Journal of Business Intelligence Management Studies
Vol. 9, No.36, Summer 2021
ims.atu.ac.ir
DOI: 10.22054/IMS.2021.49415.1661

ATU PRESS

ISSN: 1735/112X eISSN: 2538-2527 Received: 2020/01/22 Revised: 2020/09/03 Accepted: 2021/05/02 Original Research

A New Approach to Plagiarism Detection Using Cellular Learning Automatons and Semantic Role Labeling

Rezvan Yaghobi* MSc., Computer Science, Islamic Azad University, Malayer Branch, Young and Elite Researchers Club, Malayer, Iran.

Mahdi Yaghobi MSc., Information Technology, Islamic Azad University, Malayer Branch, Iran.

Hassan Khotanloue Professor, Department of Computer Science, Bu Ali Sina University, Hamadan, Iran.

Abstract

Plagiarism is removal and to put it in their own name the ideas or words of others. With the Increasing progress of the Internet and the proliferation of online articles, scientific theft has also become easier. Many systems have been developed today to detect plagiarism. Most of these systems are based on lexical structure and string matching algorithms. Therefore, these systems can hardly detect recovery robberies, placement of synonyms. This paper presents a method for identifying plagiarism based on semantic role labeling and cellular learning automata. In this paper, cellular learning automata are used to locate the processed words. Semantic role labeling specifies the role of words in sentence. Comparison operations are performed for all sentences of the original text and suspicious text. Results of the experiments on PAN-PC-11 corpus demonstrate the proposed method improves values of evaluation parameters such as recall, precision and F-measure, comparing to previous approaches in plagiarism detection.

Keywords: Plagiarism Detection, Cellular Learning Automata, Semantic Role Labeling, Semantic Similarity.

* Corresponding Author: rezvaneyaghobi2050@gmail.com

How to Cite: Yaghobi, R., yaghobi, M., khotanloue, H., (2020). A New Approach to Plagiarism Detection Using Cellular Learning Automatons and Semantic Role Labeling, *Journal of Business Intelligence Management Studies*, 9(36), 183-208.

مقاله پژوهشی
تاریخ ارسال: ۹۸/۱۱/۰۲
تاریخ بازنگری: ۹۹/۰۶/۱۳
تاریخ پذیرش: ۹۹/۱۲/۱۲
ISSN: 1735/112X
eISSN: 2538-2527

مطالعات مدیریت کسب و کار هوشمند
سال نهم، شمارهٔ ۳۶، تابستان ۱۴۰۰، ص ۱۸۳ تا ۲۰۸
ims.atu.ac.ir
DOI: 10.22054/IMS.2021.49415.1661

رویکردی جدید برای شناسایی سرقت ادبی با استفاده از آتوماتای یادگیر سلولی و برچسب‌گذاری نقش معنایی

رضوان یعقوبی * کارشناسی ارشد کامپیوتر، دانشگاه آزاد اسلامی، واحد ملایر، باشگاه پژوهشگران جوان و نخبگان، ملایر، ایران.

مهدی یعقوبی کارشناسی ارشد فناوری اطلاعات، دانشگاه آزاد اسلامی، واحد ملایر، ایران.

حسن ختن لو استاد گروه کامپیوتر، دانشگاه بوعلی سینا، همدان، ایران.

چکیده

سرقت ادبی عبارت از برداشتن و به نام خود قلمداد کردن ایده و یا کلمات دیگران است. با پیشرفت روزافزون اینترنت و گسترش مقالات آنلاین، سرقت‌های علمی آسان‌تر شده است. امروزه سیستم‌های زیادی جهت شناسایی سرقت ادبی ایجاد شده‌اند. بیشتر این سیستم‌ها بر اساس ساختار لغوی و الگوریتم‌های تطابق رشته‌ای عمل می‌کنند؛ بنابراین این سیستم‌ها به‌سختی می‌توانند سرقت‌های بازگردانی و جایگذاری مترادف‌ها را شناسایی کنند. در این مقاله روشی جهت شناسایی سرقت ادبی بر مبنای برچسب‌گذاری نقش معنایی و آتوماتای یادگیر سلولی ارائه می‌شود. در این مقاله جهت قرارگیری کلمات پردازش‌شده از آتوماتای یادگیر سلولی استفاده می‌شود. برچسب‌گذاری نقش معنایی، نقش کلمات در جمله را مشخص می‌کند. عملیات مقایسه برای تمام جملات متن اصلی و متن مشکوک به سرقت انجام می‌شود. نتایج آزمایش بر روی مجموعه داده‌های PAN-PC-11 نشان می‌دهد که روش پیشنهادی، مقدار پارامترهای ارزیابی مانند Precision،Recall و F-measure را نسبت به روش‌های قبلی ارائه‌شده در زمینه شناسایی سرقت ادبی بهبود می‌دهد.

کلیدواژه‌ها: سرقت ادبی، آتوماتای سلولی، برچسب‌گذاری نقش معنایی، شباهت معنایی.

* نویسنده مسئول: rezvaneyaghobi2050@gmail.com

مقدمه

شناسایی شباهت متون یکی از شاخه‌های متن‌کاوی است که کاربرد آن در شناسایی سرقت ادبی است. سرقت ادبی به معنای گرفتن (ایده، اسناد، کد، عکس و ...) از دیگران و انتصاب (ایده‌ها، اسناد، عکس و ...) به نام خود بدون ذکر منبع و مرجع است. دسترسی آسان به وب و پایگاه داده‌های بزرگ و به‌طورکلی ارتباطات از راه دور باعث شده که سرقت ادبی به یک مشکل بزرگ برای ناشران، محققان و مؤسسات آموزشی تبدیل شود. اگر سرقت ادبی به‌درستی شناسایی نشود، متقلبان و سارقان می‌توانند به نتایجی برسند که مستحق آن نیستند. کپی- جایگزینی، بازگردانی، سرقت ادبی ترجمه‌ای، سرقت ادبی هنرمندانه، سرقت ادبی ایده، سرقت ادبی کد و سرقت ادبی اشتباه مرجع دادن از انواع سرقت‌های ادبی به شمار می‌آیند (سیندو و همکاران[1]،۲۰۱۱).

امروزه سیستم‌های زیادی جهت شناسایی سرقت ادبی ایجاد شده‌اند. بیشتر این سیستم‌ها بر اساس ساختار لغوی و الگوریتم‌های تطابق رشته‌ای عمل می‌کنند؛ بنابراین این سیستم‌ها به‌سختی می‌توانند سرقت‌های بازگردانی و جایگذاری مترادف‌ها را شناسایی کنند. شناسایی سرقت‌های ادبی معنایی چالش بزرگی است که در بحث شناسایی سرقت ادبی وجود دارد. در این مقاله ما برای بهبود چالش موجود در مبحث شناسایی سرقت ادبی از برچسب‌گذاری نقش معنایی به همراه آتوماتای یادگیر سلولی استفاده کرده‌ایم.

برچسب‌گذاری نقش معنایی یکی از روش‌های پردازش زبان طبیعی است که در شاخه‌های مختلف متن‌کاوی استفاده می‌شود (عثمان و همکاران[2]، ۲۰۱۲). آتوماتای یادگیر سلولی مدل قدرتمند ریاضی است که برای حل بسیاری از مسائل غیرمتمرکز استفاده می‌شود (اسماعیل‌پور و همکاران،۲۰۱۲). برچسب‌گذاری نقش معنایی جهت تعیین نقش معنایی هر کلمه در جمله استفاده می‌شود. در این مقاله از فرهنگ لغت WordNet استفاده شده است. با استفاده از فرهنگ لغت WordNet مجموعه مترادف‌های هر کلمه در جمله

1. Sindhu et al
2. Osman et al

استخراج می‌شود. این روش قادر است سرقت‌های کپی- جایگزینی، بازگردانی یا جایگذاری مترادف‌ها، تغییر ساختار کلمات در جمله را شناسایی کند. قسمت‌های بعدی مقاله به‌صورت زیر سازمان‌دهی می‌شود.

