بهبود طبقه بندی مقالات در سایت‌های خبری آنلاین مبتنی بر رویکردهای فازی

مینا کاسب
گروه مهندسی کامپیوتر، واحد اردبیل، دانشگاه آزاد اسلامی، اردبیل، ایران

مسعود بکروی
گروه مهندسی کامپیوتر، واحد اردبیل، دانشگاه آزاد اسلامی، اردبیل، ایران

بابک نوری مقدم
گروه مهندسی کامپیوتر، واحد اردبیل، دانشگاه آزاد اسلامی، اردبیل، ایران

چکیده

طبقه‌بندی اخبار آنلاین در سایت‌های خبری به منظور سازماندهی محتوا و تسهیل دسترسی کاربران به مقالات مورد علاقه‌شان انجام می‌شود. این دسته‌بندی‌ها که معمولاً بر اساس موضوعات خبری مانند اقتصاد، سیاست، فرهنگ، ورزش و علم و فناوری شکل می‌گیرند، به کاربران کمک می‌کنند تا به‌راحتی اطلاعات لازم را پیدا کنند. همچنین، استفاده از روش‌های طبقه‌بندی مناسب در مقالات متنی اهمیت زیادی دارد، زیرا با ایجاد دسته‌بندی‌های منطقی و معنادار، دسترسی و جستجوی سریع‌تر اطلاعات را در موتورهای جستجو افزایش می‌دهد، و در نتیجه کاربران می‌توانند به‌طور مؤثرتری به مقالات مورد نظر خود دسترسی پیدا کنند. در این تحقیق یک رویکرد طبقه بند اخبار آنلاین بر اساس ترکیب معیار شباهت کسینوسی فازی و شبکه‌های عصبی عمیق، ارائه شده است. روش پیشنهادی در مقایسه با روش‌های قبلی که از معیارهای کلاسیک همچون Bag of Words، TF-IDF و CHI2 استفاده کرده‌اند، با استفاده از معیار شباهت کسینوسی فازی به نتایج بهتری دست یافته است. نتایج آزمایشات نشان می‌دهد روش پیشنهادی برای معیارهای ارزیابی شامل دقت، حساسیت، صحت و معیار F به ترتیب مقادیری برابر با ۹۰.۵، ۹۰.۵، ۹۳.۷۵ و ۹۱.۸۳ نسبت به سایر روش‌های طبقه بندی و روش‌های پیشین در پیش بینی اخبار آنلاین جدید، بهتر عمل کرده است.

واژگان کلیدی: سایت خبری، طبقه بندی اخبار آنلاین، معیار شباهت فازی، یادگیری عمیق.

مقدمه

توسعه مداوم فناوری اطلاعات و ارتباطات در دهه گذشته منجر به جمع آوری حجم عظیمی از محتوا در اینترنت شده است. در زمان‌های اخیر، وب سایت‌های خبری به منابع اصلی اخبار برای خوانندگان تبدیل شده اند، زیرا این سایت‌ها اخبار فوری و جدیدترین اخبار را در اختیار خوانندگان قرار می‌دهند. این محتوای تولید شده به درستی به افرادی که در جستجوی دسته بندی خبری خاصی هستند که به آن علاقه مند هستند ارائه نمی شود. محتوای شلوغ رسانه‌های اجتماعی بیشتر محتوای حمایت شده و انتخاب کمتری را در اختیار خوانندگان قرار می‌دهد(Katari & Myneni, 2020).

مقالات با توجه به گسترش و رشد مداوم محتوای دیجیتال، طبقه‌بندی خودکار متن به دسته‌های موجود یک روش مهم برای مدیریت و پردازش داده‌های متنی است که از منابع مختلفی مانند صفحات وب، ایمیل‌ها، و نشریات تولید می‌شود. بسیاری از افراد از این منابع آنلاین برای دسترسی به اطلاعات روزانه استفاده می‌کنند که اگرچه دسترسی به این اطلاعات را آسان‌تر کرده، اما سازماندهی و مدیریت آن‌ها را چالش‌برانگیز می‌سازد. طبقه‌بندی متن نقش مهمی در بازیابی اطلاعات، خلاصه‌سازی و پاسخگویی به پرسش‌ها ایفا می‌کند و به دلیل تنوع منابع و سبک‌های نوشتاری، ضروری است تا اطلاعات متنی به‌طور موثر سازماندهی شود تا در فرآیند تصمیم‌گیری مفید واقع شود (Ahmed & Ahmed, 2021; Jindal, Malhotra, & Jain, 2015).

طبقه‌بندی اخبار بر اساس پردازش زبان طبیعی (NLP[1]) و با استفاده از تکنیک‌های یادگیری ماشینی انجام می‌شود. این فرآیند با مجموعه‌ای از داده‌ها آغاز می‌شود که در آن تخصیص کلاس‌ها مشخص است و هدف اصلی آن پیش‌بینی صحیح دسته هر خبر است. در حالی که گروه‌بندی مقالات خبری به دلیل به‌روزرسانی‌های مداوم و متنوع بودن محتوا چالش‌برانگیز است، اما این دسته‌بندی به کاربران کمک می‌کند تا به سادگی و در زمان واقعی به مقالات مختلف دسترسی پیدا کنند و آن‌ها را مرور کنند(Thaipisutikul, Tuarob, Pongpaichet, Amornvatcharapong, & Shih, 2021).

با این حال، طبقه‌بندی اخبار متنی با مشکلاتی نیز مواجه است، از جمله ابهام در موضوعات و کمبود تنوع در محتوا که می‌تواند منجر به قرار گرفتن خبر در دسته‌بندی نادرست شود. تغییرات سریع در موضوعات خبری و نبود استانداردهای دقیق برای طبقه‌بندی نیز بر دقت این فرآیند تأثیر می‌گذارد. برای بهبود این چالش‌ها، استفاده از الگوریتم‌ها و تکنیک‌های هوش مصنوعی جهت تعیین دقیق‌تر موضوعات و استانداردسازی روش‌های طبقه‌بندی ضروری است. تکنولوژی‌های پیشرفته می‌توانند به تحلیل و دسته‌بندی دقیق‌تر اخبار کمک کنند و باعث بهبود دقت و کارآمدی این فرآیند شوند.

از این رو در این تحقیق یک رویکرد مبتنی بر منطق فازی به منظوربهبود دقت طبقه بندی مقالات در سایت‌های خبری آنلاین ارائه شده است. روش پیشنهادی از مجموعه داده‌های دسته بندی اخبار در مخزن داده Kaggle2 استفاده خواهد نمود. داده‌های مجموعه داده شامل متن‌های خبری و دسته‌های است که متون به آن‌ها اختصاص یافته اند. روش پیشنهادی با توجه به شباهت متن موجود در خبرها با استفاده از معیارهای شباهت با دسته‌های موجود اقدام به انتخاب ویژگی (انتخاب نشانه‌ها) در متن و طبقه بندی متون می‌نماید. در روش پیشنهادی از یک رویکرد شباهت فازی (Kumar, Singh, & Pais, 2019) به منظور بررسی تطابق متون خبری با دسته‌ها و آماده سازی برای طبقه بندی، استفاده خواهد شد. پس از سنجش شباهت و آمادگی برای طبقه بندی، در روش پیشنهادی از رویکردهای

[1] Natural Language Processing
[2] https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset

نظارت شده مانند درخت تصمیم، K نزدیک ترین همسایه، شبکه عصبی، رگرسیون لجستیک و طبقه بندی بیزین استفاده خواهد شد. روش پیشنهادی با آاده سازی داده‌های مناسب برای طبقه بندی سعی در افزایش دقت طبقه بندی اخبار متنی در سایت‌های خبری آنلان را دارد

روش‌های پیشین

اخبار زیادی به صورت آنلاین در دسترس است و همه آنها طبقه بندی نمی شوند. چند محقق در گذشته روی طبقه بندی اخبار کار کرده اند و بیشتر کارها بر روی شناسایی اخبار جعلی متمرکز شده است. بیشتر کارهای انجام شده در زمینه طبقه بندی اخبار بر روی یک مجموعه داده معیار انجام می‌شود. مشکل مجموعه داده معیار این است که مدل آموزش داده شده با آن در دنیای واقعی قابل اجرا نیست زیرا داده‌ها از قبل سازماندهی شده اند. این مطالعه از تکنیک‌های یادگیری ماشین برای دسته‌بندی مقالات خبری آنلاین استفاده کرد، زیرا این تکنیک‌ها از نظر نیازهای محاسباتی ارزان‌تر هستند و پیچیدگی کمتری دارند. در (Daud et al., 2023)، ماشین‌های بردار پشتیبان بهینه‌سازی شده با فراپارامتر (SVM[1]) را برای دسته‌بندی مقالات خبری بر اساس دسته‌بندی مربوطه پیشنهاد کرد. علاوه بر این، پنج تکنیک دیگر ML، نزول گرادیان تصادفی (SGD[2])، جنگل تصادفی (RF[3])، رگرسیون لجستیک (LR[4])، K-نزدیک‌ترین همسایه (KNN[5])، و بیزین ساده (NB[6])، برای مقایسه برای وظیفه طبقه‌بندی اخبار بهینه شدند. . نتایج نشان داد که مدل SVM بهینه شده بهتر از سایر مدل‌ها عمل می‌کند، در حالی که بدون بهینه‌سازی، عملکرد آن بدتر از سایر مدل‌های یادگیری ماشین بود.

در، (Ghadiri, Ghadiri, & Sheikholeslami, 2022) یک معماری جدید مبتنی بر استنتاج فازی و یادگیری عمیق برای طبقه‌بندی متنی اخبار پیشنهاد می‌کند که بر این محدودیت غلبه می‌کند. دنبال کردن حجم عظیمی از متن تولید شده در شبکه‌های اجتماعی و کانال‌های خبری، و به دست آوردن بینش‌های ارزشمند و قابل اعتماد از منابع مختلف اطلاعات، کاری خسته کننده است. این چالش در دوره‌های خاصی افزایش می‌یابد، به عنوان مثال، در یک رویداد همه‌گیر مانند Covid-19. هدف روش‌های طبقه‌بندی متن موجود، مانند طبقه‌بندی احساسات، کمک به افراد برای مقابله با این چالش با دسته‌بندی و خلاصه‌سازی محتوای متن است. با این حال، عدم قطعیت ذاتی متن تولید شده توسط کاربر، کارایی آنها را محدود می‌کند. روش پیشنهادی را با اعمال آن در مجموعه داده‌های متنی شناخته شده مرتبط با سلامت و مقایسه دقت با روش‌های پیشرفته ارزیابی کرده است. نتایج نشان می‌دهد که روش‌های همجوشی فازی پیشنهادی دقت را در مقایسه با مدل‌های پیش‌آموزش‌شده فردی افزایش می‌دهند. این مدل همچنین یک معماری رسا برای طبقه بندی اخبار سلامت ارائه می‌دهد.

