Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال دوم، شماره اول، بهار و تابستان ۱۴۰۳، صفحات ۱۵۴ - ۱۴۲

تشخیص خودکار متون تولیدشده توسط ماشین مبتنی بر مدل‌های زبانی و یادگیری عمیق

ابوالفضل شیرافکن*، مهدی نقوی، میثم میرزایی

دانشکده و پژوهشکده هوش مصنوعی و علوم شناختی، دانشگاه جامع امام حسین (ع)، تهران، ایران

مقاله پژوهشی

تاریخ دریافت:
۱۴۰۳/۰۹/۰۶

تاریخ پذیرش:
۱۴۰۳/۱۱/۰۲

کلیدواژه‌ها:
یادگیری عمیق، مدل مبتنی‌بر سبک نوشتار، RoBERTa، BiLSTM

نویسنده‌مسئول:
Abolfazlshirafkan00@gmail.com

چکیده

امروزه با رشد چشمگیر هوش مصنوعی و تولیدات آن، فرصت‌ها و تهدیدات زیادی به وجود آمده است. یکی از معروف‌ترین و محبوب‌ترین تولیدات هوش مصنوعی تولید متن است که به آن متن ماشینی نیز گفته می‌شود. در این پژوهش روش جدیدی معرفی می‌شود که ویژگی‌های استخراج‌شده از متن را با ویژگی‌های ساختاری آن ترکیب کرده و به این ترتیب برای تشخیص متن نوشته شده توسط انسان و متن تولیدی هوش‌مصنوعی، یک متمایزگر خودکار ایجاد می‌کند. روش معرفی شده متشکل از دو بخش است، بخش اول: مدل توسعه یافته BERT (RoBERTa) و مدل حافظه‌ای کوتاه مدت بلند مدت دو سویه (BiLSTM) است که با لایه ادغام بهبود یافته‌اند. بخش دوم: ویژگی‌های ساختاری متن با روشی مبتنی بر سبک نوشتار استخراج می‌شود. در نهایت خروجی بخش‌های مدل باهم ترکیب شده و به این ترتیب مدل متن نوشته شده توسط انسان را از متن تولیدشده توسط ماشین تشخیص می‌دهد. نتایج حاصل از این پژوهش نشان می‌دهد که روش پیشنهادی با دقت ۹۰ درصدی قابلیت تشخیص متون ماشینی را داشته و عملکرد مطلوبی را از خود به نمایش می‌گذارد.

: 10.22034/abmir.2025.22455.1078

E-ISSN: 2821-2037 /© 2023. Published by Yazd University This is an open access article under the CC BY 4.0 License (https://creativecommons.org/licenses/by/4.0/).

۱- مقدمه

با پیشرفت هوش مصنوعی، در سال‌های اخیر شاهد رشد فراگیر هوش مصنوعی در ابعاد و زمینه‌های مختلفی بوده‌ایم. یکی از عرصه‌های بسیار محبوب و پرکاربرد که با استقبال و نگرانی‌های زیادی مواجه شده است، تولید محتوای خودکار و مصنوعی است. در حال حاضر مدل‌های مولد متن[1] در تولید متونی که با سبک نوشتار و زبان انسان مطابقت زیادی دارند؛ پیشرفت‌های قابل توجه کسب کرده‌اند. نمونه بسیار مشهور آن که اخیراً توجه بی‌سابقه‌ای به خود جلب کرده، ابزاری به نام ChatGPT است که توسط شرکت Open AI تولید و عرضه شده است. با توجه به تهدیدات و فرصت‌های بسیاری که این‌چنین تولیداتی ایجاد می‌کند، پژوهش حاضر از جهت‌های مختلف به بررسی این مدل‌های مولد می‌پردازد. با بررسی جامع تهدیدات این نوع مولدها، به مواردی دست‌یافته شد که برای یکپارچگی در مفهوم، آن‌ها را مدل‌سازی کرده و موارد زیر به‌عنوان مدل‌های تهدید در این زمینه معرفی می‌شود:

- تسهیل در ساخت نرم‌افزارهای مخرب و مهندسی اجتماعی مانند ساخت چت بات های مخرب، پیام‌های فیشینگ (ایمیل، پیامک، چت، رسانه‌های اجتماعی و..) و فیشینگ صوتی.
- استفاده از نویسندگی هوش مصنوعی مانند انتشارات علمی جعلی، درخواست‌های انبوه (بورسیه تحصیلی، نامه‌های پوششی و..)، تولید محتواهای ممنوعه در شبکه‌های اجتماعی و op-ed مقاله انبوه/ ارسالی.
- ایجاد کمپین‌های نفوذ آنلاین که می‌توان آن‌ها را به دو دسته الف) کمپین‌های نفوذ سیاسی مانند انتشار اطلاعات نادرست و جنگ اطلاعاتی و تبلیغات سیاسی خاص و ب) کمپین‌های نفوذ تجاری که شامل مواردی چون بررسی‌های تقلبی کالا و تبلیغ محصول ربات اجتماعی می‌باشند، دسته‌بندی کرد.
- ایجاد هرزنامه و آزار و اذیت. که از موارد آن می‌توان به اظهار نظر غیر واقعی زیر پست‌های اینستاگرامی و فرستادن پیام‌های مستقیم به صورت انبوه اشاره کرد.

با توجه به دانستن تهدیدات موجود، اهمیت مقابله با آن‌ها بسیار ملموس است. ساخت آشکارسازهایی که می‌توانند متن تولیدشده توسط این مولدها را از متن نوشته شده توسط انسان متمایز سازند، نقش حیاتی در جلوگیری از اینگونه سؤ استفاده‌ها خواهند داشت. در حال حاضر روش‌های مختلفی برای ساخت این آشکارسازها پیشنهاد شده است؛ اما اکثر آن‌ها دارای دقت کم و هزینه‌های محاسباتی بالا و از همه مهم‌تر تعمیم پذیری پایین می‌باشند[۱]. از میان روش‌های موجود، در تحقیق حاضر مدل‌های زبان عصبی هدف تحقیقات قرار گرفت؛ به این دلیل که رویکردهای تشخیص، مبتنی‌بر شبکه‌های عصبی، مؤثر بودن خود را در تشخیص متون تولیدشده توسط ماشین نشان داده‌اند. از طرفی روش‌های مبتنی بر ویژگی با وجود قدرت در ارائه ویژگی‌های متنوع که ممکن است در شبکه‌های مولد متخاصم به وجود بیاید؛ ممکن است در انتقال ویژگی‌های خاص در بین روش‌های تولید و روش‌های نمونه‌گیری ضعیف عمل کند. روش دیگر روش‌های آماری است که برای روشن شدن روندهای آماری گسترده‌تر، تعداد نمونه‌های بیشتری لازم دارد، نتایج تحقیقات گذشته نشان می‌دهد که روش‌های آماری، زمانی مؤثرتر هستند؛ که مجموعه‌های طولانی‌تری از متن در دسترس باشد و برای متن‌های کوتاه مانند توییت‌ها و نظرات کاربران و غیره که متن تولیدشده کوتاه است، عملکرد قابل قبولی ندارند. از دیگر روش‌هایی که می‌توان به آن اشاره کرد روش‌هایی است که حاصل همکاری انسان و ماشین است. که این روش‌ها علاوه بر پایین بودن دقت، کاری طاقت فرسا و هزینه‌بر هستند؛ از طرفی ما به دنبال خودکار کردن هرچه بیشتر الگوریتم‌ها هستیم و این روش‌ها ما را از این هدف دور می‌کنند. با توجه به اهمیت موضوع مطرح شده و بررسی کلی روش‌های موجود ، که برخی از کاستی‌های آن‌ها ذکر شد؛ در این پژوهش جهت تولید آشکارسازی که قابلیت تعمیم پذیری بالا و هزینه محسباتی پایینی دارد؛ از میان روش‌های موجود روش تشخیص عصبی مبتنی بر معماری‌های ترنسفورماتورهای دوطرفه برای توسعه و کار بیشتر انتخاب شده است؛ زیرا مزایای قابل ملاحظه‌ای نسبت به دیگر روش‌ها دارند. به

