ارائه روشی برای استخراج خودکار عبارات کلیدی از اخبار وب پارسی

مریم باسره[۱]، دانشجوی کارشناسی ارشد؛ ولی درهمی[۲]، دانشیار؛ سجاد ظریف‌زاده[۳]، استادیار

۱- دانشکده مهندسی کامپیوتر - پردیس فنی و مهندسی - دانشگاه یزد - یزد - ایران - basere.m92@yazd.ac.ir
۲- دانشکده مهندسی کامپیوتر - پردیس فنی و مهندسی - دانشگاه یزد - یزد - ایران - vderhami@yazd.ac.ir
۳- دانشکده مهندسی کامپیوتر - پردیس فنی و مهندسی - دانشگاه یزد - یزد - ایران - szarifzadeh@yazd.ac.ir

چکیده: دادگان متنی و از آن جمله متون خبری از حوزه‌های مهم بازیابی اطلاعات به شمار می‌رود و استخراج اطلاعات از آن‌ها ضروری است. این امر با استخراج عبارات کلیدی اسناد که دربردارنده محتوای اصلی متن است، صورت می‌گیرد. در این پژوهش، راهکاری سه مرحله‌ای جهت استخراج عبارات کلیدی از صفحات خبری وب پارسی، با ترکیب شیوه‌های زبان‌شناختی، یادگیری با ناظر، ابتکاری و تعداد نسبتاً جامعی از شیوه‌های آماری ارائه می شود. همچنین، یک مجموعه داده خبری و لیستی از عبارات توقفی خبری ایجاد می‌گردد. در پژوهش حاضر، با توجه به ویژگی‌های دادگان، از دسته‌بند جنگل تصادفی استفاده و عملکرد خوب آن به کمک نتایج آزمایش‌ها ثابت می‌شود. به علاوه، استفاده از امتیاز تعلق‌گرفته به عبارات توسط دسته‌بند، جهت ایجاد لیستی مرتب از عبارات برای دسته‌بندی، به جای استفاده ازخروجی دسته‌بند، پیشنهاد می‌شود. نتایج، نشان‌دهنده دقت قابل‌قبول سیستم ارائه‌شده است.

واژه‌های کلیدی: عبارات کلیدی، استخراج عبارات کلیدی، اسناد خبری، شیوه‌های آماری، یادگیری با ناظر، متن‌کاوی، بازیابی اطلاعات.

A Method for Automatic Key phrase Extraction from Persian Web News

M. Basereh[1], MSc Student; V. Derhami[2], Associate Professor; S. Zarifzadeh[3], Assistant Professor

1- Faculty of Electrical and Computer Engineering, University of Yazd, Yazd, Iran, Email: basere.m92@yazd.ac.ir
2- Faculty of Electrical and Computer Engineering, University of Yazd, Yazd, Iran, Email: vderhami@yazd.ac.ir
3- Faculty of Electrical and Computer Engineering, University of Yazd, Yazd, Iran, Email: szarifzadeh@yazd.ac.ir

Abstract: Text documents, especially news, are one of the important information retrieval fields which are necessary to extract information. This job, is done by extracting key phrases which include the main context of the news. In this research, a three level approach combining lingual, supervised learning, heuristic, and a relatively comprehensive number of statistical approaches, is suggested for key phrase extraction from Persian news web pages. A news dataset and a stop word list are generated. In this research, according to the data characteristics, Random Forest classifier is used; and its good performance is proved through experiments. Furthermore, using scores given by classifier to phrases, to build an ordered list of phrases, for classification, instead of using the classifier output, is suggested. Results show an acceptable f-measure.

Keywords: Keyphrase, keyphrase extraction, news texts, statistical techniques, supervised learning, text mining, information retrieval.

تاریخ ارسال مقاله: ۱۳۹۵/۰۴/۰۸
تاریخ اصلاح مقاله: ۱۳۹۵/۰۷/۱۹
تاریخ پذیرش مقاله: ۱۳۹۵/۰۸/۲۴
نام نویسنده مسئول: ولی درهمی
نشانی نویسنده مسئول: ایران - یزد - بلوار دانشجو - دانشگاه یزد - پردیس فنی و مهندسی - دانشکده مهندسی کامپیوتر.

۱- مقدمه

امروزه با تولید انفجاری دادگان متنی روبه‌رو هستیم. به سبب اهمیت بالای این نوع دادگان، استخراج اطلاعات از آن‌ها ضروری است. اطلاعات هر متن در مهم‌ترین ویژگی آن، یعنی عبارات تشکیل‌دهنده‌اش نهفته است. بدین ترتیب، در استخراج اطلاعات از دادگان متنی، در پی یافتن عباراتی هستیم که به‌خوبی بیان‌کننده محتوای آن باشد. به چنین عباراتی، عبارات کلیدی گفته می‌شود [۱].

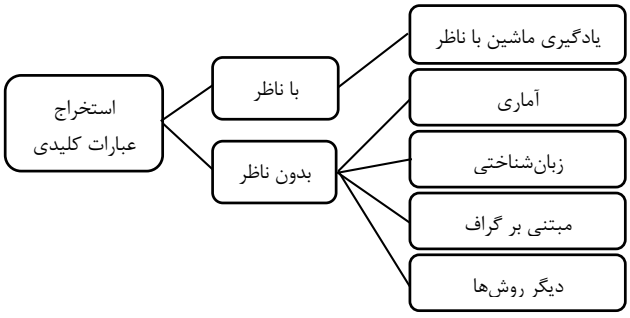
استخراج عبارات کلیدی، هسته انواع کاربردهای متن‌کاوی و بازیابی اطلاعات است [۲، ۳]. ازجمله این کاربردها می‌توان به دسته‌بندی و خوشه‌بندی متون [۴-۷]، خلاصه‌سازی متن [۵، ۷، ۸]، پیشنهاد عبارت[۱] [۹]، بسط پرس‌وجو[۲] [۱۰] و پاسخ به پرس‌وجو [۵] اشاره کرد. به‌طورکلی، استخراج عبارات کلیدی، جهت هر کاربردی که انجام گیرد، در وزن دهی و رتبه‌بندی مناسب عبارات متن، با توجه به میزان اهمیتشان، خلاصه می‌شود [۱۱].

استخراج عبارات کلیدی، به دو طریق دستی و خودکار قابل انجام است. شیوه دستی به کمک خبره، صورت می‌گیرد. کاری که در دهه ۹۰ میلادی توسط شرکت یاهو انجام می‌شد [۸]. این شیوه، صحت بالا و سرعت پایینی دارد. درنتیجه، با توجه به حجم بالای متون، امروزه به‌کارگیری آن عملاً غیرممکن است. بنابراین، بایستی به دنبال راهی برای استخراج خودکار عبارات کلیدی از دادگان متنی بود.

ا ستخراج خودکار عبارات کلیدی با چالش‌های بسیار روبه‌رو ا ست. مهم‌ترین آن‌ها پیچیدگی و ابهام زبان طبیعی اســت [۱]. وجود کلمات چندمعنایی یا دارای نقش‌های دســتوری متفاوت، نمونه‌های پیچیدگی زبان طبیعی اســت. چالش مهم دیگر، مفهوم فازی کلیدی بودن اســت. در هر جمله از متن تعدادی عبارت برجســته وجود دارد. اما، عبارات برجسته یک جمله خاص، با توجه به مفهوم و محتوای کلی متن، لزوماً برای کل آن باارزش نیست. بنابراین، نه‌تنها کل عبارات موجود در متن درجات اهمیت متفاوتی دارند، بلکه این مســئله در مورد مجموعه عبارات مهم یک متن نیز صــدق می‌کند. ازاین‌رو، ممکن اســت نظر دو خبره در تعیین عبارات کلیدی متنی یکسان متفاوت باشد [۱۱].