در قسمت دوم توصیف کاملی از کارهای مرتبط با شناسایی سرقت ادبی ارائه می‌شود. در قسمت سوم برچسب‌گذاری معنایی شرح داده می‌شود. در قسمت چهارم آتوماتای یادگیر سلولی شرح داده می‌شود. در قسمت پنجم روش پیشنهادی ما که مبتنی بر برچسب‌گذاری نقش معنایی و آتوماتای یادگیر سلولی است بیان می‌شود. طراحی آزمایش‌ها و مجموعه داده و معیار محاسبه شباهت در قسمت ششم بحث می‌شود. قسمت هفتم نتیجه‌گیری مقاله را ارائه می‌دهد.

پیشینه پژوهش

تشخیص سرقت ادبی خارجی یک کار بازیابی اطلاعات است. هدف از تشخیص سرقت ادبی خارجی مقایسه یک سند ورودی با یک مجموعه بزرگ از اسناد بیرونی و بازیابی کلیه اسنادی که دارای شباهت‌های بالاتر از یک آستانه هستند (میوسچک و همکاران[1]، ۲۰۱۹).

امروزه بسیاری از سیستم‌های شناسایی سرقت بر اساس ویژگی‌های نحوی عمل می‌کنند. در این روش با استفاده از برچسب‌گذاری قسمتی از متن و برخی از معیارهای شناسایی شباهت رشته‌ای به محاسبه شباهت بین دو متن پرداخته می‌شود. دو متن شبیه به هم از لحاظ نحوی ساختار نحوی برخی از جمله‌هایشان شبیه به هم است. الهادی و همکاران[2] (۲۰۰۸) روشی ارائه دادند که اسناد را بر اساس برچسب‌گذاری قسمتی از متن رتبه‌بندی می‌کنند.

1. Meuschke et al
2. Elhadi et al

برادر[1] (۱۹۹۷)، کریسزتی[2] (۲۰۰۰) و هینتز[3] و (۱۹۹۶) روش اثرانگشت را پیشنهاد کردند که تطابق رشته‌ها و شناسایی سرقت ادبی را بر مبنای درصد اثرانگشت پیدا می‌کرد. این روش کارایی خوبی دارد اما هنگامی‌که متنی به سرقت می‌رود و قسمت‌هایی از آن بازگردانی می‌شود یا برخی از کلمات متن با مترادف‌های کلمه جایگزین می‌شود، شناسایی سرقت ادبی با روش اثرانگشت با شکست مواجه می‌شود.

بسیاری از روش‌های شناسایی سرقت ادبی به ویژگی‌های لغوی مبتنی بر کاراکتر، ویژگی‌های لغوی مبتنی بر کلمه و ویژگی‌های نحوی وابسته‌اند. تطابق رشته بین دو رشته x و y به این معناست که آن‌ها کاراکترهای دقیقاً یکسان دارند. رشته‌های مشابه با استفاده از تطابق رشته تا حدودی می‌توانند شناسایی شوند (عثمان و همکاران،۲۰۱۲).

بیشتر روش‌های شناسایی سرقت ادبی بر اساس ویژگی‌های نحوی، لغوی یا معنایی اسناد کار می‌کنند. در روش ساختاری بر روی ویژگی‌های ساختاری مانند سرفصل، بخش، پاراگراف و ... اسناد کار می‌شود.

گیپ[4] (۲۰۱۰) روشی برای شناسایی سرقت ادبی مبتنی بر نقل‌قول ایجاد نمود. این روش برای شناسایی اسناد دانشگاهی که بدون ذکر کردن علامت نقل‌قول ایجاد می‌شوند به کارگرفته می‌شود.

گیپ و همکاران[5] (۲۰۱۱) روشی ایجاد کردند که برای شناسایی سرقت ادبی بر روی الگوهای مشابه در دنباله‌های نقل‌قول اسناد دانشگاهی تمرکز می‌کند. روش معنایی بر روی شناسایی شباهت معنایی دو متن کار می‌کند. دو جمله می‌توانند از لحاظ ساختاری و نحوی با هم فرق داشته باشند درحالی‌که از لحاظ معنایی شبیه به هم باشند.

عثمان و همکاران[6] (۲۰۱۲) روشی معنایی برای شناسایی سرقت ادبی مبتنی بر برچسب‌گذاری معنایی ارائه دادند. بسیاری از تحقیقات علمی جهت شناسایی شباهت

1. Broder
2. Kriszti
3. Heintze
4. Gipp
5. Gipp et al
6. Osman et al

معنایی از شبکه واژگان استفاده کرده‌اند. گلبوخ[1] (۲۰۰۹) شباهت معنایی بین کلمات را با محاسبه میزان ارتباط بین کلمات با استفاده از شبکه واژگان به دست آورده‌اند. چو و همکاران[2] (۲۰۱۰) روشی ارائه دادند که شباهت بین سند مشکوک به سرقت و سند اصلی را بر اساس جملات خبری محاسبه می‌کند. جملات خبری با استفاده از پارسر (تجزیه‌کننده) استنفورد استخراج می‌شود. درجه شباهت بین جملات استخراج‌شده با استفاده از فرهنگ لغت محاسبه می‌شود. الزهرانی و همکاران[3] (۲۰۱۰) روشی ارائه دادند که شباهت معنایی را با استفاده از شباهت رشته‌ای و منطق فازی محاسبه می‌کرد. وقتی دو جمله از متن کاملاً شبیه به هم بودند امتیاز یک و در غیر این صورت امتیاز صفر می‌گرفتند. سرقت‌های معنایی با استفاده از بازگردانی، جایگذاری مترادف‌ها و تبدیل جمله معلوم به مجهول و بالعکس و ... انجام می‌شود. روش ارائه‌شده در این مقاله بر اساس روش معنایی به شناسایی سرقت ادبی می‌پردازد. این روش می‌تواند اکثر سرقت‌های ادبی معنایی را تشخیص دهد. روش پیشنهادی با استفاده از برچسب‌گذاری نقش معنایی و فرهنگ لغت WordNet به تجزیه‌وتحلیل معنایی می‌پردازد. از آتوماتای یادگیر سلولی به‌عنوان ساختاری جهت قرارگیری کلمات و هم‌معنی‌های آن‌ها در سلول‌های یادگیر استفاده می‌شود. با توجه به تعداد نقش‌های موجود در جمله آتوماتای یادگیر سلولی تعریف می‌کنیم. سپس شباهت بین کلمات درون آتوماتا با استفاده از معیار شباهت محاسبه می‌شود.

روش‌شناسی پژوهش

در این قسمت بر روی ایده روش پیشنهادی‌مان صحبت می‌کنیم. در ابتدا برچسب‌گذاری قسمتی از متن انجام می‌شود. گام دوم روش پیشنهادی پیش‌پردازش است. عمل پیش‌پردازش که شامل قطعه کردن متن، حذف ایست واژه‌ها و ریشه گیری است، انجام می‌گیرد. بعد از انجام این مراحل برچسب‌گذاری نقش معنایی برای کلمات جملات متن