[1] Text generative models

این ترتیب این پژوهش با استفاده از روش‌های یادگیری عمیق، دقت در یادگیری روابط بین متون تولیدشده توسط ماشین و انسان را افزایش داده و تعمیم پذیری و دقت در تشخیص مدل معرفی شده را نسبت به مدل‌های پیشین، افزایش داده است.

۲- پیشینه تحقیق

بررسی روش‌های مختلف موجود در زمینه تشخیص متن، ما را بر آن داشت که از روش‌های تشخیص مبتنی بر شبکه عصبی در تحقیقات خود استفاده نماییم. از این رو مقالات و روش‌های این حوزه را مطالعه و مورد بررسی قرار می‌دهیم. و در جدول انتهایی روش‌ها را برای بررسی بهتر با یکدیگر مقایسه می‌کنیم.

کودوگونتا و همکاران [۲]، در سال ۲۰۱۸ در تحقیقی با عنوان «شبکه‌های عصبی عمیق برای تشخیص ربات» بر طبقه‌بندی در سطح محتوا تمرکز کرده‌اند. آن‌ها در این مقاله بر استفاده از متن توییت، از ابرداده‌های توییت مانند تعداد بازتوییت‌ها، تعداد پاسخ‌ها یا موارد دیگر، برای تقویت ویژگی‌های جاسازی GloVe[1] برای طبقه‌بندی بهتر استفاده کرده‌اند. آن‌ها یک لایه LSTM برای یادگیری ویژگی‌های متوالی متن توییت را به کار برده‌اند و قبل از اعمال لایه‌های طبقه‌بندی کاملاً متصل، آن را با ویژگی‌های ابرداده پیوند داده‌اند. آن‌ها همچنین یک امتیاز طبقه‌بندی را فقط با نمایش LSTM محاسبه کرده‌اند.

سلیمان و همکاران [۳]، در سال ۲۰۱۹ در مقاله‌ای با عنوان «استراتژی‌های انتشار و تاثیرات اجتماعی مدل‌های زبان» نوشته‌اند که مدل‌های زبانی طیف بزرگی از کاربردهای مفید را شامل می‌شوند. آن‌ها می‌توانند در نثر، شعر و برنامه‌نویسی کمک کنند. با این حال، انعطاف پذیری و قابلیت‌های تولیدی آن‌ها نگرانی‌هایی را در مورد سوء استفاده از این مدل‌ها ایجاد می‌کند. این گزارش به کار OpenAI، مربوط به انتشار مدل زبانی GPT-2[2] می‌پردازد. این انتشار مرحله‌ای را مورد بحث قرار داده است که اجازه می‌دهد با افزایش اندازه مدل، زمان بین انتشار مدل طولانی‌تر شود تا تجزیه‌وتحلیل ریسک و سود انجام شود.

روآن زلرز، آری هولتزمن و همکاران [٤]، در سال ۲۰۱۹ در تحقیقی تحت عنوان «دفاع در برابر اخبار جعلی عصبی» نشان دادند که پیشرفت‌های اخیر در تولید زبان طبیعی نگرانی‌هایی را برای استفاده‌های دو گانه ایجاد کرده است. در حالی که برنامه‌هایی مانند خلاصه‌سازی و ترجمه مثبت هستند، اما فناوری زیر بنایی ممکن است به دشمنان امکان تولید اخبار جعلی عصبی را بدهد. آن‌ها یک مدل تولید متن قابل کنترل به نام Grover ارائه کردند. آن‌ها با اعمال رویکرد شات صفر[3]، متن‌های تولیدشده توسط Grover را با مدل متمایزسازی که با همین مدل آموزش دیده شده بود را با دقت ۹۲ درصد تشخیص دادند. این در حالی بود که دیگر مدل‌ها در این مجموعه داده به دقت ۷۲ درصد رسیده بودند. این مدل، با وجود دقت بالا اما یک مشکل دارد! ما باید بدانیم که متن با کدام مدل تولیدشده؛ این در صورتی است که همیشه ما این دانش را نداریم. و این مورد ما را از استفاده کردن رویکردهای مبتنی بر شات صفر منصرف می‌کند. با این وجود آداکو اوچندو و همکاران [۵]، در سال ۲۰۲۰، برای حل این مشکل، انتساب نویسنده برای تولید متن عصبی را برای تشخیص متن ماشینی پیشنهاد دادند. که این اقدام آن‌ها در مواجه با مولدهای پیشرفته مانند GPT-2 و Groverعملکرد مطلوبی ندارد.

دیوید دوکیچ و همکاران [٦]، درسال ۲۰۲۰؛ در پژوهشی دیگر با عنوان «آیا شما انسان هستید؟» به شناسایی ربات‌های توییتر بر اساس متن منتشرشده توسط آن‌ها، با استفاده از مدل BERT[4] پرداخته‌اند. نویسنده‌گان در این مقاله، روی مجموعه داده" PAN Author Profiling" کار نمودند تا توییت‌های تولیدشده توسط ربات‌ها را شناسایی کنند. آن‌ها جهت به دست آوردن جاسازی متنی مناسب در توییت‌ها از مدل BERTBase از پیش آموزش دیده شده استفاده کردند؛ و آن را با جاسازی emoji2vec و چند ویژگی باینری پیوند دادند؛ تا با طبقه‌بندی کننده رگرسیون لجستیک یا طبقه‌بندی کننده شبکه عصبی عمیق تغذیه شود. آن‌ها با استفاده از این معماری درمعیار F1، به امتیاز ۸۳,۳۵ در کار تشخیص ربات دست یافتند.