به‌علاوه، تعیین عبارات کلیدی به نوع کاربرد نیز وابسته ا ست. این مورد، ا ستخراج قواعد را برای تشخیص عبارات کلیدی دشوار می سازد. افزون بر این، عبارات کلیدی عموماً اسمی و وصفی بوده و دارای معنای مســتقل هســتند. لذا هرچند می‌توان الگوهایی جهت تشــخیص عبارات اســمی و وصــفی اســتخراج کرد، بااین‌حال، تعیین اســتقلال معنایی و واردکردن معنا در الگوریتم‌های ماشــینی، چالشــی بزرگ اســت. خطای ابزار پیش‌پردازش نیز از دیگر چالش‌های عمده استخراج عبارات کلیدی است. مسائل مذکور نشان‌دهنده دشواری و پیچیدگی استخراج خودکار عبارات کلیدی به کمک الگوریتم‌های فاقد هوش بشری است.

تلاش برای وزن دهی به عبارات متون، از اواخر دهه ۶۰ میلادی آغاز شد [۱۲]. در ابتدا، تنها فراوانی کلمات یک سند به‌عنوان امتیاز آن‌ها در نظر گرفته می‌شد [۱]. تا این‌که تحلیل آماریِ فراوانی کلمات یک سند با در نظر گرفتن فراوانی آن‌ها در مجموعه‌ای از اسناد (TF.IDF) در ۱۹۷۲ میلادی [۱۳] ارائه شد. از آن زمان تاکنون، شیوه‌های مختلفی برای استخراج کلمات کلیدی ابداع‌شده و به‌کاررفته است. روش‌های استخراج عبارات کلیدی را می‌توان به‌صورت شکل ۱ تقسیم‌بندی کرد.



شکل ۱: تقسیم‌بندی روش‌های استخراج عبارات کلیدی

در روش‌های یادگیری ماشین با ناظر [۱۴، ۱۵] عبارات اسناد آموزشی و آزمون به کمک ویژگی‌هایی مانند فراوانی، به بردارهای ویژگی عددی تبدیل می‌شود. سپس، الگوریتم یادگیری با ناظر، به کمک بردارهای ویژگی اسناد آموزشی، که کلیدی بودن یا نبودن (برچسب) آن‌ها از پیش تعیین‌شده، آموزش‌دیده و مدلی برای تعیین عبارات کلیدی متون دیگر می‌سازد. الگوریتم‌های یادگیری مختلفی در این زمینه بکار رفته است [۴، ۱۶، ۱۷، ۱۸، ۱۹].

وابستگی به دادگان برچسب دار (دادگانی که عبارات کلیدی آن‌ها از پیش تعیین شده باشد) و سرعت پایین، به دلیل پیچیدگی محاسباتی و زمان‌بر بودن آموزش مدل، از معایب شیوه‌های یادگیری با ناظر است. از مزایای این روش‌ها نیز، می‌توان به دقت نسبتاً بالای آن‌ها اشاره کرد.

از میان روش‌های بدون ناظر استخراج عبارات کلیدی، روش‌های آماری، مبتنی بر شمارش عبارات و تحلیل آماری آن‌هاست. از پرکاربردترین و شناخته‌شده‌ترین روش‌های این دسته، می‌توان به TF.IDF [۲، ۴، ۲۰، ۲۱]، مکان اولین رخداد عبارت در متن و رخداد در عنوان اشاره کرد. سادگی [۲]، مستقل از زبان بودن [۴]، پیچیدگی محاسباتی پایین و سرعت نسبتاً بالا، از مزیت‌های این روش‌ها است. ضعف عمده این روش‌ها در نظر نگرفتن معنا است.

روش‌های زبان‌شناختی به دو دسته تقسیم می‌شوند. دسته اول، در ترکیب با سایر روش‌ها و برای پیش‌پردازش داده‌ها به کار می‌روند. ازجمله این روش‌ها می‌توان به استفاده از انواع ریشه‌یاب‌ها و تحلیلگرها اشاره کرد [۴، ۷، ۲۰، ۲۱]. دسته دوم، جهت اعمال ساختار دستوری و معنایی روی عبارات استخراج‌شده مورد استفاده قرار می‌گیرند. مانند به‌کارگیری فرهنگ لغت و مجموعه‌ای از قواعد دستوری [۲۲]. بدیهی است که این روش‌ها به زبان دادگان موردبررسی وابسته بوده و به جهت پیچیدگی زبان طبیعی، دشوار است.

در روش‌های مبتنی بر گراف، عبارات متن همچون گره‌های یک گراف، به کمک روابط معنایی، دستوری یا هم رخدادی به هم وصل می‌شوند. مانند TextRank که در آن می‌توان هم رخدادی نشانه‌ها را بر

اساس یک پنجره لغزان به طول N سنجیده و وزن نشانه‌ها را بر اساس فاصله آن‌ها کنترل کرد [۲۱]. در این الگوریتم، نشانه‌هایی که به تعداد زیادی از نشانه‌های دیگر متصل باشد، معتبر است. نشانه‌های متصل به نشانه‌های معتبر هم با احتساب یک ضریب میرایی[۳] دارای اعتبار هستند. بدین ترتیب، امتیاز در بین نشانه‌ها پخش می‌شود.

یکی از اشکالات عمده شیوه‌های مبتنی بر گراف، مسئله "ثروتمندتر شدن ثروتمندان"[۴] است. بدین معنی که نشانه‌هایی که به تعداد بیشتری از نشانه‌های دیگر وصل هستند، اعتبار بیشتری می‌یابند. در مقابل، نشانه‌های مهم کم تکرارتر، فرصتی برای دریافت امتیاز بالا نخواهد یافت. امکان تشخیص و به‌کارگیری ارتباطات معنایی، ساختاری و دستوری بین کلمات، مهم‌ترین ویژگی مثبت این شیوه‌هاست.

از روش‌های دیگر می‌توان به روش‌های اکتشافی اشاره کرد [۲۳]. در این روش‌ها، اغلب بر اساس ویژگی‌ها، محدودیت‌ها و شرایط دادگان موردبررسی، مجموعه‌ای از قواعد یا راهکارهای میانبر استخراج می‌شود. در نظر گرفتن مواردی چون فاصله بین کلمات، روابط از پیش تعیین‌شده بین کلمات و تحلیل برچسب‌های HTML در اطراف کلمات ازجمله روش‌های اکتشافی است.

هر یک از روش‌های مذکور، دارای مزایا و معایب خاص خود بوده و اغلب برای دستیابی به نتیجه بهتر و پوشش معایب هر یک به کمک مزایای دیگری، در ترکیب با هم به کار می‌رود. در مقاله [۱۹] یک سیستم یادگیری با ناظر برای استخراج عبارات کلیدی طراحی‌شده که از ترکیبی از انواع روش‌ها اعم از آماری، مبتنی بر زبان‌های طبیعی (جهت پیش‌پردازش) و مبتنی بر یادگیری، برای وزن دهی به عبارات استفاده می‌کند.

کارهای اشاره‌شده همگی در زبان انگلیسی انجام‌شده‌اند. تاکنون استخراج عبارات کلیدی از متون پارسی چندان مورد پژوهش قرار نگرفته و این حوزه نسبتاً بکر است. درنتیجه، علاوه بر چالش‌های مطرح‌شده برای استخراج عبارات کلیدی، چالش‌های دیگری نظیر نبود پیشینه پژوهش، ابزار پیش‌پردازش دقیق و دادگان استاندارد مناسب در این حوزه برای زبان پارسی وجود دارد.

