سامانه‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)

**Fall 2024, 5 (3), 115-130**
**DOR:**

**Received: 14 July 2024**
**Accepted: 27 Aug 2024**

**مقاله پژوهشی**

# A Comprehensive Review of Semantic Text Similarity for Developing Intelligent Judicial Recommendation Systems

**Nasim Gereilinia[1], Neda Abdolvand[2*], Zahra Rezaei[3], Farzad Movahedi Sobhani[4]**

1. PhD Student, Department of Industrial Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.
2. Associate Professor, Department of Management, Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran. **Corresponding Author*
3. Assistant Professor, Department of Future Studies, Judiciary Research Institute, Tehran, Iran.
4. Assistant Professor, Department of Industrial Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.

## Abstract

Intelligent judicial recommendation systems, aimed at enhancing the efficiency and consistency of judicial decision-making, rely on the ability to accurately measure semantic similarity between legal texts. This research investigates various methods for evaluating semantic similarity that can be employed in the development of judicial recommendation systems. The analysis covers a wide range of text similarity detection techniques, from basic vocabulary-based approaches to advanced natural language processing and machine learning models. By considering the unique characteristics of legal language and the challenges inherent in this domain, this study carefully analyzes the strengths and weaknesses of each method. In fact, by providing a comprehensive overview of the most advanced methods for semantic similarity of legal texts, this research paves the way for the design and development of powerful and effective intelligent decision support tools for the judicial system. The development of such tools will significantly assist judges and lawyers and play a crucial role in promoting procedural uniformity within the judicial system**.**

**Introduction:** As technology advances and governments become more digitized, a large volume of electronic structured and unstructured data is being generated in the judicial system. Analyzing this data can lead to improved operations and decision-making in the judicial system. The development of intelligent judicial recommendation systems, powered by semantic text similarity techniques, can revolutionize the way legal professionals work. Such systems can streamline legal research, improve the accuracy of judicial decisions, and ultimately enhance the administration of justice. In the common law system, case matching plays an important role and determines the outcome of the judgment of new cases. For this reason, legal experts often spend a lot of time finding similar cases. Therefore, finding similar cases in this legal system is of great importance. Finding similar cases in courts is also a major challenge in the field of legal document review. The length and complex structure of legal documents, as well as the existence of multiple legal issues in a single case, create this challenge. In the judicial system of the Islamic Republic of Iran, crimes are also graded, and crimes of similar degree should have similar punishments. In cases where there is a sharp difference of opinion between rulings, a unanimous ruling is issued. The purpose of this paper is to review methods of semantic similarity in legal texts. In studies conducted, various methods have been used for semantic similarity between texts. In recent years, most studies have focused on the use of deep learning algorithms on text data, the results of which have been better than shallow methods

**Method**: This paper reviews various methods of semantic similarity in legal texts with a focus on deep learning.

**Result**: Shallow methods for evaluating semantic similarity, despite their simplicity and efficiency, are unable to understand the complex patterns present in legal texts. In contrast, deep learning-based methods, with their high power of extracting complex concepts from texts, are recognized as the leaders in this field. Recent studies have shown that in most cases, deep learning-based methods outperform shallow approaches in semantic similarity analysis of legal texts.

**Keywords:** NLP, Deep Learning, Legal Text, Semantic Similarity**,** Intelligent recommendation systems.

سامانه‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)
انجمن علمی تجارت الکترونیکی ایران
واحد زنجان

# مروری جامع بر شباهت‌یابی معنایی متون برای توسعه سیستم‌های توصیه‌گر هوشمند قضایی

تاریخ دریافت: ۲۴/ ۰۴/ ۱۴۰۳
تاریخ پذیرش: ۰۶/ ۰۶/ ۱۴۰۳

دورهٔ پنجم، پاییز ۱۴۰۳
شمارهٔ سوم، صص: ۱۱۳-۱۳۰

**نسیم گریلی‌نیا[۱]، ندا عبدالوند[۲*]، زهرا رضایی[۳]، فرزاد موحدی‌سبحانی[۴]**

۱. دانشجوی دکتری رشته مهندسی صنایع، دانشکده فنی و مهندسی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران.
۲. دانشیار، گروه مدیریت، دانشکده علوم اجتماعی و اقتصادی، دانشگاه الزهرا، تهران، ایران. (نویسندهٔ مسئول)
۳. استادیار، گروه آینده‌پژوهی، پژوهشگاه قوه قضاییه، تهران، ایران.
۴. استادیار، گروه مهندسی صنایع، دانشکده فنی و مهندسی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران.

**چکیده:** سیستم‌های توصیه‌گر هوشمند قضایی با هدف بهبود کارایی و ثبات تصمیم‌گیری قضایی، بر سنجش دقیق شباهت معنایی بین متون حقوقی متکی هستند. در این پژوهش روش‌های مختلف ارزیابی شباهت معنایی که می‌تواند در توسعه سیستم‌های توصیه‌کننده قضایی مورد استفاده قرارگیرد، بررسی شده‌است. این تجزیه و تحلیل طیف وسیعی از تکنیک‌های تشخیص شباهت متن، از رویکردهای مبتنی بر واژگان تا پردازش زبان طبیعی پیشرفته و مدل‌های یادگیری ماشین را پوشش می‌دهد. این بررسی جامع با در نظر گرفتن ویژگی‌های خاص زبان حقوقی و چالش‌های موجود در این حوزه، نقاط قوت و ضعف هر روش را به‌دقت تحلیل‌کرده‌است. در واقع این پژوهش با ارائه یک نمای کلی از پیشرفته‌ترین روش‌های شباهت‌یابی معنایی متون قضایی، زمینه را برای طراحی و توسعهٔ ابزارهای پشتیبانی تصمیم‌گیری هوشمند قوی و مؤثر برای سیستم قضایی فراهم‌می‌کند. توسعه این ابزارها کمک شایانی به قضات و وکلا خواهدکرد و همچنین نقش مهمی در ایجاد وحدت رویه در سیستم قضایی خواهدداشت.

**واژه‌های کلیدی:** پردازش زبان طبیعی، یادگیری عمیق، متون قضایی، شباهت‌یابی معنایی، سیستم‌های توصیه‌گر هوشمند.

## ۱. مقدمه

با پیشرفت فناوری و الکترونیک شدن دولت‌ها، حجم بالایی از داده‌های ساخت‌یافته و ساخت‌نیافته الکترونیکی در سیستم قضایی تولیدمی‌شود. این داده‌ها شامل اطلاعات شاکیان، متهمان، و پرونده‌های حقوقی ثبت و ضبط شده‌است. تحلیل این داده‌ها می‌تواند به بهبود عملیات و تصمیم‌گیری در سیستم قضایی منجر شود [۱].

سیستم‌های توصیه‌گر هوشمند قضایی ابزارهایی مبتنی بر هوش مصنوعی هستند که برای کمک به متخصصان حقوقی و قوه قضاییه در فرآیندهای تصمیم‌گیری طراحی شده‌اند. این سیستم‌ها در هسته خود از تکنیک‌های پردازش زبان طبیعی و یادگیری ماشینی پیشرفته برای تجزیه و تحلیل مخازن گسترده اسناد حقوقی، قانون قضایی و سایر داده‌های متنی مرتبط استفاده‌می‌کنند. یکی از قابلیت‌های کلیدی سیستم‌های توصیه‌کننده قضایی، توانایی اندازه‌گیری دقیق شباهت معنایی بین متون مختلف قانونی است. شباهت بین اسناد قضایی، مانند قانون، دادنامه، اساسنامه، و خلاصه حقوقی، برای کارهایی مانند شناسایی سابقه، بازیابی رویه قضایی، و پیش بینی نتایج پرونده بسیار مهم است. با مدل‌سازی روابط مفهومی و زمینه‌ای بین متون حقوقی، سیستم‌های توصیه‌گر می‌توانند موارد قبلی مرتبط را نشان‌دهند، استدلال‌های قانونی مشابه بالقوه را پیشنهادکنند، و بینش‌های ارزشمندی را برای قضات و وکلا در راستای تصمیم‌های خود ارائه‌کنند. بنابراین، کارایی روش‌های تشخیص شباهت به‌کاررفته، عاملی حیاتی در اثربخشی سیستم‌های هوشمند قضایی است.

در نظام قضایی حقوق عرفی، تطبیق پرونده‌های مشابه نقش مهمی دارد و نتایج قضاوت پرونده‌های جدید را تعیین‌می‌کند. به همین دلیل، متخصصان حقوقی معمولاً زمان زیادی را برای یافتن پرونده‌های مشابه صرف‌می‌کنند. بنابراین، یافتن پرونده‌های مشابه در این سیستم حقوقی اهمیت زیادی دارد [۲]. یافتن پرونده‌های مشابه در دادگاه‌ها یک چالش مهم در حوزه بررسی اسناد حقوقی است. طولانی بودن و ساختار پیچیده اسناد حقوقی و همچنین وجود مسائل حقوقی متعدد در یک پرونده واحد، این چالش را ایجادمی‌کند [۳]. در نظام قضایی جمهوری اسلامی ایران نیز، جرایم بر اساس درجه دسته بندی می‌شوند و جرایم با درجه مشابه باید مجازات یکسانی داشته باشند. در مواردی که اختلاف نظر شدید بین آراء وجوددارد، رای وحدت رویه ارائه‌می‌شود. صدور آراء متفاوت از دادگاه‌ها در یک موضوع مشابه به جهت تفاوت‌های اصحاب دعوا، طبیعی است، اما نظام دادرسی باید تا حد امکان از بروز نابرابری در خروجی دادرسی جلوگیری‌کند و به اجرای عدالت و تامین نظم مورد انتظار جامعه تمایل داشته‌باشد [۴]. بنابراین با استفاده از شباهت یابی در متون قضایی می‌توان این متون را با یکدیگر مقایسه‌کرده و شباهت‌های آن‌ها را شناسایی کرد. این کاربرد می‌تواند در تحلیل و بررسی قضاوت‌ها و تفسیرات مختلف حقوقی مفید باشد [۵]. در واقع نظام حقوقی ایران با وجود ریشه در نظام حقوقی رومی-ژرمنی [۶]، دارای ویژگی‌های منحصر به فردی است که آن را از سایر نظام‌های حقوقی متمایز می‌کند. یکی از این ویژگی‌ها تعیین سقف و کف مشخص برای احکام صادره در دادنامه است. در نتیجه استفاده از شباهت‌یابی متون قضایی در شناسایی مشخصه‌های مهم پرونده و دسته‌بندی سطح جرم و حکم مؤثر است.

در مطالعات انجام شده، روش‌های مختلفی برای شباهت‌یابی معنایی متون استفاده شده‌است. در سال‌های اخیر، اکثر مطالعات بر استفاده از الگوریتم‌های یادگیری عمیق در پردازش متون تمرکز کرده‌اند و نتایج حاکی از آن است که در بیشتر موارد، عملکرد روش‌های مبتنی بر یادگیری عمیق نسبت به روش‌های سطحی بهتر است [۷] و [۸]. لذا در این مقاله با توجه به اهمیت شباهت‌یابی متون قضایی در نظام‌های قضایی مختلف از جمله نظام قضایی حقوق عرفی و همچنین نظام قضایی ایران و از طرفی پیچیدگی‌های متون قضایی، مروری جامع از روش‌های شباهت‌یابی متون قضایی ارائه شده‌است. با توجه به پیشرفت های اخیر در زمینه روش های یادگیری عمیق و عملکرد درخشان آن‌ها در شباهت‌یابی متون، در این مقاله تلاش‌های انجام‌شده در حوزه شباهت‌یابی متون قضایی در دو بخش روش‌های سطحی و روش‌های یادگیری عمیق مورد بررسی قرارگرفته‌است.

ساختار مقاله در ادامه به شرح ذیل می‌باشد: در بخش ۲ به متدولوژی جستجو پرداخته شده‌است و در ادامه در بخش ۳ تاریخ نگار پژوهش‌های انجام‌شده، ارائه شده‌است و بخش ۴ تحلیلی از نقشه علم‌سنجی شباهت‌یابی متون است. در بخش ۵ به سیستم‌های توصیه‌گر و نقش شباهت‌یابی در این سیستم‌ها پرداخته شده‌است. در بخش ۶ به بررسی روش‌های شباهت‌یابی در دو گروه روش‌های سطحی و روش‌های مبتنی بر یادگیری عمیق پرداخته‌شده‌است. بخش ۷ مروری بر تحقیقات پیشین انجام شده‌است، تمرکز این مرور در زمینه شباهت‌یابی متون قضایی است. در بخش۸ فرصت‌ها و محدودیت‌های شباهت‌یابی متون قضایی بررسی شده‌است. و در بخش ۹ به بیان نتایج حاصل از بررسی روش‌های شباهت‌یابی متون حقوقی پرداخته شده‌است و در نهایت در بخش ۱۰به تحقیقات آتی در حوزه شباهت‌یابی معنایی متون قضایی اشاره شده‌است.

