Journal of Emerging Defense Knowledge and Research Policy

Volume 1, Issue 1
April 2024
p.p 77-99

Shahid Sattari Air University

Designing an Intelligent Model for Detecting Plagiarism in Persian: A Structural and Semantic Recognition Approach

Ebrahim Nazari Farokhi[1], Mohammad Nazari Farokhi[2]

Abstract

Background & Purpose: The increasing of communication networks has made easy to access scientific literature of researchers. In the meantime, the profiteers are trying to copy the works and ideas of others to achieve their destinations, in the shortest time and without effort. In recent decades, this problem called plagiarism has spread to various scientific communities. The purpose of this study is to design an intelligent model for detecting plagiarism in Persian Language.

Methodology: The type of research is applied and method of research in the first step is to identify the dimensions of the model, exploratory, in second stage, comparative study of selected models and in the third stage, analytical, and in the fourth stage, due to the software testing of the model is experimental. The data collection tool is interviews with 10 experts. In the following, comparative study of selected models was performed and after analyzing the interview with experts, the final model was designed.

Findings: After testing model using the Robinson model, phrases and phrases that were similar to the original text were identified by the software and the terms were represented in four clusters and with four different colors.

Conclusion: The result is that the use of structural and semantic analysis methods, sequentially, can lead to the diagnosis of plagiarism and the prevention of repetitive content in Persian language.

Keywords: *Design, Intelligence model, Plagiarism, Structural recognition, Semantic recognition*

[1] Assistant Prof, Department of Management, Faculty of Management, Imam Ali Officers University, Tehran, Iran.

[2] Ph.D. Student in Information Technology Management, Faculty of Management and Economics, Science and Research Branch of the Islamic Azad University, Tehran, Iran.

Received: 03/09/2022
Accepted: 04/30/2022
Corresponding Author :Ebrahim Nazari Farokhi

فصلنامهٔ سیاستگذاری دانش و پژوهش‌های نوین دفاعی
دوره ۱، شماره ۱
اردیبهشت ۱۴۰۳
صص۷۷-۹۹

معاونت پژوهش و فناوری

طراحی مدل هوشمند کشف سرقت علمی در زبان فارسی: رویکرد تشخیص ساختاری و معنایی

ابراهیم نظری فرخی[۱]، محمد نظری فرخی[۲]

چکیده

زمینه و هدف: افزایش شبکه‌های ارتباطی دسترسی به آثار علمی-پژوهشی محققان را آسان ساخته است. در این میان افراد سودجو برای رسیدن به مقاصد خود در کوتاه‌ترین زمان و بدون تلاش، اقدام به کپی‌برداری از آثار و ایده‌های دیگران می‌نمایند. در دهه‌های اخیر این مشکل با نام سرقت علمی دامن‌گیر جوامع علمی مختلف شده است. هدف این پژوهش، طراحی مدل هوشمند کشف سرقت علمی در زبان فارسی است.

روش‌شناسی: نوع پژوهش، کاربردی و روش پژوهش در مرحله اول به علت شناخت ابعاد مدل، اکتشافی، در مرحله دوم، مطالعه تطبیقی مدل های منتخب و در مرحله سوم، تحلیلی و در مرحله چهارم به علت آزمون نرم‌افزاری مدل، تجربی است. ابزار گردآوری داده ها، مصاحبه با ۱۰ نفر از خبرگان است. در ادامه مطالعه تطبیقی مدل‌های منتخب انجام و پس از تحلیل مصاحبه با خبرگان، مدل نهایی، طراحی شد.

یافته‌ها: پس از آزمون مدل با استفاده از مدل رابینسون، عبارات و جملاتی که با متن مادر هم‌خوانی داشتند، توسط نرم‌افزار شناسایی شده و عبارات با توجه به تطبیق از نظر ساختاری، و معنایی در قالب چهار دسته و با چهار رنگ مختلف نمایش داده شدند.

نتیجه‌گیری: نتیجه آن که استفاده از روش‌های تحلیل ساختاری و معنایی، به صورت متوالی می‌تواند موجب تشخیص سرقت علمی و پیشگیری از تولید محتوای تکراری در زبان فارسی شده است.

واژه های کلیدی: *طراحی، مدل هوشمند، سرقت علمی، تشخیص ساختاری، تشخیص معنایی*

[۱] استادیار، دانشکده مدیریت، دانشگاه افسری امام علی(ع)، تهران، ایران.
[۲] دانشجوی دکتری مدیریت فناوری اطلاعات، دانشکده مدیریت، دانشگاه علوم و تحقیقات، تهران، ایران.

تاریخ دریافت مقاله: ۱۴۰۰/۱۲/۱۸
تاریخ پذیرش نهایی مقاله: ۱۴۰۱/۰۲/۱۰
نویسنده مسئول مقاله:ابراهیم نظری فرخی

مقدمه

سرقت علمی نه تنها مانع پیشرفت علمی شده بلکه افت کیفیت مراکز علمی و پژوهشی را نیز به دنبال داشته است. یکی از کارهایی که در کشورهای مختلف برای جلوگیری از اینگونه سرقت‌ها انجام شده، استفاده از سامانه‌های خودکار تشخیص سرقت علمی نظیر Turnitin ، Eve2، PLMC می‌باشد (لوکاشنکو[۱] و همکاران، ۲۰۰۷). سرقت در متون، به استفاده از آثار علمی-پژوهشی دیگران بدون ارجاع صحیح به آنها اشاره دارد (یوسف[۲] و همکاران، ۲۰۱۳). در حالت کلی، سرقت متون[۳] را می‌توان در محیط‌های بین زبانی[۴] یا محیط‌های تک‌زبانی[۵] بررسی کرد. در محیط بین زبانی، زبان منبع و مقصد متفاوت می‌باشند و روش‌های تشخیص سرقت در دو سطح، سطح اول دو زبان گرامر یکسانی دارند و سطح دوم دو زبان گرامر یکسانی ندارند، انجام می‌گیرند (بارون[۶] و همکاران، ۲۰۱۳). در محیط تک زبانی، زبان منبع و زبان مقصد همنوع بوده و تشخیص سرقت براساس ویژگی‌های متن به صورت لغوی، گرامری و معنایی انجام می‌شود (نواب[۷]، ۲۰۱۲)

روش‌های کلی کشف سرقت متون شامل تطبیق رشته[۸]، شباهت کلمات کلیدی[۹]، تحلیل اثر انگشت[۱۰] و تحلیل زبان[۱۱] می‌شوند. در روش تطبیق رشته، دو رشته (به طور مثال، دو جمله) با هم مقایسه می‌شوند. در روش شباهت کلمات کلیدی، کلمات کلیدی از اسناد استخراج می‌شوند. اگر دو سند از لحاظ کلمات کلیدی نزدیک به هم باشند، متون به بخش-های کوچک‌تر شکسته شده و به صورت بازگشتی روال فوق اجرا می‌شود. در روش تحلیل اثر انگشت، متون به بخش‌هایی به نام چونک[۱۲] تقسیم می‌شوند (بخش‌هایی که دارای همپوشانی هستند) سپس با اعمال توابع اثرانگشت، بخش‌های مشابه کشف می‌شوند. در روش تحلیل زبانی، متون با بهره‌گیری از قوانین موجود در زبان، از لحاظ ساختاری و معنایی باهم مقایسه می‌شوند(عبدی[۱۳] و همکاران، ۲۰۱۵).