[1] Global Vectors for Word Representation
[2] Generative Pre-trained Transformer
[3] Zero-shot
[4] Bidirectional Encoder Representations from Transformers

آدلانی و همکاران [۷]، در سال ۲۰۲۰ در پژوهشی دیگر تحت عنوان «ایجاد نظرات آنلاین جعلی برای حفظ احساسات با استفاده از مدل‌های زبانی عصبی و تشخیص مبتنی بر انسان و ماشین» مدل‌های زبان موجود را برای ایجاد نظرات جعلی با احساسات دلخواه برای شرکت‌های آمازون و یلپ[1] ترکیب کردند. آن‌ها عملکرد خوانندگان انسانی و طبقه‌بندهای مبتنی بر یادگیری ماشین را، در تشخیص ماشینی بودن متون نظرسنجی بررسی کرده‌اند. یافته‌های آن‌ها نشان می‌دهد که عملکرد خوانندگان انسانی در تشخیص صحیح ماشینی بودن متون ٥٤٪ بوده که تقریباً معادل شانسی بودن آن است. باوجود اینکه سازوکار تشخیص یادگیری ماشین، عملکرد بهتری از انسان‌ها داشته است، ولی هنوز به پیشرفت‌های بسیار بیشتری نیاز دارند.

ایپولیتو و همکارانش [۸]، در سال ۲۰۲۰ در تحقیقی با عنوان «تشخیص انسانی و خودکار متن تولیدشده» بر مقایسه انسان و ماشین در تشخیص متون جعلی عمیق[2] تمرکزکرده‌اند. آن‌ها ارزیابی‌های خود را بر اساس متن تولیدشده توسط GPT-2 انجام می‌دهند و از BERT به‌عنوان مدل تشخیص دهنده اولیه استفاده نمودند. آن‌ها نوشتند که از آنجایی که مدل‌های تولیدکننده متن برای فریب دادن انسان‌ها آموزش داده‌شده‌اند، با وجود موفقیت در دستیابی به این هدف، ناهنجاری‌هایی را معرفی کردند که کار تشخیص را برای متمایزکنندگان خودکار آسان می‌کند. آزمایشات آن‌ها همچنین نشان داد که تشخیص متون جعلی در صورتی که متن کوتاه باشد دشوارتر است.

خوان رودریگز و همکاران [۹]، در سال ۲۰۲۲، تشخیص متقابل دامنه متن فنی GPT-2-Generated را معرفی کردند. آن‌ها در این مقاله مشکل تشخیص متن تحقیقات فنی تولیدشده توسط GPT-2 را مورد بررسی قرار دادند. آن‌ها ابتدا سناریوی واقع بینانه‌ای را در نظر گرفتند که در آن مدافع، اطلاعات کاملی در مورد روش اصلی تولید متن توسط حریف ندارد، اما می‌تواند مقادیر کمی از متن واقعی و مصنوعی درون دامنه را به منظور تطبیق با توزیع هدف برچسب گذاری کند. حتی در سناریوی سخت‌تر مانند تطبیق یک آشکارساز حوزه فیزیک با یک آشکارساز زیست پزشکی، متوجه شدند که تنها چند صد برچسب برای عملکرد خوب کافی است. در نهایت، نشان دادند که آشکارسازهای سطح پاراگراف را می‌توان برای تشخیص دستکاری اسناد تحت انواع مدل‌های تهدید استفاده کرد. آن‌ها در این مقاله، مشکل تشخیص خودکار متن فنی تولیدشده توسط GPT-2 را بررسی کردند و دریافتند که آشکارسازهای مبتنی بر RoBERTa می‌توانند با موفقیت از یک رشته علمی (فیزیک) به رشته دیگر (زیست پزشکی) تطبیق داده شوند، که به مقادیر نسبتاً کمی از داده‌های برچسب گذاری شده در دامنه نیاز دارد.

رویکرد تنظیم دقیق[10]، رویکردی پیشرفته برای تشخیص عصبی متن تولیدشده توسط ماشین است. روش‌های تشخیص عصبی مبتنی بر معماری‌های ترنسفورمر[3] دو طرفه در حال حاضر نشان‌دهنده پیشرفته‌ترین مجموعه داده‌های ارزیابی GPT-2 هستند. یک روند کلی به سمت افزایش استفاده از معماری ترنسفورمر دو طرفه، به جای معماری ترنسفورمر یک طرفه، به‌ویژه RoBERTa وجود دارد، که در ادامه برخی از تحقیقات در این زمینه را بیان می‌کنیم.

جیکوب دولین و همکاران [۱۱]، در سال ۲۰۱۸ مدل معروف BERT را که پیش آموزش ترنسفورماتورهای دو جهته عمیق است معرفی کرده‌اند. مدل BERT از پیش آموزش دیده را می‌توان تنها با یک لایه خروجی اضافی تنظیم کرد تا مدل‌های پیشرفته‌ای را برای طیف وسیعی از وظایف، مانند پاسخ گویی به سؤال و استنتاج زبان، بدون تغییرات اساسی در معماری خاص ایجاد کند. BERT از نظر مفهومی ساده و از نظر تجربی قدرتمند است. نتایج پیشرفته‌ای را در یازده وظیفه پردازش زبان طبیعی به دست آورده است، از جمله افزایش امتیاز GLUE[4] به ۸۰,۵٪ (۷,۷٪ امتیاز بهبود)، دقت MultiNLI به ۸٦,۷٪، SQuAD v1.1. پاسخ به سؤال تست F1 تا ۹۳,۲ و SQuAD v2.0 تست F1 تا ۸۳,۱.

[1] Yelp
[2] Deep fake
[3] Transformer
[4] General Language Understanding Evaluation

در سال ۲۰۱۹ یینهان لئو و همکاران[۱۲]، سعی کردند مدل BERT را بهبود بخشند. آن‌ها نام مدل خود را رویکرد پیش‌آموزشی BERT بهینه شده قوی (RoBERTa[1]) نهادند. آن‌ها باور داشتند که پیش آموزش مدل زبان، به دست آوردهای عملکردی قابل توجه منجر می‌شود، ولی مقایسه دقیق بین رویکردهای مختلف هزینه محاسباتی بالایی دارد؛ آن‌ها نشان دادند که انتخاب هایپرپارامترها تأثیر زیادی بر روی نتایج نهایی می‌گذارد. آن‌ها مطالعه همانندسازی شده‌ای از BERT ارائه کردند؛ که تأثیر بسیاری از هایپرپارامترهای کلیدی و اندازه داده‌های آموزشی را به‌دقت اندازه‌گیری می‌کند. آن‌ها تغییراتی چون: حذف هدف پیش‌بینی جمله بعدی؛ آموزش در بخش‌های طولانی‌تر؛ و تغییر پویای الگوی پوشش اعمال شده بر روی داده‌های آموزشی، را در مدل BERT اعمال کردند و روش پیش آموزشی بهبودیافته BERT، که RoBERTa نامیدند را به نتایج پیشرفته‌ای در GLUE، RACE و SQuAD رساندند.