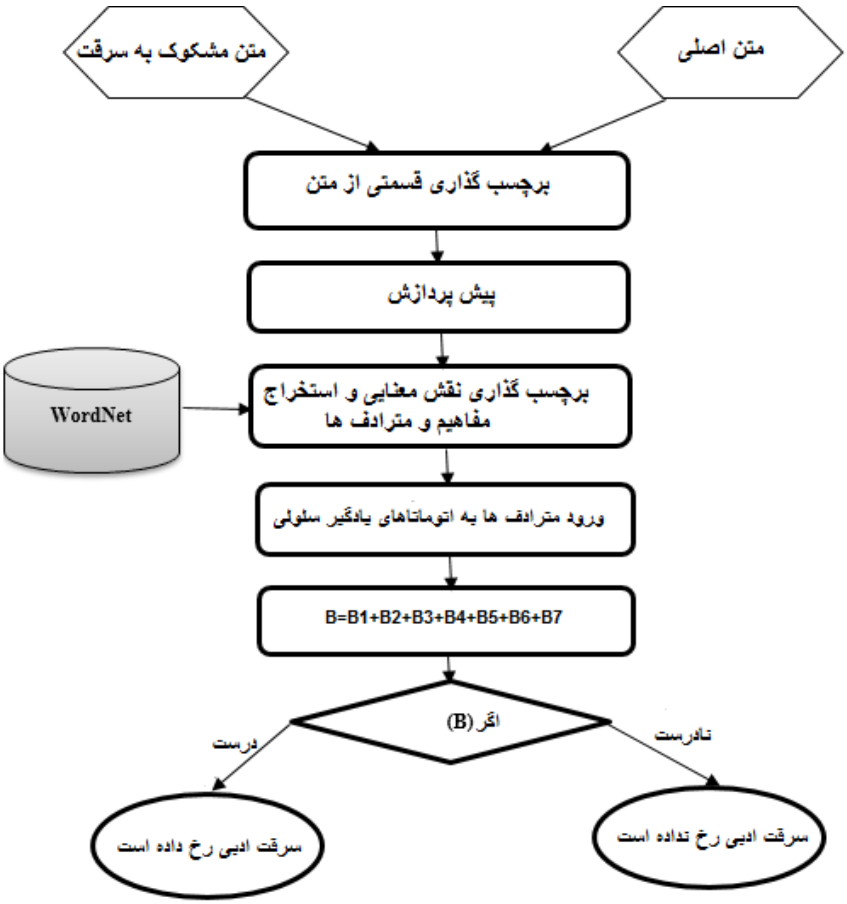
1. Gelbukh
2. Chow
3. Alzahrani et al

اصلی و متن مشکوک به سرقت انجام می‌شود.ARG0، ARG1، ARG2، V، NEG، ADJ و DIR برچسب‌هایی است که به کلمات جملات خورده می‌شود. جدول (۱) نشان می‌دهد که هرکدام از این برچسب‌ها در جمله چه نقشی را نشان می‌دهند. باید توجه داشت که برخی از کلمات با مراجعه به فرهنگ لغت ممکن است بیش از یک نقش داشته باشند مثلاً هم‌اسم باشند و هم صفت باشند. روش پیشنهادی در این موارد نقشی را در نظر می‌گیرد که با توجه به آنالیز برچسب‌گذاری اجزای کلام و برچسب‌گذاری نقش معنایی دانشگاه ایلینوی صورت می‌گیرد.

جدول ۱. برچسب کلمات در متن به همراه نقششان در جمله

برچسب	توضیحات
ARG0	فاعل
ARG1	مفعول
ARG2	نقش اضافه
V	فعل
NEG	نشانگرهای منفی
ADJ	صفت
DIR	حروف اضافه

به تعداد نقش‌های معنایی که در جمله است آتوماتای یادگیر سلولی تعریف می‌شود. سپس همه مفاهیم و مترادف‌ها را برای هر واژه برچسب خورده در جمله با استفاده از فرهنگ لغت WordNet استخراج می‌کنیم. این مفاهیم برای هر نقش از جمله وارد آتوماتای یادگیر سلولی می‌شوند. شکل (۱) معماری کلی روش پیشنهادشده را نشان می‌دهد. در بخش پنجم مقاله معماری روش پیشنهادی با جزئیات بیشتری توضیح داده می‌شود.



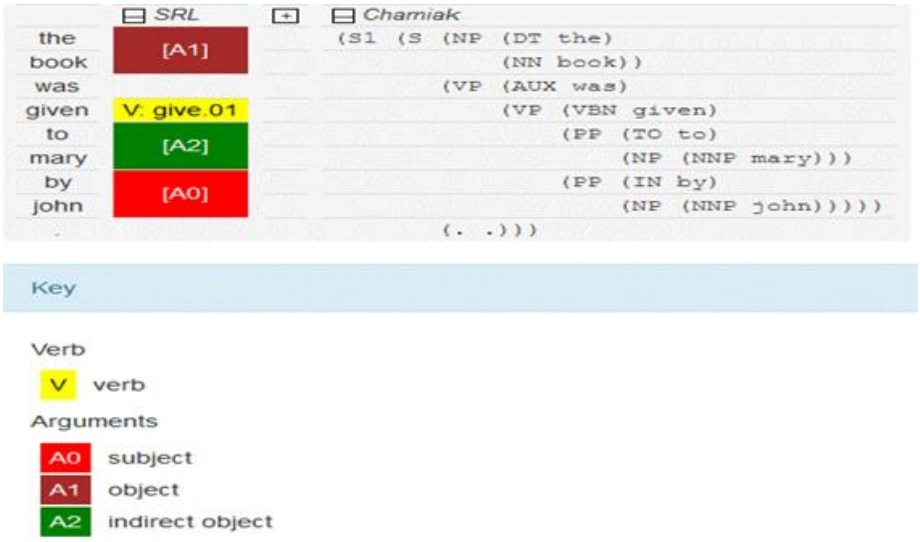
شکل ۱. معماری کلی روش پیشنهادی

برچسب‌گذاری قسمتی از متن

بخش‌های سخن، طبقه‌بندی‌هایی زبانی از کلمات هستند که رفتار نحوی یک قسمت از جمله را بیان می‌کنند. مهم‌ترین بخش‌های نحوی اسم، فعل، صفت و قید است. روش پیشنهادی ما از ابزار برخط دانشگاه ایلینوی برچسب‌گذاری قسمتی از متن استفاده می‌کند. تعداد نقش‌هایی که به یک واژه با مراجعه به فرهنگ لغت داده می‌شود ممکن است بیش از یک باشد. مثلاً هم‌اسم و هم صفت باشد. در این صورت نقشی برای کلمه در نظر گرفته می‌شود که توسط برچسب‌گذار نقش معنایی دانشگاه ایلینوی تعیین شده است. در شکل

(۲) برچسب‌گذاری نقش معنایی برای جمله "the book was given to mary by john " نشان داده شده است؛ و با توجه به اینکه یک کلمه می‌تواند در جمله هم‌اسم و هم صفت باشد، ابتدا برچسب‌گذاری قسمتی از متن و در مراحل بعدی برچسب‌گذاری معنایی انجام می‌شود. در واقع با استفاده از برچسب‌گذاری اجزای واژگانی کلام (pos) عمل انتساب برچسب‌های واژگانی به کلمات و نشانه‌های تشکیل‌دهنده یک متن است به صورتی که این برچسب‌ها نشان‌دهنده نقش کلمات و نشانه‌ها در جمله باشد.

درصد بالایی از کلمات از دیدگاه برچسب واژگانی دارای ابهام هستند زیرا کلمات در جایگاه‌های مختلف برچسب‌های واژگانی متفاوتی دارند؛ بنابراین برچسب‌گذاری واژگانی عمل ابهام‌زدایی از برچسب‌ها با توجه به زمینه (متن) مورد نظر است. برچسب‌گذاری قسمتی از متن (POS) و برچسب‌گذاری نقش معنایی (SRL) توسط سرویس دانشگاه آیلینوی انجام شد.



شکل ۲. برچسب گذاری نقش معنایی برای یک جمله

پیش‌پردازش

پیش‌پردازش یکی از گام‌های مهم در پردازش زبان طبیعی و متن‌کاوی است. پیش‌پردازش شامل مراحل قطعه‌قطعه کردن، حذف ایست واژه‌ها و ریشه گیری است (ویرمانی و همکاران[1]،۲۰۱۹).

برای متن‌کاوی نیاز است که بر روی متن عملیاتی انجام شود. یکی از گام‌های اولیه پیش‌پردازش، قطعه کردن متن است. در این گام متن به واحدهای معنی‌دار مانند جملات، کلمات و... تقسیم می‌شود. در روش پیشنهادی ما واحدهای معنی‌دار جمله‌ها هستند. جملات همان قطعه‌های معنی‌دار هستند. این گام برای متن اصلی و متن مشکوک به سرقت انجام می‌شود.

ایست واژه‌ها کلماتی هستند که در متن زیاد تکرار می‌شوند. این کلمات معنای مفیدی ندارند و سرعت پردازش را کاهش می‌دهند. فضای خالی زیادی از حذف کردن ایست واژه‌ها ایجاد می‌شود اما حذف این کلمات تأثیری در بازگردانی اطلاعات ندارد. در اغلب زمینه‌های متن‌کاوی، ایست واژه‌ها حذف می‌شوند. حذف این کلمات نتایج پردازش را بهبود می‌دهد و سبب کاهش بار محاسبات و افزایش سرعت خواهد شد. به همین دلیل، این کلمات غالباً در فاز پیش‌پردازش حذف می‌شوند (ژانگ و همکاران[2]،۲۰۱۹).

