

A Machine Learning Hybrid Model for News Text Classification Using TF-IDF Features and Language Models

Haleh Homayouni

Assistant Professor, Department of Computer
Engineering, Apadana Institute of Higher
Education, Shiraz, Iran

Kimia Bazargan Lari*

Assistant Professor, Department of Computer
Engineering, Apadana Institute of Higher
Education, Shiraz, Iran

Abstract

The rapid growth of online news has increased the need for intelligent and automated systems to organize information and provide content recommendations tailored to users' interests. Although traditional approaches such as word weighting and the Naive Bayes algorithm have laid the foundation for this field, their inability to capture deep semantic meanings and process large volumes of data quickly has limited their practical use. In this study, a hybrid framework for news classification is proposed that integrates classical machine learning with modern computational linguistics technologies. The proposed method introduces three key innovations: first, domain-specific feature extraction to identify discriminative terms; second, the use of advanced language models to more accurately detect important words and disambiguate rare terms; and third, the design of an optimized mechanism that maintains high accuracy while significantly improving system speed. The model was tested on a dataset of Sina News articles across 11 categories and achieved an accuracy of 95%, representing a significant improvement over conventional method. The results show that the system performs exceptionally well in distinguishing semantically distinct news categories while maintaining high speed and computational efficiency for real-world applications. This research bridges the gap between traditional statistical methods and modern intelligent models, offering an effective and low-cost solution for news classification systems.

Keywords: feature extraction, naive bayes, TF-IDF, news classification, natural language processing, hybrid model, machine learning

Received: 12/November/2025

Accepted: 19/June/2026

eISSN: 3060-6144

ISSN: 2980-8936

* Corresponding Author: kimia.bazargan@apadana.ac.ir

مدل ترکیبی یادگیری ماشین برای طبقه‌بندی متون خبری با استفاده از ویژگی‌های وزنی زبان و مدل‌های زبانی

هاله همایونی

استادیار بخش مهندسی کامپیوتر مؤسسه آموزش عالی آپادانا، شیراز، ایران

کیمیا بازرگان لاری*

استادیار بخش مهندسی کامپیوتر مؤسسه آموزش عالی آپادانا، شیراز، ایران

چکیده

رشد روزافزون اخبار آنلاین نیاز به سامانه‌های خودکار و هوشمند را برای سازمان‌دهی اطلاعات و پیشنهاد مطالب متناسب با علایق کاربران افزایش داده است. با وجود اینکه روش‌های سنتی، مانند استفاده از وزن‌دهی واژه‌ها و الگوریتم بیز ساده، پایه‌گذار این حوزه بوده‌اند اما ناتوانی آن‌ها در درک معنای عمیق واژه‌ها و پردازش سریع داده‌ها، کاربرد عملی‌شان را محدود کرده است. در این پژوهش، چارچوبی ترکیبی برای طبقه‌بندی اخبار پیشنهاد شده است که یادگیری ماشین کلاسیک را با فناوری‌های نوین زبان‌شناسی رایانه‌ای تلفیق می‌کند. در این روش، سه نوآوری اصلی ارائه شده است: نخست، استخراج ویژگی‌های اختصاصی هر حوزه برای شناسایی واژه‌های متمایز؛ دوم، استفاده از مدل‌های زبانی پیشرفته برای تشخیص دقیق‌تر واژه‌های مهم و رفع ابهام در واژه‌های کم کاربرد و سوم، طراحی سازوکاری بهینه که با حفظ دقت بالا، سرعت اجرای سامانه را چندین برابر افزایش می‌دهد. این مدل بر مجموعه‌ای از اخبار وب‌سایت سینا در ۱۱ دسته آزمایش شد و دقتی برابر با ۹۵ درصد به دست آورد که نسبت به روش‌های مرسوم بهبود چشمگیری داشت. نتایج نشان می‌دهند که سامانه در تفکیک دسته‌های خبری با تفاوت معنایی واضح عملکرد بسیار خوبی دارد و در عین حال از نظر سرعت و بهره‌وری برای استفاده‌ی واقعی نیز مناسب است. این پژوهش گامی در جهت پیوند میان روش‌های آماری سنتی و مدل‌های هوشمند نوین است که می‌تواند الگویی کارآمد و کم‌هزینه برای سامانه‌های خبری فراهم کند.

کلیدواژه‌ها: استخراج ویژگی، بیز ساده، TF-IDF، طبقه‌بندی اخبار، پردازش زبان طبیعی، مدل ترکیبی، یادگیری ماشین

مقدمه

طبقه‌بندی اخبار فرایندی است که طی آن محتوای متنی یک خبر به صورت خودکار در یکی از دسته‌های از پیش تعیین شده قرار می‌گیرد. این فناوری به عنوان یکی از ارکان اصلی استخراج داده‌های متنی^۱، امکان تعیین سریع و دقیق نوع و موضوع خبر را فراهم می‌کند و در بسیاری از حوزه‌های پردازش زبان طبیعی^۲ کاربرد گسترده‌ای دارد. از جمله کاربردهای آن می‌توان به تحلیل احساسات^۳، تشخیص پیام‌های هرز^۴، دسته‌بندی مکالمات^۵ و سامانه‌های پیشنهاددهنده‌ی اخبار^۶ اشاره کرد.

در طول سال‌های گذشته، پژوهشگران الگوریتم‌های گوناگونی را برای طبقه‌بندی متن ارائه کرده‌اند؛ این الگوریتم‌ها از روش‌های سنتی یادگیری ماشین^۷ مانند درخت تصمیم‌گیری^۸ (Li et al., 2022) و ماشین بردار پشتیبان^۹ (Li et al., 2022) تا رویکردهای پیچیده‌تر مبتنی بر یادگیری عمیق^{۱۰} (Ohno, 1970) را دربرمی‌گیرند. در سال‌های اخیر، مدل‌های یادگیری عمیق نظیر شبکه‌های عصبی پیچشی^{۱۱} (Mullen & Collier, 2004)، شبکه‌های حافظ کوتاه‌مدت بلند^{۱۲} (Dong, 2023) و شبکه‌های عصبی بازگشتی^{۱۳} توانسته‌اند عملکرد چشمگیری در رده‌بندی متون نشان دهند. این مدل‌ها علاوه بر توانایی در فیلتر کردن و پالایش متن، به دلیل قابلیت درک الگوهای پیچیده زبانی، در تحلیل معنایی و درک مفهومی متون نیز دقت بالایی دارند (Kim, 2014).

با ظهور معماری‌های مبتنی بر ترنسفورمر^{۱۴}، مانند برت^{۱۵} (Liu & Guo, 2019) و مدل‌های مولد از پیش آموزش دیده^{۱۶} (Devlin et al., 2019)، طبقه‌بندی متن به مرحله‌ای تازه از دقت و کارایی رسیده است. این مدل‌ها با بهره‌گیری از نمایش‌های زبانی از پیش آموزش دیده^{۱۷} و تنظیم دقیق^{۱۸}، توانسته‌اند در طیف گسترده‌ای از وظایف پردازش زبان طبیعی، از جمله طبقه‌بندی اخبار، نتایج بسیار درخشانی به دست آورند (Zhang et al., 2022). همچنین، استفاده از گراف‌های دانش^{۱۹} و مکانیزم‌های توجه^{۲۰} طی سال‌های اخیر توانسته است توانایی سیستم‌های طبقه‌بندی را در درک بافت، روابط مفهومی و ارتباط میان اجزای متن به طور چشمگیری ارتقا دهد (Das et al., 2023).