---

[1] Lukashenko
[2] Yousuf
[3] Plagiarism
[4] Cross-Lingual
[5] Monolingual
[6] Barrón
[7] Nawab
[8] String Matching
[9] Keyword Similarity
[10] Fingerprint Analysis
[11] Linguistic Analysis
[12] Chunk
[13] Abdi

در محیط‌های تک‌زبانی، ابزارهای کشف سرقت متون به سه دسته اصلی تقسیم می-شوند(مائورر[1] و همکاران، ۲۰۰۶): دسته‌ی اول، بخشهای متفاوت یک سند را از لحاظ سبک نگارشی بررسی کرده و بخش‌هایی با سبک نگارشی متفاوت را به عنوان بخش‌های تقلبی شناسایی می‌کنند. دسته‌ی دوم، صفحات وب مشابه با سند مشکوک را، به عنوان مراجعی که کپی‌برداری از آن‌ها صورت گرفته است، پیدا می‌کند. دسته‌ی سوم، سند پرس‌وجو را با اسناد موجود در پایگاه داده مقایسه می‌کند.

اختلاف نظر در فهم متون، امری رایج و واقعیتی انکارناپذیر است. علاقه‌مندان به متون ادبی، همواره شاهد تفاوت نظر ادیبان در تفسیر اشعار شاعران و سخنوران بوده‌اند. وجود تفسیرهای متفاوت و متنوع از آیه‌های قرآن و اختلاف آراء متکلمان و فقیهان، شاهد گویایی بر وجود اختلاف نظر و تفاوت فهم در متون است (حسین زاده،۱۳۸۰). جکندوف بر این باور است که معنی جمله، از مجموع واژه ها تشکیل شده است (سعید، ۲۰۰۳). او، یکی از وظایف معناشناسی را بررسی رابطه میان جمله‌ها می‌داند و معتقد است که رابطه استلزام معنایی[2] میان دو جمله از رهگذر رابطه معنایی خاصی میان دو واژه در دو جمله مذکور پدید می‌آید که عامل ایجاد این رابطه معنایی خاص، وجود مولفه معنایی سبب[3] است (زعفرانلو و همکاران، ۱۳۹۰).

بافت زبانی بر محیط زبانی یک واحد زبان، یعنی بر روابط دستوری و معنایی این واحدها با دیگر واحدهای زنجیره گفتار یا متن، دلالت دارد. در بافت زبانی معنای یک واژه را می‌توان بر اساس محیط وقوع آن عنصر، تعیین کرد (بستانی و سپه‌وند، ۱۳۹۰). شناسایی شباهت متون یکی از شاخه‌های متن‌کاوی است که کاربرد آن در شناسایی سرقت علمی می‌باشد (یعقوبی و ختنلو، ۱۳۹۴). بسیاری از نویسنده‌ها، از روش‌های استخراج اطلاعات و پردازش زبان طبیعی، به منظور استخراج داده از متن استفاده می‌کنند. متن‌کاوی، کشف به وسیله اطلاعات ناشناخته قبلی و استخراج خودکار اطلاعات از منابع نوشته شده مختلف است (آقا کاردان و کیهانی نژاد، ۱۳۹۱).

فرآیند شناسایی اسناد نزدیک به تکراری می‌تواند با بررسی محتوای هر سند انجام شود، هنگامی دو سند دارای محتوای کاملاً یکسان باشند آن‌ها را به عنوان تکراری در نظر گرفته و اسنادی که در یک بخش کوچک غیرمشابه باشند به عنوان نزدیک به تکراری در نظر

1 Maurer
2 Semantic Entailment
3 Cause

گرفته می‌شوند. از جمله این موارد می‌توان اسناد با نسخه‌های متفاوت، اسنادی که آثار ادبی دیگران را سرقت کرده‌اند، اسناد با محتوای یکسان اما فونت‌های غیرمشابه یا نوع فایل مختلف مانند Pdf، Doc، اسناد با محتوای یکسان و سایت‌هایی که با نام‌ها و آدرس‌های متفاوتی در اختیار کاربران قرار می‌گیرند را نام برد (گودرزی و همکاران، ۱۳۹۵). برای پیش-گیری از سرقت علمی در تولید محتوای دانشی، روش‌های مختلفی وجود دارد که بعضی از آن‌ها سطحی هستند. یکی از این روش‌ها، روش تطبیق رشته‌ای متون است که اگر متون از لحاظ نحوی[1] و معنایی[2] تغییر قابل توجهی داشته باشند، این روش، عملکرد خوبی ندارد. برای شناسایی این تغییرات، نیازمند تکنیک‌های زبانی هستیم که با انجام تحلیل عمیق این کار انجام می‌شود.

زبان طبیعی که افراد برای بیان منظور خود استفاده می‌کنند، خواص ویژه‌ای دارد که از سودمندی نظام بازیابی اطلاعاتِ متنی می‌کاهد. این خواص عبارتند از اختلاف زبانی و ابهام. منظور از اختلاف زبانی، استفاده از واژه‌ها یا عبارت‌های مختلف برای انتقال یک ایده واحد می‌باشد. ابهام زبانی نیز زمانی رخ می‌دهد که واژه یا عبارت دارای بیش از یک تفسیر می‌باشد. در این پژوهش از طریق تجزیه ساختار و تحلیل معنایی، نسبت به بررسی شباهت متون اصلی و دستکاری شده به منظور پیشگیری از سرقت علمی بر اساس مدل هوشمند طراحی شده، اقدام می‌شودکه این خود بر اهمیت موضوع افزوده است.

با توجه به پژوهش های انجام شده، لزوم طراحی مدلی هوشمند که توان تحلیل ساختاری و معنایی جملات را داشته باشد، ضروری به نظر می رسد. بر این اساس در این پژوهش، نخست، روش‌های تشخیص شباهت ساختاری سرقت علمی، شناسایی شده است. در ادامه روش‌های تشخیص شباهت معنایی جهت طراحی مدل پیشنهادی، مطالعه گردیدند. مطالعه روش‌های پردازش زبان طبیعی[3] جهت بررسی شباهت معنایی محتواهای دانشی، از جمله سایر اهداف این پژوهش است. در گام بعدی، با انجام مطالعه تطبیقی مدل‌های مرتبط، ابعاد اصلی مدل پیشنهادی، استخراج گردید. در بخش چهارم، روش‌ها و در بخش پنجم به یافته‌ها پرداخته شده است. در انتها نیز، بحث و نتیجه‌گیری آورده شده است.

[1] Syntax
[2] Semantic
[3] NLP

پیشینه پژوهش

کیوان‌آرا و همکاران (۱۳۹۲) پژوهشی با عنوان «گونه‌شناسی تقلب‌ها و سرقت‌های علمی» انجام داده‌اند. نتایج پژوهش حاکی از این است که مرز بین سرقت علمی و صداقت در علم بسیار نامحسوس است و بایستی نهادی در کمیته‌های پژوهشی دانشگاه‌ها به منظور آموزش افراد در زمینه پرهیز از انواع تقلب و سرقت علمی وجود داشته باشد. داروئیان و فقیهی(۱۳۹۰) پژوهشی با عنوان «بررسی انگیزه‌ها و علل انجام سرقت علمی در ایران» انجام داده‌اند. نتیجه پژوهش این است که حفاظت و حراست از مالکیت فکری پژوهشگران و نویسندگان اهمیت بیشتری پیدا نموده و رعایت اصول اخلاقی و حفظ امانت‌داری استفاده از مطالب و پژوهش‌های نویسندگان و محققان ضرورت بیشتری دارد.

