Journal of Information and Communication Technology

Volume 16, Issue 59-60, Spring and Summer 2024, pp. 47-70

A Review on Hadith Text Processing Tasks

Sepideh Baradaran-Hazaveh[1], Behrouz Minaei-Bidgoli[2*], Mohammad E. Shenassa[1], Sayyed-Ali Hossayni[3]

[1] Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
[2] School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
[3] Artificial Intelligence Laboratory of Digital Humanities and Islamic Sciences Research Institute (NOOR), Qom, Iran

Received: 16 November 2022, Revised: 31 January 2023, Accepted: 22 March 2023
Paper type: Review

Abstract

In order to facilitate and achieve higher precision and less processing time, it is recommended to evaluate the authenticity of hadith by intelligent methods. Due to the huge volume of narrative texts (hadith) and the complex concepts and relationships in them, many researches have been conducted in the field of automatic hadith processing. In this field, some researchers have evaluated intelligent methods in the fields of Matn (text) and Isnad processing, which according to the review of previous researches, about 47% of them in the field of hadith text processing and 46% in the case of Isnad processing of hadiths and 7% have done research in both fields. By examining 97 researches in the field of processing hadiths, it was found that hadiths were evaluated in the field of measuring the accuracy of the text or Isnad or both cases. Processing tasks can be classified into different categories such as ontology construction, hadith text classification, hadith similarities and hadith authentication. The most used hadith processing method has been the information retrieval method in the field of hadith text processing.

Keywords: Hadith, Text Authenticity, Isnad, Narrator, Corpus of Hadith.

* Corresponding Author's email: b_minaei@iust.ac.ir

دوفصلنامه
فناوری اطلاعات و ارتباطات ایران

سال شانزدهم، شماره های ۵۹ و ۶۰، بهار و تابستان ۱۴۰۳، صفحه ۴۷ الی ۷۰

پژوهشی مروری بر حوزه‌های پردازشی متون روایی و احادیث

سپیده برادران هزاوه[۱]، بهروز مینائی بیدگلی[۲*]، محمدابراهیم شناسا[۱]، سیدعلی حسینی [۳]

[۱] گروه کامپیوتر، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات، تهران، ایران

[۲] دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران

[۳] آزمایشگاه هوش مصنوعی پژوهشکده علوم اسلامی و انسانی دیجیتال (نور) ، قم، ایران

تاریخ دریافت: ۱۴۰۱/۰۸/۲۵ تاریخ بازبینی: ۱۴۰۱/۱۱/۱۱ تاریخ پذیرش: ۱۴۰۲/۰۱/۰۲

نوع مقاله: مروری

چکیده

جهت سهولت و رسیدن به دقت بالاتر و زمان پردازش کمتر، ارزیابی صحت حدیث به روش‌های هوشمند توصیه می‌شود. با توجه به حجم قابل توجه متون روایی و مفاهیم و روابط پیچیده موجود در آنها، تاکنون پژوهش‌های فراوانی در حوزه پردازش خودکار حدیث انجام شده است. در این حوزه، عده‌ای از محققان در زمینه‌های پردازش متن و سند، شیوه‌های هوشمندی را آزمایش کرده‌اند، که با توجه به مرور تحقیقات پیشین، حدود ۴۷٪ از آنان در خصوص پردازش متن احادیث و ۴۵٪ در مورد پردازش سند احادیث و ۸٪ در هر دو حوزه پژوهش نموده‌اند. با بررسی ۱۰۱ پژوهش در حوزه پردازش احادیث، مشخص شد که احادیث در حوزه سنجش صحت متن یا سند یا هر دو مورد، ارزیابی شده‌اند. وظایف پردازش را می‌توان به دسته‌های مختلفی از جمله ساخت هستان‌شناسی، رده‌بندی متن حدیث، تشابهات حدیثی و اعتبارسنجی احادیث طبقه‌بندی نمود. پرکاربردترین روش پردازشی حدیث، روش رده‌بندی در حوزه پردازش متن حدیث بوده است.

کلیدواژگان: حدیث، صحت متن، سند، راوی، پیکره حدیث.

* رایانامهٔ نویسنده مسؤول: b_minaei@iust.ac.ir

۱- مقدمه

قرآن و حدیث از اصلی‌ترین منابع دینی در اسلام است. حدیث از فرمایشات معصومین صلوات الله علیهم اجمعین، خصوصا به نقل از رسول الله، حضرت محمد، صلی الله علیه و آله و سلم بیان شده و از دو قسمت تشکیل شده است: قسمت اول، اسناد روایت است که در بردارنده سلسله راویانی هستند که حدیث را نقل کرده‌اند و قسمت دوم، متن روایت است. تاریخ حیات فکری مسلمانان نشانگر استفاده از رهنمودهای پیشوایان دینی در همه عرصه‌های دانش است، از این رو انتقال صحیح و دقیق فرمایشات ایشان منجر به گسترش حقیقت علوم اسلامی در متن جامعه خواهد بود، لذا هر گونه تحریف و نقصانی در این حوزه، ضربات جبران‌ناپذیری بر پیکره حیات اسلامی وارد خواهد آورد.

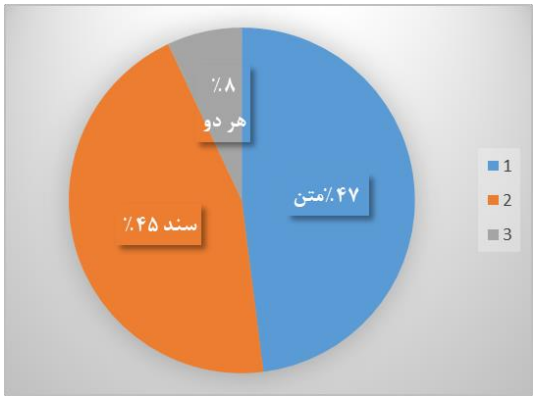
پدیده جعل حدیث توسط معاندین و منافقین، از زمان پیامبر (ص) تاکنون سابقه داشته و نشان‌دهنده نفوذ سودجویان فرصت طلب در صفوف مسلمانان است. پیامبر اکرم (ص) نیز برای پیشگیری از آثار سوء این پدیده شوم، در زمان حیات خود هشدارهایی داده‌اند و صاحبان عقل و انصاف را از آن آگاه نموده‌اند. ائمه طاهرین (ع) و به تبع آنان علمای بزرگوار نیز همواره متوجه این خطر بزرگ بوده و شیوه‌هایی برای مقابله با آن ابداع نموده‌اند. دو نوع جعل حدیث وجود دارد: ۱) جعل سند و متن حدیث توامان ۲) جعل متن حدیث با سند به ظاهر صحیح (سندی که مربوط به روایت دیگری است)، در پی آن، پالایش احادیث نیز دو نوع می‌باشد. الف) پالایش متنی و سندی، ب) پالایش متنی.

وجود شیوه‌های گوناگون دستیابی به درستی حدیث، گواه گستره پهناور این تلاش جهت دستیابی به این مهم است. ملاک‌های نقد متن حدیث عبارتند از: عرضه حدیث بر قرآن کریم، عرضه حدیث بر سنت مقطوع، نقد متن حدیث بر پایه ضروریات مذهب، نقد حدیث بر اساس سیاق آن، نقد حدیث بر اساس عقل عرفی و غیره. اما ملاک‌های نقد سند حدیث عبارتند از: نقد حدیث بر پایه طبقه راویان، نقد سند بر پایه روش‌های غیرمعمول در بین امامیه و غیره [۱].

جهت سهولت و رسیدن به دقت بالاتر و زمان پردازش کمتر، ارزیابی صحت حدیث به روش‌های هوشمند توصیه می‌شود. در این حوزه، عده‌ای از محققان در زمینه‌های پردازش متن و سند، شیوه‌های هوشمندی را آزمایش کرده‌اند، که با توجه به مرور تحقیقات پیشین، حدود ۴۷٪ از آنان در خصوص پردازش متن احادیث و ۴۵٪ در مورد پردازش سند احادیث و ۸٪ در هر دو حوزه پژوهش نموده‌اند. نسبت پردازش متن به سند حدیث یا پردازش همزمان آنها در شکل ۱ نمایش داده شده است.

پژوهش‌های مرتبط با ارزیابی صحت متن احادیث را می‌توان به وظایف پردازشی مختلف دسته‌بندی کرد که در شکل ۲ شرح داده شده است. همچنین تحقیقات زیادی در رابطه با وظایف پردازشی سند احادیث وجود دارد که در شکل ۳ شرح داده شده است.

برخی پژوهش‌ها، از دو منظر پردازش صحت متن و سند، احادیث را بررسی نموده‌اند؛ که در شکل ۴ نشان داده شده‌اند.



شکل ۱. نسبت پردازش صحت متن با سند و با هر دو

شکل ۲. پژوهش‌های مبتنی برمتن احادیث

شکل ۳. پژوهش‌های مبتنی بر سند احادیث

شکل ۴. مطالعات مبتنی بر متن و سند احادیث

۱-۱- ضرورت پژوهش

با توجه به خطر تحریف احادیث که دومین منبع آسمانی و موثق مسلمانان بعد از قرآن کریم است؛ که در بخش مقدمه، مفصلا تشریح شد؛ این پژوهش درصدد بررسی حوزه‌های پردازشی احادیث، برآمده است تا بر اساس یک منبع علمی، بتوان تفکیک کرد که کدام شیوه پردازشی هوشمند در ارزیابی صحت متن یا سند احادیث پرکاربردتر یا بهینه‌تر است. رسیدن به چنین مطلوبی، یک ضرورت علمی برای کسانی است که خواهان دریافت منبع حدیثی موثق و به دور از تحریف، آن هم براساس شیوه‌های هوشمند ارزیابی و با حداقل درصد خطا می‌باشند.

۲- تحقیقات پیشین

پژوهش‌های پیشین بررسی شده، از سه منظر پردازش متون و اسناد حدیث یا هر دو، مورد توجه قرار گرفتند. در هر منظر دسته‌بندی‌هایی وجود دارد که به وسیله روش‌های مرتبط، آنها را شرح می‌دهیم.

۲-۱- پژوهش‌های مرتبط با متن احادیث

در این قسمت به بررسی تحقیقات مربوط به متون احادیث می‌پردازیم.

۲-۱-۱- ساخت هستان‌شناسی

نه مقاله در حوزه هستان‌شناسی متن حدیث، یافت شده است که دو تای آنها صرفا به استخراج الگو برای قرآن با طرحی برای افزودن حدیث در آینده اقدام کرده است و ۳۷۴ مفهوم یا نمونه را کشف کرده‌اند. این کار رویکردی از نسل خودکارسازی نمونه‌های هستان‌شناسی است که بر مجموعه اسناد بدون ساختار یعنی قرآن کریم محقق شده است. روش ارائه شده بر اساس ترکیبی از تکنیک‌های پردازش زبان طبیعی (NLP)[1]، استخراج اطلاعات[2] و تکنیک‌های متن‌کاوی[3] طراحی شده است. بر اساس سامانه‌های استخراج اطلاعات سنتی، نویسندگان یک قانون دستور زبان و استخراج را برای بدست آوردن نمونه‌های هستان‌شناسی اعمال و تعریف می‌کردند اما این سامانه سعی کرده است با ترکیب کلمات و موجودیت‌هایی که در متن وجود دارند، برای نمونه‌های صحیح و کامل، نمونه‌های جزئی صحیحی را شکل دهد. نتیجه‌گیری نشان می‌دهد که نمونه‌های استخراج شده در صورت خوشه‌بندی می‌توانند به شش قسمت تقسیم شوند [۲] و [۳]. یک تحقیق در حوزه ایجاد قوانین انجمنی هستان‌شناسی برای فقه اسلامی بر روی منبع احادیث انجام شده است. رویکرد این مقاله، ارزیابی استفاده از قوانین انجمنی برای شناسایی موارد تکراری مفاهیمی است که به فقه اسلامی مربوط می‌شوند و از پیکره بخاری جهت محاسبه روابط مشابه آنها با استفاده از الگوریتم استقرایی استفاده شده است [۴].

در مقاله چهارم چارچوبی برای ایجاد هستان‌شناسی از متون عربی بر اساس حدیث ارائه می‌شود. این چارچوب بر اساس پردازش زبان طبیعی، روش‌های آماری و داده‌کاوی[4] برای استخراج مفاهیم و روابط معنایی است. نویسنده معتقد است که یک چالش قابل توجه در ساخت هستان‌شناسی عربی، فقدان روش‌های ارزیابی منظم و

[1] Natural Language Processing
[2] Information Extraction
[3] Text Mining
[4] DataMining

استانداردهای مرجع است. بنابراین، به یک استاندارد طلایی پیکره و هستان‌شناسی نیاز است. در این کار، چهار فاز، پیش پردازش پیکره، استخراج مفهوم، اکتشاف رابطه مفهوم و ساختار هستان‌شناسی طراحی شده است [۵]. در مقاله پنجم هستان‌شناسی برای قوانین اسلامی مبتنی بر استخراج الگوی خودگردان‌سازی[1] طراحی شده است [۶]. در مقاله ششم هستان‌شناسی تفسیر حدیث، طراحی و ایجاد شده است. مروری بر پورتال‌های حدیث موجود و کمبود اطلاعات تفسیری در آنها، انگیزه طراحی چنین هستان‌شناسی معرفی شده است. این هستان‌شناسی برمبنای هدف هستان‌شناسی که توانایی تولید نتیجه یا پاسخگویی به سوالات است، آزمایش شده است. این هستان‌شناسی می‌تواند انباره تفسیر حدیث را پشتیبانی کند و ارتباطات غیرمستقیم حدیث و آیات قرآن را به گونه‌ای مستند برای استفاده اهداف مختلف، ذخیره‌سازی کند [۷]. مقاله هفتم یک واژه‌نامه کامل به همراه توضیحاتی برای هر اصطلاح، طراحی و پیاده‌سازی کرده است. اصطلاحات این واژه‌نامه شامل مفاهیم، نمونه‌ها و خصوصیات است. هستان‌شناسی در این تحقیق، به عنوان مخزنی عمل می‌کند که URL[2] آن، جایی که هر کلمه قرار دارد را ثبت می‌کند. هستان‌شناسی نبوی شامل ۱۲۳۰ عبارت عربی است. در این پژوهش از زبان هستان‌شناسی وب[3] استفاده شده است که یک زبان استاندارد در هستان‌شناسی است و توسط W3C[4] برای طراحی آن توصیه می‌شود [۸]. مقاله هشتم با طراحی یک فرهنگ لغت نرمال‌سازی شده، مفاهیم کلمه را در متن حدیث، اشکال‌زدایی کرده است [۹]. مقاله آخر، روشی را برای طراحی و توسعه مجموعه داده WordNet حدیث، پیشنهاد می‌کند. WordNet یک منبع زبانی قدرتمند است که به کاربران اجازه می‌دهد از طریق روابط واژگانی و معنایی-مفهومی به کلمات، مترادف‌ها و رابطه بین آنها دسترسی داشته باشند. یک منبع WordNet برای عربی استاندارد مدرن وجود دارد، اما حدیث منقول به عربی کلاسیک چنین منبعی ندارد. در مجموع ۲۶۷۱ حدیث در ۲۴ فصل رده‌بندی شده است و میانگین امتیاز F برابر با ۹۴ درصد گزارش شده است [۱۰]. با جمع‌بندی مقالات این حوزه به نظر می‌رسد وجود یک هستان‌شناسی خاص حدیث کلاسیک در این حوزه ضروری باشد که البته در آن رابطه میان مفاهیم آیات قرآن و احادیث مشخص شده باشند.