## ۲. متدولوژی جستجو

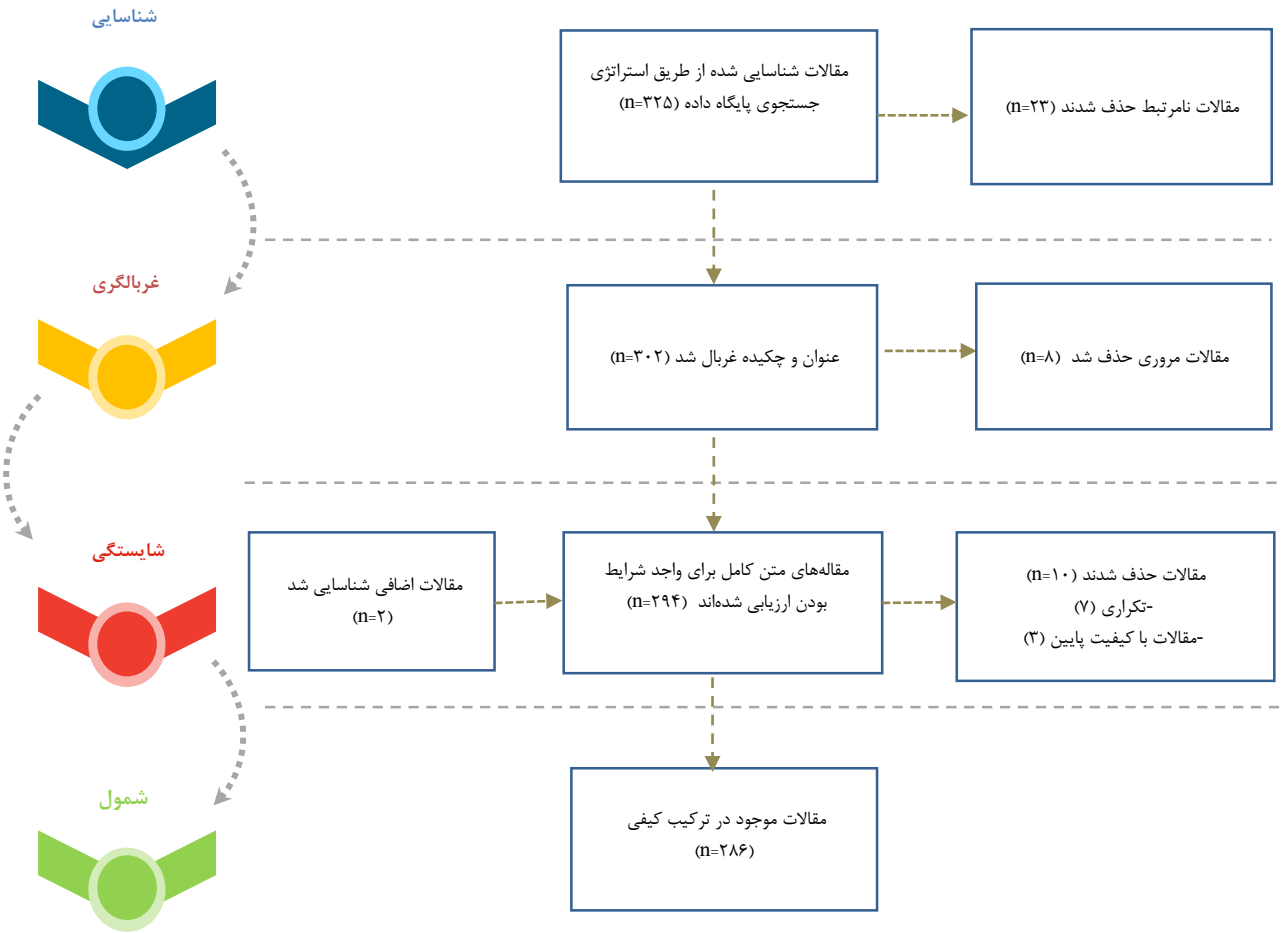
هدف اصلی این تحقیق مروری بر شباهت‌یابی متون قضایی است. برای دستیابی به این هدف، جستجوی گسترده‌ای با استفاده از مجموعه‌ی جامعی از کلیدواژه‌های فهرست‌شده در جدول (۱) در پایگاه داده ScienceDirect و Google Scholar انجام‌شد تا مقالاتی که مستقیما با زمینه تحقیق مرتبط هستند، بازیابی‌شود. در ابتدا ۳۲۵ مقاله استخراج‌شد. سپس ۲۸۶ مقاله برای تجزیه تحلیل دقیق‌تر انتخاب‌شد که شامل اطلاعات به‌روز و مرتبط بودند. این مقالات در سه حوزه کاربردی شباهت‌یابی متون شامل تشخیص سرقت ادبی، شباهت‌یابی متون پزشکی و شباهت‌یابی متون حقوقی و قضایی بود. در نهایت با توجه به هدف پژوهش، ۱۴ مقاله که در حوزه شباهت‌یابی متون قضایی در ۸ سال اخیر بود انتخاب و مورد بررسی دقیق‌تر قرارگرفت.

در شکل ۱ سه مرحله جستجو شامل شناسایی منابع، بررسی ارتباط با موضوع، ارزیابی واجد شرایط بودن و بررسی برای گنجاندن در تحقیق آمده‌است. جدول ۱ خلاصه‌ای از فیلترهای به‌کاررفته در استراتژی جستجو را ارائه‌می‌کند.

**جدول ۱: فیلترهای مورد استفاده در استراتژی جستجو**

| جزئیات | ملاک | مناطق کلیدی جستجو |
|---|---|---|
| Google Scholar ScienceDirect | پایگاه داده | همه مناطق جستجو |
| عنوان و چکیده | فیلد جستجو | همه مناطق جستجو |
| پردازش زبان طبیعی، متن‌کاوی، شباهت‌یابی متون، روش‌های سطحی، روش‌های عمیق | مطالعه | همه مناطق جستجو |
| حقوقی، پزشکی، سرقت ادبی | حوزه | موضوع کلی در مورد شباهت معنایی متون |
| از ۲۰۱۷ تا ۲۰۲۴ | سال | موضوع کلی در مورد شباهت معنایی متون |

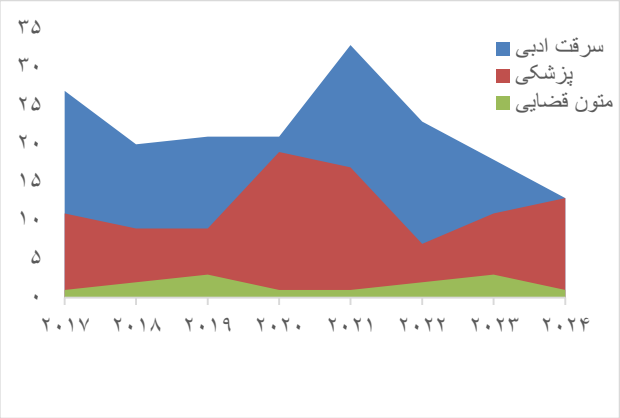
این فیلترها شامل معیارهایی مانند پایگاه داده، فیلد جستجو (عنوان و چکیده) و کلمات کلیدی مورد استفاده برای شناسایی محدودیت‌های بررسی می‌باشد. بررسی بین مقالات منتشرشده از سال ۲۰۱۷ تا ۲۰۲۴ انجام‌شده و حوزه‌های جستجو مرتبط با شباهت‌یابی متون قضایی، شباهت‌یابی متون پزشکی و سرقت ادبی است که روش‌های یادگیری عمیق و روش‌های سطحی را پوشش‌می‌دهد.




**شکل ۱: گردش کار جستجوی مروری**

## ۳. تاریخ‌نگار پژوهش‌های انجام‌شده

در شکل ۲ تعداد مقالات منتشر شده در سه حوزه شباهت‌یابی متون قضایی، شباهت‌یابی متون پزشکی و تشخیص سرقت ادبی از سال ۲۰۱۷ تا ۲۰۲۴ آمده‌است که بیانگر تمرکز بیشتر تحقیقات در حوزه تشخیص سرقت ادبی و شباهت‌یابی متون پزشکی است. با توجه به اهمیت شباهت‌یابی متون قضایی از منظر زبانی و کاربردی، انتظار می‌رود پژوهش‌های متن‌کاوی در آینده به شباهت‌یابی عمیق‌تر متون قضایی در زبان‌های مختلف بپردازند.



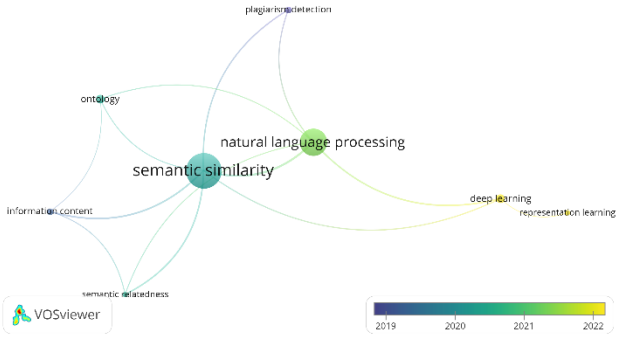

**شکل ۲: پراکندگی انتشارات شباهت‌یابی متون در سه حوزه کاربردی**

## ۴. علم‌سنجی

نقشه علمی شکل ۳ یک نمای کلی از پژوهش‌های انجام‌شده در حوزه شباهت‌یابی معنایی را ارائه‌می‌دهد. مفهوم مرکزی این حوزه، شباهت معنایی است که مستقیماً با مفاهیم پردازش زبان طبیعی و یادگیری عمیق در ارتباط است. این ارتباط نشان‌می‌دهد که برای محاسبه شباهت بین متون، از تکنیک‌های پیشرفته پردازش زبان طبیعی و مدل‌های یادگیری عمیق استفاده‌می‌شود.

یکی از نکات قابل‌توجه در این نقشه، ارتباط نزدیک بین شباهت‌یابی معنایی و مفهوم "سرقت ادبی" است. این ارتباط نشان‌می‌دهد که یکی از کاربردهای مهم شباهت‌یابی معنایی، تشخیص موارد مشابهت بیش از حد بین متون، به‌ویژه در زمینه‌های علمی و آموزشی است. همچنین، مفاهیم دیگری مانند "محتوای غنی" و "نمایش معنایی" نیز در این نقشه برجسته‌شده‌اند. این برجستگی بیانگر این است که محققان به‌دنبال روش‌هایی برای استخراج اطلاعات معنایی از متن‌های پیچیده و غنی هستند.

رنگ‌آمیزی گره‌ها در این نقشه نیز اطلاعات جالبی را ارائه می‌دهد. رنگ زرد در نقشه بیانگر مقالات و تحقیقات جدیدتر است که نشان می‌دهد حوزه شباهت‌یابی معنایی، به‌ویژه با تمرکز بر یادگیری عمیق، در سال‌های اخیر رشد قابل‌توجهی داشته‌است. به‌طور خلاصه، این نقشه یک تصویر کلی از روند تحقیقات در این حوزه ارائه‌می‌دهد و با توجه به اهمیت تشخیص شباهت بین متون در حوزه‌های مختلف مانند پزشکی و حقوقی و بسیار دیگری از حوزه‌ها، انتظارمی‌رود که تحقیقات در این حوزه‌ها در آینده گسترش‌یابد و به توسعه ابزارهای قدرتمندتر برای تحلیل متن منجر شود.




**شکل ۳: نقشه علم‌سنجی شباهت‌یابی متون**

## ۵. سیستم‌های توصیه‌گر

سیستم‌های توصیه‌گر ابزارهای هوشمندی هستند که با تحلیل رفتار، علایق و تاریخچه تعامل کاربران با داده‌ها، اقدام به ارائه پیشنهادات شخصی‌سازی شده، می‌کنند. این سیستم‌ها با استفاده از الگوریتم‌های پیچیده یادگیری ماشین، از جمله یادگیری عمیق، قادرند الگوهای پیچیده و پنهان در داده‌ها را شناسایی‌کرده و بر این اساس، به کاربران پیشنهاداتی ارائه‌دهند که احتمالاً مورد علاقه آن‌ها باشد [۹].