۳- روش پیشنهادی

در این بخش به بررسی روش پیشنهادی جهت تشخیص متن تولیدشده توسط ماشین از متن نوشته‌شده به‌وسیله انسان پرداخته‌شده است. ابتدا مجموعه‌داده‌ای که آموزش و آزمایش مدل بر روی آن انجام‌شده، موردبررسی قرارگرفته و پیش‌پردازش داده‌ها شرح داده‌شده است، سپس نحوه تقسیم‌بندی داده‌های این مجموعه‌داده بیان شده است؛ همچنین در ادامه این بخش به بیان و توضیح روش پیشنهادی پرداخته‌شده و روش کلی تشخیص به طور کامل بررسی می‌شود.

۳-۱ مجموعه‌داده

انتخاب مجموعه‌داده استفاده‌شده در این پژوهش بر اساس نیاز موجود در زمینه تشخیص متن‌های کوتاه، صورت گرفت؛ همچنین یک عامل مهم در انتخاب مجموعه‌داده در این پژوهش، کیفیت متن و شباهت توییت‌های تولیدشده توسط ربات به توییت‌های تولیدشده توسط انسان بود. به‌عبارت‌دیگر، یک مجموعه‌داده متنی عمیق جعلی که شامل نمونه‌هایی از متن است که توسط مدل‌های پیشرفته اخیر در زمینه تولید متن ماشینی، تولیدشده‌اند. در حوزه متن جعلی عمیق، مدل‌های زبانی مانند GPT، RNN، LSTM، GROVER و غیره به قابلیت تولید متن با کیفیت بالا رسیده‌اند و برخی مطالعات گزارش می‌دهند که میزان تشخیص انسان در برابر این نمونه‌های متن تقریباً شانسی است[۱۳, ۱۴]، مجموعه‌داده انتخابی، مجموعه‌داده « TweepFake » است که توسط تیزیانوفاگنی و همکاران [15]، در سال ۲۰۲۱ منتشر شد. این مجموعه‌داده شامل ۲۵۵۷۲ توییت است که ۱۲۷۸۶ توییت مربوط به توییت‌های نوشته شده توسط انسان و ۱۲۷۸۶ توییت نیز به متن‌های تولیدشده توسط ربات اختصاص دارد. متن‌های موجود در این مجموعه‌داده از ۲۳ ربات و ۱۷ حساب کاربری انسانی جمع‌آوری‌شده است. ربات‌های تولیدکننده متن در این مجموعه‌داده از سه فناوری اصلی برای تولید متن خود استفاده کرده‌اند. ۱۱ ربات که ۳۸۶۱ توییت زده‌اند، از فناوری GPT-2، ۷ ربات با مجموع توییت ۴۱۸۱ از فناوری تولید متن RNN و ۵ ربات دیگر نیز با مجموع ۴۸۷۶ توییت تولیدشده، از دیگر روش‌های تولید متن مانند: زنجیره مارکوف، RNN + زنجیره مارکوف، LSTM و CharRN استفاده نموده‌اند.

۳-۲ پیش‌پردازش داده‌ها

برای آماده‌سازی مجموعه‌داده جهت ورود به مدل، نیاز است داده‌ها پیش‌پردازش شوند. ازآنجاکه هدف در این پژوهش، تشخیص متن نوشته شده توسط انسان از متن تولیدشده توسط ماشین است؛ پیش‌پردازش بیش از حد می‌تواند دقت در تشخیص را تحت‌الشعاع قرار دهد چرا که برخی از موارد موجود در متن‌های ماشینی و انسانی خود می‌توانند شاخصی برای تشخیص باشند. به‌عنوان مثال غلط‌های رایجی که در دست نوشته‌های انسانی وجود دارد، در صورت اصلاح، کار را برای تشخیص و شناسایی متن نوشته شده توسط انسان سخت می‌کند. در این پژوهش از مجموعه‌داده Tweepfake استفاده‌شده است که در سال ۲۰۲۱ منتشر شد. آن‌ها مجموعه‌داده خود را جهت تحقیقات بیشتر در سایت kaggle قرار دادند. این فایل منتشرشده دارای نواقصی است که برای آماده سازی آن نیاز به پیش‌پردازش‌هایی بر روی داده‌های مجموعه‌داده است.

[1] Robustly Optimized BERT pretraining Approach

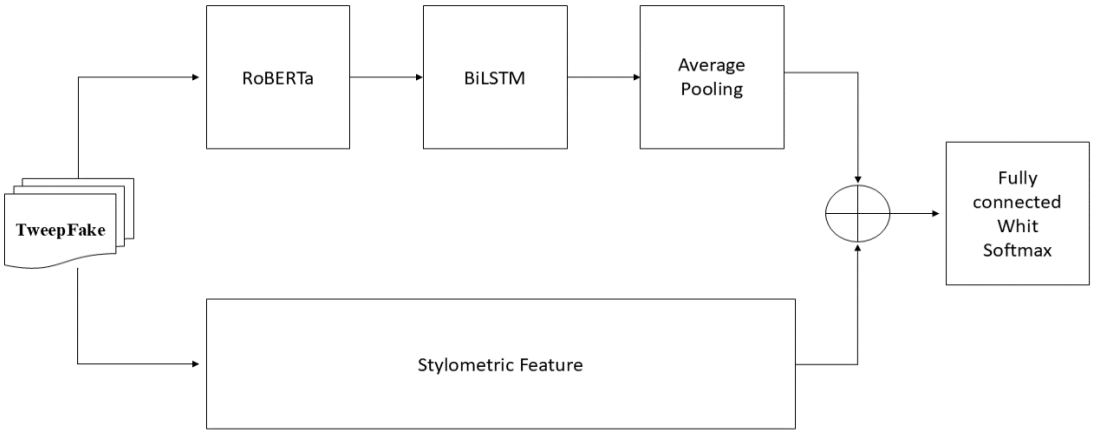
در این مجموعه‌داده، سطرهایی وجود دارند که خالی بوده و یا برچسب گذاری نشده است. از این رو در این پژوهش ابتدا این مجموعه‌داده پاک‌سازی شده و سپس برای ورود به مدل تقسیم‌بندی شده است.

۳-۳ تقسیم‌بندی مجموعه‌داده

در این پژوهش داده‌های موجود در مجموعه‌داده به ۳ دسته: داده‌های آموزش، داده‌های آزمایش و داده‌های ارزیابی، تقسیم‌بندی شده‌اند. به این صورت که ابتدا ۹۰ درصد داده‌ها به داده‌های آموزش اختصاص پیداکرده و ۱۰ درصد باقی‌مانده به‌عنوان داده‌های آزمایش درنظر گرفته شده است. سپس ۱۰ درصد از داده‌های آموزش از مجموعه داده‌های آموزشی جداشده تا مجموعه‌داده ارزیابی را ایجاد کنند.