۲-۱-۲- رده‌بندی متن حدیث

شانزده مقاله در حوزه رده‌بندی متن حدیث، یافت شده است که در مقاله اول به گروه‌بندی متن حدیث با استفاده از ترکیب الگوریتم‌های استخراج متن و الگوریتم میانگین C فازی[5] پرداخته شده است. هدف این مطالعه، کشف گروه‌های جدید برای جستجوی بهتر کاربران است. الگوریتم میانگین C فازی، به عنوان روش گروه‌بندی استفاده شده است. با مقایسه نتایج محاسبات دستی با نتایج محاسبات با استفاده از نرم افزار Rstudio، صحت[6] بدست آمده ۸۰٪ گزارش شده است [۱۱]. در حوزه SVM[7] در متن حدیث، مقاله دوم با استفاده از سه روش، SVM، رگرسیون خطی و پرسپترون چند لایه به متن‌کاوی متن حدیث پرداخته است [۱۲]. در مقاله سوم با پرس‌وجو از اسناد مرتبط در مالایی، نمونه اولیه‌ای را برای تحقیق در مورد داده‌های مربوط به حلال ۱۰ تهیه کردند. با استفاده از الگوریتم نمایه‌سازی معنای پنهان LSI[8] و تحلیل فرکانس به توسعه پرس‌وجوی متن احادیث مالایی دست یافتند. تکنیک شباهت کسینوس برای اندازه‌گیری شباهت بین پرس‌وجو و اسناد استفاده شد. پنج مجموعه پرس‌وجو درباره محصولات حلال ایجاد شد. برای ارزیابی تکنیک، مجموعه داده به صورت دستی مورد تحلیل قرار گرفت و لیستی از قضاوت‌های مرتبط تهیه شد. این آزمایش ثابت کرد که LSI نتایج بهتری را ارائه می‌دهد اما به زمان پردازش بیشتری نیز نیاز دارد. بهترین نتیجه گزارش شده $P = 0.37$ و $R = 1.0$ است [۱۳]. در مقاله چهارم با استفاده از روش فراوانی واژه- معکوس فراوانی سند (TF-IDF)[9]، به طور گسترده در بازیابی اطلاعات استفاده کردند و به رده‌بندی متون حدیث پرداختند و صحت ۸۳٪ را گزارش کردند [۱۴]. مقاله پنجم یک مطالعه تطبیقی را در مورد رده‌بندی متون عربی با استفاده از معیارهای مختلف انجام دادند. با روش‌های شباهت کسینوسی، جاکارد، دایس، ضرب داخلی و NB[10] رده‌بندی متون حدیث را با معیار ارزیابی $F = 0.85$ برای بیزین ساده انجام دادند [۱۵]. در مقاله ششم آزمایشی را برای رده‌بندی احادیث مبتنی بر درجه شباهت آنها با پرس‌وجوی کاربرانجام دادند. آنها ابزار متن‌کاوی تهیه کردند تا برای مجموعه داده‌های حدیث براساس مدل فضای برداری (VSM)[11]، معیار شباهت کسینوس و TF-IDF استفاده شود. هنگامی که کاربر موضوعی را جستجو می‌کند، سامانه مجموعه‌ای از احادیث مرتبط با

[1] Bootstrapping
[2] Uniform Resource Locator
[3] Web Ontology Language
[4] World Wide Web Consortium
[5] Fuzzy C-Means Method
[6] Accuracy
[7] Support Vector Machine
[8] LSI: Latent Semantic Indexing
[9] Term Frequency- Inverse Document Frequency
[10] Naïve Bayes
[11] Vector Space Model

پرس‌وجوی کاربر را که به صورت نزولی مرتب شده است، برمی‌گرداند. عملکرد سامانه به ترتیب ۶۶ و ۸۰ درصد برای دقت[1] و فراخوانی[2] گزارش شده است [۱۶]. مقاله هفتم با استفاده از شبکه عصبی و SVD به رده‌بندی متون می‌پردازند. برای شناسایی مناسب‌ترین ویژگی‌های رده‌بندی، آنها از تکنیک تجزیه ارزش واحد[3] استفاده کردند. معیار ارزیابی F را برای شبکه عصبی به تنهایی ۸۵٪ و به همراه SVD، ۸۸٪ گزارش کرده‌اند [۱۷]. مقاله هشتم نتیجه کار قبلی رده‌بندی اسناد متنی عربی (مقاله پنجم) را با استفاده از درخت تصمیم[4] نیز ارائه دادند. به ترتیب، مقادیر معیار ارزیابی F را برای پیکره علمی۷۰٪ و برای پیکره حدیث ۴۰٪ گزارش کرده‌اند. بسیاری از احادیث غلط رده‌بندی شده حاوی تعداد زیادی کلمه بودند که نمایانگر سایر دسته‌ها بودند و این یکی از دلایل اصلی ضعف عملکرد بود. DT هنگامی که روی پیکره علمی آزمایش شد، از سایر رده‌بندها بهتر عمل کرد. معیار عملکرد F برابر ۷۰٪ برای DT، به دنبال آن ۶۸٪ برای NB، و ۶۳٪ برای حداکثر آنتروپی بود. بدترین نتیجه نیز برای دایس گزارش شد که F برابر ۴۲٪ است [۱۸]. در مقاله نهم در مورد تأثیر مکانیسم کاهش ابعاد در رده‌بندی متن عربی با استفاده از الگوریتم شبکه عصبی پس انتشار[5] بحث کردند. مولفان پنج تکنیک کاهشی مختلف را مقایسه کردند: ریشه‌یابی، ریشه‌یابی سطحی، فرکانس سند[6]، TF-IDF و نمایه‌سازی معنایی نهفته[7]. نتیجه نشان داد که تکنیک‌های DF، TF-IDF و LSI نسبت به ریشه‌یابی و ریشه‌یابی سطحی برتر بودند. بهترین میانگین معیار ارزیابی F را برای BPNN[8] و TF-IDF برابر با ۵۶٪ گزارش کرده‌اند [۱۹]. در پژوهش دهم، مطالعه‌ای برای کشف دانش در متن حدیث با هدف رده‌بندی حدیث انجام شده است. سامانه پیشنهادی شامل چهار مرحله است: پیش‌پردازش پیکره، وزن‌دهی ویژگی‌ها، پردازش پرس‌وجو و گسترش علاوه بر رده‌بندی، و آخرین مرحله تحلیل نتایج است. با استفاده از TF-IDF، ریشه‌یابی، ضرب داخلی، شباهت کسینوس، جاکارد و دایس به رده‌بندی متون پرداخته شده و F برای هر فصل به طور جداگانه بین ۳۵٪ تا ۹۵٪ گزارش شده است [۲۰]. در مقاله یازدهم، با استفاده از NB و Bagging و LogiBoost به رده‌بندی موضوعی متون حدیث پرداختند و معیار ارزیابی F را ۶۰٪ بیان کردند [۲۱]. در مقاله دوازدهم همان نویسندگان، به روش‌های مقاله قبلی SVM را اضافه کردند و F را برای SVM، ۵۸٪ گزارش کردند [۲۲]. در کار سیزدهم، با استفاده از ریشه‌یاب‌ها و رده‌بندهای متفاوت به ارزیابی ابزارهای NLP برای رده‌بندی متون حدیث پرداختند با استفاده از اعتبارسنجی متقاطع ۱۰ لایه[9]، بهترین نتیجه با استفاده از ریشه‌یاب خوجه[10] و رده‌بند SVM با صحت[11] ۵۷٪ بدست آمد. جالب اینجاست که بدترین عملکرد SVM با استفاده از هر یک از ریشه‌یاب‌های آزمایش شده با بهترین نتیجه حاصل شده توسط رده‌بند NB قابل مقایسه است [۲۳]. در کار آخر، مطالعات مبتنی بر یادگیری ماشین را که منحصرا بر حوزه حدیث متمرکز بود، کشف کردند. ارزیابی صحیح این مطالعات با توجه به مجموعه داده‌های متنوع استفاده شده، دشوار است. در این کار، مولفان به زحمت پیاده‌سازی مجدد و ارزیابی روش‌های مختلف را در یک مجموعه داده[12]، انجام دادند. نتایج نشان داد که رده‌بند شبکه‌های عصبی مصنوعی[13] با دقت ۹۴٪ بهترین در بین سایرین است. همچنین، این مطالعه با استفاده از مدل VSM و شباهت کسینوس علاوه بر پرس‌وجوی غنی شده، اثر بازیابی حدیث را منعکس می‌کند [۲۴]. در پژوهش پانزدهم، به مقایسه و ارزیابی عملکرد چهار مدل تشخیص موجودیت نامدار عربی (CAMeLBERT-CA، Marefa-NER،Hatmimoha و Stanz) برای مجموعه داده بخاری پرداخته شده است. هدف اصلی این مطالعه یافتن بهترین عملکرد ابزارهای ذکر شده برای استفاده در سایر مجموعه داده‌های حدیث است. مدل‌های Stanza و Marefa-NER بهترین هستند چون برای معیار F1 به ترتیب مقادیر ۰/۸۳ و ۰/۸۱ را به‌دست آوردند. در این پژوهش یک مجموعه داده جدید در حدود ۵۰۰۰ کلمه بر اساس حاشیه‌نویسی CANER-Corpus[14] ایجاد شده است. ابتدا مجموعه داده صحیح البخاری مورد استفاده قرار گرفت که از مجموعه حدیث دانشگاه لیدز و ملک سعود (LK) دانلود شد. مجموعه حدیث LK مجموعه‌ای دو زبانه از حدیث انگلیسی-عربی است که شامل ۹۷ فایل است که بیش از هفت هزار حدیث را پوشش می‌دهد. سپس، CANER-Corpus در این مطالعه برای بازنگری و تصحیح مجموعه داده آزمایش استفاده شد. با اینکه

[1] Precision
[2] Recall (یا نرخ یادآوری)
[3] SVD: Singular Value Decomposition
[4] DT: Decision Tree
[5] BPNN: Back Propagation Neural Network
[6] DF: Document Frequency
[7] LSI: Latent Semantic Indexing
[8] Back-Propagation Neural Network
[9] Ten-Fold Cross Validation
[10] خوجه و گرساید، ۱۹۹۹
[11] Accuracy
[12] ۳۱۵۰ حدیث از صحیح بخاری
[13] ANN: Artificial Neural Network
[14] یک مجموعه کلاسیک NER عربی است که توسط متخصصان انسانی به صورت دستی حاشیه نویسی شده است. این مجموعه شامل بیش از ۷۰۰۰ حدیث از صحیح البخاری است.

مدل‌های Stanza و Marefa-NER بهترین بودند اما نتایج متفاوتی هنگام آزمایش تمام مدل‌های قبلی در مجموعه آزمایشی جدید ایجاد شد. این به دلیل تعداد کم کلمات حاشیه نویسی است که محدودیتی برای این اثر محسوب می‌شود. Hatmimoha بهترین امتیاز را در مقایسه با Marefa-NER و Stanza به دست آورد. مولفان معتقدند که اگر مدل دارای کلاس‌های موجودیت نام‌گذاری‌شده زیادی باشد و با تگ‌های CANERCorpus مطابقت داشته باشد، امتیاز بالایی نتیجه می‌دهد. متون اسلامی دارای کلمات منحصر به فردی است که با متن استاندارد عربی مدرن متفاوت است زیرا حاوی نام خدا و پیامبر است. بنابراین، برای کارهای آینده، طرح بهبود مدل جدیدی برای متن کلاسیک عربی، به ویژه برای متون اسلامی پیشنهاد می‌شود [۲۵]. هدف از پژوهش آخر، کشف احادیث ساختگی‌ای است که بیشتر از همه از سوی دانشمندان مسلمان رد شده است. در این پژوهش به جای تمرکز بر سلسله راویان حدیث، از متن و محتوای حدیث، استفاده شده است. به منظور انجام این کار، اولین مجموعه داده اختصاصی احادیث ساختگی ایجاد و منتشر شد که MAHADDAT نام دارد. علاوه بر این، یک سیستم تشخیص حدیث جعلی[۱] را بر اساس یک مدل زبان مبدل، یعنی BERT[۲] راه‌اندازی گردید که نرخ معیار F1 برابر با ۹۲,۴۷٪ شد. در مقام مقایسه با سایر مدل‌های BERT عربی که در مجموعه داده‌های بسیار کوچک‌تری آموزش دیده‌اند، این مجموعه که مبتنی بر CAMeLBERT_CA است، یک مدل مبتنی بر BERT و متخصص در نوع عربی کلاسیک می‌باشد. یک مطالعه مقایسه‌ای کامل در احراز هویت حدیث بین الگوریتم‌های متعدد ML[۳] کلاسیک و همه TLM[۴] های عربی موجود نیز انجام شد. چنین مقایسه‌ای نشان می‌دهد که تمام TLMهای عربی بر همه مدل‌های کلاسیک ML برتری دارند [۲۶]. با بررسی مقالات موجود در این زمینه می‌توان گفت که استفاده از شبکه عصبی پیچشی[۵] (CNN) و شبکه عصبی بازگشتی[۶] (RNN) و شبکه‌ی مبتنی بر BERT در این زمینه به نتایج بهتری منجر خواهد شد. از آنجا که احادیث لزوما متعلق به یک باب موضوعی نیستند و ممکن است یک حدیث به چند موضوع مختلف اشاره کند، لازم است مجموعه دادگانی از احادیث در اختیار باشد که در آن موضوعات مختلف مربوط به یک حدیث مشخص شده باشد تا رده‌بندی با دقت بالاتری انجام گیرد.

۲-۱-۳- قطعه‌بندی متن حدیث

یک مقاله در حوزه قطعه‌بندی متن حدیث، یافت شده است که در آن به منظور بهینه‌سازی پردازش متن حدیث، یک ابزار قطعه‌بندی برای پیکره حدیث مبتنی بر رمزگذاری TEI[۷] ایجاد شده است. این ابزار یکپارچه در نمونه اولیه رمزگذاری TEI برای تقسیم‌بندی پیکره حدیث طراحی و ایجاد شده و سپس در پیکره بخاری که شامل ۷۵۶۳ حدیث در ۹۴ فصل است، آزمایش شده است. مقادیر معیارهای ارزیابی شامل دقت، فراخوانی و F نشان می‌دهد که نتایج به دست آمده از ابزار تقسیم‌بندی حدیث برای سه معیار، برابر با ۹۶٪ و دلگرم‌کننده است [۲۷].

۲-۱-۴- پرسش و پاسخگویی

دو مقاله در حوزه پرسش و پاسخ متن حدیث، یافت شده است که در مقاله اول، چالش‌های اصلی سامانه پاسخگویی به سوالات کاربران[۸] را بررسی کرده است. هدف این کار، افزایش دقت سامانه پاسخگویی به سوالات کاربران جهت یافتن احادیث مرتبط با استفاده از روش‌های مفید از قبیل روش‌های پیش‌پردازش مانند رمزنگاری و حذف کلمات توقف[۹] برای شناسایی مفاهیم اصلی پرسش کاربران، N-gram، WordNet، CS[۱۰] و LCS[۱۱] برای به روزرسانی و غنی‌سازی مفاهیم استخراج شده از پرسش کاربران و روش‌های بردار پشتیبان و تشخیص موجودیت نامدار[۱۲] برای رده‌بندی اسناد حدیث بر اساس موضوعات و انواع سوالات مرتبط به منظور کاهش دامنه جستجو است. میانگین صحت پاسخ به ترتیب با استفاده از تکنیک CS برابر با ۶۷٪، روش LCS برابر با ۶۶٪، ترکیب روش‌های CS و LCS برابر با ۷۰٪ و با استفاده از CS، LCS و SVM[۱۳] میانگین صحت پاسخگویی برابر با ۸۰٪ است. سهم اصلی این تحقیق، استفاده از روش SVM برای کاهش دامنه جستجوی اسناد احادیث براساس موضوعات مختلف و انواع سوالات در کنار تحلیل موثر نیاز پرسش کاربران با استفاده از روش‌های پردازش زبان طبیعی است. SVM پاسخ دقیق‌تری از استخراج پاسخ را فقط با استفاده از تکنیک‌های تشابه مانند CS و LCS ارائه می‌دهد [۲۸]. در مقاله

[1] Mawdu Hadith (MH)
[2] Bidirectional Encoder Representations from Transformers
[3] Machine learning
[4] The transmission-line matrix
[5] Convolutional Neural Network
[6] Recurret Neural Network
[۷] Text Encoding Initiative (TEI) یک انجمن است که به طور کلی، استانداردی برای بازنمایی متون به شکل دیجیتال تهیه و نگهداری می‌کند.
[8] QAS: Question Answering System
[9] Stop words
[10] Cosine
[11] Longest Common Subsequence
[12] Name Entity Recognition
[13] Support Vector Machine

آخر، یک سامانه پاسخگو به پرس‌وجوهای کاربران با هدف بازیابی اطلاعات دقیق از مجموعه بزرگ حدیث طراحی و پیاده سازی شده است. مشکل روش‌های موجود، این است که آنها هنگام مقایسه یک جمله و پرسش کاربر، نمی‌توانند معنی آن را بدست آورند؛ بنابراین اغلب بین جملات استخراج شده و نیازهای کاربران تعارض وجود دارد. روش پیشنهادی ASHLK[1] با موفقیت این مشکل را حل کرده است: اول اینکه از استخراج عبارت مشابه با پرس‌وجو اما با مفهوم متفاوت جلوگیری میکند؛ دوم اینکه شباهت معنایی و نحوی جمله با جمله و جمله با پرس‌وجو را محاسبه می‌کند و سوم اینکه کلمات را هم در پرس‌وجو و هم در جملات گسترش می‌دهد تا مشکل اساسی عدم تطابق اصطلاحات بین جملات و پرس‌وجو کاربران حل شود. به منظور کاهش متون حدیث زائد، روش پیشنهادی با استفاده از الگوریتم حریصانه، جریمه متنوع را برای جملات اعمال می‌کند. خروجی‌های روش پیشنهادی، برای محاسبه دقت، فراخوان و معیار F با مرجع انسانی مقایسه شده است. نتایج تجربی نشان می‌دهد که عملکرد روش پیشنهادی در مقایسه با سایر روش‌ها بسیار رقابتی است [۲۹]. با بررسی مدل‌های یادگیری عمیق که در این مقالات بکار گرفته شده است، استفاده از مدل‌های یادگیری دنباله به دنباله[2] مبتنی بر رمزگذار-رمزگشا[3] به بهبود نتایج خواهد انجامید. در حال حاضر مجموعه دادگان جامعی در مورد پرسش و پاسخ‌های موجود پیرامون یک حدیث وجود ندارد و این امکان وجود دارد که از تعاملاتی که انسان مراجعه کننده و خبره حدیث با هم داشته‌اند، چنین دادگانی ایجاد و استفاده شود.

۲-۱-۵- شباهت‌یابی حدیث

دو مقاله در حوزه شباهت‌یابی متن حدیث، یافت شده است که مقاله اول با به کارگیری مفاهیم شباهت معنایی اسناد بر مدل فضای برداری (VSM)[4] به شبیه‌سازی سند به سند اقدام کرده و جهت ارزیابی از پرسشنامه پیمایشی استفاده نموده است. هنوز خروجی داده‌ها به شکل لیست شده، تک بعدی و خطی ارایه می‌شود که به سختی می‌توان اطلاعات مربوط به درخواست‌ها را پیدا کرد؛ ایده این است که با استفاده از مدل فضای برداری، روابط احادیث در مفهوم شباهت معنایی اسناد، تعیین گردد. این رابطه بین احادیث را می‌توان به طور بصری در قالب گراف بیان کرد (دو بعدی). روش مورد استفاده در ایجاد مدل فضای برداری، رابطه واژه - سند، TF IDF - و روش تشابه کسینوس است که برای رابطه سند به سند استفاده می‌شود. جعبه ابزار Prefuse به عنوان ابزار تجسم استفاده می‌گردد. برای ارزیابی، آنها یک پرسشنامه پیمایشی با یازده سوال که بیشتر مربوط به قابلیت استفاده از سامانه است را توزیع کردند. از بین شرکت کنندگان، ۹۰٪ موافقت کردند که نگرش دو بعدی، اسناد مرتبط‌تری را در مقایسه با اسنادی که با جستجوی ساده پیدا شده‌اند، ارائه می‌دهد [۳۰]. در مقاله آخر، چارچوبی جهت شناسایی متون مشابه در پیکره حدیث معرفی شده است. این سامانه جدید برای تشخیص تشابه متن در پیکره بزرگ حدیث اسلامی مرکز تحقیقات رایانه‌ای علوم اسلامی طراحی شده است و از روش N-gram و اندازه‌گیری کسینوس استفاده می‌کند. با توجه به نتیجه ارزیابی، سامانه‌های تشخیص تشابه رایانه‌ای می‌توانند نسبت به کار قبلی در این حوزه کارآمدتر باشند. معیار F برای این سامانه ۹۷٪ ارزیابی شده است. امید است که این سامانه بتواند ضمن یافتن احادیث یکپارچه، امکان تشخیص چگونگی تقسیم یک حدیث بزرگ به چندین قطعه کوچک حدیث مانند آنچه در کتاب‌های مختلف حدیث سنتی تقسیم کرده‌اند، را دارا باشد [۳۱].