سهم علمی مقاله حاضر بدین شرح است: نظر به اهمیت بالای متون خبری، با ترکیب شیوه‌های زبان‌شناختی (نشانه‌گذاری، نرمال‌سازی، حذف عبارات توقفی، حذف افعال و ریشه‌یابی اسمی)، ۱۸ شیوه آماری، دو شیوه ابتکاری و الگوریتم یادگیری با ناظر جنگل تصادفی، راهکاری برای استخراج عبارات کلیدی از صفحات خبری وب پارسی ارائه‌شده است. به‌علاوه، یک مجموعه دادگان خبری با برچسب‌زنی دستی عبارات کلیدی، به همراه یک لیست عبارات توقفی خبری، به شکل اصولی تولیدشده است. همچنین، مسئله عدم توازن توزیع کلیدی-غیر کلیدی دادگان موردبررسی قرارگرفته و استفاده از امتیاز تعلق‌گرفته به عبارات توسط دسته‌بند و ایجاد یک لیست مرتب جهت دسته‌بندی به‌جای استفاده از خروجی دسته‌بند پیشنهادشده است. نتایج نشان‌دهنده دقت قابل‌قبول سیستم ارائه‌شده است. مهم‌ترین مزیت پژوهش حاضر این است که برای اولین بار (لااقل در میان کارهای به چاپ رسیده در حوزه زبان پارسی) بررسی جامعی جهت استخراج عبارات کلیدی از اخبار پارسی انجام گرفته و دادگان و ابزار پیش‌پردازش تولید و شخصی‌سازی‌شده همچنان قابل‌بهبود و استفاده جهت تحقیقات آتی هستند. بدین ترتیب تحقیق حاضر می‌تواند نقطه شروعی برای کار درزمینه اختصاصی استخراج عبارات کلیدی از اسناد پارسی باشد.

ساختار مقاله بدین‌صورت است که در بخش دوم، راه‌حل پیشنهادی شرح داده می‌شود. در بخش سوم، نتایج آزمایش‌ها آمده است. بخش چهارم، به ارائه بحث و تحلیل نتایج اختصاص‌یافته و درنهایت در بخش پنجم، نتیجه‌گیری ارائه خواهد شد.

۲- راه‌حل پیشنهادی

در این مقاله راهکاری سه مرحله‌ای برای استخراج عبارات کلیدی از صفحات خبری وب پارسی ارائه‌شده است. مراحل راهکار ارائه‌شده عبارت است از پیش‌پردازش، محاسبه ویژگی‌های آماری و اکتشافی برای عبارات و درنهایت دسته‌بندی. در ادامه ابتدا شرح مختصری از مفاهیم پایه‌ای به‌کاررفته در این مقاله می‌آید؛ و سپس مراحل مذکور به ترتیب شرح داده می‌شود.

۲-۱- تعریف‌ها

به دنباله‌ای از یک یا چند کلمه یک عبارت یا n-gram گفته می‌شود. عبارات توقفی به مجموعه عباراتی گفته می‌شود که به‌وفور در متون دیده‌شده و حاوی معنای خاص یا نشان‌دهنده محتوای متن نیستند. ریشه‌یابی عبارت است از حذف پسوندهای تصریفی کلمات. گاهی در متون پارسی از برخی کاراکترهای زبان عربی هم استفاده می‌شود، که علی‌رغم تشابه با کاراکترهای پارسی، یونیکدهای متفاوت دارند. درنتیجه، همه حروف متن طی عملیاتی با عنوان نرمال‌سازی کاراکتری یکپارچه می‌شود.

به تعیین برچسب‌های اسناد، برچسب‌زنی دادگان گفته می‌شود. توزیع انتشار انفجاری بدین معناست که موضوعات داغ روز، یک‌باره توسط خبرگزاری‌های زیادی مخابره می‌شود؛ و در فاصله زمانی نسبتاً کوتاهی نیز از صدر اخبار خارج می‌شود.

۲-۲- پیش‌پردازش

یکی از مراحل مهم داده‌کاوی، پیش‌پردازش دادگان است. پیش‌پردازش، داده را به قالب مناسب برای داده کاوی تبدیل کرده [۲۴] و روند محاسبات و استخراج اطلاعات را سریع و ساده می‌کند [۲۵]. در دادگان متنی به دلیل وجود انواع علائم و نشانه‌ها، عبارات اضافه و ربط و عبارات دارای نقش ها و معانی مختلف، این مرحله چالش‌برانگیز و حساس است. در این مقاله، پیش‌پردازش شامل پنج مرحله به شرح زیر، روی دادگان اعمال‌شده است.

۱- **نرمال‌سازی کاراکتری:** در این مرحله، نرمال‌سازی حرفی و عددی انجام می‌شود. همچنین، برخی کاراکترها مانند حرکات و تنوین‌ها و همزه، از متن خبر حذف می‌گردد.

۲- **نشانه‌گذاری:** در این مرحله، متن به جملات و کلماتش شکسته می‌شود، تا امکان اعمال الگوریتم‌ها و محاسبه ویژگی‌ها فراهم شود. از مهم‌ترین نشانه‌های جداکننده کلمات می‌توان به فاصله و علائم نگارشی مانند نقطه، ویرگول، نقطه‌ویرگول و دونقطه اشاره کرد.

۳- **حذف عبارات توقفی:** بالای ۳۰ تا ۵۰ درصد متون پارسی را عبارات توقفی تشکیل می‌دهند [۲۲]. ازجمله کلمات توقفی، می‌توان به ضمایر، حروف اضافه و ربط اشاره کرد [۲۵]. در این مرحله عبارات توقفی از متن اخبار حذف می‌گردد.

۴- **حذف افعال:** عبارات کلیدی اغلب اسمی و صفی بوده و شامل افعال نمی‌شود. درنتیجه، می‌توان افعال را از متن حذف کرد.

۵- **ریشه‌یابی:** در این مرحله، کلمات باقیمانده ریشه‌یابی می‌شود.

پس از طی مراحل فوق که با کاهش مجموعه کلمات تحت بررسی، باعث افزایش قدرت تشخیص می‌شود، جهت استخراج عبارات کاندیدای کلیدی، به ازای هر کلمه باقیمانده همه عبارات ممکن تا طول چهار کلمه استخراج می‌شود.

۲-۳- محاسبه ویژگی‌های آماری و اکتشافی

پس از استخراج عبارات از متن اخبار، ابتدا تعدادی ویژگی برای عبارات استخراج شده محاسبه می شود. در روش ارائه شده، برای هر عبارت استخراج‌شده، ۱۸ ویژگی آماری محاسبه می‌شود. ویژگی‌های آماری مذکور، در جدول ۱ و توضیحات آن در جدول ۲ ذکرشده است.

عنوان و خلاصه اخبار معمولاً حاوی عبارات مهم هستند. بدین‌جهت، در روش ارائه‌شده، به‌عنوان شیوه‌هایی اکتشافی، عبارات این دو بخش نیز استخراج‌شده و امتیاز عبارات موجود در عنوان اخبار دو برابر در نظر گرفته می‌شود.