زمانی و همکاران (۱۳۹۲) پژوهشی با عنوان «شناسایی و اولویت‌بندی عوامل موثر بر سرقت علمی دانشجویان دانشگاه اصفهان» انجام داده‌اند. یافته‌های این پژوهش نشان می‌دهد که مدرک‌گرایی و توجه زیاد به نمره، اولین و مهمترین عامل تاثیرگذار بر سرقت علمی دانشجویان است؛ سایر عوامل به ترتیب اهمیت عبارتند از: نبود احساس خودکارآمدی در دانشجویان در انجام دادن پژوهش و نوشتن گزارش‌های علمی، نبود سازوکارهای مناسب برای تشخیص و تنبیه سارقان علمی، عوامل اجتماعی–فرهنگی، آموزش‌های ناکافی درباره چگونگی اسناددهی و تشخیص ندادن سرقت علمی دانشجویان از سوی استادان دانشگاه. شریفی‌راد و همکاران (۱۳۹۱) پژوهشی با عنوان«سوء رفتار علمی: سرقت علمی از خود» انجام داده‌اند. در این پژوهش اشاره شده است که: سرقت علمی، یکی از مباحث جدی جامعه علمی محسوب می‌شود و از نظر محققین امری نادرست و غیراخلاقی تلقی می‌شود؛ چرا که به منزله سرقت از کار فکری افراد است.

در انتهای دهه ۲۰۰۰ میلادی، برای تشخیص سرقت علمی متون، رویکردهای تجاری این سیستم‌ها مطرح شدند. لتروپ[1] و فس[2]، از طریق ارائه دو رویکرد به صورت آنلاین[3] و آفلاین[4] و با انجام مقایسه‌ی متون با پایگاه داده‌هایی که خودشان طراحی کرده بودند، اقدام به جستجوی شباهت متون کردند. رویکرد تشخیص سرقت علمی دیگری در سال ۲۰۰۲

[1] Lathrop
[2] Foss
[3] Online
[4] Offline

توسط فولام[۱] و پارک[۲] تبیین گردیده است. هر کلمه به ریشه اصلی خودش، معنا می‌شود. مقایسه شباهت در سطح جمله با استفاده از سنجه Cosine انجام می‌گردد. اگر امتیازِ شباهت یک زوج جمله بالاتر از سطح آستانه تعریف شده باشد، جمله به عنوان«سرقت‌شده» مارک می‌شود. در سال ۲۰۱۰، پژوهشی توسط تاستسارونیس[۳] و همکاران انجام شد. این پژوهش نشان داد که هرچند سنجه‌های آماری در روش‌های شناسایی سرقت علمی به کار رفته و پیاده‌سازی ساده‌تر و اثربخش‌تر در برابر سرقت علمیِ کلمه به کلمه دارند، اما این روش‌ها به تحلیل معنایی اطلاعات متنی و غیرمتنی کمک نمی‌کنند.

سیلوا[۴] و همکاران(۲۰۱۰) پژوهشی به انجام رسانده‌اند. نتایج این پژوهش نشان داد که جایگزینی واژگان با واژگان معنایی که از لحاظ معنا، مرتبط هستند و جایگزینی مترادف‌ها، مشخصه‌های اصلی هستند که یک مطالعه در خصوص بررسی «عبارت به عبارت» را پیشنهاد می‌دهد. پژوهشی در حوزه تشخیص سرقت علمی توسط چونگ[۵] (۲۰۱۳) انجام شد. نتایج نشان داد که برای تشخیص سرقت علمی، نیازمند رویکردهای محاسباتی به دارایی‌های فکری افراد هستیم و تحلیل زبانی عمیق باعث بهبود دسته‌بندی[۶] متن‌های سرقت شده می شود.

محمودی[۷] و ورنامخواستی(۲۰۱۴) برای کشف تقلب در متون فارسی روش دو مرحله‌ای پیشنهاد کرده‌اند. مرحله اول با نام پیش‌پردازش شامل نرمال‌سازی (حذف فاصله اضافی)، حذف کلمات ایست، تبدیل متن به جملات، توکن‌بندی، ریشه‌یابی، لم‌یابی، جایگزینی اعداد، شناسایی مترادف کلمات و برچسب‌گذاری کلمات است. پس از اتمام مرحله اول، دو توالی از کلمات موجود است. یک توالی برای متن اصلی و توالی دیگر برای متن پرس‌وجو. در مرحله دوم، این توالی‌ها با ترکیبی از روش‌های مشابهت‌یابی بایکدیگر مقایسه می‌شوند. پیکره‌ی مورد استفاده برای ارزیابی کار، توسط نویسنده تهیه شده و دقت روش پیشنهادی ۲۵٪ گزارش شده است.

[1] Fullam
[2] Park
[3] Tsatsaronis
[4] Silva
[5] Chong
[6] Classification
[7] mahmoodi

راکیان[1] و همکاران(۲۰۱۵) برای تشخیص تقلب، نمایش سطح جمله را استفاده کرده‌اند. روش ارائه شده‌ی آن‌ها شامل مراحل پیش پردازش، بازیابی اسناد کاندیدا و تشخیص تقلب است.

رفعیان[2] (۲۰۱۶) روش اثر انگشت را برای تشخیص تقلب بکار برده است. روش پیشنهادی وی برای نیل به هدف، مراحل زیر را طی می‌نماید:

پیش‌پردازش: توکن‌بندی متن، جداسازی جملات پاراگراف‌ها، تبدیل جملات به توکن، جایگزینی اعداد با #، نرمال‌سازی کلمات (حذف سه نقطه از متن، درج نیمفاصله بین وندها، حذف فاصله‌های اضافی)، حذف کلمات ایست، ریشه‌یابی توکن‌های جمع بدون علامت، جایگزینی مترادف‌ها، تعیین برچسب کلمات، ریشه‌یابی، حذف علائم نگارشی، لم‌یابی.

اثر انگشت: برای اعمال اثر انگشت، نمایش درختی سند استفاده شده است. ریشه درخت شامل سند بوده و فرزندان آن پاراگراف‌های سند و فرزندان هر یک از نودهای پاراگراف، جملات پاراگراف است و در نهایت هر یک از جملات به صورت 3-Gram نمایش داده شده-اند.

روش پژوهش

دراین پژوهش، نسبت به تشکیل پیکره متنی به منظور بررسی متون اصلی و دست‌کاری شده پرداخته شد. در مرحله بعدی، پیش‌پردازش متن[3] انجام می‌شود. در این مرحله، متن اصلی و متن دست‌کاری شده را با استفاده از تکنیک‌های پردازش زبان، آماده می‌شوند. حذف واژه-های عمومی و استخراج ویژگی‌های آماری و نحوی و ریخت‌شناسانه به عنوان مجموعه ویژگی‌های شخصی یا ترکیب شده مورد استفاده قرار می‌گیرند. شباهت متون از طریق تکنیک‌های نرمال‌سازی و برچسب‌زنی و روش‌های دیگر شامل تشکیل ماتریس متون[4]، وزن‌دهی به واژه‌ها و استفاده از مترادف و چندمعنایی واژه‌ها و همچنین بر اساس روش‌های جدیدی که در حین مطالعه ادبیات موضوع به دست خواهد آمد، بدست می‌آیند.

