مجله‌ی علمی پژوهش در علوم رایانه ، شماره ۱۹، پاییز ۱۳۹۹، ص ۵۴- ۴۲

ISSN: 2745-3673
http://www.Rcsj.ir

خلاصه سازی متون بر استفاده از الگوریتم جستجوی فاخته و ترکیب آن با تحلیل احساسات

آرزو حمیدی[۱]، عبدالرضا حاتم لو[۲]
[۱] گروه مهندسی کامپیوتر، واحد خوی، دانشگاه آزاد اسلامی، خوی، ایران.
[۲] گروه مهندسی کامپیوتر، واحد خوی، دانشگاه آزاد اسلامی، خوی، ایران.

نام نویسنده مسئول:
عبدالرضا حاتم لو

تاریخ دریافت: ۱۳۹۹/۵/۱۲
تاریخ پذیرش: ۱۳۹۹/۷/۸

چکیده

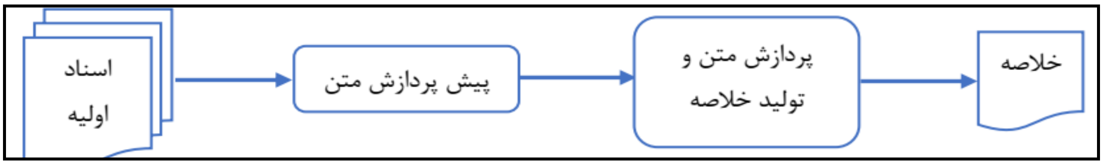
امروزه با توجه به فراوانی اسناد در اینترنت که بیشتر آنها محتوی اطلاعات غیر ضروری می باشند، اهمیت خلاصه سازی متون به منظور کاهش زمان مطالعه از اهمیت خاصی برخوردار است. افزایش چشمگیر این نوع اطلاعات، اهمیت وجود ابزارهایی برای خلاصه سازی خودکار منابع متنی را بیش از هر زمان دیگری آشکار می کند. خلاصه سازی در اصل، فرآیند فشرده سازی یک منبع بوده، به نحوی که نتیجه کار شامل اطلاعات مهم آن منبع باشد. به عبارت دیگر، استخراج قسمت های مهم از یک یا چند متن را خلاصه سازی می گویند. با وجود تکنیک های متعددی که برای خلاصه سازی متون وجود دارد، هنوز یک راه حل بهینه که برای تمامی کاربردها مناسب باشد وجود ندارد. در این مقاله با بکارگیری الگوریتم جستجوی فاخته و ترکیب آن با تحلیل احساسات، دقت خلاصه های استخراجی افزایش یافته است. فرآیند تحقیق شامل مراحل پیش پردازش، توزیع گوسی، تابع برازش و اجرای الگوریتم خلاصه سازی متن می باشد. مرحله پیش پردازش شامل تقسیم بندی جملات، توکنیزاسیون یا نشانه گذاری، حذف کلمه توقف و زیر وظایف شامل حذف ریشه لغات می باشد. الگوریتم پیشنهادی این تحقیق بر روی ۱۰ سند متفاوت جمع آوری شده از مجموعه داده های DUC2007، مورد ارزیابی قرار گرفته است. داده های DUC2007 در مجموع شامل ۴۵ موضوع بوده که هر کدام شامل ۲۵ سند می باشد. بسته ROUGE-1.5.5 برای ارزیابی خلاصه متن با معیارهای ارزیابی مورد استفاده قرار گرفته است. سیستم خلاصه سازی پیشنهادی در مقایسه با سایر سیستم های خلاصه سازی با استفاده از بسته ROUGE-1.5.5 در معیارهای Recall, Precision و F-measure نتایج بهتری را نشان می دهد.

واژگان کلیدی: خلاصه سازی متن، تحلیل احساسات، الگوریتم بهینه سازی فاخته، ابزار ارزیابی ROUGE.

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

با توجه به رشد روز افزون اطلاعات در عصر اینترنت نیاز به یک ابزار دقیق و کارآمد جهت استخراج اطلاعات مهم و مرتبط احساس می گردد. این مساله با مکانیزم خلاصه سازی خودکار متن حل می گردد. خلاصه سازی می تواند به دو صورت خلاصه سازی استخراجی و خلاصه سازی چکیده انجام شود. در روش اول یعنی خلاصه سازی استخراجی جملات مهم به همان که صورتی در متن اصلی بیان شدند، شناسایی و لفظ به لفظ در متن خلاصه کپی می شوند. در روش خلاصه سازی چکیده، جملات متن خلاصه برگرفته از متن اولیه هستند. به عبارت دیگر جملات خلاصه عینا در متن اصلی وجود ندارند. از دیدگاه دیگر، خلاصه سازی را می توان به دو روش کلاسیک و نیز روش مبتنی بر یادگیری پیاده سازی نمود.در روش خلاصه سازی کلاسیک، پس از پیش پردازش متن ورودی، با توجه به نشانگرهای جمله مانند نقطه، آن را به جملات موجود تقسیم بندی می کنیم. سپس هر جمله به صورت برداری از (ویژگی ها و مقدار آن ویژگی در جمله) نمایش داده می شود. این ویژگی ها مانند موقعیت جمله، متوسط تکرار کلمات در آن جمله و تعداد کلمات موجود در عنوان از پیش تعریف شده هستند. همچنین ارزش و اهمیت هر ویژگی نیز ثابت می باشد. بنابراین به هرجمله با توجه به مقدار و ارزش آن ویژگی، امتیازی داده می شود و در نهایت با توجه به مجموع امتیاز هر جمله، جملات دارای با ارزش بالاتر در خلاصه قرار می گیرند. در روش خلاصه سازی بر اساس تکنیک های یادگیری ماشین پس از پیش پردازش متن ورودی، با توجه به نشانه های خاص، آن را به جملات و زیرجملات می شکند. سپس هر جمله با یک مجموعه ویژگی مانند موقعیت جمله، متوسط تکرار کلمات در آن جمله و تعداد کلمات موجود در عنوان بیان می شود. این ویژگی ها توسط یک بردار ویژگی برای آن جمله در نظر گرفته می شوند. چالش مهمی که در اینجا وجود دارد انتخاب ویژگی های صحیح برای هر جمله و نیز تعیین اهمیت هر ویژگی می باشد. در خلاصه ساز کلاسیک به هر یک از ویژگی ها ارزش ثابتی داده می شود، اما در این نوع خلاصه ساز ها به یک الگوریتم یادگیری ناظر و همچنین مجموعه آموزشی جهت آموزش طبقه بند و ارزش ویژگی ها نیاز است. به عبارت دیگر در ابتدا یک طبقه بندی کننده مانند دسته بندی کننده بیز و یا شبکه عصبی را انتخاب می شود. سپس با ورود یک مجموعه آموزشی از جملات که به صورت بردار ویژگی نمایش داده شده اند به طبقه بند انتخابی، و با توجه به اینکه جمله مورد نظر در خلاصه موجود است یا خیر، به هر ویژگی ضریب اهمیت پویایی داده می شود. حال به ازای هر متن اولیه که به عنوان مجموعه تست شناخته می شود، ابتدا آن متن به جملاتی تقسیم شده و هر جمله به صورت بردار ویژگی نمایش داده می شود. سپس هر بردار ویژگی به خلاصه ساز آموزش دیده وارد شده و با توجه به مقدار کلاس خروجی، تعیین می شود آیا آن جمله در خلاصه وجود دارد یا خیر. همچنین با استفاده از این روش می توان با ورود انواع ویژگی و بررسی خلاصه به دست آمده ویژگی ها های نامرتبط را تعیین و از مجموعه ویژگی حذف نمود. به عبارت دیگر با استفاده از تکنیک های یادگیری ماشین میتوان ضرایب و اهمیت هر یک از ویژگیهای جملات را تعیین نمود .خلاصه سازی به این روش، Precision و Recall بیشتری نسبت به تکنیک کلاسیک خلاصه ساز که در آن ضرایب ثابتی به عنوان اهمیت ویژگی های تعریف شده برای جملات در نظر گرفته می شود داراست. همچنین حذف ویژگی های زائد می تواند سبب افزایش دقت خلاصه ساز شود. در حالت کلی گام های خلاصه سازی خودکار متن به شکل شماره ۱ می باشد:



شکل ۱- گام های خلاصه سازی خودکار متن

پارامترهای سیستم خلاصه ساز

یک سیستم خلاصه ساز مانند بسیاری از سیستم های دیگر، شامل پارامترهایی می باشد که در ذیل به تفضیل بیان می گردد:

- نرخ فشرده سازی خلاصه ساز
- مخاطب، منظور خلاصه عمومی یا مبتنی بردرخواست کاربر می باشد.
- ارتباط خلاصه با منبع اصلی؛ چکیده یا گزینشی بودن خلاصه مشخص می شود.
- ارجاع؛ اخباری یا اطلاع بودن خلاصه را مشخص می نماید.
- همبستگی؛ مشخص می کند که واحد خلاصه کلمه، عبارت یا جمله یا پاراگراف هست.
- گستره؛ مشخص می کند که تولید خلاصه از یک سند و یا چند سند صورت می پذیرد .
- زبان؛ سیستم خلاصه ساز ممکن است قادر به تولید خلاصه از متن های به زبان های مختلف باشد. یا فقط برای یک زبان خاص طراحی شده باشد. همچنین ممکن است که خلاصه تولیدی و متن اصلی در دو زبان مختلف باشند یا اینکه هر دو در یک زبان باشند.
- رسانه؛ منبع ورودی را مشخص می کند که متن، ویدئو، عکس و... می باشد.
- دامنه؛ در نوع مستقل از دامنه، متن می تواند از حوزه ای به سامانه داده شود اما در نوع وابسته به دامنه، سامانه تنها قادر به خلاصه سازی موثر متون مربوط به یک حوزه ی معین است.

مساله تحقیق و ابعاد آن

سوالی که در این پژوهش مطرح است این است که آیا با ترکیب تحلیل احساسات و الگوریتم بهینه سازی فاخته می توان متنی را به صورت بهینه خلاصه سازی نمود؟ شروع تحقیقات در زمینه خلاصه سازی خودکار متن به سال ۱۹۵۰ بر می گردد. در ابتدا محققان به تمامی جنبه های خلاصه سازی توجه نداشتند و آن را امری بسیار راحت تر از آنچه اکنون به آن نگریسته می شود، تصور می کردند. اما با شروع تحقیقات و ساخت خلاصه سازی خودکار دریافتند که ماشینی کردن خلاصه سازی کار چندان راحتی نیست. بنابراین بعد از تحقیقات اولیه، این زمینه تحقیقاتی برای حدود یک دهه بی فروغ شد. خلاصه سازی همانگونه که در بخش قبلی به آن اشاره شد، از دو نوع استخراجی و چکیده است. در خلاصه های استخراجی با استفاده از ویژگی های سطحی متن جمله های مهم تشخیص داده می شوند. این جمله های مهم خلاصه را تشکیل می دهند. در خلاصه های چکیده ای معمولا نیاز به درک متن می باشد. البته می توان بدون درک متن نیز خلاصه هایی از این نوع چکیده ها را تولید کرد. زیرا طبق تعریف، چکیده به خلاصه ای گفته می شود که شامل جملاتی باشد که در متن اولیه وجود ندارند. ادغام و فشرده سازی جملات، باعث تولید جملات جدیدی می شوند بدون آنکه متن اولیه به وسیله ماشین درک شود. هر کدام از دو دسته ذکر شده در بالا دارای مشکلاتی می‌باشد که در ذیل به تعدادی از آنها اشاره می شود:

- **مشکلات روش‌های چکیده:** بزرگترین چالش برای این روش‌ها، مساله نمایش دوباره مفاهیم اصلی متن در قالب کلمات و جملات می‌باشد. رسیدن به راه‌حلی برای مساله، نیازمند تحلیل معنایی وسیع متون می‌باشد. بنابراین، سیستم خلاصه-ساز متن باید کلیه قابلیت‌های مرتبط با زبان طبیعی را داشته باشد.
- **مشکلات روشهای استخراجی:** امکان دارد که جملات استخراج شده طولانی تر از متوسط طول جملات باشند و این باعث افزایش فضای مصرفی می شود. ممکن است که تناقض بین موضوعات به خوبی نشان داده نشود. همچنین ممکن است که بعد از خلاصه سازی، با حذف تعدادی از جملات بین ضمایر و اسم های مرجع آن ها تداخلاتی پیش آید.

با وجود تکنیک های متعددی که برای خلاصه سازی متون وجود دارد، هنوز چالشی اساسی برای تولید یک راه حل بهینه وجود دارد. در این پژوهش با بکارگیری الگوریتم جستجوی فاخته و ترکیب آن با تحلیل احساسات دقت خلاصه های استخراجی افزایش یافته و روشی برای حل مشکل بهینه سازی در این زمینه ارائه می‌شود.

ISSN: 2745-3673
http://www.Rcsj.ir

مرور ادبیات تحقیق

سیستم های خلاصه ساز در دنیای امروز کاربردهای فراوانی دارند. تولید خلاصه های کتب مختلف و مقالات علمی، تولید خلاصه اخبار و انتقال آن از طریق سیستم های نظیر تلفن همراه، نمایش خلاصه سند یافته شده توسط موتور جستجو، تولید سیستم های پاسخ گویی به سوال و ... همگی از کاربردهای این سیستم می باشند. همزمان با پیشرفت در حوزه داده کاوی و فناوری اطلاعات، روش ها و الگوریتم های متعددی برای خلاصه سازی خودکار متن ارائه شده است. در سالهای اخیر نیز تحقیقات متعددی در داخل و خارج کشور برای تولید سیستم های خلاصه ساز انجام گرفته است.