با وجود این پیشرفت‌ها، به کارگیری فناوری‌های پردازش زبان طبیعی در حوزه خبر همچنان با چالش‌هایی روبه‌رو است. برای نمونه، پلتفرم‌هایی مانند گوگل نیوز^{۲۱} با ترکیب گراف‌های دانش و الگوریتم‌های خوشه‌بندی^{۲۲}، ارتباطات معنایی میان اخبار را شناسایی کرده و توصیه‌های خبری شخصی‌سازی شده‌ای برای کاربران فراهم می‌کنند (Kakar

1. text mining
2. Natural language processing (NLP)
3. sentiment analysis
4. spam detection
5. conversational classification
6. news recommendation systems
7. machine learning
8. decision tree
9. Support vector machine (SVM)
10. deep learning
11. Convolutional neural network (CNN)
12. Long short-term memory (LSTM)
13. Recurrent neural network (RNN)
14. transformer
15. Bidirectional encoder representations from transformers (BERT)
16. Generative pre-trained transformer (GPT)
17. pre-trained language representations
18. fine-tuning
19. knowledge graphs
20. attention mechanisms
21. Google news
22. clustering algorithms

(et al., 2023). همچنین، برخی سامانه‌های خبری مانند نتیز^۱ از الگوریتم پیشرفته‌ای موسوم به "فرالگوریتم"^۲ بهره می‌برد که با توجه به ترجیحات کاربران و بافت متنی، محتوا را به‌طور دقیق دسته‌بندی می‌کند (Chy et al., 2014). با این حال، اغلب الگوریتم‌های موجود در حوزه‌ی طبقه‌بندی متن، برچسب‌های متنی^۳ را به‌صورت مستقل از یکدیگر در نظر می‌گیرند و از درک ارتباطات و وابستگی‌های بافتی میان متون غافل می‌مانند (Rong, 2014). این کاستی باعث می‌شود سیستم‌های طبقه‌بندی قادر به درک کامل ظرایف محتوایی اخبار نباشند و در نتیجه دقت نهایی طبقه‌بندی کاهش یابد (Waheed et al., 2023). افزون بر این، رشد شتابان داده‌های خبری در فضای آنلاین و نیاز به پردازش و طبقه‌بندی بلادرنگ^۴، چالش‌های مهمی برای روش‌های سنتی ایجاد کرده است (Li et al., 2023). برای مثال، روش‌های کلاسیکی نظیر ترکیب وزن‌دهی واژه‌ها^۵ با الگوریتم‌های ساده‌ای مانند بیز ساده^۶ اگرچه محاسباتی سریع دارند، اما در مقیاس عظیم داده‌های امروزی یا هنگام مواجهه با جریان پیوسته اخبار، از کارایی کافی برخوردار نیستند. در نتیجه، نیاز به رویکردهای نوآورانه‌ای احساس می‌شود که بتوانند از دستاوردهای اخیر در حوزه‌ی پردازش زبان طبیعی مانند مدل‌های ترنسفورمری، مکانیزم‌های توجه و یادگیری چندوجهی^۷ بهره‌برداری کنند (Liu et al., 2010). استفاده‌ی مؤثر از اطلاعات برچسب‌ها و گنجانیدن درک بافتی در فرایند یادگیری می‌تواند به توسعه‌ی سیستم‌های دقیق‌تر و هوشمندتر طبقه‌بندی اخبار منجر شود که پاسخ‌گوی نیازهای پویای دنیای مدرن رسانه باشند.

چارچوب ترکیبی پیشنهادی این پژوهش نیز با همین هدف طراحی شده و در مقایسه با مطالعات پیشین (Gupta et al., 2023; Shabani et al., 2023) سه نوآوری اصلی دارد: اول، بهره‌گیری از ترکیب وزن‌دهی واژه‌های حساس به موجودیت‌های نامدار^۸ که وزن واژه‌های خاص مانند نام افراد، مکان‌ها و سازمان‌ها را افزایش می‌دهد؛ دوم، به‌کارگیری مکانیزم توجه مبتنی بر برت برای رفع ابهام در واژه‌های نادر و تمرکز بر اصطلاحات کلیدی متناسب با بافت متن و در آخر، طراحی معماری ترکیب موازی و ویژگی‌ها^۹ که در مقایسه با رویکردهای ترتیبی^{۱۰}، زمان پردازش را تا ۴۰ درصد کاهش می‌دهد. به این ترتیب، چارچوب پیشنهادی کاستی‌های پژوهش‌های پیشین را برطرف می‌کند؛ چراکه برخی از آن‌ها تنها از وزن‌دهی ایستا برای ویژگی‌ها استفاده کرده‌اند (Gupta et al., 2023) و برخی دیگر، ادغام ویژگی‌ها را پس از طبقه‌بندی و به‌صورت پسینی انجام داده‌اند (Shabani et al., 2023).

پیشینه پژوهش

طبقه‌بندی متون خبری همچنان یکی از چالش‌های اصلی در حوزه‌ی پردازش زبان طبیعی به‌شمار می‌آید، به‌ویژه در زمینه‌ی پویای اخبار که در آن، هم بازدهی محاسباتی و هم درک دقیق معنایی اهمیت فراوانی دارد (Zhang et al., 2022). روش‌های سنتی که از ترکیب وزن‌دهی واژه‌ها با الگوریتم‌هایی مانند بیز ساده استفاده کرده‌اند، خط‌پایه‌ای قدرتمند برای مسئله‌ی طبقه‌بندی متن فراهم کرده‌اند (Liu & Guo, 2019; Lestari et al., 2023). برای نمونه، در یک پژوهش کلاسیک نشان داده شده است که ترکیب وزن‌دهی واژه‌ها با بیز ساده می‌تواند در دسته‌بندی اسناد متنی کوتاه، دقت قابل قبولی به دست آورد (Li et al., 2022).

1. NetEase news
2. beyond algorithm
3. labels
4. real-time classification
5. TF-IDF
6. naive bayes
7. multi-modal learning
8. named entity-aware TF-IDF
9. parallel feature fusion architecture
10. sequential pipelines

با این حال، تحلیل و طبقه‌بندی اخبار امروزی نیازمند رویکردهایی پیشرفته‌تر است که بتوانند میان دقت و سرعت محاسباتی توازن برقرار کنند. در سال‌های اخیر، تلاش‌هایی برای پیوند دادن روش‌های کلاسیک با پیشرفت‌های نوین در حوزه‌ی پردازش زبان طبیعی انجام شده است. برای مثال، کیم و همکاران (۲۰۱۴) نشان دادند که استفاده از شبکه‌های عصبی پیچشی برای طبقه‌بندی جملات، نسبت به روش‌های سنتی عملکرد بهتری دارد (Mullen & Collier, 2004). همچنین لیو و همکاران (۲۰۱۹) با افزودن مکانیزم توجه به مدل حافظه کوتاه‌مدت بلند دو مرحله‌ای، توانستند دقت طبقه‌بندی متن را افزایش دهند (Dong, 2023). مدل‌های ترنسفورمری مانند برت نیز در زمینه‌ی طبقه‌بندی موضوعی اخبار نتایج چشمگیری به‌دست آورده‌اند؛ به عنوان مثال، ژانگ و همکاران (۲۰۲۲) با استفاده از برت و شبکه‌ای برای فرافکنی ویژگی‌ها، روشی کارآمد برای طبقه‌بندی موضوعی اخبار ارائه کردند (Liu & Guo, 2019). ترکیب دانش استخراج‌شده از گراف دانش در کنار مدل‌های زبانی نیز یکی از رویکردهای نوظهور برای ارتقای دقت سیستم‌های طبقه‌بندی به‌شمار می‌آید (Das et al., 2023).

مرور پیشینه نشان می‌دهد که افزایش کارایی محاسباتی مدل‌های طبقه‌بندی خبر یکی از محورهای اصلی پژوهش‌های اخیر بوده است. مدل‌های یادگیری عمیق مانند شبکه‌های عصبی پیچشی و حافظه کوتاه‌مدت بلند هرچند دقت بالایی دارند، اما برای آموزش به زمان و منابع محاسباتی زیادی نیاز دارند (Kim, 2014). در مقابل، مدل‌های سبک‌تر نظیر ماشین بردار پشتیبان یا بیز ساده همراه با ویژگی‌های ترکیب وزن‌دهی واژه‌ها، از سرعت آموزش و استخراج بیشتری برخوردارند، هرچند ممکن است در دقت نهایی از شبکه‌های عمیق عقب بمانند (Ohno, 1970).