در حوزه شباهت‌یابی حدیث، برای رسیدن به موفقیت بیشتر و پردازش بهینه‌تر، موارد زیر توصیه می‌گردد:

- پیش‌پردازش مناسب حدیث
- برگرداندن کلمات به ریشه[5] برای ساخت بردار کلمات مناسب
- استفاده از بردار تعبیه‌ی[6] مناسب برای بیان بهتر معنای حدیث
- استفاده از معیارهای شباهت برداری بهتر

بررسی مدل‌های یادگیری عمیق در این حوزه نشان می‌دهد که استفاده از مدل‌های مبتنی بر BERT علی‌الخصوص ROBERTA که با داده بیشتری آموزش دیده، می‌تواند منجر به نتایج بهتری شود.

۲-۱-۶- بازیابی اطلاعات

چهارده پژوهش در حوزه بازیابی اطلاعات متن حدیث، یافت شده است که اولی به استخراج دانش با استفاده از مبدل با حالت متناهی (FST)[7] پرداخته است که شاخص ارزیابی F برابر با ۶۷٪ برای فصول و ۷۷٪ برای زیرفصل و ۳۳٪ برای اسناد و ۴۵٪ برای متن حدیث گزارش شده است [۳۲]. مقاله دوم با طراحی برنامه‌های کاربردی در تلفن همراه به جستجوی حدیث به زبان مالایی می‌پردازد [۳۳]. در مقاله سوم با استفاده از TF-IDF و مترادف‌ها به توسعه پرس‌وجو

[1] Question Answering System in al-Hadith using Linguistic Knowledge
[2] Sequence-to-Sequence
[3] Enocder-Decoder
[4] VSM: Vector Space Model
[5] Lemma
[6] Word Embedding
[7] FST: Finite-State Transducer

پرداخته شده است [۳۴]. در مقاله چهارم یک پیکره ویژه به نام سامانه بازیابی حدیث آنلاین به مالایی برای توسعه پرس‌وجو طراحی شده است [۳۵]. در مقاله پنجم با طرح عبارات منظم به بازیابی اطلاعات متن حدیث پرداخته است. سامانه از سه نوع جستجو پشتیبانی می‌کند. ۱. جستجویی مبتنی بر ریشه که در آن سامانه با داشتن ریشه سه حرفی کلمه (کتب: ktb) سامانه می‌تواند تمام احادیث را بر مبنای کلماتی که از ریشه داده شده استخراج می‌کند، انتخاب کند. ۲. قابلیت جستجوی دوم این امکان را برای فرد فراهم می‌کند تا تمام احادیث را با دو کلمه در فاصله خاص پیدا کند. ۳. قابلیت جستجوی سوم، اجازه جستجوی همه احادیث را می‌دهد که کلمه خاصی دارند در حالی که کلمه دیگر غایب است. برای ساده نگه داشتن پایگاه داده، مولفان برای امکان قابلیت جستجوی پیشرفته به عبارات منظم (RE)[1] متوسل شدند. این سامانه قبل از جستجوی واقعی از الگویی برای ساخت RE مناسب استفاده می‌کند [۳۶]. مقاله ششم برای بهبود عملکرد بازیابی متن حدیث مالایی، از الگوریتم نمایه‌سازی معنای پنهان (LSI) موازی استفاده کردند. LSI یکی از روش‌های معروف جستجو است که پرس‌وجوها را در برنامه‌های بازیابی اطلاعات با اسناد مطابقت می‌دهد. ثابت شده است که LSI عملکرد بازیابی را بهبود می‌بخشد، با این وجود، هر چه اندازه اسناد بزرگتر می‌شود، پیاده‌سازی‌های فعلی به اندازه کافی سریع نیستند که بتوانند نتیجه را در یک رایانه شخصی استاندارد محاسبه کنند. در این مقاله، الگوریتم موازی LSI جدیدی در رایانه‌های شخصی استاندارد با پردازنده‌های چند هسته‌ای پیشنهاد شده است تا عملکرد بازیابی اسناد مربوطه را بهبود بخشد. LSI موازی پیشنهادی برای اجرای خودکار محاسبه ماتریس در الگوریتم‌های LSI به عنوان شیارهای موازی با استفاده از پردازنده‌های چند هسته‌ای طراحی شده است. روش Fork-Join[2] برای اجرای برنامه‌های موازی استفاده می‌شود. چون مولفان، زمان بازیابی اسناد مربوطه را در سامانه موازی خود در مقابل یک سامانه ترتیبی اندازه‌گیری کردند. نتیجه مطابق انتظار بود، الگوریتم LSI موازی پیشنهادی در مقایسه با الگوریتم LSI متوالی، زمان جستجو را بهبود بخشیده است [۳۷]. در مقاله هفتم، رابطه بین کلمات در فصل‌های حدیث را در سطح کلمات کلیدی بررسی کردند. برای این منظور، مولفان از ترکیبی از متن کاوی و تحلیل خوشه‌ای برای کشف فراوانی کلمات کلیدی در اسناد حدیث در یک فصل و شباهت‌هایشان با فصل دیگر استفاده کردند. مجموعه داده‌ها قبل از پردازش به زبان انگلیسی ترجمه شده است. نویسندگان کار خود را فقط به ۲۶ کلمه کلیدی محدود کردند که از مجموعه کلمات کلیدی استخراج شده، انتخاب شده‌اند. خوشه‌های تولید شده، احادیث مشابهی را نشان می‌دهند اما هیچ روش ارزیابی دقیق نشان داده نشده است [۳۸]. مقاله هشتم به بررسی اثربخشی و کارایی بازیابی حدیث بدون خروجی کاربرپسند پرداخته است [۳۹]. در مقاله نهم، یک مجموعه جامع حدیث چند زبانه (MHC)[3] یعنی یک ابزار جستجوی مفهوم برای حدیث طراحی شده است. چون جستجوی حدیث، درک مفاهیم و معانی این علم مهم است. علاوه بر جستجو، استفاده از مفاهیم مدنظر است که این روشی است که از داده‌های بزرگ و هدف به منظور دستیابی به نتایج مرتبط و دقیق‌تر استفاده می‌کند. ایده ساختن یک ابزار جستجو برای حدیث با مفاهیم جهت تسهیل جستجوی کاربران در وب و دسترسی به حدیث از چند طریق طراحی شده است. همین مولفان، مطالعه جداگانه‌ای را در زمینه طراحی ابزار جستجوی مفاهیم حدیث منتشر کردند. ابزار پیشنهادی چند زبانه، که احادیث را با مفاهیم آنها پیوند می‌دهد تا کار جستجوی کاربر را تسهیل کند. از آنجا که هدف اصلی ایجاد MHC است، مولفان مفاهیم را به چهار زبان عربی، انگلیسی، فرانسوی و روسی ترجمه کردند. مولفان ۱۰۰٪ را برای هر دو معیار (دقت و فراخوانی) در مقابل ابزار جستجوی حدیث آنلاین[4] گزارش کردند، که هر دو معیار آن سامانه کمتر از ۵۰٪ بود [۴۰]. در مقاله دهم، یک سامانه بازیابی حدیث به زبان اندونزیایی ایجاد شده است که قادر به نمایش نتایج جستجوی کلمات کلیدی وارد شده توسط کاربر است. برای بازیابی متن حدیث از الگوریتم‌های بنیادی Nazief و Andriani جهت نمایش نتایج جستجو بر اساس کلمات کلیدی وارد شده توسط کاربر و طرح[5] XML به عنوان اساس مخزن اطلاعات استفاده شده است. تحلیل نتایج آزمون نشان می‌دهد که این سامانه می‌تواند در روند بازیابی، تعداد زیادی از اسناد مربوطه را برگرداند چون نمره کامل ۱۰۰٪ را برای فراخوانی و دقت ۹۶٪ کسب کرده است [۴۱]. در مقاله یازدهم، مخزن ترجمه مقاله قبلی را با استفاده از طرح XML پیاده‌سازی کردند. برای بررسی عملکرد مخزن، از نمایش وب، با استفاده از PHP به وسیله الگوریتم‌های تطبیق رشته‌ای brute-force جهت نمایش نتایج جستجو بر اساس کلمات کلیدی وارد شده توسط کاربر استفاده می‌شود. نتیجه آزمون

[1] Regular Expressions

[2] صفی است که کارهای ورودی به چند بخش تقسیم می‌شوند تا سرورها بتوانند به کارهای ورودی سرویس دهند، و در انتها ادغام می‌شوند. این مدل بیشتر برای محاسبات موازی یا در سامانه‌هایی که برای تولید محصول چندین تامین‌کننده نیاز است (کارگاه‌های تولیدی)، استفاده می‌شود: در این مدل‌ها مسئله‌ای که مورد بررسی قرار می‌گیرد معمولاً زمانی است که طول می‌کشد تا یک کار به اتمام برسد.

[3] Multi-Language Hadith Corpora

[4] www.muhaddith.org

[5] Schema

سامانه، این است که ذخیره حدیث با استفاده از یک فایل XML ساخت‌یافته، نتایج جستجو را سریعتر از استفاده از فایل‌های XML بدون ساختار نشان می‌دهد. یک فایل XML ساخت‌یافته، برچسب‌گذاری شده است که می‌تواند به روند جستجو کمک کند. بنابراین، مخزن، مستقیما برچسب‌گذاری مورد نظر را جستجو می‌کند. در حالی که در فایل XML بدون ساختار فقط از یک برچسب‌گذاری کلی استفاده می‌شود، بنابراین مخزن، جستجوهای بیشتری نسبت به استفاده از فایل‌های ساخت‌یافته XML انجام می‌دهد که زمان جستجو را افزایش می‌دهد. میانگین زمان جستجو مخزن ترجمه، ۰,۸۵ میلی ثانیه است که در مقایسه با مخزن بدون ساختار سریعتر است [۴۲]. در رساله دکتری که کار دوازدهم است، طراحی و پیاده‌سازی یک پیکره موازی حدیث چندزبانه زبان عربی، انگلیسی، فرانسوی و روسی مبتنی بر روش بازیابی اطلاعات بررسی می‌شود. یک مشکل مهم در بازیابی اطلاعات متون، تاکید بر تطبیق دقیق کلمه یا کلمات مورد جستجو و کلمات مشابه در یک فایل متنی خاص است. این مساله در بسیاری از موارد منجر به از دست دادن نتایجی می‌شود که حاوی مترادف کلمات مورد پرس‌وجو است و احتمالا برای کاربر مفید می‌باشد. این معضل در اکثر سامانه‌های بازیابی اطلاعات برای داده‌های متنی بدون ساختار و در اکثر زبان‌ها خصوصا در زبان عربی وجود دارد. الگوریتم تطبیقی سامانه، از داده‌های فرایند بازیابی استفاده کرده، وزن کلمات پرس‌وجو را بر اساس اهمیت آنها محاسبه می‌کند و سپس آنها را با اسناد موجودی که برای محاسبه اهمیت کلمات در هر سند پردازش شده‌اند، مقایسه می‌کند. سپس ضریب تشابه از پرس‌وجویی خاص و مدارک موجود آن محاسبه می‌شود. برای بهبود عملکرد، سامانه دارای یک فرهنگ لغت از کلمات با قابلیت شناسایی تمام فایل‌هایی است که حاوی آن کلمات به عنوان یک شاخص معکوس است. یک پرتال وب برای سامانه ایجاد شده است تا امکان جستجوی کاربر از طریق شبکه جهانی وب فراهم شود. نتیجه ارزیابی، هم دقت متوسط و هم فراخوان متوسط را برای هر زبان نشان می‌دهد. میانگین دقت و متوسط فراخوان زبان عربی ۹۷٪ و ۸۲٪، برای زبان انگلیسی ۹۸٪ و ۹۰٪، زبان فرانسه ۹۸٪ و ۹۲٪ و زبان روسی ۹۸٪ و ۹۱٪ بودند [۴۳]. در مقاله سیزدهم، آزمایشی برای بررسی تأثیر بخش‌بندی موضوع در بازیابی اطلاعات عربی انجام شد. سامانه بازیابی اطلاعات سنتی، لیستی از اسناد[1] را به عنوان پاسخ به پرس و جوی کاربر برمی‌گرداند، لیست بزرگی که هیچ کاربری نمی‌توانست به طور کامل کاوش کند. امید بود که ساماندهی اسناد بازیابی شده به صورت موضوعی، تأثیر مثبتی بر دقت عملکرد داشته باشد. برای بخش‌بندی موضوع، مولفان با تطبیق کد دسترسی آزاد شخص ثالث از الگوریتم‌های بخش‌بندی، پیاده‌سازی خود را با تکنیک‌های TextTiling و C99 اعمال کردند آنها سامانه را با استفاده از چهار پرس‌وجو در هر مجموعه داده آزمایش کردند. مولفان نتیجه گرفتند که تقسیم‌بندی موضوع تأثیر قابل توجهی در سامانه بازیابی دارد. آنها برای معیار دقت، بهبود ۰,۴۴+ و برای معیار فراخوانی بهبود ۰,۵+ را برای بازیابی اطلاعات حدیث گزارش کردند [۴۴]. در مقاله آخر، با استفاده از لغت‌نامه‌ای جهت گسترش بازیابی احادیث در یک محیط غیرعربی، توسعه پرس‌وجو را آزمایش کرد. کل سامانه، متن حدیث و پرس‌وجو به زبان مالایی است. از طریق حذف کلمات توقف، برای ریشه‌یابی کلمات کلیدی، و ریشه کلمات را پردازش کردند و قاموس‌نامه مالایی برای کلمات معادل استفاده شد. پرس‌وجوهای گسترده برای جستجوی اسناد مرتبط در پایگاه داده مورد استفاده قرار گرفت. عملکرد پایین و بازیابی موثر با افزایش ۴ درصدی را گزارش کردند [۴۵]. از آنجا که یکی از مهمترین مولفه‌های بازیابی اطلاعات، استفاده از جستجوی معنایی واژگان می‌باشد، می‌توان موتور جستجویی طراحی نمود که با استفاده از ترکیب روش‌ها و بردارهای تعبیه معنایی پیشرفته‌تر مانند جستجوی کشسان[2] و BERT به جستجوی دقیق‌تر واژه‌ها و عبارات بپردازند.

۲-۱-۷- نمایه‌سازی مولف[3]

دو پژوهش در حوزه نمایه‌سازی مولف متن حدیث، یافت شده است که در مقاله اول با استفاده از n-grams کلمات و حروف، شیوه dis-legomena و رده‌بندهای متفاوت به تشخیص نویسندگی قرآن و احادیث پرداختند و نتیجه گرفتند که قرآن و احادیث نویسنده‌های متفاوتی دارند [۴۶]. در مقاله دوم، به مساله تفاوت نویسندگی در دو کتاب مذهبی قرآن و حدیث با روش اعتبارسنجی LOO[4] با ویژگی‌های ۴ گرمی، مبتنی بر ماشین بردار پشتیبانی پرداخته شده است. این تکنیک اعتبارسنجی، متشکل از ۳۷ آزمایش مختلف انتساب تألیف است که به صورت چرخشی انجام می‌شود، به استثنای هر بار یک نمونه جدید (به عنوان مثال پیکربندی پویای LOO). در هر آزمایش مجزا، امتیاز انتساب ۱۰۰٪ بوده است که منجر به صحت کامل اعتبارسنجی متقاطع به عدد ۱۰۰٪ بین این دو کتاب می‌شود. این تحقیق نشان می‌دهد که دو کتاب مورد تحلیل از لحاظ سبک‌شناسی متفاوت هستند و نظریه دو نویسنده متفاوت را تأیید می‌کند. این نتیجه مهم، موید آن کلام نورانی پیامبر اکرم (ص) است

[1] Documents
[2] Elastic Search
[3] Author profiling
[4] Leave-One-Out

که قرآن فقط برای او نازل گردیده، و او فقط راوی بوده است نه اینکه نویسنده آن باشد. این نتیجه‌گیری همچنین فرضیه‌ها و ادعاهای برخی افراد را که قرآن را اختراع پیامبر (ص) می‌دانستند، انکار می‌کند [۴۷]. با بررسی نتایج بدست آمده در این حوزه، علی‌الخصوص مقالاتی که نویسنده‌های مورد بررسی متعدد باشند به نظر می‌رسد این حوزه هنوز به نتایج قابل قبول و حد اشباع نرسیده است و جای تحقیقات بیشتر در این حوزه زیاد است.

۲-۱-۸- تحلیل ریخت‌شناسی[1]

دو پژوهش در حوزه تحلیل ریخت‌شناسی متن حدیث، یافت شده است که مقاله اول در حوزه هستان‌شناسی نیز بررسی شده بود و همزمان از شیوه ریخت‌شناسی هم استفاده نموده است که با طراحی فرهنگ لغت نرمال شده قصد در ابهام‌زدایی مفهوم کلمه در متن حدیث داشته است[۹]. در مقاله آخر با طرح رده بندی احتمالی به ابهام‌زدایی ریخت‌شناسی متون پرداخته‌اند [۴۸].