بعد از حذف ایست واژه‌ها گام بعدی ریشه‌یابی است. در این مرحله به‌منظور یکسان‌سازی اشکال مختلف یک کلمه، یکپارچه‌سازی انجام می‌شود. ریشه‌یابی به فرایند تبدیل کلمات به فرم ریشه‌ای و پایه‌ای آن‌ها اشاره می‌کند. این مرحله در پردازش متن اهمیت زیادی دارد زیرا باعث می‌شود کامپیوتر با کلمات هم‌خانواده که ظاهراً با هم متفاوت هستند مانند دوکلمه‌ای که از لحاظ ریشه‌ای هیچ ارتباطی با هم ندارند، برخورد ننماید. الگوریتم‌های مختلفی برای ریشه‌یابی لغات پیشنهاد شده است. روش پیشنهادی ما از الگوریتم پورتر برای ریشه‌یابی لغات متن استفاده می‌کند (تاتا و همکاران[3]،۲۰۲۰).

---

1. Virmani et al
2. Zhang et al
3. Thatha et al

برچسب‌گذاری نقش معنایی جمله‌های پیش‌پردازش شده

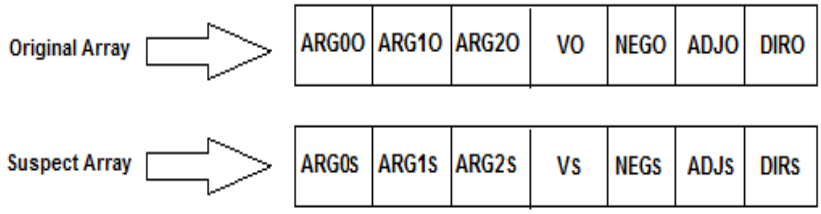
در این مرحله برچسب‌گذاری نقش معنایی را برای جملات متن اصلی و متن مشکوک به سرقت انجام می‌دهیم. با استفاده از برچسب‌گذاری نقش معنایی نقش هر کلمه به آن برچسب می‌خورد. برچسب‌هایی که به کلمات متن اصلی و متن مشکوک به سرقت می‌خورد با توضیحات کامل در جدول (۲) نشان داده شده است.

جدول ۲. برچسب‌های کلمات جمله اصلی و مشکوک به سرقت

توضیح	برچسب کلمات	توضیح	برچسب کلمات
فاعل در متن مشکوک	ARG0S	فاعل در متن اصلی	ARG0O
مفعول در متن مشکوک	ARG1S	مفعول در متن اصلی	ARG1O
نقش اضافه در متن مشکوک	ARG2S	نقش اضافه در متن اصلی	ARG2O
فعل در متن مشکوک	VS	فعل در متن اصلی	VO
نشانگرهای منفی در متن مشکوک	NEGS	نشانگرهای منفی در متن اصلی	NEGO
صفت در متن مشکوک	ADJS	صفت در متن اصلی	ADJO
حروف اضافه در متن مشکوک	DIRS	حروف اضافه در متن اصلی	DIRO

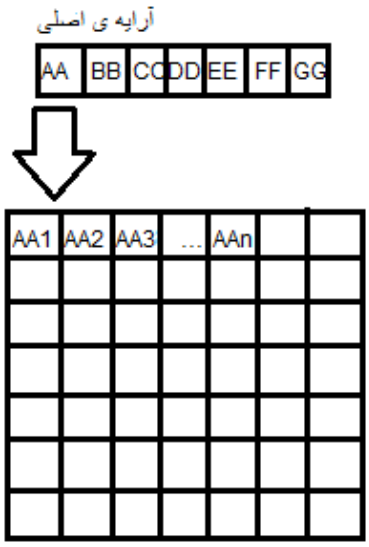
گروه‌بندی کلمات برچسب خورده و استخراج مفاهیم با استفاده از فرهنگ لغت WordNet

در این مرحله نقش‌های جملات متن اصلی و متن مشکوک به سرقت گروه‌بندی می‌شود. برای گروه‌بندی دو آرایه Suspect, Original با اندازه ۱ * ۷ در نظر می‌گیریم. آرایه Original را به نقش‌های جمله اصلی و آرایه Suspect را به نقش‌های جمله مشکوک به سرقت اختصاص می‌دهیم. سپس هرکدام از درایه‌های آرایه را به یکی از برچسب‌ها اختصاص می‌دهیم. در شکل (۳) این اختصاص نشان داده شده است.

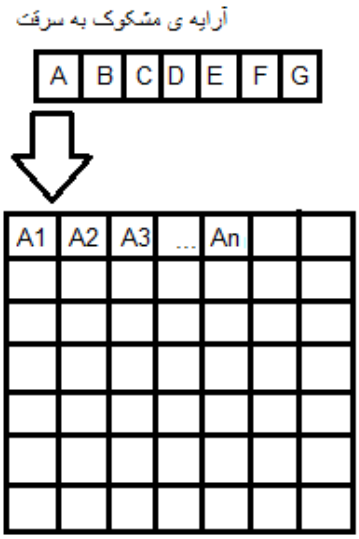


شکل ۳. اختصاص درایه‌های آرایه‌ها به برچسب‌های جملات

بعدازاینکه هر دو آرایه ۷*۱ با برچسب‌ها پر شد با استفاده از فرهنگ لغت WordNet به استخراج مفاهیم می‌پردازیم. مفاهیم استخراج‌شده برای هر برچسب از آرایه ۷*۱ را در یک آرایه ۷*۷ ذخیره می‌شود. سپس با توجه به روشی که در بخش ۵-۵ ذکر می‌شود، مفاهیم و مجموعه مترادف‌های ذخیره‌شده در آرایه ۷*۷ را وارد هفت آتوماتای یادگیر سلولی می‌کنیم. این تخصیص در شکل (۴) برای نقش‌های جمله اصلی و شکل (۷) برای نقش‌های جمله مشکوک به سرقت نشان داده شده است. A1,A2,..., An و $AA_1,AA_2,\ldots,AA_n$ همان مجموعه مترادف‌ها و مفهوم‌ها برای جمله اصلی و مشکوک هستند.



شکل ٤. نحوه تخصیص مفاهیم استخراج‌شده هر درایه آرایه به ماتریس ۷*۷ در متن اصلی



شکل ۵. نحوه تخصیص مفاهیم استخراج‌شده هر درایه آرایه به ماتریس ۷*۷ در متن مشکوک به سرقت

ورود مجموعه مترادف‌ها به آتوماتاهای یادگیر سلولی

با توجه به اینکه هر جمله از متن هفت نقش یا هفت برچسب می‌تواند داشته باشد، هفت آتوماتای یادگیر سلولی تعریف می‌کنیم؛ یعنی برای هرکدام از خانه‌های آرایه ۷*۱ یا به‌عبارت‌دیگر یک نقش در جمله و مجموعه مترادف‌ها یک آتوماتا تعریف می‌کنیم. پر کردن هر آتوماتای یادگیر سلولی به این صورت است که اولین مجموعه مترادف از متن مشکوک به سرقت را در مرکز آتوماتای یادگیر سلولی قرار می‌دهیم. همه مجموعه مترادف‌ها از جز جمله اصلی را به‌صورت تصادفی در آتوماتای یادگیر سلولی پخش می‌کنیم.

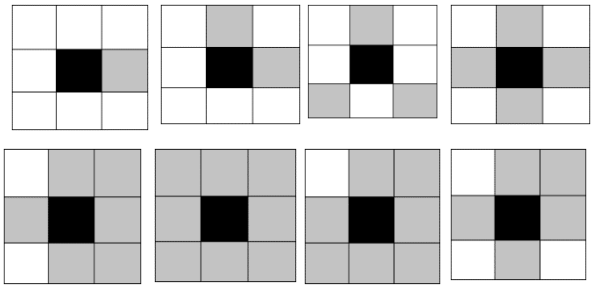
در این مقاله برای محاسبه شباهت از معیار WU-PALMER استفاده می‌شود. با توجه به اینکه هرکدام از آتوماتاهای یادگیر آتوماتای یادگیر سلولی از نوع P می‌باشند بنابراین نرخ پاداش و جریمه را تعریف می‌کنیم.