(داستانی داکدره، مریم و فاطمه احمدی آبکناری، ۱۳۹۵) مقاله ای تحت عنوان خلاصه سازی چند سندی با استفاده از متن کاوی و راهکار گراف های رویداد را ارائه داده اند. در این مقاله راهکاری جدید مبتنی بر گراف رویداد به منظور بازیابی اطلاعات و خلاصه سازی چند سندی معرفی می گردد. در ابتدا از روش فضای بردار وزنی برای تشخیص عبارات تکراری استفاده شده و سپس میزان شباهت نمونه های خبری درقالب اسنیپت های خبری از پیکره متن اصلی و پرس جو با استفاده از ضریب تشابه دوبردار محاسبه می گردد. سپس با استفاده از گراف رویداد، یک مدل نمایش سند مبتنی بر رویداد برای معناشناسی رویدادهای سطح جمله محاسبه می گردد که بر اساس آن اطلاعات مرتبط با رویدادهای توصیف شده در متن فیلتر شده و بازسازی می شود. در این روش با استفاده از هسته گراف ضرب تنسور و کونرمال، شباهت بین پرسوجوها و سندها اندازه گیری می شود. با توجه به کامل نبودن مدل های موجود، راهکار معرفی شده در این مقاله با تکیه بر گراف رویداد شباهت بین پرسوجوها و سندها با تفکیک هم رخدادی رئوس غیرمتناظر با استفاده از روش استخراج روابط معنایی موجود درمتن و تکنیک های برچسب زنی معنایی لغات، اندازه گیری می شود و همچنین روابط زمانی بین آنها نیز تعیین می گردد. در گام بعد اسناد بر اساس نمرات شباهت بدست آمده رتبه بندی شده. نتایج ارزیابی چهار روش فوق دلالت بر افزایش چشمگیرصحت و دقت مدل پیشنهادی این مقاله در مقایسه با مدلهای فضای بردار وزنی، گراف ضرب کونرمال و گراف ضرب تنسور بر روی مجموعه های آزمایشی رویدادگرای خبری دارد. (ولی ها، شهرزاد؛ بهروز معصومی و اسماعیل زینالی، ۱۳۹۵) پژوهشی تحت عنوان خلاصه سازی استخراجی متون با استفاده از الگوریتم بهینه سازی ازدحام جوجه ها انجام داده اند. در این مقاله محققان روی استخراج جملات، که یک نوع از خلاصه سازی است، تمرکز کرده اند. در اسناد با حجم زیاد مساله خلاصه سازی استخراجی به صورت یک مساله NP-Complete مطرح می شود. الگوریتم های فرا اکتشافی مبتنی بر هوش جمعی در حل این گونه مسایل توانا هستند. در این مقاله، روشی بر پایه الگوریتم فرااکتشافی مبتنی بر هوش جمعی به نام الگوریتم بهینه سازی ازدحام جوجه ها است که در آن از رفتار غذایابی گروهی خروس ها، مرغ ها و جوجه ها برای یافتن جملات خلاصه الهام گرفته شده است. روش های پیشنهادی در این مقاله بر روی مجموعه اسناد استاندارد DUC 2002 آزمایش شده است و توسط ابزار ROUGE مورد ارزیابی قرار گرفته اند. نتایج بدست آمده از آزمایش های انجام گرفته در مقایسه با سایر روش ها گویای عملکرد بهتر روش پیشنهادی نسبت به سایر روش هاست.

(شریفی نژاد, میلاد و فرشید کی نیا، ۱۳۹۵) بهبود عملکرد خلاصه سازی استخراجی متن با استفاده از روش بهینه سازی غذایابی نهنگ ها را ارائه داده اند. در این مقاله روشی برای بهبود عملکرد خلاصه سازی استخراجی متن بر پایه روش بهینه سازی غذایابی نهنگ ها ارایه شده است که در آن از رفتار گروه نهنگ های کوهان دار در پیدا کردن غذا الهام گرفته شده است. روش پیشنهادی در این مقاله بر روی مجموعه اسناد استاندارد DUC ۲۰۰۲ آزمایش و توسط نرم افزار ارزیاب ROUGE مورد تحلیل قرار گرفته است. نتایج بدست آمده از آزمایش های انجام گرفته در مقایسه با سایر روش ها نشان می دهد که این روش عمکرد بهتری نسبت به سایر روش ها دارد. (فتاح زاده، شهره و زهرا رضایی، ۱۳۹۶) روش های خلاصه سازی متون را مورد بررسی قرار داده اند. در این مقاله، محققین تکنیک و روش های خلاصه سازی متن را از لحاظ سرعت اجرایی و پیچیدگی مورد بحث و مقایسه قرار داده اند. (بلوچیان, حسین و مریم نظری، ۱۳۹۷) بهبود و خلاصه سازی متن بر اساس الگوریتم هوش جمعی را ارائه کرده اند. این پژوهش با هدف بهبود و خلاصه سازی متن مبتنی بر خوشه بندی بر اساس الگوریتم هوش جمعی صورت گرفته است. در این پژوهش یک روش جدید ترکیبی با استفاده از دو الگوریتم TF-IDF در کنار خوشه بندی چند عامله PSO که بدنه اصلی آن را پوشش داده وجهت استخراج متن استفاده شده است. نتایج حاکی از آن بود که ارزیابی

الگوریتم ترکیبی از کارایی خوبی برخوردار بوده و الگویتم خوشه بند پیشنهادی بر روی مجموعه داده مورد آزمایش قرار گرفته است. همچنین این پژوهش شاهد بهبود چند درصدی نسبت به کارهای پیشین بوده است. (احمدی، بختیار و سمیه دهقان، ۱۳۹۷) مقاله خلاصه سازی موضوعی اخبار کانال های مختلف خبری تلگرام با استفاده از تکنیک خوشه بندی را ارائه نموده اند. در این مقاله یک سیستم خلاصه ساز موضوعی اخبار با استفاده از تکنیک خوشه بندی و شباهت یابی ارایه شده است که اخبار جمع آوری شده از کانال های مختلف خبری فارسی تلگرام را بطور اتوماتیک خوشه بندی و سپس خلاصه سازی می کند.

