



کاربردهای داده کاوی در علوم اسلامی



حسین عابدینی

کارشناس گروه پژوهشی متن کاوی نور مرکز تحقیقات کامپیوتری علوم اسلامی
habedini@noornet.net

دکتر بهروز مینایی*

عضو هیأت علمی دانشگاه علم و صنعت ایران
دبیر هیأت علمی مرکز تحقیقات کامپیوتری علوم اسلامی
b_menaei@iust.ac.ir

است که در دانش آمار غالباً فرضیه‌ای طرح می‌شود و با استفاده از تحلیل‌های آماری به اثبات یا رد آن فرضیه می‌پردازند؛ اما داده کاوی نیازمند فرضیه نیست. در تعبیر این موضوع، گاهی چنین گفته می‌شود که در داده کاوی خود داده کاو هم ممکن است نداند که به دنبال چه می‌گردد یا چه نتیجه و کاربردهایی برایش حاصل خواهد شد. به این ترتیب، با کاوش میان داده‌ها، ارتباطات مخفی، گوناگون و معناداری میان آنها کشف خواهد شد. گذشته از اینکه دستمایه دانش آمار، بیشتر داده‌های عددی است؛ اما دستمایه داده کاوی، منحصر به داده‌های عددی نیست. مهم‌ترین مراحل داده کاوی یا همان کشف دانش از میان داده‌ها، اینها است:

۱. استخراج و گردآوری داده‌ها از چندین منبع؛
۲. یکپارچه‌سازی و حذف داده‌های زاید (تصفیه)؛
۳. تشکیل انبار داده بر پایه داده‌های تصفیه شده؛
۴. اجرای عملیات داده کاوی؛
۵. تفسیر و نمایش نتایج.

داده کاوی چیست؟

«داده کاوی» یا فرایند کشف دانش در پایگاه داده، زمینه‌ای نسبتاً نوظهور است. داده کاوی، پل ارتباطی میان علوم آمار، مهندسی کامپیوتر، هوش مصنوعی، شناسایی آماری الگو، یادگیری ماشین و محاسبات رایانشی نرم است. داده کاوی، فرایند یا پروژه‌ای نسبتاً پیچیده برای شناسایی الگوها و مدل‌های صحیح، قابل استناد و مفید در حجم وسیعی از داده است؛ به گونه‌ای که این الگوها و مدل‌ها برای انسان‌ها قابل درک باشند. داده‌های در اختیار ما اغلب حجیم هستند و به‌تنهایی قابل استفاده نیستند؛ بلکه دانش نهفته در داده‌ها است که قابل استفاده است. در اینجا است که فرایند داده کاوی به ما امکان شناسایی الگوها، مدل‌ها و ارتباط میان عناصر مختلف در پایگاه داده را می‌دهد تا دانش نهفته در داده‌ها را کشف و نهایتاً این دانش را به اطلاعات تبدیل کنیم.

گاهی داده کاوی را با دانش آمار مقایسه می‌کنند. ابتدایی‌ترین تفاوت‌ها و در حقیقت، مزایای داده کاوی نسبت به دانش آمار این

* دکترای هوش مصنوعی از دانشگاه ایالتی میشیگان آمریکا.

مهم‌ترین کارکردهای متن کاوی در علوم اسلامی

۱. تشخیص ماشینی نوع کلمه و تجزیه آن (۱) (برچسب‌گذاری (۲) صرفی کلمات متن)

با استفاده از متون برچسب‌گذاری شده و قواعد استخراج شده از متن، می‌توان سیستم خبره‌ای طراحی کرد که به صورت هوشمند به تشخیص لایه‌های صرفی کلمات عربی بپردازد. این سیستم در بردارنده متون برچسب‌گذاری شده، بانک کلمات جامد غیرمصدری، کلمات جمع مکسر، صفت‌های مشبیه، ریشه‌ها و قواعد تشخیص اعلال، ابدال و ادغام خواهد بود.

برنامه‌هایی از این دست، با ترکیب دو روش «آماری (۳)» و «قانون محوری (۴)» می‌توانند نتایج مناسبی را در اختیار کاربران قرار دهند.