۳-٤ مدل پیشنهادی

در این پژوهش، ابتدا از میان مجموعه‌داده‌های موجود، مجموعه‌داده مناسب و استانداردی برای پژوهش انتخاب شد، مجموعه‌داده انتخابی از نظر محتوایی پاک‌سازی شد و سپس مجموعه‌داده به گونه‌ای از هم تفکیک شد که هیچ کاربر مشترکی بین زیر مجموعه آموزش و تست وجود نداشته باشد؛ این کار به این جهت انجام گرفت که فقط با گرفتن یک متن از یک کاربر، تشخیص نویسنده انسان از ماشین انجام پذیرد. در مرحله آموزش از مدل توسعه یافته BERT یعنی RoBERTa که به نتایج قابل قبول و مطلوبی دست‌یافته است؛ جهت آموزش مدل مورد استفاده قرار گرفت. در این مرحله، آموزش در سطح کارکتر به جای کلمه انجام شد زیرا ممکن است متون از نشانه‌هایی مانند ایموجی و غیره استفاده کرده باشند و این اقدام یعنی استفاده از token به جای word باعث می‌شود این نشانک‌ها در مرحله آموزش و آزمایش، پوشش داده شوند. در قدم بعدی خروجی مدل به لایه‌ای بر پایه BiLSTM متصل می‌شود، تا وابستگی زمانی میان اجزاء مختلف جمله افزایش پیدا کند؛ سپس لایه ادغام مناسبی برای بهبود نتایج و جلوگیر از بیش برازش در نظر گرفته می‌شود. از آن‌سو ویژگی‌های موجود در داده‌ها با استفاده از یک روش مبنی بر سبک سنجی نوشتار، از آن‌ها استخراج‌شده و با ویژگی‌هایی که از مدل قبلی (RoBERTa+BiLSTM+Pooling) به دست آمد، ترکیب و سپس به لایه بعدی که شامل چند لایه کاملاً متصل به همراه طبقه بند softmax است، برای دسته‌بندی و تشخیص نهایی، داده شد تا طبقه‌بندی نهایی صورت پذیرد. ساختار کلی روش پیشنهادی در شکل (۱) به‌وضوح قابل مشاهده است.



شکل (۱): ساختار کلی مدل پیشنهادی

در این بخش به بیان فنی مسئله پرداخته شده است. مجموعه‌داده انتخابی ابتدا پیش‌پردازش شد و سپس داده‌ها به شکل مناسبی برای آموزش و آزمایش تقسیم‌بندی شدند. درنهایت مجموعه داده به مدل دوبخشی پیشنهادی واردشده و شبکه آموزش لازم را دیده وبا داده‌های آزمایش، میزان دقت مدل پیشنهادی ارزیابی شد.

٤- نتایج و ارزیابی

در این بخش، محیط اجرایی مدل پیشنهادی برای تشخیص متن نوشته شده توسط انسان و متن تولیدشده توسط هوش مصنوعی بررسی‌شده است. در این بررسی، مشخصات نرم‌افزاری و سخت‌افزاری این محیط ارائه شده است. در ادامه مدل پیشنهادی با استفاده از معیارها ارزیابی معروف، ارزیابی‌شده، و در پایان مقایسه‌ای بین عملکرد مدل پیشنهادی و مدل‌های موجود انجام شده است.

٤-١ سخت‌افزار و محیط پیاده‌سازی مدل

برای پیاده‌سازی مدل پیشنهادی از سیستمی با مشخصات سخت‌افزاری زیر استفاده‌شده است:

- **GPU**: GEFORCE RTX 4060
- **CPU**: Core™ i7-13700K
- **RAM**: G64

زبان برنامه‌نویسی، جهت پیاده‌سازی این مدل پایتون بوده و این مدل در محیط ژوپیتر کد نویسی شده است.

٤-٢ نتایج مدل پیشنهادی

در ادامه، داده‌های به دست آمده از مدل پیشنهادی در قالب داده‌های دو کلاسه و چهار کلاسه، مورد بررسی قرار می‌گیرند.

٤-٢-١ نتایج مدل در دو کلاس:

برای بررسی نتایج مدل پیشنهادی و مقایسه آن با کارهای مشابه، داده‌ها در دو کلاس تقسیم‌بندی شده، و به مدل آموزش داده می‌شود. در این تقسیم‌بندی داده‌ها به دو کلاس ربات و انسان تقسیم‌بندی می‌شوند و مدل برای تشخیص اینکه متن را ربات نوشته یا انسان، آموزش می‌بینند. مشخصات هایپرپارامترهای اعمالی برای آموزش و تست مدل، در جدول (۱) نمایش داده‌شده است.

جدول (۱): هایپرپارامترهای اعمالی بر مدل (دو کلاسه)

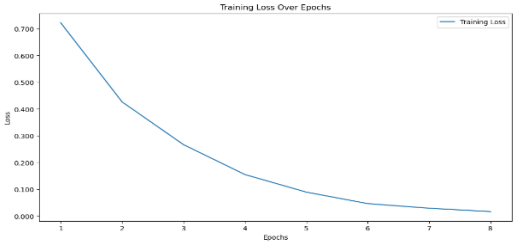
Hyperparameter	Value
Num.of training epochs	**8**
Initial learning rate	**10^{-5}**
Batch size	**2**
Dropout rate	**0.3**
Embedding size for RoBERTa	**768**
Max tokens to use	**512**

مقادیر به دست آمده از مدل پیشنهادی در مواجهه با داده‌هایی که به دو کلاس تقسیم شدن در جدول (۲) ارائه شده است. این مقادیر قدرت تشخیص مدل پیشنهادی را جهت تمایز بین متنی که ربات تولید کرده و متنی که توسط انسان نوشته شده است را به نمایش می‌گذارد.

نتایج به دست آمده از مدل، نشان می‌دهد که این مدل، دقت مطلوبی در تشخیص متن تولیدشده توسط ماشین از متن نوشته شده توسط انسان دارد.