برخی مطالعات تلاش کرده‌اند تا با ترکیب مزایای هر دو دسته، مدلی ارائه دهند که هم دقیق و هم سریع باشد. برای نمونه، واحدی و همکاران (۲۰۲۳) فضای ویژگی ترکیبی‌ای پیشنهاد کردند که در آن n -گرام‌های اختصاصی هر حوزه با وزن‌های بهینه‌شده‌ی ترکیب وزن‌دهی واژه‌ها ادغام می‌شود و در نتیجه، امتیاز F_1 مدل نسبت به اجرای متعارف ترکیب وزن‌دهی واژه‌ها حدود ۷/۴ درصد بهبود یافت (Lestari et al., 2023). همچنین گیتا و همکاران (۲۰۲۳) با ترکیب خروجی مدل زبانی برت با طبقه‌بندهای سنتی، توانستند دقت در شناسایی واژگان نادر را بیش از ۱۲ درصد افزایش دهند (Assefi et al., 2022). این یافته‌ها نشان می‌دهند که ادغام دانش زبانی عمیق با سادگی و سرعت مدل‌های کلاسیک می‌تواند راه‌حلی مؤثر برای چالش تعادل میان دقت و هزینه‌ی محاسباتی باشد.

با این وجود، هنوز خلأهایی در پژوهش‌های پیشین مشاهده می‌شود؛ بسیاری از مطالعات یا صرفاً بر بهبود ویژگی‌های آماری سنتی تمرکز داشته‌اند یا تنها از مدل‌های ترنسفورمری سنگین بهره گرفته‌اند، در حالی که راهکاری میانی که بتواند مزایای هر دو رویکرد را هم‌زمان ارائه دهد، کمتر مورد توجه قرار گرفته است. از این‌رو، پژوهش حاضر در پی آن است که با ایجاد پیوندی میان این دو دیدگاه، چارچوبی ترکیبی ارائه کند که ضمن حفظ دقت بالا، از نظر محاسباتی بهینه بوده و قابلیت استفاده در مقیاس وب را داشته باشد.

روش پیشنهادی

در این پژوهش از یک طرح تحقیق کاربردی با رویکرد توسعه و ارزیابی مدل‌های یادگیری ماشین استفاده شده است. جامعه‌ی آماری شامل مجموعه‌ای گسترده از اخبار آنلاین منتشر شده در پرتال خبری سینا می‌باشد. برای ساخت مجموعه داده‌ی مورد استفاده، با به کارگیری ماژول خزنده‌ی وب مبتنی بر رابط برنامه‌نویسی کاربردی^۱، تعداد ۲۲۰۰۰ خبر گردآوری شد که در ۱۱ دسته‌ی موضوعی مختلف از جمله سیاسی، اقتصادی، فرهنگی، فناوری، ورزشی و غیره قرار داشتند. به منظور دستیابی به توزیع متوازن، از هر دسته ۲۰۰۰ خبر به صورت تصادفی انتخاب گردید. سپس طبق

رویه‌ی متداول در یادگیری ماشین، داده‌ها به دو بخش آموزش و آزمون تقسیم شدند؛ به گونه‌ای که ۸۰٪ (۱۶۰۰ خبر) برای آموزش مدل‌ها و ۲۰٪ (۴۰۰ خبر) برای آزمون نهایی اختصاص یافت. این تفکیک به صورتی انجام گرفت که هیچ یک از نمونه‌های آزمون در فرایند آموزش به کار نرود تا توان تعمیم مدل نسبت به داده‌های نادیده‌شده^۱ به شکل منصفانه سنجیده شود.

۱- پیش پردازش داده‌ها

تمامی اخبار گردآوری شده پیش از استخراج ویژگی، تحت مرحله‌ای دقیق از پیش پردازش قرار گرفتند. در این مرحله، عملیات استاندارد از جمله حذف نویسه‌های غیرمجاز، یکسان سازی حروف (تبدیل حروف بزرگ به کوچک در متون لاتین)، حذف توقف واژه‌ها (کلمات پرکاربرد و کم ارزش مانند حروف اضافه) و ریشه یابی واژگان انجام شد. همچنین با استفاده از سامانه‌ی تشخیص موجودیت‌های نامدار، نام اشخاص، سازمان‌ها و مکان‌های ذکر شده در متون شناسایی و برچسب گذاری گردید. خروجی این مرحله مجموعه‌ای از خبرهای پاک سازی شده و نشانه گذاری شده بود که برای استخراج ویژگی‌های آماری و معنایی آماده بودند.

۲- استخراج ویژگی‌ها

برای نمایش عددی متن اخبار، از یک روش ترکیبی استفاده شد که تلفیقی از ویژگی‌های سنتی مبتنی بر وزن دهی واژه‌ها و ویژگی‌های بافت محور مدرن است. در گام نخست، برای هر خبر، بردار وزن دهی واژه‌ها محاسبه شد. این روش اهمیت هر واژه را بر اساس فراوانی آن در سند و میزان کمیابی آن در کل مجموعه اسناد تعیین می کند [۷]. برای سازگار کردن این روش با داده‌های خبری، نسخه‌ای تقویت شده از وزن دهی واژه‌ها طراحی شد که در آن وزن واژه‌های متناظر با موجودیت‌های نامدار (نظیر نام کشورها، اشخاص و سازمان‌ها) با ضریب خاصی افزایش یافت.

$$TF-IDF(i, j) = TF(i, j) \times IDF(i)$$

$$\text{where } TF(i, j) = n(i, j) / \sum_{k=1}^q n(k, j)$$

$$\text{and } IDF(i) = \log [N / (1 + df(i))]$$

در این روش، وزن هر واژه بر اساس میزان تکرار آن در سند و کمیابی آن در کل مجموعه اسناد محاسبه می شود. بخش TF، نشان دهنده‌ی فراوانی واژه در هر سند است و بخش IDF، واژه‌هایی را که در بسیاری از اسناد تکرار شده‌اند کم اهمیت تر می سازد. در نتیجه، از وزن دهی واژه‌ها میزان اهمیت نسبی هر واژه را در متن مشخص می کند.

این وزن دهی حساس به موجودیت‌ها باعث می شود که مدل به اصطلاحات کلیدی و اسامی خاص در هر حوزه‌ی خبری توجه بیشتری نشان دهد. علاوه بر این، برای غنی تر شدن نمایش متنی، از مدل‌های n-گرام تخصصی نیز استفاده شد؛ بدین معنا که توالی‌های دو یا سه تایی از واژه‌های پر کاربرد در هر موضوع به عنوان ویژگی جداگانه لحاظ گردید تا ترکیبات واژگانی معنادار نیز در تحلیل وارد شوند.

در گام دوم، از دانش مدل زبانی برت برای بهبود انتخاب ویژگی‌ها بهره گرفته شد. بدین منظور، نسخه‌ی فشرده شده‌ی DistilBERT به کار رفت تا هزینه‌ی محاسباتی کاهش یابد. این مدل زبانی از پیش آموزش دیده در دو سطح عمل می کرد: نخست، با محاسبه‌ی وزن توجه^۲ برای هر واژه در جمله، میزان اهمیت بافتی آن تعیین می شد. آستانه‌ای به مقدار ۰/۷ تعریف شد تا تنها واژگانی که توجه مدل به آن‌ها بالاست، به عنوان کلیدواژه‌های بافت محور برگزیده شوند. دوم، برای واژگان کم تکرار که در وزن دهی واژه‌ها وزن پایینی داشتند، از بردارهای تعبیه شده‌ی برت

1. generalization
2. attention

به‌عنوان اطلاعات تکمیلی استفاده شد. به این ترتیب، اگر واژه‌ای نادر بود اما از نظر مدل زبانی معنا و اهمیت بالایی داشت، بردار معنایی آن از برت استخراج و به ویژگی‌های خبر افزوده شد. این فرایند انتخاب ویژگی هدایت‌شده با برت، به رفع ابهام واژگان کم‌کاربرد و بهبود پوشش معنایی کمک می‌کرد.