در حوزه قضایی، سیستم‌های توصیه‌گر می‌توانند به عنوان یک دستیار هوشمند برای قضات و وکلا عمل‌کنند. برای مثال، هنگامی که یک پرونده جدید ثبت می‌شود، سیستم می‌تواند با مقایسه آن با پرونده‌های مشابه قبلی، پیشنهاداتی در مورد قوانین مرتبط، آرای قضایی و استدلال‌های حقوقی ارائه‌دهد. درواقع یکی از کلیدی‌ترین تکنیک‌های مورد استفاده در این سیستم‌ها، شباهت‌یابی معنایی متون است. این تکنیک به سیستم اجازه‌می‌دهد تا با مقایسه معنایی متون مختلف، میزان شباهت آن‌ها را تعیین‌کند [۱۰].

شباهت‌یابی متن، به‌ویژه در سیستم‌های توصیه‌گر مبتنی بر تگ، از اهمیت ویژه‌ای برخوردار است. همان‌طور که در پژوهش اخیر بازرگانی و همکاران در سال ۲۰۲۴ نشان داده شده‌است، بهبود دقت در پیشنهاد برچسب‌ها، به‌ویژه در شرایط شروع سرد[1] کاربران، می‌تواند به‌طور قابل‌توجهی عملکرد سیستم را بهبود بخشد. برای مثال، در صورتی که کاربری جدید به پلتفرمی بپیوندد، سیستم توصیه‌گر با چالش شروع سرد مواجه‌می‌شود و نمی‌تواند علایق او را به‌دقت تشخیص‌دهد. بهبود دقت در پیشنهاد برچسب‌ها در این شرایط، می‌تواند به سیستم کمک‌کند تا به‌سرعت علایق کاربر را درک‌کرده و پیشنهادات مرتبط‌تری ارائه‌دهد [۱۱].

در حوزه قضـــایی نیز، بهبود دقت در پیشـــنهاد تگ‌ها برای پرونده‌های جدید می‌تواند منجر به بهبود کارایی قضات و وکلا شود. با استفاده از تکنیک‌های شباهت‌یابی متن پیشرفته، سیستم‌های توصیه‌گر می‌توانند پرونده‌های مرتبط را با دقت بیشتری شناسایی‌کرده و به این ترتیب، به تصمیم‌گیری‌های بهتر و سریع‌تر کمک‌کنند.

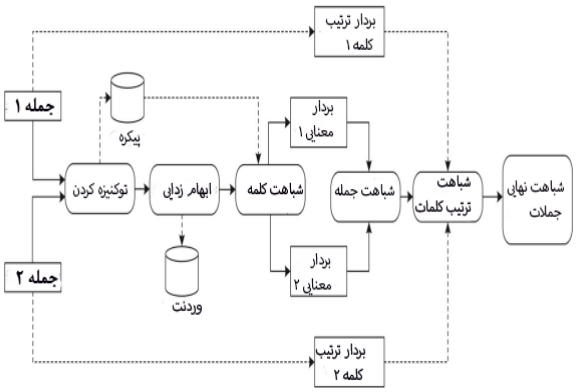
## ۶. روش‌های شباهت‌یابی

ایده تشـــخیص دو ســـند مشـــابه اولین بار در مقاله منبر [۱۲] مطرح شده‌است. دو سند مشابه هستند اگر تعداد واژه‌های مشترک بین آن‌ها چه از نظر معنایی و چه از نظر لغوی از آســـتانه تعریف‌شـــده‌ای بیشـــتر باشـــد [۱۳]. به‌طور کلی می‌توان شـــباهت‌یابی را در دو حوزه شـــباهت لغوی و شـــباهت معنایی بررســـی‌کرد. کلماتی که یک متن را تشـــکیل‌می‌دهند از نظر واژگانی مشـــابه هســـتند اگر توالی کاراکترهای مشـــابهی داشـــته‌باشـــند [۱۴]. شـــباهت‌یابی معنایی میزان هم ارزی معنایی بین دو کلمه، جمله یا ســـند را پیدامی‌کند [۱۵]. در حال حاضر، روش‌های مختلفی برای محاسبه شباهت معنایی بین دو سند متنی وجوددارد.

برخی تکنیک‌ها مانند کیسه کلمات و فراوانی اصطلاح- معکوس فراوانی متن[۲] برای معرفی متن به صـــورت بردارهای اعداد حقیقی به منظور محاسبه شباهت معنایی استفاده‌می شوند، البته این روش‌ها به این حقیقت توجه ندارند که کلمات، معانی مختلفی دارند و کلمات مختلف می‌توانند به یک مفهوم مشابه اشاره‌کنند. در حقیقت این روش ها جنبه‌های لغوی متن را در نظر می‌گیرند و به‌راحتی قابل پیاده سازی هستند اما همان‌طور که اشاره‌شد دارای نقاط ضعف می‌باشند، به منظور رفع این نقاط ضعف، روش‌های شباهت‌یابی معنایی بسیاری در سه دهه اخیر معرفی‌شده‌اند. روش‌های شباهت‌یابی معنایی غالباً به جای تصمیم صفر و یک در این خصوص که آیا دو متن شبیه به هم هستند یا خیر، یک رتبه‌بندی و یا درصدی از شباهت بین متون را ارائه‌می‌دهند . البته در ســـال‌های اخیر روش‌های مبتنی بر یادگیری عمیق نیز در شـــباهت یابی معنایی عملکرد درخشانی داشته‌است.

همچنین بســـیاری از محققان به منظور بهره‌برداری از مزایای روش‌های مختلف به ساخت مدل‌های ترکیبی روی آورده‌اند [۸].

به عنوان نمونه در شـــکل شـــماره (۴) یک رویکرد ترکیبی برای شباهت‌یابی جملات نشان داده شده‌است. این رویکرد در سال ۲۰۱۸ توســـط پاوار و مگو معرفی‌شـــد [۱۶]. در این رویکرد برای محاســـبه شباهت جملات هم به معنای کلمات و هم به ترتیب قرارگیری آن‌ها در جمله توجه‌می‌شود.




**شکل ٤ : رویکرد شباهت‌یابی جملات [۱۷]**

این رویکرد (شکل ۴) با تقسیم جمله‌های ورودی به لیستی از نشانه‌ها (توکنیزه‌کردن) آغازمی‌شـــود. ســـپس، برای برچســـب‌گذاری تک‌تک نشـــانه‌ها/کلمات بر اســـاس نوع کلمه آن‌ها، تگ‌گذاری دســـتوری انجام‌می‌گیرد. برای هر جمله، یک بردار معنایی ســـاخته‌می‌شـــود که مقدار تشـــابه کلمه را برای هر کلمه با تمام کلمات دیگر از جمله دوم (برای مقایسه) در خود جای‌می‌دهد. محاسبه تشابه کلمه با استفاده از وردنت به عنوان یک شـــبکه معنایی انجام‌می‌شـــود. این محاســـبه با در نظرگرفتن کوتاه‌ترین مسیر بین کلمات و عمق کمترین جد مشترک آن‌ها در وردنت به عنوان یک ساختار سلسله‌مراتبی اندازه‌گیری‌می‌شود. این فرایند ســـاخت بردار معنایی همچنین شـــامل محتوای اطلاعاتی استخراج شده از وردنت به عنوان یک بدنه است. این روش، شباهت معنایی را از این دو بردار معنایی محاســـبه‌می‌کند. به عنوان یک قابلیت اختیاری، می‌توان یک بردار ترتیب کلمه برای محاسبه شباهت ترتیب کلمه تشکیل‌داد. در نهایت، شباهت جمله با ترکیب شباهت معنایی و ترتیب کلمه اندازه‌گیری‌می‌شود [۱۶].

بر اســـاس مفهوم شـــباهت یابی متون، روش های مختلفی برای اندازه‌گیری درجه شـــباهت متون توســـعه یافته‌اند. در ادامه روش‌های شباهت‌یابی متون در دوبخش روش‌های سطحی و روش‌های مبتنی بر یادگیری عمیق مورد بررســـی قرارمی‌گیرد. لازم به ذکر اســـت که برخی از این روش‌ها در سطح بررسی کلمه، برخی در سطح بررسی جمله و برخی دیگر در ســـطح بررســـی متن کاربرد دارند. البته طبق شـــکل ۵ نشـــان داده‌شـــده‌اســـت برخی از روش‌ها در چند ســـطح مختلف کاربرد دارند.

### ۱,۶. روش‌های سطحی[۳]

روش‌های سطحی شباهت‌یابی متون بر اساس مقایسه ویژگی‌های سطحی متون مانند تعداد کلمات مشترک، توالی کلمات، فراوانی-n gram ها و شاخص‌های شباهت مانند جاکارد عمل‌می‌کنند. همان‌طور که اشاره‌شد این روش‌ها در سه سطح بررسی کلمه، جمله و سند به بررسی شباهت می‌پردازند. در ابتدا روش‌هایی را مورد بررسی

قرارمی‌دهیم که شباهت را در سطح کلمه بررسی‌می‌کنند. این روش‌ها عبارتند از:

**شباهت کوسینوسی**[4]: در سنجش شباهت بین دو بردار، از روش‌های مختلفی استفاده‌می‌شود که یکی از رایج‌ترین آن‌ها، روش شباهت‌یابی کسینوسی است. این روش در حوزه‌های متعددی مانند پردازش زبان طبیعی، بازیابی اطلاعات و بینایی ماشین کاربرد دارد. در روش شباهت‌یابی کسینوسی، زاویه بین دو بردار محاسبه‌می‌شود.

هرچه زاویه بین دو بردار کوچکتر باشد، شباهت آن‌ها بیشتر خواهدبود. برای محاسبه این شباهت، از فرمول زیر استفاده‌می‌شود [۱۸].

$$sim(s_a, s_b) = cos\Theta = \frac{\overrightarrow{s_a}.\overrightarrow{s_b}}{\|s_a\|.\|s_b\|} \qquad (۱)$$

**فاصله اقلیدسی**: از نظر ریاضی فاصله اقلیدسی خط مستقیم بین دو نقطه در فضای اقلیدسی است. یک فاصله اقلیدسی برابر با ۰ نشان‌دهنده این است که دو بردار یکسان هستند، در حالی که یک فاصله اقلیدسی بیشتر نشان‌دهنده این است که دو بردار از هم دورتر هستند. این فاصله طبق فرمول زیر محاسبه‌می‌شود [۱۸].

$$d(s_a, s_b) = \sqrt{\sum_{i=1}^{n}(s_a^{(i)} - s_b^{(i)})^2} \qquad (۲)$$

**شباهت جاکارد:** این روش، شباهت بین دو مجموعه از کلمات را محاسبه‌می‌کند. شباهت جاکارد، نسبت تعداد کلمات مشترک دو مجموعه به تعداد کل کلمات دو مجموعه است .یک شباهت جاکارد برابر با ۱ نشان‌دهنده این است که دو مجموعه یکسان هستند، در حالی که یک شباهت جاکارد برابر با ۰ نشان‌دهنده این است که دو مجموعه هیچ کلمه‌ای مشترک ندارند [۱۹].

**شباهت مبتنی بر وردنت**[5]: در این روش از یک پایگاه داده لغوی، برای اندازه‌گیری ارتباط معنایی بین کلمات استفاده‌می‌شود. وردنت یک پایگاه داده بزرگ از کلمات و روابط آن‌هاست. شباهت وردنت بین دو کلمه، نشان‌دهنده میزان نزدیکی ارتباط معنایی دو کلمه است [۲۰].

**محتوای اطلاعاتی**[6]: این روش بر اساس تئوری اطلاعات عمل می‌کند و میزان اطلاعات مشترک بین دو متن را محاسبه می‌کند. این روش، مقدار اطلاعاتی که یک کلمه حمل می‌کند را اندازه‌گیری می‌کند. محتوای اطلاعاتی یک کلمه، نشان‌دهنده میزان تعجب ما در دیدن آن کلمه در یک متن است. کلمات با محتوای اطلاعاتی پایین، کلمات رایج هستند، در حالی که کلمات با محتوای اطلاعاتی بالا، کلمات نادر هستند. شباهت مبتنی بر اطلاعات بین دو کلمه، نشان‌دهنده میزان اشتراک اطلاعاتی آن‌ها است [۲۱].

**فاصله لون‌اشتاین**[7]: این روش، حداقل تعداد ویرایش‌های تک‌کاراکتری (درج، حذف، جایگزینی) را که لازم است برای تبدیل یک رشته به رشته دیگر انجام‌شود، اندازه‌گیری‌می‌کند. این فاصله نشان‌دهنده میزان شباهت دو رشته از نظر املای آن‌ها است [۲۲].