شایان ذکر است که اگرچه در دنیا در مورد برچسب‌گذاری ماشینی متون عربی کارهایی انجام شده است، ولی با توجه به ضعف‌های موجود در موتورهای برچسب‌زن فعلی، ضرورت ساخت موتور جدیدی به عنوان پایه کار و پردازش بر روی متون اسلامی، احساس می‌شود. امتیاز این موتور بر دیگر موتورها را در دو مورد اصلی می‌توان بیان کرد:

الف) غالب موتورهای موجود، فقط قدرت پردازش بر متون با حرکت و اعراب را دارد؛ در حالی که بیش‌تر متون اسلامی، هیچ حرکت و اعرابی ندارند. موتور ساخته شده توسط گروه پژوهشی متن کاوی نور، قابلیت پردازش متون بی‌حرکت را دارا است.

ب) غالب موتورهای موجود، بر اساس سیستم‌هایی طراحی شده‌اند که بیشتر بر روی متون محاوره‌ای جدید عربی مانند روزنامه‌ها و... کار می‌کنند. بسیاری از پژوهشگران معاصر در حوزه علوم اسلامی، تنها قرآن کریم را محور کار خود قرار داده‌اند؛ در حالی که پردازش و برچسب‌گذاری متون فقهی و اصولی و روایی، نیاز به برنامه‌ای جامع‌تر با دایره لغات و قواعد بیشتری دارد. موتور برچسب‌گذار مورد استفاده در این تحقیق، با محور قرار دادن متون اسلامی، قابلیت تجزیه کلمات متون اسلامی را با درصد صحت بالایی خواهد داشت.

۲. یافتن مسند و مسندالیه (پیش‌نیاز تعیین نقش کلمات)

با توجه به اینکه شناخت روابط میان کلمات، مهم‌ترین مباحث در درک ماشینی معنا است، شناخت مسند و مسندالیه که از ارکان عبارات عربی است، برای تحقق این مرحله اهمیت بسیاری دارد. از این رو، با استخراج قواعد تشخیص مبتدا، خبر، فاعل و... علاوه بر استفاده از متون برچسب‌گذاری شده و اجرای سیستم‌های آماری بر روی متن، می‌توان به سیستمی برای تشخیص ماشینی مسند و مسندالیه دست یافت.



داده کاوی، فرآیند یا پروژه‌ای نسبتاً پیچیده برای شناسایی الگوها و مدل‌های صحیح، قابل استناد و مفید در حجم وسیعی از داده است؛ به گونه‌ای که این الگوها و مدل‌ها برای انسان‌ها قابل درک باشند



استخراج اسنادهای موجود در یک روایت و یا یک متن فقهی، به شناخت بسیاری از ابعاد مبهم آن روایت و یا متن کمک می‌کند و کاربر را برای رسیدن به نتایج صحیح جستجو راهنمایی می‌کند.

تشخیص مسند و مسندالیه نیز مانند برچسب‌گذاری، از کاربردهای پایه‌ای متن کاوی است که هم به عنوان کاربرد مستقل در پژوهش‌های علوم اسلامی و هم به عنوان ابزاری برای تحقق دستاوردهای دیگر داده کاوی مورد استفاده قرار می‌گیرد.

۳. تشخیص ماشینی نقش کلمات (برچسب‌گذاری نحوی متون)

پس از مرحله برچسب‌گذاری صرفی کلمات یک متن، نیازمند تعیین و تشخیص نقش کلمات یک متن برای شناخت و استخراج روابط لفظی و معنوی میان کلمات هستیم. به این ترتیب، برای انجام این کار از متون برچسب‌گذاری شده نحوی و نیز قواعد استخراجی از میان کتب نحوی و رویه‌های استعمال کلمات نزد عرب زبان‌ها استفاده می‌شود.

همان طور که در مورد برچسب‌گذاری صرفی بیان شد، موتورهای برچسب‌گذار موجود در زبان عربی (صرفی و نحوی) دارای نواقصی هستند که قادر به پاسخگویی در متون اسلامی نمی‌باشند. از طرفی، می‌دانیم جهت هر گونه پردازش بر روی متن، اعم از نمایه‌زنی و یا درختواره‌سازی و...، از اساسی‌ترین پیش‌نیازها، تشخیص نقش‌های کلمات آن متن می‌باشد.