۲-۴- دسته‌بندی

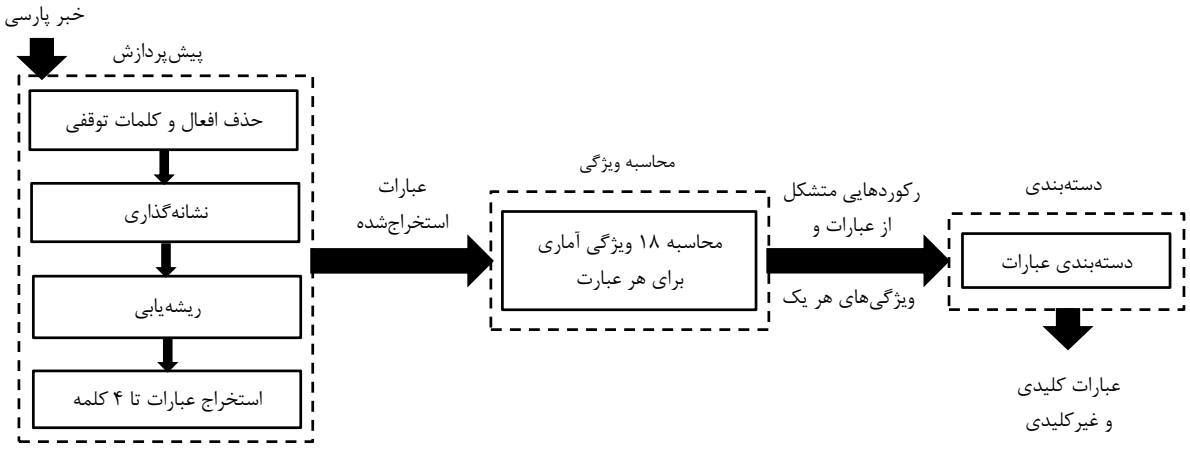
در آخرین مرحله، نوبت به تصمیم‌گیری درباره کلیدی بودن یا نبودن عبارات استخراج‌شده به کمک مقایسه مقادیر ویژگی‌های آن‌ها می‌رسد. ازآنجاکه تعداد ویژگی‌های محاسبه شده برای عبارات نسبتاً زیاد است، بررسی و مقایسه آن‌ها به کمک الگوریتم‌های یادگیری با ناظر انجام می‌شود.

در پژوهش حاضر، با توجه به اینکه عبارات استخراج‌شده مجموعه داده بزرگی ایجاد می‌کنند، و با توجه به اینکه تعداد متغیرها (ویژگی‌ها) نسبت به موارد مشابه بیشتر است، دسته‌بند جنگل تصادفی، جهت دسته‌بندی دادگان، مناسب به نظر می‌رسد. این دسته‌بند، مقیاس‌پذیری بالایی داشته و به شکلی کارا دادگان بزرگ با متغیرهای زیاد را با صحت بالا دسته‌بندی می‌کند [۲۶]. البته، عملکرد این دسته‌بند از بین شش دسته‌بند به‌کاررفته در زمینه استخراج عبارات کلیدی آزموده شده و برتری آن با توجه به نتایج موجود در بخش ۳ به اثبات رسیده است. نمای سیستم ارائه‌شده در شکل ۲ آمده است.

۳- آزمایش‌ها

۳-۱- تولید و آماده‌سازی دادگان

در این پژوهش، تولید دادگان طی سه مرحله انجام شد. حجم دادگان برای پژوهش درزمینه استخراج عبارات کلیدی، بازه‌ای حدودی بین ۳۵ تا ۳۱۲ سند بوده است [۲۳، ۲۷، ۲۸، ۲۹]. در تحقیق حاضر، مشابه مجموعه داده SemEval-2010 [۳۰] که مجموعه داده نسبتاً پرکاربردی در حوزه استخراج عبارات کلیدی در زبان انگلیسی است، ابتدا تعداد ۲۴۴ سند خبری منتشرشده در تاریخ ۱۳۹۳/۷/۸ با گستره وسیع موضوعی (جهت ارزیابی سیستم برای انواع توزیع‌های کلمات)، به‌طور تصادفی انتخاب شد.



شکل ۲: نمای کلی سیستم طراحی‌شده

جهت در نظر گرفتن توزیع انتشار انفجاری اخبار، در انتخاب دادگان از اخبار ۳۲ خبرگزاری مختلف استفاده شد؛ و متون با موضوع تکراری از بین دادگان حذف نشد. در گام دوم، برچسب‌زنی دادگان با تعیین حداقل ۱۰ عبارت کلیدی به ازای هر خبر انجام شد. به دلیل پیچیدگی و چالش‌برانگیز بودن برچسب‌زنی، جهت حصول اطمینان از صحت برچسب‌های دادگان، برچسب‌های آن‌ها طی گام سوم مورد بازبینی قرار گرفته و اصلاحات لازم اعمال گردید.

جدول ۱: ویژگی‌های آماری به‌کاررفته در سیستم پایه [۱۹]

تعداد	ویژگی	توضیح
۱	$Len(t)=\|t\|$	تعداد کلمات عبارت
۲	$TF(t,d)=\frac{f(t,d)}{\|d\|}$	فراوانی نرمال‌سازی شده عبارت
۳	$IDF(t,\ d,\ D)=\log\left(\frac{\|D\|}{df(t,D)}\right)$	معکوس فراوانی سند
۴	$TFIDF(t,d,D)=\log TF(t,d)$ $.\max\left(0,\frac{\log(\|D\|-df(t,D))}{df(t,D)}\right)$	تغیر TFIDF
۵	$FP(t,d)=npos_0(t,d)=\frac{pos_0(t,d)}{\|d\|}$	اولین مکان رخداد عبارت
۶	$FS(t,d)=\frac{sent_0(t,d)}{\|S_d\|}$	اولین جمله رخداد عبارت
۷	$HF(t,\ d,\ r)=\frac{f(t,\ head(d,\ r))}{r\|d\|},(r=0.25)$	فراوانی عبارت t در یک‌چهارم اول سند d
۸	$ASL(t,d)=\frac{\sum_{s\in S_d(t)}\left(\frac{\|s\|}{\|S_d(t)\|}\right)}{\sum_{s\in d}\left(\frac{\|s\|}{\|S_d\|}\right)}$	متوسط طول جمله
۹	$SFS(t,d)=\frac{\sum_{s\in sub(t)}f(s,d)}{\|d\|}$	مجموع فراوانی‌های زیررشته‌ها
۱۰	$GDC(t,d)=\frac{\|t\|\log(f(t,d))f(t,d)}{\sum_{c\in comp(t)}f(c,d)}$	ضریب طاس تعمیم‌یافته
۱۱	$MLE(t,d)=p(t,d)$	تخمین احتمال حداکثر
۱۲	$KLD(t,d,D)=p(t,d)\log\left(\frac{p(t,d)}{p(t,D)}\right)$	واگرایی Kullback-Leibler
۱۳	$DPM-index(t,d)=1-max_{s\in\sup(t,d)}\left(\frac{f(s,d)}{f(t,d)}\right)$	اندیس حداکثری عبارت سند
۱۴	$DPM.TFIDF(t,d,D)=DPM-index(t,d,D)$ $.TFIDF(t,d,D)$	حاصل‌ضرب DPM-index و TFIDF
۱۵	$TFIDFRatio(t,d,D)=$ $\frac{TFIDF(t,d,D)}{max_{c\in comp(t)}(TFIDF(c,d,D))}$	نرخ TFIDF عبارت
۱۸-۱۶	Position mean and 2-means	2-Means و میانگین مکان‌های نرمال‌سازی شده

۳-۲- ابزار پیش‌پردازش

جهت تولید لیست عبارات توقفی خبری پارسی از ۱۵۰ عبارت توقفی عمومی به‌عنوان پایه استفاده شد [۲۲]. در ادامه، بامطالعه دادگان پژوهش، تعداد ۱۰۲۶ عبارت (شامل برخی عبارات اضافه، افعال، نام ماه‌های میلادی، روزهای هفته و ...) به لیست اضافه و تعداد کلمات توقفی به ۱۱۷۶ عبارت رسید. پس از ایجاد لیست عبارات توقفی خبری، این عبارات شناسایی و از متن اخبار حذف گردید. به‌علاوه، با بررسی دادگان مشخص شد که اغلب اعداد موجود در اخبار غیر از برخی موارد خاص (مانند «۸ سال دفاع مقدس» و «۲۲ بهمن»)، بی‌اهمیت است. درنتیجه، اعداد (حروفی و عددی) نیز به‌عنوان عبارات توقفی شناخته‌شده و به‌جز فهرستی ده‌تایی از موارد خاص، از متن اخبار حذف گردید.

برخی عبارات نیز وجود دارد که علی‌رغم با اهمیت بودن شامل یک یا چند کلمه توقفی است (مانند «از بین بردن»، «از» و «بین» توقفی است). در این موارد، حذف کلمات توقفی، موجب از دست رفتن معنای آن خواهد شد که مطلوب نیست. درنتیجه، فهرستی ۴۸ تایی از این عبارات استخراج و از حذف آن‌ها جلوگیری شد.