در ادامه از سایر ابزار پردازش زبان طبیعی از قبیل کتابخانه nhazm نیز استفاده شد. در نهایت با انجام مراحل بالا و بر اساس اخذ نظر خبرگان از طریق انجام مصاحبه، طراحی مدل هوشمند کشف سرقت علمی در زبان فارسی با رویکرد تشخیص ساختاری و معنایی صورت

[1] Rakian
[2] Rafieian
[3] Pre-Processing
[4] TDM

می‌پذیرد تا نسبت به پیشگیری از سرقت علمی در حوزه پژوهشگری اقدام گردد. نوع پژوهش، کاربردی است. روش پژوهش در مرحله اول به علت شناخت ابعاد مدل، اکتشافی، در مرحله دوم، مطالعه تطبیقی مدل‌های منتخب، در مرحله سوم، تحلیلی و در نهایت مدلی ارائه می‌شود که اعتبارسنجی آن از طریق طراحی سیستم نرم‌افزاری صورت می‌گیرد؛ بر این اساس، روش پژوهش در این مرحله، تجربی خواهد بود.

تجزیه و تحلیل یافته‌ها

گام اول: مطالعه کلیات و روش‌های به کار گرفته شده در طراحی مدل: با توجه به مطالعه تطبیقی صورت گرفته در مورد نرم‌افزارهای تشخیص سرقت علمی، شاخص‌های مشترک جهت اخذ نظر خبرگان به منظور طراحی مدل پژوهش، تنظیم گردید.

گام دوم: محدود نمودن معنا: پس از تشکیل پیکره متنی و فیلتر کردن حوزه بررسی از طریق باکس زمینه[1]، حذف واژه‌های عمومی و پیش‌پردازش صورت می گیرد. برای محدود کردن دامنه و حوزه تشخیص علوم و محتواهای تکراری در سایر مدل‌ها، باکس زمینه، در نظر گرفته شده است. به عنوان مثال کلمه fine دو معنا را به ذهن القاء می‌نماید؛ معنای خوب در علم مدیریت و معنای جریمه در علم حقوق را به ذهن می رساند. در مدل طراحی شده در این باکس علاوه بر در نظر گرفتن عنوان متن و یا کلمات کلیدی، از طریق فراوانی کلمات در متن، فیلتر صورت گرفته تا زمینه اصلی محتوای تولید شده در حیطه رشته‌ی مدنظر بررسی و تشخیص داده شود تا به اصطلاح، مدل، در دام مقایسه بی‌مورد نیفتد.

گام سوم: تشخیص شباهت ساختاری محتوا: هر واژه از چهار منظر قابل بررسی است. اول از لحاظ آواشناسی، دوم از نظر لغوی، سوم از بعد ساختاری و در نهایت از لحاظ معنایی. در مدل پیشنهادی، حذف واژه‌های عمومی و پیش‌پردازش انجام می‌شود تا از این طریق، مرحله دوم فیلتر پس از باکسِ زمینه، صورت گیرد. سپس ماتریس واژگان بر اساس فعل، فاعل، مفعول و قید تنظیم خواهد شد تا از این طریق به بررسی میزان شباهت ساختاری محتواهای تولید شده پرداخته شود.

گام چهارم: تشخیص میزان شباهت معنایی محتوا: به منظور تشخیص میزان شباهت معنایی محتوا، از ریشه یابی کلمات در زبان فارسی استفاده می‌شود. هدف از انجام این روش، جداسازی کلمات از متن و بازگردانیدن کلمات به ریشه اصلی تشکیل‌دهنده آن‌ها است.

[1] Context

تفاوت اصلی این روش با سایر پژوهش‌های انجام شده در زمینه ریشه‌یابی، قابلیت بازگرداندن کلمات به ریشه بدون از بین رفتن معنای آن‌ها در جمله است. یکی از جنبه‌های نوآوری در این مدل، بررسی شباهت ساختاری و معنایی جملات به صورت توامان است. پس از حذف واژه های اضافی، حروف ربط و بررسی میزان شباهت ساختاری، برگرداندن هر کلمه به ریشه اصلی خود، می تواند به عنوان راهکاری جهت تشخیص میزان شباهت معنایی محتوا در مدل طراحی شده، موثر واقع شود. برای این کار در مدل پیشنهادی، نخست پایگاه داده‌ای از کلمات با مترادف‌هایشان تشکیل داده و سپس هر کلمه در داخل متن را جستجو نموده و به معنای اصلی‌اش که در لغت‌نامه تعریف شده است، تبدیل می نماید (مثل کلمه نظارت، مانیتور و نمایشگر). وقتی متن با سندهای پایگاه داده مقایسه می‌شود، کافی است هر جا که درجه شباهت بالا رفت، در پایگاه داده، ذخیره شود که در کدام سند این اتفاق رخ داده، سپس با مقایسه درجه شباهت‌ها می‌توان پی برد که متن تکراری بیشتر از چه منبعی استفاده کرده است.

مطالعه تطبیقی

برای طراحی مدل، نیاز به انجام مطالعه تطبیقی است. بر این اساس، ماتریس تطبیقی تدوین گردید که این ماتریس، ابعاد اصلی مدل ها و نرم افزارهای منتخبِ حاصل از مطالعه تطبیقی است. در ماتریس تطبیقی، شاخص های اصلی مدل، شناسایی می شود که در گام بعدی این شاخص ها با خبرگان مطرح شد. نتیجه حاصل از انجام مصاحبه خبرگی، منجر به استخراج شاخص های مناسب جهت طراحی مدل در این پژوهش گردید. در نهایت بر اساس شاخص های منتخب توسط خبرگان، ابعاد کلی مدل، تعیین شد. در جدول‌های (۱) و (۲)، به ترتیب ماتریس تطبیقی و مطالعه تطبیقی حاصل از پژوهش ها و مدل های مرتبط، آورده شده است.

جدول ۱. ماتریس تطبیقی نرم‌افزارهای تشخیص شباهت متون در زبان فارسی

ردیف	عنوان مقاله و سال چاپ	نام نویسندگان	روش / تکنیک
۱	به سوی کشف استفاده مجدد کدمنبع چندزبانه/ ۲۰۱۱ میلادی	Enrique Flores Alberto Barron Code no Paolo Rosso Lidia Moreno	روش: مدل مقایسه ویژگی ها N-Grams روش به کار رفته در: کامنت ها و لغات ذخیره شده

۲	کشف کد منبع استفاده مجدد بیت زبان های برنامه نویسی چندگانه/ ۲۰۱۲ میلادی	Enrique Flores Alberto Barron Code no Paolo Rosso Lidia Moreno	روش: پردازش زبان طبیعی ابزارها: مقایسه کد منبع در حد توابع و روش
۳	کشف سرقت نرم افزاری با رویکردی مبتنی بر گراف/۲۰۱۳ میلادی	Dong-Kuy Chae giwoon Ha, sang-Wook Kim, Boo Joong Kang, Elu Gyulm	برچسب API و گراف جریان کنترل (A-CFG)
۴	تجربه کشف تقسیم کدها در مایکروسافت/۲۰۱۱ میلادی	Yin Gong Dang Song Ge, Ray Huang and Dong mei Zhang	ابزارها: ابزارهای XIAO توسط پژوهشگران مایکروسافت ساخته شد و برای آزمون به اشتراک گذاشته شد
۵	موتور کشف سرقت علمی برای کدمنبع جاوا: یک رویکرد مبتنی بر خوشه بندی/ ۲۰۰۷ میلادی	Ameera Jadalla Ashraf Elnagar	تکنیک: داده کاوی خوشه بندی توکنیزاسیون N-Gram
۶	سرقت علمی در تکالیف برنامه نویسی/ ۱۹۹۹ میلادی	Mike Joy Michael Luck	تکنیک: محاسبه و مقایسه مقادیر ویژگی و مقایسه برنامه ها بر اساس ساختارشان
۷	کشف سرقت علمی با استفاده از ابزارهای نرم افزاری: مطالعه ای در علوم کامپیوتر	Bugarin Carreria, M Lama, X.M Pardo	ابزار: ترنیتن و Jplag
۸	اطلاعات به اشتراک گذاشته شده و برنامه کشف سرقت/ ۲۰۰۴ میلادی	Xian Chen Brent Francia Ming Li, Brian Mc Kinnon Amiti Seker	روش: پیچیدگی کولموگروف
۹	سیستم مشابهت کدمنبع برای کشف سرقت علمی	Zoran Djuric Dragan Gasevic	روش: فرآیند توکنیزاسیون