برای هر آتوماتای یادگیر سلولی عنصر مرکز آتوماتای یادگیر سلولی را با تمام مجموعه مترادف‌های پخش‌شده در آتوماتای یادگیر سلولی مقایسه می‌کنیم. مترادف‌هایی که امتیاز شباهتشان بزرگ‌تر مساوی ۰/۵ است، پاداش می‌گیرند. این مترادف‌ها در مرحله بعد همسایه سلول مرکز آتوماتا می‌شوند. مترادف‌هایی که امتیاز شباهتشان کوچک‌تر از ۰/۵ است جریمه می‌شوند و در مکان خود باقی می‌مانند.

اگر امتیاز شباهت آتوماتای یادگیر در آتوماتای یادگیر سلولی در یک ضرب شود به این معناست که آتوماتای یادگیر پاداش گرفته است. اگر امتیاز شباهت آتوماتای یادگیر در آتوماتای یادگیر سلولی در صفر ضرب شود به این معناست که آتوماتای یادگیر جریمه شده است. قوانین به‌روز شدن هرکدام از آتوماتاهای یادگیر آتوماتای یادگیر سلولی با توجه به مقدار امتیازشان در معادله (۱) نشان داده شده است. به‌عبارت‌دیگر نرخ پاداش یک و نرخ جریمه صفر است.

$$\begin{cases} Score(\acute{x}) == Score * 0, & Score(x) < 0.5 \\ Score(\acute{x}) == Score * 1, & Score(x) \geq 0.5 \end{cases} \quad (1)$$

نحوه قرار گرفتن مجموعه مترادف‌هایی که امتیاز شباهتشان بزرگ‌تر مساوی ۰/۵ است به دور سلول مرکز آتوماتا در شکل (۶) نشان داده شده است. سلول‌های مشکی مرکز آتوماتای سلولی و سلول‌های خاکستری همسایه‌های مرکز آتوماتا محسوب می‌شوند.



شکل ٦. همسایگی‌های تعریف‌شده در آتوماتای یادگیر سلولی

بعدازاینکه آتوماتای یادگیر سلولی به‌روز شد همسایگان سلول مرکز آتوماتای یادگیر سلولی که امتیاز شباهتشان برابر یک است به‌عنوان هم‌معنای عنصر مرکز آتوماتا تشخیص داده می‌شود. این عملیات برای تمام مجموعه مترادف‌های جز جمله مشکوک به سرقت انجام می‌شود. هر بار یکی از مجموعه مترادف‌های جز مشکوک به سرقت در مرکز آتوماتا قرار می‌گیرد و محاسبه شباهت انجام می‌شود زیرا ما نمی‌دانیم دقیقاً کدام مفهوم از جز جمله مشکوک به سرقت مدنظر است؛ بنابراین محاسبه شباهت را با تمام مجموعه مترادف‌های جز جمله مشکوک به سرقت انجام می‌دهیم.

برای هر آتوماتای یادگیر سلولی یک متغیر بولی (B_i) تعریف می‌کنیم. اگر عنصر مرکز آتوماتا با یکی از مجموعه مترادف‌های پخش‌شده در آتوماتای یادگیر سلولی هم‌معنا باشد مقدار این متغیر بولی مقدار TRUE را نشان می‌دهد در غیر این صورت مقدار FALSE را نشان می‌دهد. زمانی که متغیر بولی همه آتوماتاهای یادگیر سلولی مقدار را داشته باشند به این معنا است که دو جمله از متن مشکوک به سرقت با جمله از متن اصلی از نظر معنایی شبیه به هم هستند. مقدار متغیر B با استفاده معادله (۲) به دست می‌آید. در این مرحله به تعداد جملات به سرقت رفته یکی افزوده می‌شود.

$$B = B_1 + B_2 + B_3 + B_4 + B_5 + B_{6+}B_7 \qquad (2)$$

متغیرهای B که در شکل (۲) وجود دارند نتیجه وجود شباهت یا عدم وجود شباهت بین سند اصلی و سند مشکوک به سرقت است. اگر به شکل (۹) توجه کنید به ازای هر نقش در جمله‌ک اتوماتای سلولی یادگیر در نظر گرفته شده است؛ که بعد از قرارگیری مترادف‌ها از متن اصلی و متن مشکوک به سرقت اگر شباهت بالای پنجاه‌درصد به دست بیاید مقدار متغیرهای B یک خواهد شد و در غیر این صورت صفر خواهد شد.

متغیر B نهایی حاصل or متغیرهای $B_1...B_7$ است. متغیرB_1 برای آتوماتای ARG0-CLA، متغیرB_2 برای آتوماتای ARG1-CLA، متغیرB_3 برای آتوماتای ARG2-CLA ، متغیرB_4 برای آتوماتایV، متغیرB_5 برای آتوماتایNEG ، متغیرB_6 برای آتوماتایADG ، متغیرB_7 برای آتوماتایDIR است.

زمانی که از مقدار متغیرهای B1 تا B7 صحبت می‌کنیم منظورمان این است که هرکدام از این متغیرها متعلق به ارزیابی یک آتوماتا است. طبق معادله (۱) امتیاز شباهت برای هر آتوماتا به دست می‌آید که این امتیازها صفر یا یک است. درنهایت هفت متغیرB خواهیم داشت که می‌توانند بر اساس شباهت بین سند مشکوک و سند اصلی مقدار یک یا صفر اخذ کنند. مقدار این هفت متغیر را با هم or می‌کنیم تا اگر حتی در یکی از اتوماتاها سرقت رخ داده باشد، سرقت به‌راحتی مشخص گردد.

در شکل (۷) تشخیص سرقت ادبی بین دو جمله متن مشکوک به سرقت و جمله متن اصلی با جزییات بیشتری نشان داده شده است. همان‌طور که در شکل (۷) می‌بینید ابتدا سند اصلی و سند مشکوک به سرقت را از بین متون مجموعه داده‌ها انتخاب می‌کنیم. ابتدا عمل برچسب‌گذاری اجزای کلام انجام شده و نقش هر کلمه معین می‌گردد. سپس با تجزیه متن به جهت استخراج واحدهای معنایی ابتدا جملات هر متن استخراج می‌گردند و درون هشت‌ضلعی‌ها قرار می‌گیرند. ممکن است تعداد جملات متون مختلف با هم متفاوت باشد. مثلاً متنی ۱۰۰ جمله داشته باشد و متنی دیگر ۲۰ جمله داشته باشد. که برای شناسایی جملات متن، در قسمت پیاده‌سازی علائم نگارشی مانند نقطه، نقطه –ویرگول و ... را به معنای اتمام جمله در نظر گرفته‌ایم.

بعد از این‌که جملات متن مشخص شدند، کلمات مهم استخراج می‌شوند. بسیاری از ایست واژه‌ها حذف می‌شوند و فقط کلمات مهم باقی می‌مانند. کلمات مهم در شش‌ضلعی‌هایی که در شکل (۷) نشان داده شده است قرار می‌گیرند. با توجه به اینکه تعداد کلمات یک جمله متفاوت است تعداد شش‌ضلعی‌ها هم متفاوت خواهد بود و آنچه در شکل (۷) رسم شده است صرفاً جهت ارائه معماری روش پیشنهادی است.

بعد از اینکه کلمات جمله مشکوک به سرقت و جمله اصلی مشخص شدند، هر کلمه با توجه به نقشی که دارد در ارائه مخصوص متن اصلی و متن مشکوک به سرقت قرار می‌گیرد. پیش‌تر در جدول (۲) این نقش‌ها و خصوصیات آن‌ها تعریف شد.

بعد از اینکه هر کلمه در جایگاه خود قرار گرفت، نیاز است که مترادف‌های آن کلمه جهت محاسبه شباهت استخراج شوند.