جدول (۲): نتایج مدل پیشنهادی

Model	human			bot			Accuracy
ours	precision	recall	f1-score	precision	recall	f1-score	**0/90**
	0/91	0/87	0/89	0/89	0/92	0/90	

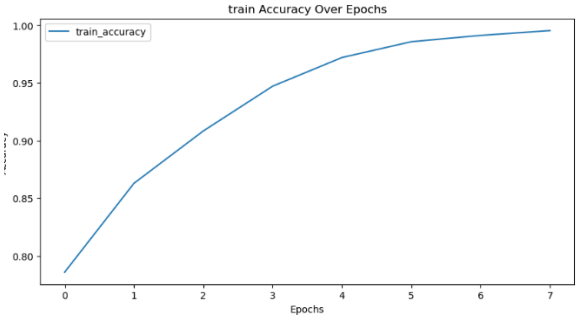


نمودار (۱): خطای آموزش در هر تکرار

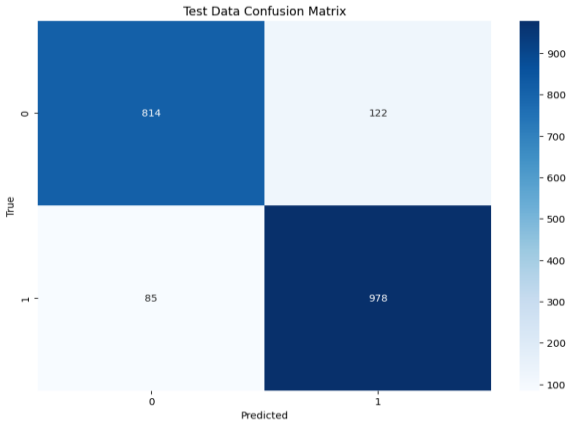
نمودار (۱) مقادیر خطای آموزش در هر تکرار را نمایش می‌دهد. بررسی نتایج نشان می‌دهد که با هر تکرار خطای آموزش کاهش‌یافته و مقادیر داده‌های آموزش به‌خوبی در مدل آموزش می‌بینند. همچنین نمودار زیر که بیانگر صحت در هر تکرار آموزش است، یادگیری مطلوب در مدل پیشنهادی را تأیید می‌کند.

در بحث دسته‌بندی یک مجموعه‌داده با استفاده از روش‌های مختلف دسته‌بندی، هدف دستیابی به بالاترین دقت و صحت ممکن در دسته‌بندی و تشخیص دسته‌ها است. در برخی از مسائل ممکن است تشخیص صحیح نمونه‌های مربوط به یکی از دسته‌ها برای ما اهمیت بالاتری داشته باشد. برای مثال در پژوهش حاضر که مربوط

به تشخیص متون تولیدشده توسط ربات و متمایزسازی آن با متنی که توسط انسان نوشته شده، تشخیص درست متن ماشینی از اهمیت بالاتری برخوردار است، زیرا با توجه به آنچه که در اهداف تحقیق بیان شد، متون تولیدشده توسط ماشین می‌توانند بسیار مخرب و تأثیرگذار ظاهر شوند، درنتیجه تشخیص درست آن‌ها از اهمیت بالاتری برخوردار است. در چنین مواقعی که دقت و صحت تشخیص یک دسته در مقایسه با دقت و صحت کلی، اهمیت بیشتری دارد، ماتریس درهم‌ریختگی می‌تواند به ما کمک کند تا بررسی بهتری از عملکرد مدل داشته باشیم. در ادامه ماتریس درهم‌ریختگی مدل پیشنهادی ترسیم شده است.



نمودار (۲): صحت آموزش مدل در هر تکرار



شکل (۲): ماتریس سردرگمی مدل پیشنهادی

با بررسی این ماتریس، مشاهده می‌شود همان‌طور که تشخیص متون تولیدشده توسط ماشین، برای ما از اهمیت بالاتری برخوردار است، این مدل قابلیت تشخیص بالاتری در این متون دارد و نیاز و هدف این پژوهش را تأمین می‌نماید.

- **توجه:** در تصویر فوق کلاس ۰ مربوط به کلاس انسان و کلاس ۱ مربوط به کلاس ربات است.

در ادامه به بررسی نتایج حاصل از آموزش مدل پیشنهادی، بر روی داده‌های کلاس‌بندی شده در قالب چهار کلاس پرداخته شده است. هدف از این دسته‌بندی ایجاد توان مقایسه نتایج حاصل از تشخیص مدل، برای هر یک از مدل‌های تولید متن ماشینی، است. این مقایسه به‌خوبی نشان می‌دهد که مدل پیشنهادی در مواجهه با هر یک از مدل‌های مولد متنی، چقدر توانایی تشخیص دارد؛ بدهی است که هر چقدر مدل مولد متن قوی‌تر باشد تشخیص سخت‌تر شده و دقت پایین‌تر می‌آید. مدل پیشنهادی اما با وجود نزدیک‌تر شدن متن تولیدشده توسط مدل مولد متن GPT-2 و RNN به متن نوشته شده توسط انسان، عملکردی بسیار مطلوب در تشخیص متون این مولدها داشته و در پایان با مقایسه تشخیص‌های این مدل با مدل‌های مشابه در این چهار کلاس، پی به برتری مدل پیشنهادی در این زمینه خواهیم برد.

۴-۲-۲ نتایج مدل در چهار کلاس:

به منظور رسیدن به اهدافی که در بخش قبلی ذکر شد، مجموعه‌داده با توجه به نوع کلاس آن‌ها به چهار دسته تقسیم‌بندی و به مدل آموزش داده شد. در این تقسیم‌بندی داده‌ها به چهار کلاس: انسان، متون تولیدشده توسط مدل مولد GPT-2، RNN و دیگر مدل‌های مولد متن، تقسیم‌بندی شده و به مدل پیشنهادی آموزش داده می‌شوند.

جدول (۳): هایپرپارامترهای اعمالی بر مدل (چهار کلاسه)

Hyperparameter	Value
Num.of training epochs	**8**
Initial learning rate	**10^{-6}**
Batch size	**2**
Dropout rate	**0.3**
Embedding size for RoBERTa	**768**
Max tokens to use	**512**

مشخصات هایپرپارامترهای اعمالی برای آموزش و تست در مدل در جدول (۳) نشان داده‌شده است.

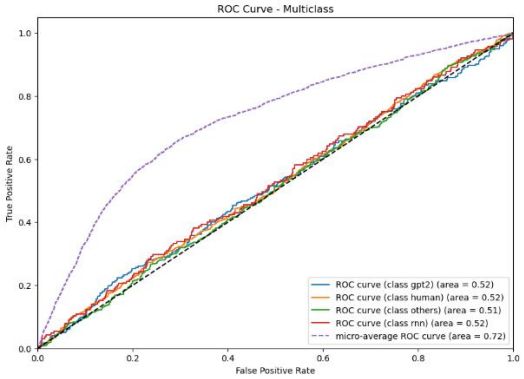
مقادیر به دست آمده از مدل پیشنهادی در مواجهه با داده‌هایی که به چهار کلاس تقسیم شدن در جدول (٤) ارائه شده است. این

مقادیر قدرت تشخیص مدل پیشنهادی را جهت تمایز بین این متون به نمایش می‌گذارد. نتایج حاصل‌شده از مدل، نشان می‌دهد که این مدل دارای دقت مطلوبی در تشخیص متون تولیدشده توسط مولدهای تولید متن قوی ماشینی دارد، همچنین نشان می‌دهد که این مدل تعمیم پذیری بالایی در مواجهه با مدل‌های مختلف تولید متن را داراست.