در (Timmerman & Bronselaer, 2022)، یک روش اساسی جدید و خودکار برای نظارت بر صحت اخبار آنلاین پیشنهاد شده است. این رویکرد بر این واقعیت متکی است که مقالات خبری آنلاین اغلب پس از انتشار اولیه به روز می‌شوند و در نتیجه اشتباهات را نیز تصحیح می‌کنند. مشاهده خودکار تغییرات ایجاد شده در مقالات آنلاین و تشخیص خطاهای اصلاح شده ممکن است بینش مفیدی در مورد صحت اخبار ارائه دهد. پتانسیل مدل تشخیص خودکار تصحیح خطا ارائه شده با ساخت مدل‌های طبقه بندی نظارت شده برای تشخیص خطاهای عینی، ذهنی و زبانی به ترتیب در به روز رسانی اخبار آنلاین نشان داده شده است. این مدل‌ها با استفاده از مجموعه داده‌های به‌روزرسانی خبری بزرگ ساخته شده‌اند که طی دو سال متوالی برای شش روزنامه مختلف فلاندری آنلاین جمع‌آوری می‌شوند. سپس یک زیر مجموعه از ۲۱۱۲۹ تغییر با استفاده از ترکیبی از حاشیه نویسی خودکار و انسانی از طریق یک پلت فرم حاشیه نویسی آنلاین حاشیه نویسی می‌شود. در نهایت، ویژگی‌های دستی ساخته‌شده و جاسازی‌های متن به‌دست‌آمده توسط چهار مدل زبان مختلف (TF-IDF، word2vec، BERTje و SBERT) به سه الگوریتم یادگیری ماشین نظارت‌شده (رگرسیون لجستیک، ماشین‌های بردار

[1] Hyper-parameter-optimized support vector machines
[2] Stochastic Gradient Descent
[3] Random Forest
[4] Logistic Regression
[5] K-Nearest Neighbor
[6] Naïve Bayes

پشتیبانی و درخت‌های تصمیم) و عملکرد مدل‌های به‌دست‌آمده تغذیه می‌شوند. متعاقبا مورد ارزیابی قرار می‌گیرد. نتایج نشان می‌دهد که تفاوت‌های کوچکی در عملکرد بین الگوریتم‌های مختلف یادگیری و مدل‌های زبان وجود دارد.

در (Dogra et al., 2022)، چندین الگوریتم یادگیری ماشین و یادگیری عمیق مورد استفاده در طبقه بندی متن را با مزایا و کاستی‌های آنها خلاصه شده است. با پیشرفت سریع فناوری اطلاعات، اطلاعات آنلاین روز به روز به طور تصاعدی در حال رشد است، به ویژه در قالب اسناد متنی مانند رویدادهای خبری، گزارش‌های شرکت، بررسی محصولات، گزارش‌های مربوط به سهام، گزارش‌های پزشکی، توییت‌ها و غیره. . به همین دلیل، نظارت آنلاین و متن کاوی به یک کار برجسته تبدیل شده است. در طول دهه گذشته، تلاش‌های قابل توجهی در استخراج اسناد متنی با استفاده از مدل‌های یادگیری ماشینی و عمیق مانند نظارت، نیمه نظارت و بدون نظارت صورت گرفته است. حوزه بحث در این مقاله مدل‌های یادگیری پیشرفته برای متن کاوی یا حل مسائل چالش برانگیز NLP (پردازش زبان طبیعی) با استفاده از طبقه‌بندی متون را پوشش می‌دهد. این مقاله همچنین به خوانندگان کمک می‌کند تا وظایف فرعی مختلف، همراه با ادبیات قدیمی و جدید مورد نیاز در فرآیند طبقه بندی متن را درک کنند. ما معتقدیم که خوانندگان می‌توانند زمینه ای برای پیشرفت‌های بیشتر در زمینه طبقه بندی متن پیدا کنند یا تکنیک‌های جدید طبقه بندی متن را پیشنهاد کنند که در هر حوزه مورد علاقه خود قابل اجرا است.

در(Fayaz, Khan, Bilal, & Khan, 2022)، تکنیک‌های مختلف یادگیری ماشینی برای شناسایی و طبقه بندی اخبار جعلی استفاده شده است. با این حال، این رویکردها از نظر دقت محدود هستند. این مطالعه از طبقه‌بندی‌کننده جنگل تصادفی برای پیش‌بینی اخبار جعلی یا واقعی استفاده کرده است. برای این منظور، بیست و سه (۲۳) ویژگی متنی از مجموعه داده اخبار جعلی ISOT استخراج شده است. چهار بهترین تکنیک انتخاب ویژگی مانند chi2، تک متغیره، افزایش اطلاعات و اهمیت ویژگی برای انتخاب چهارده بهترین ویژگی از بیست و سه مورد استفاده می‌شود. مدل پیشنهادی و سایر تکنیک‌های معیار بر روی مجموعه داده‌های معیار با استفاده از بهترین ویژگی‌ها ارزیابی می‌شوند. یافته‌های تجربی نشان می‌دهد که مدل پیشنهادی از نظر دقت طبقه‌بندی از تکنیک‌های پیشرفته‌ی یادگیری ماشین مانند GBM، XGBoost و مدل رگرسیون Ada Boost بهتر عمل می‌کند.

در (Naredla & Adedoyin, 2022)، از مجموعه داده‌های دو مقاله ای منتشر شده در SEMEVAL-2019 استفاده شده است. تعبیه‌های کلمه ELMo که بر روی یک طبقه‌بندی جنگل تصادفی آموزش داده می‌شوند، دقت ۰.۸۸ دارد که بسیار بهتر از سایر مدل‌های هنری است. دقت مدل BERT و Word2vec برابر با ۰.۸۳ است. این تحقیق طول‌های مختلف ورودی جمله را برای BERT امتحان کرد و ثابت کرد که BERT می‌تواند متن را از کلمات محلی استخراج کند. با شواهد مدل‌های توصیف‌شده ML، این مطالعه به دولت‌ها، خوانندگان اخبار و سایر ذینفعان سیاسی کمک می‌کند تا هرگونه اخبار فراحزبی را شناسایی کنند، و همچنین به سیاست‌ها برای ردیابی، و تنظیم اطلاعات نادرست درباره احزاب سیاسی و رهبران آنها کمک می‌کند. در جدول ۱ مقایسه روش‌های پیشین نشان داده است.

جدول ۱. مقایسه روش‌های پیشین

روش	شماره مرجع	مزایا	معایب
ماشین‌های بردار پشتیبان(SVM)	(Daud et al., 2023)	-دقت بالا در طبقه‌بندی -عملکرد عالی بر روی داده‌های غیرخطی با استفاده از کرنل‌های مختلف - مقاوم به overfitting در داده‌های کوچک	-زمان آموزش طولانی -نیاز به انتخاب مناسب‌هایپرپارامترها -پیچیدگی در پیاده‌سازی و تفسیر نتایج
نزول گرادیان تصادفی (SGD)	(Daud et al., 2023)	-سادگی و سرعت در آموزش -امکان کار با حجم بالای داده‌ها -قابلیت انطباق با روش‌های مختلف یادگیری	-حساس به مقیاس داده‌ها -نتیجه نهایی وابسته به انتخاب‌هایپرپارامترها
جنگل تصادفی(RF)	(Daud et al., 2023)	-دقت بالا و مقاوم به overfitting- قابلیت پردازش ویژگی‌های متنوع -تفسیر ساده برای نتایج	-زمان آموزش طولانی و مصرف بالای حافظه - ممکن است نتایج غیرقابل تفسیر تولید کند

