal of Digital Library Systems*, 1(4).
53. Mao, S., Kim, J., & Thoma, G. (2004). A dynamic feature generation system for automated metadata extraction in preservation of digital materials. In *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.* (pp. 225-232). doi:10.1109/DIAL.2004.1263251
54. Marinai, S. (2009). Metadata Extraction from PDF Papers for Digital Library Ingest. In *2009 10th International Conference on Document Analysis and Recognition* (pp. 251-255). doi:10.1109/ICDAR.2009.232
55. Meng, B., Hou, L., Yang, E., & Li, J. (2018). Metadata Extraction for Scientific Papers. *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*.
56. Mohseni, M., Ghofrani, j., & Faili, H. (2018). Persianp: A Persian Text Processing Toolbox. In Gelbukh A. (eds) *Computational Linguistics and Intelligent Text Processing. CICLing 2016. Lecture Notes in Computer Science*. Springer, Cham.
57. Naderi, M., & Khamsehchi, E. (2017). Well placement optimization using metaheuristic bat algorithm. *Journal of Petroleum Science and Engineering*, 348 - 354.
58. Nasar, Z., Jaffry, S., & Malik, M. (2018). Information extraction from scientific articles: a survey. *Scientometrics*, 117(3), 1931–1990.
59. Ojokoh, B., Adewale, O., & Samuel oluwole, F. (2009). Automated document metadata extraction. *J. Information Science*, 35, 563-570. doi:10.1177/0165551509105195
60. Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A Survey of the Usages of Deep Learning for. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*.
61. Peanho, C., Stagni, H., & da Silva, F. (2012). Semantic information extraction from images of complex documents. *Applied Intelligence*, 37(4), 543–557.
62. Peng, F., & McCallum, A. (2006). Information extraction from research papers using conditional random fields. *Information Processing & Management*, 42(4), 963-979.
63. Piskorski, J., & Yangarber, R. (2013). Information Extraction: Past, Present and Future. In *Multi-source, Multilingual Information Extraction and Summarization* (pp. 23-50). Berlin: Springer.
64. Qiu, S., & Zhou, T. (2019). A Method of Extracting Metadata Information in Digital Books. *10th International Conference on Information Technology in Medicine and Education (ITME)*.
65. Rahal, N., Tounsi, M., Benjlaiel, M., & Alimi, A. (2018). Information Extraction from Arabic and Latin scanned invoices. *IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*.
66. Rahat, M., Talebpour, A., & monemian, S. (2018). A recursive algorithm for open information extraction from Persian texts. *International Journal of Computer Applications in Technology*, 57(3), 193-206. doi:10.1504/IJCAT.2018.092978
67. Rizvi, S., Dengel, A., & Ahmed, S. (2019). DeepBiRD: An Automatic Bibliographic Reference Detection Approach.
68. Ronzano, F., & Saggion, H. (2015). Dr. Inventor Framework: Extracting Structured Information from Scientific Publications. In *Discovery Science* (pp. 209-220). Cham: Springer International Publishing.
69. Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbühler, A., Fabry, P., Veuthey, A.-L. (2007). Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics*, 76, 195 - 200.

70. Safder, I., Hassan, S.-U., Visvizi, A., Noraset, T., Nawaz, R., & Tuarob, S. (2020). Deep Learning-based Extraction of Algorithmic Metadata in Full-Text Scholarly Documents. *Information Processing & Management*, 57(6).
71. Sarkhel, R., & Nadi, A. (2019). Visual Segmentation for Information Extraction from Heterogeneous Visually Rich Documents. *Proceedings of the 2019 International Conference on Management of Data*.
72. Souza, A., Moreira, V., & Heuser, C. (2014). ARCTIC: metadata extraction from scientific papers in pdf using two-layer CRF. *DocEng '14: Proceedings of the 2014 ACM symposium on Document engineering*.
73. Teufel, S., & Moens, M. (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Comput. Linguist.*, 28, 409–445. doi:10.1162/089120102762671936
74. Teufel, S., Siddharthan, A., & Batchelor, C. (2009). Towards Discipline-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3* (pp. 1493–1502). USA: Association for Computational Linguistics.
75. Tharwat, A., Hassanien, A., & Elnaghi, B. (2017). A BA-based algorithm for parameter optimization of Support Vector Machine. *Pattern Recognition Letters*, 93, 13-22.
76. Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P., & Bolikowski, Ł. (2015). CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJ DAR)*, 18, 317–335.
77. Toledo, J., Carbonell, M., Fornés, A., & Lladós, J. (2019). Information extraction from historical handwritten document images with a context-aware neural model. *Pattern Recognition*, 86, 27-36.
78. Wu, J.-C., Chang, Y.-C., Liou, H.-C., & Chang, J. S. (2006). Computational Analysis of Move Structures in Academic Abstracts. In *Proceedings of the {COLING}/{ACL} 2006 Interactive Presentation Sessions* (pp. 41-44). Sydney, Australia: Association for Computational Linguistics.
79. Yang, X.-S. (2010). A New Metaheuristic Bat-Inspired Algorithm. *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010)*, 284, 65-74.
80. Zhao, M., Fu, C., Ji, L., Tang, K., & Zhou, M. (2011). Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes. *Expert Systems with Applications*.

Abstract

Automatic metadata extraction is increasingly becoming a vital factor in modern digital libraries, and widely considered to be the most important in the process of indexing and information retrieval. Traditionally in the field of information extraction from academic literature, the focus has generally been on scientific articles, specifically English papers. This study seeks to address how to obtain header level metadata from Persian thesis documents. The aim is to extract metadata from Persian theses based on machine learning technique in two phases. The proposed method automatically detects contents of the front matter of Persian theses, and in the following, classifies paragraphs into predefined metadata. Additionally, novel features are introduced based on textual and structural properties, in order to tackle this issue. In an effort to carry out classification, we used Support Vector Machine (SVM) classifier in both two steps. The evaluation of the proposed technique is employed with Iranian theses and dissertations, collected from IranDoc Institution. The results reveal that the method has satisfactory performance and predicts the type of paragraphs with 98.1 percent accuracy.

Keywords: Information extraction, Metadata extraction, Scientific literature analysis, Content classification, Support Vector Machines (SVM), Persian (Farsi) text processing.



**Iranian Research Institute for Information Science and Technology
(Irandoct)**

Research Project

Machine learning based metadata extraction from Persian theses

By:

Jalal A. Nasiri

Co-authors:

Nasrollah Pakniat

April 2021