برای شناسایی افعال و حذف آن‌ها از متن اخبار و نیز جهت ریشه‌یابی اسمی کلمات، به ترتیب، از دو ریشه‌یاب فعلی و اسمی به‌کاررفته در پژوهش [۳۱] استفاده شد.

۳-۳- معیار ارزیابی

نتایج آزمایش‌های انجام‌شده با معیار F1 سنجیده شده است. معیار مذکور به کمک رابطه (۱۸) محاسبه می‌شود.

$$precision = \frac{tp}{(tp+fp)} \qquad (۱۶)$$

جدول ۲: شرح علائم روابط ۱ تا ۱۵ [۱۹]

نماد	توضیح
d	سند، $\|d\|$ تعداد کلمات موجود در آن
D	مجموعه اسناد یا پیکره، $\|D\|$ تعداد اسناد موجود در D
T	مجموعه کلیه عبارات انتخاب‌شده از پیکره اسناد پس از گام پیش‌پردازش. $\|T\|$ اندازه آن
T_d	مجموعه کلیه عبارات انتخاب‌شده از سند d پس از گام پیش‌پردازش. $\|T_d\|$ اندازه آن
t	عبارتی از T، $\|t\|$ تعداد کلمات آن
s	یک جمله، $\|s\|$ تعداد کلمات موجود در آن
S_d	جملات d، $\|s_d\|$ اندازه آن
$S_d(t)$	جملات d، $\|S_d(t)\|$ اندازه آن
$head(d,r)$	بخش ابتدایی سند d با اندازه $0<R<1$ ، $R\|d\|$
$f(t,d)$	فراوانی عبارت t در سند d
$f(t,D)$	فراوانی عبارت t در پیکره D
$df(t,D)$	تعداد اسناد D که t در آن‌ها رخداده است (فراوانی سند)
$p(t,d)$	تخمینی از احتمال t با داشتن d: $p(t,d)=f(t,d)/\sum_{t'\in T_d} f(t',d)$
$p(t,D)$	تخمینی از احتمال t با داشتن D : $p(t,D)=f(t,D)/\sum_{t'\in T} f(t',D)$
$pos_n(t,d)$	n امین رخداد عبارت t در سند d (تعداد کلمات پیش از آن)
$npos_n(t,d)$	n امین رخداد نرمال شده: $npos_n(t,d)=pos_n(t,d)/\|d\|$
$sent_n(t,d)$	شماره جمله حاوی n امین رخداد عبارت t در سند d
$comp(t)$	اجزای t، کلمات عبارت t
$sub(t)$	زیر عبارات عبارت t، کلیه $m-gram$ هایی که در $n-gram$ t موجودند
$\sup(t,d)$	سرعبارات عبارت t در سند d، کلیه عبارات s انتخاب‌شده از سند d که حاوی t هستند به استثنا خود t
$\sup(t,D)$	سرعبارات عبارت t در پیکره D، کلیه عبارات s انتخاب‌شده از پیکره D که حاوی t هستند به استثنا خود t
$TF(t,d)$	فراوانی نرمال‌سازی شده یک عبارت در سند D : $TF(t,d)=f(t,d)/\|d\|$
$IDF(t,D)$	معکوس فراوانی سند عبارت t در پیکره D : $IDF(t,D)=\log(\|D\|/df(t,D))$
$TFIDF(t,d,D)$	$TFIDF(t,d)=TF(t,d)\times IDF(t,D)$

$$recall = \frac{tp}{(tp + fn)} \quad (۱۷)$$

$$F1 = \frac{2 \times recall \times precision}{(recall + precision)} \quad (۱۸)$$

دقت[۵]، عبارت است از نسبت دادگانی که به‌درستی کلیدی تشخیص داده‌شده‌اند (*tp*) به تمام دادگانی که کلیدی تشخیص داده‌شده‌اند (مجموع دادگانی که به‌درستی کلیدی تشخیص داده شده‌اند و دادگانی که به‌اشتباه کلیدی تشخیص داده‌شده‌اند (*fp*)). فراخوانی[۶] عبارت است از نسبت کلماتی که به‌درستی کلیدی تشخیص داده‌شده‌اند به تمام کلمات کلیدی موجود (مجموع دادگانی که به‌درستی کلیدی تشخیص داده‌شده‌اند و دادگانی که به‌اشتباه غیرکلیدی تشخیص داده‌شده‌اند (*fn*)). *F1* نیز نوعی میانگین وزن‌دار از دو معیار نام‌برده است.

۳-۴- دسته‌بند مناسب

در این پژوهش، جهت اثبات فرض برتری دسته‌بند جنگل تصادفی، شش شیوه دسته‌بندی رگرسیون منطقی (LR)[۷]، درخت تصمیم (DT)، LB[۸]، جنگل تصادفی (RF)، K - نزدیک‌ترین همسایه (KNN) و بیز ساده (NB)، که پیش‌ازاین در حوزه استخراج عبارات کلیدی از زبان انگلیسی به‌کاررفته‌اند، به کمک شیوه اعتبارسنجی متقابل[۹] (CV) و API های ابزار داده‌کاوی WEKA [۲۲] در جاوا روی دادگان خبری پارسی، به‌کار گرفته شد. نتایج در جدول ۳ آمده که برتری شیوه دسته‌بندی جنگل تصادفی، با توجه به نتایج آن اثبات می‌شود. چنانکه پیداست F1 سیستم، برابر ۲۵/۸۳ درصد است که برای سامانه‌های مبتنی بر زبان طبیعی معمول و قابل‌قبول است.

۳-۵- نحوه دسته‌بندی

دو راه برای دسته‌بندی عبارات وجود دارد؛ استفاده از خروجی دسته‌بند و استفاده از امتیاز تعلق‌گرفته به هر عبارت توسط دسته‌بند برای ایجاد یک لیست رتبه‌بندی شده و انتخاب n عبارت برتر به‌عنوان عبارات کلیدی. در شیوه حاضر راه دوم به کار گرفته شده است. در ادامه، دلیل انتخاب این نحوه دسته‌بندی بررسی می‌شود.

استخراج عبارات کلیدی یک مسئله دسته‌بندی دو کلاسه بوده که معمولاً از تعداد زیاد عبارات کاندیدا، تعداد بسیار کمی کلیدی است. در چنین شرایطی، دادگان به‌اصطلاح نامتوازن است. در شرایط عدم توازن، با دسته‌بندی به کمک خروجی دسته‌بند، ازآنجاکه، تعداد عبارات کلیدی دادگان آموزشی بسیار کم است، دسته‌بند با دقت بالایی (از ۹۰ تا ۱۰۰ درصد) دادگان غیرکلیدی را شناسایی می‌کند. اما، دقت شناسایی عبارات کلیدی بسیار پایین (بین ۰ تا ۱۰ درصد) بوده و تقریباً کلیه عبارات، غیرکلیدی تشخیص داده می‌شود [۳۲].

بر اساس بررسی‌های انجام‌شده، در پژوهش‌های مربوط به استخراج عبارات کلیدی، تاکنون عدم توازن دادگان مطرح نشده است. درحالی‌که، این مسئله به‌ویژه زمانی که از خروجی دسته‌بندها استفاده می‌شود، می‌تواند دقت سیستم را به‌شدت تحت تأثیر قرار دهد. جهت نمایش عدم توازن دادگان، آمار عبارات استخراج‌شده در هر بخش از شیوه اعتبارسنجی متقابل پنج‌بخشی، در جدول ۴ آمده است. تعداد اسناد آزمون و آموزشی در هر بخش، به ترتیب ۴۸ و ۱۹۲ سند است.