رگرسیون لجستیک (LR)	(Daud et al., 2023)	-ساده و قابل تفسیر -سرعت بالای آموزش -مناسب برای داده‌های خطی	-کارایی پایین در داده‌های غیرخطی -نیاز به پیش‌پردازش مناسب داده‌ها
K-نزدیک‌ترین همسایه (KNN)	(Daud et al., 2023)	-سادگی در پیاده‌سازی و تفسیر -دقت بالا با تنظیم مناسب K- نیازی به آموزش اولیه ندارد	-زمان مشاهده بالا -حساس به مقیاس ویژگی‌ها -عملکرد ضعیف در داده‌های بزرگ
بیزین ساده(NB)	(Daud et al., 2023)	-کارایی خوب در طبقه‌بندی متن‌های بزرگ -سرعت بالا در آموزش و پیش‌بینی -نیاز به تنظیم کم	-فرض استقلال ویژگی‌ها (که در عمل ممکن است نقض شود -(عملکرد ضعیف با داده‌های نامتعادل
معماری مبتنی بر استنتاج فازی و یادگیری عمیق	(Ghadiri et al., 2022)	-قابلیت مدل‌سازی عدم قطعیت -دقت بالاتر در مقایسه با مدل‌های سنتی -انطباق با داده‌های پیچیده	-پیچیدگی در پیاده‌سازی -نیاز به تنظیم دقیق معیارهای مختلف
روش نظارت بر صحت اخبار آنلاین	(Timmerman & Bronselaer, 2022)	-توانایی شناسایی و تصحیح خطاهای خبری - استفاده از داده‌های به‌روز برای بهبود دقت -ارائه بینش‌های مفید	-نیاز به پایش مداوم داده‌ها -ممکن است به متون با تغییرات جزئی دقت کمتری داشته باشد
تحلیل انواع الگوریتم‌های یادگیری ماشین و عمیق	(Dogra et al., 2022)	-پوشش گسترده الگوریتم‌های مختلف و بررسی مزایا و معایب آنها -کمک به شناسایی تکنیک‌های جدید و بهبود روش‌های موجود	-ممکن است به جزئیات تعدادی از روش‌ها نپردازد -نیاز به تحقیق بیشتر برای تدوین یک روش مناسب بر اساس نیازهای خاص
جنگل تصادفی (Random Forest)	(Fayaz et al., 2022)	-دقت بالا در پیش‌بینی اخبار جعلی -مقاوم در برابر overfittingبه علت استفاده از چند درخت تصمیم -قابلیت کار با داده‌های پیچیده و ویژگی‌های زیاد	-زمان آموزش طولانی -مصرف بالای حافظه - ممکن است پیچیدگی مدل منجر به تفسیر سخت شود
تقویت گرادیان (GBM)	(Fayaz et al., 2022)	-دقت بسیار بالا -قابلیت بهینه‌سازی و پردازش داده‌های پیچیده -انعطاف‌پذیری در انتخاب ویژگی‌ها	-زمان آموزش طولانی و مصرف بالای منابع - حساس به noise در داده‌ها -قابل تنظیم برای overfitting
XGBoost	(Fayaz et al., 2022)	-سرعت و عملکرد بالا در دقت -بهینه‌سازی شده برای سرعت و کارایی -قابلیت مدیریت داده‌های گمشده	-نیاز به تنظیم دقیق‌هایپرپارامترها -پیچیدگی در پیاده‌سازی و تفسیر نتایج -حساس بهoutlier ها
رگرسیون آدا (AdaBoost)	(Fayaz et al., 2022)	-دقت بالا با ترکیب چندین طبقه‌بند ساده -قابلیت تسهیل و بهینه‌سازی در یادگیری ماشین	-آسیب‌پذیری در برابر داده‌های noisy و outlier-نیاز به انتخاب مناسب تعداد طبقه‌بندهای ترکیبی
تعبیه‌های کلمهELMo	(Naredla & Adedoyin, 2022)	-قابلیت استخراج ویژگی‌های غنی از متن -دقت بالا در طبقه‌بندی با آموزش بر روی جنگل تصادفی	-نیاز به منابع محاسباتی زیاد -وابستگی به ساختار وابسته به معماری مدل

روش پیشنهادی

در این تحقیق به بررسی و ارائه یک رویکرد نوین برای طبقه‌بندی اخبار آنلاین در وب‌سایت‌های خبری می‌پردازد، به‌طوری‌که خبرها به چهار دسته اصلی جهانی، ورزشی، تجاری و تکنولوژی تقسیم‌بندی می‌شوند. برای این منظور، یک مجموعه داده بزرگ و جامع شامل بیش از ۱ میلیون مقاله خبری مورد استفاده قرار گرفته است. این مقالات از بیش از ۲۰۰۰ منبع خبری مختلف جمع‌آوری شده‌اند و فرآیند جمع‌آوری آن‌ها توسط پلتفرم ComeToMyHead، که یک موتور جستجوی اخبار دانشگاهی است و از جولای ۲۰۰۴ به فعالیت خود ادامه می‌دهد، انجام شده است. این پلتفرم به‌عنوان منبعی معتبر در ارائه اخبار و مقالات تحقیقاتی به شمار می‌رود و داده‌های گردآوری‌شده از آن به‌ویژه برای جامعه دانشگاهی ارزشمند بوده و برای اهداف تحقیقاتی در زمینه‌هایی مانند داده‌کاوی (خوشه‌بندی و طبقه‌بندی)، بازیابی اطلاعات (رتبه‌بندی و جستجو)، و فشرده‌سازی داده‌ها مورد استفاده قرار می‌گیرد. این تحقیق تلاش می‌کند با استفاده از این مجموعه داده گسترده، الگوریتم‌ها و تکنیک‌های مختلفی را برای بهبود دقت و کارایی طبقه‌بندی اخبار به‌کار گیرد، که

می‌تواند به ترویج روش‌های نوین در تحلیل داده‌های خبری و افزایش دسترسی به اطلاعات مفید در این حوزه کمک کند (Yousafzai et al., 2024).

در روش پیشنهادی، از یک رویکرد ترکیبی که شامل منطق فازی، معیار شباهت فازی و شبکه‌های عصبی عمیق است، برای تحلیل و دسته‌بندی اخبار آنلاین در وب‌سایت‌های خبری استفاده می‌شود. اساس این رویکرد بر تعیین و شناسایی تیترهای مرتبط با هر خبر است که به یکی از چهار دسته اصلی جهانی، ورزشی، تجاری و تکنولوژی تخصیص می‌یابند. این فرایند شامل پردازش و تحلیل مجموعه داده‌هایی است که حاوی تیترهای خبری به همراه توصیفاتی متنی درباره محتوا هستند. برای استخراج اطلاعات مفید و طبقه‌بندی صحیح این اخبار، نیاز به تکنیک‌های پیشرفته متن‌کاوی احساس می‌شود. در این فرآیند، منطق فازی به عنوان ابزاری برای مدیریت عدم قطعیت‌ها و ناپایداری‌های داده‌های متنی استفاده می‌شود و معیار شباهت فازی به تحلیل روابط و شباهت‌های بین تیترها و محتوا کمک می‌کند. همچنین، شبکه‌های عصبی عمیق به عنوان ابزار قدرتمند یادگیری ماشین، قابلیت بالایی در یادگیری الگوها و ویژگی‌های پیچیده داده‌های متنی دارند. به‌طور کلی، این رویکرد به دنبال بهینه‌سازی دقت و کارایی در فرایند طبقه‌بندی اخبار، با استفاده از ترکیب ابزارهای متنوع هوش مصنوعی و یادگیری ماشین است(Tandel, Jamadar, & Dudugu, 2019).

در روش‌های متن‌کاوی، سنجش شباهت بین متون یکی از مراحل کلیدی به شمار می‌رود که تأثیر زیادی بر نتایج طبقه‌بندی دارد. به‌ویژه در فرایندهای طبقه‌بندی متن، معیارهای شباهت نظیر معیار شباهت فازی در مراحل پیش‌پردازش و استخراج ویژگی‌ها به کار گرفته می‌شوند. در این مراحل، متون به فضای ویژگی‌های عددی تبدیل می‌شوند که می‌توان به‌واسطه آن‌ها شباهت میان متون مختلف را محاسبه کرد. محاسبه شباهت‌ها نه تنها برای تعیین کیفیات و ویژگی‌های متون ضروری است، بلکه در انتخاب و بهینه‌سازی ویژگی‌ها، مشخص کردن کلاس‌های مشابه و تسهیل فرآیند آموزش مدل نیز نقش حیاتی دارد. به‌عنوان مثال، استفاده از معیارهای شباهت می‌تواند منجر به گروه‌بندی متون مشابه شود، به‌گونه‌ای که الگوریتم‌های طبقه‌بندی در طول فرایند یادگیری، بر روی الگوهای خاص و مشابه تمرکز بیشتری داشته باشند. بنابراین، این معیارها به‌عنوان ابزارهایی اساسی در مراحل ابتدایی طبقه‌بندی متن عمل می‌کنند و در تعیین کیفیت و دقت مدل‌های آموزشی پیش از مرحله نهایی آموزش، تأثیرگذار هستند. این فرآیند نه تنها قابلیت‌های الگوریتم‌های یادگیری را بهبود می‌بخشد، بلکه باعث افزایش دقت در شناسایی و طبقه‌بندی دقیق گروه‌های مختلف متون می‌شود.

در این تحقیق، از معیار شباهت کسینوسی فازی به عنوان روشی نوآورانه برای استخراج و تحلیل داده‌های متنی بهره‌گیری شده است. این رویکرد به طور خاص بر مبنای منطق فازی و فاصله‌های محاسبه‌شده طراحی شده که ویژگی‌های منحصر به فردی را در مقایسه با روش‌های قبلی که از معیارهای فاصله‌ای سنتی مانند BoW[1]، TF-IDF[2] و CHI2[3] استفاده می‌کردند، در اختیار قرار می‌دهد(Sunagar et al., 2021). روش‌های پیشین به دلیل ناتوانی در درک تعاملات پیچیده معنایی بین کلمات و جملات، به دقت پایینی در محاسبه شباهت‌ها منجر می‌شدند و اغلب نتایج ناامید کننده‌ای را ارائه می‌کردند. در عوض، استفاده از معیار فاصله کسینوسی فازی به دلیل قابلیت‌های بالای آن در تجزیه و تحلیل و تشخیص شباهت‌های ظریف بین متون، به محققین این امکان را می‌دهد که نتایج دقیق‌تری ارائه کنند. این معیار به نحوی طراحی شده که می‌تواند تفاوت‌های ظریف و نوانس‌های معنایی را بهتر درک کند و در نهایت بر کیفیت و دقت دسته‌بندی متون تأثیر مثبت بگذارد. به همین دلیل، این تحقیق می‌تواند به عنوان یک پیشرفت مهم در حوزه تحلیل متون و داده‌کاوی به شمار آید و ابزارهای کارآمدتری را برای دسته‌بندی داده‌های متنی فراهم کند. اما در روش پیشنهادی، از معیارهای فاصله کسینوسی فازی استفاده می‌شود که به طور خاص برای محاسبه شباهت بین متون طراحی شده‌اند.