جدول (٤): نتایج مدل پیشنهادی در داده‌های چهارکلاسه

Model	human			Gpt-2			rnn			others			Accuracy
ours	precision	recall	f1	precision	recall	f1	precision	recall	f1	precision	recall	f1	0/89
	0/66	0/82	0/73	0/92	0/86	0/89	0/91	0/92	0/92	0/96	0/86	0/91	

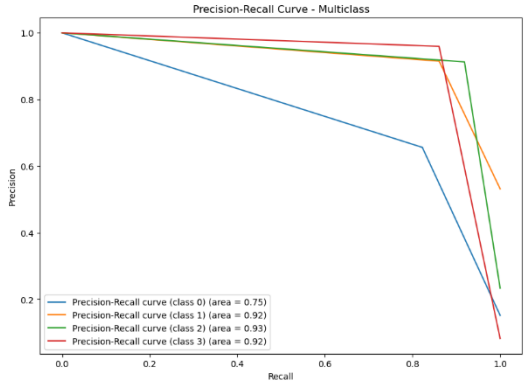
یکی از روش‌های بررسی و ارزیابی عملکرد دسته‌بندی، «نمودار مشخصه عملکرد[1]» یا به اختصار منحنی ROC است. منحنی ROC توسط ترسیم نسبت یا نرخ «مثبت صحیح[2]» که به اختصار TPR نامیده می‌شود برحسب «نرخ مثبت کاذب[3]» با نام اختصاری FPR، ایجاد می‌شود. در ادامه جهت بررسی دقیق‌تر عملکرد مدل پیشنهادی در کلاس‌بندی چهار کلاسه، منحنی ROC مربوطه ترسیم شده است. که بیانگر عملکرد مطلوب مدل است. نمودار (۳) منحنی ROC را نمایش می‌دهد.



نمودار (۳): ROC مربوط به مدل پیشنهادی در داده‌های چهارکلاسه

منحنی ROC برای هر کلاس نشان می‌دهد که چگونه مدل می‌تواند آن کلاس خاص را از کلاس‌های دیگر متمایز کند. به طور خلاصه، نمودار منحنی ROC یک ابزار قدرتمند برای نمایش و ارزیابی عملکرد مدل طبقه‌بندی چند کلاسه است که بینش‌هایی را در مورد اینکه چگونه مدل می‌تواند بین هر کلاس تمایز قائل شود، به خواننده می‌دهد و به انتخاب آستانه‌های بهینه کمک می‌کند.

همچنین می‌توان منحنی PR[4] را برای ارزیابی بیشتر مدل ترسیم نمود. منحنی PR برای ارزیابی عملکرد یک مدل طبقه‌بندی، به‌ویژه از نظر دقت (مقدار پیش‌بینی مثبت) و recall آن استفاده می‌شود. این به‌ویژه در سناریوهایی که کلاس‌ها نامتعادل هستند، مفید است.



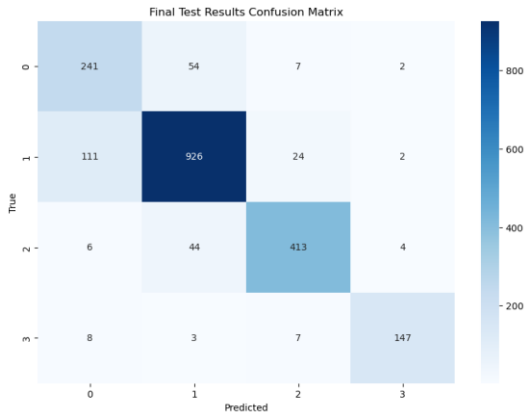
نمودار (٤): PR مدل پیشنهادی در داده‌های چهار کلاسه

- **توجه:** در نمودارهای این بخش کلاس ۰ مربوط به کلاس انسان، کلاس ۱ مربوط به کلاس GPT-2، کلاس ۲ مربوط به کلاس RNN و کلاس ٤ مربوط به کلاس دیگر مولدها است.

محور افقی در این نمودار (٤)، نشان‌دهنده نرخ مثبت واقعی است که به‌درستی توسط مدل شناسایی شده است. محور عمودی نشان‌دهنده ارزش اخباری مثبت یا نسبت موارد مثبت پیش‌بینی‌شده است که درواقع درست هستند. بررسی دقیق این نمودار به مطلوب بودن دقت مدل پیشنهادی اشاره می‌کند و عملکرد خوب مدل را در داده‌های چهار کلاسه به مانند داده‌های دو کلاسه، نشان می‌دهد. در ادامه ماتریس سردرگمی مربوط به داده‌های چهار کلاسه ترسیم شده است. ماتریس سردرگمی خلاصه‌ای از عملکرد مدل را با نشان

[1] Receiver Operating Characteristic
[2] True Positive Rate
[3] False Positive Rate
[4] Precision-Recall Curve

دادن تعداد پیش‌بینی‌های مثبت درست، مثبت کاذب، منفی درست و منفی کاذب برای هر کلاس ارائه می‌دهد. ردیف‌ها در این ماتریس نشان‌دهنده کلاس‌های واقعی هستند و ستون‌ها نشان‌دهنده کلاس‌های پیش‌بینی است.



شکل (۳): ماتریس درهم‌ریختگی مربوط به مدل پیشنهادی در داده‌های چهارکلاسه

شکل (۳) ماتریس درهم‌ریختگی مدل پیشنهادی را در داده‌های چهار کلاسه نشان می‌دهد. بررسی نتایج حاصل از مدل پیشنهادی، نشان از مطلوبیت مدل در شناسایی و دسته‌بندی متون تولیدشده توسط ماشین، دارد. در ادامه پژوهش مقایسه‌ای بین مدل پیشنهادی با سایر مدل‌هایی که در این زمینه با مجموعه داده مشابه، معرفی شده‌اند، صورت گرفته است.

۵- مقایسه نتایج

داده‌های موجود در مجموعه‌داده Tweepfake با روش‌های زیادی جهت تشخیص متن تولیدشده توسط ماشین از متن نوشته شده توسط انسان، آموزش داده‌شده و آزمایش شده‌اند. در این بخش در قالب جدول زیر نتایج صحت مدل پیشنهادی در این پژوهش با دیگر روش‌های ارائه شده جهت تشخیص، با یکدیگر مقایسه شده است. نتایج نشان از برتری مدل پیشنهادی نسبت به دیگر روش‌های ارائه شده، دارد. معیارهایی که با آن روش پیشنهادی و دیگر روش‌ها مقایسه می‌شوند عبارت است از: Precision، Recall، F1 و Accuracy. این نتایج حاصل از آموزش و آزمایش مجموعه‌داده در دو کلاس متن نوشته شده توسط انسان و متن تولیدشده توسط ربات هوش مصنوعی است. در این مقایسه مدل‌های مختلف پردازش زبان طبیعی مورد استفاده قرارگرفته و از آن رو می‌توان جدول (۵) مقایسه‌ای را جدولی با تعمیم پذیری بالا و قابل استناد دانست.

