سایبربولینگ و تاثیر هوش مصنوعی

فائزه اسکندری، فاطمه اصغری بیرامی

nfaezeeskandari@gmail.com

fatemehasgharib@gmail.com

۱. چکیده:

سایبر بولیینگ به‌عنوان یکی از مشکلات جدی عصر دیجیتال، اثرات منفی عمیقی بر سلامت روانی، اجتماعی و اقتصادی افراد و جوامع دارد. با گسترش استفاده از شبکه‌های اجتماعی، هوش مصنوعی به ابزاری کلیدی برای شناسایی و مقابله با این معضل تبدیل شده است. این مقاله مروری با رویکردی سیستماتیک، علمی معتبر از پایگاه‌هایی نظیر PubMed، Google Scholar و IEEE Xplore را بررسی کرده است. مقالات انتخابی بر اساس ارتباط موضوعی، تاریخ انتشار (۲۰۱۸ به بعد)، و کیفیت علمی تحلیل شدند. سپس، تکنیک‌های هوش مصنوعی، دقت سیستم‌ها، و چالش‌های فنی و اخلاقی استخراج و ارزیابی شدند تا تصویری جامع از وضعیت فعلی و نیازهای پژوهشی ارائه شود.این مقاله مروری به بررسی جامع نقش الگوریتم‌های هوش مصنوعی در شناسایی محتوای آزاردهنده می‌پردازد. ابتدا، تکنیک‌های پیشرفته مورد استفاده در پردازش زبان طبیعی (NLP)، یادگیری عمیق و سیستم‌های چندمدلی برای تحلیل متن، تصویر و ویدئو معرفی می‌شوند. سپس، عملکرد و دقت این سیستم‌ها از طریق تحلیل آماری و مطالعات موردی در پلتفرم‌های اجتماعی نظیر توییتر، فیسبوک و اینستاگرام ارزیابی می‌شود. چالش‌های فنی نظیر درک زمینه در متون، تحلیل زبان‌های عامیانه و چندگانه، و شناسایی محتوای پیچیده بصری بررسی شده و محدودیت‌های کنونی تشریح می‌گردد. علاوه بر این، به چالش‌های اخلاقی مانند نقض حریم خصوصی، سوگیری الگوریتم‌ها، و شفافیت تصمیم‌گیری پرداخته می‌شود. در پایان، راهکارهایی شامل توسعه الگوریتم‌های توضیح‌پذیر، ترکیب هوش مصنوعی و نیروی انسانی، و ایجاد قوانین اخلاقی پیشنهاد می‌شود. نتایج نشان می‌دهد که به‌رغم پیشرفت‌های چشمگیر، برای مقابله با سایبر بولیینگ به تحقیقات بیشتر، داده‌های متنوع‌تر و همکاری‌های بین‌المللی نیاز است.

کلیدواژه ها: سایربولینگ، هوش مصنوعی، پردازش زبان طبیعی، چالش های اخلاقی، شناسایی محتوای آزاردهنده

۲. مقدمه:

بولینگ یا قلدری نوعی رفتار نابهنجار و همراه با خشونت مداوم است که به منظور تهدید و ترساندن قربانی است که معمولا آدم ضعیف تری نسبت به فرد آسیب رسان می باشد؛ فرد بولی کننده، خود را به لحاظ موقعیت اجتماعی، روانی و فیزیکی از قربانی برتر میبیند و میخواهد با رفتارهای خشونت بار این را نشان دهد. این رفتار علاوه بر آسیب های فیزیکی میتواند همراه با ضربه های روانی نیز باشد (کنت و همکاران، ۲۰۰۸). قلدری، همچنین میتواند شامل توهین های زبانی مثل مسخره کردن و طرد اجتماعی علاوه بر زد و خوردهای فیزیکی باشد (کووالسکی و لیمبر، ۲۰۰۷). گسترش چشمگیر فناوری در چند دهه ی اخیری، ابزاری جدید برای بولی و آسیب رساندن به دیگران ارائه کرده است، این پدیده ی سایبربولینگ یا حملات سایبری به حدی فراگیر شده است که تعاریف متفاوت و راهکارهای مختلفی جهت پیشگیری و حل کردن آن ارائه شده است. سایبربولینگ به نوعی قلدری اشار می کند که از طریق پیامک، پست الکترونیکی، اتاقهای گفتگو، صفحات وب یا از طریق تصاویر و پیامهای ارسالی از طریق تلفن های همراه رخ می دهد (دیلماک و آیدوگان، ۲۰۱۱). کورکوران و همکاران(۲۰۱۵) نیز تعریف مشابهی از قلدری مجازی ارائه دادند، این تعریف به ابعاد وسیع تری اشاره دارد که در آن هر نوع رفتار مجرمانه ای را از طریق رسانه های ارتباط دیجیتال نشان میدهد. با توجه به این تعاریف و پیشرفت فناوری، یکی از راه حل های قابل توجه برای حل مشکل سایبربولینگ را میتوان هوش مصنوعی نام برد. الگوریتم های هوش مصنوعی برای تشخیص دقیق تر و سریع تر محتوا و پیام های آزار دهنده در شبکه های ارتباط مجازی بزرگ از جمله تلگرام و اینستاگرام استفاده می شوند، اما همچنان نیاز به اصلاح و نیرومند سازی دارند. هوش مصنوعی به عنوان «توانایی یک سیستم برای پردازش صحیح داده های خارجی، یادگیری از همان داده ها، استفاده از آن دانش از طریق سازگاری و انعطاف پذیری برای دستیابی به اهداف و وظایف خاص» تعریف میشود (هانلین و کاپلان، ۲۰۲۴). در این راستا، استفاده از هوش مصنوعی میتواند به عنوان یک ابزار پیشگیرانه و همچنین واکنشی در برابر سایبربولینگ عمل کند. به عنوان مثال، الگوریتم های یادگیری ماشین میتوانند الگوهای رفتاری مشکوک را شناسایی کرده و به سرعت به مدیران شبکه های اجتماعی گزارش دهند. علاوه بر این، سیستم های هوش مصنوعی میتوانند به کاربران کمک کنند تا محتوای آزار دهنده را فیلتر کنند و از تجربه های منفی در فضای مجازی جلوگیری کنند. با توجه به مطالعات انجام گرفته، گروه های سنی کودکان و نوجوانان، در معرض آسیب پذیری بیشتری نسبت به سایر گروه های سنی در برابر سایبربولینگ، قرار دارند. نوجوانان به دلیل تغییرات دوران بلوغ و هویت یابی، ممکن است شبکه های مجازی مثل اسنپ چت را راهی برای رهایی و دوری از دنیای واقعی پیدا کنند، اما با نبود پیشگیری لازم و مناسب، این گروه سنی میتواند درمعرض خطرات مجازی از جمله فحاشی و تهدید شدن قرار بگیرد. همچنین، مشخص شده است که کودکان و نوجوانانی که در اینترنت خود را با هویت جعلی معرفی میکنند، معمولا از اعتماد به نفس کمتر و درگیر سطح بالایی از اضطراب و خشونت اجتماعی برخوردار هستند (آریکاک، ۲۰۰۹). آموزش کاربران، به ویژه کودکان و نوجوانان، در مورد خطرات و پیامدهای سایبربولینگ و همچنین راه های مقابله با آن میتواند به کاهش اثرات سایربولینگ کمک کند. برنامه های آموزشی میتوانند شامل کارگاه ها، سمینارها، و منابع آنلاین باشند که به کاربران اطلاعات لازم را در مورد نحوه شناسایی و گزارش سایبربولینگ ارائه میدهند. همچنین، ایجاد قوانین و مقررات سختگیرانه تر در فضای مجازی میتواند به کاهش سایبربولینگ کمک کند. این قوانین باید به گونه ای طراحی شوند که هم از حقوق کاربران محافظت کنند و هم به سرعت و به طور موثر با موارد سایبربولینگ برخورد کنند. همکاری بین المللی نیز در این زمینه ضروری است، زیرا فضای مجازی مرزهای جغرافیایی را نمیشناسد و برای مقابله با سایبربولینگ نیاز به همکاری جهانی است. حمایت روانی و اجتماعی از قربانیان سایبربولینگ نیز بسیار مهم است. ارائه خدمات مشاوره و حمایت روانی به

قربانیان میتواند به آنها کمک کند تا با اثرات منفی سایبربولینگ مقابله کنند و به زندگی عادی خود بازگردند. ایجاد شبکه های حمایتی و گروه های پشتیبانی نیز میتواند به قربانیان کمک کند تا احساس تنهایی نکنند و بدانند که تنها نیستند. با توجه به پیچیدگی و گستردگی مسئله سایبربولینگ، همکاری بین متخصصان فناوری، روانشناسان، و سیاستگذاران ضروری است تا راهکارهای جامع و موثری برای مقابله با این پدیده ارائه شود. هوش مصنوعی میتواند نقش مهمی در این فرآیند ایفا کند، اما نیاز به نظارت و بهبود مستمر دارد تا بتواند به طور موثر با چالشهای جدید و پیچیده سایبربولینگ مقابله کند.