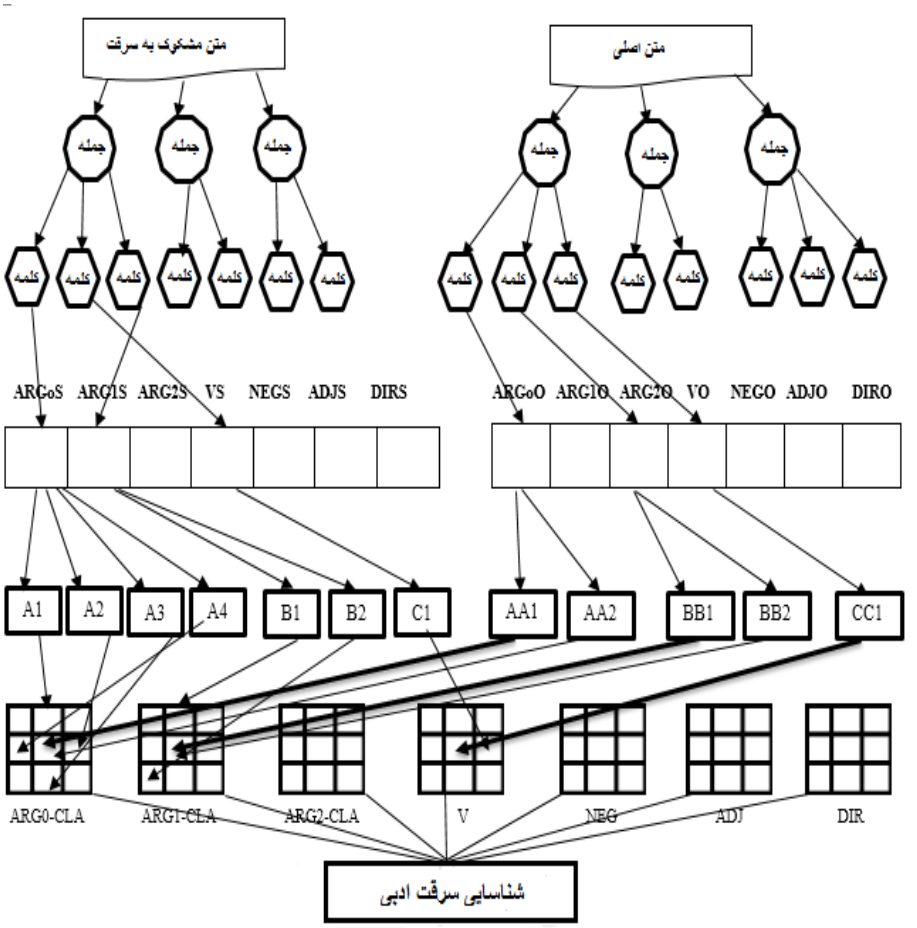
با توجه به شکل (۷) مترادف‌های کلمه‌ای که نقش ARG0S را دارد عبارت‌اند از A_1,A_2,A_3,A_4 و مترادف‌های کلمه‌ای که نقش ARG0O را دارند عبارت‌اند از AA_1,AA_2. این مترادف‌ها با توجه به اینکه نقششان در جمله ARG0 بوده است وارد آتوماتای ARG0-CLA می‌شوند و بر اساس فرمول شباهت، به محاسبه شباهتشان پرداخته می‌شود. بقیه آتوماتاهای موجود هم به همین روش پر می‌شوند.

با توجه به نقش‌هایی که در جمله وجود دارد برای هر نقشی که در جدول (۲) ذکر شده یک آتوماتا در نظر گرفته شده است که مترادف‌ها هر کلمه با توجه به نقش آن کلمه در آتوماتای متعلق به خودشان ذخیره می‌گردند.

این آتوماتاها با توجه به نقش هر کلمه شامل، ADJ,NEG, ARG2-CLA, ARG1-CLA, ARG0-CLA DIR و V می‌باشند.

یعنی هر آتوماتا با توجه به اسمی که دارد مترادف‌های کلمه‌ای که نقش آن به آتوماتا نسبت داده شده است درون آتوماتا قرار می‌گیرند؛ و با توجه به معادله (۱) محاسبه امتیاز شباهت برای هر آتوماتا صورت می‌گیرد.

همان‌طور که در قبلاً ذکر شد بعد از برچسب‌گذاری کلمات مترادف‌های آن‌ها با کمک فرهنگ لغت Wordnet استخراج می‌گردد و در یک آرایه ذخیره می‌شود. این مترادف‌های ذخیره‌شده با توجه به اینکه با چه کلمه‌ای مترادف هستند و آن کلمه چه نقشی در جمله دارد وارد آتوماتای خاص خودشان می‌شوند. البته این مترادف‌ها قبل از ورود به آتوماتاها در یک آرایه ذخیره می‌گردند. مترادف‌ها برای کلمه از متن اصلی و متن مشکوک به سرقت استخراج می‌شوند و درون آرایه ذخیره می‌گردند. تعداد مترادف‌های هر کلمه متغیر است.



شکل ۷. مقایسه دو متن جهت تشخیص سرقت ادبی

یافته‌ها

برای ارزیابی روش پیشنهادی لازم است که از مجموعه داده‌های استاندارد استفاده کنیم. در این مقاله ما از مجموعه داده‌های PAN-PC-11 استفاده کرده‌ایم. برای ارزیابی روش نیاز به یک معیار جهت محاسبه شباهت بود، بنابراین از معیار شباهت WU-PALMER استفاده می‌کنیم. برای ارزیابی روش پیشنهادی از سه پارامتر Recall، Precision و F-measure استفاده کرده‌ایم. برای بار اول ارزیابی سه متن مشکوک به سرقت را با متن اصلی مقایسه می‌کنیم. برای بار دوم ارزیابی را با شش سند مشکوک به سرقت و متن اصلی

انجام می‌دهیم. برای آخرین بار نه سند مشکوک به سرقت را با متن اصلی مقایسه می‌کنیم و نتایج را ارزیابی می‌کنیم.

شناسایی شباهت

معیارهای مختلفی برای محاسبه شباهت معنایی وجود دارد. در این مقاله ما از معیار WU-PALMER که بر مبنای طول مسیراست (چو[1]،۲۰۱۰)، استفاده کردیم. شمارش تعداد گره‌های درخت یا پیوند بین گره‌ها، یکی از راه‌های ممکن جهت محاسبه شباهت معنایی است. کمترین فاصله بین دو مفهوم به معنای بیشترین شباهت بین آن‌ها است.

معیار WU-PALMER بر اساس طول عمق و نزدیک‌ترین رده بند مشترک مجموعه مترادف‌ها، به محاسبه شباهت می‌پردازد. معادله (۳) نحوه محاسبه شباهت را بر اساس WU-PALMER نشان می‌دهد.

$$Sim(S,T) = \frac{2 * Depth(LCS)}{Depth(S) + Depth(T)} \qquad (3)$$

در معادله (۳)،T را به مجموعه مترادف‌های کلمه از متن اصلی و S به مجموعه مترادف‌های کلمه از متن مشکوک به سرقت اشاره می‌کند. Depth (T) به کوتاه‌ترین فاصله از گره ریشه تا گره اشاره می‌کند.

Depth (S) به کوتاه‌ترین فاصله از گره ریشه تا گره اشاره می‌کند. LCS به نزدیک‌ترین رده بند مشترک بین T و S اشاره می‌کند. از این معیار محاسبه شباهت بین سلول مرکز آتوماتای یادگیر سلولی و مجموعه مترادف‌های پخش‌شده در آتوماتای یادگیر سلولی استفاده کردیم.

روش ما بر روی تعداد سندهای مختلف (۹،۶،۳) آزمایش شد. برای ارزیابی روش پیشنهادی سه پارامتر که معمولاً در شناسایی سرقت ادبی کاربرد دارند، در معادله (۴)، معادله (۵) و معادله (۶) بیان شده‌اند.

1. Chow

عملکرد یک سیستم متن‌کاوی، از طریق پارامترهای متفاوتی نظیر، "صحت "(Recall)،"دقت " (Precision) و "امتیاز " (F-measure) سنجیده می‌شود. درک این معیارها، به کاربران اجازه می‌دهد تا بفهمند که یک مدل دسته‌بندی توسعه داده شده تا چه حد در تحلیل داده‌های متنی خوب عمل می‌کند.

معیار "صحت " (Recall)، بیان‌کننده نسبت تعداد داده‌های متنی درست دسته‌بندی شده در یک کلاس خاص، به تعداد کل داده‌هایی است که باید در همان کلاس خاص دسته‌بندی شوند. مقدار بالا برای معیار صحت، بیانگر تعداد کم داده‌هایی است که به اشتباه، در آن کلاس خاص دسته‌بندی نشده‌اند. استفاده از این معیار، به‌تنهایی، برای ارزیابی عملکرد سیستم درست نیست و باید در کنار معیار "دقت " (Precision) مورد استفاده قرار بگیرد؛ زیرا به‌راحتی می‌شود مدل‌های دسته‌بندی متنی طراحی کرد که صحت بالایی داشته باشند و این لزوماً به معنای "دقت "(Precision) بالا نیست.