			زبان ها: کدمنبع جاوا
۱۰	مطالعه منحصر بفرد نبودن کد منبع/ ۲۰۱۰ میلادی	Mark Gabel Zhen dong Su	روش: توکنیزاسیون تحلیل لغوی جدول بندی زبان ها: جاوا، C و C++
۱۱	کشف کد منبع استفاده شده در زبان های برنامه نویسی چندگانه	Enrique Flores Alberto Barron Code no Paolo Rosso Lidia Moreno	روش: ویژگی N-Gram زبان: به کار گرفته شده در پایتون، جاوا و C++
۱۲	رویکردی به کشف سرقت کد منبع و بررسی با به کارگیری LSA	Georgian cosma Mike Joy	روش: LSA تحلیل معنایی پنهان ابزارها: PlaGate ابزارهای نوآورانه
۱۳	تحلیل لغوی برای ارزیابی دوتایی مفهومی بین برنامه C/ ۲۰۱۳	Akhil Gupta Dr.sukhvir Singh	روش: تحلیل لغوی
۱۴	ابزاری بر اساس یادگیری ماشین برای کشف سرقت علمی کد منبع/ ۲۰۱۱	Upul Bandara Gamini Wijayarathna	روش: تکنیک یادگیری ماشین
۱۵	سیستم کشف سرقت علمی کد زبان C	N. Haritha M. Bhavani K. Thammi Reddy	ابزار: توکنیزاسیون روش اثربخشی مشابهت JACCARD تکنیک N-Gram
۱۶	شناسایی سرقت علمی کد منبع اتربخشی بر اساس قطعه رشته های حجیم	Tabassam Nawaz Sami ud Din Ali Javed	روش: قطعه رشته حجیم

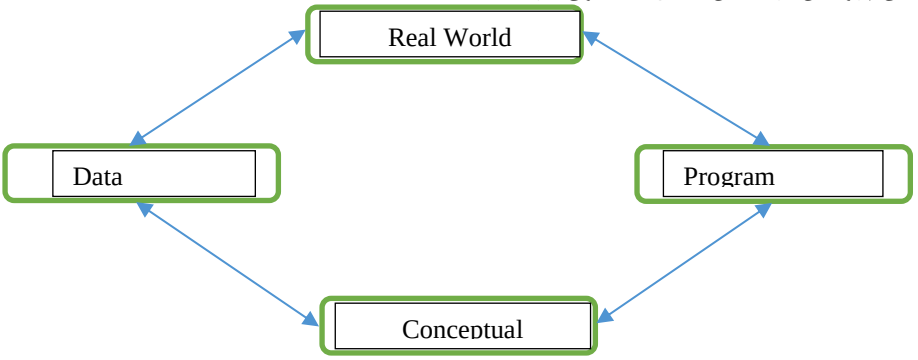
جدول ۲. خلاصه‌ای از مطالعه تطبیقی پژوهش‌های انجام شده متفاوت با ویژگی‌ها و وظایف مربوطه (Gondaliya et.al. 2014)

شاخص-ها/پارادایم	مدل SIM	مدلJPlag	مدل PMDSCPD	مدل YAP3	مدل Sherlock	مدل MOSS
نوع اجرا	کنسول	اجرای وب، جاوا	اجرای وب، جاوا	---	اجرای جاوا	سرویس وب، پرل اسکریپت
زبان پشتیبانی	C, Java, pascal,Lisp, Miranda and Natural language	Java, c#, c, c++, Scheme and Natural Language	Java, JSP, c, c++, Fortran and PHP	Pascal, c, LISP	Programming Language and Natural language	C, c++, Java, Javascript, Pascal, Ada, Lisp, Python, C#, Perl
روش ها/ الگوریتم ها	توکنیزاسیون	توکنیزاسیون + کنارهم گذاری رشته ای گریدی	توکنیزاسیون	توکنیزاسیون + کنار هم گذاری رشته ای گریدی	مقایسه افزایشی دو فایل	الگوریتم وینوونگ
GUI	خیر	بله	بله	خیر	بله	خیر
روش اندازه گیری شباهت	جز در درصد، تعداد ردیف های مشابه	جز در درصد، هیستوگرام، گروه فایل های مشابه	تعداد ردیف های مشابه، نشان ها	---	جمع درصدها، گراف شباهت	جز در درصد،
آنلاین/ آفلاین	آفلاین	آنلاین	آفلاین	آفلاین	آفلاین	آنلاین

شاخص-ها/پارادایم	مدل SIM	مدل JPlag	مدل PMDSCPD	مدل YAP3	مدل Sherlock	مدل MOSS
نوع اجرا	کنسول	اجرای وب، جاوا	اجرای وب، جاوا	---	اجرای جاوا	سرویس وب، پرل اسکریپت
زبان پشتیبانی	C, Java, pascal,Lisp, Miranda and Natural language	Java, c#, c, c++, Scheme and Natural Language	Java, JSP, c, c++, Fortran and PHP	Pascal, c, LISP	Programming Language and Natural language	C, c++, Java, Javascript, Pascal, Ada, Lisp, Python, C#, Perl
روش ها/ الگوریتم ها	توکنیزاسیون	توکنیزاسیون + کنارهم گذاری رشته ای گریدی	توکنیزاسیون	توکنیزاسیون + کنار هم گذاری رشته ای گریدی	مقایسه افزایشی دو فایل	الگوریتم وینوونگ
GUI	خیر	بله	بله	خیر	بله	خیر
روش اندازه گیری شباهت	جز در درصد، تعداد ردیف های مشابه	جز در درصد، هیستوگرام، گروه فایل های مشابه	تعداد ردیف های مشابه، نشان ها	---	جمع درصدها، گراف شباهت	جز در درصد،
آنلاین/ آفلاین	آفلاین	آنلاین	آفلاین	آفلاین	آفلاین	آنلاین

آزمون مدل

آزمون مدل بر اساس مدل رابینسون انجام شد. بر این اساس و مطابق شکل ۱ چهار بعد اصلی در خصوص آزمون نرم‌افزار مورد بررسی قرار گرفت که در ادامه، مدل طراحی شده در این پژوهش بر اساس شکل ۱ آزمون گردید.