۳. سایبربولینگ و هوش مصنوعی

شناسایی محتوای آزاردهنده و آزار مجازی با الگوریتم های هوش مصنوعی میتواند سرعت بیشتری بگیرد اما همچنان این سیستم نیاز به بهبودسازی دارد تا بتواند در بهینه ترین و سریع ترین حالت ممکن، از میان دنیای وسیع تبادل اطلاعات به صورت مجازی، محتوای آسیب زا را حذف کند.

۳.۱. سایبربولینگ

ویلارد (۲۰۰۵)، سایبربولینگ را به شکل ارسال مضامین ناراحت کننده و بی رحمانه از طریق ابزارهای ارتباطی دیجیتال تعریف کرد و راه هایی که سایبربولینگ ممکن است رخ بدهد را به این صورت معرفی کرد:

الف) عصبانیت: ارسال پیامهای خشمگینانه، بی ادبانه یا زشت برای افراد خاص ، به شکل خصوصی یا گروه های آنلاین

ب) آزار: ارسال مکرر یک پیام ناراحت کنند به یک فرد یا آزار از طریق تهدید به آسیب یا ارسال پیام های رعب آور

ج) بی آبرویی : ارسال عباراتی ناراحت کننده ، غیرواقعی و بی رحمانه در مورد افراد دیگر

د) تظاهر: وانمود کردن به اینکه فرد دیگری هستید و ارسال مواردی برای بد نام کردن فرد قربانی یا دچار مشکل ساختن او

ج) تفریح و فریب: ارسال مواردی شامل اطلاعات شخصی یا خجالت آور در مورد یک فرد ، درگیر شدن در فریبکاری برای بدست آورن اطلاعات شخصی و سپس عمومی ساختن آن و نیز ارسال تصاویر و پیامهای شخصی برای دیگران

د) بیرون سازی: طرد و خارج ساختن عمدی یک فرد از اجتماع یا یک گرو آنلاین

قلدری مجازی برای مشخص شدن نوع جرم و مجازات، نیاز دارد که در ابعاد قانونی و روانشناختی به خوبی تعریف شود؛ برای مثال ایالت میشیگان در آمریکا در سال ۲۰۱۹، یک تعریف مشخص برای سیستم قضایی خود نسبت به سایبربولینگ ارائه داد: " سایبربولینگ شامل بارگذاری عامدانه عبارات و پیامهای آسیب زا در فضای مجازی است که ترس از صدمه بدنی شدید یا خشونت یا مرگ را در یک فرد متعارف به وجود می آورد."

در تحقیقات شکل گرفته در زمینه ی سایبربولینگ در میان نوجوانان و قشر جوان مشخص شد که افرادی که معمولا دست به قلدری مجازی میزنند، درون گراتر و عصبی تر از افرادی هستند که دست به خشونت مجازی نمیزنند. همچنین، این افراد اضطراب بالاتری را نسبت به همسن و سالان خود تجربه میکنند. دلیل این یافته ها ممکن است این باشد که افراد مضطرب از طریق مورد سایبربولینگ قرار دادن دیگران و انتقال اضطراب خود به آنها، سعی می کنند تا حدودی از فشار ها و تنش های

درونی خود بکاهند. از آنجایی که سایبربولینگ شکل جدیدی از پرخاشگری است که توسط فناوری ابراز می شود، پس ممکن است افرادی که پرخاشگر تر هستند بیشتر به سایبربولینگ روی بیاورند و از این طریق پرخاشگری خود را ابراز کنند، بدون اینکه نیاز باشد به صورت رو در رو با افراد مواجه شوند.

۳.۲. هوش مصنوعی

شناسایی محتوای آزاردهنده در محیط‌های دیجیتال به دلیل تنوع و پیچیدگی محتوا یکی از چالش‌های اصلی در علم داده و امنیت سایبری است. تکنیک‌های هوش مصنوعی در این زمینه نقش کلیدی ایفا می‌کنند و شامل پردازش زبان طبیعی، یادگیری عمیق، و سیستم‌های چندمدلی هستند که در ادامه به تفصیل بررسی می‌شوند.