(رودباری مونجی, زینب و رضا طاولی، ۱۳۹۷) یک روش خلاصه سازی خودکار متون تک و چند سندی بر پایه روش گراف ارائه نموده اند. در این پروژه یک روش جدید خلاصه سازی مبتنی بر گراف پیاده سازی شده است. در این مقاله سعی بر آن بوده تا متون بررسی و خلاصه شوند .برای خلاصه کردن در ابتدا متون پیش پردازش شده،کلمات اضافی حذف و ریشه یابی انجام شده است و در نهایت جایگاه کلمات را مشخص می شود. برای یافتن ویژگی ها و به دست آوردن ماتریس کلمات، از روش TF-ISF استفاده شده و وزن دهی اجرا شده است. برای بدست آوردن جملات هم از گراف استفاده شده برای جملات بر اساس شباهت کلمات مشترکی که دارند امتیاز قائل شده اند و بر اساس آن امتیازات نود های گراف ترسیم شده،که نود هایی که بیشترین یال را دارا بودند همان جملات منتخب بوده اند. (قانع, ساره؛ قدرت سپیدنام و احسان جعفری، ۱۳۹۷) خلاصه سازی متون فارسی با استفاده از الگوریتم یادگیری عمیق را ارائه داده اند. هدف ما در این مقاله، ایجاد خلاصه سازی استخراجی از متون و انتخاب بهترین جملات درون متون برای ایجاد خلاصه بهتر با استفاده از یادگیری عمیق می باشد. این خلاصه سازی بر روی متون فارسی صورت می گیرد و پیش پردازش ها و محدودیت های خاصی را دارا می باشد. همچنین بررسی هایی انجام شده که بتوان بهترین حالت نمایش یک سند متن فارسی را به حالت عددی به دست آورد به طوری که معنای کلمات درون متون آنها حذف نشود. نتایج نشان می دهد که مدل درخت، تصمیم با میزان دقت، ۱۰۰ ٪ بهترین مدل اعمال شده است و مدل شبکه عصبی با دقت، ۹۸ ٪ درصد در رتبه دوم قرار دارد. (صفایی, علیرضا و محمدعلی جوادزاده، ۱۳۹۸) خلاصه سازی متون فارسی به روش استخراجی با استفاده از گراف را ارائه نموده اند. در این مقاله ضمن برشمردن روشها و مجموعه های داده آماده برای زبان فارسی، به کمک نظریه گراف روشی استخراجی برای خلاصه سازی متون فارسی پیشنهاد شده است. در این روش پس از واکشی متن از مجموعه داده، جملات تفکیک شده و هر جمله به عنوان یک گره از گراف در نظر گرفته میشود. در ادامه ضمن پیش پردازش روی متن، مقدار ویژگی هر یال و گره ها محاسبه شده و بر این اساس گره ها رتبه بندی می شوند. خلاصه متن از بین گره های با امتیاز بالاتر ارائه می شود .در پایان ضمن پیاده سازی روش ارائه شده در زبان جاوا بر اساس معیارهای دقت، صحت و F-Measure روش ارائه شده مورد ارزیابی قرار گرفت که نشان از عملکرد مناسب آن دارد. (یادگاری, الهام؛ هادی خسروی و محمدعلی عرب زاده، ۱۳۹۸) روشی جدید برای خلاصه سازی چکیده تک سند فارسی با استفاده از یادگیری عمیق ارائه داده اند. در این پژوهش تلاش شده با استفاده از روشهای جدید یادگیری عمیق، به توسعه ی مدلی مبتنی بر شبکه های عصبی بازگشتی جهت نگاشت دنباله به دنباله پرداخته شود. مدل پیاده سازی شده از روش چکیده و ابزار تنسورفلو یکی از ابزارهای یادگیری عمیق در بستر زبان برنامه نویسی پایتون استفاده کرده است. در یادگیری عمیق برای آموزش شبکه به تعداد داده های زیاد احتیاج است به همین علت جهت آموزش شبکه از مجموعه دادگانی شامل بیش از صدهزار سند خبری استفاده گردید. به منظور تطبیق داده ها با مدل از پردازش زبان طبیعی استفاده شد. نتایج پیاده سازی نشان دهنده نرخ یادآوری ۲/۱۹، دقت ۴۵/۱۰ و کیفیت خلاصه ۶/۱۳ است. (محمود یوسفی آذر و همکاران، ۲۰۱۷) خلاصه سازی متن را با استفاده از یادگیری عمیق بدون ناظر پیشنهاد داده اند. محققین یک روش خلاصه سازی تک سندی استخراجی مبتنی بر جستجو با استفاده از با استفاده از رمزگذار خودکار عمیق برای محاسبه فضای ویژگی از ورودی فرکانس دوره را مورد بررسی قرار داده اند. نویسندگان واژگان محلی و عمومی را مورد کاوش قرار داده و تاثیر اضافه کردن نویز تصادفی کوچکی را به ورودی فرکانس دوره رمزگذار خودکار عمیق بررسی نموده و اثر کلی چنین نویزی را بر روی رمزگذارهای خودکار عمیق طرح کرده اند. همچنین از معیار ارزیابی Rouge برای بررسی تمام آزمایشات استفاده کرده اند.

(راسمیتا روتری و همکاران، ۲۰۱۸) یک چارچوب تکاملی برای خلاصه سازی متون چند سندی با استفاده از روش جستجوی فاخته ارائه نموده اند. روش ارائه شده با دو روش خلاصه سازی الهام گرفته از طبیعت یعنی خلاصه سازی مبتنی بر

بهینه سازی ازدحام ذرات و خلاصه سازی مبتنی بر بهینه سازی ازدحام گربه مورد مقایسه قرار گرفته است. همچنین عملکرد این روش ها با استفاده از معیار ارزیابی Rouge ، تشابه درون جمله ای و معیار خوانایی برای تایید عدم افزونگی و انسجام مورد مقایسه قرار گرفته است. نتایج ارزیابی به وضوح برتری این روش را نسبت به سایر روش های مقایسه شده نشان داده است. (چیرانتانا مالیک و همکاران، ۲۰۱۹) خلاصه سازی متن مبتنی بر گراف را با استفاده از رتبه متنی ارائه داده اند. این روش با استفاده از تغییر محاسبه رتبه متنی بر اساس مفهوم رتبه صفحه در هر صفحه وب توسعه یافته است. روش پیشنهادی گرافی را با جملات به عنوان گره ایجاد می نماید و تشابه بین جملات را به عنوان وزن لبه مابین گراف ها در نظر می گیرد. تشابه فرکانس کسینوسی جملات معکوس تغییر یافته برای دادن وزن به کلمات مختلف در جمله مورد استفاده قرار می گیرد حال آنکه در روش سنتی، تشابه کسینوسی با کلمات یکسان رفتار می نماید و کلمات وزن یکسانی دارند. گراف به صورت پراکنده و مجزا در خوشه های مختلف تقسیم می شود با این پیش فرض که جملات داخل هر خوشه با یکدیگر یکسان بوده و جملات در خوشه های مختلف تشابهی با هم ندارند. ارزیابی عملکرد با معیارهای مختلف نشان از اثربخشی این روش دارد. (موداسیر مد و همکاران، ۲۰۲۰) خلاصه سازی متون را با استفاده از جاسازی کلمات ارائه نموده اند. خلاصه سازی خودکار متن اساسا یک سند طولانی را به یک سند کوتاهتر با حفظ محتوای اطلاعاتی و همان مفهوم مختصر می نماید. اینکار یک راه حل بالقوه برای جلوگیری از انباشت اطلاعات می باشد.

چندین روش خلاصه سازی خودکار در ادبیات تحقیق وجود دارند که قادر به تولید خلاصه های با کیفیت هستند اما در این پژوهش ها، به معنی و مفهوم متن توجهی نشده است. در این تحقیق، مفهوم متن به عنوان یک ویژگی اساسی برای خلاصه سازی متون در نظر گرفته شده است. یک خلاصه ساز خودکار با استفاده از توزیع مدل معنایی پیشنهاد شده است که این مدل معنایی خلاصه هایی با کیفیت بالا تولید می نماید. نتایج ارزیابی با ابزار ROUGE نشان از کارآیی این روش دارد.