معیار "دقت " (Precision)، نسبت تعداد "پیش‌بینی‌های صحیح انجام شده " برای نمونه‌های یک کلاس خاص، به تعداد "کل پیش‌بینی‌ها " برای نمونه‌های همان کلاس خاص را (این تعداد، مجموع تمامی پیش‌بینی‌های صحیح و پیش‌بینی‌های نادرست را شامل می‌شود) ارزیابی می‌کند. مقدار بالا برای معیار دقت، بیانگر تعداد کم داده‌هایی است که به‌اشتباه، در کلاس خاص دسته‌بندی شده‌اند. شایان توجه است که معیار دقت، فقط برای مواردی ارزیابی می‌شود که در آن‌ها، مدل دسته‌بندی تعلق یک نمونه به یک کلاس خاص را پیش‌بینی کرده باشد.

معیار " F-measure "پارامترهای "دقت " (Precision) و "صحت "(Recall) را با هم ترکیب می‌کند تا مشخص شود یک مدل دسته‌بند تا چه حد عملکرد خوبی از خود نشان می‌دهد. به این معیار،"میانگین متوازن "دو معیار دقت (Precision) و صحت (Recall) نیز گفته می‌شود. این معیار، نسبت به معیار صحت، تصویر دقیق‌تری از نحوه عملکرد مدل دسته‌بند روی تمامی کلاس‌های موجود در داده‌ها ترسیم می‌کند.

$$Recall = \frac{Number\ of\ Detect\ Sentences}{Total\ Number\ of\ Sentences} \quad (4)$$

$$Precision = \frac{Number\ of\ Plagiarized\ Sentences}{Number\ of\ Detect\ Sentences} \quad (5)$$

$$F - mesure = \frac{2 * Recall * Precision}{Recall + Precision} \quad (6)$$

معادله (۴) نحوه محاسبه مقدار پارامتر "صحت " که حاصل تقسیم جملات شناسایی‌شده به کل جملات متن است نشان می‌دهد.

چگونگی محاسبه پارامتر "دقت " که حاصل تقسیم جملات شناسایی‌شده سرقتی تقسیم‌بر تعداد کل جملات است در معادله (۵) آمده است.

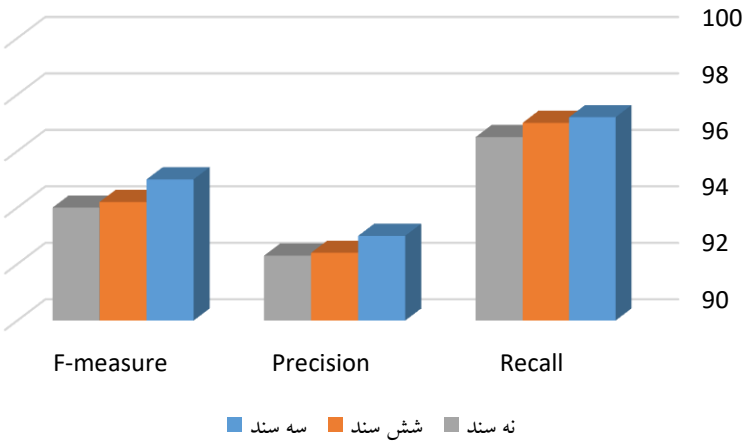
در معادله (۶) نحوه محاسبه پارامتر "امتیاز " آمده است.

جدول (۳) به ترتیب میانگین پارامترهای ارزیابی (صحت، دقت و امتیاز) را برای شناسایی شباهت سه، شش و نه سند مشکوک به سرقت را با متن اصلی نشان می‌دهد. جدول (۳) نتایج شباهت بین اسناد مشکوک و اصلی را برای هر مجموعه از اسناد نشان می‌دهد. ما دریافتیم که همه مقادیر اندازه‌گیری صحت بالاتر از ۰/۹۶ و همه مقادیر دقت بالاتر از ۰/۹۲ و مقادیر امتیاز بیشتر از ۰/۹۴ بود. در واقع این مقادیر نتایج خوبی به نظر می‌رسد زیرا از ۰/۹۰ بیشتر است.

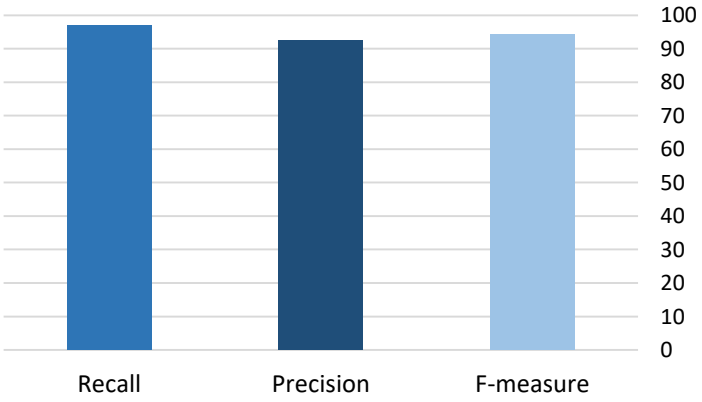
جدول ۳. میانگین پارامترهای ارزیابی برای اسناد مشکوک به سرقت مقایسه شده با متن اصلی

تعداد سند مشکوک به سرقت	Recall	Precision	F-measure
۳	۰,۹۷۲	۰,۹۳۰	۰,۹۵۰
۶	۰,۹۷۰	۰,۹۲۴	۰,۹۴۲
۹	۰,۹۶۵	۰,۹۲۳	۰,۹۴۰

شکل (۸) مقدار پارامترهای ارزیابی روش پیشنهادی ما را به صورت درصد نشان می‌دهد. در شکل (۹) میانگین پارامترهای ارزیابی به صورت درصدی نشان داده شده است.



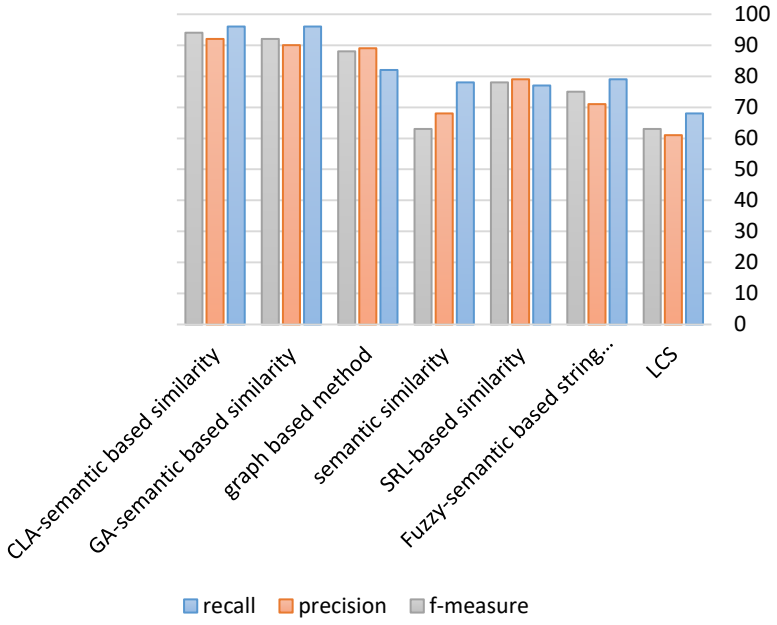
شکل ۸. پارامترهای ارزیابی روش پیشنهادی بر اساس درصد



شکل ۹. میانگین پارامترهای ارزیابی روش پیشنهادی

شکل (۱۰) مقایسه بین روش پیشنهادی ما با روش‌های الزهرانی و همکاران[1] (۲۰۱۰)، چو و همکاران[2] (۲۰۱۰)، کنت و همکاران[3] (۲۰۱۰)، عثمان و همکاران[4] (۲۰۱۱)، وایت و جوی[5] (۲۰۰۴)، عثمان و همکاران[6] (۲۰۱۲) و یعقوبی (۱۳۹۴) را نشان می‌دهد.