**فاصله جارو- وینکلر**[8]: این روش، شباهت بین دو رشته را با درنظرگرفتن تطابق کاراکترها و انتقال‌ها محاسبه‌می‌کند. این روش به رشته‌هایی با پیشوندهای مشابه امتیاز بالاتری می‌دهد، که آن را برای تشخیص تغییرات املایی و کپی‌های ناقص مفید می‌کند [۲۳].

**فاصله همینگ**[9]: این روش، تفاوت بین دو رشته با طول مساوی را با شمارش تعداد موقعیت‌هایی که کاراکترهای مربوط با هم متفاوت هستند، اندازه‌گیری‌می‌کند [۲۴].

**شباهت‌یابی برداری مبتنی بر ماتریس هم‌بستگی لغوی**[10]: روشی برای محاسبه شباهت معنایی بین کلمات است. این روش بر اساس یادگیری ماشین عمل‌می‌کند و از ماتریس‌های هم‌بستگی لغوی برای یادگیری رابطه بین کلمات استفاده‌می‌کند [۲۵].

**مدل فضای برداری**[11]: این روش برای نمایش و سنجش شباهت معنایی بین کلمات است. در این روش به هر کلمه یک بردار عددی اختصاص داده‌می‌شود. این بردار، بر اساس تعداد دفعاتی که آن کلمه در یک مجموعه متن بزرگ ظاهرشده، ساخته‌می‌شود. در نهایت شباهت معنایی بین دو کلمه با محاسبه‌ی فاصله بین بردارهای آن‌ها اندازه‌گیری‌می‌شود. به عبارت دیگر، روش مدل فضای برداری مستقیما به ساختار جمله یا سند توجه‌نمی‌کند، بلکه بر روی روابط آماری بین کلمات در یک مجموعه متن بزرگ تمرکز دارد .این موضوع باعث می‌شود که این روش برای کارهایی مانند دسته‌بندی متون، بازیابی اطلاعات و ترجمه ماشینی که در آن‌ها بر روی معنای کلی کلمات در متن تمرکز می‌شود، مناسب باشد [۲۶]..

**تحلیل ارتباطی نهفته**[12]: این روش برای اندازه‌گیری روابط معنایی بین کلمات است. این روش درواقع، توسعه‌یافته مدل فضای برداری است که برای غلبه بر مشکلات موجود در اندازه‌گیری روابط معنایی بین جفت کلمات به‌کارمی‌رود [۲۷].

برخی دیگر از روش‌های شباهت‌یابی سطحی در سطح جمله به بررسی شباهت می‌پردازند. این روش‌ها عبارتند از:

**کوتاه‌ترین طول مسیر**: روش کوتاهترین طول مسیر یکی از روش‌های متداول برای سنجش شباهت معنایی بین جملات است. در این روش، شباهت دو جمله با محاسبه کوتاهترین مسیر بین گره‌های مربوط به کلمات در آن جملات تعیین‌می‌شود. هرچه طول مسیر کوتاه‌تر باشد، نشان‌دهنده شباهت بیشتر بین دو جمله است. مزیت این روش سادگی و کارایی آن است. با این حال، معایبی نیز دارد، از جمله عدم درنظرگرفتن روابط معنایی عمیق‌تر بین کلمات و عدم در نظر گرفتن وزن کلمات .به همین دلیل، روش کوتاهترین طول مسیر به تنهایی برای سنجش دقیق شباهت معنایی جملات کافی نیست و غالباً در کنار سایر روش‌های شباهت‌سنجی استفاده‌می‌شود [۲۸].

**کوتاه‌ترین طول مسیر وزن‌دار**: این روش بر اساس روش کوتاه‌ترین طول مسیر بنا شده‌است و ضعف‌های آن را برطرف می‌کند.

در این روش، به هر ارتباط (لینک) بین کلمات یک وزن اختصاص داده‌می‌شود. این وزن نشان‌دهنده درجه نزدیکی و اهمیت آن ارتباط است. سپس شباهت جملات با در نظر گرفتن این وزن‌ها و محاسبه کوتاهترین مسیر بین کلمات مشابه در دو جمله سنجیده‌می‌شود.

به‌عبارتی، هرچه مسیر کوتاه‌تر و وزن لینک‌های آن بیشتر باشد، شباهت معنایی بین جملات بالاتر است. این روش نسبت به روش کوتاهترین طول مسیر کارآمدتر است، چون اهمیت روابط بین کلمات را در نظر می‌گیرد [۲۹].

**مقیاس‌گذاری نسبی عمیق:** در این روش، شباهت دو جمله بر اساس روابط موجود در یک طبقه‌بندی و با در نظر گرفتن عمق مفاهیم در آن طبقه‌بندی محاسبه‌می‌شود[۳۰]. برای این منظور، هر جمله به صورت یک گراف معنایی که در آن مفاهیم به عنوان گره و روابط بین آن‌ها به عنوان لینک نمایش داده‌می‌شوند، تبدیل‌می‌شود. سپس از طریق الگوریتم مقیاس‌گذاری نسبی عمق، شباهت بین گره‌های هم‌معنا در دو جمله محاسبه‌می‌شود. با جمع‌آوری این شباهت‌های موضعی بین گره‌ها، شباهت کلی دو جمله در سطح معنایی به‌دست‌می‌آید [۲۹].

**الگوریتم اسمیت-واترمن**[۱۳]: این روش یک الگوریتم برنامه‌نویسی پویا است که همترازی محلی دو رشته را محاسبه‌می‌کند. این الگوریتم معمولاً در زیست‌شناسی مولکولی برای همترازی توالی و جستجو در شباهت استفاده‌می‌شود. این الگوریتم می‌تواند برای شباهت‌یابی متون با در نظر گرفتن ترتیب کلمات، غلط‌های املایی و نگارشی و نویز موجود در متن استفاده‌شود [۳۱].

**طولانی‌ترین زیررشته مشترک**[۱۴]: این روش طولانی‌ترین رشته کاراکترهای مشترک بین دو رشته را پیدامی‌کند .این الگوریتم بر اساس طول طولانی‌ترین زیر رشته مشترک، شباهت را اندازه‌گیری می‌کند و در کاربردهای مختلف مانند تشخیص سرقت ادبی و مقایسه توالی DNA استفاده می‌شود [۳۲].

علاوه بر روش‌های مذکور برخی روش‌های شباهت‌یابی سطحی مانند روش‌های ذیل در سطح سند به بررسی شباهت می‌پردازند.

**تحلیل معنایی نهفته**[۱۵]: این روش، بر اساس تحلیل ماتریس برداری اسناد عمل‌می‌کند و مفاهیم نهفته در متون را استخراج‌می‌کند، سپس بر اساس این مفاهیم، شباهت بین متون را محاسبه‌می‌کند [۳۳].

**تحلیل معنایی نهفته تعمیم‌یافته**: این روش بر اساس رویکرد معنایی نهفته ساخته‌شده‌است [۳۴]. تفاوت اصلی این روش با روش تحلیل معنایی نهفته در این است که تحلیل معنایی نهفته تعمیم‌یافته به جای تمرکز روی کل سند و ماتریس سند-واژه، روی بردارهای واژه (کلمه) تمرکزمی‌کند. به عبارت دیگر، این روش سعی‌می‌کند بردارهایی برای هر واژه بسازد که نشان‌دهنده معنای آن واژه در متن است [۲۹].

**تخصیص دریکله پنهان**[۱۶]: این روش، یک مدل مولد احتمالی است که اسناد را به عنوان مخلوطی از موضوعات نشان‌می‌دهد. روش تخصیص دریکله پنهان، احتمال اینکه هر کلمه در یک سند به هر موضوع تعلق داشته‌باشد را یادمی‌گیرد و شباهت بین دو سند در این روش نشان‌دهنده میزان شباهت دو سند از نظر توزیع موضوعی آن‌ها است [۳۵].

**فراوانی اصطلاح-معکوس فراوانی سند**[۱۷] : این روش بیشتر در سطح سند به بررسی شباهت می‌پردازد. این روش بر اساس دو اصل کلیدی تعداد دفعات ظهور هر کلمه در یک متن و لوگاریتم معکوس تعداد اسنادی که هر کلمه در آن‌ها ظاهر می‌شود، کارمی‌کند. به‌طور خلاصه، در این روش به کلماتی که در یک متن خاص بیشتر ظاهر می‌شوند و در اسناد دیگر کمتر ظاهر می‌شوند، وزن بیشتری داده‌می‌شود [۳۶].

**تحلیل معنایی صریح**[۱۸]: روشی برای سنجش شباهت معنایی بین متون است که از منابع عظیم دانش مانند ویکی‌پدیا و فهرست پروژه دایرکتوری باز[۱۹] استفاده‌می‌کند. این روش، در مقایسه با روش‌های دیگر، بهبود قابل‌توجهی در ارزیابی و سنجش میزان نزدیکی معنایی متون ارائه‌می‌دهد. این روش با استفاده از یک تکنیک یادگیری ماشینی به نام مترجم معنایی، متن زبان طبیعی را به یک دنباله وزنی از مفاهیم موجود در ویکی‌پدیا تبدیل‌می‌کند. مترجم معنایی از یک طبقه‌بندی‌کننده مبتنی بر مرکز جرم استفاده‌می‌کند. بردارهای وزنی این مفاهیم، بردارهای تفسیر نامیده‌می‌شوند. مترجم معنایی، مفاهیم مرتبط با متن را از یک فهرست وارونه بازیابی‌می‌کند و آن‌ها را با وزن مرتبط با آن‌ها ترکیب‌کرده و در نهایت بردار تفسیر متن را می‌سازد [۳۷].

تحلیل معنایی صریح چندزبانه که تعمیمی از روش تحلیل معنایی صریح برای چند زبان است، از منابع مرجع چندزبانه هم‌راستا با سند، مانند ویکی‌پدیا، برای نمایش یک سند به عنوان یک بردار مفهومی مستقل از زبان استفاده می‌کند [۳۸].

### ۲,۶. روش‌های مبتنی بر یادگیری عمیق

در طول دهه گذشته، پردازش زبان طبیعی پیشرفت چشمگیری داشته‌است که علت اصلی آن استفاده از یادگیری عمیق در این حوزه است [۳۹]. روش‌های شباهت‌یابی مبتنی بر یادگیری عمیق از شبکه‌های عصبی مصنوعی برای محاسبه شباهت بین متون استفاده‌می‌کنند. این روش‌ها نیز مانند روش‌های سطحی در سه سطح کلمه، جمله و سند به بررسی شباهت می‌پردازند. البته برخی از این روش‌های شباهت‌یابی مانند روش‌های مبتنی بر ترنسفورمر، شبکه‌های کانولوشنال و شبکه‌های بازگشتی در هر دو سطح جمله و سند کاربرد دارند. در ادامه به معرفی روش‌های شباهت‌یابی مبتنی بر یادگیری عمیق پرداخته‌می‌شود.

**تعبیه کلمه**[۲۰]: روشی مبتنی بر یادگیری عمیق برای شباهت‌یابی متون است. این روش کلمات را به بردارهای عددی با طول ثابت تبدیل‌می‌کند. این بردارها، معنای کلمات را در فضای برداری نشان می‌دهند. تعبیه‌های کلمه از شبکه‌های عصبی مصنوعی برای یادگیری روابط معنایی بین کلمات از طریق متن استفاده‌می‌کنند. این روابط می‌تواند شامل هم‌معنی بودن، متضاد بودن، یا قرارگیری در یک زمینه مشابه

باشد. مزیت اصلی تعبیه کلمه این است که می‌تواند معنای کلمات را به‌طور ضمنی از طریق متن یادبگیرد [۴۰]. روش‌هایی مانند Word2Vec [۴۱] و فست تکست [۴۲] در این گروه قرارمی‌گیرند. روش word2vec کلمات را به بردارهای عددی تبدیل‌می‌کند. این بردارها نمایش معنایی کلمات هستند. هرچه بردار دو کلمه به هم نزدیک‌تر باشد، شباهت معنایی آن دو کلمه بیشتر است [۴۱]. روش فست تکست نیز روشی کارآمد برای محاسبه شباهت بین کلمات و جملات است. این روش از شبکه‌های عصبی عمیق برای یادگیری بردارهای کلمه استفاده‌می‌کند. بردارهای کلمه فست تکست، نمایش‌های عددی از کلمات هستند که معنای کلمات را ذخیره‌می‌کنند [۴۲].