شکل ۱. آزمون مدل

مرحله اول: اعتبارسنجی داده[1]: شناخت داده: مرحله اول، رویه انتخاب داده‌ها برای کسب مناسب‌ترین داده‌های مورد نیاز است. یکی از مشکلاتی که در این زمینه وجود دارد، این است که ممکن است داده‌ها دقیق نباشند و همچنین داده‌ها، ناکامل بوده و برخی مقادیر موجود نباشند. این مرحله، قبل از مرحله پیش‌پردازش صورت می‌گیرد؛ زیرا در مجموعه داده جمع-آوری شده برای پردازش غالبا هر داده با یک ویژگی خاص در نظر گرفته می‌شود تا از بروز خطا در داده‌های ورودی به سیستم اجتناب شود. این مرحله از آزمون مدل، شامل اطمینان از داده‌هاست که پاک‌سازی داده‌ها صورت گرفته باشد تا از کیفیت داده‌ها اطمینان حاصل شود. در این مرحله، فرآیند حذف و اصلاح داده‌های نامطمئن صورت می‌گیرد.

مرحله دوم: اعتبارسنجی مفهومی[2]: نحوه استفاده و به کارگیری روش‌های انجام کار در مدل طراحی شده مورد قضاوت خبرگان قرار گرفت که نتیجه در همین فصل ارائه گردیده است. با تایید خبرگان، برای تشخیص شباهت ساختاری در مدل طراحی شده، بررسی ساختار جمله از نظر فاعل، فعل، مفعول و قید کفایت می نماید. همچنین برای تشخیص شباهت

[1] Data validation
[2] Conceptual Validation

معنایی، استفاده از ریشه‌یابی کلمه‌ها و بردن کلمات به ریشه اصلی از جمله راه‌کارهایی است که مورد تایید خبرگان قرار گرفت. در ادامه به منظور تشخیص فرکانس کلمات و پی بردن به زمینه اصلی محتوای مورد بررسی، الگوریتم I-Match مورد تایید خبرگان قرار گرفت. به منظور تشخیصِ شباهت ساختاری و معنایی از بین الگوریتم‌های متعدد، الگوریتم مدل فضای برداری[1] مورد تایید خبرگان قرار گرفت و بر این اساس مدل نهایی پژوهش از طریق نرم‌افزار طراحی شده، مورد آزمون قرار گرفت.

اعتبارسنجی برنامه[2]: این بخش از آزمون با مرحله اول و سوم آزمون یعنی اعتبارسنجی داده و اعتبارسنجی مفهومی، ارتباط مستقیم و دوسویه‌ای دارد. به این معنا که اگر داده‌ها در مرحله ورود به درستی شناسایی، انتخاب و پاک‌سازی نشوند، باعث بُروز خطا در محاسبات الگوریتم‌های برنامه خواهد شد. تایید منطقِ برنامه، خطایابی، اجرای الگوریتم و بررسی معیارهای صحت و دقت الگوریتم از مراحل این بخش می‌باشند.

اعتبارسنجی در دنیای واقعی[3]: یکی از مشکلات مدل‌سازی، مواجهه شدن مدل در عمل و با دنیای واقعی است. به عبارتی دیگر، مدل طراحی شده زمانی معتبر است که در یک نمای واقعی، جواب بدهد. برای این منظور در آزمون مدل طراحی شده در این پژوهش، سه سناریوی متفاوت به شرح زیر در نظر گرفته شد تا از اعتبار مدل در دنیای واقعی جهت بررسی شباهت ساختاری و معنایی اطمینان حاصل نمود:

الف- تطبیق و پردازش داده‌های متنی کاملا مشابه

ب- تطبیق و پردازش داده‌های متنی کاملا مخالف

ج- تطبیق و پردازش داده‌های متنی مرتبط در یک حوزه علمی خاص.

تشخیص زمینه (بافت) متن مورد بررسی

برای محدود کردن دامنه و حوزه تشخیص علوم و محتواهای تکراری در سایر مدل‌ها، در مدل طراحی شده در این پژوهش، باکسِ زمینه در نظر گرفته شده است. در این بخش از مدل، بر اساس عنوان متن و یا کلمات کلیدی، فیلتر صورت گرفته تا زمینه اصلی محتوای

[1] Vector Space Model
[2] Program validation
[3] Real-World Validation

تولید شده در حیطه علم مدنظر بررسی شود و به اصطلاح، مدل در دام مقایسه بی مورد نیفتد. همچنین از طریق فراوانی کلمات پرتکرار که معمولا از بین ضمیرها، اسم‌ها، قیدها و صفات هستند، حوزه و زمینه سندِ مورد بررسی، تشخیص داده می‌شود.

پس از اینکه تمامی کارهای نرمال‌سازی، تجزیه و تحلیل جمله‌های فایل‌های متن اصلی (پایان‌نامه) و فایل مورد بررسی (مقاله استخراج شده) و برچسب‌زنی و ریشه‌یابی انجام شد، نتیجه کار در آرایه‌هایی ذخیره شده و برای فرم بعدی ارسال می‌شود. در نرم‌افزار طراحی شده برای این منظور، در قسمت Keyword کلمات پرتکرار که معمولا از بین ضمیرها، اسم‌ها، قیدها و صفات تشخیص داده می‌شود، ۸ کلمه اول را انتخاب و در قالب Checkboxهایی نمایش داده می‌شود. حال کاربر بایستی از بین کلمات مورد نظرِ خود، انتخاب انجام دهد و از طریق دکمه Find Subject اگر موضوعی مرتبط با موضوع خود، در بانک اطلاعاتی موجود باشد را نشان دهد، در غیر این صورت، اگر موضوعی منطبق با آن پیدا نشد، بایستی مشخصات موضوع خود را وارد نماید تا در بانک ثبت گردیده و در مراحل بعدی جستجو، سیستم آن را به حافظه سپرده و به نوعی هوشمندی حاصل شود.

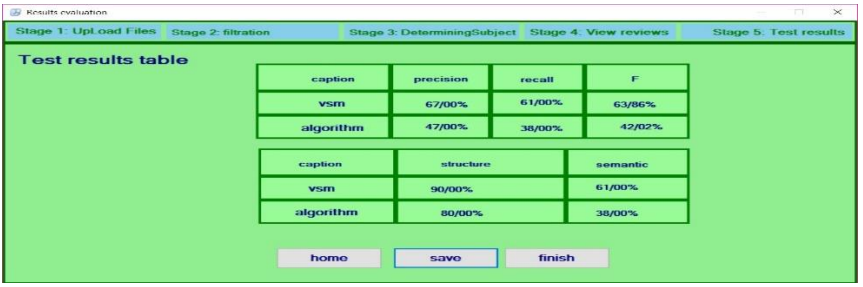
در گام بعدی، متن پردازش‌شده، نمایش داده می‌شود. نمایش متن پردازش‌شده حاصل از مراحل قبلی تست مدل به صورت زیر است:

جملاتی که به رنگ قرمز هستند؛ یعنی از لحاظ ساختاری با جملاتی در پایان‌نامه (متن اصلی) تطبیق دارند. اگر به رنگ مشکی باشد، یعنی نه از لحاظ ساختاری و نه از لحاظ معنایی، تطبیقی وجود ندارد. اگر جملات با رنگ سبز مشخص شد، یعنی از لحاظ معنایی تطبیق دارند و اگر جمله مورد نظر با رنگ آبی مشخص شده باشد، یعنی هم از لحاظ ساختاری و هم از لحاظ معنایی با جمله‌ای در پایان‌نامه (متن اصلی) تطابق وجود دارد.