در این راستا، گروه متن کاوی نور، اولویت کاری خود را برای تهیه پیکره‌های متنی برچسب‌گذاری شده و استخراج قواعد صرفی و نحوی قرار داده است.

۴. ریشه‌یابی (۵)

استخراج ریشه و بن کلمات، در شناسایی حالات و ویژگی‌های هر کلمه اهمیت فراوانی دارد. امکان جستجو و پژوهش از طریق اشتراک ریشه نیز روش بسیار سودمند و پربازدهی است. شناخت ریشه در زبان عربی از طریق بانک ادات حرفی که شامل پیشوندها و پسوندها است و همین‌طور استخراج قواعد



اکنون در مراکز تحقیقاتی گروه‌هایی فعالیت می‌کنند که وظیفه اصلی آنها حرکت‌گذاری دستی متون اسلامی است. با توجه به تنوع حالات کلمات در عربی از حیث حرکت آخر و حرکت غیر آخر، منجر به رجوع زیاد به فرهنگ‌های لغت و منابع مرجع می‌باشد. حرکت‌گذار ماشینی به راحتی می‌تواند با استفاده از سیستم‌های آماری، به حرکت‌گذاری متون اسلامی بپردازد



مربوط به چگونگی اتصال هر پیشوند و پسوند و حالات مختلف آن، امکان‌پذیر است. این امر هم با یک فعالیت دادگانی قابل دستیابی است.

شناسایی ریشه‌های کلمات متون اسلامی، در جستجوهای پژوهشی بسیار کاربرد دارد. تشخیص کلماتی که اشتراک ریشه‌ای دارند و رده‌بندی و خوشه‌بندی آنها در راستای درختواره‌سازی و یافتن متون مشابه و...، بسیار اهمیت دارد.

۵. تشخیص اشتقاق کلمه

در زبان عربی از هر ریشه‌ای، تعدادی از مشتقات ساخته می‌شود و ریشه‌هایی که به تمام ابواب ثلاثی مجرد و مزید رفته باشد، بسیار کم است. بر اساس پایگاه داده‌ای جامعی از کتب لغت و استعمالات عرب می‌توان الگوریتمی داشت که هر کلمه را استعمال‌یابی کند و اعلام نظر کند که آیا کلمه مورد نظر در زبان عربی استعمال می‌شود یا غیر مستعمل است. همچنین این برنامه می‌تواند در ابهام‌زدایی بین مشترکات هم استفاده شود.

بیشتر محققان معجم‌های موضوعی، نیازمند ساخت واژه و ایجاد چکیده‌های مفهومی متن هستند. برای ساخت واژه از کلمات یک متن، نیاز به دانستن استعمالات ریشه آن کلمه می‌باشد؛ به عنوان مثال، اگر از کلمه‌ای قصد ساخت واژه‌ای داریم که از باب استفعال باشد، باید استعمالات ریشه آن کلمه را بررسی کنیم که آیا در زبان عرب چنین ریشه‌ای در باب استفعال استعمال شده است یا نه؟ در صورت عدم استعمال باید گزینه دیگری به عنوان واژه برای آن کلمه پیشنهاد شود. این الگوریتم می‌تواند جهت تصحیح و اشکال‌زدایی واژه‌های دستی ساخته شده توسط محققان نیز مورد استفاده قرار گیرد.

شایان ذکر است که با توجه به پویا بودن زبان، باید بانک کلمات استعمال شده در زبان، همواره به روز نگه داشته شود؛ زیرا چه بسا واژه‌ای در بانک مستعمالات استخراج شده از متون اسلامی، غیر مستعمل تشخیص داده شود، اما در زبان جاری روز، کاربرد داشته باشد. نکته دیگر اینکه برای پیشنهاد واژه صحیح، علاوه بر در اختیار داشتن بانک مستعمالات از یک ریشه، می‌توان به بانک مترادفات نیز مراجعه کرد؛ زیرا شاید نتوان آن معنای

درخواستی توسط پژوهشگر را در هیچ کدام از مشتقات ریشه مورد نظر یافت، اما همان معنا از مشتقی با ریشه مترادف دیگری مورد استعمال باشد.