**تعبیه جمله**[۲۱]: روشی برای تبدیل جملات به بردارهای عددی است که معنا و زمینه جملات را رمزگذاری‌می‌کند. این بردارها را می‌توان برای مقایسه جملات و محاسبه شباهت آن‌ها استفاده‌کرد. روش‌های مختلفی برای تعبیه جمله وجود دارد، اما بسیاری از روش‌های مدرن از یادگیری عمیق استفاده‌می‌کنند [۴۳]. که از آن جمله می‌توان به روش برت جمله [۴۴] اشاره‌کرد. روش برت جمله، روش دقیقی برای محاسبه شباهت بین جملات است و می‌تواند برای تبدیل جملات به بردارهای عددی استفاده‌شود. این روش از مدل زبانی بزرگی مانند برت برای تبدیل جملات به بردارهای عددی استفاده‌می‌کند. این بردارها نمایش‌های عددی از معنای جملات هستند. در نهایت شباهت بین جملات با محاسبه میزان تشابه بردارهای آن‌ها محاسبه‌می‌شود.

**تعبیه سند**[۲۲]: بردارهایی عددی هستند که معنای کلی و محتوای یک سند را رمزگذاری می‌کنند. این بردارها را می‌توان برای مقایسه اسناد، دسته‌بندی اسناد، و بازیابی اطلاعات استفاده‌کرد. روش‌های مختلفی برای ایجاد تعبیه سند وجوددارد، اما بسیاری از روش‌های مدرن از یادگیری عمیق استفاده‌می‌کنند [۴۵].

**شبکه‌های سیامی**[۲۳]: شبکه‌های سیامی نوعی شبکه عصبی مصنوعی هستند که از دو شبکه مشابه با وزن‌های مشترک تشکیل-شده‌اند. این شبکه‌ها برای مقایسه دو ورودی مختلف و تعیین شباهت یا تفاوت آن‌ها به‌کارمی‌روند. درواقع، هر یک از شبکه‌های سیامی جداگانه روی یکی از ورودی‌ها عمل‌می‌کنند و سپس خروجی‌های آن‌ها با هم مقایسه‌می‌شوند. این مقایسه می‌تواند برای انجام وظایف مختلفی مانند تشخیص چهره، دسته‌بندی متن، و تطبیق تصاویر به کار رود. یکی از مزیت‌های شبکه‌های سیامی این است که می‌توانند از حجم کمی داده برای آموزش استفاده کنند. همچنین، این شبکه‌ها به دلیل استفاده از وزن‌های مشترک، از نظر محاسباتی کارآمد هستند [۴۶].

**شبکه‌های عصبی کانولوشنال**[۲۴]: نوعی شبکه عصبی عمیق هستند که برای پردازش داده‌های تصویری، صوتی و متنی به کار می‌روند. این شبکه‌ها از لایه‌های متعددی تشکیل‌شده‌اند که هر کدام وظایف خاصی را انجام‌می‌دهند. یکی از مهم‌ترین لایه‌های شبکه‌های عصبی کانولوشنال، لایه کانولوشن است. این لایه از فیلترهایی برای استخراج ویژگی‌های مهم از داده‌ها استفاده‌می‌کند. پس از استخراج ویژگی‌ها، آن‌ها به لایه‌های بعدی شبکه ارسال و در نهایت، طبقه‌بندی یا شباهت‌یابی انجام‌می‌شود [۴۷]. از جمله روش‌های شباهت‌یابی مبتنی بر شبکه‌های عصبی کانولوشنال می‌توان به شبکه‌های کانولوشنال ایستا و پویا اشاره‌کرد. روش‌های ایستا برای متون با طول ثابت و روش‌های پویا برای متون با طول متغییر مناسب است [۴۸].

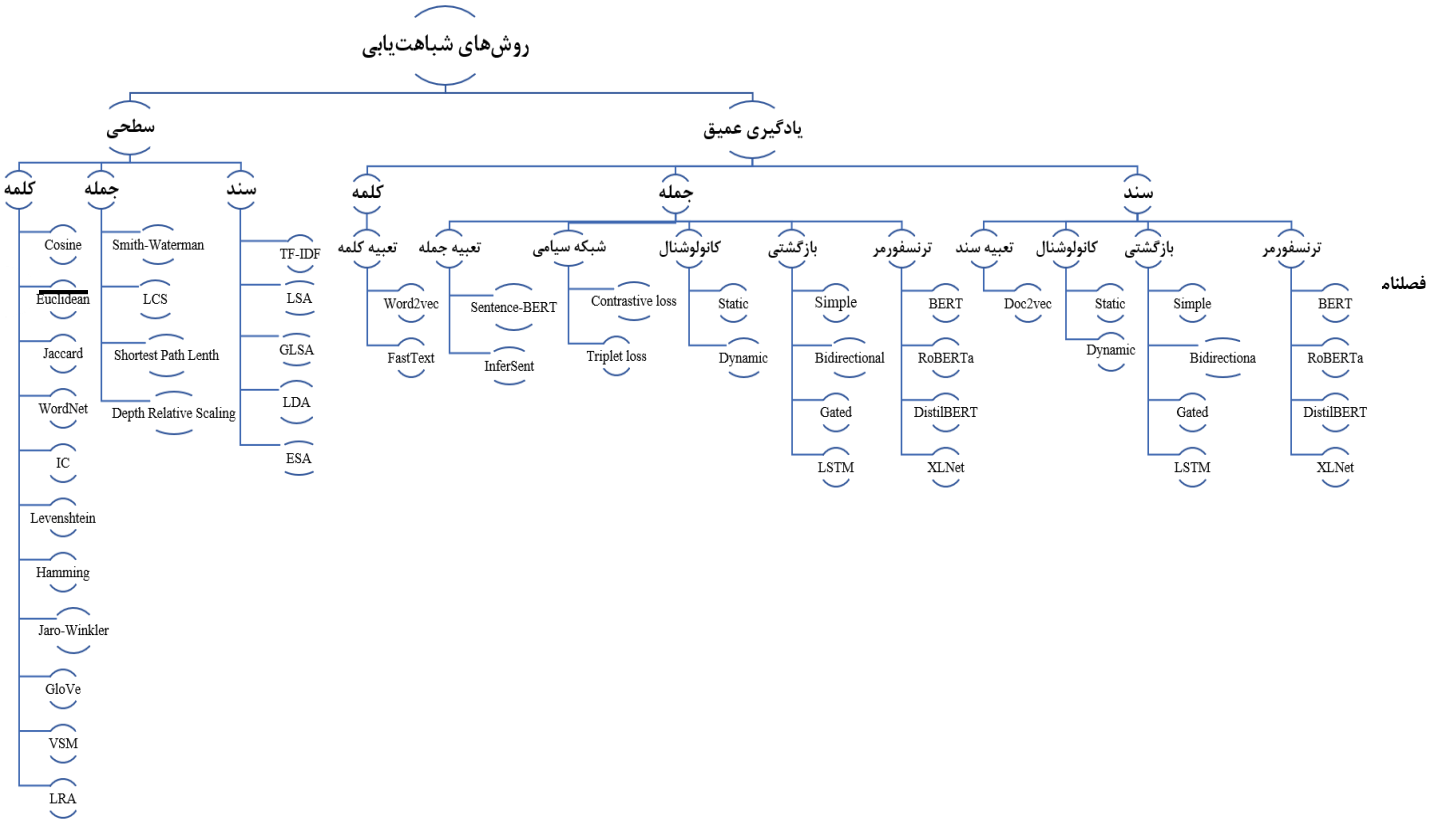
**شبکه‌های عصبی بازگشتی**[۲۵]: شبکه‌های عصبی بازگشتی نوعی شبکه عصبی مصنوعی هستند که برای پردازش داده‌های ترتیبی مانند متن، گفتار و سری‌های زمانی به‌کارمی‌روند. این شبکه‌ها از حلقه‌های بازخورد استفاده‌می‌کنند که به آن‌ها اجازه‌می‌دهد اطلاعات را از ورودی‌های قبلی به خاطر بسپارند و در پردازش ورودی‌های فعلی استفاده‌کنند. این شبکه‌ها به دلیل توانایی یادگیری وابستگی‌های دوربرد بین عناصر یک دنباله، در بسیاری از وظایف پردازش زبان طبیعی مانند ترجمه ماشینی، تشخیص گفتار و تولید متن به‌کارمی‌روند [۴۹]. یکی از انواع این شبکه‌ها، حافظه طولانی‌مدت و کوتاه‌مدت[۲۶] است. این شبکه‌ها نوعی شبکه عصبی بازگشتی هستند که به‌طور خاص برای حل مشکل وابستگی‌های طولانی‌مدت در داده‌ها طراحی‌شده‌اند. درواقع، این شبکه‌ها می‌توانند اطلاعات زمینه‌ای را به طور مؤثر حفظ‌کنند و به همین دلیل، در بسیاری از وظایف پردازش زبان طبیعی مانند تجزیه و تحلیل احساسات و تولید زبان طبیعی، عملکرد بسیار بهتری نسبت به شبکه‌های بازگشتی ساده دارند [۵۰].

**مدل‌های مبتنی بر ترنسفورمر**[۲۷]: مدل‌های مبتنی بر ترنسفورمر نوعی از مدل‌های یادگیری عمیق هستند که برای پردازش زبان طبیعی و به‌طور خاص، برای شباهت‌یابی متون، بسیار کارآمد و مفید هستند. این مدل‌ها برخلاف مدل‌های سطحی که بر اساس دنباله‌ها کار می‌کنند، از مکانیزم توجه استفاده‌می‌کنند که به آن‌ها اجازه‌می‌دهد تا به‌طور مستقیم به روابط بین کلمات در یک متن توجه‌کنند. این مکانیزم توجه، ترنسفورمرها را قادرمی‌سازد تا به‌طور مؤثرتری وابستگی‌های بین کلمات را درک‌کنند و به همین دلیل، در مقایسه با مدل‌های سطحی، دقت و کارایی بالاتری در شباهت‌یابی متون ارائه‌می‌کنند. علاوه بر این، ترنسفورمرها به دلیل ساختار موازی خود، می‌توانند به‌طور کارآمدتر بر روی پردازنده‌های گرافیکی[۲۸] اجراشوند که سرعت پردازش را به‌طور قابل‌توجهی افزایش‌می‌دهد [۵۱]. از جمله مدل‌های مبتنی بر ترنسفورمر می‌توان به روش برت [۵۲] اشاره‌کرد که با استفاده از مکانیزم توجه، به‌طور مستقیم به روابط بین کلمات در یک متن توجه‌می‌کند و شباهت بین متون را با دقت بالایی محاسبه‌می‌کند.

در سال‌های اخیر مدل زبانی برت برای استخراج بردارهای ویژگی عددی کاربرد گسترده‌ای پیداکرده‌است زیرا این مدل‌ها می‌توانند معنای جملات را در نظر بگیرند [۵۳] .

مدل دیگری که در این حوزه وجود دارد ربرتا[۲۹] [۵۴] می‌باشد که نسخه‌ای بهبودیافته از برت است که از آموزش بدون نظارت برای پیش‌آموزش مدل استفاده‌می‌کند .این مدل به‌طور کلی عملکردی بهتر

از برت در وظایف شباهت‌سنجی دارد. دیستیل برت[۳۰] [۵۵] نسخه‌ای فشرده و سبک از برت است که برای استفاده در دستگاه‌های با منابع محدود مناسب است. این مدل عملکردی مشابه برت دارد، اما به محاسبات و حافظه کمتری نیاز دارد. و مدل ایکس ال نت [۵۶] از یک مکانیزم جدید برای توجه به توکن‌ها استفاده‌می‌کند که به آن اجازه-می‌دهد تا وابستگی‌های دوربرد بین توکن‌ها را بهتر از برت و ربرتا یادبگیرد. این مدل به‌طور کلی بهترین عملکرد را در بین این چهار مدل در وظایف شباهت‌سنجی دارد.




**شکل ۵: روش‌های شباهت‌یابی متون**

## ۷. پژوهش‌های پیشین در شباهت‌یابی متون قضایی

در سال‌های اخیر، متون حقوقی به حوزه رو به رشدی در پژوهش‌های شباهت‌یابی بدل‌شده‌اند.

با توجه به جایگاه تحقیقات در حوزه شباهت‌یابی متون حقوقی، تحقیقات انجام‌شده در این حوزه در بازه سال‌های ۲۰۱۷ تا ۲۰۲۴ مورد بررسی قرارگرفت. خلاصه این تحقیقات در جدول (۲) نشان داده‌شده-است.