معرفی الگوریتم بهینه سازی فاخته

الگوریتم بهینه سازی فاخته یکی از الگوریتم‌های توسعه داده شده برای حل مسائل بهینه‌سازی غیرخطی و مسائل بهینه سازی پیوسته محسوب می‌شود[۱]. این الگوریتم، از زندگی خانواده‌ای از پرندگان به نام فاخته الهام گرفته شده است. الگوریتم بهینه سازی فاخته براساس شیوه زندگی بهینه و ویژگی‌های جالب این گونه، نظیر تخم‌گذاری و تولید مثل آن‌ها ساخته شده است. فاخته‌های بالغ و تخم‌های فاخته، جمعیت اولیه الگوریتم بهینه سازی فاخته را تشکیل می‌دهند. فاخته‌های بالغ در لانه پرندگان دیگر تخم‌گذاری می‌کنند. در صورتی که تخم‌های فاخته توسط پرندگان میزبان بالغ شناسایی نشوند و از بین نروند، رشد کرده و به فاخته‌های بالغ تبدیل خواهند شد. فاخته‌های بالغ تحت تأثیر ویژگی‌های محیطی و به امید یافتن محیط بهینه برای زندگی و تولید مثل، به صورت گروهی مهاجرت می‌کنند. در این الگوریتم، محیط بهینه همان بهینه سراسری در تابع هدف مسأله بهینه‌سازی خواهد بود. این الگوریتم تاکنون در سناریوهای بهینه‌سازی مختلف و کاربردهای جهان واقعی، عملکرد خوبی از خود نشان داده است. ویژگی شاخص این الگوریتم، شبیه‌سازی مفهوم بقا، مهاجرت برای یافتن منابع غذایی و انتخاب محیط بهینه برای زندگی است. جمعیت الگوریتم بهینه سازی فاخته را فاخته‌های بالغ و تخم‌های فاخته تشکیل می‌دهند.

تحلیل احساسات

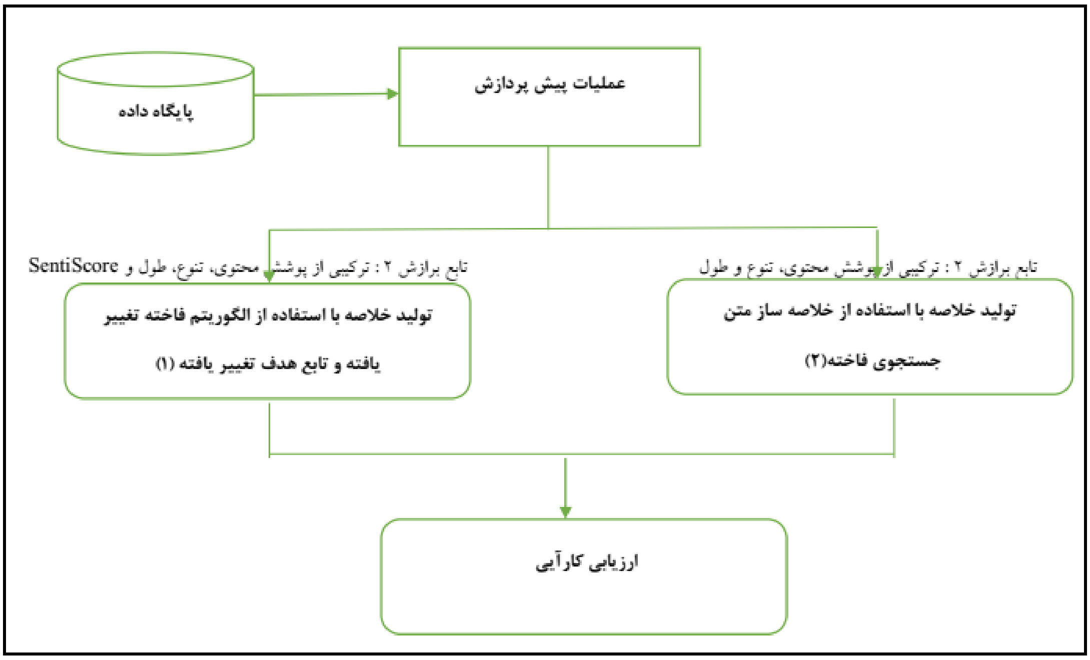
کلمه احساس برای بیان احساس قطعی برای یک موضوع یا شخص خاص استفاده می شود که تاثیر خود را حتی در نوشتن متن یک وب سایت، بلاگ و یا انجمن ها گذاشته باشد[۲]. تحلیل احساسات یک فرآیند ضروری در افکارسنجی (استخراج افکار) است که نقش بسیار مهمی را در بازیابی اطلاعات، جمع بندی و خلاصه سازی اسناد و زبان شناسی ایفاء می نماید[۳]. تحلیل احساسات می تواند در هر سطحی بخصوص در سطح سند، در سطح جمله و در سطح کلمه انجام گیرد.
SentiWordNet یک منبع و فرهنگ لغت برای افکار سنجی و داده کاوی اندیشه و نظرات می باشد. این فرهنگ لغت به هر synset از WordNet (شبکه لغات) سه نمره احساس را اختصاص می دهد که این سه نمره عبارتند از:

- مثبت بودن یا positivity
- منفی بودن یا negativity
- عینیت داشتن یا بی طرفی objectivity

در این پژوهش SentiWordNet برای محاسبه امتیاز دهی احساس در سطح جمله برای افکار سنجی و داده کاوی اندیشه و نظرات مورد استفاده قرار گرفته است. SentiWordNet به عنوان یک فرهنگ لغت منبع باز Open Source محسوب می شود که می تواند نقش مهمی در خلاصه سازی متون داشته باشد.

نمای کلی معماری سیستم پیشنهادی

شکل ۲ نمای کلی معماری سیستم پیشنهادی را نشان می دهد. معماری پیشنهادی، سیستم جریان داده را به صورت مرحله ای نشان می دهد. پس از پیش پردازش، داده برای خلاصه سازی به دو روش مختلف فراهم می گردد. نتایج خروجی هر دو نوع تکنیک خلاصه سازی در نهایت مورد مقایسه قرار می گیرد.



شکل ۲ – نمای کلی معماری سیستم پیشنهادی

پیش پردازش

روش های پیش پردازش شامل تقسیم بندی جملات، نشانه گذاری، حذف کلمات توقف و زیر وظیفه ریشه یابی لغات می باشد.

- **تقسیم بندی جمله:** تحت تقسیم بندی جمله، باقیمانده متن به صورت جملات اختصاصی متفاوت نمایش داده می شوند. سند ورودی نمونه به صورت $D = \{s_1,s_2,...,s_k\}$ است که s_i به عنوان جمله i ام و s_k به عنوان جمله آخر k ام در سند می باشد.
- **نشانه گذاری:** نشانه گذاری فرآیندی است که برای جدا کردن هر اصطلاح (کلمه) از جمله مورد استفاده قرار گرفته وتنها اصطلاحات یکتای جدا شده را به صورت $as = \{t_1, t_2,...,t_m\}$ نمایش می دهد که در آن t_k به مفهوم همه اصطلاحات یکتا و $k = 1, 2, 3... m$ می باشد.