۶. یافتن کلمات کلیدی (۶)

به دنبال تشخیص ماشینی نقش کلمات، می‌توان با سهولت بیشتری کلمات اصلی و ارکان جملات را شناسایی کرد. یافتن کلمات کلیدی، کاربردهای زیادی در پژوهش متون مختلف دارد که در موارد بعدی به آنها اشاره خواهیم کرد.

از موارد کاربرد کلمات کلیدی در علوم اسلامی می‌توان به: شناسایی متون مشابه، طبقه‌بندی متون (۷)، پیراسته‌سازی (۸) و... اشاره کرد. یافتن کلمات کلیدی یک متن می‌تواند در جستجوی سریع و ایجاد ارتباط معنوی بین متون علوم اسلامی، به‌ویژه در روایات، محقق را یاری دهد.

۷. نمایه‌زنی ماشینی (۹)

یکی از روش‌های پژوهشی رایج در علوم اسلامی، نمایه‌زنی دستی بر کتاب‌های تخصصی است. نمایه‌زنی بر متن را می‌توان به دو قسم نمایه‌های «پیش‌هم‌آرا» و نمایه‌های «پس‌هم‌آرا» تقسیم کرد. این روش بسیار زمان‌بر و هزینه‌بر است. و با روش‌های هوشمند ماشینی و با تشخیص روابط میان کلمات متن از طریق شناسایی نقش‌ها و عبارات متنی و اعمال قواعد استخراجی و نیز پیوسته‌سازی متون، می‌توان نمایه‌زنی را با روشی نوین انجام داد و نمایه‌های احتمالی به کاربر پیشنهاد شوند.

برای بهینه‌سازی این نمایه‌زنی، باید بانکی غنی و غیر تکراری از نمایه‌ها داشته باشیم و پژوهشگر موظف باشد نمایه خود را از میان موارد موجود انتخاب کرده و تنها بعد از طی مراحل خاصی، امکان افزودن نمایه‌ای جدید به بانک نمایه‌ها وجود داشته باشد. نکته قابل توجه این است که استخراج مستقیم نمایه از متن که به عنوان نمایه‌های «پیش‌هم‌آرا (۱۰)» شناخته می‌شود، یکی از روش‌های نمایه‌زنی است. ولی روش ماشینی می‌تواند علاوه بر استخراج نمایه‌های «پیش‌هم‌آرا»، با استفاده از برچسب‌گذاری‌های موجود و با ترکیب کلمات کلیدی و پیراسته‌های استخراجی، نمایه‌های «پس‌هم‌آرا (۱۱)» را نیز به

محقق پیشنهاد دهد که این مسأله در بازیابی مفاهیم موجود در متن و توسعه دامنه پاسخ‌های مناسب، بسیار مؤثر است.

۸. حرکت‌گذاری ماشینی

حرکت‌گذاری متون عربی، کاری بسیار وقت‌گیر و هزینه‌بر است. با وجود داده‌های حرکت‌گذاری شده بسیاری که در دسترس است و با بهره‌گیری از سیستم‌های خبره و آماری می‌توان به راحتی و با درصد بالایی از صحت، متون بدون اعراب و حرکت را حرکت‌گذاری و اعراب‌گذاری نمود. ارتباط حرکت‌گذاری و اعراب‌گذاری، ارتباطی از نوع عام و خاص است. به حرکت‌گذاری آخر کلمه، اعراب‌گذاری می‌گویند. دقت ماشین هم‌اکنون در حرکت‌گذاری غیر آخر، بسیار بالاتر از حرکت آخر که اصطلاحاً اعراب نام دارد، می‌باشد.

اکنون در مراکز تحقیقاتی گروه‌هایی فعالیت می‌کنند که وظیفه اصلی آنها حرکت‌گذاری دستی متون اسلامی است. با توجه به تنوع حالات کلمات در عربی از حیث حرکت آخر و حرکت غیر آخر، منجر به رجوع زیاد به فرهنگ‌های لغت و منابع مرجع می‌باشد. حرکت‌گذار ماشینی به راحتی می‌تواند با استفاده از سیستم‌های آماری، به حرکت‌گذاری متون اسلامی بپردازد.