برخی از این تحقیقات فقط از روش‌های سطحی و برخی از روش‌های یادگیری عمیق و برخی نیز از هر دو روش برای شباهت‌یابی متون قضایی استفاده‌کرده‌اند که در ادامه به بررسی این مقالات پرداخته‌می‌شود.

### ۱،۷ تحقیقات در حوزه روش‌های سطحی

در این بخش تحقیقاتی مورد بررسی قرارمی‌گیرد که از روش‌های سطحی برای شباهت‌یابی متون حقوقی استفاده‌کرده‌اند. البته لازم به ذکر است که در برخی از این تحقیقات علاوه بر روش‌های سطحی از روش‌های مبتنی بر یادگیری عمیق نیز استفاده‌شده‌است و نتایج حاصل از این روش‌ها با یکدیگر مقایسه‌شده‌است.

سال ۲۰۱۷ مندل و همکاران بر بهبود رویکردهای مبتنی بر متن برای محاسبه درجه مشابهت بین دو سند حقوقی تمرکز کردند. در این پژوهش آن‌ها هم از روش‌های سطحی مانند فراوانی اصطلاح-معکوس فراوانی سند و تخصیص دیریکله نهفته و هم از روش‌های یادگیری عمیق مانند روش‌های موجود در تعبیه کلمه و تعبیه سند برای شباهت‌یابی پرونده‌های دادگاه عالی هند استفاده‌کردند [۳]. در این راستا در سال ۲۰۱۸ ما و همکاران [۵۷] رویکرد جدیدی مبتنی بر یادگیری

ماشین برای محاسبه شباهت معنایی آرای قضایی چین معرفی‌کردند. روش آن‌ها از هستی‌شناسی‌های حوزه‌ای برای استخراج مفهوم‌شناسی‌های بنیادی آرای قضایی چین و سپس خلاصه‌سازی آن‌ها در قالب بلوک‌های دانش استفاده‌می‌کند. سپس، الگوریتم فاصله کلمه متحرک[۳۱] برای اندازه‌گیری درجه مشابهت بین این بلوک‌های دانش به کار گرفته‌می‌شود. در سال ۲۰۱۸ نیر و واق [۵۸] استخراج قواعد باهم‌آیی را به عنوان روشی جدید برای استخراج دانش از اسناد حقوقی و تحلیل مرتبط بودن آن‌ها پیشنهادکردند .آن‌ها این تکنیک را روی تحلیل آراء دادگاه‌ها و به‌طور خاص با تمرکز بر استنادات انجام‌شده در آراء به کار بردند. آن‌ها دریافتند که استخراج باهم‌آیی می‌تواند برای شناسایی الگوهایی در استنادات استفاده‌شود که برای تعیین بیشترین ارتباط بین اسناد به‌کارمی‌آید.

## ۲,۷ تحقیقات در حوزه روش‌های یادگیری عمیق

در این بخش به بررسی تحقیقاتی پرداخته‌می‌شود که از روش‌های مبتنی بر یادگیری عمیق برای شباهت‌یابی متون حقوقی استفاده‌کرده‌اند. البته در این بخش نیز مانند بخش قبلی تحقیقاتی وجود دارد که علاوه بر روش‌های مبتنی بر یادگیری عمیق از روش‌های سطحی نیز برای شباهت‌یابی متون حقوقی استفاده کرده و نتایج را با یکدیگر مقایسه‌کرده‌اند.

در سال ۲۰۱۹ رنجیت و آیدیکولا [۵۹] با استفاده از روش تعبیه سند به شباهت‌یابی اسناد قضایی هند پرداختند. در سال ۲۰۱۹ ایکسیا و همکاران [۶۰] رویکرد جدیدی را برای کار با اسناد حقوقی پیشنهادکردند روشی مبتنی بر تعبیه کلمه به منظور افزایش دقت در تحلیل شباهت اسناد حقوقی چین ارائه‌دادند. سال ۲۰۱۹ سوگاتاداسا و همکاران [۶۱] پژوهشی را در زمینه بازیابی اسناد حقوقی با استفاده از مدل‌های تعبیه سند انجام‌دادند. آن‌ها سه مدل نوآورانه شامل یک مدل برداری سند با استفاده از ارجاعات خام سند، یک مدل مبتنی بر شبکه عصبی، و یک مدل مبتنی بر رتبه‌بندی جمله، توسعه‌دادند. این مطالعه بر ماهیت حوزه‌محور بازیابی اطلاعات، به‌ویژه در حوزه حقوق، متمرکز است و به‌دنبال رفع چالش‌های نمایش اسناد پرونده‌های حقوقی در فضاهای برداری مختلف با ادغام معیارهای معنایی واژگان و تکنیک‌های پردازش زبان طبیعی است. این پژوهش همچنین بر اهمیت گنجاندن معیارهای مشابهت معنایی حوزه‌محور در فرآیند بازیابی اطلاعات تأکید دارد. در سال ۲۰۲۰ باتاچاریا و همکاران [۶۲] در مقاله‌ای به محاسبه شباهت بین دو سند حقوقی تمرکزکردند. چالش مطرح در این پژوهش خودکار کردن محاسبه این شباهت است. تا قبل از این پژوهش مقایسه جامعی از روش‌های موجود در یک پلت فرم مشترک وجود نداشت. در این مقاله اولین تحلیل سیستماتیک از روش‌های موجود انجام‌گرفت، علاوه بر این دو روش محاسباتی شباهت جدید مورد بررسی قرارگرفت که یکی مبتنی بر متن و دیگری مبتنی بر تعبیه شبکه بود. در سال ۲۰۲۱ ماندال و همکاران [۶۳] بر اهمیت خودکارسازی فرآیند یافتن سوابق مشابه در پرونده‌های حقوقی برای افزایش کارایی و اثربخشی متخصصان حقوقی تأکیدکردند. نویسندگان با بررسی رویکردهای نظارتی مختلف برای اندازه‌گیری مشابهت متن و شناسایی مناسب‌ترین روش‌ها برای سیستم‌های بازیابی پرونده‌های دادگاه، به این حوزه کمک شایانی کرده‌اند. در این پژوهش علاوه بر روش‌های یادگیری عمیق از روش‌های سطحی نیز برای شباهت‌یابی پرونده‌های دادگاه عالی هند استفاده شد. در سال ۲۰۲۲ شیتال و همکاران [۶۴] به بررسی تشابه بین متون حقوقی پرداخته و از استخراج تشابه موضوعی بر اساس نقش‌های سخنرانی استفاده‌کرده‌اند. در سال ۲۰۲۲ دی اولیویرا و نسکیمنتو [۶۵] بر تحلیل شباهت‌های بین اسناد دادگاه با استفاده از مدل‌های مبتنی بر ترنسفورمر تمرکزکردند. محققان شش تکنیک پردازش زبان طبیعی مبتنی بر معماری ترنسفورمر را بر روی یک مطالعه موردی از فرآیندهای حقوقی در سیستم قضایی برزیل اعمال‌کردند. در سال ۲۰۲۳ سی سودیا [۶۶] به بررسی تشابه متنی معنایی در قراردادهای حقوقی پرداخت. او از مدل‌های زبانی پیش‌آموزش دیده برای این کار استفاده‌کرد و آن‌ها را بر روی قراردادهای حقوقی بهبود داد. در سال ۲۰۲۳ ژو و همکاران [۶۷] یک سیستم بازیابی اطلاعات حقوقی مقابله با کووید-۱۹ را با استفاده از تطابق معنایی توسعه‌دادند.

این سیستم از یک شبکه عصبی کانولوشنال و روش تطابق معنایی برای گرفتن روابط بین جملات و پیش بینی‌ها استفاده‌می‌کند. در سال ۲۰۲۳ لی و ونگ [۶۸] با هدف توانمندسازی پژوهشگران حقوقی و ارتقاء خدمات حقوقی در مدیریت ریسک‌های عمومی، سامانه هوشمند تحلیل و بازیابی متون حقوقی مبتنی بر مدل برت را طراحی‌کردند. مدل پیشنهادی از برت برای استخراج اطلاعات معنایی از متون حقوقی استفاده‌می‌کند.

سال ۲۰۲۴ نصری و همکاران [۶۹] یک روش جدید برای شتاب-دادن به فرآیند تشریعات از طریق تحلیل همبستگی معنایی با استفاده از تکنیک‌های یادگیری عمیق مبتنی بر مدل برت ارائه‌داده‌اند. آن‌ها از مدل برت برای تشخیص همبستگی معنایی بین جملات در اسناد استفاده‌کرده و عملکرد این روش را با روش‌های موجود مقایسه‌کرده‌اند. نتایج نشان‌داده‌اند که این روش دقت و صحت بالایی در تشخیص همبستگی معنایی اسناد حقوقی مجلس شورای اسلامی ایران دارد.

**جدول ۱: تحقیقات انجام‌شده در زمینه شباهت‌یابی متون قضایی**

| ردیف | نویسندگان | سال | مجموعه داده | شاخص ارزیابی | روش شباهت‌یابی | دسته‌بندی روش‌ها |
|---|---|---|---|---|---|---|
| ۱ | مندل و همکاران [۳] | ۲۰۱۷ | پرونده‌های دادگاه عالی هند | Pearson correlation | TF IDF<br>Word2Vec<br>LDA<br>Doc2Vec | سطحی و یادگیری عمیق |
| ۲ | ما و همکاران [۵۷] | ۲۰۱۸ | CTA, CDD | Accuracy | WMD | سطحی |
| ۳ | نیر و واق [۵۸] | ۲۰۱۸ | CTA, CDD | Accuracy | WMD | سطحی |
| ٤ | رنجیت و آیدیکولا [۵۹] | ۲۰۱۹ | AILA Track dataset (اسناد قضایی هند) | Precision<br>Mean Average Precision<br>Binary Preference<br>Reciprocal Rank | Doc2Vec | یادگیری عمیق |
| ٥ | ایکسیا و همکاران [۶۰] | ۲۰۱۹ | اسناد قضایی چین | Accuracy | Word2Vec | یادگیری عمیق |
| ٦ | سوگاتاداسا و همکاران [۶۱] | ۲۰۱۹ | مجموعه داده ای شامل بیش از ۲۵۰۰ پرونده قضایی جمع آوری شده از وبسایت FindLaw | Accuracy<br>Recall | Doc2Vec | یادگیری عمیق |
| ۷ | باتاچاریا و همکاران [۶۲] | ۲۰۲۰ | پرونده‌های دادگاه عالی هند | Pearson correlation | Node2Vec<br>Thematic Similarity | یادگیری عمیق |
| ۸ | ماندال و همکاران [۶۳] | ۲۰۲۱ | پرونده‌های دادگاه عالی هند | Accuracy<br>Pearson correlation | word2vec<br>Doc2Vec<br>TF IDF<br>LDA<br>Law2Vec<br>BERT | سطحی و یادگیری عمیق |
| ۹ | دی اولیویرا و نسکیمنتو [۶۵] | ۲۰۲۲ | ۲۱۰،۰۰۰ پرونده قضایی از نوع تجدیدنظر عادی در سیستم قضایی برزیل | میانگین تشابه گروه‌های اسناد<br>میانگین شباهت کسینوسی در بین تمام اسناد گروه<br>میانگین شباهت اسناد هر گروه<br>میانگین شباهت مراکز اسناد گروه | BERT<br>GPT-2<br>ROBERTa | یادگیری عمیق |
| ۱۰ | شیتال و همکاران [۶۴] | ۲۰۲۲ | پرونده‌های دادگاه عالی هند | Micro F1 Score<br>Precision<br>Recall | TF-IDF<br>Doc2Vec<br>BERT | سطحی و یادگیری عمیق |
| ۱۱ | ژو و همکاران [۶۷] | ۲۰۲۳ | مجموعه‌ای از پرونده‌های حقوقی مربوط به همه‌گیری کووید-۱۹ | Accuracy<br>Precision<br>Recall<br>F1 Score | CNN<br>BERT | یادگیری عمیق |
| ۱۲ | سی سودیا [۶۶] | ۲۰۲۳ | LEDGAR | Pearson<br>spearman | Roberta | یادگیری عمیق |
| ۱۳ | لی و ونگ [۶۸] | ۲۰۲۳ | مجموعه داده انتزاعی CAIL2020، که از ۹۸۴۸ حکم دادگاه مدنی تشکیل شده‌است | Accuracy<br>F1 Score<br>Training duration | Bleem | یادگیری عمیق |
| ۱٤ | نصری و همکاران [۶۹] | ۲۰۲۴ | ۲۵۰ سند مربوط به مجلس شورای اسلامی | Precision<br>Recall<br>F1 Score<br>Accuracy<br>Roc | BERT | یادگیری عمیق |

## ۸. شباهت‌یابی متون قضایی: فرصت‌ها و محدودیت‌ها

یافتن پرونده‌های مشابه دادگاهی یکی از چالش‌های مهم در حوزه بررسی اسناد حقوقی است. دلیل این چالش ساختار پیچیده و طولانی اسناد حقوقی و نیز وجود مسائل حقوقی مختلف در یک پرونده واحد قضایی است [۳]. پس از تحلیل و بررسی در نظام قضایی ایران موارد ذیل به عنوان فرصت‌ها و محدودیت‌های شباهت‌یابی متون قضایی شناسایی شد که در ادامه به بررسی آن‌ها پرداخته‌می‌شود.