جدول ۴. نمایش متن پردازش‌شده حاصل از مراحل قبلی تست مدل

رنگ جمله	نوع تشخیص
قرمز	تطبیق ساختاری وجود دارد.
مشکی	نه ساختاری و نه معنایی تطبیق وجود ندارد.
سبز	تطبیق معنایی وجود دارد.
آبی	هم ساختاری و هم معنایی تطبیق وجود دارد.

شکل ۲. خروجی حاصل از پردازش متن در نرم‌افزار



شکل ۳. نتایج دو سند با الگوریتم‌های پیشنهادی و مدل فضای برداری بر اساس معیارهای پریسیژن، ریکال و معیار F

نتیجه‌گیری

یکی از جنبه‌های نوآوری در این مدل، بررسی هر دو جنبه شباهت ساختاری و معنایی جملات است که در سایر پژوهش های مرتبط در حوزه سرقت علمی در ایران، انجام نشده است. بررسی میزان شباهت ساختاری، از طریق برگرداندن هر کلمه به ریشه اصلی خود، می تواند به عنوان راهکاری جهت تشخیص میزان شباهت معنایی محتوا در مدل طراحی شده، موثر واقع شود که این با نتیجه مطالعه فولام و پارک(۲۰۰۲) هم راستا می باشد. مغایرت دیگری که این پژوهش با سایر پژوهش های انجام شده دارد، استفاده از باکس کانتکس است تا از این طریق بافت زبانی هر پژوهش جهت بررسی میزان شباهت، مشخص گردد و مدل در دام مقایسه بی مورد نیفتد. استفاده از روش های تحلیل معنایی در این پژوهش منطبق با پژوهش سوسا و همکاران(۲۰۱۰) است که در آن مطالعه نیز برای بررسی سرقت علمی، جایگزینی واژگان با واژگان معنایی، انجام شده است.

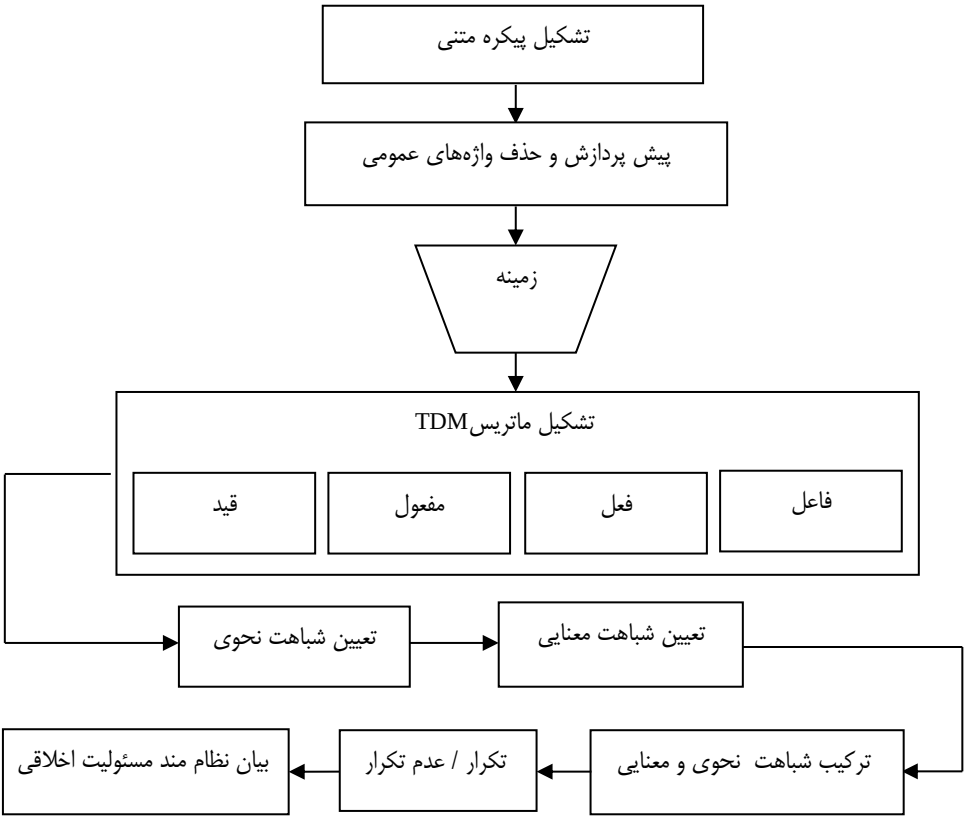
نتیجه بررسی نرم‌افزاری شاخص فعل در زبان فارسی: معیار بررسی فعل در تشخیص شباهت ساختاری و معنایی، معیار مناسبی نیست؛ چون بعد از این که ریشه‌یابی افعال انجام می‌شود، ریشه اکثر افعال مشابه است و این خود درصد خطا را بالا می‌برد؛ در صورتی که در تمامی مقاله‌ها و کتاب‌ها از افعال یکسانی در زبان فارسی استفاده می‌شود.

تعریف سرقت علمی حاصل از انجام پژوهش حاضر: مبتنی بر مدل ارائه شده و آزمون مدل بر اساس نرم‌افزار طراحی شده، سرقت علمی یعنی جمله‌های یک متن از یک سند با جمله‌های موجود در سندی دیگر، بدون ارجاع‌دهی و رعایت اخلاق پژوهشی به لحاظ ساختاری/معنایی تطبیق داشته و یا اینکه این جمله‌ها، هم به صورت ساختاری و هم معنایی با هم تطبیق داشته باشند. نتیجه حاصل از مصاحبه با خبرگان، به شرح جدول ۵ است تا در مدل طراحی شده، از طریق هر کدام از این شاخص های احصا شده، به منظور تشخیص شباهت ساختاری و معنایی اقدام شود.

جدول۵. نتیجه مصاحبه با خبرگان

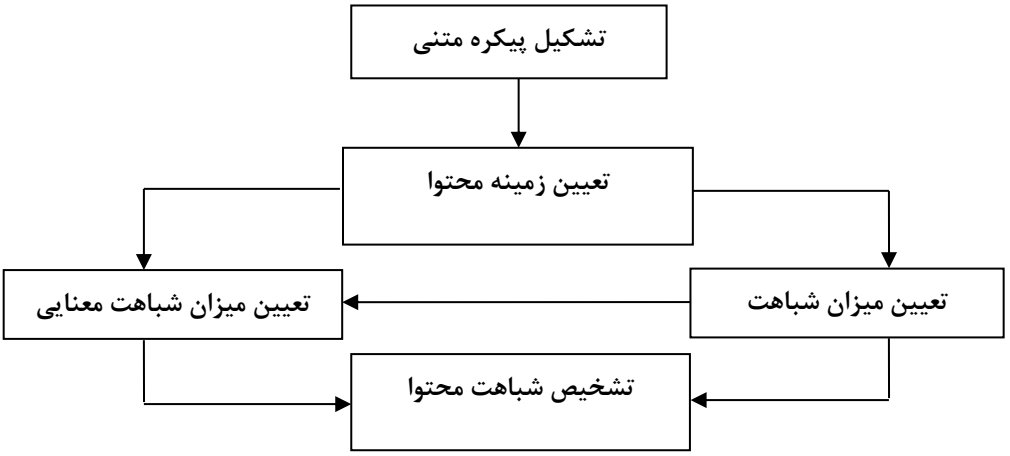
نوع اجرا	زبان پشتیبانی	روش ها/ الگوریتم ها	GUI	بررسی نتیجه	روش اندازه گیری شباهت	آنلاین/ آفلاین
تحت وب	C یا C#	توکنیزاسیون	بله	شباهت ساختاری و معنایی	جز در درصد، تعداد ردیف های مشابه	بله

پس از تشکیل پیکره متنی و فیلتر کردن حوزه بررسی از طریق باکس کانتکس، حذف واژه‌های عمومی و پیش‌پردازش صورت می گیرد. در ادامه، شباهت اسناد بر اساس نام نویسنده، سازمان و مکان انجام پژوهش، بررسی می‌شود. سپس ماتریس واژگان بر اساس فعل، فاعل، مفعول، صفت و قید تنظیم خواهد شد تا از این طریق به بررسی میزان شباهت ساختاری محتواهای تولید شده پرداخته شود.