۶- نتیجه‌گیری

مدل ارائه شده در این پژوهش، مدلی ترکیبی است، که دارای دقتی ۹۰ درصدی در تشخیص متون ماشینی است. همچنین بررسی دقت این مدل در داده‌های چهارکلاسه، نشان از تعمیم‌پذیری مطلوب مدل پیشنهادی دارد. مقایسه دقت مدل پیشنهادی با دیگر مدل‌هایی که روی مجموعه‌داده استفاده‌شده در این پژوهش، پیاده‌سازی شده‌اند، نشان از برتری این مدل نسبت به کارهای گذشته دارد. به‌طورکلی مدل پیشنهادی می‌تواند عملکرد مطلوبی در تشخیص متون ماشینی از خود به نمایش بگذارد و در این زمینه مورد استفاده بیشتری قرار گیرد.

۷- کارهای آتی

وجود مجموعه‌داده مناسب برای تشخیص و تمایز بین متون تولیدشده توسط هوش مصنوعی و متون نوشته شده توسط انسان، بسیار لازم و ضروری است. با توجه به ظهور مدل‌های زبانی مولد متن ماشینی جدید و پیچیده، نیاز به مجموعه‌داده‌ای که این متون را شامل شود، ضرورت دارد. همچنین با توجه به فقدان مجموعه‌داده‌ای این‌چنینی در زبان فارسی، لازم است که با توجه به‌ضرورت موضوع، جهت جلوگیری از آثار مخربی که این متون جعلی ماشینی برای جامعه ایجاد می‌کند، مجموعه‌داده کامل و جامعی گردآوری شود. تا بتوان این مدل تشخیص را در داده‌های زبان فارسی امتحان نمود و ارتقاع داد. همچنین می‌توان مدل پیشنهادی را بر روی دیگر مجموعه داده‌های منتشرشده موجود آموزش و آزمایش کرد تا بیش‌ازپیش کارایی مدل پیشنهادی مورد تجزیه‌وتحلیل قرار گیرد.

جدول (۵): مقایسه مدل پیشنهادی با سایر مدل‌ها

روش		انسان			ربات		نتایج کلی
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	Accuracy
LOG_REG_BOW	0/84	0/74	0/79	0/77	0/85	0/81	**0/80**
RAND_FOREST_BOW	0/75	0/79	0/77	0/78	0/74	0/76	**0/77**
SVC_BOW	0/85	0/75	0/80	0/77	0/86	0/82	**0/81**
LOG_REG_BERT	0/84	0/82	0/83	0/82	0/85	0/83	**0/83**
RAND_FOREST_BERT	0/86	0/77	0/81	0/79	0/87	0/83	**0/82**
SVC_BERT	0/86	0/81	0/83	0/82	0/86	0/84	**0/84**
CHAR_CNN	0/89	0/79	0/84	0/81	0/90	0/85	**0/85**
CHAR_GRU	0/89	0/74	0/81	0/78	0/91	0/84	**0/83**
CHAR_CNNGRU	0/84	0/82	0/83	0/82	0/85	0/83	**0/83**
BERT_FT	0/89	0/88	0/89	0/88	0/90	0/89	**0/89**
DISTILBERT_FT	0/89	0/88	0/88	0/88	0/89	0/88	**0/88**
ROBERTA_FT	0/90	**0/89**	0/89	0/89	0/90	0/89	**0/89**
XLNET_FT	0/91	0/83	0/87	0/84	0/92	0/88	**0/87**
OURS	**0/91**	0/87	**0/89**	**0/89**	**0/92**	**0/90**	0/90

References

[1] Jawahar, G., M. Abdul-Mageed, and L.V. Lakshmanan, Automatic detection of machine generated text: A critical survey. arXiv preprint arXiv:2011.01314, 2020.

[2] Kudugunta, S. and E. Ferrara, Deep neural networks for bot detection. Information Sciences, 2 : 467 . 018p. 312-322.

[3] Solaiman, I., et al., Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203, 2019.

[4] Zellers, R., et al., Defending against neural fake news. Advances in neural information processing systems, 2019. 32.

[5] Uchendu, A., et al. Authorship attribution for neural text generation. in Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). 2020.

[6] Dukić, D., D. Keča, and D. Stipić. Are you human? Detecting bots on Twitter Using BERT. in 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). 2020. IEEE.

[7] Adelani, D.I., et al. Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. in Advanced information networking and applications: Proceedings of the 34th international conference on advanced information networking and applications (AINA-2020). 2020. Springer.

[8] Ippolito, D., et al., Automatic detection of generated text is easiest when humans are fooled. arXiv preprint arXiv:1911.00650, 2019.

[9] Rodriguez, J.D., et al. Cross-domain detection of GPT-2-generated technical text. in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: human language technologies. 2022.

[10] Crothers, E.N., N. Japkowicz, and H.L. Viktor, Machine-generated text: A comprehensive survey of threat models and detection methods. IEEE Access, 2023. 11: p. 70977-71002.

[11] Jacob, D., et al., BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). arXiv preprint arXiv:1810.04805, 2018.

[12] Liu, Y., Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[13] Adelani, D.I., et al. Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. in Advanced information networking and applications: Proceedings of the 34th international conference on advanced information networking and applications (AINA-2020).

[14] Ippolito ,D., et al., Automatic detection of generated text is easiest when humans are fooled. arXiv preprint arXiv:1911.00650, 2019.

[15] Fagni, T., et al., TweepFake: About detecting deepfake tweets. Plos one, 2021. 16(5): p. e0251415.

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 2, No. 1, Spring & Summer 2024, pp. 142 - 154

Automatic detection of machine-generated texts based on linguistic models and deep learning

AbolfazlShirafkan,* MahdiNaghavi, MeysamMirzai

Faculty and Research Institute of Artificial Intelligence and Cognitive Sciences, Imam Hossein (AS) University, Tehran, Iran

Article Information

Original Research Paper

Received:
2024 November 26

Accepted:
2025 January 21

Keywords:
Deep learning,writing style based model, RoBERTa, BiLSTM

Corresponding Author*:
Abolfazlshirafkan00@gmail.com

Abstract

Today, with the significant growth of artificial intelligence and its products, many opportunities and threats have emerged. One of the most famous and popular products of artificial intelligence is text generation, also called machine text. In this research, a new method is introduced that combines features extracted from the text with its structural features, thus creating an automatic discriminator to distinguish between human-written text and artificial intelligence-generated text. The introduced method consists of two parts, the first part: the extended BERT (RoBERTa) model and the bidirectional long-term short-term memory (BiLSTM) model, which are improved with the fusion layer. The second part: the structural features of the text are extracted using a writing style-based method. Finally, the output of the model parts is combined together, and in this way, the model distinguishes human-written text from machine-generated text. The results of this research show that the proposed method is capable of recognizing machine texts with 90% accuracy and exhibits good performance.

: 10.22034/abmir.2025.22455.1078

E-ISSN: 2821-2037 /© 2023. Published by Yazd University This is an open access article under the CC BY 4.0 License (https://creativecommons.org/licenses/by/4.0/).