۲-۱-۹- اعتبارسنجی (تصدیق) حدیث[2]

یک پژوهش در حوزه اعتبارسنجی (تصدیق) متن حدیث، یافت شده است که در آن، روشی برای استخراج متن حدیث از صفحات وب اسلامی ارائه شده است، سپس با یک پایگاه داده رایزنی می‌شود تا درجه صحت آن مشخص شود. این قضاوت درباره یک حدیث بر اساس کار شیخ البانی[3]، مجموعه کتاب‌های صحیح و ضعیف وی، انجام شده است. مولفان برای ارزیابی سامانه خود، یک خزنده ساختند تا متن حدیث را از پنج صفحه وب جمع‌آوری و پردازش کند و دقت برابر ۳۹٪ و فراخوان برابر ۵۱٪ را گزارش کردند [۴۹].

در جدول ۱ جمع‌بندی پژوهش‌های موجود در رابطه با متن احادیث مشاهده می‌گردد.

۲-۲- پژوهش‌های مرتبط با اسناد حدیث

در این قسمت به بررسی تحقیقات مربوط به سند احادیث می‌پردازیم.

۲-۲-۱- قطعه‌بندی حدیث

یک پژوهش در حوزه قطعه‌بندی سند حدیث، یافت شده است که در آن، طراحی و ساخت یک پیکره موازی دو زبانه عربی- انگلیسی با ۳۳۳۵۹ حدیث مطرح شده است. در این مقاله، یک ابزار بخش‌بندی خودکار جهت تفکیک متن از اسناد طراحی شده است که توانسته است با صحت ۹۲٪ بخش‌بندی مولفه‌های حدیث و حاشیه‌نویسی آن را انجام دهد. این ابزار، هزینه ایجاد منابع زبانی را به حداقل می‌رساند و اثر تجربیات فردی در حاشیه‌نویسی را کاهش می‌دهد. این ابزار پس از پیش پردازش حدیث، آن را به کلمات تبدیل کرده و سپس دو گرمی آنها را در نظر می‌گیرد و سپس با استفاده از رده‌بند بیز ساده، هر توکن را به عنوان سند یا متن برچسب‌گذاری می‌کند. نهایتا یک رویکرد باقاعده[4] برای یافتن نقطه تقسیم اقدام می‌کند [۵۰].

۲-۲-۲- تحلیل زنجیره راویان

یازده پژوهش در حوزه تحلیل زنجیره راویان سند حدیث، یافت شده است که در مقاله اول به تحلیل شبکه اجتماعی[5] راویان حدیث پرداخته شده است. شبکه‌های روایی توسط زنجیره‌های روایت از یک شخص به شخص دیگر شکل می‌گیرد. شبکه‌های روایی به دلیل عدم در دسترس بودن داده‌ها در قالب یک شبکه، تا به حال مورد کاوش قرار نگرفته است. هدف، کشف راویان مرکزی، الگوهای تعامل و خصوصیات ساختاری چنین شبکه‌هایی از طریق برخی رویکردهای کلاسیک و ارائه یک روش رتبه‌بندی راوی است. بعلاوه، ابزاری برای تحلیل شبکه روایی حدیث ایجاد شده است که به محققان و مورخان کمک خواهد کرد. گره‌های شبکه، نمایانگر راویان و لبه‌ها، نماینده انتقال حدیث بین دو راوی است. متوسط کوتاهترین طول مسیر شبکه راویان ۳/۶۲ است [۵۱]. در مقاله دوم، یک مطالعه منظم و جامع در تعیین سهم حفاظت راویان از روایات نبوی مبتنی بر تحلیل شبکه اجتماعی انجام شده است. محققان توانستند لیستی از راویان تاثیرگذار بر حراست از احادیث را در ۱۶ مجموعه شناسایی کنند. نمودار راویان اثرگذار، در قرون ۲ و ۳ از مکه و مدینه به سمت کوفه و بغداد و سپس آسیای میانه تغیر جهت داده است. نهایتا نتیجه رتبه‌بندی راویان پیکره مسلم با پیکره بخاری مقایسه شد [۵۲]. در مقاله سوم، AUBSarf، تحلیلگر ریخت‌شناسی عربی، معرفی می‌شود. به عنوان یک مطالعه موردی، مولفان از AUBSarf برای کشف زنجیره راویان حدیث استفاده کردند. برای این کار، آنها

[1] Morphological Analysis
[2] Hadith authentication
[3] الالبانی، سنن بن ماجه را انتخاب کرد و مجموعه حدیث خود را براساس آن، در دو کتاب به نام‌های صحیح بن ماجه و ضعیف بن ماجه تالیف کرد. هر حدیثی در سنن ابن ماجه به هر یک از دو کتاب ختم می‌شود. اگر او، به عنوان یک حدیث شناس، حدیث را صحیح قضاوت کرد، آن را در صحیح بن ماجه قرار داد. و اگر حدیث ضعیف تلقی شد، آن را در ضعیف بن ماجه قرار داد. الالبانی همین کار را برای سایر مجموعه‌ها مانند سنن الترمذی و غیره نیز انجام داد.
[4] Rule-Based Approach
[5] SNA: Social Network Analysis

اتوماتای حالت محدود غیر قطعی[1] را طراحی کردند که AUBSarf را هدایت می‌کرد. با توجه به مجموعه‌ای از احادیث، سامانه، شروع و پایان سند حدیث تکی را تعیین می‌کند. برای شناسایی راویان، مولفان به پایگاه داده نام راویان AUBSarf استناد کردند. چهار حالت در اتوماتای حالت محدود وجود دارد. حالتی که دستگاه تشخیص می‌دهد، در حال حاضر سامانه خارج از محدوده اسناد حدیث است، دو حالت برای خواندن یک نام، و آخرین حالت برای بیان به دستگاه که در حال گذر از محدوده اصطلاحات روایت است [۵۳]. در مقاله چهارم، iTree، ابزاری نرم‌افزاری برای تولید خودکار تصویری گرافیکی از زنجیره‌های انتقال کامل حدیث را ارائه دادند. از آنجا که ورودی حدیث خام بود، یعنی متن ساده و بدون ساختار، اولین قدم شناسایی تک تک راویان بود. برای این منظور، مولفان از روش تجزیه کم عمق و مدل یادگیری مبتنی بر حافظه استفاده کردند. علاوه بر این، آنها گرامر ویژه دامنه را در فرم Backus-Naur توسعه یافته (EBNF) تعبیه کردند [۵۴] و مقاله پنجم با استفاده از CFG، تجزیه کم‌عمق و یادگیری مبتنی بر حافظه به بررسی صحت اسناد حدیث پرداخته و معیار موفقیت ۸۷٪ بیان شده است [۵۵]. اما در مقاله ششم، از روش‌های مشابه در دو مقاله قبلی استفاده شده است ولی مکانیزم هستان‌شناسی معنایی وب را بر متن حدیث اعمال کردند. به طور خاص، آنها از مکانیزم تحول هستان‌شناسی وب معنایی استفاده کردند تا زنجیره روایت حدیث را نشان دهند و درخت کامل آن را به صورت گرافیکی ارائه دهند. چارچوب توصیف منابع[2] برای تولید بازنمایی زنجیره راویان استفاده شد. همان معیار موفقیت کار قبلی مطرح شده است [۵۶]. مقاله هفتم، به استخراج و بصری سازی زنجیره راویان با استفاده از شناسایی و رده‌بندی موجودیت‌های نامداری که به صورت دستی تفسیر شده بود، می‌پردازد. یک استخراج کننده گراف راوی خودکار حاوی حاشیه‌نویسی، ANGE را ارائه دادند. ANGE، گراف روایت را از اسناد حدیث و بیوگرافی، با استفاده از ترکیبی از فن‌آوری‌های مختلف (به عنوان مثال، مورفولوژی، دستگاه حالت محدود، سند متقاطع) ایجاد می‌کند. اساسا این سامانه زنجیره‌های روایت حدیث را می‌سازد و سپس زنجیره‌های مختلف روایت را ادغام می‌کند، و نهایتا متن حدیث را در انتهای یک زنجیره ضمیمه می‌کند. برای ادغام زنجیره‌ها، مولفان معیار فاصله را برای حل مشکل ناهنجاری اسامی راویان تعریف کردند. از دو دسته رده‌بند ایجابی (بیز ساده) و افتراقی (نزدیکترین همسایگی و درخت تصمیم) برای رده‌بندی موجودیت‌های نامدار استفاده شده و p برابر با ۹۰٪ و R برابر با ۸۲٪ گزارش شده است. اسامی راویانی که از هر حدیث مشخص شده‌اند برای ایجاد زنجیره‌ای از روایات به هم پیوستند و به عنوان گراف راویان بازنمایی شدند [۵۷]. در مقاله هشتم، سامانه‌ای را برای نشان دادن زنجیره روایت حدیث به عنوان یک گراف شبکه ایجاد کردند که گامی در جهت احراز هویت حدیث بود. در واقع یک سلسله روایت از احادیث مسلم مبتنی بر تئوری گراف با صحت ۶۰٪ طراحی شده است [۵۸]. در کار نهم که پایان‌نامه مرتبه کارشناسی ارشد است؛ پایگاه اطلاعاتی خبره‌ی علم رجال طراحی و پیاده‌سازی شده است. هدف، به‌کارگیری شیوه نوین مهندسی دانش و اطلاعات در برنامه‌های رایانه‌ای فقه و علوم مقدماتی آن است. در این کار، ابتدا امکان‌سنجی ایجاد سامانه خبره فقه، ضرورت و روش دستیابی به آن مورد بررسی قرار می‌گیرد. سپس طبقه‌بندی‌های مختلف اطلاعات و دانش علم رجال مورد بررسی قرار داده می‌شود، علم رجال یکی از علوم مقدماتی فقه است که عهده‌دار بازشناسی روایات، اعتبار آنها و تعیین درجه اعتبار راویان احادیث می‌باشد. بعد از شناخت قلمرو محیط مساله‌ی علم رجال، بخش اصلی پروژه یعنی تحلیل نیازمندی‌های نرم‌افزار پایگاه اطلاعاتی خبره علم رجال و روش‌های اخذ دانش و نمایش آن و مدل اطلاعاتی و پردازه‌ای نرم‌افزار مطرح می‌گردد. در بخش پایانی، مراحل طراحی و پیاده‌سازی نرم‌افزار و زیرسامانه ارزیابی ماشینی اسناد و تعیین هویت راویان تبیین می‌شوند [۵۹]. مقاله دهم، به درجه‌بندی راویان حدیث با استفاده از رده‌بندهای SVM و BPM پرداخته است یعنی با استفاده از مفهوم تحلیل حساسیت تعیین شود که آیا یک راوی قابل اعتماد است یا خیر. و معیار ارزیابی F را برای SVM برابر ۹۵٪ و برای BPM برابر ۵۲٪ گزارش کرده است [۶۰]. مقاله آخر، یک ابتکار ساده را بر اساس زنجیره روایت ارائه داده است تا حدیث را به سه کلاس صحیح، حسن و ضعیف درجه‌بندی کند. این ایده برای استفاده از طرح قطعیت خالص و بدون یادگیری ماشینی برای تعیین صحت حدیث بود. طرح قطعیت اینگونه است که به هر راوی در زنجیره نوعی وزن اختصاص داده شود، که به ویژگی عمومی راوی بستگی دارد. مجموع نرمال شده وزن همه راویان در اسناد، صحت حدیث را تعیین می‌کند.

صحت گزارش شده برای صحیح بخاری ۹۹/۶ درصد و برای ترمیذی ۹۳/۶۲ درصد گزارش شده است [۶۱].

۲-۲-۳- اعتبارسنجی حدیث

چهارده پژوهش در حوزه اعتبارسنجی سند حدیث، یافت شده است که در مقاله اول، یک طرح مقدماتی مطرح شده است که منطق

[1] NFA: Nondeterministic Finite Automaton

[2] RDF: Resource Description Framework

فازی را برای تقلید از نحوه مواجهه حدیث‌شناسان با الجرح و التعدیل استفاده نموده است [۶۲]. مقاله دوم، برای تعیین اعتبار حدیث، یک سامانه خبره فازی ابداع شد که آن را بر اساس مجموعه قوانین و نظر خبرگان بنا نهادند. مولفان از دو موتور استنتاج استفاده کردند. اولی، هر یک از راویان را در حدیث رتبه‌بندی می‌کرد، خروجی آن به موتور استنتاج دوم منتقل می‌شد که میزان اعتبار حدیث را تعیین کند. صحت گزارش شده سامانه ۹۴٪ است [۶۳]. در مقاله سوم، آمارسنجی اصطلاحات حدیث با استفاده از قواعدکاوی انجام شده است [۶۴]. در مقاله چهارم، یک کار مقدماتی برای رده‌بندی صحت حدیث با استفاده از قوانین انجمنی انجام شده است [۶۵]. در مقاله پنجم، تحلیلی بر تولید سلسله مراتبی با سطوح مختلف مطالعات مرتبط برای پیوند با احراز هویت محاسباتی علم اسناد الحدیث مورد بحث قرار گرفته است. نتیجه حاصل از تحلیل، عمیق‌ترین سطح تصدیق حدیث است که بر اساس اصول تأیید حدیث در علم حدیث ارائه شده است [۶۶]. در سه مقاله‌ی ششم تا هشتم، با استفاده از درخت تصمیم (DT) صحت حدیث در چهار کلاس (به عنوان مثال، صحیح و حسن) بر اساس زنجیره‌ای از راویان رده‌بندی می‌شود [۶۷].

برای رده‌بندی حدیث، محققان مقالات، پنج ویژگی بولین در نظر گرفتند: اتصال، معیوب، بی‌قاعده، درجه اطمینان و درجه نگهداری. ویژگی‌های معیوب و بی‌قاعده مربوط به متن است، در حالی که سه مورد دیگر مربوط به اسناد است. اتصال درست است اگر هیچ وقفه‌ای (شکاف) در اسناد وجود نداشته باشد، و در غیر این صورت نادرست است. مولفان MDD[1] را که روشی برای رسیدگی به داده‌های از دست رفته در مجموعه داده حدیث است را ارائه دادند.

وظیفه رده‌بندی با استفاده از دو روش مختلف انجام شد: C4.5 (تولید DT) و بیز ساده (NB). صحت گزارش شده بدون MDD برابر با ۵۰٪ است و پس از به کارگیری MDD، صحت برای DT به ۹۸٪ رسیده است، که کمی بهتر از رده‌بند NB بود [۶۸] و [۶۹]. مقاله نهم، یک مطالعه مقدماتی در رده‌بندی و قاعده‌کاوی انجمنی انجام شده است [۷۰]. در مقاله دهم، مولفان کار دیگری برای درجه‌بندی حدیث در چهار کلاس ارئه دادند. مولفان از VSM[2] برای نمایش احادیث به عنوان بردار راویان استفاده کردند، با هر راوی به عنوان یک واژه[3] رفتار می‌شود. ترتیب راویان هنگام رده‌بندی حدیث ضروری است. نویسندگان پیشنهاد کردند که برای رده‌بندی از LVQ[4]، یک الگوریتم رده بندی باسرپرست مبتنی بر نمونه اولیه استفاده شود. بهترین عملکرد برای این سامانه در رده‌بندی صحیح و موضوع احادیث، به ترتیب ۸۰ و ۱۰۰ درصد گزارش شده است. برای حسن و ضعیف، دقت نسبی ۲۰٪ و ۰٪ بوده است [۷۱]. در مقاله یازدهم، با استفاده از الگوریتم ژنتیک رویکرد جدیدی در پردازش سند حدیث مطرح شده است. در این مقاله از ادات اسناد که نماینده انواع موجودیت‌های اسناد از قبیل نام راوی هستند، برای برچسب‌گذاری استفاده شده است. برای قضاوت درباره صحت حدیث، ابتدا باید نام راویان در اسناد استخراج شود و سپس قوانین داوری بر روی آنها اعمال شود. بسیاری از تحقیقات روش‌های مختلفی را برای استخراج نام راویان از اسناد ارائه داده‌اند. که در این تحقیق، با استفاده از الگوریتم‌های ژنتیک پردازش اسناد انجام می‌شود. این روش با هدف پیش بینی نام راویان و سایر POIها برای اسناد پیش بینی شده است. روش پیشنهادی به دقت ۸۱٪ رسیده است [۷۲]. در مقاله دوازدهم، تکنیک تحلیل حساسیت از پردازش زبان طبیعی برای ساخت رده‌بندی متن جهت کشف صحت حدیث در سه کلاس، صحیح، حسن و ضعیف استفاده شده است. صحت (صحیح و حسن) و سقم (ضعیف) احادیث ناشناخته بر اساس میزان یادگیری مجموعه داده‌های سفارشی اسناد، پیش‌بینی می‌شود. نسبت داده‌های آموزشی به داده‌های آزمون، ۱۹ به یک است یعنی فقط ۵٪ داده‌ها برای آزمون در نظر گرفته شدند. از روش اعتبارسنجی متقاطع ۵ لایه برای تخمین میزان مهارت یادگیری ماشین استفاده شده است. از میان ۶ رده‌بند، SVC[5] خطی بهترین عملکرد را نسبت به بقیه با عدد ۸۰٪ کسب کرده است اما رگرسیون لجستیک، درصد اصالت حدیث را می‌رساند. نتیجه آزمون رده‌بندی، صحت ۸۶٪ را نشان می‌دهد [۷۳]. در مقاله سیزدهم، یک طرح پیشنهادی برای استفاده از یادگیری عمیق جهت پردازش اسناد حدیث مطرح شده است.