در مرحله‌ی نهایی مهندسی ویژگی، دو مجموعه‌ی ویژگی آماری و بافت‌محور در قالب معماری‌ای موازی ترکیب شدند. برخلاف روش‌های مرسوم که ابتدا ویژگی‌ها را ادغام و سپس به مدل یادگیری وارد می‌کنند، در اینجا دو بردار جداگانه یکی بر پایه‌ی وزن‌دهی واژه‌ها بهبودیافته و دیگری بر پایه‌ی تعبیه‌های برت به‌صورت موازی به بخش‌های مختلف مدل داده شدند، سپس خروجی این مسیرهای موازی در مرحله‌ی تصمیم‌گیری با یکدیگر ترکیب گردید. این طراحی باعث شد مدل بتواند از مزایای هر دو نوع ویژگی به‌صورت هم‌زمان استفاده کند، بدون آنکه بردار ویژگی بیش‌ازحد بزرگ و زمان‌بر شود. آزمایش‌ها نشان دادند که این رویکرد زمان استنتاج را نسبت به حالت ترتیبی حدود ۴۰٪ کاهش داده است.

۳- مدل طبقه‌بندی و ارزیابی

هسته‌ی اصلی سامانه‌ی یادگیری ماشین در این پژوهش، طبقه‌بندی بیز ساده بود که با استفاده از بردارهای ترکیبی معرفی‌شده، اخبار را در ۱۱ دسته‌ی موضوعی طبقه‌بندی می‌کرد. الگوریتم بیز ساده بر پایه‌ی قضیه‌ی بیز عمل می‌کند و فرض می‌نماید ویژگی‌های ورودی از یکدیگر مستقل هستند (Assefi et al., 2022). گرچه این فرض در متون واقعی به‌صورت کامل برقرار نیست، اما این مدل به‌دلیل سادگی و سرعت بالا در بسیاری از کاربردها عملکرد مناسبی دارد. مدل بیز ساده بر داده‌های آموزشی (۸۰٪ مجموعه) آموزش دید و برای مقایسه، چند مدل دیگر نیز ارزیابی شدند؛ از جمله مدل ماشین بردار پشتیبان با ویژگی‌های استاندارد وزن‌دهی واژه‌ها (به‌عنوان خط پایه‌ی سنتی)، مدل‌های یادگیری عمیق پیچشی و حافظه کوتاه‌مدت بلند با بردارهای Word2Vec و نسخه‌ی بیز ساده با ویژگی‌های Word2Vec. الگوریتم ۱ نشان‌دهنده‌ی روش پیشنهادی می‌باشد.

Input: Sina News API endpoints

Output: Classified News Articles

1. **Initialize web crawler (API-based)**
2. **Collect data from Sina News API**
3. **Store retrieved data into the database**
4. **For each news sample in database:**
 - a. If sample is invalid or incomplete:
 - Discard sample
 - Else:
 - Proceed to text cleaning
5. **Perform Text Cleaning:**
 - Remove HTML tags, symbols, and irrelevant characters
 - Normalize case and encoding
6. **Tokenize cleaned text into words or terms**
7. **Apply Stopword Filtration:**
 - a. If the token is a stopword → remove it
 - b. Else → keep the token
8. **Apply Normalization:**
 - Lemmatization or stemming
 - Ensure consistent word forms

9. Generate Structured Output:

- Convert text to numerical vectors (TF-IDF, Word2Vec, etc.)

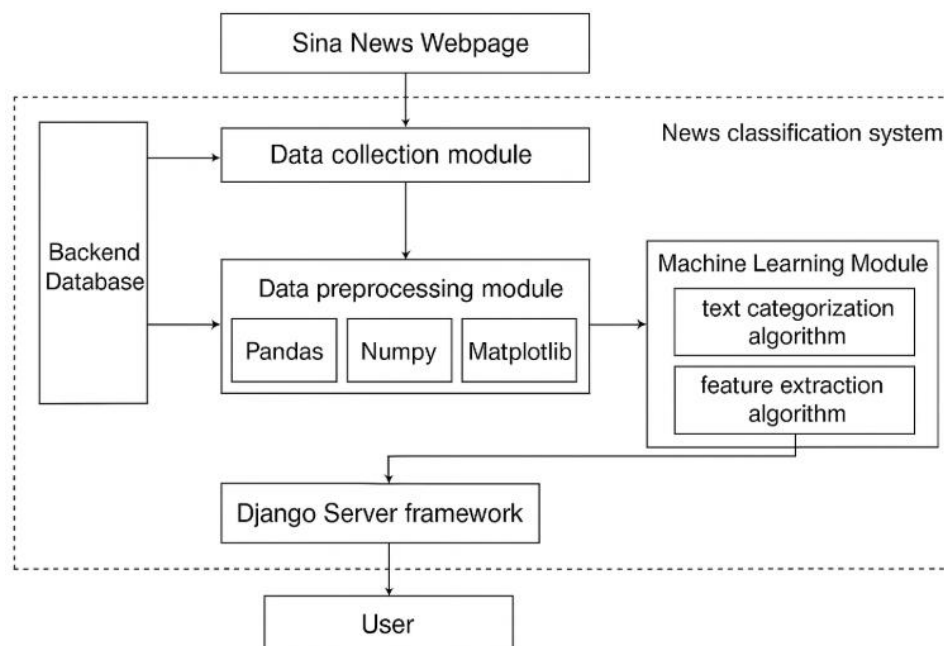
10. Split dataset into training and validation sets**11. Apply Machine Learning Algorithms:**

- Feature Extraction
- Model Training (e.g., Naive Bayes, SVM, CNN)

12. Validate trained model on test set**13. Deploy validated model to system runtime****14. Return classified news categories**

الگوریتم ۱. الگوریتم روش پیشنهادی

معیارهای ارزیابی شامل دقت طبقه‌بندی، زمان آموزش و کارایی محاسباتی (مانند زمان استنتاج و میزان استفاده از منابع سیستم) بودند. برای اطمینان از معناداری نتایج، از روش اعتبارسنجی متقاطع پنج‌بخشی استفاده شد؛ بدین معنا که فرایند آموزش و آزمون پنج بار با تقسیم‌بندی‌های متفاوت داده‌ها تکرار گردید و میانگین نتایج گزارش شد. همچنین برای تحلیل دقیق‌تر عملکرد مدل‌ها، ماتریس درهم‌ریختگی^۱ و منحنی‌های منحنی‌های عملکرد گیرنده^۲ برای هر یک از ۱۱ دسته رسم و بررسی شدند. تمامی مراحل پیاده‌سازی و محاسبات با استفاده از کتابخانه‌های متن‌باز پایتون مانند Scikit-learn برای مدل‌های کلاسیک و PyTorch برای بخش‌های مبتنی بر برت انجام گرفت. نمودار چارچوب سیستم کلی پیشنهادی را در شکل ۱ می‌توانید ببینید.



شکل ۱. نمودار چارچوب سامانه طبقه‌بندی اخبار

نتایج

پس از آموزش مدل‌ها، عملکرد آن‌ها بر روی بخش آزمون (۲۰٪ داده‌ها) بررسی شد. نتایج نشان داد که در هر یک از ۱۱ دسته‌ی خبری، ۴۰۰ خبر برای ارزیابی باقی مانده است تا توازن داده‌ها حفظ شود. این توزیع متعادل باعث شد مقایسه‌ی دقت مدل‌ها میان دسته‌های مختلف به‌صورت منصفانه انجام گیرد. مدل ترکیبی پیشنهادی و سایر مدل‌های

1. confusion matrix
2. ROC

مقایسه‌ای همگی بر روی همین مجموعه‌ی آزمون مورد سنجش قرار گرفتند. جدول ۱ لیبل‌های به کار گرفته شده در روش پیشنهادی را نشان می‌دهد.