- **حذف کلمات توقف** : لیستی از کلمات زبان انگلیسی است که به صورت مکرر مورد استفاده قرار گرفته و اهمیت آن از سایر کلمات کمتر می باشد. بنابراین پیش از شروع فرآیند خلاصه سازی، این کلمات از بدنه جملات و مستندات حذف می شوند.
- **ریشه یابی لغات:** این فرایند برای پیدا کردن ریشه لغات در شکل پایه مورد استفاده قرار می گیرد. این الگوریتم طبق یک سری قاعده‌ی منظم (مثلاً حذف حرف s در آخر کلماتِ جمع) می‌تواند ریشه‌ی کلمات را با دقتِ خوبی به دست آورد. ریشه یابی کمک می‌کند که کلمات را به صورت ریشه کلمه یا حالت پایه استانداردسازی نماییم . خروجی متن نهایی پس از ریشه یابی(که معمولاً روش های مشخص و از مجموعه قواعد ثابتی دارد) لزوماً کلمات با معنا و موجود در لغت نامه نخواهد بود اما پیشوندها و پسوندها حذف شده و در نهایت ساده ترین حالت کلمه و ریشه آن، به عنوان خروجی خواهد بود.

الگوریتم تغییر یافته فاخته یا فاخته گوسی

الگوریتم فاخته تغییر یافته که توسط ژنگ و همکاران [۴] ارائه شده، به عنوان الگوریتم فاخته گوسی نام گذاری شده است. در الگوریتم فاخته گوسی، توزیع لوی با توزیع گوسی جایگزین شده است که این توزیع گوسی منجر به بهبود همگرایی و دقت جستجوی فاخته می شود. به همین دلیل معادله ۱، برای جایگزینی معادله ۲ استفاده شده است:

$$\sigma_s = \sigma_0 \exp(-\mu_k) \qquad (۱)$$

که در این معادله σ_0 و μ به عنوان ثوابت و k به نسل اشاره دارد.

$$X_i^{(t+1)} = x_i t + \alpha \oplus L\acute{e}vy(\gamma) \qquad (۲)$$

در معادله ۲، α به معنی اندازه گام بوده و بایستی با اندازه مساله بهینه سازی مرتبط باشد. به صورت نرمال، از مقدار $\alpha=1$ استفاده می شود. ضرب $\oplus$ گام ورودی در طی ضرب و $L\acute{e}vy(\gamma)$ به مفهوم توزیع لوی می باشد. برای تولید تخم جدید $X^{(t+1)}$ برای فاخته i، معادله ۳ استفاده می شود:

$$X_i^{(t+1)} = x_i t + \alpha \oplus \sigma_s \qquad (۳)$$

تابع برازش

برای دستیابی به هدف خلاصه سازی متن، بایستی تعدادی از پارامترها نظیر پوشش محتوی، تنوع، خوانایی و طول مورد بحث و بررسی قرار گیرد[۸] . تابع پوشش محتوی ($f_{cov}(s)$) نشان می دهد که چه اندازه از کل سند در خلاصه نهایی قرار گرفته است. تابع پوشش محتوی ($f_{cov}(s)$) می تواند با معادله ۴ برای ارزیابی اهمیت جمله نمایش داده شود.

$$f_{cov}(s) = sim(s_i, O) \quad i = 1, 2, \ldots, n \qquad (۴)$$

که در این معادله O نقطه مرکزی بدنه سند مورد استفاده می باشد. مقدار بالای تابع ($f_{cov}(s)$) به مفهوم پوشش بالای متنی می باشد. تابع دیگر تابع تنوع ($f_{div}(s)$) می باشد که برای حداقل سازی همپوشانی جملات درون متن خلاصه با کاهش افزونگی می باشد. تابع تنوع ($f_{div}(s)$) می تواند با معادله زیر نمایش داده شود.

$$f_{div}(s) = 1 - sim(s_i, s_j) \quad i \neq j = 1, 2 \ldots n \qquad (۵)$$

مقدار بیشتر $f_{cov}(s)$ نشانگر شانس بیشتر جایگیری جمله داخل گروه می باشد. طول خلاصه نیز پارامتر بسیار مهمی است که با تابع $f_{len}(s)$ مورد محاسبه قرار می گیرد.

$$f_{len}(s) = sim(s_i, s_j) \quad i \neq j = 1, 2 \ldots n \qquad (۶)$$

سودمندی بالای $f_{len}(s)$ به مفهوم خوانایی بیشتر می باشد.

این سه تابع با یکدیگر برای شناسایی تخم های لقاح یافته (بارور شده) در معادله ۷ ترکیب شده است.

$$f(s) = f_{cov}(s) + f_{div}(s) + f_{len}(s) \qquad (۷)$$

در این تحقیق، سعی شده است از الگوریتم جستجوی فاخته برای خلاصه سازی متن استفاده شود. در استخراج روش خلاصه سازی متن، فرایند انتخاب جمله وظیفه بسیار حساسی می باشد. بنابراین برای انجام این وظیفه حساس، احساس جمله می تواند نقش بسیار مهمی را بازی نماید. برای محاسبه نمره احساس جمله، از sentiWordNet با استفاده از معادله ۸ استفاده شده است.

$$fsenti\ score(i) = \sum_{k=1}^{n} Senti(tk) \qquad (۸)$$

که در این معادله S_i در واقع i امین جمله و t_k در واقع k امین اصطلاح (کلمه) جمله می باشد. معادله ۷ برای فرمولاسیون تابع برازش به معادله ۹ تغییر یافته است.

$$f\ (s) = f_{cov}(s) + f_{div}\ (s) + f_{len}\ (s) + f_{Senti\ Score}\ (s) \qquad (۹)$$

مراحل الگوریتم خلاصه ساز پیشنهادی

مراحل الگوریتم خلاصه سازی به صورت موارد ذیل می باشد:

مرحله ۱- فرض شود سند منبع D با جملات متعدد وجود دارد که $D = \{S_1, S_2, \ldots, S_i\}$ و $S_i = \{t_1, t_2, \ldots t_k\}$ اصطلاحات (کلمات) متعددی دارند.

مرحله ۲- پیش پردازش بیان شده در بالا به سند منبع D اعمال می شود.

مرحله ۳- اکنون، نمایش ورودی با محاسبه امتیاز مفید جمله مدیریت می شود. مجموع تناوب (تکرار) اصطلاحات به عنوان امتیاز مفید جمله شناخته می شود. مقادیر بالای امتیاز مفید جمله، بیانگر اهمیت بیشتر جمله می باشد. امتیاز مفید توسط معادله زیر محاسبه می شود:

$$IS_{ik} = TF_{ik} \times \log\ (n|n_k) \qquad (۱۰)$$

که در آن TF_{ik} تناوب (تکرار) اصطلاحات می باشد که نشانگر تعداد کلی اصطلاح t_k ارائه شده در جمله S_i می باشند. n_k مجموع کلی جملاتی است که در آن t_k به صورت اصطلاح ظاهر شده است. عبارت $\log\ (n|n_k)$ عموما تناوب (تکرار) جمله معکوس نامیده می شود و در حالت کلی برای محاسبه وزن اصطلاح t_k درون کل سند مورد استفاده قرار می گیرد.