در این مقاله یک چارچوب کلی، در واقع فرصت استفاده از یادگیری عمیق تعریف شده است که به رده‌بندی منظم احادیث مبتنی بر دو کلاس صحیح و ضعیف (نادرست) کمک می‌کند [۷۴]. مقاله آخر، مجموعه داده جدیدی را پیشنهاد می‌کند که شامل زنجیره روایات (اسناد) با راویان مشخص است. مجموعه داده AR-Sanad 280K حدود ۲۸۰ هزار سند مصنوعی دارد که می‌تواند برای شناسایی ۱۸۲۹۸ راوی استفاده شود. پس از ایجاد مجموعه داده AR-Sanad 280K، ابهام‌زدایی راوی در چندین گام آزمایشی مورد بررسی قرار گرفت. ابهام‌زدایی راوی حدیث به‌عنوان یک مسئله رده‌بندی چند

[1] برای تشخیص مفقودی داده

[2] مدل فضای برداری

[3] Term

[4] کمی‌سازی بردار یادگیری

[5] C-Support Vector Classifier

برچسبی با ۱۸۲۹۸ کلاس برچسب، مدل‌سازی شده است. بهترین نتایج با تنظیم دقیق مدل یادگیری عمیق مبتنی بر BERT (AraBERT) به دست آمد. در مجموعه اعتبارسنجی مجموعه داده AR-Sanad 280K امتیاز Micro F1 برابر با ۹۲,۹ و نرخ خطای سند[1] برابر با ۳۰,۲ بدست آمد. علاوه بر این، مجموعه آزمون واقعی از اسناد شش کتاب حدیث معروف اهل سنت استخراج شد. در ارزیابی داده‌های آزمون واقعی، بهترین مدل، امتیاز ۸۳,۵ را برای Micro F1 و ۶۰,۶ درصد برای نرخ خطای سند کسب کرد [۷۵].

با بررسی مدل‌های یادگیری عمیق مورد استفاده برای تشخیص اسناد حدیث از جمله صحیح و ضعیف بودن آنها و زنجیره راویان، مدل‌های مبتنی بر برت خصوصا مانند AraBERT که خاص زبان عربی طراحی شده‌اند سبب بهبود نتایج در این حوزه شده‌اند. این حوزه یکی از حوزه‌های پردازشی پرچالش است، چرا که تشخیص اعتبار یک حدیث به پارامترهای بسیار زیادی وابسته است و حتی افراد خبره در این زمینه نیز دچار اشتباه می‌شوند، بنابراین اعتبارسنجی صحت حدیث به دقت دادگان مورد نظر وابستگی زیادی دارد.

۲-۲-۴- ساخت هستان‌شناسی

دو مقاله در حوزه هستان‌شناسی سند حدیث، یافت شده است که در مقاله اول و مقاله دوم هستان‌شناسی وابسته به دامنه، به نام سامانه داوری اسناد مبتنی بر هستان‌شناسی[2] طراحی شده است. این هستان‌شناسی برای کمک به احراز هویت اسناد است. اساس این هستان‌شناسی، RDF خودکار راویان حدیث[3] است که مولفان سعی داشتند با خصوصیات، روابط و صفات بیشتر، سامانه آنها را غنی‌سازی کنند.

جدول ۱. جمع‌بندی روش‌های صحت‌سنجی متن حدیث

ردیف	حوزه‌های پردازشی	پژوهش	روش‌ها و الگوریتم‌ها	مجموعه دادگان	مزایا و محدودیت‌ها
۱	**ساخت هستان‌شناسی**	نه مقاله [۲-۱۰]	تکنیک‌های NLP و الگوهای متون اسلامی در ترکیب با روش‌های آماری و قوانین انجمنی	پیکره بخاری و WordNet و مسلم	خودکارسازی نمونه‌های هستان‌شناسی و استانداردسازی آنها، پشتیبانی از انباره تفسیر حدیث، ذخیره‌سازی ارتباطات غیرمستقیم حدیث و آیات قرآن، فرهنگ لغت نرمال‌سازی شده
۲	**قطعه‌بندی متن حدیث**	یک مقاله [۲۷]	ابزار قطعه‌بندی برای پیکره حدیث مبتنی بر رمزگذاری TEI	پیکره بخاری	بهینه‌سازی پردازش متن حدیث
۳	**پرسش و پاسخ**	دو مقاله [۲۸-۲۹]	الگوریتم جستجوی حریصانه، با بکارگیری WordNet و موجودیت نامدار	بخاری	استخراج عبارت مشابه از نظر معنایی با بسط کلمات پرس‌وجو
۴	**شباهت‌یابی حدیث**	دو مقاله [۳۰-۳۱]	یافتن شباهت معنایی اسناد بر مدل فضای برداری (VSM) و تشابه کسینوس	ترمذی و جامع الاحادیث	چارچوبی جهت شناسایی متون مشابه در پیکره حدیث، امکان تشخیص چگونگی تقسیم یک حدیث بزرگ به چندین قطعه کوچک حدیث
۵	**رده‌بندی متن حدیث**	چهارده مقاله [۱۱-۲۶]	الگوریتم‌های یادگیری عمیق و فازی	سنن نسائی و بخاری	عملکرد بهتر درخت تصمیم از سایر رده‌بندها، کشف دانش در متن حدیث با هدف رده‌بندی حدیث، ایجاد مجموعه داده احادیث ساختگی به نام MAHADDAT
۶	**بازیابی اطلاعات**	چهارده مقاله [۳۲-۴۵]	عبارات منظم(RE)، الگوریتم نمایه‌سازی معنای پنهان (LSI) موازی، الگوریتم‌های بنیادی Nazief و Andriani، الگوریتم‌های تطبیق رشته‌ای	بخاری و مسلم و ابدعود	طراحی برنامه‌های کاربردی جستجوی حدیث مالایی، هیچ روش ارزیابی دقیق نشان داده نشده است، خروجی کاربرپسند ندارد، عملکرد پایین
۷	**نمایه‌سازی مولف**	دو مقاله [۴۶-۴۷]	n-grams کلمات و حروف، شیوه dis-legomena و رده‌بندهای متفاوت، روش اعتبارسنجی LOO با ویژگی‌های ۴گرمی، مبتنی بر ماشین بردار پشتیبانی	بخاری	تشخیص نویسندگی قرآن و احادیث، بررسی تفاوت نویسندگی در دو کتاب مذهبی قرآن و حدیث، تفاوت در سبک‌شناسی دو کتاب مورد تحلیل (قرآن و حدیث) و تایید نظریه دو نویسنده متفاوت برای آنها
۸	**تحلیل ریخت‌شناسی**	دو مقاله [۹] و [۴۸]	رده‌بندهای درخت تصمیم، بیز ساده Naïve Possibilistic Network, SVM	کتب سته	ابهام‌زدایی مفهوم کلمه در متن حدیث با طراحی فرهنگ لغت نرمال شده، ابهام‌زدایی ریخت‌شناسی متون با طرح رده بندی احتمالی
۹	**اعتبارسنجی حدیث**	یک مقاله [۴۹]	HTML cleaner Java package ، نمایه‌سازی خودکار	کار شیخ البانی (صحیح بن ماجه و ضعیف بن ماجه)	طراحی روشی برای استخراج متن حدیث از صفحات وب اسلامی

[1] SER: Sanad Error Rate
[2] IJS: ontology-based Judging Hadith Isnad system
[3] عظمی و بن بدیع، ۲۰۱۰

IJS حدیث را به عنوان صحیح، حسن و ضعیف درجه‌بندی می‌کند. مولفان اعلام کردند که نتیجه کار آنها، ۳۷/۵٪ با قضاوت الالبانی و ۸۱٪ با قضاوت سایر حدیث‌شناسان تطابق داشته است [۷۶] و [۷۷].

۲-۲-۵- تشخیص اسامی راویان

ده مقاله در حوزه تشخیص اسامی راویان سند حدیث، یافت شده است که مقاله اول، به شناسایی اسامی راویان حدیث مبتنی بر روش قاعده‌کاوی پرداخته است. یک روش قاعده‌محور برای تشخیص ویژگی‌های نام راویان حدیث در متن حدیث مالایی ابداع شد. هدف این بود که مسئله هجی‌های مختلف نام راویان، که نام اصلی آنها به زبان عربی است، حل شود. در این زمینه دو مشکل وجود دارد: (الف) فقدان آوانگاری استاندارد نام‌ها بین زبان‌های مختلف، و (ب) ناهنجاری نام راویان. مورد دوم یک مشکل جهانی در ادبیات حدیث است که اشکال مختلفی برای یک راوی یکسان دارد [۷۸]. در مقاله دوم، یک روش ترکیبی مبتنی بر قاعده‌کاوی و معیارهای آماری جهت تشخیص نام راویان در حدیث مطرح شده است. روش مبتنی بر قاعده متکی به مجموعه‌ای از کلمات کلیدی است که شروع و موقعیت پایانی کاندید نام راوی را مشخص می‌کند و پس از مشخص شدن کاندید نام راوی، برای ارزیابی درستی احتمال نام کاندید به عنوان نام راوی، این امر به روش تحلیل آماری تایید می‌شود. نتایج اعلام شده برای روش قاعده‌کاوی ۸۶٪ =F در حالی که دقت برای LLR برابر با ۸۵٪ گزارش شده است. در نتیجه اعلام شده است، رویکرد ترکیبی در شناخت نام راوی حدیث نتیجه بهتری دارد [۷۹]. در مقاله سوم، با استفاده از تکنیک‌های چندگانه، به عنوان مثال ریخت‌شناسی، مستندسازی متقاطع و غیره، به تشخیص راویان حدیث پرداخته شده و معیار ارزیابی F برای تشخیص راویان در حدیث ۷۰٪ و در مجموعه بیوگرافی ۸۷٪ بیان شده است [۸۰]. در مقاله چهارم، نمایه‌ای[1] از اسامی راویان حدیث در مجموعه حدیث اندونزیایی بررسی شده است. اسامی نمایه‌سازی شده از روش تشخیص موجودیت نامدار[2] استفاده می‌کند زیرا اسامی نمایه‌سازی شده فقط به موجودیت‌هایی به شکل اسامی افراد نیاز دارند. در واقع، موجودیت تشخیص داده شده، صرفا یک موجودیت نام شخص است. برای ایجاد اسامی نمایه‌سازی شده در این تحقیق، از مدل مخفی مارکوف[3] استفاده شده است. استفاده از روش مارکوف و ترکیب چندین ویژگی دیگر، منجر به دستیابی سامانه به مقدار عملکرد[4] مناسب برابر با ۸۶٪ شد. اما با بکارگیری اعتبارسنجی متقاطع مبتنی بر پارامترها، مقدار عملکرد ۲٪ افزایش می‌یابد [۸۱].

در پژوهش پنجم، مولفان به بازیابی اطلاعات مبتنی بر پرس‌وجو با FST و CRF و استخراج دانش با استفاده از مجموعه داده‌های حدیث پرداختند. در این مقاله، یک چارچوب استخراج دانش برای استخراج موجودیت‌های نامدار از ترجمه اردویی صحیح البخاری پیشنهاد شده است. چارچوب پیشنهادی مبتنی بر سامانه مبدل حالت محدود برای استخراج موجودیت‌ها و پردازش محتوای حدیث با استفاده از برچسب‌گذاری ادات سخن[5] طراحی شده است. زمینه مشروط تصادفی[6]، یک الگوریتم کل نگر است که اسامی استخراج شده را برای NER[7] و رده‌بندی پردازش می‌کند [۸۲]. در مقاله ششم، زنجیره روایت حدیث از منظر اعتبار اطلاعات مطالعه شده است. ایده پیشنهادی، توسعه سامانه‌ای است که ورودی آن زنجیره انتقال حدیث باشد که از اعتبار زنجیره برخوردار است. برای حل این مساله، مولفان، سامانه زنجیره انتقال را تجزیه کردند و مجبور شدند نام کامل راویان موجود در آن را ساماندهی کند. با استفاده از یک فراداده از راویان که حاوی اطلاعات آنها و روابط بین آنها است، سامانه می‌تواند اصالت راویان موجود و سلسله احتمالی زنجیره را تشخیص دهد. برای اعتبار انتقال، مولفان کلمات و اصطلاحات مورد استفاده در سند حدیث را بررسی کردند. آنها از رده‌بند NB به دلیل سادگی و صحت آن استفاده کرده‌اند. مولفان از میزان موفقیت خوبی، F برابر ۸۹٪ برای تشخیص هویت خبر دادند [۸۳]. در مقاله هفتم، تحقیقی در FST[8] مبتنی بر استخراج‌گر موجودیت نامدار گزارش شده است. نویسندگان بر استخراج اطلاعات سطح، متمرکز شده‌اند که به پردازش زبانی پیچیده‌ای نیاز ندارد. هدف، شناسایی مناطق مرتبط متن و برچسب‌گذاری آنها با استفاده از مجموعه محدودی از برچسب‌ها است، به عنوان مثال شماره فصل، عنوان فصل، اسناد، متن و غیره. نتیجه نشان می‌دهد که FST با عناوین فصل و زیر فصل بهتر نتیجه می‌دهد و این ضعفی است برای اسناد و متن. برای مثال معیارهای (P, R, F) = (۱۰۰٪, ۵۰٪, ۶۷٪) برای عنوان فصل، در حالی که برای سند (P, R, F) = (۴۴٪, ۲۶٪, ۳۳٪) گزارش شده است [۸۴].

در مقاله هشتم، به استخراج نام راویان بر اساس مدل زبانی N-gram به جای استفاده از برچسب‌گذاری POS پرداخته شده است. ایده این است که لیستی از اصطلاحات روایت (اولین کلمه در عبارات روایی) ترکیب شود، سپس یک مدل n گرم از این عبارات فرموله

[1] Index
[2] Named Entity Recognition (NER)
[3] Hidden Markov Model
[4] Performance
[5] POS: part of speech
[6] Conditional Random Field
[7] Named Entity Recognition
[8] Finite State Transducer

شود. برای دریافت نام بیشتر افراد، نویسندگان از مدل ۱۰گرمی استفاده کردند. معیارهای ارزیابیP و R برای n برابر با ۳-۱۵، به ترتیب برابر با ۸۵٪ و در رنج ۱۹ تا ۶۱٪ گزارش شده است [۸۵]. در مقاله نهم، مدلی جهت استخراج نام افراد از متون عربی-اسلامی قدیمی مطرح شده است. اخیرا تحقیقات زیادی در زمینه شناسایی موجودیت‌های نامدار، در زبان انگلیسی و سایر زبان‌های اروپایی با موفقیت قابل قبولی انجام شده است؛ در حالی که نتایج در زبان‌های دیگر مانند عربی، فارسی و بسیاری از زبان‌های آسیای جنوبی قانع‌کننده نیست. یکی از مهمترین و مشکل‌ترین وظایف فرعی در شناسایی موجودیت نامدار، استخراج نام شخص است. در این مقاله با استفاده از مفهوم پیشنهادی «تزریق نام مناسب نامزد» در مدل زمینه‌های تصادفی مشروط، سامانه‌ی برای استخراج نام افراد در متون دینی عربی معرفی شده است. همچنین از متون دینی عربی باستانی یک پیکره ایجاد شده است. آزمایشات نشان می‌دهد که بر اساس این روش نتایج بسیار کارآمدی بدست آمده است. معیار F برای سه پیکره صفین، الارشاد و شرایع به ترتیب ۱۰۰٪، ۹۴٪ و ۷۶٪ است [۸۶]. در مقاله دهم، یک تکنیک پیشرفته پردازش زبان طبیعی (NLP) را برای شناسایی و احراز اعتبار راوی حدیث به عنوان بخشی از سند، با استفاده از شناسایی موجودیت نامدار (NER) برای پرداختن به ضرورت احراز هویت حدیث ارائه می‌کند. هدف این پژوهش شناسایی راوی حدیث با استفاده از رویکرد NER است. تکنیک NER که در این تحقیق توضیح داده شد، یک رده‌بند پیش‌خور فوق‌العاده[1] به آخرین لایه مدل BERT[2] از پیش آموزش‌دیده اضافه می‌کند. اکثر مطالعات روی NER عملکرد آن را با استفاده از نرخ F1 اندازه‌گیری می‌کنند. توزیع هر تگ NER قابل پیش بینی نیست و ممکن است داده‌های نامتعادل داشته باشد. یک امتیاز F1 برای گرفتن میانگین هارمونیک دقت و نرخ یادآوری لازم است. در فرآیند آزمایشی با استفاده از Cahya/bert-base-indonesian-1.5G، راه‌حل پیشنهادی، نرخ کلی معیار F1 را برابر با ۹۹/۶۳ درصد دریافت کرد. در شناسایی راوی حدیث با استفاده از سایر قسمت‌های حدیث، آزمون نهایی ۹۸/۲۷ درصد امتیاز F1 را به دست آورد. این نتایج نشان می‌دهد که وقتی برای شناسایی راویان حدیث در متون حدیثی اندونزیایی از مدل NER پیشنهادی استفاده می‌شود، این مدل در این نوع کار بهترین عملکرد را دارد [۸۷].