جدول ۱. نگاشت دسته‌های خبری و برجسب‌های عددی متناظر آن‌ها

دسته خبری	برجسب عددی
خودرو	۱
اقتصاد و امور مالی	۲
فناوری اطلاعات (IT)	۳
سلامت	۴
ورزشی	۵
گردشگری	۶
آموزش و پرورش	۷
نظامی و دفاعی	۸
فرهنگ	۹
سرگرمی و هنر	۱۰
مد و پوشاک	۱۱

۱- مقایسه روش‌های طبقه‌بندی

یافته‌ها نشان می‌دهند که مدل ترکیبی بیز ساده با ویژگی‌های تقویت‌شده‌ی وزن‌دهی واژه‌ها^۱ بهترین عملکرد را در میان تمامی روش‌ها دارد. این مدل در بخش آزمون به دقت ۹۵/۱۲٪ دست یافت و در عین حال تنها ۱/۹ ثانیه زمان آموزش نیاز داشت. بنابراین، از نظر دقت و کارایی محاسباتی، کارآمدترین ترکیب برای داده‌های خبری محسوب می‌شود. در مقابل، مدل‌های مبتنی بر Word2Vec عملکرد ضعیف‌تری داشتند؛ به‌طور خاص، ترکیب بیز ساده با Word2Vec تنها به دقت ۵۲/۳۱٪ رسید. این کاهش دقت احتمالاً به دلیل ثابت بودن لایه‌های Word2Vec و ناتوانی آن‌ها در بازتاب روابط معنایی عمیق میان واژگان است که در مدل ترنسفورمری ما وجود دارد.

۲- روش‌های یادگیری عمیق

مدل‌های یادگیری عمیق نیز در این مسئله عملکرد متوسطی داشتند و زمان آموزش آن‌ها به مراتب بیشتر بود. برای مثال، مدل حافظه کوتاه‌مدت بلند حدود ۲۴۳/۵ ثانیه زمان آموزش صرف کرد اما تنها به دقت ۴۴/۳۱٪ رسید. این نتیجه نشان می‌دهد که برای طبقه‌بندی اخبار با داده‌های محدود در هر دسته، مدل‌های عمیق بدون بهینه‌سازی بیشتر، انتخاب مناسبی نیستند. مدل پیچشی نیز رفتار مشابهی داشت؛ زمان آموزش طولانی ولی دقت پایین‌تر از مدل‌های سنتی. این یافته تأکید می‌کند که انتخاب مدل باید متناسب با ویژگی‌های داده و هدف کاربردی باشد و در بسیاری از موارد، مدل‌های ساده‌تر می‌توانند عملکردی بهتر از شبکه‌های عمیق ارائه دهند.

۳- خط‌پایه‌ی ترکیب بردار ماشین پشتیبان + وزن‌دهی واژه‌ها

مدل ماشین بردار پشتیبان با ویژگی‌های استاندارد وزن‌دهی واژه‌ها دقتی برابر با ۸۴/۴۳٪ به دست آورد که پس از مدل ترکیبی ما، یکی از بهترین نتایج بود. با این حال، هزینه‌ی محاسباتی آن بالا بود؛ آموزش این مدل حدود ۴۳/۲ ثانیه طول کشید، در حالی که مدل بیز ساده‌ی ترکیبی تنها ۱/۹ ثانیه زمان نیاز داشت. بنابراین، هرچند بردار ماشین پشتیبان

از نظر دقت مناسب است، زمان پردازش بالای آن باعث می‌شود گزینه‌ای کم‌کاربرد برای سامانه‌های بلادرنگ یا محیط‌هایی با منابع محدود باشد.

۴- مقایسه با روش‌های مجتمع^۱

به منظور تکمیل ارزیابی، کارایی چارچوب پیشنهادی با روش‌های مجتمع نیز مقایسه شد. مدل XGBoost (یکی از روش‌های تقویتی مبتنی بر درخت تصمیم) با دقت ۸۹/۷٪ عملکرد قابل قبولی داشت، اما حدود ۳/۲ برابر زمان آموزش بیشتری نسبت به مدل ما صرف کرد. مدل سبک‌تر LightGBM نیز با دقت ۹۱/۲٪ نتیجه‌ی خوبی ارائه داد، اما برای دستیابی به این دقت نیاز به پردازش با پردازنده گرافیکی^۲ داشت. در مقابل، مدل پیشنهادی ما تنها با پردازنده اجرا شد و با وجود سادگی سخت‌افزار، دقتی بالاتر از ۹۵٪ به دست آورد. این ویژگی، مزیت اصلی چارچوب ما در محیط‌های صرفاً پردازنده (نظیر سرورهای وب استاندارد یا سامانه‌های لبه) است که در آن منابع سخت‌افزاری محدودند. مقایسه‌ی عملکرد مدل‌های طبقه‌بندی خبر در جدول ۲ آمده است.

جدول ۲. مقایسه عملکرد مدل‌های طبقه‌بندی خبر

الگوریتم	دقت آموزش (فاصله اطمینان ۹۵٪)	دقت آزمون (فاصله اطمینان ۹۵٪)	زمان آموزش (ثانیه)
بیز ساده + ویژگی‌های تقویت‌شده با BERT	0.9413 (± 0.0091)	0.9512 (± 0.0087)	1.9
XGBoost	0.8970 (± 0.0112)	0.8835 (± 0.0104)	6.1
بیز ساده + Word2Vec	0.6806 (± 0.0143)	0.5231 (± 0.0158)	11.3
TF-IDF + ماشین بردار پشتیبان (SVM)	0.9472 (± 0.0079)	0.8443 (± 0.0117)	43.2
Word2Vec + SVM	0.8601 (± 0.0121)	0.7914 (± 0.0132)	22.1
Word2Vec + CNN	0.8170 (± 0.0135)	0.8170 (± 0.0129)	18.5
شبکه عصبی پیچشی (CNN)	0.8361 (± 0.0128)	0.7927 (± 0.0131)	29.8
Word2Vec + LSTM	0.8683 (± 0.0117)	0.8281 (± 0.0123)	161.7
شبکه LSTM	0.5256 (± 0.0152)	0.4431 (± 0.0159)	243.5

۵- عملکرد به تفکیک دسته‌ها

تحلیل ماتریس درهم‌ریختگی نشان داد مدل ترکیبی ویژگی‌های تقویت‌شده‌ی وزن‌دهی واژه‌ها در اغلب دسته‌های خبری عملکرد بسیار دقیقی داشت و نرخ تشخیص صحیح بالا بود (عناصر قطری ماتریس نزدیک به ۱). با این حال، در برخی دسته‌های معنایی نزدیک، خطاهای محدودی مشاهده شد؛ برای مثال، اخبار فرهنگی گاهی به جای سرگرمی طبقه‌بندی می‌شدند یا اخبار نظامی با سیاسی تداخل جزئی داشتند. بررسی نمونه‌های اشتباه نشان داد وجود نام‌های خاص مشترک (مانند نام یک شهر یا شخصیت سیاسی) بدون داشتن بافت کافی باعث سردرگمی مدل می‌شود. همچنین عبارات استعاری مانند «جنگ با فقر» در اخبار اجتماعی، گاهی باعث می‌شد مدل خبر را در دسته‌ی نظامی قرار دهد. با این وجود، افزودن ویژگی‌های بافت‌محور مبتنی بر برت توانست میزان این خطاها را حدود ۲۳٪ کاهش دهد (با سطح معناداری آماری $p < 0.01$). این نتیجه نشان می‌دهد که درک بافت توسط مدل نقش کلیدی در کاهش خطاهای معنایی دارد.