با توجه به جدول ۴، توزیع عبارات غیرکلیدی در مقابل عبارات کلیدی در دادگان آموزشی به‌طور میانگین حدود ۹۷ درصد در مقابل ۳ درصد و نشان‌دهنده عدم توازن شدید دادگان است.

روش‌های برخورد با عدم توازن دادگان را می‌توان به چهار دسته کلی روش‌های نمونه‌برداری، روش‌های حساس به هزینه[۱۰]، روش‌های مبتنی بر هسته[۱۱] و یادگیری فعال[۱۲] و روش‌های جنبی تقسیم‌بندی کرد [۳۲]. در این مقاله، با توجه به‌سادگی روش‌های نمونه‌برداری، تأثیر پنج روش نمونه‌برداری، با دو توزیع ۵۰ درصد کلیدی / ۵۰ درصد غیرکلیدی و ۲۵ درصد کلیدی / ۷۵ درصد غیرکلیدی بررسی و نتایج حاصل در جدول ۵ آمده است.

در نمونه‌برداری هدف، ایجاد توازن در دادگان به کمک حذف یا اضافه کردن نمونه‌های آن‌ها است. شیوه‌های نمونه‌برداری آزمایش‌شده عبارت است از بیش‌نمونه‌برداری ساده[۱۳] (SOS)، کم‌نمونه‌برداری ساده[۱۴](SUS)، کم‌نمونه‌برداری K تا نزدیک‌ترین همسایه[۱۵] (- KNN US)، ترکیبی ساده[۱۶] (EE)، نمونه‌برداری مصنوعی انطباقی[۱۷] (ADASYN)، بیش‌نمونه‌برداری مصنوعی[۱۸] (SYN - OS) [۳۲].

در شیوه SYN - OS، به دلیل استفاده از اعداد تصادفی، نتایج نسبتاً ناپایدار است. بدین‌جهت، میانگین نتایج سه بار اجرای آن در جدول ۵ آمده است. واضح است که نمونه‌برداری تأثیری در بهبود نتایج نداشته است. جهت بررسی اثر نحوه دسته‌بندی بر میزان اثر روش‌های نمونه‌برداری، میانگین نتایج سه بار اجرای شیوه SYN - OS با توزیع ۵۰-۵۰ با خروجی دسته‌بند، در جدول ۶ آمده است.

جدول ۳: عملکرد سیستم استخراج عبارات کلیدی به کمک شش دسته‌بند مختلف

معیارهای ارزیابی	الگوریتم‌های دسته‌بندی به‌کاررفته					
	LR	DT	LB	RF	KNN	NB
tp	۹۳	۱۰۴	۱۱۲	۱۲۴	۴۶	۸۵
fp	۳۸۶	۳۷۵	۳۶۷	۳۵۶	۴۳۳	۳۹۴
tn	۱۵۳۳۷	۱۵۳۴۸	۱۵۳۵۶	۱۵۳۶۸	۱۵۲۹۰	۱۵۳۲۹
fn	۳۸۶	۳۷۵	۳۶۷	۳۵۶	۴۳۳	۳۹۴
F1 (%)	۱۹/۴۲	۲۱/۷۱	۲۳/۵	**۲۵/۸۳**	۹/۷۵	۱۷/۷۵

جدول ۴: توزیع دادگان آموزشی در بخش‌های مختلف شیوه اعتبارسنجی متقابل پنج‌بخشی

آمار دادگان	شماره بخش‌ها					میانگین مقادیر
	۱	۲	۳	۴	۵	
تعداد نمونه‌های آموزشی	۵۹۷۱۴	۶۰۴۱۵	۶۷۸۱۲	۶۵۱۱۹	۷۱۰۲۰	۶۴۸۱۶
تعداد نمونه‌های کلیدی نمونه‌های آموزشی	۱۷۱۸	۱۷۰۶	۱۶۲۶	۱۶۷۱	۱۶۶۷	۱۶۷۸

جدول ۵: نتایج به‌کارگیری شیوه‌های نمونه‌برداری در سیستم بهبودیافته

شیوه‌های نمونه‌برداری	توزیع دادگان غیرکلیدی - کلیدی									
	۵۰-۵۰					۲۵-۷۵				
	معیارهای ارزیابی									
	tp	*fp*	*tn*	*fn*	*F1*	*tp*	*fp*	*tn*	*fn*	*F1*
SOS	۱۰۱	۳۷۸	۱۲۵۹۴	۳۷۸	۲۱/۲۱	۱۰۰	۳۷۹	۱۲۵۹۳	۳۷۹	۲۱
SUS	۴۹	۴۳۰	۱۵۲۹۳	۴۳۰	۱۰/۲۹	۶۰	۴۱۹	۱۵۳۰۴	۴۱۹	۱۲/۵۸
KNN - US	۳۶	۴۴۳	۱۵۲۸۰	۴۴۳	۷/۵۴	۵۲	۴۲۷	۱۵۲۹۶	۴۲۷	۱۰/۹۶
EE	۴۳	۴۳۶	۱۵۲۸۷	۴۳۶	۹/۰۴	۵۸	۴۲۱	۱۵۳۰۲	۴۲۱	۱۲/۲۵
SYN - OS	۱۱۹	۳۶۰	۱۵۳۶۳	۳۶۰	۲۴/۷۹	۱۱۶	۳۶۳	۱۵۳۶۰	۳۶۳	۲۴/۱۸

جدول ۶: نتایج سیستم به کمک خروجی دسته‌بند، با و بدون نمونه‌برداری

معیارهای ارزیابی	خروجی سیستم	خروجی سیستم با SYN - OS ۵۰-۵۰
precision	۵۴/۸۹	۳۲/۲۴
recall	۴/۲۹	۱۶/۴۸
F1 (%)	۷/۹۷	۲۱/۸

با توجه به جدول ۶، روش نمونه‌برداری SYN - OS میزان F1 سیستم را حدود سه برابر افزایش داده است. بدین ترتیب، روش‌های نمونه‌برداری برخلاف دسته‌بندی به کمک امتیاز نمونه‌ها، بر دسته‌بندی به کمک خروجی دسته‌بند بسیار تأثیرگذار است. چراکه، در روش اول، احتمال کلیدی بودن نمونه‌ها نسبت به یکدیگر سنجیده می‌شود. درحالی‌که، در روش دوم احتمال کلیدی بودن هر نمونه با احتمال غیرکلیدی بودن همان نمونه مقایسه می‌شود. نمونه‌برداری با دست‌کاری توزیع کلیدی - غیرکلیدی دادگان، امتیاز یا احتمال کلیدی بودن در مقابل غیرکلیدی بودن نمونه‌ها را تغییر می‌دهد؛ و تأثیر چندانی بر امتیاز کلیدی بودن نمونه‌ها نسبت به یکدیگر ندارد.

در سیستم حاضر، ازآنجاکه دسته‌بندی به کمک امتیاز نمونه‌ها امکان کنترل تعداد عبارات استخراجی از هر متن را فراهم می‌کند؛ و استفاده از خروجی دسته‌بند (حتی با نمونه‌برداری) دقت پایین‌تری نسبت به استفاده از امتیاز نمونه‌ها دارد، از امتیاز نمونه‌ها جهت دسته‌بندی استفاده شده است.