۲-۲-۶- بصری‌سازی اطلاعات

هشت مقاله در حوزه بصری‌سازی اطلاعات سند حدیث، یافت شده است که در مقاله اول، یک مطالعه مقدماتی با استفاده از الگوریتم DAG[3] و طراحی پایگاه داده بیوگرافی راویان صورت گرفته که منجر به بازنمایی و تحلیل اسناد حدیث شده است [۸۸]. در مقاله دوم، با تطبیق تکنیک تاییدیه حدیث بر احراز هویت شواهد دیجیتال به تشخیص راویان حدیث پرداخته است [۸۹]. در مقاله سوم، مولفان نمونه اولیه‌ای را برای مصورساز زنجیره راویان حدیث[4] تهیه کردند. نویسندگان ادعا می‌کنند، ابزار بصری‌سازی اطلاعات از روند یادگیری علوم حدیث پشتیبانی خواهد کرد. آنها از فنون بصری‌سازی گراف برای نمایش راویان حدیث و پیوندهای بین آنها استفاده کردند. برای ارزیابی، بیست دانشجوی علم حدیث نمونه اولیه را آزمایش کردند. سپس، پرسشنامه‌ای برای مقایسه تجربه جدیدشان با روش سنتی به آنها ارائه شد. با توجه به مقیاس ۱ (بسیار دشوار) تا ۵ (ساده‌ترین)، میانگین نمره CHN[5] (مصورساز زنجیره راویان حدیث) برابر با ۴,۹۴ شد، در حالی که در روش سنتی برابر با ۲,۹۱ بود. اگرچه نتایج مثبت است، اما این نوع کار به مجموعه داده‌های آزمایش بزرگتر و اندازه‌گیری عملکرد دقیق‌تر نیاز دارد [۹۰]. در چهار مقاله‌ی چهارم تا هفتم، طراحی پایگاه داده حدیث با استفاده از XML یعنی طراحی واژه‌نامه‌ای برای علوم حدیث با استفاده از HPSG بررسی می‌شود [۹۱] و [۹۲]. هدف، سازماندهی و انتشار دانش علوم حدیث در قالبی واحد است که دستیابی به استفاده مجدد از محتوا و همکاری بین بخش‌های مختلف را میسر می‌کند. ویژگی متمایز طراحی HPSG اطلاعات یکپارچه سامانه است، که به آنها امکان می‌دهد چندین فراداده در مورد اسناد حدیث را در یک ساختار واحد به نام ماتریس مقدار ویژه[6] محصور کنند. با توجه به اهمیت اسناد، به ویژه برای قضاوت درباره صحت حدیث، مولفان ایده خود را بر روی اسناد آزمایش کردند. آنها برای تجزیه اسناد یک گرامر بدون متن ابداع کردند. جزئیات زنجیره روایت همراه با سایر اطلاعات در ساختار AVM ذخیره می‌شود [۹۳] و [۹۴]. در مقاله یک سامانه چند عاملی با پنج عامل تشکیل شده است که برای تحلیل و تولید درخت اسناد با هم همکاری می‌کنند. به عنوان مثال، عامل واژگانی، اسناد را به توکن‌هایی (به عنوان مثال نام راوی، لقب راوی) تجزیه می‌کند که نویسنده آن را اصطلاحاً ادات اسناد[7] می‌نامد. این سامانه از روش XML برای ذخیره داده‌های اسناد در پایگاه داده استفاده می‌کند

[1] Extra Feed-Forward
[2] Bidirectional Encoder Representations from Transformers
[3] گراف جهت دار بدون دور
[4] CHN: Chain of Hadith Narrators Visualizer
[5] Chain of Hadith Narrators Visualizer
[6] AVM: Attribute Value Matrix
[7] POI: Parts of Isnad

[۹۳]. در تحقیق هفتم، کار گسترش داده شده و در مورد استفاده از مدل مخفی مارکوف[1] برای شناسایی POI بحث شده است. این سامانه، POI را به چندین دسته طبقه‌بندی می‌کند، عمدتا به: نام راوی، پیشوند نام راوی، عنوان، عبارت روایت و نام پیامبر (ص) [۹۴]. در مقاله آخر، مولفان سه روش مختلف یادگیری ماشین SVM، NB و k-NN را بررسی کردند. هدف، رده‌بندی اسناد حدیث این است که آنها به کدام مجموعه تعلق دارند. SVM بهترین صحت را ۸۲٪ نشان می‌دهد، NB نزدیک به آن است، در حالی که مقدار k-NN با فاصله زیاد از آنها برابر با ۶۲٪ است [۹۵].

جدول ۲ جمع‌بندی پژوهش‌های مرتبط با صحت‌سنجی سند احادیث را نمایش می‌دهد.

۲-۳- پژوهش‌های مرتبط با هر دو حوزه متن و سند حدیث

در این قسمت به بررسی تحقیقات مربوط به هر دو حوزه متن و اسناد از وظایف پردازشی حدیث می‌پردازیم.

۲-۳-۱- ساخت هستان‌شناسی

یک مقاله در حوزه هستان‌شناسی به بررسی هر دو حوزه‌ی متن و سند حدیث، پرداخته است که در آن پیکره الحدیث ورد نت مبتنی بر الحدیث الشریف طراحی و پیاده‌سازی شده است. در این پژوهش، برای درک بهتر روابط معانی میان کلمات در حدیث، با استفاده از فرهنگ لغت‌های عربی سنتی و هستان‌شناسی حدیث، از روش جدیدی برای ایجاد الحدیث وردنت استفاده شده است. این پیکره، به شباهت معنایی و مترادف‌ها برای هر مجموعه، رابطه معنایی بین مجموعه‌ها و کتاب‌ها که هر کلمه به آن تعلق دارد، می‌پردازد. ارزیابی این پیکره مبتنی بر الگوریتم رده‌بندی PN[2] انجام شده است که PNها با در نظر گرفتن ۱٪ برای هر کلاس به عنوان کاهش ویژگی، از دقت عملکرد بسیار خوبی برخوردارند. متوسط دقت برابر با ۹۵,۴٪، فراخوان ۹۳,۵٪ و معیار F برابر با ۹۴,۵٪ است [۱۰].

جدول ۲. جمع‌بندی روش‌های صحت‌سنجی سند حدیث

ردیف	حوزه‌های پردازشی	پژوهش	روش‌ها و الگوریتم‌ها	مجموعه دادگان	مزایا و محدودیت‌ها
۱	**ساخت هستان‌شناسی**	دو مقاله [۷۶-۷۷]	RDF خودکار راویان حدیث و IJS	الالبانی	---
۲	**قطعه‌بندی سند حدیث**	یک مقاله [۵۰]	n-gram و رویکرد باقاعده	صحاح سته (بخاری، مسلم، ترمذی، ابن ماجه، سنن ابی داود، نسائی)	طراحی و ساخت یک پیکره موازی دو زبانه عربی-انگلیسی، -طراحی ابزار بخش‌بندی، خودکار جهت تفکیک متن از اسناد
۳	**تشخیص اسامی راویان**	۱۰ مقاله [۷۸-۸۷]	قاعده‌کاوی، مبدل حالت محدود، برچسب‌گذاری ادات سخن و الگوریتم‌های یادگیری عمیق مبتنی بر BERT	بخاری، اصول کافی، مسلم و ابن حنبل، صحاح سته و موطا مالک و التهذیب	حل مسئله هجی‌های مختلف نام راویان برای مالایی‌ها، طراحی سامانه‌ای که قادر است با یک فراداده، راویان موجود و سلسله احتمالی زنجیره را تشخیص دهد دقت ناکافی شناسایی موجودیت‌های نامدار در زبان‌هایی مانند عربی، فارسی
۴	**تحلیل زنجیره راویان**	۱۱مقاله [۵۱-۶۱]	تحلیل شبکه اجتماعی راویان حدیث، اتوماتای حالت محدود، چارچوب توصیف منابع برای تولید بازنمایی زنجیره راویان، تئوری گراف، سامانه خبره فقه	بخاری و مسلم اصول کافی، ابن حنبل کتب اربعه شیعه (کافی، تهذیب، استبصار و من لایحضره الفقیه) و ترمیذی	کشف راویان مرکزی، تعیین سهم حفاظت راویان از روایات نبوی، طراحی ابزاری نرم‌افزاری برای تولید خودکار تصویری گرافیکی از زنجیره‌های انتقال کامل حدیث
۵	**اعتبارسنجی حدیث**	۱۴ مقاله [۶۲-۷۵]	منطق فازی، قاعده‌کاوی، رده‌بندی صحت حدیث با الگوریتم‌های یادگیری عمیق	اصول کافی، الرساله الشافعی، الالبانی، کتب سته	تحلیلی بر تولید سلسله مراتبی با سطوح مختلف، پیش بینی نام راویان و سایر POIها برای اسناد
۶	**بصری‌سازی اطلاعات**	۸ مقاله [۸۸-۹۵]	الگوریتم DAG، احراز هویت شواهد دیجیتال، مصورساز زنجیره راویان حدیث	بخاری و ترمذی	طراحی پایگاه داده بیوگرافی راویان و بازنمایی و تحلیل اسناد حدیث، سازماندهی و انتشار دانش علوم حدیث در قالبی واحد

[1] HMM: Hidden Markov Model

[2] Polynomial Network Classification Phase

۲-۳-۲- بازیابی اطلاعات

دو مقاله در حوزه بازیابی اطلاعات به بررسی هر دو حوزه‌ی متن و سند حدیث، پرداخته است که در مقاله اول، محققان پیاده‌سازی سامانه منطق فازی سلسله مراتبی را با استفاده از مدل BM25 در سامانه بازیابی اطلاعات حدیث مالایی ارائه دادند. مدل BM25 الحاقی مدل احتمالی است. مدل احتمالی یکی از مدل‌های بازیابی اطلاعات کلاسیک است. برخلاف مفاهیم اولیه از مدل احتمالی کلاسیک، مدل BM25 را می‌توان بدون هیچ‌گونه اطلاعات مربوط به کاربر ارائه داد. در این تحقیق، از روش خلاصه‌سازی خودکار متن برای ایجاد خلاصه‌ای از هر حدیث در پیکره استفاده شده است که می‌تواند در محاسبه نرخ مثبت استفاده شود. پس از آن، نرخ مثبت یکی از چهار ورودی کنترل‌کننده منطق فازی سلسله مراتبی سیستم استنتاج فازی از نوع ممدانی براساس مطالعه قبلی محققان در مورد ورودی مانند (امتیاز هستان‌شناسی BM25، نرخ ساخت حدیث و میزان شیعه بودن حدیث) خواهد بود. مدل پیشنهادی مقاله، در کل مقادیر چهار ورودی ذکر شده و چهار خروجی امتیاز نهایی رتبه‌بندی را بررسی می‌کند که از سه تابع عضویت مثلثی تشکیل شده‌اند. در نتیجه این مقاله شامل دو مرحله کلی است: الف- نمره BM25 هستان‌شناسی و نرخ مثبت حدیث به عنوان شاخص رتبه‌بندی مثبت جهت فراخوان اسناد مشاهده نشده و ارتقای اسناد مثبت به بالای لیست رتبه‌بندی و ب- حدیث شیعی ترجمه شده به مالایی به عنوان شاخص رتبه بندی منفی در معادله جدید تابع رتبه‌بندی، برای تنزل سند منفی به پایین نتایج رتبه‌بندی. نرخ موضوع حدیث، نرخ حدیث شیعه و نرخ مثبت حدیث از قوانینی مشتق شده از ویژگی‌های منحصر به فرد مثل نام راویان از اسناد، اصطلاحات خاص از متن هم‌چنین خلاصه متن به ۱۰ کلمه برای هر حدیث در پیکره انتخابی بدست آمده است. در این تحقیق، یک کنترل‌کننده منطق فازی سلسله مراتبی از سامانه استنتاج فازی از نوع ممدانی برای تعریف تابع رتبه‌بندی بر اساس مدل BM25 ساخته شده است. این مدل چهار ورودی را بررسی می‌کند که عبارتند از: امتیاز BM25 هستان‌شناسی، نرخ موضوع حدیث، نرخ حدیث شیعه از کارهای قبلی محققان و نرخ مثبت جدید اضافی حدیث. محققان از ۳۰ پرس‌وجوی کلی در هشت موضوع برای آزمایش استفاده کردند. براساس نتایج و ارزیابی تجارب این کار، سامانه پیشنهادی از امتیاز اصلی BM25 و مدل فضای برداری در ۵ موضوع پرس‌وجو و ۲۶ پرس وجو در اصطلاح پرس‌وجوهای فردی بهتر بوده است. نتایج نشان می‌دهد سامانه پیشنهادی، توانایی پایین آوردن اسناد منفی و بالا بردن اسناد مربوطه در لیست رتبه‌بندی با شاخص مثبت و قابلیت یادآوری سند دیده نشده با استفاده از هستان‌شناسی در بازیابی متن را دارد [۹۶]. در مقاله آخر، چارچوبی چندزبانه برای استخراج داده‌های حدیث از منابع معتبر حدیث ارائه شده است. این مقاله در مورد تهیه انباره پایگاه داده بحث می‌کند و توضیح می‌دهد که چگونه داده‌ها در قالب مجموعه داده یا پایگاه داده معمولی استخراج می‌شوند تا بتوانند متن و تحلیل داده‌های بیشتری را انجام دهند. معیارهای ارزیابی دقت، فراخوانی و F جهت سنجش عملکرد روشهای این مطالعه استفاده شد که به دلیل تفاوت ساختار کتاب‌های مختلف، سطح دقت متفاوت است. معیار ارزیابی دقت بین ۹۶٪ تا ۱۰۰٪ برای سنن مسلم، بخاری، ابوداود و مالک است، معیار ارزیابی فراخوانی بین ۹۱٪ تا ۱۰۰٪ برای سنن مسلم، بخاری، ابوداود و مالک است و معیار ارزیابی F بین ۹۳٪ تا ۱۰۰٪ برای سنن مسلم، بخاری، ابوداود و مالک است [۹۷].

۲-۳-۳- قطعه‌بندی حدیث

سه مقاله در حوزه قطعه‌بندی به بررسی هر دو حوزه‌ی متن و سند حدیث، پرداخته است که در مقاله اول، از روش‌های پردازش زبان طبیعی، شیوه N-gram برای طراحی سامانه‌ای استفاده شد که به طور خودکار حدیث را به دو جز متن و سند تقسیم می‌کند. یافته‌های این مطالعه به وضوح نشان می‌دهد که استفاده از دو گرم[1] برای تقسیم‌بندی[2] حدیث بهتر از سه گرم[3] عمل می‌کند زیرا داده‌های آموزشی مقاله، محدود است. حدیث را می‌توان به بخش‌های جزئی‌تری تقسیم کرد که فراتر از دو بخش متن و سند باشد. چون برخی از احادیث حاوی اطلاعاتی در سند هستند که توسط این سامانه به عنوان بخش‌های متن شناسایی شده‌اند. به عنوان مثال، سند یک حدیث ممکن است شامل اطلاعاتی در میان سلسله راویان در خصوص محل تولد و زندگی راوی خاصی باشد. هدف اصلی این مطالعه، ایجاد سامانه‌ای است که مولفه‌های حدیث شامل اسناد و متن را تفکیک و حاشیه‌نویسی کند. نتیجه نشان می‌دهد که دو گرم در شناسایی بخش‌های حدیث با صحت ۹۲,۵٪ موثر است [۹۸]. در مقاله دوم، نگرشی جهت تشخیص شباهت احادیث با استفاده از الگوریتم Doc2vec مطرح شده است.این مطالعه، یک مدل رده‌بندی با قابلیت شناسایی احادیث مشابه برحسب متن و سند احادیث، با استفاده از سناریوهای مختلف ایجاد کرده است. از کل ۹ مجموعه کتاب مورد استفاده به عنوان داده‌های

[1] Bi-grams
[2] Segmentation
[3] Tri-grams

این مطالعه، ۸ کتاب حدیث به همراه متن و سند، جهت آموزش و کتاب نهم با متن و سند، برای آزمون مورد استفاده قرار گرفته است. در این مقاله، الگوریتم بردار پاراگراف[1] معروف به Doc2Vec به عنوان تکنیک تشابه‌یابی استفاده می‌شود. یکی از مزایای این مدل این است که از داده‌های بدون برچسب آموزش می‌بینند. یکی از اشکالاتی که در مدل Doc2Vec وجود دارد این است که برای آموزش به تعداد زیادی متن احتیاج دارد. نتایج این مطالعه نشان می‌دهد که این مدل قادر است تا شباهت ۸۰٪ بین احادیث را تشخیص دهد [۹۹]. در مقاله آخر که در شیوه تشخیص اسامی راویان از حوزه سند حدیث بررسی شد؛ از شیوه قطعه‌بندی نیز استفاده شده است؛ در شیوه اخیر، یک سامانه بازیابی اطلاعات با استفاده از FST[2] جهت استخراج موجودیت‌ها و پردازش محتوای حدیث با استفاده از برچسب‌گذاری ادات سخن[3]، CRF[4] برای برچسب‌گذاری متن و سند احادیث و NER[5] جهت استخراج موجودیت‌های نامدار استفاده شده و نهایتا این موجودیت‌های نامدار در رده‌های مختلف رده‌بندی شده است. نتایج ارزیابی برچسب‌گذاری POS برای دقت ۹۶٪ و فراخوانی ۸۹٪ و معیار F برابر با ۹۲٪ گزارش شده است [۸۲].

۲-۳-۴- رده‌بندی حدیث

در دو مقاله، مدل رده‌بندی متن و سند حدیث با توجه به حافظه و قابلیت اطمینان راویان حدیث ارائه شده است که می‌تواند حدیث را به طور خودکار بر اساس روش‌های یادگیری ماشین تشخیص داده و به کلاس‌های صحیح، حسن، ضعیف و موضوع (مجعول) رده‌بندی کند. این رده‌بندی مانند سایر متون عربی فقط به متن حدیث بستگی ندارد، بلکه به سند حدیث نیز بستگی دارد. نتایج تجربی نشان داد که افزودن سند حدیث به متن در فرآیند رده‌بندی تأثیر قابل توجهی در افزایش صحت رده‌بندی دارد. صحت رده‌بند درخت تصمیم مبتنی بر سند حدیث تا ۹۳٪ گزارش شده است. در این مطالعه چندین الگوریتم یادگیری استفاده شده است، اما بهترین آنها، سه رده‌بند LinearSVC، SGDClassifier و رگرسیون لجستیک است که با صحت بالاتری به ترتیب به ۹۴٪ ، ۹۴٪ و ۹۲ ٪ رسیده‌اند [۱۰۰]. در مقاله دوم، به ضعف کم توجهی به پردازش‌های مبتنی بر حوزه سند حدیث پرداخته شده است. از آنجایی که اکثر مطالعات در رده‌بندی متون حدیثی بر قسمت محتوایی به نام حوزه رده‌بندی موضوعی متمرکز است. برعکس، بخش سند در حوزه رده‌بندی متن کمتر مورد توجه قرار می‌گیرد. در این تحقیق قواعد رده‌بندی را برای رده‌بندی متن حدیث بر اساس قسمت سند با استفاده از مفهوم فهرست لبه طراحی شده است. رده‌بندی مبتنی بر قاعده با استفاده از ۷۰۰ متن حدیثی به عنوان مجموعه داده‌های آموزشی و ۳۰۰ متن حدیثی برای مجموعه داده‌های آزمایشی طراحی شده است. نتایج، با استفاده از معیارهای دقت و نرخ یادآوری (فراخوانی) ارزیابی شده است. ارزش معیار دقت با اجرای رده‌بندی مبتنی بر قاعده بر اساس نظر خبرگان ۱,۰۰ است. اما نرخ یادآوری ۰,۱۱ شده است، زیرا فهرست‌های ترتیب زمانی متن حدیث در مجموعه آموزشی، همه فهرست‌های زمانی در حدیث بخاری را نشان نمی‌داد با اینکه بیش از ۷۰۰۰ است. در آینده، می‌توان این مطالعه را برای تحلیل رده‌بندی کلی متن بخاری گسترش داد [۱۰۱].