۹. پیراسته‌سازی متون

پیراسته‌سازی متون، یکی از روش‌های نوین پژوهشی و در عین حال، بسیار کاربردی است. منظور از پیراسته‌سازی، حذف پسوندها و پیشوندهای کلمات، تبدیل افعال و مشتقات به مصادر اصلی خود همراه با قرار دادن (ال) در ابتدای آن مصدر، حذف حروف بی‌بار از متن، تبدیل جمع‌ها به مفرد و مؤنث‌ها به مذکر و... است که تحوّل چشمگیری در سیستم جستجوی نرم‌افزارهای علوم اسلامی ایجاد خواهد کرد و پس از آن، معجمی از واژه‌ها جایگزین معجم الفاظ خواهد شد.

اکنون در نرم‌افزارهای رایج جهت جستجو از معجم الفاظ استفاده می‌شود. از معایب این روش می‌توان به پاسخ‌های زاید در دامنه جستجو اشاره کرد؛ به عنوان مثال، اگر کاربری بخواهد در مورد «استغفار» تحقیق کند، باید تمامی حالات ممکن (استغفره، یستغفر، یستغفرون، لیستغفر، لاستغفار و...) را جستجو کند تا به نتیجه مورد نظر برسد و یا از طریق ریشه «غفر» جستجو کند که نتیجه آن، دامنه پاسخ‌های زیاد و زاید (غفران، غفار، مغفرة، مغفور و...) می‌باشد؛ در حالی که با روش پیراسته‌سازی ماشینی و جایگزینی معجم واژه‌ها به جای معجم الفاظ، کاربر با جستجوی کلمه «لاستغفار» و در کمترین زمان، به پاسخ‌های مورد نظر خود دست می‌یابد.

این امر صرفاً بخشی از کاربردهای پیراسته‌سازی در پژوهش‌های اسلامی است.

۱۰. یافتن متون مشابه (۱۲)

با طراحی یک سیستم خبره و در نظر گرفتن قوانین با هم‌آیی کلمات می‌توان متون مشابه را به صورت ماشینی شناسایی کرد. در صورتی که بخش قواعد معنوی و بانک‌های مترادفات و مشترکات نیز ضمیمه چنین برنامه‌ای شود، بر صحت پاسخ‌های برنامه افزوده خواهد شد.

این برنامه می‌تواند دامنه پاسخ‌های مناسب پژوهش‌های کاربر را وسعت دهد؛ به عنوان مثال، در متون روایی اگر کاربری با روایتی مواجه شود و بخواهد تمام روایاتی که نوعی تشابه از نظر لفظی و یا معنوی با این روایت را دارند، مشاهده کند، با استفاده از این برنامه به راحتی می‌تواند با تعیین دامنه جستجوی خود در کتب روایی، تمام احادیث مشابه روایت مورد نظر را بیابد و مورد بررسی قرار دهد.

۱۱. تدوین درختواره (هستان‌شناسی) (۱۳) برای علوم اسلامی

یکی از راه‌های رایج برای انتقال مفاهیم متنی و پژوهش در متون، نمایش سلسله مراتبی مفاهیم موجود در متن، یا روش نمایش درختی مفاهیم و خوشه‌بندی (۱۴) آنها است. در حال حاضر، این کار به صورت دستی و با بهره‌گیری از محققان انجام می‌شود. اما ماشین می‌تواند با به کارگیری سیستم‌های آماری و پردازش داده‌ها با پیوست به پایگاه‌های غنی شده مترادفات، مشترکات و مرتبطات، حجم بالایی از این عملیات را به صورت هوشمند پوشش دهد.

نکته قابل ذکر این است که روش نمایش درختی (هستان‌شناسی)، یکی از بهترین و جامع‌ترین روش‌های عرضه مفاهیم یک متن می‌باشد که این روش اکنون در مورد برخی متون علوم اسلامی به صورت دستی در حال انجام است. اما نکته مهم در درختواره کردن متون، اتصال این درخت‌ها به یکدیگر و استخراج روابط معنوی موجود در بین آنها است که نیاز به تتبع و احاطه کامل به تمامی موارد استخراجی دارد و این کار با روش ماشینی و پردازش داده‌ها به راحتی و با درصد صحت بالایی میسر می‌شود.