۳-۶- خلاصه الگوریتم ارائه‌شده

۱- پیش‌پردازش و استخراج عبارات کاندیدا (جداسازی جملات از عنوان، خلاصه و متن اخبار، حذف عبارات توقفی و افعال، نشانه‌گذاری کلمات، ریشه‌یابی، استخراج تمام عبارات ممکن تا طول چهار کلمه به ازای هر کلمه باقیمانده)

۲- محاسبه ویژگی (محاسبه ۱۸ ویژگی آماری به ازای هر عبارت کاندیدا)

۳- تصمیم‌گیری در مورد کلیدی بودن یا نبودن کاندیداها (دسته‌بندی عبارات کاندیدا به کمک دسته‌بند جنگل تصادفی به دو دسته عبارات کلیدی و غیرکلیدی)

۴- بحث و تحلیل نتایج

عملکرد ابزار پیش‌پردازش از عوامل مهم مؤثر بر دقت سیستم است. عبارات استخراج‌شده، پیش از دسته‌بندی در مرحله پیش‌پردازش حرص می‌شود. هرچقدر عبارات غیرکلیدی بیشتر و عبارات کلیدی کمتری در این مرحله حذف شود (خطای پیش‌پردازش کمتر باشد)، عملکرد ابزار پیش‌پردازش بهتر بوده است. آمار عبارات استخراج‌شده در جدول ۷ آمده است.

جدول ۷: آمار عبارات استخراج‌شده

مرحله پیش‌پردازش	آمار عبارات
۸۱۷۷۶	تعداد کل عبارات استخراجی
۲۱۳۵	تعداد عبارات کلیدی شناسایی‌شده
۴۲۴	اختلاف با تعداد برچسب‌های دستی (۲۵۵۹)

با توجه به جدول ۷، حدود ۱۶ درصد از عبارات کلیدی در مرحله پیش‌پردازش از سیستم حذف می شود. نظر به برابری تقریبی عبارات کلیدی مشخص شده در دادگان (برچسب‌های دستی) با تعداد عبارات کلیدی موردنظر برای استخراج از هر متن (۱۰ عبارت)، ریزش عبارات کلیدی پیش از دسته‌بندی تأثیر زیادی بر دقت سیستم دارد.

یکی از دلایل نبود برخی عبارات کلیدی در مجموعه عبارات حاصل از مرحله پیش‌پردازش، وجود برچسب‌های کلیدی با طول بیش از چهار کلمه است. جهت تعیین سهم این برچسب‌ها در ریزش عبارات کلیدی، آمار تعداد برچسب‌های دستی به تفکیک طول آن‌ها در جدول ۸ آمده است.

جدول ۸: تعداد عبارات کلیدی به تفکیک طول آن‌ها

تعداد عبارات کلیدی	طول عبارات استخراج‌شده
۹۱۸	۱
۹۴۹	۲
۲۹۸	۳
۱۵۶	۴
۲۳۸	+۴

چنانکه از جدول ۸ پیداست، حدود ۵۶ درصد از ۴۲۴ عبارت غیرقابل‌شناسایی، طولی بالای چهار کلمه داشته‌اند.

یکی دیگر از دلایل عدم شناسایی برخی برچسب‌های دستی، وجود حروف ربط و اضافه در تعدادی از آن‌ها است. حروف ربط و اضافه در مرحله پیش‌پردازش از سیستم حذف می‌شود. درنتیجه، استخراج عبارات حاوی آن‌ها امکان‌پذیر نیست. آمار عبارات کلیدی حاوی حروف ربط و اضافه به تفکیک طول آن‌ها در جدول ۹ آمده است.

جدول ۹: آمار عبارات کلیدی حاوی حروف ربط و اضافه به تفکیک طول

تعداد عبارات کلیدی دارای حروف ربط و اضافه	طول عبارات کلیدی
۱	۲
۲۷	۳
۵۹	۴
۱۸۶	+۴

با توجه به جدول ۹، از مجموع ۲۷۳ عبارت کلیدی حاوی حروف ربط و اضافه در مجموعه دادگان، ۸۷ عبارت با طول چهار کلمه و کمتر از آن دارای حروف ربط و اضافه بوده و در سیستم حاضر قابل‌شناسایی نیستند.

بدین ترتیب، از ۴۲۴ عبارت ناموجود در مجموعه عبارات ورودی دسته‌بند، ۳۲۵ عبارت (مجموع ۸۷ عبارت کلیدی دارای حروف ربط و اضافه با طول چهار کلمه و کمتر از آن و ۲۳۸ عبارت کلیدی با طول بالای چهار کلمه) یعنی حدود ۷۶/۶۵ درصد کل عبارات ناموجود در مجموعه عبارات ورودی دسته‌بند، قابل‌شناسایی نیستند. برای نبود ۹۹ عبارت باقیمانده در مجموعه دادگان ورودی دسته‌بند می‌توان به وجود برخی عبارات توقفی (غیر از حروف ربط و اضافه) در برچسب‌های دستی اشاره کرد.

۵- نتیجه‌گیری

به‌طورکلی، استخراج عبارات کلیدی از دادگان متنی یکی از مهم‌ترین و پرکاربردترین مسائل در متن‌کاوی و بازیابی اطلاعات بوده و در اغلب امور مرتبط با محتوای متن، تشخیص عبارات کلیدی نقش تعیین‌کننده‌ای دارد. در این پژوهش، راهکاری سه مرحله‌ای (شامل پیش‌پردازش، محاسبه ویژگی و دسته‌بندی) با ترکیب انواع شیوه‌های آماری، زبان‌شناختی، یادگیری با ناظر و اکتشافی برای استخراج عبارات کلیدی از اخبار پارسی ارائه شد. همچنین، کاربرد شش شیوه دسته‌بندی با ناظر روی اخبار پارسی موردبررسی قرار گرفت.

به‌علاوه، وجود عدم توازن در توزیع کلیدی - غیرکلیدی دادگان، تأثیر پنج شیوه نمونه‌برداری در دو توزیع متفاوت، در کنار شیوه دسته‌بندیِ استفاده از خروجی دسته‌بندها و شیوه دسته‌بندیِ امتیاز نمونه‌ها، مورد بررسی قرار گرفت. در کنار موارد مذکور، یک مجموعه دادگان خبری حاوی ۲۴۴ خبر از خبرگزاری‌های مختلف بدین منظور انتخاب و به‌صورت دستی برچسب‌زنی شده و یک لیست عبارات توقفی خبری حاوی ۱۷۶۰ عبارت توقفی ایجاد شد. همچنین، جهت کاهش خطای ناشی از حذف عبارات توقفی، دو لیست استثنائات عددی و حرف اضافه‌ای ایجاد، و از حذف آن‌ها جلوگیری شد.

آزمایش‌های انجام شده، نشان‌دهنده برتری عملکرد دسته‌بند جنگل تصادفی در میان شش دسته‌بند آزموده شده می‌باشد. به علاوه، دسته‌بندی به کمک امتیاز تعلق‌گرفته به نمونه‌ها توسط دسته‌بند مذکور نتایج بهتری نسبت به استفاده از خروجی دسته‌بند داشته است.

مقدار معیار *F1* به‌دست‌آمده برای روش ارائه‌شده با استفاده از امتیاز تعلق‌گرفته به نمونه‌ها توسط دسته‌بند جنگل تصادفی و شیوه اعتبارسنجی متقابل پنج‌بخشی، ۲۵/۸۳ درصد است که برای سامانه‌های مبتنی بر زبان طبیعی معمول است. بااین‌حال، بهبود دقت ابزار پیش‌پردازش و پژوهش هر چه بیشتر در این زمینه می‌تواند باعث بهبود دقت استخراج عبارات کلیدی از متون پارسی شود.

ضمناً، از کل عبارات کلیدی و مهم موجود در متون تنها تعداد معینی (حدود ۱۰ عبارت به ازای هر متن) برچسب‌زنی شده‌اند. بدین معنی که تعدادی از عباراتی که به‌درستی کلیدی تشخیص داده می‌شوند، به دلیل عدم حضور در برچسب‌های دستی دادگان جز مواردی قرار می‌گیرند که به‌اشتباه کلیدی تشخیص داده‌شده‌اند. در این شرایط، می‌توان گفت معیارهای ارزیابی به‌درستی نشان‌دهنده عملکرد سیستم نیستند.