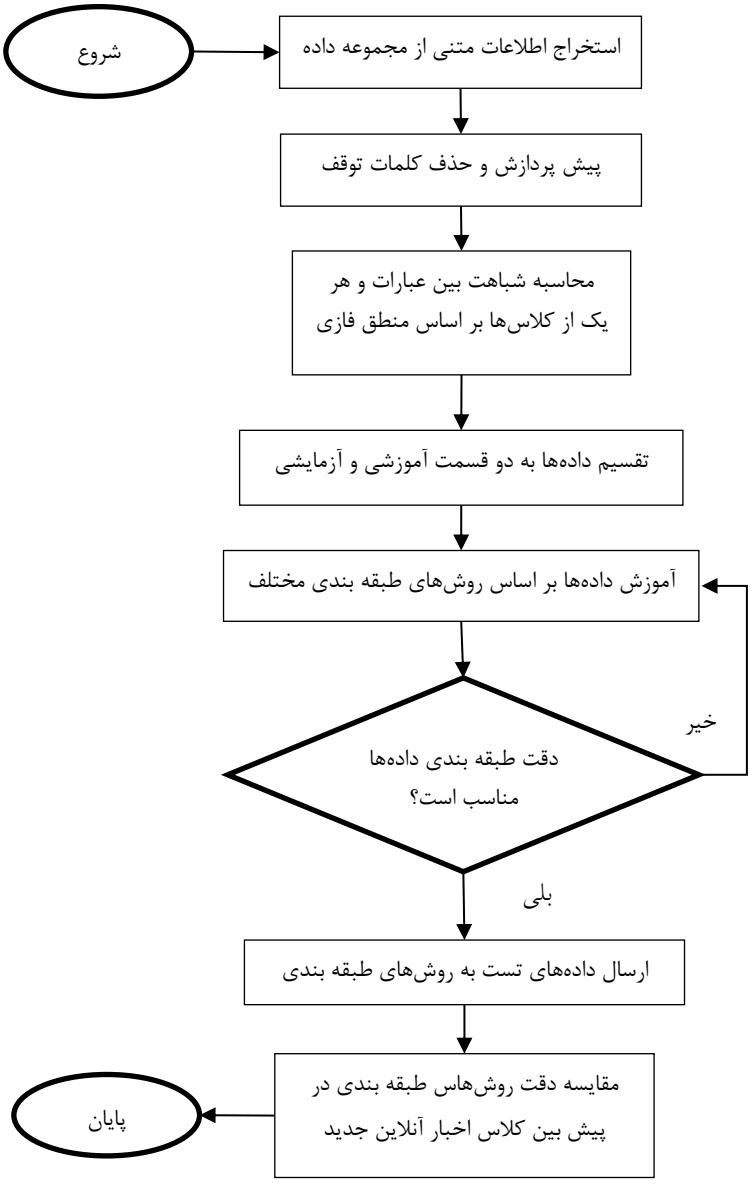
در رویکرد پیشنهادی، ابتدا مرحله پیش‌پردازش بر روی متون صورت می‌گیرد که هدف آن حذف کلمات بی‌معنی و غیرمؤثر است. این مرحله اهمیت زیادی دارد، زیرا کلمات غیرضروری می‌توانند باعث کاهش دقت تجزیه و تحلیل و برچسب‌گذاری شوند. پس از این مرحله، با استفاده از معیار شباهت کسینوسی فازی، میزان شباهت هر متن در مجموعه داده با متونی که در تمامی دسته‌ها وجود دارند، اندازه‌گیری می‌شود. این فرآیند به محققین این امکان را می‌دهد که صحت برچسب‌گذاری را بررسی کنند و برچسب هر یک از متون

[1] Bag of Words
[2] Term Frequency-Inverse Document Frequency
[3] Chi-Squared

آموزشی را بر اساس بیشترین شباهت به هر یک از کلاس‌ها تعیین نمایند. پس از برچسب‌گذاری، داده‌های برچسب‌گذاری‌شده به مرحله بعدی داده‌کاوی منتقل می‌شوند. در این مرحله، از روش‌های یادگیری نظارت‌شده جهت شناسایی و دسته‌بندی اخبار آنلاین استفاده می‌شود. به منظور ارزیابی و مقایسه کارایی روش پیشنهادی، چندین الگوریتم مختلف نظیر شبکه‌های عصبی عمیق(DNN[1])، درخت تصمیم(DT[2])، K– نزدیک ترین همسایه (KNN[3])، جنگل تصادفی (RF[4]) و بیزین ساده (NB[5]) به کار گرفته می‌شوند. هر یک از این الگوریتم‌ها به صورت مستقل برای دسته‌بندی اخبار آنلاین جدید ارزیابی می‌شوند که نتایج آن‌ها نه تنها به سنجش عملکرد هر الگوریتم کمک می‌کند بلکه امکان مقایسه بین آن‌ها را فراهم می‌سازد تا بهترین روش برای طبقه‌بندی اخبار آنلاین مشخص شود. این فرآیند، به محققین این قابلیت را می‌دهد که با توجه به نتایج به دست آمده، تکنیک‌های بهینه‌تری را برای تجزیه و تحلیل داده‌ها انتخاب و تنظیم کنند.

در نهایت، این رویکرد ترکیبی که به طور همزمان از تکنیک‌های متن‌کاوی و الگوریتم‌های یادگیری نظارت‌شده استفاده می‌کند، به منظور افزایش دقت و کارایی در طبقه‌بندی و پیش‌بینی اخبار آنلاین طراحی شده است. با ادغام این دو روش، قادر است ویژگی‌های معنایی و ساختاری متون را استخراج کرده و این اطلاعات را در فرایند یادگیری الگوریتم‌های طبقه‌بندی به کار گیرد. این ترکیب به محققان این امکان را می‌دهد که نسبت به روش‌های قبلی، به درک بهتری از داده‌های متنی برسند و به همین دلیل انتظار می‌رود که نتایج بهتری در زمینه شناسایی و دسته‌بندی اخبار آنلاین به دست آید. از آنجا که اطلاعات موجود در اخبار آنلاین به سرعت و به طور مداوم در حال تغییر است، این تحقیق به راستی می‌تواند سرعت واکنش به اخبار و صحت اطلاعات طبقه‌بندی‌شده را بهبود بخشد. در نتیجه، توسعه ابزارهای مؤثرتر و کارآمدتر در این حوزه نه تنها به طبقه‌بندی بهتر اخبار کمک خواهد کرد بلکه به تامین اطلاعات دقیق و به‌روز برای کاربران و پژوهشگران نیز یاری می‌رساند. این پیشرفت‌ها می‌توانند به شکل‌گیری مدل‌های جدیدی منجر شوند که در طول زمان به بهبود مستمر سیستم‌های اطلاعاتی و خبری کمک خواهند کرد. در شکل ۱ فلوچارت روش پیشنهادی ارائه شده است.

[1] Deep Neural Network
[2] Decision Tree
[3] K- Nearest Neighbors
[4] Random Forest
[5] Naïve Bayesian



شکل ۱. فلوچارت روش پیشنهادی

پیش‌پردازش متن

فرآیند پیش‌پردازش متن شامل نشانه‌گذاری و گسسته‌سازی اطلاعات، تقسیم جریان کاراکترها به واحدهای معنادار مانند کلمات و نشانه‌هاست. این مرحله با چالش‌هایی در تفکیک صحیح کلمات و عبارات به دلیل عدم نشانه‌گذاری دقیق در زبان‌های طبیعی مواجه است. همچنین، این فرآیند شامل شناسایی کلمات، حذف کلمات توقف و ریشه‌یابی است که به کاهش حجم متن و بهبود عملکرد سیستم‌های پردازش متن کمک می‌کند. حذف کلمات توقف به تمرکز بر کلمات اصلی منجر شده و دقت سیستم‌های پردازش زبان طبیعی را افزایش می‌دهد.

نرمال‌سازی متن

نرمال‌سازی متن شامل تمیز کردن و استانداردسازی متن برای بهبود کیفیت داده‌های ورودی است. این فرآیند شامل نشانه‌گذاری، تبدیل حروف بزرگ به کوچک، تصحیح اشتباهات املایی و حذف کلمات غیرمرتبط است. تکنیک‌های ریشه‌سازی و واژه‌سازی نیز به

تحلیل و تبدیل کلمات به شکل استاندارد خود کمک می‌کنند. این مراحل باعث می‌شوند که داده‌های متنی با دقت و کارایی بیشتری توسط مدل‌های NLP پردازش شوند.

برداری‌سازی متن

برداری‌سازی متن شامل تبدیل داده‌های غیر ساختاریافته مانند متن خام به داده‌های ساختاریافته عددی است. این فرآیند از روش‌هایی مانند "کیسه کلمات"، "TF-IDF"، و "بردارهای تعبیه شده" استفاده می‌کند تا متن‌ها را به نمایه‌های عددی تبدیل کرده و به الگوریتم‌های یادگیری ماشین کمک کند الگوها و روابط معنایی را شناسایی کنند. این فرآیند به مدل‌های یادگیری ماشین امکان می‌دهد تا از داده‌های متنی بهره‌برداری کرده و عملکرد بهتری داشته باشند.

بردارهای تعبیه شده

بردارهای تعبیه شده تکنیکی برای نمایش کلمات در فضای چندبعدی هستند که به درک روابط معنایی بین کلمات کمک می‌کنند. مدل‌هایی مانند Word2Vec، GloVe، و FastText به تحلیل دقیق‌تری از کلمات و ارتباطات بین آنها می‌پردازند. این تکنیک‌ها در کاربردهایی مانند ترجمه ماشینی و تحلیل احساسات به‌طور مؤثر استفاده می‌شوند و به الگوریتم‌های یادگیری ماشین کمک می‌کنند تا از اطلاعات معنایی بهینه‌تری استفاده کنند.

محاسبه شباهت

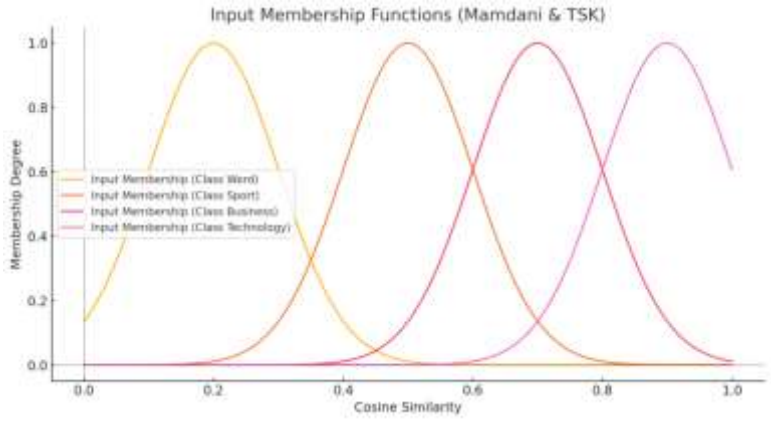
در این روش، از معیار شباهت کسینوسی فازی برای سنجش شباهت بین نیازمندی‌های نرم‌افزاری استفاده می‌شود. این معیار، اسناد متنی را به بردارهای عددی تبدیل کرده و شباهت بین آنها را محاسبه می‌کند. مقادیر حاصل از این فرآیند نشان‌دهنده میزان شباهت بین نیازمندی‌های جدید و موجود هستند که به تحلیل دقیق‌تر داده‌های متنی کمک می‌کند. این روش به سیستم‌های پردازش متن اجازه می‌دهد تا با دقت بیشتری نیازمندی‌های مختلف را طبقه‌بندی و تحلیل کنند.