۱۲. خطایابی (۱۵) متن و پیشنهاد کلمات صحیح

با توجه به بانک‌های غنی اطلاعاتی و همچنین پردازش‌گر هوشمند ماشینی می‌توان به خطایابی مقالات و یا متون اسلامی پرداخت. با ورود متن به سیستم، برنامه با نگاه دقیق به کلمات و مشتقات و اسنادهای استفاده شده در متن، اشکالات احتمالی موجود در متن را شناسایی کرده و به جای آن، کلمه صحیح را در لایه صرف و یا عبارت صحیح را در لایه نحو به کاربر پیشنهاد می‌دهد.

با این شیوه می‌توان ابتدا متون اسلامی را از جهت لغت و دستور زبان، تصحیح، و سپس در برنامه مورد پردازش قرار داد تا در ضمن اجرای برنامه، مشکلی در جهت پیاده‌سازی قواعد و تشخیص نتایج حاصل نشود.

۱۳. چکیده‌گیری (۱۶)

همواره یکی از خواسته‌های علمی و تحقیقی محققان در بیان مقاصد و اهداف خود به دیگران، بیان مطالب در کوتاه‌ترین و مختصرترین قالب بوده است. چکیده‌گیری از متون، در حال حاضر به شیوه‌ای دستی انجام می‌شود که با دستیابی به اهداف گفته شده‌ای مانند: تجزیه و ترکیب ماشینی، یافتن کلمات کلیدی، تشخیص مسند و مسندالیه و... می‌توان این کار را توسط ماشین انجام داد. چکیده‌گیری ماشینی متون اسلامی، مقدمه‌ای است برای موضوع‌برداری و سپس نمایه‌زنی، که اکنون در بسیاری از مؤسسات تحقیقاتی این طرح توسط محققان و به شیوه دستی انجام می‌شود.

۱۴. تشخیص خودکار محدوده جمله (۱۷)

تشخیص محدوده جملات و عبارات در یک متن، به درک معنا و مفهوم آن کمک زیادی می‌کند. تشخیص حد جمله، یکی از اهدافی است که می‌توان آن را مقدمه بسیاری از پروژه‌های داده‌کاوی دانست. این کار، یکی از فرایندهایی است که به دلیل بار معنایی جملات و قواعد پیچیده و حالات و استثنائات فراوان، اندکی مشکل به نظر می‌رسد؛ اما با استخراج قواعد فراوان از استعمال زبانی روزمره اعراب و تحلیل‌های نحوی و سیستم‌های آماری می‌توان تا حد فراوانی به ماشینی شدن آن امیدوار بود. تشخیص محدوده جملات متن، کمک بسیاری به نمایه‌زنی و موضوع‌برداری می‌کند. روند موضوع‌برداری کنونی مرکز تحقیقات کامپیوتری علوم اسلامی به این شیوه می‌باشد که ابتدا محقق قسمتی از متن را انتخاب نموده، سپس آن را چکیده کرده و موضوعات مربوطه را استخراج می‌کند. همچنین یکی از کاربردهای مهم تشخیص محدوده جملات را می‌توان در درک مفهوم صحیح از روایات و تعیین دامنه قول معصوم دانست.

۱۵. ترجمه ماشینی (۱۸)

تبدیل متن از زبانی به زبان دیگر که در اصطلاح به آن ترجمه گفته می‌شود، نیازمند شناخت کلمات و ویژگی‌های آنها اعم از اسم، فعل و حرف و همچنین تشخیص نقش و روابط میان کلمات است. اکنون برای ترجمه یک متن توسط مترجم انسانی مرحله‌ای طی می‌شود که عیناً همین مراحل را می‌توان برای سیستم تعریف کرد. با به کارگیری اولویت‌های پژوهشی قبلاً گفته شده (تجزیه و ترکیب ماشینی، یافتن کلمات کلیدی و...)