با توجه به شکل (۱۰) به نظر می‌رسد که در روش پیشنهادی پارامترهای ارزیابی شناسایی سرقت ادبی نتایج بهتری را کسب می‌کنند. روش پیشنهادی می‌تواند انواع سرقت ادبی را شناسایی کند. روش پیشنهادی می‌تواند کپی - جایگزینی، جایگذاری مترادف‌ها، تغییر ساختار جملات و سرقت‌های معنایی را به‌خوبی و با دقت بالا شناسایی کند.



شکل ۱۰. مقایسه روش پیشنهادی و روش‌هایی ارائه‌شده قبلی

1. Alzahrani et al
2. Chow et al
3. Kent et al
4. Osman et al
5 White and joy
6. Osman et al

بحث و نتیجه‌گیری

در این مقاله ما یک روش جهت بهبود روش‌های شناسایی سرقت ادبی مبتنی بر آتوماتای یادگیر سلولی و برچسب‌گذاری نقش معنایی ارائه دادیم. در روش پیشنهادی ما، برچسب‌گذاری نقش معنایی به همراه آتوماتای یادگیر سلولی می‌تواند کارایی خوبی در شناسایی سرقت ادبی داشته باشد. آتوماتای یادگیر سلولی ساختاری است که با استفاده از آن توانستیم پیش‌پردازش و پس پردازش متون را به‌خوبی انجام دهیم. استفاده هم‌زمان از برچسب‌گذاری نقش معنایی و قسمتی از متن باعث شد که دقت روش پیشنهادی نسبت به روش‌هایی که قبلاً ارائه شده‌اند بیشتر باشد.

آزمایش‌ها را بر روی متن‌های مجموعه داده‌های PAN-PC-11 انجام دادیم. روش پیشنهادی ما در مقایسه با روش‌های قبلی که مبتنی بر برچسب‌گذاری نقش معنایی بوده‌اند، کارایی بهتری دارد. روش پیشنهادی می‌تواند سرقت‌های کپی- جایگزینی، جایگذاری مترادف‌ها، تغییر ساختار جملات و سرقت‌های معنایی و ... را تشخیص دهد.

برای تکمیل این پژوهش و تحقیق بیشتر در این موضوع، قصد داریم در آینده بر روی سرقت ادبی دوزبانه تحقیق کنیم. در کارهای آینده تمرکزمان در جهت شناسایی سرقت‌های ادبی ترجمه‌ای است. محدودیت‌هایی که برای تکمیل این پژوهش با آن روبرو شدیم نبود مجموعه داده استاندارد جهت کار بر روی مجموعه داده‌های فارسی بود. بنابراین یکی از محدودیت‌های این پژوهش این بود که مجبور شدیم روی داده‌های با زبان انگلیسی کار کنیم؛ و نیاز به یک مجموعه داده متون فارسی جهت پژوهش‌های آینده وجود دارد.

ORCID
Rezvan Yaghobi **http://orcid.org/**
Mahdi yaghobi **http://orcid.org/**
Hassan khotanloue **http://orcid.org/**

منابع

رضوان، یعقوبی و حسن ختنلو. (۱۳۹٤). شناسایی سرقت ادبی مبتنی بر الگوریتم ژنتیک و برچسب‌گذاری نقش معنایی در مقالات علمی. *فصلنامه صنایع الکترونیک*,٦(۳)، ٦۷-۷۹.

مهدی، شاه آبادی و محمدرضا، میبدی.(۱۳۸۲). *الگوریتم‌های مرتب سازی جدید برای اتوماتای سلولی دو بعدی*. کنفرانس ملی سالانه انجمن کامپیوتر ایران.

References

A.H. Osman, N. S. (2011). Conceptual similarity and graph -based method for plagiarism detection. *Journal of Theoretical and Applied Information Technology*, 32(2), 135-145.

A.H. Osman, N. S. (2012). *An improved plagiarism detection scheme based on semantic role label*ing. 12, 1493-1502.

A.Z, B. (1997). *On the resemblance and containment of documents*. in: Compression and Complexity of Sequences Proceedings.

B. Gipp, J. B. (2010). *Citiation based plagiarism detection:a new approach to identify plagiarized work language independently*. 273-274.

B. Gipp, N. M. (2011). *Citation pattern matching algorithms for citation-based plagiarism detection:greedy citation tiling, citation chunking and longest common citation sequence*. Conference: Proceedings of the 2011 ACM Symposium on Document Engineering, Mountain View, CA, USA, 19-22.

D.R. White, M. J. (2004). Sentence-based natural language plagiarism detection. *Journal of Education Resources in Computing*, 4(4), 2-3.

Gelbukh, S. (2009). Computing Similarity Measures for Original WSD Lesk Algorithm. *Advances in Computer Science and Application*, 43, 155-166.

Heintze, N. (1996). *Scalable document fingerprinting*. in:UNIX Workshop on Electronic Commerce, (pp. 191-200).

K.K. Chow, N. S. (2010). *Web based cross language plagiarism detection. in: Second International Conference on Computational Intelligence, Modelling and Simulation, (pp. 199-204).*

Kent, N. C. (2010). Features based text similarity detection. *Journal of Computing*, 2(1), 53-57.

Kriszti, e. (2000). *Document overlap detection system for distributed digital libraries.* in: Proceedings of fifth ACM conference on Digital libraries, (pp. 226-227). San Antonio, TX, United States.

M. Elhadi, A. A.-T. (2008). *Use of text syntactical structures in detection of document duplicates.* in:Digital Information Management Third International Conference on ICDIM.

M. Esmaeilpour, V. N. (2012). *Cellualr Learning Automata for Mining Customer Behavior in Shopping Activity*. 8(4), 2491-2511.

M. Esnaashari, M. M. (2010). Dynamic point coverage problem in wireless sensor networks:a cellular laerning automata approach. *Journal of Ad Hoc and Sensors Wireless Networks*, 10(2-3), 193-234.

Meuschke, N. S. (2019). *Improving Academic Plagiarism Detection for STEM Documents by Analyzing Mathematical Content and Citations.* ACM/IEEE-CS Joint Conf. on Digital Libraries (JCDL).

Mohamed, M. &. (2019). SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis. *Information Processing & Management*, 56(4), 1356-1372.

S. Alzahrani, N. S. (2010). *Fuzzy Semantic-based Sring Similarity for Extrinsic Plagiarism Detection*. CLEF(Notebook papers/LABs/Workshops).

Savargiv, M., Masoumi, B., & Keyvanpour, M. R. (2020). A new ensemble learning method based on learning automata. *Journal of Ambient Intelligence and Humanized Computing*, 1-16.

Sindhu.L, B. T. (2011). A Study of Plagiarism Detection Tools and Technologies. *Interrnational Journal of Research In Technology*, 1(1), 64-70.

Thatha, V. N. (2020). *An Enhanced Feature Selection for Text Documents.* In Smart Intelligent Computing and Applications, 21-29.

The Stanford NLP Group. (2014). Retrieved from https://nlp.stanford.edu/software/lex-parser.shtml#Download

Virmani, D. &. (2019). *A text preprocessing approach for efficacious information retrieval*. In Smart Innovations in Communication and Computational Sciences, 13-22.

Zhang, F. F. (2019). *Construction site accident analysis using text mining and natural language processing techniques*. Automation in Construction, 238-248.

References [In Persian]

Yaghobi, R., A & khotanloue, H. (2015). Plagiarism detection in the scientific papers using semantic role labeling and Genetic algorithm. *Electronics Industries*, 6(3),67-79 .[In Persian]

Shahabadi M., & Meybodi, M. R. (2003). *New sorting algorithms for two-dimensional cellular automation.* Annual National Conference of the Iranian Computer Association.[In Persian]

استناد به این مقاله: یعقوبی، رضوان، یعقوبی، مهدی، ختن لو، حسن. (۱۴۰۰). رویکردی جدید برای شناسایی سرقت ادبی با استفاده از آتوماتای یادگیر سلولی و برچسب‌گذاری نقش معنایی، *مطالعات مدیریت کسب وکار هوشمند*، ۹(۳۶)، ۱۸۳-۲۰۸.

DOI: 10.22054/IMS.2021.49415.1661