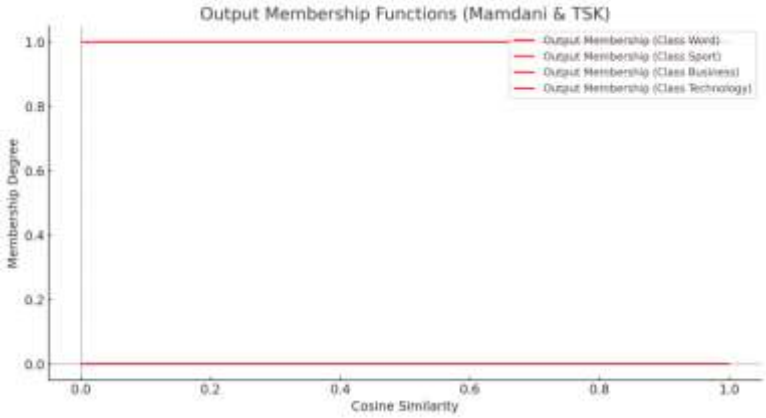
طبقه‌بندی فازی

طبقه‌بندی فازی از درجات مختلف شباهت برای تحلیل و دسته‌بندی داده‌های متنی استفاده می‌کند. برخلاف روش‌های سنتی که نمونه‌ها را به‌طور قطعی به یک دسته اختصاص می‌دهند، طبقه‌بندی فازی امکان تعلق یک نمونه به چندین دسته با درجات مختلف را فراهم می‌کند. این رویکرد در مواجهه با ابهامات متنی عملکرد بهتری دارد و به بهبود فرآیند تصمیم‌گیری در سیستم‌های پردازش زبان طبیعی کمک می‌کند.

نوع منطق فازی

طبقه‌بندی فازی از منطق فازی برای مدل‌سازی سیستم‌های حاوی ابهام و عدم قطعیت استفاده می‌کند. برخلاف منطق کلاسیک، منطق فازی از درجات عضویت برای تعیین تعلق نمونه‌ها به دسته‌ها بهره می‌برد. این ویژگی در تحلیل متن بسیار مفید است، زیرا متون اغلب پیچیده و دارای معانی چندگانه هستند. دو نوع رایج منطق فازی در طبقه‌بندی متن عبارتند از: منطق فازی Mamdani و Takagi-Sugeno-Kang (TSK) که در شکل ۲ قابل مشاهده است.





شکل ۲. نمونه ای از سیستم فازی ممدانی و TSK برای طبقه بندی متن

تابع عضویت برای طبقه‌بندی متن

در طبقه‌بندی متنی بر اساس منطق فازی، تابع عضویت نقش مهمی در تعیین میزان تعلق نمونه‌ها به دسته‌ها ایفا می‌کند. توابع عضویت مختلفی بسته به نیاز سیستم به کار گرفته می‌شوند. تابع عضویت گوسی به دلیل ساختار صاف و انعطاف‌پذیرش برای تحلیل داده‌های متنی پیچیده مناسب است، زیرا می‌تواند تغییرات جزئی در داده‌ها را به خوبی مدل کند.

نحوه تعلق نمونه‌ها به دسته‌ها

در منطق فازی، تعلق یک نمونه به یک دسته به صورت یک مقدار بین ۰ و ۱ بیان می‌شود، که نشان‌دهنده درجه شباهت نمونه به آن دسته است. این انعطاف‌پذیری به سیستم کمک می‌کند تا داده‌های مبهم را بهتر تحلیل کند. نمونه‌ها می‌توانند به‌طور نسبی به چندین دسته اختصاص داده شوند که این ویژگی برای مسائلی با مرزهای نامشخص میان دسته‌ها بسیار کاربردی است.

سنجش شباهت نمونه‌ها نسبت به سایر دسته‌های متنی

معیار شباهت کسینوسی یکی از روش‌های متداول برای محاسبه شباهت بین بردارهای ویژگی متون است. این معیار میزان شباهت دو نمونه را بین -۱ و ۱ بیان می‌کند و برای تحلیل دقیق شباهت بین متون بسیار مفید است. با استفاده از این معیار، می‌توان تعیین کرد که یک متن چقدر به دسته‌های مختلف نزدیک است.

ترکیب معیار شباهت کسینوسی با منطق فازی

در این روش، ابتدا شباهت کسینوسی میان متن ورودی و دسته‌ها محاسبه می‌شود و سپس این مقدار به تابع عضویت فازی اعمال می‌گردد. این تبدیل به تحلیل دقیق‌تر و انعطاف‌پذیرتر داده‌های متنی کمک می‌کند، زیرا نمونه‌ها به صورت فازی به دسته‌ها تعلق می‌گیرند و سیستم بر اساس این اطلاعات تصمیم‌گیری می‌کند. . اگر T_1 و T_2 اسناد متنی باشند، هر کدام به صورت بردارهای فازی V_{T_1} و V_{T_2} نمایش داده می‌شوند، که به در روابط ۱ و ۲ تعریف شده است:

$$V_{T_1} = \left(\mu_{w_1}^{T_1}, \mu_{w_2}^{T_1}, \dots \mu_{w_n}^{T_1}\right) \quad (۱)$$

$$V_{T_2} = \left(\mu_{w_1}^{T_2}, \mu_{w_2}^{T_2}, \dots \mu_{w_n}^{T_2}\right) \quad (۲)$$

که در آن $\mu_{w_i}^{T_1}$ و $\mu_{w_i}^{T_2}$ درجات عضویت کلمه w_i در اسناد T_1 و T_2 هستند. شباهت کسینوسی فازی بین دو سند T_1 و T_2 به شکل رابطه ۳ محاسبه می‌شود:

$$Sim_{cos}(T_1,T_2) = \frac{\sum_{i=1}^{n} \mu_{w_i}^{T_1} . \mu_{w_i}^{T_2}}{\sqrt{\sum_{i=1}^{n}\left(\mu_{w_i}^{T_1}\right)^2} . \sqrt{\sum_{i=1}^{n}\left(\mu_{w_i}^{T_2}\right)^2}} \quad (۳)$$

در این فرمول، $\sum_{i=1}^{n} \mu_{w_i}^{T_1} . \mu_{w_i}^{T_2}$ ضرب داخلی دو بردار فازی را نشان می‌دهد و $\sqrt{\sum_{i=1}^{n}\left(\mu_{w_i}^{T_1}\right)^2}$ و $\sqrt{\sum_{i=1}^{n}\left(\mu_{w_i}^{T_2}\right)^2}$ اندازه‌گیری‌های برداری هستند که نشان‌دهنده طول بردارهای فازی در فضا هستند.

سنجش شباهت فازی بین جملات

در این روش، ویژگی‌هایی مانند تعداد کلمات و شباهت معنایی جملات برای ارزیابی شباهت متنی استفاده می‌شود. معیار شباهت کسینوسی فازی با مدل‌سازی هر سند متنی به عنوان یک بردار فازی، امکان تحلیل دقیق‌تر شباهت‌ها و ارتباطات بین اسناد را فراهم می‌کند. این رویکرد به تحلیل‌های پیشرفته‌تر در پردازش زبان طبیعی کمک می‌کند.

روند تولید بردارهای تعبیه‌شده

ایجاد بردارهای تعبیه‌شده در پردازش زبان طبیعی شامل مراحل تولید، نرمال‌سازی و محاسبه شباهت کسینوسی فازی است. این بردارها به‌عنوان نمایشی عددی از متون عمل می‌کنند و روابط معنایی بین متون را نمایش می‌دهند. استفاده از این تکنیک‌ها، تحلیل دقیق‌تر و مؤثرتری از متون را امکان‌پذیر می‌کند.

قواعد تصمیم‌گیری فازی

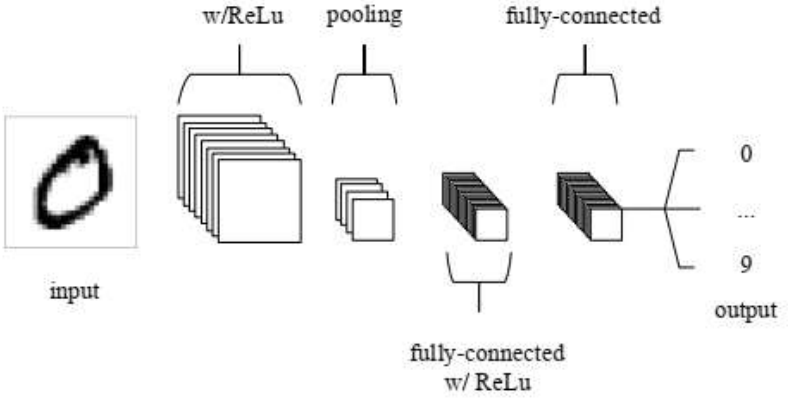
پس از تعیین درجات عضویت نمونه‌ها به دسته‌های مختلف، قواعد فازی برای تصمیم‌گیری نهایی به کار می‌روند. این قواعد به صورت قوانین شرطی تعریف می‌شوند که به سیستم کمک می‌کنند بر اساس درجات عضویت تصمیم‌گیری کند. این رویکرد انعطاف‌پذیری بالایی در تحلیل نمونه‌های پیچیده فراهم می‌کند.

نحوه تخصیص نمونه‌ها به دسته‌ها

تخصیص نمونه‌ها به دسته‌ها بر اساس درجه عضویت آن‌ها انجام می‌شود. هر نمونه به دسته‌ای اختصاص داده می‌شود که بیشترین درجه عضویت را دارد. این نوع تخصیص به سیستم امکان می‌دهد که نمونه‌ها را حتی در شرایط مرزهای نامشخص میان دسته‌ها به درستی دسته‌بندی کند.