شکل ۴. مدل استخراج شده اولیه

بر اساس مدل استخراج شده بالا، مصاحبه مجدد با خبرگان به عمل آمد و ابعاد اصلی مدل نهایی پژوهش، به شرح شکل ۵ استخراج و مدل نهایی پژوهش، طراحی گردید. پیشنهاد می شود که نتیجه بررسی شباهت ساختاری با نتیجه بدست آمده از طریق بررسی میزان شباهت معنایی، در قالب طراحی مدلی، مورد مطالعه قرار گیرد.



شکل۵. مدل نهایی پژوهش

منابع

آقاکاردان، احمد.؛ کیهانی نژاد، سینا. (۱۳۹۱). ارائه مدلی برای استخراج اطلاعات از مستندات متنی, مبتنی بر متن کاوی در حوزه یادگیری الکترونیکی. فناوری اطلاعات و ارتباطات ایران, ۴(۱۱-۱۲), ۴۷-۵۴.

بستانی، قاسم.؛ سپه‌وند، وحید. (۱۳۹۰). کاربرد روش های معناشناسی نوین در پژوهش های قرآنی. مطالعات قرآن و حدیث, ۵(۱ (پیاپی ۹), ۱۶۵-۱۸۶.

بهمن آبادی، سمیه.؛ جعفرآبادی کلاته، طاهره، و بختیار شعبانی ورکی. (۱۳۹۳). رعایت اخلاق پژوهش، مجله راهبرد فرهنگ.(۲۵):۱۵۲-۱۲۹.

حسین زاده.؛ محمد. (۱۳۹۵). مبانی معرفت دینی، انتشارات موسسه آموزشی و پژوهشی امام خمینی (ره).(۳).

داروئیان، سهیلا.؛ مهدی فقیهی. (۱۳۹۰). بررسی انگیزه‌ها و علل انجام سرقت علمی در ایران، فصلنامه رسالت مدیریت دولتی.۲(۱) :۱۵۴-۱۳۷.

عبدالکریمی، سپیده.؛ زعفرانلو کامبوزیا، عالیه کرد، آقا گل‌زاده، فردوس، و گلفام، ارسلان. (۱۳۹۰). مجهول-سازی افعال مرکب فارسی از منظر معنایی و نظریه معنی‌شناسی مفهومی. فصلنامه پژوهش‌های زبان و ادبیات تطبیقی.۲(۲).

زمانی، بی‌بی عشرت.؛ عظیمی، سید امین، و نسیم سلیمانی. (۱۳۹۲). شناسایی و اولویت‌بندی عوامل موثر بر سرقت علمی دانشجویان دانشگاه اصفهان. فصلنامه پژوهش و برنامه ریزی در آموزش عالی. ۱۹(۱).

شریفی‌راد، غلامرضا.؛ شهنازی، حسین، کامران، عزیز، و عباسی، محمدهادی. (۱۳۹۱). سوء رفتار علمی، سرقت علمی از خود. مجله تحقیقات نظام سلامت. ۸(۶): ۹۲۸-۹۲۲.

گودرزی، الهام.؛ صالح پور، نرگس، نظری فرخی، محمد، و نظری فرخی، ابراهیم. (۱۳۹۵). ارائه راهکاری جهت تشخیص اسناد نزدیک به تکراری.

ودادهیر، ابوعلی.؛ فرهود، داریوش، طباطبایی، قاضی، محمود، و توسلی، غلامعباس. (۱۳۸۷). معیارهای رفتار اخلاقی در انجام کار علمی، تاملی بر جامعه شناسیِ اخلاق در علم–فناوری مِرتن و رِزنیک، فصلنامه اخلاق در علوم و فناوری. ۳(۴و۳).

یعقوبی، رضوان.؛ و حسن ختنلو. (۱۳۹۴). شناسایی سرقت ادبی مبتنی بر الگوریتم ژنتیک و برچسب گذاری نقش معنایی در مقالات علمی.

Abdi, A., Idris, N., Alguliyev, R. M., & Aliguliyev, R. M. (2015). PDLK: Plagiarism detection using linguistic knowledge. Expert Systems with Applications, 42(22), 8936-8946.

Barrón-Cedeño, A., Gupta, P., & Rosso, P. (2013). Methods for cross-language plagiarism detection. Knowledge-Based Systems, 50, 211-217.

Chong, Man Yan Miranda (2013). 'A Study on Plagiarism Detection a Plagiarism Direction Identification Using Natural Language Processing Techniques'.

G. Tsatsaronis, G., Varlamis, I., Giannakoulopoulos, A., & Kanellopoulos, N. (2010). Identifying free text plagiarism based on semantic similarity. In Proceedings of the 4th International Plagiarism Conference.

K.Fullam, and J. Park. (2002). Improvements for scalable and accurate plagiarism detection in digital documen ts. http://www.lips.utexas.edu/~ kfullam/pdf/DataMiningReport. Pdf.

Lathrop, Ann and Kathleen Foss. (2000). 'The Student cheatin plagiarism in the Internet era: a wake - up call'. Englewood, CO: Libraries Unlimited.

Lukashenko, R., Graudina, V., & Grundspenkis, J. (2007, June). Computer-based plagiarism detection methods and tools: an overview. In Proceedings of the 2007 international conference on Computer systems and technologies (pp. 1-6).

Mahmoodi, M., & Varnamkhasti, M. M. (2014). Design a Persian automated plagiarism detector (AMZPPD). arXiv preprint arXiv:1403.1618.

Maurer, H. A., Kappe, F., & Zaka, B. (2006). Plagiarism-A survey. J. Univers. Comput. Sci., 12(8), 1050-1084.

Nawab, R. M. A. (2012). Mono-lingual paraphrased text reuse and plagiarism detection (Doctoral dissertation, University of Sheffield).

Rafieian, S., & Baraani Dastjerdi, A. (2016). Plagiarism checker for Persian (PCP) texts using hash-based tree representative fingerprinting. Journal of AI and Data Mining, 4(2), 125-133.

Rakian, S., Safi, E. F., & Rastegari, H. (2015). A Persian fuzzy plagiarism detection approach.

Saeed, John I. (2003). 'Semantics, 2nd. edn. Blackwell, Oxford, UK.

Sousa Silva, R., Grant, T., & Maia, B. (2010). I didn't mean to steal someone else's words!: a forensic linguistic approach to detecting intentional plagiarism. In *4th International Plagiarism Conference*.

Yousuf, S., Ahmad, M., & Nasrullah, S. (2013, October). A review of plagiarism detection based on Lexical and Semantic Approach. In 2013 International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA) (pp. 1-5). IEEE.