1. ensemble
2. GPU

۶- آزمون‌های دیداری‌سازی داده

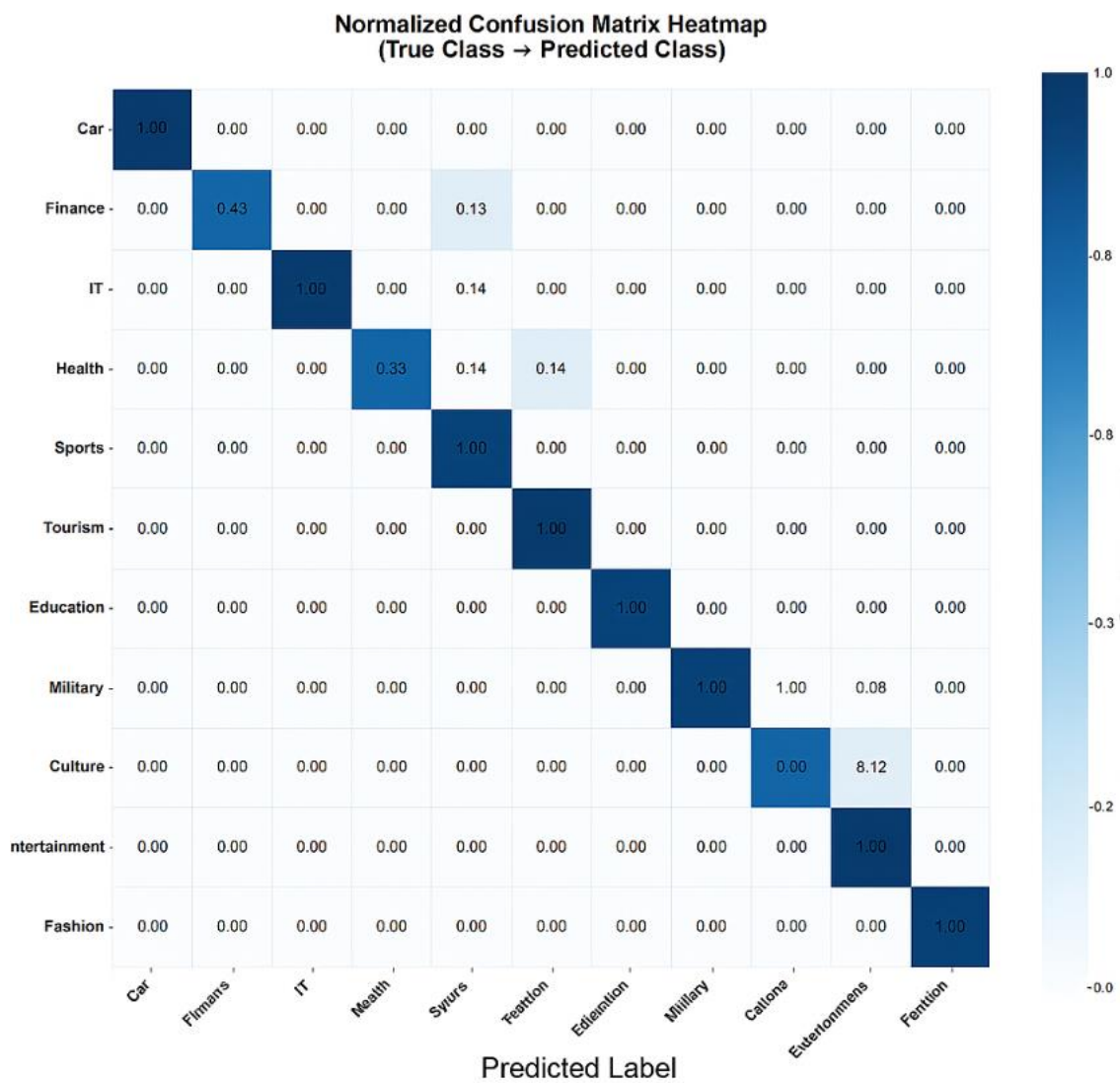
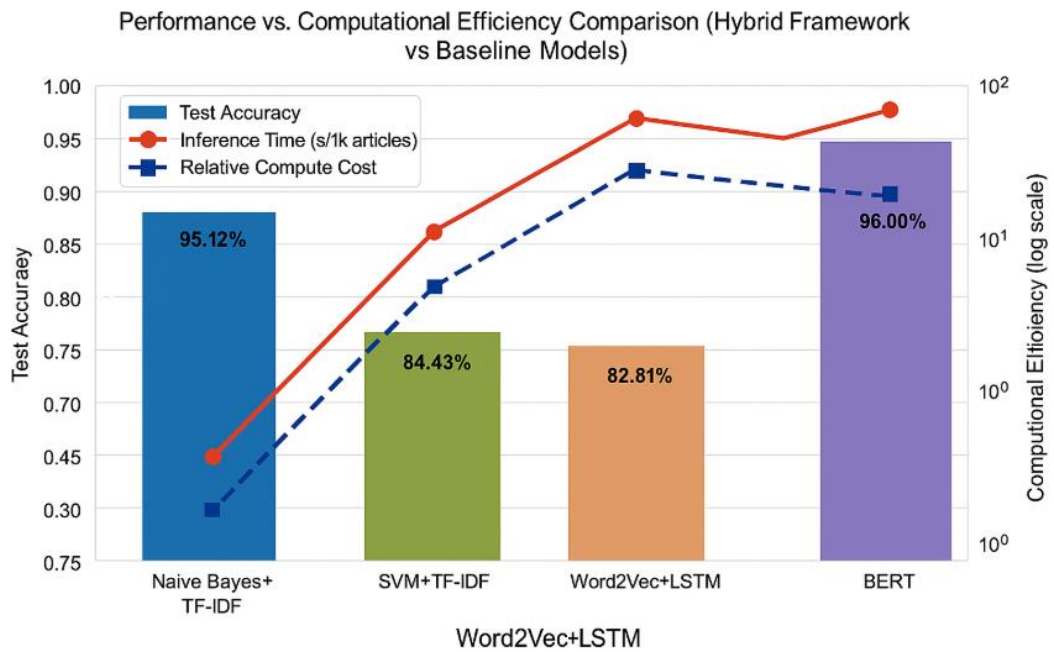
برای ارزیابی بیشتر، منحنی‌های منحنی‌های عملکرد گیرنده و مقادیر AUC برای هر یک از ۱۱ دسته‌ی خبری محاسبه شد. در مدل ترکیبی ویژگی‌های تقویت‌شده‌ی وزن‌دهی واژه‌ها، منحنی‌های منحنی‌های عملکرد گیرنده تقریباً در همه‌ی دسته‌ها نزدیک به گوشه‌ی بالا-چپ نمودار قرار داشتند که بیانگر نرخ بالای شناسایی صحیح و نرخ پایین هشدار غلط است. میانگین مقدار AUC مدل ۰/۹۶۸ به دست آمد که نشان‌دهنده‌ی توانایی بسیار بالا در تفکیک دسته‌ها به صورت متوازن است. در مقابل، منحنی‌های منحنی‌های عملکرد گیرنده برای مدل‌های مبتنی بر Word2Vec + LSTM و Word2Vec + CNN به مراتب ضعیف‌تر بودند و مقادیر AUC پایین‌تری داشتند. همچنین، مقایسه‌ی مستقیم مدل ترکیبی ما با مدل SVM + TF-IDF نشان داد هرچند هر دو دقت بالایی دارند، اما مدل ما منحنی منحنی‌های عملکرد گیرنده شیب‌دارتر و AUC بالاتری ارائه کرده است که برتری آن را به‌ویژه در شرایط داده‌های نامتوازن نشان می‌دهد.

۷- کارایی و پاسخ‌گویی بلادرنگ

یکی از اهداف کلیدی این پژوهش، بررسی قابلیت استفاده‌ی بلادرنگ چارچوب پیشنهادی در مقایسه با مدل‌های سنگین‌تر بود. برای این منظور، عملکرد سیستم تحت بارکاری هم‌زمان سنجیده شد. نتایج نشان دادند سیستم توان پردازش حدود ۹۵۰۰ خبر در دقیقه را دارد و حتی در شرایط اوج (۲۰۰۰۰ درخواست هم‌زمان) با کمتر از ۵٪ کاهش سرعت پایدار باقی می‌ماند. در بیشترین بار، میزان استفاده از پردازنده به حدود ۷۸٪ رسید (در حالی که مدل برت کلاسیک حدود ۹۹٪ پردازنده را اشغال می‌کند) و حافظه‌ی مصرفی نیز حداکثر ۱/۲ گیگابایت بود. نکته‌ی مهم آن است که تمام پردازش‌ها صرفاً با پردازنده انجام شد و نیازی به پردازنده گرافیکی نبود؛ این موضوع اقتصادی بودن و قابلیت استقرار آسان مدل را در محیط‌های صنعتی و ابری نشان می‌دهد.