می‌توان بانک لغات جامعی در اختیار گرفت و سیستمی را طراحی کرد که با الگویی از پردازش فکری انسان در ترجمه یک متن، به ترجمه ماشینی متون بپردازد.

۱۶. تشخیص اسامی خاص (۱۹)

یکی از مهم‌ترین پیش‌نیازهای برچسب‌گذاری صرفی، جداسازی اسامی خاص یک متن است. اسامی خاص، اسامی جامد غیر مصدری هستند که همیشه برچسب مخصوصی دارند و غالباً در بیشتر برچسب‌ها با هم مشترک می‌باشند که با شناسایی آنها می‌توان لایه‌های ابهام برنامه برچسب‌گذار صرفی را کاهش دهیم. به عنوان مثال، چون مبنای کار ما بر متون بی‌حرکت است، وقتی ماشین برچسب‌گذار به کلمه «علی» برخورد کند، دو برچسب با مشخصات حرف جر و اسم جامد بیان می‌کند. حال اگر ما قبل از شروع کار برچسب‌گذاری، متن مورد نظر را از جهت اسامی خاص پاک‌سازی کنیم و با استفاده از بانک‌های غنی موجود و قواعد استقرایی استخراجی، موارد اسم خاص را کاملاً شناسایی کنیم، در این صورت لایه‌های ابهامی بسیار کمتر خواهد شد.

مثلاً در کلمه «علی» قواعد ما می‌گوید کلمه بعد و قبل را نگاه کن؛ اگر «بن» بود و یا کلمه قبل از آن فعل «قال» بود، احتمال اسم خاص بودن این کلمه زیاد خواهد بود.

اسامی خاص موجود در یک متن، شامل: نام شخص، مکان، زمان، کتاب و اسم جنس است که موارد تک‌کلمه‌ای را بیشتر از طریق بانک‌ها می‌توان شناسایی کرد؛ ولی موارد ترکیبی را با استفاده از قواعدی که به وسیله استقرا استخراج کرده‌ایم، می‌توانیم شناسایی کنیم.

۱۷. رده‌بندی (طبقه‌بندی) متون (۲۰)

دسته‌بندی متون، از لوازم اولیه پژوهش‌های علوم اسلامی است. زمانی که محقق می‌خواهد در مورد موضوعی تحقیق کند، جمع‌آوری و دسته‌بندی متونی که جهت پژوهش خود به آن نیاز دارد، برای او ضروری است. در سایت‌های تخصصی علوم اسلامی که مقالات رده‌بندی از پیش تعیین شده‌ای بر آن حاکم است، برای به روزرسانی ماشینی و جای دادن خودکار متون جدید، نیازمند این دسته از الگوریتم‌ها می‌باشیم.

طبقه‌بندی ماشینی متون که یکی از دستاوردهای مهم متن‌کاوی است، بعد از استخراج کلمات کلیدی متن و با کمک پیراسته‌سازی می‌توانیم به آن دست یابیم.

گفتنی است که دسته‌بندی متون الزاماً موضوعی نیست؛ بلکه از جهات مختلف مانند: تاریخی، ادبی و... نیز قابل طبقه‌بندی است.

۱۸. ابهام‌زدایی (۲۱) و تمیز مشترکات

از آنجایی که محور کار ما بر متون بی‌حرکت است، به‌ناچار

پاسخ‌های موتور برچسب‌گذار دارای ابهام‌هایی خواهد بود که جهت رسیدن به جواب مطلوب باید موارد مبهم، ابهام‌زدایی شود. شیوه ارائه بانک‌های اولیه مانند: بانک اسامی خاص، حروف، اسامی جامد و صفات مشبیه، در کم کردن لایه‌های ابهام بسیار مؤثر خواهد بود؛ به عنوان مثال، گاهی دیده شده، اگر ابتدا بانک اسامی خاص عرضه شود و بعد بانک صفات مشبیه، مواردی به اشتباه شناسایی شده و مواردی هم دارای ابهام خواهند شد. در نتیجه، مدیریت ارائه بانک‌ها به متن مورد نظر بسیار مهم خواهد بود. ابهام‌زدایی ماشینی با کمک متن‌های برچسب‌گذاری شده دستی که توسط محققان تهیه شده و با استفاده از یادگیری ماشین امکان‌پذیر خواهد بود.