### ۸,۱. فرصت‌ها

دسترسی به اطلاعات حقوقی: شباهت‌یابی متون حقوقی می‌تواند یافتن قوانین و مقررات مرتبط را تسهیل‌کند، حتی اگر این قوانین در منابع مختلف و به زبان‌های متفاوت پراکنده شده‌باشند [۵۷].

**تحلیل پرونده‌ها:** این تکنیک می‌توانند با شناسایی پرونده‌های مشابه در یک حوزه حقوقی خاص، به قضات و متخصصان حقوقی کمک-کند تا از تجربیات پرونده‌های مشابه بیاموزند [۵۸].

**پیش‌بینی نتایج پرونده‌ها:** با تشخیص عواملی که بر نتایج پرونده‌ها تاثیرگذارند، می‌توان از شباهت‌یابی متون حقوقی برای پیش‌بینی نتایج پرونده‌های قضایی استفاده‌کرد [۵۹].

**شناسایی مسائل حقوقی جدید:** با مقایسه متون حقوقی جدید و متون موجود، امکان شناسایی مسائل حقوقی نوظهور فراهم می‌شود. به عنوان مثال می‌توان به بررسی آماری موضوعات پرونده‌ها و شناسایی داده‌های پرت در پرونده‌های قضایی و همچنین شناسایی چالش‌های حقوقی و قضایی در محتوای پرونده پس از تحلیل پرونده‌های مشابه اشاره‌کرد.

**شناسایی دادنامه‌های مشابه**: شباهت‌یابی متون حقوقی می‌تواند با شناسایی دادنامه‌های مشابه در یک حوزه حقوقی خاص، به قضات و متخصصان حقوقی کمک‌کند تا فرآیند رسیدگی به پرونده‌ها را سریعتر و کارآمدتر پیش ببرند [۶۰].

**شناسایی نقاط ابهام و یا همپوشان روندها و فرآیندهای قانونی:** مقایسه‌ی قوانین و مقررات با یکدیگر، امکان شناسایی نقاط ابهام و یا همپوشان روندها و فرآیندهای قانونی را فراهم می‌کند.

### ۸,۲. محدودیت‌ها

با وجود تمام مزایا و فرصت‌های پیش رو در شباهت‌یابی متون حقوقی استفاده از شباهت‌یابی متون حقوقی خالی از چالش نیست. از جمله:

**محدودیت‌های زبانی:** شباهت‌یابی متون حقوقی بر پایه زبان استوار است [۶۱] و این در حالی است که متون حقوقی اغلب پیچیده و مبهم هستند. این امر می‌تواند بر دقت این تکنیک اثر بگذارد [۵۸] همچنین متون حقوقی فارسی معمولاً ترکیبی از واژگان دو زبان فارسی و عربی هستند.

**کمبود داده‌های استاندارد:** برای ارزیابی الگوریتم‌های شباهت‌یابی متون حقوقی به داده‌های استاندارد نیاز است ولی متأسفانه در حال حاضر در بسیاری از زبان‌ها مانند زبان فارسی چنین داده‌های استانداردی موجود نیست. این مسئله می‌تواند مانعی برای توسعه‌ی این فناوری در بسیاری از زبان‌ها باشد [۶۰].

با توجه به چالش‌های مطرح‌شده، برای مقابله با آن‌ها می‌توان از راه‌کارهای ذیل استفاده‌کرد:

**توسعه روش‌های جدید:** با توسعه‌ی روش‌های جدید شباهت‌یابی متون، می‌توان دقت این فناوری را افزایش‌داد [۵۵].

**ایجاد پیکره‌های زبانی استاندارد:** ایجاد پیکره‌های زبانی استاندارد حقوقی در زبان فارسی می‌تواند توسعه این فناوری را در حوزه حقوق تسهیل‌کند. در نتیجه، شباهت‌یابی متون حقوقی علی رغم محدودیت‌هایی که دارد، فرصت‌های قابل‌توجهی را برای افزایش کارایی، دقت و شفافیت در حوزه حقوق فراهم‌می‌کند .انتظار می‌رود با توسعه بیشتر این فناوری و با در نظر گرفتن چالش‌های موجود، شاهد پیشرفت گسترده‌تری از شباهت‌یابی متون حقوقی باشیم.

## ۹. نتیجه

تحلیل شباهت معنایی متون حقوقی به عنوان یک حوزه تحقیقاتی پویا و نوظهور، در حال رشد و توسعه است و پتانسیل قابل توجهی برای توسعه سیستم‌های توصیه گر هوشمند قضایی دارد. این سیستم‌ها با بهره‌گیری از توانایی سنجش دقیق شباهت بین متون حقوقی، می‌توانند به‌طور قابل‌توجهی به بهبود کارایی و دقت نظام قضایی از طریق تسریع و تسهیل روند رسیدگی به دعاوی، کاهش اطاله دادرسی، ارائه خدمات حقوقی دقیق‌تر و تخصصی‌تر به مردم، افزایش دسترسی به عدالت و بهبود کیفیت زندگی شهروندان کمک‌کنند.

در این مقاله، به بررسی جامع روش‌های مختلف شباهت‌یابی معنایی متون حقوقی پرداخته‌شده‌است. این روش‌ها به‌طور کلی به دو دسته روش‌های سطحی و روش‌های مبتنی بر یادگیری عمیق تقسیم و نتایج ذیل حاصل‌گردید:

**محدودیت‌های روش‌های سطحی در شباهت‌یابی متون حقوقی:** روش‌های سطحی علی‌رغم سادگی و کارایی، در درک الگوهای پیچیده موجود در متون حقوقی ناتوان هستند.

**عملکرد برتر مدل‌های مبتنی بر یادگیری عمیق در شباهت‌یابی متون حقوقی:** روش‌های مبتنی بر یادگیری عمیق با قدرت بالایی که در استخراج مفاهیم پیچیده از متون دارند، به عنوان پیشتازان این عرصه شناخته‌می‌شوند. مطالعات اخیر نشان‌داده‌اند که روش‌های مبتنی بر یادگیری عمیق، در بسیاری از موارد، عملکرد به مراتب بهتری نسبت به رویکردهای سطحی در شباهت‌یابی معنایی متون حقوقی دارند این روش‌ها بر محدودیت‌های رویکردهای سطحی غلبه‌می‌کنند و توانایی بالایی در درک زمینه، استدلال منطقی و تشخیص روابط پیچیده بین مفاهیم را دارند.

**انعطاف‌پذیری در انتخاب روش:** با وجود برتری‌های روش‌های مبتنی بر یادگیری عمیق، انتخاب روش مناسب برای شباهت‌یابی

معنایی متون حقوقی به‌طور کامل به شرایط خاص هر مسئله بستگی دارد. در برخی موارد، استفاده از روش‌های یادگیری عمیق ممکن است بهترین راهکار باشد، در حالی‌که در موارد دیگر، استفاده از روش‌های سطحی ممکن است کافی باشد.

## ۱۰.تحقیقات آتی

با وجود پیشرفت‌های قابل‌توجه در حوزه شباهت‌یابی متون حقوقی، هنوز چالش‌های زیادی در این حوزه و به‌خصوص متون حقوقی فارسی باقی مانده‌است. با تداوم تحقیقات در این زمینه و تلاش برای رفع چالش‌های موجود، امید است شاهد پیشرفت‌های قابل‌توجهی در حوزه شباهت‌یابی معنایی متون حقوقی باشیم. این پیشرفت‌ها می‌تواند به ارتقای کارایی و دقت نظام قضایی، افزایش دسترسی به عدالت و بهبود کیفیت زندگی شهروندان کمک‌کند.

برخی از زمینه‌های تحقیقاتی امیدوارکننده در این حوزه عبارتند از:

- **توسعه پیکره‌های حقوقی زبان فارسی:** این موضوع به مدل‌های یادگیری عمیق کمک‌می‌کند تا ظرافت‌های زبان فارسی و پیچیدگی‌های متون حقوقی را به‌طور کامل درک‌کنند.
- **طراحی روش‌های جدید یادگیری عمیق:** این روش‌ها باید به‌طور خاص برای شباهت‌یابی معنایی متون حقوقی طراحی شده باشند و بتوانند از اطلاعات موجود در متون حقوقی به شیوهٔ مؤثرتری استفاده‌کنند.
- **ایجاد ابزارهای کاربرپسند برای استفاده از روش‌های یادگیری عمیق:** این ابزارها باید به گونه‌ای طراحی‌شوند که استفاده از روش‌های یادگیری عمیق برای افراد متخصص و غیرمتخصص آسان باشد.
- **گسترش کاربردهای شباهت‌یابی معنایی در حوزه حقوق:** با توجه به گستردگی موضوعات حقوقی و قضایی، می‌توان از روش‌های شباهت‌یابی معنایی متون حقوقی در زمینه‌های مختلفی مانند بازیابی اطلاعات حقوقی، تحلیل پرونده‌های حقوقی، تشخیص تناقضات و مغایرت‌ها در متون حقوقی، ترجمه متون حقوقی، خلاصه‌سازی متون حقوقی و طبقه‌بندی متون استفاده‌کرد. این موضوع می‌تواند به بهبود کیفیت خدمات حقوقی ارائه‌شده به مردم و ارتقای عدالت در جامعه کمک‌کند.

در مجموع، می‌توان انتظار داشت که سیستم‌های توصیه‌گر هوشمند قضایی در آینده به‌طور گسترده در سیستم‌های قضایی مورد استفاده قرارگیرند و به عنوان دستیارانی هوشمند برای قضات و وکلا عمل‌کنند. این سیستم‌ها می‌توانند به بهبود کارایی، دقت و شفافیت تصمیم‌گیری‌های قضایی کمک شایانی کنند. با این حال، در راستای تحقق این پتانسیل، نیاز به تحقیقات بیشتر در زمینه توسعه روش‌های دقیق‌تر و کارآمدتر برای شباهت‌یابی معنایی متون حقوقی و همچنین ایجاد زیرساخت‌های لازم برای پیاده‌سازی این سیستم‌ها در محی‌طهای قضایی واقعی است.

## References


[1] Talib, M. R., Hanif, M. K., Nabi, Z., Sarwar, M. U., and Ayub, N. (2017). "Text mining of judicial system's corpora via clause elements", International Journal on Information Technologies & Security, 9(3).

[2] Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). "How does NLP benefit legal system: A summary of legal artificial intelligence", *arXiv preprint arXiv:2004.12158*.

[3] Mandal, A., Chaki, R., Saha, S., Ghosh, K., Pal, A., and Ghosh, S. (2017, November). "Measuring similarity among legal court case documents", In Proceedings of the 10th annual ACM India compute conference (pp. 1-9).

[4] Ghanbari, N., Nasirannajabadi, D., and Soltani, R. (2018). "Disparities in Outcomes of Similar Cases in Civil Litigation", International Legal Research, 11(39), 353-373. [Persian]

[5] Farhadishad, M., Kazemifard, M., & Rezaei, Z. (2023). "Predicting Court Judgment in Criminal Cases by Text Mining Techniques", Journal of Information Technology Management, 15(2), 204-222.

[6] Herissinejad, K. (2010). "Study of the factors of Roman-German legal system's impact on Iranian modern law", Journal of Legal Research, 39(2), 245-264. [Persian]

[7] Chandrasekaran, D., & Mago, V. (2021). "Evolution of semantic similarity—a survey", ACM Computing Surveys *(CSUR)*, *54*(2), 1-37.

[8] Yao, H., Liu, H., and Zhang, P. (2018). "A novel sentence similarity model with word embedding based on convolutional neural network", Concurrency and Computation: Practice and Experience, 30(23), e4415.