شبکه‌های عصبی عمیق پیشنهادی

شبکه‌های عصبی عمیق از سه نوع لایه تشکیل شده‌اند. این‌ها لایه‌های کانولوشن، لایه‌های ترکیبی و لایه‌های کاملاً متصل هستند. هنگامی‌که این لایه‌ها روی‌هم قرار می‌گیرند، یک معماری شبکه‌های عصبی عمیق شکل می‌گیرد. یک معماری ساده‌شده شبکه‌های عصبی عمیق برای طبقه‌بندی در شکل ۳نشان داده شده است.



شکل ۳. معماری ساده شبکه عصبی عمیق متشکل از پنج لایه (O'Shea & Nash, 2015)

با توجه به شکل ۳ لایه‌های شبکه‌های عصبی عمیق به‌صورت زیر نشان داده شده است.

لایه ورودی

لایه ورودی در شبکه عصبی نقشی حیاتی در پردازش و بهینه‌سازی داده‌های تصویری دارد. هدف این لایه کاهش مقدار میانگین داده‌ها به صفر برای تسهیل پردازش اطلاعات توسط نورون‌هاست. پس از آن، داده‌ها در بازه [۰،۱] نرمال‌سازی می‌شوند تا از نوسانات بزرگ جلوگیری شده و سرعت هم‌گرایی الگوریتم یادگیری افزایش یابد. تکنیک‌هایی مانند سفید کردن داده‌ها وابستگی‌های غیرضروری را کاهش می‌دهند و بهبود عملکرد شبکه را تسهیل می‌کنند. همچنین، استفاده از تجزیه‌وتحلیل اجزای اصلی (PCA) برای کاهش ابعاد و تمرکز بر عوامل کلیدی انجام می‌شود که داده‌ها را فشرده‌تر کرده و قابلیت پردازش و تحلیل را بهبود می‌بخشد. در مجموع، این مراحل بهینه‌سازی، ویژگی‌های داده‌ها را شفاف و متمرکز می‌کند و عملکرد مدل در یادگیری را ارتقا می‌دهد(Du, 2018).

لایه کانولوشن (CONV)

لایه کانولوشنال در شبکه‌های عصبی عمیق نقش حیاتی در پردازش و تحلیل داده‌های ورودی، به‌خصوص تصاویر، دارد. این لایه با استفاده از هسته‌های کانولوشنال یا فیلترها، ویژگی‌های محلی را از داده‌ها استخراج می‌کند؛ هر نرون به گروهی از داده‌های همبسته توجه کرده و با عملیات ضرب و جمع، نتایج کانولوشنی تولید می‌کند. این فرآیند به استخراج ویژگی‌ها کمک کرده و به آن پیچیدگی قاعده‌گذاری می‌گویند. وزن‌های کانولوشنال به‌صورت یکسان در تمام نقاط تصویر اعمال می‌شوند که قابلیت شناسایی ویژگی‌های مشابه در نقاط مختلف را فراهم می‌آورد. پارامترهای تنظیم‌شده در این لایه شامل اندازه هسته، عمق (تعداد فیلترها)، اندازه گام و دیگر تنظیمات مربوط به فیلترها هستند. به این ترتیب، لایه کانولوشنال به عنوان ابزاری قوی در تحلیل داده‌ها به شبکه کمک می‌کند تا الگوها را به‌طور مؤثری تشخیص دهد. الگوریتم محاسبه اندازه خروجی بر اساس رابطه ۴ تعریف می‌شود(Du, 2018):

$$H_{out} = 1 + \frac{H_{in}+(2*pad)-K_{height}}{S};\ W_{out} = 1 + \frac{W_{in}+(2*pad)-K_{width}}{S} \quad (۴)$$

لایه فعال‌ساز

غیرخطی کردن خروجی لایه کانولوشنال با استفاده از لایه‌های فعال‌سازی در شبکه‌های عصبی اهمیت زیادی دارد، زیرا این لایه‌ها به رفع مشکل گرادیان ناپدید شده کمک می‌کنند که می‌تواند مانع یادگیری مؤثر شود. توابع فعال‌سازی متنوعی مانند Sigmoid، Tanh، ReLU، Leaky ReLU، ELU و Maxout وجود دارد که هرکدام مزایای خاص خود را دارند. به ویژه، Leaky ReLU به‌دلیل سرعت بالای همگرایی و محاسبات ساده‌تر نسبت به Sigmoid و Tanh، به عنوان گزینه‌ای محبوب در سال‌های اخیر شناخته شده است. این تابع فاقد ناحیه مرده است که در برخی توابع دیگر وجود دارد و می‌تواند به بهبود عملکرد و سرعت آموزش شبکه‌های عمیق کمک کند. بنابراین، انتخاب مناسب تابع فعال‌سازی تأثیر زیادی در عملکرد شبکه‌های عصبی دارد(Du, 2018).

لایه ادغام

لایه ادغام در شبکه‌های عصبی کانولوشنال (CNN) به عنوان یک اجزای کلیدی برای کاهش ابعاد ویژگی‌های استخراج‌شده از لایه‌های کانولوشنال عمل می‌کند و به کاهش پیچیدگی محاسباتی و جلوگیری از بیش برازش کمک می‌کند. سه نوع ادغام اصلی شامل ادغام عمومی، ادغام همپوشانی و ادغام هرمی فضایی (SPP) وجود دارد. ادغام عمومی با اندازه گام حرکت می‌کند و شامل ادغام حداکثر و ادغام میانگین است، در حالی که ادغام همپوشانی با نواحی همپوشان، اطلاعات بیشتری را حفظ می‌کند. SPP با قابلیت‌های منحصر به فرد خود، ویژگی‌های تصاویر را با اندازه‌های ورودی مختلف به ابعادی یکنواخت ترجمه کرده و از دست دادن اطلاعات را جلوگیری می‌کند. به همین دلیل، SPP به عنصر اصلی در طراحی شبکه‌های عصبی عمیق (DNN) تبدیل شده و کارایی و دقت این شبکه‌ها را به طرز چشمگیری افزایش می‌دهد.(Du, 2018).

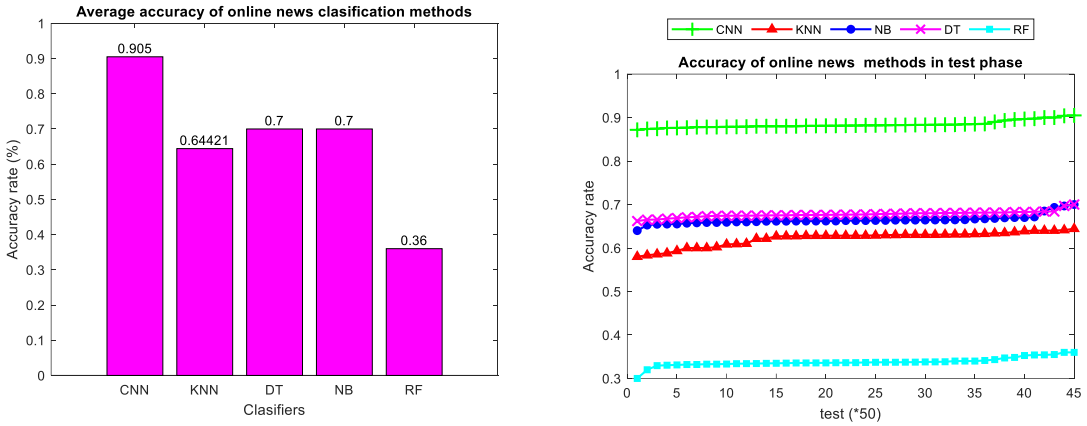
لایه کاملاً متصل

لایه‌های کاملاً متصل در انتهای شبکه‌های عصبی عمیق قرار دارند و وظیفه انتقال اطلاعات استخراج‌شده از لایه‌های قبلی به خروجی نهایی را بر عهده دارند. این لایه‌ها به گونه‌ای طراحی شده‌اند که هر نورون در آن به تمامی نورون‌های لایه قبلی متصل است، که این ارتباط کامل باعث می‌شود تا تمام ویژگی‌های استخراج‌شده به‌طور همزمان استفاده شوند و به دقت خروجی‌ها افزوده شود. به‌خصوص پس از چندین لایه کانولوشنال که ویژگی‌های مختلفی را از تصاویر استخراج کرده‌اند، لایه‌های کاملاً متصل می‌توانند این اطلاعات را تجزیه و تحلیل کنند و نمایشی جامع‌تر از داده‌ها ایجاد کنند. این لایه‌ها در شرایطی که نیاز به ترکیب اطلاعات از منابع مختلف وجود دارد، به ساده‌سازی محاسبات و افزایش سرعت پردازش کمک می‌کنند و به همین دلیل در طراحی و عملکرد نهایی شبکه‌های عصبی عمیق (DNN) اهمیت زیادی دارند. در نهایت، استفاده از لایه‌های کاملاً متصل بهبود قابلیت‌های شبکه در وظایفی مانند طبقه‌بندی و پیش‌بینی را تضمین می‌کند(Wu, 2017).

ارزیابی روش پیشنهادی

در این بخش، برای ارزیابی روش پیشنهادی از معیارهای مبتنی بر ماتریس آشفتگی استفاده شده است که یکی از اصلی‌ترین آن‌ها "دقت" است. دقت به نسبت تعداد اخبار آنلاین به‌درستی شناسایی‌شده به کل اخبار در داده‌های تست اشاره دارد و به عنوان یک افراز

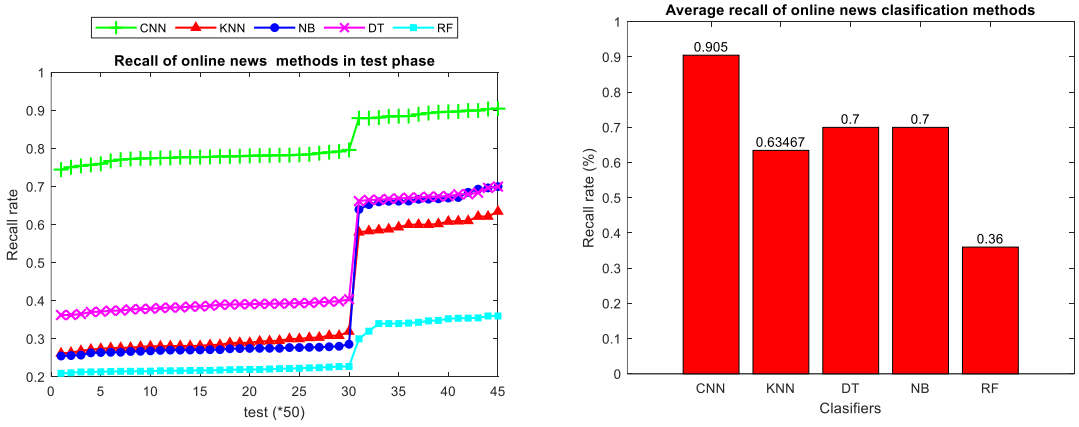
کیفی، کارایی کلی سیستم در تشخیص اخبار آنلاین را نشان می‌دهد. شکل ۴، نمودار مقایسه و میله ای دقت را برای روش‌های مختلف طبقه‌بندی در مجموعه داده‌های تست نمایش می‌دهد.