نتیجه‌گیری و کارهای آینده

با توجه به مطالب بیان‌شده روشن است که استفاده از شیوه‌های نو در علوم اسلامی بسیار ضروری است و تحول اساسی در ارائه مفاهیم و معارف اسلامی به مخاطبان را در پی دارد. به کارگیری متن‌کاوی در پردازش متون اسلامی، دریچه‌ای جدید در بازیابی معارف اسلامی بوده و امکان خلق ایده‌های بدیع و راهکارهای مفید پژوهشی را در ذهن مخاطبان ایجاد خواهد کرد. در ادامه، علاوه بر هیجده کاربرد مذکور می‌توان به موارد دوازده‌گانه ذیل نیز به عنوان نمونه‌هایی از کاربرد متن‌کاوی و هوش مصنوعی در علوم اسلامی اشاره کرد که در مقالات بعد به آنها می‌پردازیم:

۱۹. تشخیص خودکار متن فارسی از عربی؛
۲۰. تشخیص ماشینی آیات قرآن؛
۲۱. شناسایی نویسنده یک متن؛
۲۲. شناخت دست‌خط نویسندگان کتب خطی؛
۲۳. تشخیص هوشمند شعر از نثر؛
۲۴. یافتن هوشمند وزن عروضی اشعار؛
۲۵. تفکیک سند از متن روایت؛
۲۶. تفکیک راوی از تعبیر واسطه در اسناد روایات؛
۲۷. یافتن طبقه، مشایخ و شاگردان راوی؛
۲۸. یافتن مرجع ضمیر در راویان مضمّر؛
۲۹. حل مشکل تحویل و تعلیق در اسناد روایات؛
۳۰. طراحی سیستم فقیه‌یار.

الگوریتم‌های هوش ماشین، اعم از یادگیری ماشین و یافتن آماری الگوها و دیگر الگوریتم‌های متن‌کاوی، راه‌هایی برای رسیدن به این مقاصد متعدد کاربردی فوق‌الذکر عرضه کرده‌اند؛ اما مهم این است که نقشه راه بهینه‌ای برای پیمایش مسیر در

نظر گرفته شود که از جهت زمان، دقت و کارایی راه حل برتر باشد. امید است این فعالیت‌ها در راستای نشر معارف الهی مورد رضایت حضرت حق قرار گیرد. ■

پی‌نوشت‌ها:

1. Morphology.
2. Tagging.
3. statistical.
4. Rule-based.
5. stemming.
6. eyword Extraction.
7. Text Classification.
8. Lemmatization.
9. Automatic Indexing.
۱۰. نمایه پیش هم‌آرا: نوعی استخراج مفاهیم ترکیبی از متن است؛ به این صورت که از موضوعات استخراجی متن، ترکیباتی انتزاع می‌شود که بیانگر حیثیات آن متن می‌باشند. به این ترکیبات استخراجی، نمایه‌های پیش هم‌آرا می‌گویند.
- مثال: موضوع: وجوب الوضوء الصلاة الواجبة والطواف الواجب والمس المس لكتابة القرآن و استحبابه لغیرها.
۱۱. نمایه پس هم‌آرا: نوعی استخراج مفاهیم ترکیبی از متن است؛ به این صورت که ابتدا کلمات کلیدی و پیراسته‌ها از متن استخراج شده و سپس با ترکیب آن کلمات استخراجی، نمایه‌هایی تولید می‌شود که حیثیات پنهان متن را بیان می‌کند.
- مثال: کلمات کلیدی استخراجی از متن: الصلاة، الاستسقاء، الکيفية، الركعة، المعطر.
- نمایه‌های پیشنهادی: صلاة الاستسقاء، كيفية صلاة الاستسقاء، ركعة صلاة الاستسقاء و...
12. Text similarity detection.
13. Ontology.
14. Clustering.
15. Spell checking.
16. Summerization.
17. Sentence Boundry Detection.
18. Machine Translation.
19. Name Entity Recognition.
20. Text Classification.
21. Disambiguation.