۲-۳-۵- اعتبارسنجی حدیث

یک مقاله در شیوه اعتبارسنجی حدیث از حوزه بررسی متن و سند، تحقیقات مفصلی را در مورد راه‌های تشخیص خودکار صحت حدیث در متون حدیثی عربی ارائه می‌کند. این پژوهش به بررسی استفاده از رده‌بندهای مبتنی بر یادگیری عمیق، پیش‌بینی مبتنی بر تطبیق جزئی (PPM) و فشرده‌سازی می‌پردازد، که قبلاً در تشخیص اعتبارسنجی حدیث استفاده نشده‌اند. روش‌های پیشنهادی با جدیدترین روش مورد استفاده که یادگیری ماشینی است، مقایسه شد. علاوه بر این، شرح مفصلی از مجموعه حدیث عربی جدید (پیکره حدیث غیر معتبر) که برای این مطالعه و آزمون‌های مولفان ایجاد شده است، وجود دارد که از پیکره حدیث دانشگاه لیدز و دانشگاه ملک سعود نیز استفاده کرده است. طبق آزمایش‌ها، اعتبارسنجی بر اساس اسناد صحتی در محدوده ۸۴٪ تا ۹۳٪ است. اعتبارسنجی بر اساس متن، محدوده صحت ۵۵٪ تا ۹۳٪ را به دست آورد، در حالی که محدوده صحت این آزمایش از ۵۵٪ تا ۸۵٪ بود و به این معنی است که اسناد موثرترین قسمت حدیث، برای تشخیص خودکار اعتبارسنجی است. علاوه بر این، آزمایش ثابت کرد که می‌توان از متن برای قضاوت در اعتبارسنجی حدیث با صحت ۸۵ درصد استفاده کرد. این مطالعه همچنین نشان داد که رده‌بندهای PPM و یادگیری عمیق، ابزارهای مؤثری برای تشخیص خودکار احادیث معتبر هستند [۱۰۲]. جدول ۳ جمع‌بندی روش‌های صحت‌سنجی در هر دو حوزه‌ی متن و سند حدیث را نمایش می‌دهد.

[1] Paragraph Vector
[2] Finite State Transducer
[3] Part of Speech
[4] Conditional Random Field
[5] Named Entity Recognition

جدول ۳. جمع‌بندی روش‌های صحت‌سنجی متن و سند حدیث

ردیف	حوزه‌های پردازشی	پژوهش	روش‌ها و الگوریتم‌ها	مجموعه دادگان	مزایا و محدودیت‌ها
۱	**ساخت هستان‌شناسی**	یک مقاله [۱۰]	الگوریتم رده‌بندی PN	ورد نت مبتنی بر الحدیث الشریف	استفاده از روش جدیدی برای ایجاد الحدیث وردنت جهت درک بهتر روابط معانی میان کلمات در حدیث، با استفاده از فرهنگ لغت‌های عربی سنتی و هستان‌شناسی حدیث، بررسی شباهت معنایی و مترادف‌ها برای هر مجموعه، رابطه معنایی بین مجموعه‌ها و کتاب‌ها
۲	**قطعه‌بندی متن حدیث**	۳مقاله [۹۸-۹۹] و [۸۲]	N-gram -شباهت احادیث با استفاده از الگوریتم Doc2vec FST، CRF و NER	۹ مجموعه کتاب	تقسیم‌بندی بهتر با دو گرمی در مقابل سه گرمی برای داده‌های محدود
۳	**رده‌بندی متن حدیث**	۲مقاله [۱۰۰-۱۰۱]	درخت تصمیم، رده‌بند LinearSVC، SGDClassifier و رگرسیون لجستیک	بخاری	-رده‌بندی متن و سند حدیث با توجه به حافظه و قابلیت اطمینان راویان حدیث
۴	**بازیابی اطلاعات**	دو مقاله [۹۶-۹۷]	-منطق فازی سلسله مراتبی با استفاده از مدل BM25	کتب سته	ایجاد خلاصه‌ای از هر حدیث در پیکره با روش خلاصه‌سازی خودکار متن جهت محاسبه نرخ مثبت قابلیت یادآوری سند دیده نشده با استفاده از هستان‌شناسی در بازیابی متن طراحی انباره پایگاه داده چندزبانه از منابع معتبر حدیث
۵	**اعتبارسنجی حدیث**	یک مقاله [۱۰۲]	رده‌بندهای مبتنی بر یادگیری عمیق، پیش‌بینی مبتنی بر تطبیق جزئی (PPM) و فشرده‌سازی	پیکره حدیث دانشگاه لیدز و ملک سعود	بررسی استفاده از رده‌بندهای مبتنی بر یادگیری عمیق، پیش‌بینی مبتنی بر تطبیق جزئی (PPM) و فشرده‌سازی اسناد موثرترین قسمت حدیث، برای تشخیص خودکار اعتبارسنجی است

۳- یافته‌ها

از ۱۰۱ پژوهش بررسی شده، ۴۷٪ آنها در حوزه پردازش متن حدیث از روش‌های ساخت هستان‌شناسی، قطعه‌بندی متن حدیث، پرسش و پاسخ، شباهت‌یابی حدیث، رده‌بندی متن حدیث، بازیابی اطلاعات، نمایه‌سازی مولف، تحلیل ریخت‌شناسی و اعتبارسنجی حدیث استفاده نموده‌اند. ۴۵٪ پژوهش‌ها نیز در حوزه پردازش اسناد حدیث با روش‌های ساخت هستان‌شناسی، قطعه‌بندی حدیث، تحلیل زنجیره راویان، اعتبارسنجی حدیث، تشخیص اسامی راویان و بصری‌سازی اطلاعات تحقیق نموده‌اند. ۸٪ پژوهش‌ها نیز از هر دو منظر پردازش صحت متن و سند، احادیث را بررسی نموده‌اند و از روش‌های ساخت هستان‌شناسی، قطعه‌بندی حدیث، بازیابی اطلاعات، رده‌بندی حدیث و اعتبارسنجی حدیث بهره جسته‌اند. جزئیات تخصیص هر روش پردازشی حدیث از میان ۱۰۱ پژوهش بررسی شده به هر حوزه پردازش آن در جدول ۴ به تفصیل شرح داده شده است.

در ۶ پژوهش، به طور همزمان، از دو روش متفاوت برای پردازش هوشمند احادیث استفاده شده است که منجر به اختلاف جمع تعداد پژوهش‌های بررسی شده با جمع روش‌ها در همه‌ی حوزه‌های پژوهشی شده است.

جدول ۴. جزئیات تخصیص هر روش پردازشی به حوزه پردازش آن

روش‌های پردازشی حدیث	حوزه پردازشی متن	حوزه پردازشی سند	حوزه پردازشی متن و سند	جمع
هستان‌شناسی	۹	۲	۱	۱۲
رده‌بندی	۱۶	-	۲	**۱۸**
قطعه‌بندی	۱	۱	۳	۵
پرسش و پاسخگویی	۲	-	-	۲
شباهت‌یابی	۲	-	-	۲
بازیابی اطلاعات	۱۵	-	۲	۱۷
نمایه‌سازی مولف	۲	-	-	۲
تحلیل ریخت‌شناسی	۲	-	-	۲
اعتبارسنجی حدیث	۱	۱۴	۱	۱۶
تحلیل زنجیره راویان	-	۱۱	-	۱۱
تشخیص اسامی راویان	-	۱۰	-	۱۰
بصری‌سازی اطلاعات	-	۱۰	-	۱۰
جمع	**۵۰**	**۴۸**	**۹**	**۱۰۷**
درصد	**٪۴۷**	**٪۴۵**	**٪۸**	**٪۱۰۰**

با تحلیل جدول فوق، مشخص می‌شود که روش رده‌بندی بیشترین کاربرد را در پردازش هوشمند احادیث از میان سایر روش‌ها داشته است و این روش، بیشتر در حوزه پردازش متن حدیث استفاده شده است.

البته روش بازیابی اطلاعات حدیث در رتبه بعد از پرکاربردها قرار دارد که آن هم بیشتر در حوزه متن مورد بهره‌برداری قرار گرفته است. روش‌های پرسش و پاسخگویی، شباهت‌یابی، نمایه‌سازی مولف و تحلیل ریخت‌شناسی، کمترین کاربرد را در پردازش هوشمند احادیث از میان سایر روش‌ها داشته‌اند. از میان ۳ حوزه پردازشی احادیث، محققان بیشتر به بررسی متن یا سند احادیث به طور مجزا اکتفا کرده‌اند و فقط ۸٪ مقالات همزمان هر دو حوزه متن و سند را لحاظ نموده است.

۴- جمع‌بندی و پیشنهادات

در این پژوهش به بررسی روش‌های هوشمند پردازشی حدیث مبتنی بر حوزه‌های پردازش متن و سند احادیث و یا هر دو مورد پرداخته شده است. روش‌های پردازشی حدیث مبتنی بر هر حوزه‌ای به طور مختلف دسته‌بندی گردیده است. از دوازده روش پردازشی هوشمند احادیث، روش رده‌بندی بیشترین کاربرد را از میان سایر روش‌ها داشته است و این روش، بیشتر در حوزه پردازش متن حدیث مورد استفاده قرار گرفته است. روش‌های پرسش و پاسخگویی، شباهت‌یابی، نمایه‌سازی مولف و تحلیل ریخت‌شناسی، کمترین کاربرد را در پردازش هوشمند احادیث از میان سایر روش‌ها داشته‌اند. از میان ۳ حوزه پردازشی احادیث، محققان بیشتر به بررسی متن یا سند احادیث به طور مجزا اکتفا کرده‌اند و فقط ۸٪ مقالات همزمان هر دو حوزه متن و سند را لحاظ نموده است. لذا برای کارهای آتی، پیشنهاد می‌شود که محققان از روش‌های کم‌کاربرد ذکر شده و هم چنین بررسی توامان حوزه متن و سند استفاده کنند.

مراجع

[۱] ک. ایزدی مبارکه and م. مجتبی, "ملاکهای نقد حدیث از منظر استاد علی اکبر غفاری," *پژوهش دینی*, vol. 12, pp. 151–169, 1384.

[2] S. Saad, N. Salim, and H. Zainal, "Islamic knowledge ontology creation," *Int. Conf. Internet Technol. Secur. Trans. ICITST 2009*, no. November, 2009.

[3] S. Saad, N. Salim, H. Zainal, and S. A. M. Noah, "A framework for Islamic knowledge via ontology representation," *Proc. - 2010 Int. Conf. Inf. Retr. Knowl. Manag. Explor. Invis. World, CAMP'10*, no. July 2014, pp. 310–314, 2010.

[4] [A. S. Harrag F, Alothaim A, Abanmy A, Alomaigan F, "Ontology Extraction Approach for Prophetic Narration (Hadith) using Association Rules," *Int. J. Islam. Appl. Comput. Sci. Technol.*, vol. 1, no. 2, pp. 48–57, 2013, Accessed: Apr. 19, 2021.

[5] A.-S. A. Al-Arfaj A, "Towards ontology construction from Arabic texts – a proposed framework," *IEEE Int. Conf. Comput. Inf. Technol.*, pp. 737–742, 2014.

[6] M. Ghanem, A. Mouloudi, and M. Mourchid, "Creation and populating of an Islamic knowledge ontology using extraction pattern bootstrapping," in *Third National Day on Engineering, Networks and Telecommunications (NDENT 2015)*, 2015, pp. 36–39.

[7] A. H. Jaafar, N. C. Pa, A. Hamzah Jaafar, and N. Che Pa, "Hadith Commentary Repository: An Ontological Approach," *Proc. 6th Int. Conf. Comput. Informatics*, no. 167, pp. 191–198, 2017.

[8] H. A. Al-Sanasleh and B. H. Hammo, "Building domain ontology: Experiences in developing the prophetic ontology form Quran and Hadith," in *Proceedings - 2017 International Conference on New Trends in Computing Sciences, ICTCS 2017*, 2017, vol. 2018-Janua, pp. 223–228.

[9] N. Soudani, I. Bounhas, B. Elayeb, and Y. Slimani, "Toward an Arabic ontology for Arabic word sense disambiguation based on normalized dictionaries," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8842, pp. 655–658, 2014.

[10] M. Alkhatib, A. A. Monem, K. Shaalan, and S. K. Alkhatib M, Monem AA, "A rich Arabic WordNet resource for al-hadith al-shareef," in *Procedia Computer Science 117*, 2017, vol. 117, pp. 101–110.

[11] E. D. Sri Mulyani, N. Nelis Febriani SM, A. Darmawan, R. A. Wiyono, R. Deli Saputra, and D. Rohpandi, "Keyword-Based Hadith Grouping Using Fuzzy C-Means Method," in *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, Oct. 2020, pp. 1–6.

[12] H. Sayoud, "Automatic authorship classification of two ancient books: Quran and Hadith," *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA*, vol. 2014, pp. 666–671, 2014.

[13] H. M. Hanum, Z. A. Bakar, N. A. Rahman, M. M. Rosli, and N. Musa, "Using Topic Analysis for Querying Halal Information on Malay Documents," *Procedia - Soc. Behav. Sci.*, vol. 121, no. March, pp. 214–222, 2014.

[14] A.-M. R. Al-Kabi MN, Kanaan G, Al-Shalabi R, Al-Sinjilawi SI, M. N. A.- Kabi, G. Kanaan, R. A.- Shalabi, S. I. A.- Sinjilawi, and R. S. A.- Mustafa, "Al-hadith text classifier," *J. Appl. Sci.* 5, vol. 5, no. 3, pp. 584–587, Feb. 2005.

[15] A.-S. S. Al-Kabi MN, M. N. Al-kabi, and S. I. A.- Sinjilawi, "A comparative study of the efficiency of different measures to classify Arabic text," *Univ. Sharjah J. Pure Appl. Sci.*, vol. 4, no. 2, pp. 13–26, 2007.

[16] F. Harrag and A. Hamdi-Cherif, "UML Modeling of Text Mining in Arabic Language Application to the Prophetic Traditions 'Hadiths,'" *1st Int. Symp. Comput. Arab. Lang.*, no. August, 2007.

[17] F. Harrag and E. El-Qawasmah, "Neural network for Arabic text classification," *2nd Int. Conf. Appl. Digit. Inf. Web Technol. ICADIWT 2009*, pp. 778–783, 2009.

[18] F. Harrag, E. El-Qawasmeh, and P. Pichappan, "Improving Arabic text categorization using decision trees," in *2009 1st International Conference on Networked Digital Technologies, NDT 2009*, 2009, no. August, pp. 110–115.

[19] F. Harrag, E. El-Qawasmah, and A. M. S. Al-Salman, "Comparing dimension reduction techniques for Arabic text classification using BPNN algorithm," in *Proceedings - 1st International Conference on Integrated Intelligent Computing, ICIIC 2010*, 2010, pp. 6–11.

[20] K. Jbara, "Knowledge Discovery in Al-Hadith Using Text Classification Algorithm," *J. Am. Sci.*, vol. 6, no. 11, pp. 485–494, 2010.

[40] S. Mohamed, O. Hassan, and E. Atwell, "Concept Search Tool for Multilingual Hadith Corpus," *Int. J. Sci. Res.*, vol. 5, no. 4, pp. 1326–1328, 2016.

[41] A. Aulia, D. Khairani, and N. Hakiem, "Development of a retrieval system for Al Hadith in Bahasa (case study: Hadith Bukhari)," *2017 5th Int. Conf. Cyber IT Serv. Manag. CITSM 2017*, Oct. 2017.

[42] A. Aulia, D. Khairani, R. B. Bahaweres, and N. Hakiem, "WatsaQ: Repository of Al Hadith in Bahasa (Case study: Hadith Bukhari)," in *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Dec. 2017, vol. 2017-Decem.

[43] S. M. O. Hassan, "Building the Multilingual Hadith Corpus to Enhance Performance of Information Retrieval System for Hadith," Sudan University of Science and Technology, 2017.

[44] F. Harrag, A. Hamdi-Cherif, A. Al-Salman, and E. ElQawasmeh, "Experiments in Improvement of Arabic Information Retrieval," in *Third International Conference on Arabic Language Processing (CITALA '09)*, 2009, pp. 71–81.

[45] S. T. Abd Rahman N, Abu Bakar Z, "Query expansion using thesaurus in improving Malay Hadith retrieval system," *IEEE Int. Symp. Inf. Technol.*, vol. 3, pp. 1404–1409, 2010.

[46] H. Sayoud, "Author discrimination between the holy Quran and Prophet's statements," *Lit. Linguist. Comput.*, vol. 27, no. 4, pp. 427–444, 2012.