[9] Rezaei, Z., Farzaneh Kondori, N., Farhadishad, M., Hosseini, S., & Takrimi, A. (2017). Enhancing Product Recommendations and Sales Productivity with Advanced Deep Learning Models. *Mohammad and Hosseini, Sajjad and Takrimi, Ali and esmaili, sajjad and Amini, Mohammad amin, Enhancing Product Recommendations and Sales Productivity with Advanced Deep Learning Models*.

[10] Dhanani, J., Mehta, R., & Rana, D. (2022). Effective and scalable legal judgment recommendation using pre-learned word embedding. *Complex & Intelligent Systems*, *8*(4), 3199-3213.

[11] Bazargani, M., Alizadeh, S. H., & Masoumi, B. (2024). A deep neural network-based approach in tag recommender system to overcome users' Cold Start. *International Journal of Nonlinear Analysis and Applications*, *15*(7), 197-214.

[12] Manber, U. (1994, January). "Finding Similar Files in a Large File System", In Usenix winter (Vol. 94, pp. 1-10).

[13] Geravand, S., and Ahmadi, M. (2014). "An efficient and scalable plagiarism checking system using bloom filters", Computers & Electrical Engineering, 40(6), 1789-1800.

[14] Deza, E., Deza, M. M., Deza, M. M., and Deza, E. (2009). *Encyclopedia of distances*, (pp. 1-583). Springer Berlin Heidelberg.

[15] Sunilkumar, P., and Shaji, A. P. (2019, December). "A survey on semantic similarity", In 2019 International Conference on Advances in Computing, Communication and Control (ICAC3) (pp. 1-8). IEEE.

[16] Pawar, A., and Mago, V. (2018). "Calculating the similarity between words and sentences using a lexical database and corpus statistics", arXiv preprint arXiv:1802.05667.

[17] Wibisono, P. D., Asad, A., and Chintan, A. (2021). “Short text similarity measurement methods: a review”, Soft Computing, 1-25.

[18] Norouzi, M., Fleet, D. J., and Salakhutdinov, R. R. (2012). “Hamming distance metric learning”, Advances in neural information processing systems, 25.

[19] Momtaz, M., Bijari, K., Salehi, M., & Veisi, H. (2016, December). “Graph-based Approach to Text Alignment for Plagiarism Detection in Persian Documents”, In FIRE (working notes) (pp. 176-179).

[20] Atoum, I., and Otoom, A. (2016). “Efficient hybrid semantic text similarity using WordNet and a corpus”, International Journal of Advanced Computer Science and Applications, 7(9).

[21] Resnik, P. (1995). "Using information content to evaluate semantic similarity in a taxonomy", arXiv preprint cmp-lg/9511007 .

[22] Po, D. K. (2020). “Similarity based information retrieval using Levenshtein distance algorithm”, Int. J. Adv. Sci. Res. Eng, 6(04), 06-10.

[23] Leonardo, B., and Hansun, S. (2017). “Text documents plagiarism detection using Rabin-Karp and Jaro-Winkler distance algorithms”, Indonesian Journal of Electrical Engineering and Computer Science, 5(2), 462-471.

[24] Jiang, J. Y., Cheng, W. H., Chiou, Y. S., and Lee, S. J. (2011, July). “A similarity measure for text processing”, In 2011 International Conference on Machine Learning and Cybernetics (Vol. 4, pp. 1460-1465). IEEE.

[25] Pennington, J., R. Socher and C. D. Manning (2014). “Glove: Global vectors for word representation”, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) .

[26] Tous, R., & Delgado, J. (2006, September). A vector space model for semantic similarity calculation and OWL ontology alignment. In International Conference on Database and Expert Systems Applications (pp. 307-316). Berlin, Heidelberg: Springer Berlin Heidelberg.

[27] Turney, P. D. (2005). Measuring semantic similarity by latent relational analysis. arXiv preprint cs/0508053.

[28] Li, Y., Bandar, Z. A., & McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on knowledge and data engineering, 15(4), 871-882.

[29] Akila, D., & Jayakumar, C. (2014). Semantic similarity-a review of approaches and metrics. Int. J. Appl. Eng. Res, 9(24), 27581-27600.

[30] Sussna, M. (1993, December). Word sense disambiguation for free-text indexing using a massive semantic network. In Proceedings of the second international conference on Information and knowledge management (pp. 67-74).

[31] Su, Z., Ahn, B. R., Eom, K. Y., Kang, M. K., Kim, J. P., and Kim, M. K. (2008, June). “Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm”, In 2008 3rd International Conference on Innovative Computing Information and Control (pp. 569-569). IEEE.

[32] Chen, Y., Lu, H., and Li, L. (2017). “Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity”. PloS one, 12(3), e0173410.

[33] Landauer, T. K., and Dumais, S. T. (1997). “A solution to Plato's problem: The latent semantic analysis theory of acquisition”, induction, and representation of knowledge. Psychological review, 104(2), 211.

[34] Matveeva, I., Levow, G., Farahat, A., & Royer, C. (2007). Term representation with generalized latent semantic analysis. Amsterdam Studies in the Theory and History of Linguistic Science Series 4, 292, 45.

[35] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). “Latent dirichlet allocation”, Journal of machine Learning research, 3(Jan), 993-1022.

[36] Rong, X. (2014). “word2vec parameter learning explained”, arXiv preprint arXiv:1411.2738.

[37] Gabrilovich, E., & Markovitch, S. (2007, January). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In IJcAI (Vol. 7, pp. 1606-1611).

[38] Potthast, M., Stein, B., & Anderka, M. (2008, March). A wikipedia-based multilingual retrieval model. In European conference on information retrieval (pp. 522-530). Berlin, Heidelberg: Springer Berlin Heidelberg.

[39] Manthouri, M., Tehranipour, S., & Yazdani, S. (2023). “ParsAirCall: Automated Conversational IVR in Airport Call Center using Deep Transfer Learning”, Intelligent Multimedia Processing and Communication Systems (IMPCS), 2 (4).

[40] Kenter, T., and De Rijke, M. (2015, October). “Short text similarity with word embeddings”, In Proceedings of the 24th ACM international on conference on information and knowledge management (pp. 1411-1420).

[41] Deza, E., et al. (2009). Encyclopedia of distances, Springer .

[42] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). “Enriching word vectors with subword information”, Transactions of the association for computational linguistics, 5, 135-146.

[43] Blagec, K., Xu, H., Agibetov, A., and Samwald, M. (2019). “Neural sentence embedding models for semantic similarity estimation in the biomedical domain”, BMC bioinformatics, 20, 1-10.

[44] Reimers, N., and Gurevych, I. (2019). “Sentence-bert: Sentence embeddings using siamese bert-networks”, arXiv preprint arXiv:1908.10084.

[45] Alshammeri, M., Atwell, E., and ammar Alsalka, M. (2021). “Detecting semantic-based similarity between verses of the Quran with Doc2vec”, Procedia Computer Science, 189, 351-358.

[46] Chicco, D. (2021). “Siamese neural networks: An overview”, Artificial neural networks, 73-94.

[47] Zheng, T., Gao, Y., Wang, F., Fan, C., Fu, X., Li, M., ... and Ma, H. (2019). “Detection of medical text semantic similarity based on convolutional neural network”, BMC medical informatics and decision making, 19, 1-11.

[48] Kim, S. H., Nam, H., and Park, Y. H. (2022, May). “Temporal dynamic convolutional neural network for text-independent speaker verification and phonemic analysis”, In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6742-6746). IEEE.

[49] Han, M., Zhang, X., Yuan, X., Jiang, J., Yun, W., and Gao, C. (2021). “A survey on the techniques, applications, and performance of short text semantic similarity”, Concurrency and Computation: Practice and Experience, 33(5), e5971.

[50] Memory, L. S. T. (2010). “Long short-term memory”. Neural computation, 9(8), 1735-1780.

[51] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. and Polosukhin, I. (2017). “Attention is all you need”, Advances in neural information processing systems, 30.

[52] Devlin, J., et al. (2018). “Bert : Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805 .

[53] Gonbadi, L., & Ranjbar, N. (2022). “Sentiment Analysis of People’s opinion about Iranian National Cars with BERT”,

Intelligent Multimedia Processing and Communication Systems (IMPCS), 4(3), 51-60

[54] Liu, Y., et al. (2019). “Roberta: A robustly optimized bert pretraining approach”, arXiv preprint arXiv:1907.11692 .

[55] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”, arXiv preprint arXiv:1910.01108.

[56] Rajapaksha, P., Farahbakhsh, R., and Crespi, N. (2021). “Bert, xlnet or roberta: the best transfer learning model to detect clickbaits”, IEEE Access, 9, 154704-154716.

[57] Ma, Y., Zhang, P., and Ma, J. (2018). “An Efficient Approach to Learning Chinese Judgment Document Similarity Based on Knowledge Summarization”, arXiv preprint arXiv:1808.01843.

[58] Nair, A. M., and Wagh, R. S. (2018). “Similarity analysis of court judgements using association rule mining on case citation data-a case study”, Int J Eng Res Technol, 11(3), 373-381.

[59] Renjit, S., and Idicula, S. M. (2019). “CUSAT NLP@ AILA-FIRE2019: Similarity in Legal Texts using Document Level Embeddings”, In FIRE (Working Notes) (pp. 25-30).

[60] Xia, C., He, T., Li, W., Qin, Z., and Zou, Z. (2019, July). “Similarity analysis of law documents based on Word2vec”, In 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C) (pp. 354-357). IEEE.

[61] Sugathadasa, K., Ayesha, B., de Silva, N., Perera, A. S., Jayawardana, V., Lakmal, D., and Perera, M. (2019). “Legal document retrieval using document vector embeddings and deep learning”, In Intelligent Computing: Proceedings of the 2018 Computing Conference, Volume 2 (pp. 160-175). Springer International Publishing.

[62] Bhattacharya, P., Ghosh, K., Pal, A., and Ghosh, S. (2020). “Methods for computing legal document similarity: A comparative study”, arXiv preprint arXiv:2004.12307.

[63] Mandal, A., Ghosh, K., Ghosh, S., and Mandal, S. (2021). “Unsupervised approaches for measuring textual similarity between legal court case reports”, Artificial Intelligence and Law, 1-35.

[64] Sheetal, S., Veda, N., Prabhu, R., Pruthv, P., & Mamatha, H. R. R. (2022, December). Knowledge Graph-based Thematic Similarity for Indian Legal Judgement Documents using Rhetorical Roles. In Proceedings of the 19th International Conference on Natural Language Processing (ICON) (pp. 154-160).

[65] de Oliveira, R. S., and Nascimento, E. G. S. (2022). “Analysing similarities between legal court documents using natural language processing approaches based on Transformers”, *arXiv preprint arXiv:2204.07182*.

[66] Sisodia, Y. (2023). Semantic Textual Similarity on Contracts: Exploring Multiple Negative Ranking Losses for Sentence Transformers. Authorea Preprints.

[67] Li, B., and Wang, M. (2023). “Design of intelligent legal text analysis and information retrieval system based on BERT model”. (preprint)

[68] Zhu, J., Wu, J., Luo, X., and Liu, J. (2023). “Semantic matching based legal information retrieval system for COVID-19 pandemic”, Artificial intelligence and law, 1-30.

[69] Naseri, J., Hasanpour, H., & Ghanbari Sorkhi, A. (2024). Accelerating Legislation Processes through Semantic Similarity Analysis with BERT-based Deep Learning. International Journal of Engineering, 37(6), 1050-1058.

## پی‌نوشت

[1] Cold Start

[2] Term Frequency-Inverse Document Frequency (TF-IDF)

[3] Shallow Methods

[4] Cosin similarity

[5] WordNet-based Similarity

[6] Information Content (IC)

[7] Levenshtein Distance

[8] Jaro-Winkler Distance

[9] Hamming Distance

[10] Global Vectors for Word Representation (Glove)

[11] Vector Space Model (VSM)

[12] Latent Relational Analysis (LRA)

[13] Smith-Waterman Algorithm

[14] Longest Common Subsequence (LCS)

[15] Latent Semantic Analysis (LSA)

[16] Latent Dirichlet Allocation (LDA)

[17] term frequency–inverse document frequency (TF-IDF)

[18] Explicit Semantic Analysis (ESA)

[19] Open Directory Project (ODP)

[20] Word Embeddings

[21] Sentence Embeddings

[22] Document Embeddings

[23] Siamese Networks

[24] Convolutional Neural Networks (CNNs)

[25] Recurrent Neural Networks (RNNs)

[26] Long Short-Term Memory (LSTM)

[27] Transformer-based Models

[28] GPU

[29] RoBERTa

[30] DistilBERT

[31] Word Mover's Distance (WMD)