مرحله ۴- برای انتخاب حداقل جملات مشابه، تشابه تقاطعی را با استفاده از معادله تشابه کسینوسی محاسبه نمایید.

$$Sim\ (s_i, s_j) = \frac{\sum_{k=1}^{m} w_{ik}\, w_{jk}}{\sqrt{\sum_{k=1}^{m} w_{ik}^2 \sum_{k=1}^{m} w_{jk}^2}} \quad i = j = 1, 2 \ldots n \qquad (۱۱)$$

مرحله ۵- برخی پارامترها که در سراسر الگوریتم کاربردی می باشند نظیر اندازه جمعیت، نرخ تخم همتراز، فاکتور گام (S_f) و توان گوسی σ_s را مقدار دهی اولیه نمایید.

مرحله ۶- هر لانه از فاخته را درون ناحیه جستجو با استفاده از وزن جمله نشان دهید.

مرحله ۷- مقدار برازش f_i را با استفاده از معادله شماره ۹ محاسبه نمایید.

مرحله ۸- جمعیت جدیدی از لانه با استفاده از توزیع گوسی توضیح داده شده در معادله ۳ بدست می آید.

مرحله ۹- مقدار برازش لانه جدید f_j محاسبه کرده و آن را با مقدار قبلی لانه مورد مقایسه قرار دهید. در صورتی که $f_j > f_i$ باشد، مقدار i را با راه حل جدید جایگزین نمایید.

مرحله ۱۰- برای انتخاب بدترین لانه ها، مقدار $Pa \in (0, 1)$ را انتخاب نمایید. سپس لانه های انتخابی حذف شده و یا دور ریخته می شوند و مکان جدیدی در فضای جستجوی مشخص شده، ساخته می شود.

مرحله ۱۱- مقدار برازش برای لانه جدید به دست آمده از آخرین مرحله محاسبه می شود.

مرحله ۱۲- بهترین عملکرد لانه را از جمعیت فعلی انتخاب کنید. اگر عملکرد لانه فعلی بهتر از لانه قبلی است، آن را جایگزین نمایید.

مرحله ۱۳- اگر معیار توقف راضی کننده نیست، به مرحله ۷ بروید.

مرحله ۱۴- جملات را از اسناد ورودی با در نظر گرفتن آستانه مولد خلاصه، انتخاب نمایید.

آزمایشات و تحلیل نتایج

در این بخش، آزمایشاتی جهت تست و ارزیابی سیستم پیشنهادی انجام می گیرد. سپس خروجی خلاصه ساز با امتیاز داده شده با بعضی از خلاصه های تولید شده توسط سیستم های موجود با استفاده از مفاهیم کلیدی، اهمیت جمله [۷] و پایه UDC [۶] مورد ارزیابی قرار می گیرد.

ابزار ارزیابی Rouge

ابزار Rouge معروفترین ابزار برای ارزیابی در خلاصه سازی خودکار می باشد که البته از آن در دیگر کاربردهای پردازش زبان طبیعی و بازیابی اطلاعات هم استفاده شده است. Rouge مخفف جمله ی "Recall-Oriented Understudy for Gisting Evaluation" به معنای "ارزیابی مبتنی بر یادآوری برای خلاصه" می باشد. این ابزار شامل معیارهایی برای تعیین کیفیت خلاصه ها به صورت خودکار، از طریق مقایسه آنها با خلاصه های تولید شده توسط انسان (خلاصه های ایده آل) می باشد. این معیار ها تعداد واحدهایی که بین خلاصه های سیستمی و خلاصه های انسانی هم پوشانی دارند نظیر n تایی ها، رشته ی کلمات و جفت کلمات را محاسبه می نمایند. از جمله این معیار ها به ROUGE-N، ROUGE-L، ROUGE-W و ROUGE-S می توان اشاره کرد.

معیارهای Precision، Recall و F-measure در ROUGE

معیار Recall به صورت زیر محاسبه می‌گردد:

$$Recall = \frac{Number\_of\_Overlapping\_Words}{Total\_Words\_in\_reference\_Summary}$$

برای ارزیابی با ابزارROUGE ، امتیازات داده شده توسط ROUGE و امتیازات داده شده توسط انسان را برای تعدادی خلاصه های سیستمی با هم مقایسه می شود. یک سیستم مناسب باید به خلاصه های خوب امتیاز بالا و به خلاصه های بد امتیاز پایین دهد. با استفاده از داده های DUC ، ضریب همبستگی لحظه ای پیرسون، ضریب هبستگی درجه ای اسپیرمن و ضریب همبستگی کندال بین میانگین امتیازات داده شده توسط ROUGE و امتیازات داده شده توسط انسان (امتیازی که به میزان پوشش دادن مطالب توسط خلاصه ها به آنها داده شده است) برای خلاصه های سیستمی محاسبه می گردد. همچنین به منظور ارزیابی تاثیر ریشه یابی و حذف stopword ها، آزمایشی ترتیب داده می شود. این ابزار معیارهایی برای ارزیابی خلاصه‌سازی خودکار متن‌ها، مثل ترجمه ماشین را دارد. با مقایسه‌ی خلاصه‌های استخراجی (خلاصه‌های تولیدشده به صورت خودکار) و خلاصه‌های چکیده‌ای (خلاصه‌های تولیدشده توسط انسان) ارزیابی انجام می‌شود. مثالی برای خلاصه استخراجی و خلاصه‌ی چکیده‌ای را در زیر می‌بینیم:

خلاصه‌ی استخراجی:

The cat was found under the bed

خلاصه‌ی چکیده‌ای:

The cat was under the bed

اگر هر کلمه را به تنهایی در نظر بگیریم، در این صورت کلمات مشترک بین خلاصه‌ی استخراجی و خلاصه‌ی چکیده‌ای، ۶ کلمه می‌باشد. اما این کلمات مشترک به تنهایی معیار ارزیابی نمی‌باشد پس بهتر است از معیارهایی مثل Precision و Recallاستفاده کنیم.