[47] H. Sayoud, "AUTHORSHIP DISCRIMINATION ON QURAN AND HADITH USING DISCRIMINATIVE LEAVE-ONE-OUT CLASSIFICATION," 2017.

[48] R. Ayed, I. Bounhas, B. Elayeb, N. B. Ben Saoud, F. Evrard, and E. F. Ayed R, Bounhas I, Elayeb B, Saoud NBB, "Improving Arabic texts morphological disambiguation using a possibilistic classifier," in *19th International Conference on Application of Natural Language to Information Systems*, 2014, vol. 8455 LNCS, pp. 138–147.

[49] M. Q.Shatnawi, Q. Q. Abuein, and O. Darwish, "Verifying Hadith Correctness in Islamic Web Pages using Information Retrieval Techniques," *Int. J. Comput. Appl.*, vol. 44, no. 13, pp. 47–50, 2012.

[50] S. Altammami, E. Atwell, and A. Alsalka, "Constructing a bilingual hadith corpus using a segmentation tool," *Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, no. May, pp. 3390–3398, 2020.

[51] S. Saeed, S. Yousuf, F. Khan, and Q. Rajput, "Social network analysis of Hadith narrators," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, Feb. 2021.

[52] T. Alam and J. Schneider, "Social Network Analysis of Hadith Narrators from Sahih Bukhari," Nov. 2020.

[53] J. Makhlouta and H. Harkous, "AUBSarf: Compositional Non-deterministic Finite-state Automata for Arabic Morphological Analysis," 2010.

[54] A. Azmi and N. Bin Badia, "iTree - Automating the construction of the narration tree of Hadiths (prophetic traditions)," *Proc. 6th Int. Conf. Nat. Lang. Process. Knowl. Eng. NLP-KE 2010*, no. September 2010.

[55] A. N. Azmi A, A. Azmi, and N. Al Badia, "Mining and Visualizing the Narration Tree of Hadiths (Prophetic Traditions)," *Cross-Disciplinary Adv. Appl. Nat. Lang. Process. Issues Approaches*, no. January 2011, pp. 493–510, 2012.

[56] A. M. Azmi and N. Bin Badia, "e-Narrator - an application for creating an ontology of Hadiths narration tree semantically and graphically," *Arab. J. Sci. Eng.*, vol. 35, no. 2 C, pp. 51–68, 2010.

[57] M. A. Siddiqui, M. E. Saleh, and A. A. Bagais, "Extraction and Visualization of the Chain of Narrators from Hadiths using Named Entity Recognition and Classification," *Int. J. Comput. Linguist. Res.*, vol. 5, no. 1, pp. 14–25, 2014.

[58] N. Alias, N. A. Rahman, N. K. Ismail, Z. M. Nor, and M. N. Alias, "Searching Algorithm of Authentic Chain of Narrators' in Shahih Bukhari Book," in *2016 International Conference on Applied*

[21] A. I. Al-Kabi MN, Wahsheh HA, "A topical classification of hadith Arabic text," 2014.

[22] A.-A. A. Al-Kabi MN, Wahsheh HA, Alsmadi IM, "Extended topical classification of hadith Arabic text," *Int. J. Islam. Appl. Comput. Sci. Technol.*, vol. 3, no. 3, pp. 13–23, 2015.

[23] K. Faidi, R. Ayed, I. Bounhas, and B. Elayeb, "Comparing Arabic NLP tools for Hadith Classification," *Comput. Sci.*, 2015.

[24] M. A. Saloot, N. Idris, R. Mahmud, S. Ja'afar, D. Thorleuchter, and A. Gani, "Hadith data mining and classification: a comparative analysis," *Artif. Intell. Rev.*, vol. 46, no. 1, pp. 113–128, 2016.

[25] I. Khalaf Alshammari, E. Atwell, M. Ammar Alsalka, H. Al-Batin, and K. M. of Saudi Arabia, "Evaluation of Arabic Named Entity Recognition Models on Sahih Al-Bukhari Text." EasyChair, Jan. 16, 2023.

[26] K. Gaanoun and M. Alsuhaibani, "Fabricated Hadith Detection: A Novel Matn-Based Approach With Transformer Language Models," *IEEE Access*, vol. 10, pp. 113330–113342, 2022.

[27] H. Maraoui, K. Haddar, and L. Romary, *Segmentation Tool for Hadith Corpus to Generate TEI Encoding*, vol. 845. Springer International Publishing, 2019.

[28] N. Neamah and S. Saad, "Question answering system supporting vector machine method for hadith domain," *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 7, pp. 1510–1524, 2017.

[29] A. Abdi, S. Hasan, M. Arshi, S. M. Shamsuddin, and N. Idris, "A question answering system in hadith using linguistic knowledge," *Comput. Speech Lang.*, vol. 60, 2020.

[30] N. K. Ismail, N. H. M. Saad, S. B. S. Omar, and T. M. T. Sembok, "2D visualization of terms and documents in Malay language," in *2013 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M)*, Mar. 2013, pp. 1–6.

[31] H. Juzi, A. R. Zadeh, E. Barati, and B. Minaei-Bidgoli, "A new framework for detecting similar texts in Islamic Hadith Corpora," *Lr. Lang. Resour. Eval. Relig. Texts*, pp. 38–41, 2012.

[32] F. Harrag, "Text mining approach for knowledge extraction in Sahîh Al-Bukhari," *Comput. Human Behav.*, vol. 30, pp. 558–566, 2014.

[33] M. K. A. B. Zainudin and R. M. Rias, "M-Hadith: Retrieving Malay Haditli text in a mobile application," *ISCAIE 2012 - 2012 IEEE Symp. Comput. Appl. Ind. Electron.*, no. Iscaie, pp. 60–63, 2012.

[34] A. R. Saeed and S. W. Jaffry, "Information Mining from Muslim Scriptures," *4th Work. South Southeast Asian NLP (WSSANLP), Int. Jt. Conf. Nat. Lang. Process.*, no. October, pp. 66–71, 2013.

[35] I. Rasyidi, A. Romadhony, and A. T. Wibowo, "Indonesian Hadith Retrieval System using thesaurus," *Proceeding - 2013 Int. Conf. Comput. Control. Informatics Its Appl. "Recent Challenges Comput. Control Informatics", IC3INA 2013*, pp. 285–288, 2013.

[36] A. Azmi, F. Alkhalifah, A. Alsaeed, and Y. Barnawi, "Using non-conventional search schemes to retrieve Hadiths," in *5th International Conference on Arabic Language Processing (CITALA '14)At: Oujda, Morocco*, 2014, no. November, pp. 125–129.

[37] N. A. Rahman, Z. Mabni, N. Omar, H. F. M. Hanum, N. N. A. T. Mohamad Rahim, and R. N. Abd Rahman N, Mabni Z, Omar N, Hanum HFM, "A parallel latent semantic indexing (LSI) algorithm for Malay hadith translated document retrieval," *Int. Conf. Soft Comput. Data Sci. Springer*, vol. 545, pp. 154–163, 2015.

[38] P. N. E. Nohuddin and J. M. Zainol Z, Chao KF, Nordin AI, "Keyword based clustering technique for collections of Hadith chapters," *Int. J. Islam. Appl. Comput. Sci. Technol.*, vol. 4, no. 3, pp. 11–18, 2016.

[39] Nurul Syeilla Syazhween, Nurazzah Abdul Rahman, and Zainab Abu Bakar, "Analyzing search retrieval results on Malay ranslated Hadith text documents," *Int. Conf. Appl. Comput. Math. Sci. Eng. May 2016(ACME)*, no. June, 2016.

Computing, Mathematical Sciences and Engineering (ACME2016), 2016, no. May, pp. 60–66.

[۵۹] ب. م. بیدگلی, "پایگاه اطلاعاتی خبرهء علم رجال," وزارت علوم، تحقیقات و فناوری - دانشگاه علم و صنعت ایران, ۱۳۷۶.

[60] T. Helmy and A. Daud, "Intelligent agent for information extraction from arabic text without machine translation," *CEUR Workshop Proc.*, vol. 687, no. February, 2010.

[61] A. M. Azmi, "A novel method to automatically pass hukm on Hadith," *5th Int. Conf. Arab. Lang. Process.*, no. August, pp. 118–124, 2014.

[62] A. HM, "The use of fuzzy logic for exploring the words of the critics of the men of hadith (in Arabic)," in *Islamiyyat Al-Ma'rifa 48*, 2008, pp. 103–132.

[63] M. Ghazizadeh, M. H. Zahedi, M. Kahani, and B. Minaei Bidgoli, "Fuzzy expert system in determining hadith validity," in *Advances in Computer and Information Sciences and Engineering*, 2008, pp. 354–359.

[64] H. M. Alrazou, "Data mining application on the resources of Islamic knowledge (in Arabic)," *Alukah*, 2008.

[65] Z. A. Aldhaln KA, Zeki AM, "Datamining and Islamic knowledge extraction: al-hadith as a knowledge resource," *IEEE Int. Conf. Inf. Commun. Technol. Muslim World (ICT4M '10)*, pp. 11–21, 2010.

[66] N. K. Ibrahim, M. F. Noordin, S. Samsuri, M. S. A. Seman, and A. E. M. B. Ali, "Isnad Al-hadith computational authentication: An analysis hierarchically," *Proc. - 6th Int. Conf. Inf. Commun. Technol. Muslim World, ICT4M 2016*, pp. 344–349, 2017.

[67] A. H. Aldhaln K, Zeki A, Zeki A, "Improving knowledge extraction of hadith classifier using decision tree algorithm," *Int. Conf. Inf. Retr. Knowl. Manag. (CAMP '12)*, pp. 148–152, 2012.

[68] K. Aldhaln, A. Zeki, and A. Zeki, "Knowledge Extraction In Hadith Using Data Mining Technique," *Int. J. Inf. Technol. Comput. Sci.*, vol. 2, pp. 13–21, 2012.

[69] A. H. Aldhlan KA, Zeki AM, Zeki AM, "Novel mechanism to improve Hadith classifier performance," in *International Conference on Advanced Computer Science Applications and Technologies (ACSAT '12)*, 2012, pp. 512–517.

[70] M. M. Najeeb, "Towards Innovative System for Hadith Isnad Processing," *Int. J. Comput. Trends Technol.*, vol. 18, no. 6, pp. 257–259, 2014.

[71] M. Ghanem, A. Mouloudi, and M. Mourchid, "Classification of Hadiths using LVQ based on VSM Considering Words Order," *Int. J. Comput. Appl.*, vol. 148, no. 4, pp. 25–28, 2016.

[72] M. M. Ahmad and Najeeb, "A Novel Hadith Processing Approach Based on Genetic Algorithms," *IEEE Access*, vol. 8, pp. 20233–20244, 2020.

[73] F. Haque, A. H. Orthy, and S. Siddique, "Hadith Authenticity Prediction using Sentiment Analysis and Machine Learning," no. March 2021, pp. 1–6, 2021.

[74] M. M. A. Najeeb, "Towards a Deep Leaning-based Approach for Hadith Classification," *Eur. J. Eng. Technol. Res.*, vol. 6, no. 3, pp. 9–15, Mar. 2021.

[75] S. Mahmoud, O. Saif, E. Nabil, M. Abdeen, M. Elnainay, and M. Torki, "AR-Sanad 280K: A Novel 280K Artificial Sanads Dataset for Hadith Narrator Disambiguation," *Inf. 2022, Vol. 13, Page 55*, vol. 13, no. 2, p. 55, Jan. 2022.

[76] Y. M. Dalloul, "An Ontology-Based Approach to Support the Process of Judging Hadith Isnad," *2012 Int. Conf. Adv. Comput. Sci. Appl. Technol.*, no. March, pp. 1–108, 2013.

[77] Rebhi S. Baraka; Yehya M. Dalloul, "Building Hadith Ontology to Support the Authenticity of Isnad," *Int. J. Islam. Appl. Comput. Sci. Technol.*, vol. 2, no. 1, pp. 25–39, 2014.

[78] N. Abd Rahman, N. Alias, N. K. Ismail, Z. Bin Mohamed Nor, and M. N. B. Alias, "An identification of authentic narrator's name features in Malay hadith texts," in *ICOS 2015 - 2015 IEEE Conference on Open Systems*, Jan. 2016, pp. 79–84.

[79] S. S. Balgasem and L. Q. Zakaria, "A hybrid method of rule-based approach and statistical measures for recognizing narrators name in hadith," *Proc. 2017 6th Int. Conf. Electr. Eng. Informatics Sustain. Soc. Through Digit. Innov. ICEEI 2017*, vol. 2017-Novem, pp. 1–5, 2018.

[80] F. Zaraket and J. Makhlouta, "Arabic cross-document NLP for the hadith and biography literature," 2012, Accessed: Apr. 22, 2021.

[81] W. P. Sari, M. A. Bijaksana, and A. F. Huda, "Indexing name in hadith translation using hidden markov model (HMM)," *2019 7th Int. Conf. Inf. Commun. Technol. ICoICT 2019*, pp. 1–5, Jul. 2019.

[82] A. Mahmood, H. U. Khan, Zahoor-Ur-Rehman, and W. Khan, "Query based information retrieval and knowledge extraction using Hadith datasets," *Proc. - 2017 13th Int. Conf. Emerg. Technol. ICET2017*, vol. 2018-Janua, no. December, pp. 1–6, 2018.

[83] I. Bounhas, B. Elayeb, F. Evrard, and Y. Slimani, "Toward a computer study of the reliability of arabic stories," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 8, pp. 1686–1705, Aug. 2010.

[84] F. Harrag, E. El-Qawasmeh, and A. M. Salman Al-Salman, "Extracting named entities from prophetic narration texts (Hadith)," *Commun. Comput. Inf. Sci.*, vol. 180 CCIS, no. PART 2, pp. 289–297, 2011.

[85] M. Alhawarat, "A domain-based approach to extract Arabic person names using n-grams and simple rules," *Asian J. Inf. Technol.*, vol. 14, no. 8, pp. 287–293, 2015.

[86] M. Bidhendi, "Extracting person names from ancient Islamic Arabic texts," *... Lang. Resour. ...*, pp. 1–6, 2012.

[87] E. T. Luthfi, Z. Izzah, M. Yusoh, and B. M. Aboobaider, "BERT based Named Entity Recognition for Automated Hadith Narrator Identification," *IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 1, p. 2022, Accessed: Feb. 22, 2023.

[88] U. Relational and S. I. Hyder, "Towards a Database Oriented Hadith Research Using Relational, Algorithmic and Data-Warehousing Techniques," *Islam. Cult. Q. J. Shaikh Zayed Islam. Cent. Islam. Arab. Stud.*, vol. 19, no. March, p. 14, 2008.

[89] Y. Yusoff, R. Ismail, and Z. Hassan, "Adopting hadith verification techniques in to digital evidence authentication," *J. Comput. Sci.*, vol. 6, no. 6, pp. 613–618, 2010.

[90] Z. Shukur, N. Fabil, J. Salim, and S. A. Noah, "Visualization of the hadith chain of narrators," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7067 LNCS, no. PART 2, pp. 340–347, 2011.

[91] M. Najeeb, A. Abdelkader, M. Al-Zghoul, and A. Osman, "A Lexicon for Hadith Science Based on a Corpus," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 1336–1340, 2015.

[92] M. M. A. Najeeb, "XML database for Hadith and narrators," *Am. J. Appl. Sci.*, vol. 13, no. 1, pp. 55–63, 2016.

[93] M. M. Najeeb, "Multi-agent system for hadith processing," *Int. J. Softw. Eng. its Appl.*, vol. 9, no. 9, pp. 153–166, 2015.

[94] Najeeb MMA, "Processing of 'Hadith Isnad' based on hidden Markov model," *Int. J. Eng. Technol.*, vol. 6, no. 2, pp. 50–55, 2016.

[95] S. R. Mohammad Najib, N. Abd Rahman, N. Kamal Ismail, N. Alias, Z. Mohamed Nor, and M. N. Alias, "Comparative Study of Machine Learning Approach on Malay Translated Hadith Text Classification based on Sanad," *MATEC Web Conf.*, vol. 135, pp. 1–9, 2017.

[96] S. B. Bin Rodzman *et al.*, "Experiment with text summarization as a positive hierarchical fuzzy logic ranking indicator for domain specific retrieval of Malay translated hadith," in *ISCAIE 2019 - 2019 IEEE Symposium on Computer Applications and Industrial Electronics*, 2019, pp. 299–304.

[97] A. Mahmood, H. U. Khan, M. Ramzan, H. U. Khan, F. K. Alarfaj, and M. Ilyas, "A Multilingual Datasets Repository of the Hadith Content," *Artic. Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 2, 2018.

[98] S. Altammami, E. Atwell, and A. Alsalka, "Text segmentation using n-grams to annotate Hadith corpus," in *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, 2019, no. July, pp. 31–39.

[99] A. M. Abdelghany, H. M. Abdelaal, A. M. Kamr, and P. M. Elkafrawy, "Doc2Vec□: An approach to identify Hadith Similarities Doc2Vec□: An approach to identify Hadith Similarities," *Aust. J. Basic Appl. Sci.*, vol. 14, no. 12, pp. 46–53, 2021.

[100] H. M. Abdelaal, A. M. Ahmed, W. Ghribi, and H. A. Youness Alansary, "Knowledge Discovery in the Hadith According to the Reliability and Memory of the Reporters Using Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 157741–157755, 2019.

[101] N. Alias, N. Abdul Rahman, N. K. Ismail, Z. Mohamed Nor, M. N. Alias, and M. S. Kamis, "Hadith Text Classification on Sanad Part Using Edge List," *Fundam. Appl. Sci. Asia*, pp. 145–156, 2022.

[102] T. Tarmom, E. Atwell, and M. Alsalka, "Deep Learning vs Compression-Based vs Traditional Machine Learning Classifiers to Detect Hadith Authenticity," *Commun. Comput. Inf. Sci.*, vol. 1577 CCIS, pp. 206–222, 2022.