شکل ۴. نمودار معیار دقت در روش‌های طبقه‌بندی مختلف در مجموعه داده تست

بر اساس نتایج به‌دست‌آمده از شکل ۴ می توان دید روش ترکیبی مبتنی بر شبکه عصبی عمیق از نظر دقت، عملکرد بهتری نسبت به سایر روش‌ها دارد. تعداد اخبار جدید در مجموعه داده تست ۲۲۵۰ خبر بوده و نمودارها به گونه‌ای طراحی شده‌اند که وضوح نتایج را افزایش داده و تحلیل و تفسیر را آسان‌تر کنند. نمودارهای مشاهده‌شده در شکل‌های ۴ نشان می‌دهند که شبکه عصبی عمیق نسبت به روش‌های دیگر مانند درختان تصمیم و ماشین‌های بردار پشتیبانی، دقت بالاتری در شناسایی اخبار آنلاین به‌دست می‌آورد. این امر تأکید می‌کند که استفاده از این تکنیک قادر است فرآیند طبقه‌بندی و شناسایی اخبار را به‌طور چشمگیری بهبود بخشد و تحلیلگران را قادر می‌سازد با اطمینان بیشتری به این فعالیت بپردازند

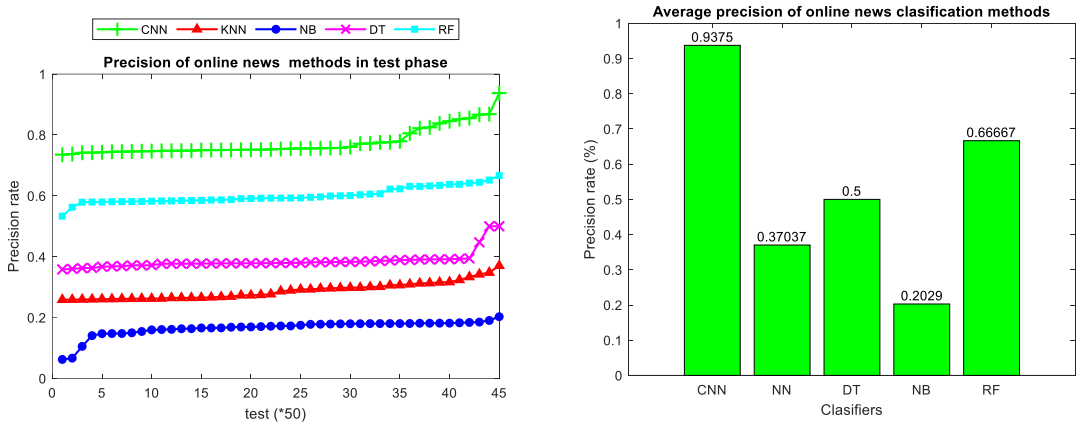
معیار حساسیت نیز یکی از شاخص‌های کلیدی در این روش است که به تحلیل دقت شناسایی اخبار آنلاین مهم و مرتبط کمک می‌کند. شکل ۵ نمودار مقایسه ای و میله ای مربوط به معیار حساسیت را در روش‌های مختلف طبقه‌بندی نمایش می‌دهد و این امکان را فراهم می‌آورد که توانایی‌های گوناگون این روش‌ها در شناسایی اخبار آنلاین مقایسه شود.



شکل ۵. نمودار معیار حساسیت در روش‌های طبقه‌بندی مختلف در مجموعه داده تست

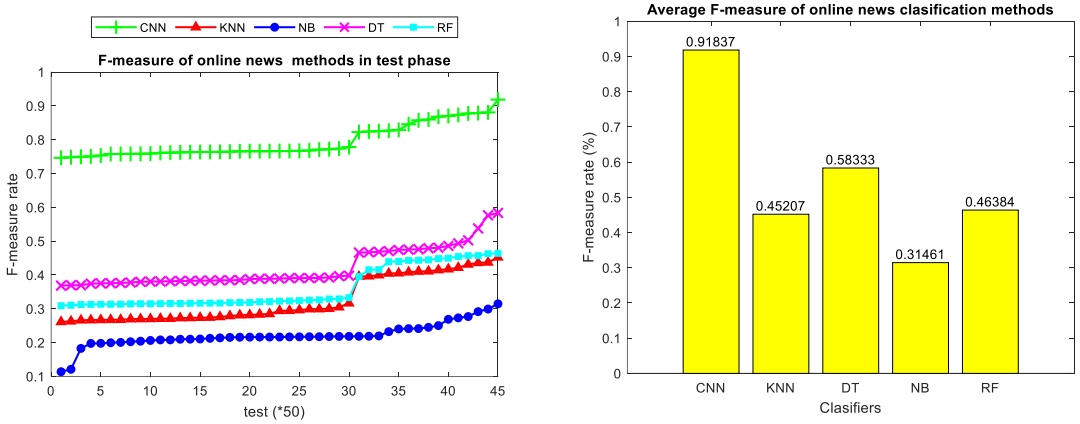
بر اساس شکل ۵، روش پیشنهادی که بر پایه اطلاعات اخبار آنلاین و ترکیب آن با شبکه‌های عصبی عمیق و سایر الگوریتم‌های طبقه‌بندی طراحی شده، در زمینه حساسیت عملکردی برجسته‌ای دارد. علاوه بر حساسیت، معیار صحت نیز به‌عنوان یکی از شاخص‌های کلیدی در این روش لحاظ شده است و نشان‌دهنده دقت شناسایی اخبار آنلاین واقعی و دیگر دسته‌ها در مجموعه داده تست است. این معیار اهمیت زیادی در ارزیابی توانایی سیستم در تشخیص صحیح اخبار دارد و نقش حیاتی در سنجش عملکرد

الگوریتم‌های طبقه‌بندی ایفا می‌کند. شکل ۶، نمودار مقایسه و میله‌ای صحت را برای روش‌های مختلف طبقه‌بندی در مجموعه داده‌های تست نمایش می‌دهد.



شکل ۶. نمودار معیار حساسیت در روش‌های طبقه‌بندی مختلف در مجموعه داده تست

نتایج بررسی‌های انجام‌شده در شکل‌های ۶ نشان می‌دهد که روش پیشنهادی ما، با تمرکز بر تحلیل اطلاعات اخبار آنلاین و ادغام آن با شبکه‌های عصبی عمیق و سایر روش‌های طبقه‌بندی، عملکرد برتری را نسبت به سایر روش‌ها دارد. شبکه عصبی عمیق به‌عنوان یک مدل پیشرفته در یادگیری ماشین، توانسته است دقت بالاتری برای معیار صحت برقرار کند. این کارایی به دلیل قابلیت این شبکه‌ها در یادگیری الگوهای پیچیده و استخراج ویژگی‌های مهم از داده‌هاست که به‌ویژه در شرایط داده‌های اخبار آنلاین به‌خوبی نمایان می‌شود.
معیار F، که به‌عنوان میانگین‌هارمونیک حساسیت و صحت شناخته می‌شود، در این روش به‌کار رفته و ابزاری کارآمد برای ارزیابی عملکرد کلی طبقه‌بندی اخبار آنلاین در مجموعه داده تست فراهم می‌آورد. این معیار ما را قادر می‌سازد تا توانایی مدل در شناسایی صحیح اخبار آنلاین را بهتر درک کنیم، به‌خصوص در شرایطی که داده‌ها ناهماهنگ یا نامتعادل هستند. نمودارهای ارائه شده در شکل ۷، نمودارهای مقایسه ای و میله ای مربوط به معیار F را در روش‌های مختلف به‌صورتی واضح نشان می دهد.

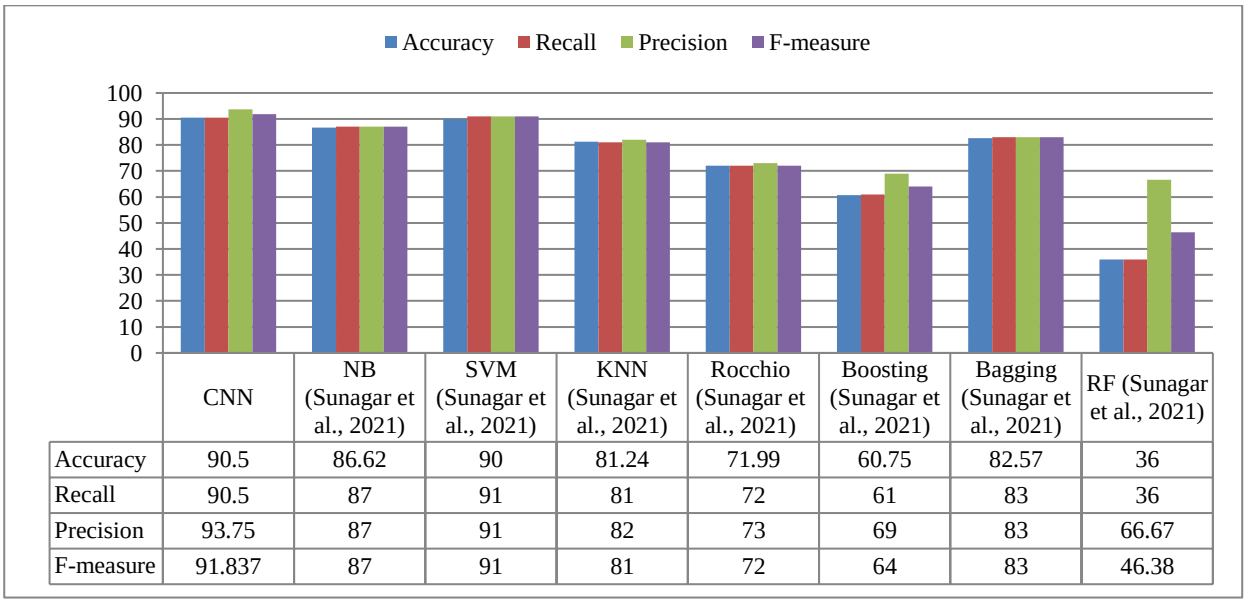


شکل ۷. نمودار معیار حساسیت در روش‌های طبقه‌بندی مختلف در مجموعه داده تست

در شکل۷ مشاهده می‌شود، روش پیشنهادی این تحقیق با ترکیب اطلاعات اخبار آنلاین و استفاده از شبکه‌های عصبی عمیق در طبقه‌بندی اخبار، کارایی بهتری نسبت به سایر روش‌ها از خود نشان داده است. نتایج به‌دست آمده از معیار F، که نشان‌دهنده دقت و کارایی مدل‌هاست، به‌طور قابل‌توجهی حاکی از بهبود عملکرد این روش است. به‌ویژه در شرایط داده‌های ناهماهنگ و پیچیده، شبکه عصبی عمیق توانسته است نتایج چشمگیری ارائه دهد و به شناسایی و طبقه‌بندی بهتر اخبار آنلاین کمک کند.