محاسبه‌ی Recall برای مثال مطرح شده:

$$\text{Recall} = \frac{6}{6} = 1.0$$

معیار Precision به صورت زیر محاسبه می‌گردد:

$$\text{Precision} = \frac{\text{Number\_of\_Overlapping\_Words}}{\text{Total\_Words\_in\_System\_Summary}}$$

محاسبه‌ی Precision برای مثال مطرح شده:

$$\text{Precision} = \frac{6}{7} = 0.86$$

حال اگر برای همین مثال خلاصه‌ی استخراجی به صورت زیر باشد

The tiny little cat was found under the big funny bed

دقت به صورت زیر محاسبه می‌شود:

$$\text{Precision} = \frac{6}{11} = 0.55$$

از ترکیب Precision و Recall نیز معیار F-measures شود که رابطه آن به صورت زیر می باشد:

F-measures= 2*Recall* Precision/ (Recall + Precision)

داده های استاندارد DUC

کنفرانس DUC از سال ۲۰۰۱ زیر نظر NIST شروع به انتشار داده های مورد نیاز برای خلاصه سازی متون کرده است و تا کنون ۷ مجموعه از داده ها را تحت عنوان DUC2001 تا DUC2007 ارائه نموده است. هر کدام از این مجموعه ها با اهداف خاصی انتشار یافته اند. هدف اصلی این کنفرانس کمک در ارزیابی روش های خلاصه سازی خودکار متون و بررسی روش های ارزیاب خلاصه سازی می باشد. مجموعه داده های DUC2001 تا DUC2004 برای خلاصه سازی تک سندی و چند سندی تولید شده اند. مجموعه داده های DUC2005 تا DUC2007 هم فقط برای خلاصه سازی چند سندی تولید شده اند. با توجه به اینکه مجموعه داده DUC2007 آخرین مجموعه از این داده ها و کاملترین آنها می باشد، در حال حاضر اکثر مقالات این مجموعه مورد ارجاع قرار می گیرد. داده های DUC2007 در مجموع شامل ۴۵ موضوع بوده که هر کدام شامل ۲۵ سند می باشد. ۱۰ نفر از اعضای NIST وظیفه نوشتن خلاصه های دستی برای این مجموعه را بر عهده داشته اند به طوری که برای هر موضوع ۴ نفر به صورت تصادفی انتخاب شده و خلاصه های چکیده ای تولید کرده اند.

یافته های تحقیق

الگوریتم پیشنهادی این تحقیق بر روی ۱۰ سند متفاوت جمع آوری شده از مجموعه داده های DUC2007، مورد ارزیابی قرار گرفته است. داده های DUC2007 در مجموع شامل ۴۵ موضوع بوده که هر کدام شامل ۲۵ سند می باشد. این مجموعه داده، یک مجموعه داده منبع باز می باشد. بسته ROUGE-1.5.5 برای ارزیابی خلاصه متن با معیارهای ارزیابی مورد استفاده قرار گرفته است. جدول شماره ۱ و ۲ میانگین نتایج آزمایشات را نشان می دهد که خلاصه مورد انتظار، بدون حذف کلمات توقف و با حذف کلمات توقف به ترتیب نشان می دهد. نتایج مدل مبتنی بر جستجوی فاخته برای خلاصه سازی متن از روی مقاله [۸] برای ROUGE-1 و ROUGE-2 به ترتیب ۴۳,۱۱ و ۱۳,۹۸ می باشد.

جدول ۱ - میانگین recall, precision وf-measure سیستم پیشنهادی بدون حذف کلمات توقف

System	ROUGE-1			ROUGE-2		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Proposed system	49.67	47.86	48.7482	24.97	21.46	23.08

جدول ۲ - میانگین recall, precision وf-measure سیستم پیشنهادی با حذف کلمات توقف

System	ROUGE-1			ROUGE-2		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Proposed system	49.67	47.86	48.7482	24.97	21.46	23.08

همچنین سیستم پیشنهادی با سیستم پیشنهادی Kamal Sarkar مورد مقایسه قرار گرفت است. این سیستم خلاصه ها را با گرفتن ورودی از یک سند واحد تولید می کند .این متدولوژی مبتنی بر مفاهیم کلیدی و جملات مهم اسناد ورودی می باشد[۶].

سیستم پایه پیشنهادی در [۷] در مقایسه با کارهای قبلی بسیار قوی تر بوده است. مقایسه این سیستم در مقابل سیستم پیشنهادی این تحقیق در جدول ۳ نشان داده شده است.

جدول ۳- مقایسه سیستم پیشنهادی با سیستم های موجود

System	**ROUGE-1** F-measure		**ROUGE-2** F-measure	
	Score with Stop words	Score without Stop words	Score with Stop words	Score without Stop words
Proposed system	**48.7482**	**43.18**	**23.08**	**21.84**
[۲۶]Article	48.55	43.08	23.04	21.76
DUC baseline	47.51	41.82	22.40	21.38

نتیجه گیری و کارهای آینده

این پژوهش عمدتا سیستم استخراج خلاصه سازی متن تک سندی را با استفاده از الگوریتم جستجوی فاخته و ترکیب آن با تحلیل احساسات به عنوان یک تکنیک نو ظهور نشان می دهد. الگوریتم اصلاح شده فاخته در این تحقیق از توزیع گوسی به جای توزیع لوی استفاده کرده و تابع هدف را با استفاده از امتیاز احساسی فرموله می نماید. این سیستم خلاصه سازی در مقایسه با سایر سیستم های خلاصه سازی از لحاظ امتیاز ROUGE در معیارهای recall, precision وf-measure نتایج بهتری را نشان می دهد. در این تحقیق، از امتیازات ROUGE-1 و ROUGE-2 برای آزمایشات استفاده شده است.

یکی از اشکالاتی که در اغلب روش های استخراج متن و خلاصه سازی متون وجود دارد، عدم خوانایی متن تولیدی می باشد. در اکثر سیستم ها، ترتیب جملات مبتنی بر یک هیوریستک ساده (به عنوان مثال مکان جمله در متن اصلی) می باشد که برای تولید یک متن منسجم کافی نمی باشد. تحقیقات آینده می تواند با هدف یافتن روش های جدید برای تولید متون منسجم تر انجام گیرد. ضمن اینکه تولید یک مدل سند عمومی و پایه می تواند به این امر کمک نماید.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[1] Rajabioun, Ramin. "Cuckoo optimization algorithm." Applied soft computing 11, no. 8 (2011): 5508-5518.

[2] Richmond, W. K. (1965). *Teachers and machines: an introduction to the theory and practice of programmed learning*. Collins.

[3] Shaikh, M. A. M., Prendinger, H., & Mitsuru, I. (2007, September). Assessing sentiment of text by semantic dependency and contextual valence analysis. In *International conference on affective computing and intelligent interaction* (pp. 191-202). Springer, Berlin, Heidelberg.

[4] Zheng, H., & Zhou, Y. (2012). A novel cuckoo search optimization algorithm based on Gauss distribution. *Journal of Computational Information Systems*, *8*(10), 4193-4200.

[5] Rautray, R., & Balabantaray, R. C. (2018). An evolutionary framework for multi document summarization using Cuckoo search approach: MDSCSA. *Applied computing and informatics*, *14*(2), 134-144.

[6] Sarkar, K. (2013). Automatic Single Document Text Summarization Using Key Concepts in Documents. *JIPS*, *9*(4), 602-620.

[7] Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference

[8] Mosa, Mohamed Atef, Arshad Syed Anwar, and Alaa Hamouda. "A survey of multiple types of text summarization based on swarm intelligence optimization techniques." (2018).

[9] Rouane, Oussama, Hacene Belhadef, and Mustapha Bouakkaz. "Combine clustering and frequent itemsets mining to enhance biomedical text summarization." Expert Systems with Applications 135 (2019): 362-373.