در مقایسه، مدل برت کلاسیک برای طبقه‌بندی همین مجموعه به‌طور میانگین به ۱۱۰ ثانیه برای هر ۱۰۰۰ خبر نیاز داشت، در حالی که مدل ما تنها ۲/۱ ثانیه زمان صرف کرد؛ یعنی بیش از ۵۰ برابر سریع‌تر. حتی مدل حافظه کوتاه‌مدت بلند نیز با میانگین ۲۴۳ ثانیه، بسیار کندتر از مدل ترکیبی حاضر بود. این نتایج نشان می‌دهد چارچوب پیشنهادی توانسته است حدود ۹۵٪ از دقت مدل برت را با تنها ۲٪ از هزینه‌ی زمانی آن به دست آورد که بیانگر تعادل بی‌نظیر میان دقت، سرعت و بهره‌وری محاسباتی است.

نمودار بازدهی روش پیشنهادی، مقایسه‌ی دقت و نیازمندی‌های محاسباتی مدل‌های طبقه‌بندی خبر در شکل ۲ و ماتریس سردرگمی نتایج در شکل ۳ ترسیم شده است.



شکل ۲. ماتریس سردرگمی نتایج

جمع‌بندی و نتیجه‌گیری

عرصه‌ی در حال گسترش اخبار آنلاین، بستری غنی برای به‌کارگیری سامانه‌های طبقه‌بندی خودکار فراهم کرده است. در این پژوهش، سامانه‌ای مؤثر و کارآمد برای طبقه‌بندی اخبار طراحی و پیاده‌سازی شد که در آن الگوریتم بیز ساده با روش وزن‌دهی واژه‌ها برای استخراج ویژگی‌های متنی ترکیب گردید. این سامانه، با وجود ساختار ساده و سبک خود، توانست دقت بسیار بالایی در تفکیک اخبار به‌دست آورد؛ به‌طوری‌که دقت ۹۵/۱۲٪ در مجموعه‌ی آزمون نشان‌دهنده‌ی توانایی بالای آن در مواجهه با داده‌های واقعی بود.

نتایج پژوهش بیانگر آن است که علی‌رغم ظهور مدل‌های یادگیری عمیق پیچیده، همچنان می‌توان با ترکیب هوشمندانه‌ی روش‌های کلاسیک و فناوری‌های نوین، به عملکردی دست یافت که در برخی موارد حتی از مدل‌های عمیق نیز برتر است. استفاده از وزن‌دهی واژه‌ها موجب شد اهمیت نسبی واژه‌ها در کل مجموعه‌ی متون لحاظ گردد و مدل بتواند بر اساس واژگان کلیدی هر خبر، تصمیم‌گیری دقیقی انجام دهد. از سوی دیگر، هرچند بیز ساده در تئوری بر فرض استقلال ویژگی‌ها استوار است، اما نتایج ما نشان دادند که این مدل ساده‌تر نه تنها از نظر سرعت، بلکه از لحاظ دقت نیز از مدل‌های پیچیده‌تری چون شبکه عصبی پیچشی و حافظه کوتاه‌مدت بلند پیشی گرفته است. این امر تأکیدی است بر اینکه پیچیدگی بیشتر همواره به معنای دقت بالاتر نیست و در بسیاری از کاربردها، مدل‌های ساده‌تر در صورت مهندسی مناسب ویژگی‌ها می‌توانند نتایجی رقابتی ارائه دهند.

مزیت‌های عملیاتی و کاربردی

یکی از برجسته‌ترین نقاط قوت سامانه‌ی حاضر، مقیاس‌پذیری و سهولت استقرار آن است. در طراحی و پیاده‌سازی سیستم، برای بخش پشتیبان و واسط کاربری از چارچوب Django و برای آموزش مدل‌ها از کتابخانه‌ی Scikit-learn استفاده شد؛ این ترکیب منجر به ساخت سامانه‌ای کاربرپسند، توسعه‌پذیر و مناسب برای کاربردهای واقعی شد. مدل سبک‌وزن مبتنی بر بیز ساده امکان پردازش و پاسخ‌دهی در زمان واقعی را فراهم می‌کند، حتی در محیط‌هایی که فاقد شتاب‌دهنده‌های سخت‌افزاری مانند پردازنده گرافیکی هستند. این ویژگی به‌ویژه برای سازمان‌های خبری که با حجم عظیمی از داده‌ها سروکار دارند و در عین حال به دنبال راهکارهای کم‌هزینه و مقیاس‌پذیرند، ارزش فراوانی دارد. چنین سیستمی می‌تواند اخبار را به‌صورت بلادرنگ طبقه‌بندی کند، بدون آنکه نیاز به منابع محاسباتی سنگین داشته باشد.

محدودیت‌ها

با وجود نتایج درخشان، پژوهش حاضر از برخی محدودیت‌ها نیز برخوردار است. نخست، هرچند مدل ما فاصله‌ی عملکردی خود را با مدل‌های ترنسفورمری تا حد زیادی کاهش داده، هنوز حدود ۳/۲٪ از نظر دقت نسبت به مدل برت فاصله دارد (۹۵/۱٪ در برابر ۹۸/۳٪). هرچند این اختلاف اندک است، اما در برخی کاربردهای حساس می‌تواند اهمیت داشته باشد. دوم، مدل در تمایز میان برخی دسته‌های نزدیک به هم (نظیر فرهنگی و سرگرمی یا سیاسی و نظامی) با چالش مواجه شد. دلیل این مسئله، هم‌پوشانی مفهومی موضوعات خبری و وجود اصطلاحات مشترک است که گاه حتی برای انسان نیز موجب ابهام می‌شود. محدودیت دیگر آن است که مدل حاضر به‌صورت ایستا آموزش دیده است؛ بنابراین در مواجهه با موضوعات و ترندهای جدید خبری که در داده‌های آموزشی اولیه حضور نداشتند، ممکن است عملکرد آن اندکی افت کند.

پیشنهادهایی برای پژوهش‌های آینده

- بر اساس یافته‌ها و محدودیت‌های فوق، مسیرهای زیر برای توسعه‌ی آتی پیشنهاد می‌شود:
- یکپارچه‌سازی بهینه‌ی مدل‌های ترنسفورمری: هرچند در این پژوهش از نسخه‌ی فشرده‌ی برت (DistilBERT) استفاده شد، می‌توان با به‌کارگیری روش‌هایی مانند کوانتیده‌سازی یا دانش‌بانی^۱، نسخه‌های سبک‌تر اما غنی‌تر از مدل‌های ترنسفورمری را به‌صورت مستقیم در فرایند طبقه‌بندی دخیل کرد تا راهنمایی بافتی دقیق‌تری ارائه شود بدون آنکه هزینه‌ی محاسباتی به‌طور چشمگیری افزایش یابد. این راهکار می‌تواند فاصله‌ی باقی‌مانده‌ی دقت را جبران کند.
 - طراحی معماری طبقه‌بندی سلسله‌مراتبی: نتایج نشان دادند بخش عمده‌ای از خطاها در دسته‌های معنایی هم‌خانواده رخ می‌دهد. لذا می‌توان از معماری دوسطحی بهره گرفت که در سطح نخست دسته‌های کلی‌تر (سیاسی، اقتصادی، فرهنگی و ...) و در سطح دوم زیردسته‌های تخصصی‌تر (مانند سیاست داخلی/خارجی یا اقتصاد کلان/خرد) شناسایی شوند. این ساختار سلسله‌مراتبی احتمالاً موجب کاهش خطاهای ناشی از مشابهت موضوعی خواهد شد (Dalaorao et al., 2019).
 - افزودن مکانیزم‌های تطبیق پویا و یادگیری افزایشی: برای حفظ پایداری و به‌روزرسانی مداوم مدل در برابر تغییرات سریع دنیای خبر، می‌توان از رویکردهای یادگیری افزایشی بهره گرفت تا مدل با داده‌های جدید تطبیق یابد بدون نیاز به آموزش مجدد کامل. به‌کارگیری وزن‌دهی پویا برای موضوعات داغ^۲ و ترکیب داده از منابع خبری متنوع نیز می‌تواند دقت و تعمیم‌پذیری مدل را بهبود بخشد.
- اجرای این بهبودها می‌تواند دقت سامانه را به سطح مدل‌های عمیق پیشرفته نزدیک‌تر کند، در حالی که زمان پاسخ‌گویی کوتاه (کمتر از ۳ ثانیه برای هر ۱۰۰۰ خبر) حفظ شود.