مقایسه روش پیشنهادی با روش‌های پیشین

مقایسه روش پیشنهادی با روش‌های پیشین (Sunagar et al., 2021) در طبقه‌بندی اخبار آنلاین اهمیت زیادی دارد، زیرا این مقایسه به انتخاب شیوه‌های بهینه‌تر، شناسایی نقاط قوت و ضعف هر روش، و ارزیابی دقیق کارایی و عملکرد روش جدید نسبت به گذشته کمک می‌کند. این ارزیابی می‌تواند تأثیر روش‌های مختلف بر دقت و صحت نتایج طبقه‌بندی را نشان دهد و اعتبار نتایج به‌دست آمده را تأیید کند. در نهایت، این فرایند منجر به بهبود کارایی سیستم‌ها و تقویت اعتماد به نتایج علمی پژوهش‌ها می‌شود. شکل ۸ مقایسه روش پیشنهادی با سایر روش های پیشین از نظر معیارهای ارزیابی را نشان می دهد.



	CNN	NB (Sunagar et al., 2021)	SVM (Sunagar et al., 2021)	KNN (Sunagar et al., 2021)	Rocchio (Sunagar et al., 2021)	Boosting (Sunagar et al., 2021)	Bagging (Sunagar et al., 2021)	RF (Sunagar et al., 2021)
Accuracy	90.5	86.62	90	81.24	71.99	60.75	82.57	36
Recall	90.5	87	91	81	72	61	83	36
Precision	93.75	87	91	82	73	69	83	66.67
F-measure	91.837	87	91	81	72	64	83	46.38

شکل ۸. مقایسه روش پیشنهادی با روش‌های پیشین از نظر معیارهای ارزیابی

شکل ۸ نشان می‌دهد که استفاده از معیار شباهت کسینوسی و شبکه‌های عصبی عمیق تأثیر قابل توجهی در طبقه‌بندی اخبار آنلاین دارد و به افزایش دقت و بهبود میانگین معیارهای ارزیابی کمک می‌کند. این روش‌ها به ویژه در پیش‌بینی اخبار جدید، حتی با داده‌های مشابه، عملکرد خوبی از خود نشان می‌دهند. یادگیری عمیق به دلیل ساختار پیچیده و لایه‌ای خود، قادر است الگوها و ویژگی‌های نهفته در داده‌ها را شناسایی کند و زمانی که با معیار شباهت کسینوسی ترکیب شود، نتایج امیدوارکننده‌ای به دست می‌آورد.

بحث و نتیجه‌گیری

در این تحقیق یک رویکرد طبقه بند اخبار آنلاین بر اساس ترکیب معیار شباهت کسینوسی فازی و شبکه‌های عصبی عمیق، ارائه شد. روش پیشنهادی در مقایسه با روش‌های قبلی که از معیارهای کلاسیک همچون Bag of Words، TF-IDF و CHI2 استفاده کرده‌اند، از معیار شباهت کسینوسی فازی استفاده نموده است. تحلیل‌ها و نتایج به‌دست‌آمده نشان می‌دهد که معیار شباهت کسینوسی به‌عنوان ابزاری مؤثر در شناسایی شباهت‌های متنی در اخبار آنلاین است و ترکیب آن با روش‌های پیشرفته مانند شبکه‌های عصبی به عملکرد بهتری نسبت به روش‌های گذشته منجر شده است. این یافته‌ها نه‌تنها اعتبار و کارایی معیار شباهت کسینوسی را تأیید می‌کند، بلکه نشان‌دهنده توانایی آن در بهبود فرآیندهای مرتبط با پیش‌بینی و دسته‌بندی اخبار آنلاین در مجموعه‌های داده تست است. نتایج آزمایشات نشان می‌دهد روش پیشنهادی برای معیارهای ارزیابی شامل دقت، حساسیت، صحت و معیار F به ترتیب مقادیری برابر با ۹۰.۵، ۹۰.۵، ۹۳.۷۵ و ۹۱.۸۳ نسبت به سایر روش‌های طبقه بندی و روش‌های پیشین در پیش بینی ابار آنلاین جدید، بهتر عمل کرده است.

مراجع

Ahmed, J., & Ahmed, M. (2021). Online news classification using machine learning techniques. *IIUM Engineering Journal*, *22*(2), 210-225.

Daud, S., Ullah, M., Rehman, A., Saba, T., Damaševičius, R., & Sattar, A. (2023). Topic classification of online news articles using optimized machine learning models. *Computers, 12*(1), 16.

Dogra, V., Verma, S., Chatterjee, P., Shafi, J., Choi, J., & Ijaz, M. F. (2022). A complete process of text classification system using state-of-the-art NLP models. *Computational Intelligence and Neuroscience, 2022*.

Du, J. (2018). *Understanding of object detection based on CNN family and YOLO.* Paper presented at the Journal of Physics: Conference Series.

Fayaz, M., Khan, A., Bilal, M., & Khan, S. U. (2022). Machine learning for fake news classification with optimal feature selection. *Soft Computing, 26*(16), 7763-7771.

Ghadiri, N., Ghadiri, A., & Sheikholeslami, A. (2022). *A fuzzy deep learning approach to health-related text classification.* Paper presented at the Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation: Proceedings of the INFUS 2021 Conference, held August 24-26, 2021. Volume 2.

Jindal, R., Malhotra, R., & Jain, A. (2015). Techniques for text classification: Literature review and current trends. *webology, 12*(2).

Katari, R., & Myneni, M. B. (2020). *A survey on news classification techniques.* Paper presented at the 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA).

Kumar, A., Singh, M., & Pais, A. R. (2019). *Fuzzy string matching algorithm for spam detection in Twitter.* Paper presented at the Security and Privacy: Second ISEA International Conference, ISEA-ISAP 2018, Jaipur, India, January, 9–11, 2019, Revised Selected Papers 2.

Naredla, N. R., & Adedoyin, F. F. (2022). Detection of hyperpartisan news articles using natural language processing technique. *International Journal of Information Management Data Insights, 2*(1), 100064. doi:https://doi.org/10.1016/j.jjimei.2022.100064

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Sunagar, P., Kanavalli, A., Nayak, S. S., Mahan, S. R., Prasad, S., & Prasad, S. (2021). *News Topic Classification Using Machine Learning Techniques.* Paper presented at the International Conference on Communication, Computing and Electronics Systems: Proceedings of ICCCES 2020.

Tandel, S. S., Jamadar, A., & Dudugu, S. (2019). *A survey on text mining techniques.* Paper presented at the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS).

Thaipisutikul, T., Tuarob, S., Pongpaichet, S., Amornvatcharapong, A., & Shih, T. K. (2021). *Automated classification of criminal and violent activities in Thailand from online news articles.* Paper presented at the 2021 13th International Conference on Knowledge and Smart Technology (KST).

Timmerman, Y., & Bronselaer, A. (2022). Automated monitoring of online news accuracy with change classification models. *Information Processing & Management, 59*(6), 103105. doi:https://doi.org/10.1016/j.ipm.2022.103105

Wu, J. (2017). Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China, 5*(23), 495.

Yousafzai, S. N., Shahbaz, H., Ali, A., Qamar, A., Nasir, I. M., Tehsin, S., & Damaševičius, R. (2024). X-News dataset for online news categorization. *International Journal of Intelligent Computing and Cybernetics*.

Improving the classification of articles in online news sites based on fuzzy approaches

Mina Kaseb
Department of Computer Engineering, Ardabil branch, Islamic Azad University, Ardabil, Iran
Masoud Bekravi
Department of Computer Engineering, Ardabil branch, Islamic Azad University, Ardabil, Iran

Babak Nouri-Moghaddam
Department of Computer Engineering, Ardabil Branch, Islamic Azad University Ardabil, Iran

Abstract

The classification of online news in news websites is conducted to organize content and facilitate user access to their preferred articles. These classifications, typically based on news topics such as economics, politics, culture, sports, and science and technology, assist users in easily locating the information they need. Furthermore, the implementation of appropriate classification methods in textual articles is highly significant, as it enhances rapid access and search efficiency for information within search engines by creating logical and meaningful categories. Consequently, users can more effectively access their desired articles. This study presents an online news classification approach based on the combination of fuzzy cosine similarity and deep neural networks. Compared to the previous methods that used classic criteria such as Bag of Words, TF-IDF and CHI2, the proposed method has achieved better results by using the fuzzy cosine similarity criterion. Experimental results indicate that the proposed method yields superior performance for evaluation metrics including accuracy, sensitivity, precision, and F-measure, with respective values of 90.5, 90.5, 93.75, and 91.83 compared to other classification methods and prior approaches in predicting new online news articles.

Keywords: “news website”, “online news classification”, “fuzzy similarity measure”, “deep learning”.