به نظر می‌رسد نبود ابزار معنایی همچون برچسب‌زن ادات سخن[۱۹]، در کاهش دقت سیستم مؤثر بوده است. گاهی کلمات متن به‌طور جداگانه دارای محتوایی که نشان‌دهنده موضوع خبر باشد، نیستند. در این حالت، آنچه اهمیت دارد معنای جمعی عبارات در هم رخدادی با یکدیگر است. بنابراین، ممکن است، در نظر گرفتن هم رخدادی کلمات نیز بتواند در افزایش دقت سیستم مؤثر باشد. بررسی این موارد و بهبود ابزار پیش‌پردازش و برچسب‌های دادگان به کارهای آینده موکول می‌شود.

مراجع

[1] B. Lott, "Survey of keyword extraction techniques," Technical Report, University of New Mexico, 2012, http://www.cs.unm.edu/~pdevineni/papers/Lott.pdf (Accessed 8/20/2014).

[2] J. Kaur and V. Gupta, "Effective approaches for extraction of keywords," *IJCSI Int. J. Comput. Sci. Issues*, vol. 7, no. 6, pp. 144–148, 2010.

[3] M. Rajman and R. Besan, "Text mining - knowledge extraction from unstructured textual data," in *Proceedings of the 6th Conference of International Federation of Classification Societies*, 1997.

[4] D. B. Bracewell, "Category classification and topic discovery of Japanese and English news articles," *Electron. Notes Theor. Comput. Sci.*, vol. 225, pp. 51–65, 2009.

[5] V. Gupta, L. C. Science, and G. S. Lehal, "A survey of text mining techniques and applications," *J. Emerg. Technol. WEB Intell.*, vol. 1, no. 1, pp. 60–76, 2009.

[۶] محمدعلی زارع چاهوکی، سیده زهرا آفتابی، «کاهش شکاف معنایی در دسته‌بندی پرسش‌ها با بهره‌گیری از قوانین طبقه‌بندی»، *مجله مهندسی برق دانشگاه تبریز*، دوره ۴۶، شماره ۳، صفحه ۱۳-۲۴، ۱۳۹۵.

[7] H. Frigui and O. Nasraoui, "Simultaneous clustering and dynamic keyword weighting for text documents," in *Survey of Text Mining Clustering, Classification, and Retrieval Scanned by Velocity*, M. W. Berry, Springer New York, pp. 45–72, 2004.

[8] M. Castellanos, "HotMiner: discovering hot topics from dirty text," in *Survey of Text Mining Clustering, Classification, and Retrieval Scanned by Velocity*, M. W. Berry, Ed. Springer-Verlag New York, p. 123 to 157, 2004.

[9] Y.-H. Tseng, "Multilingual keyword extraction for term suggestion," in Proceedings of the 21st annual international conference on Research and development in information retrieval, pp. 377–378, 1998.

[۱۰] رضا خدایی، محمدعلی بالافر، سیدناصر رضوی، «اثربخشی بسط پرس‌وجو مبتنی بر خوشه‌بندی اسناد شبه‌بازخورد با الگوریتم K-NN»، *مجله مهندسی برق دانشگاه تبریز*، دوره ۴۶، شماره ۱، صفحه ۱۴۳-۱۵۱، ۱۳۹۵.

[11] C. Wartena, "Keyword extraction using word co-occurrence," in *Workshop on Database and Expert Systems Applications*, 2010.

[12] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988.

[13] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Doc.*, vol. 60, no. 5, pp. 493–502, 2004.

[14] D. Peter, "Learning algorithms for keyphrase extraction," *Inf. Retr. Boston.*, vol. 2, no. 4, pp. 303–336, 2000.

[15] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-manning, "Domain-specific keyphrase extraction," in *Proceeding of International Joint Conference on Artificial Intelligence*, no. 16, pp. 668–673, 1999.

[16] J. Hong, and M. Fang "Keyword extraction and semantic tag prediction," *unpublished(http://cs229. stanford. edu/proj2013/FangHong-Keyword% 20Extraction% 20and% 20Semantic% 20Tag% 20Prediction. pd f)*.

[17] B. Fortuna, C. Galleguillos, N. Cristianini, Detection of bias in media outlets with statistical learning methods in *Text Mining Classification, Clustering, and Applications*, A. N. Srivastava and M. Sahami, Eds. Taylor and Francis Group, LLC, p. 27-50, 2009.

[18] T. K. Yasubumi Sakakibara and Kazuo Misue, "Text classification and keyword extraction by learning decision trees," in *Conference on Artificial Intelligence for Applications*, no. 9, p. 466, 1993.

[19] M. Haddoud and S. Abdeddaim, "Accurate keyphrase extraction by discriminating overlapping phrases," *J. Inf. Sci.*, vol. 40, no. 4, pp. 488–500, Apr. 2014.

[20] C. Wang, M. Zhang, L. Ru, and S. Ma, "An automatic online news topic keyphrase extraction system," in *International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 1–6, 2008.

[21] W. Bohemia and C. Republic, "Automatic keyphrase extraction based on NLP and statistical methods," in *Annual International Workshop on DAtabases, TExts, Specifications and Objects, Pisek, Czech Republic*, no. 11, pp. 140–145, *2011*.

[۲۲] محمدرضا میبدی و مسعود تشکری، «ساخت یک نمایه‌ساز خودکار برای متون فارسی»، *یازدهمین کنفرانس مهندسی برق*، ۱۳۸۲.

[23] C. B. Ali, R. Wang, and H. Haddad, "A two-level keyphrase extraction approach," in *Computational Linguistics and Intelligent Text Processing*, vol. 9042, pp. 390–401, 2015.

[24] C.Ramasubramanian and R.Ramya, "Effective pre-processing activities in text mining using improved porter ' s stemming algorithm," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 2, no. 12, pp. 4536–4538, 2013.

[25] M. Saraswathi and V. Balu, "Preprocessing techniques for effective data extraction and computation," *IUP J. Comput. Sci.*, vol. 7, no. 3, 2013.

[26] L. Breiman, "Random forests," Machine Learning Journal, vol. 45, no. 1, 2001.

[27] K. S. Hasan and V. Ng, "Automatic keyphrase extraction : a survey of the state of the art," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*,no. 52, pp. 1262–1273, 2014.

[28] N. G. Ali and N. Omar, "A hybrid of statistical and machine learning methods for arabic keyphrase extraction," *Asian J. Appl. Sci.*, vol. 8, no. 4, pp. 269–276, 2015.

[29] K. Zhang, H. Xu, J. Tang, and J. Li, "Keyword extraction using support vector machine," in *Proceedings of the international conference on Advances in Web-Age Information Management*, no. 7, pp. 85–96, 2006.

[30] S. N. Kim, O. Medelyan, M. Kan, and T. Baldwin, "SemEval-2010 task 5 : automatic keyphrase extraction from scientific articles," in *Proceedings of the International Workshop on Semantic Evaluation*, no. 5, pp. 21–26, July, 2010.

[۳۱] مریم باسره، ولی درهمی، سجاد ظریف‌زاده و صادق طاهرزاده، «به‌کارگیری شیوه‌های آماری و تصمیم‌گیری چند معیاره در استخراج عبارات کلیدی از صفحات خبری وب پارسی»، *کنفرانس بین‌المللی فناوری اطلاعات و دانش*، هفتمین دوره، دانشگاه ارومیه، ۱۳۹۴.

[32] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

زیرنویس‌ها

[1] Term suggestion
[2] Query Expansion
[3] Damping Factor
[4] Rich get richer
[5] Precision
[6] Recall
[7] Logistic regression
[8] Logit Boost
[9] Cross Validation
[10] Cost-sensitive Methods
[11] Kernel-based Methods
[12] Active Learning Methods
[13] Simple Over Sampling
[14] Simple Under Sampling
[15] K Nearest Neighbors Under Sampling
[16] Simple Ensemble
[17] Adaptive Synthetic sampling
[18] Synthetic Sampling
[19] Part Of Speech tagger