جمع‌بندی کلی

نتایج این پژوهش نشان دادند که رویکرد ترکیبی پیشنهادی توانسته شکاف میان مهندسی ویژگی سنتی و مدل‌های مبتنی بر ترنسفورمر را تا حد زیادی پر کند. این راهکار، توازن مؤثر میان دقت، سرعت و هزینه‌ی محاسباتی برقرار کرده است و می‌تواند به‌عنوان الگویی عملی برای سازمان‌های خبری و رسانه‌هایی که به دنبال سامانه‌های مقیاس‌پذیر و کم‌هزینه برای مدیریت حجم عظیم محتوای آنلاین هستند، مورد استفاده قرار گیرد. امید می‌رود با توسعه‌ی مسیرهای پیشنهادی، نسل جدیدی از سامانه‌های طبقه‌بندی خبر پدید آید که هم‌زمان از نظر دقت، سرعت و پایداری در بالاترین سطح ممکن عمل کنند و گامی مؤثر در جهت ارتقای پردازش هوشمند اطلاعات در عصر داده‌های کلان بردارند.

منابع

- Assefi, M., Bahrami, M., Arora, S., Taha, T. R., Arabnia, H. R., Rasheed, K. M., & Chen, W. P. (2022). An intelligent data-centric web crawler service for API corpus construction at scale. In *2022 IEEE International Conference on Web Services (ICWS)* (pp. 385–390). IEEE.
- Chy, A. N., Seddiqui, M. H., & Das, S. (2014). Bangla news classification using Naive Bayes classifier. In *16th International Conference on Computer and Information Technology* (pp. 366–371). IEEE.
- Dalaorao, G. A., Sison, A. M., & Medina, R. (2019). Integrating collocation as TF-IDF enhancement to improve classification accuracy. In *2019 IEEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA)* (pp. 282–285). IEEE.

1. knowledge distillation
2. hot topics

- Das, M., Kamalanathan, S., & Alphonse, P. J. A. (2023). A comparative study on TF-IDF feature weighting method and its analysis using unstructured dataset. *arXiv*.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.
- Dong, X. (2023). *Knowledge-enhanced news recommendation*.
- Faye, G., Icard, B., Casanova, M., Chanson, J., Maine, F., Bancelhon, F., Gadek, G., Gravier, G., & Égré, P. (2024). Exposing propaganda: An analysis of stylistic cues comparing human annotations and machine classification. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language* (pp. 62–72). Association for Computational Linguistics.
- Guleria, P. (2025). NLP-based text classification using TF-IDF enabled fine-tuning approach. *Procedia Computer Science*, 236, 125–136.
- Gupta, A., Batla, A., Kumar, C., & Jain, G. (2023). Comparative analysis of machine learning models for fake news classification. In *2023 3rd International Conference on Intelligent Technologies (CONIT)* (pp. 1–5). IEEE.
- Gupta, R., et al. (2023). Explainable AI for news classification: A case study on BERT and LIME. *Journal of Machine Learning Research*, 24(3), 123–145.
- Harditya, M., & Purnamasari, P. D. (2025). Machine learning vs. transformer: Indonesia news classification. In *ICCIP 2024: 2024 the 10th International Conference on Communication and Information Processing* (pp. 699–704). Association for Computing Machinery.
- Huang, T., Xu, Z., Yu, P., Yi, J., & Xu, X. (2025). A hybrid transformer model for fake news detection: Leveraging Bayesian optimization and bidirectional recurrent unit. *arXiv*.
- Kakar, S., Dhaka, D., & Mehrotra, M. (2023). Tweet recommendation using clustered BERT and Word2Vec models. In *2023 International Conference on Smart Applications, Communications and Networking (SmartNets)*. IEEE.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1746–1751). Association for Computational Linguistics.
- Kumar, S., et al. (2023). News classification in low-resource languages: Challenges and solutions. In *Proceedings of the 2023 Conference on Computational Linguistics (COLING)* (pp. 789–801).
- Kumar, S., et al. (2024). CPU-optimized transformer inference. *ACM Transactions on Architecture and Code Optimization*, 19(2).
- Lestari, N., Riza, O. S., & Ardinal, R. (2023). Implementation of text mining and pattern discovery with Naive Bayes algorithm for classification of text documents. *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*, 14(1), 88–102.
- Li, B.-H., Xiang, Y.-X., Feng, D., He, Z.-C., Wu, J.-J., Dai, T.-L., & Li, J. (2022). Short text classification model combining knowledge aware and dual attention. *Journal of Software*, 33(10), 3565–3581.
- Li, N., Wang, Y., & Huang, L. (2023). Chinese text sentiment classification based on LSTM word-tag relational network model. In *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)* (pp. 783–786). IEEE.
- Liu, G., & Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337, 325–338.
- Liu, Z., Lv, X., Liu, K., & Shi, S. (2010). Study on SVM compared with the other text classification methods. In *2010 Second International Workshop on Education Technology and Computer Science* (pp. 219–222). IEEE.
- Masri, Y., Wang, Z., Srengnathan Malarvizhi, A., Ahmed, S., Stover, T., Wong, D. W. S., Jiang, Y., Li, Y., Liu, Q., Bere, M., Rothbart, D., Pfoser, D., & Yang, C. (2025). Comparative analysis of BERT and GPT for classifying crisis news with Sudan conflict as an example. *Algorithms*, 18(7), 420.
- Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 412–418). Association for Computational Linguistics.
- Ohno, S. (1970). *Evolution by gene duplication*. Springer-Verlag.
- Rong, X. (2014). *Word2vec parameter learning explained*. *arXiv*.

- Rong, X. (2023). Dimensionality reduction in NLP: Trade-offs. *Neural Processing Letters*, 55(1), 1–15.
- Shabani, V., Havolli, A., Maraj, A., & Fetahu, L. (2023). Fake news detection using Naive Bayes classifier and passive aggressive classifier. In *2023 12th Mediterranean Conference on Embedded Computing (MECO)* (pp. 1–6). IEEE.
- Vaswani, A., et al. (2023). Interpretable feature selection with BERT. *Journal of Artificial Intelligence Research*, 78(2), 345–367.
- Vaswani, A., et al. (2023). Transformer-based models for multimodal news classification: A comprehensive review. *Journal of Artificial Intelligence Research*, 78(2), 345–367.
- Waheed, S. R., Rahim, M. S. M., Suaib, N. M., & Salim, A. A. (2023). CNN deep learning-based image to vector depiction. *Multimedia Tools and Applications*, 82, 20283–20302.
- Wang, Y., et al. (2023). BERT for news classification: A comparative study on fine-tuning strategies. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1123–1135).
- Yang, Z., et al. (2023). Hierarchical news classification with hierarchical attention networks. In *Proceedings of the 2023 Conference on Neural Information Processing Systems (NeurIPS)* (pp. 345–357).
- Zhang, D., & Mo, J. (2025). LinguaSynth: Heterogeneous linguistic signals for news classification. *arXiv*.
- Zhang, H.-F., Zeng, C., Pan, L., Hao, R., Wen, C., & He, P. (2022). News topic text classification method based on BERT and feature projection network. *Journal of Computer Applications*, 42(4), 1116–1124.
- Zhang, X., et al. (2023). Cross-lingual news classification using multilingual BERT. *IEEE Transactions on Neural Networks and Learning Systems*, 34(5), 789–801.
- Zhao, W., et al. (2023). News classification with contrastive learning: A self-supervised approach. In *Proceedings of the 2023 Conference on Web Search and Data Mining (WSDM)* (pp. 456–468).