۳.۲.۱. تکنیک های مورد استفاده ی هوش مصنوعی

برای مثال، پردازش زبان طبیعی (Natural Language Processing) یا NLP به طور گسترده برای شناسایی محتوای متنی آزاردهنده به کار می‌رود. این روش‌ها شامل استفاده از مدل‌های زبانی پیشرفته مانند Transformer-based models ، نظیر BERT وGPT، است که توانایی تحلیل دقیق متن، استخراج مفاهیم و تشخیص الگوهای زبانی توهین‌آمیز یا آزاردهنده را دارند. یکی از روش‌های مرسوم، استفاده از تحلیل احساسات (Sentiment Analysis) است که در آن جملات بر اساس بار عاطفی مثبت، منفی یا خنثی طبقه‌بندی می‌شوند. مدل‌های مبتنی بر یادگیری عمیق، نظیر BiLSTM وCNN ، نیز در ترکیب با ویژگی‌های معنایی استخراج‌شده از ابزارهای NLP برای بهبود دقت استفاده می‌شوند .(Kim et al., 2021) علاوه بر این، تشخیص موجودیت‌های نامدار (Named Entity Recognition) و تحلیل وابستگی‌های نحوی (Dependency Parsing)به شناسایی محتوای خاصی نظیر تهدیدات مستقیم یا زبان توهین‌آمیز کمک می‌کند .(Zhou et al., 2022) استفاده از داده‌های بزرگ برای آموزش این مدل‌ها نقش مهمی در بهبود کارایی و دقت آنها دارد.

۳.۲.۲. استفاده از یادگیری عمیق برای تحلیل تصاویر و ویدئوهای آزاردهنده

در زمینه شناسایی محتوای بصری، یادگیری عمیق (Deep Learning) نقش حیاتی دارد. روش‌های مبتنی بر شبکه‌های عصبی پیچشی Convolutional Neural Networks) یا (CNNs به طور گسترده برای تحلیل تصاویر استفاده می‌شوند. این شبکه‌ها توانایی شناسایی الگوهای پیچیده در تصاویر و تمایز محتوای آزاردهنده، مانند تصاویر خشونت‌آمیز یا پورنوگرافی، را دارند. برای تحلیل ویدئوها، مدل‌های پیشرفته‌تر نظیر شبکه‌های عصبی سه‌بعدی (3D CNNs) و ترکیب آنها با شبکه‌های بازگشتی (RNNs) جهت مدل‌سازی توالی فریم‌های ویدئویی به کار می‌روند. مدل‌هایی مانند I3D (Inflated 3D CNN) و TSN (Temporal Segment Networks) توانسته‌اند عملکرد قابل‌توجهی در شناسایی ویدئوهای آزاردهنده ارائه دهند (Nguyen et al., 2023). علاوه بر این، روش‌های یادگیری انتقالی (Transfer Learning) و مدل‌های پیش‌آموزش‌دیده، نظیرVision Transformers (ViTs) ، برای بهبود دقت در دسته‌بندی محتوای بصری مورد استفاده قرار می‌گیرند (Dosovitskiy et al., 2021). این روش‌ها به دلیل قابلیت تعمیم بالا در محیط‌های چندرسانه‌ای مختلف کارآمد هستند.

۳.۲.۲.۱. سیستم‌های ترکیبی برای تحلیل چندمدلی (متن، تصویر، ویدئو)

سیستم‌های ترکیبی (Multimodal Systems) از قدرت ترکیبی روش‌های NLP و یادگیری عمیق برای شناسایی محتوای آزاردهنده در قالب‌های مختلف بهره می‌برند. این سیستم‌ها داده‌های متنی، تصویری و ویدئویی را به صورت همزمان تحلیل می‌کنند. مدل‌های چندوجهی، مانند(CLIP (Contrastive Language–Image Pretraining ، با ترکیب متن و تصویر در یک فضای تعبیه‌شده مشترک، عملکرد بی‌نظیری در شناسایی محتوای ترکیبی آزاردهنده دار (Radford et al.,2021) . همچنین، استفاده از مدل‌های Transformer-based Multimodal Architectures نظیر MERT و VideoBERT امکان استخراج ویژگی‌های پیچیده از داده‌های چندرسانه‌ای را فراهم کرده است. این مدل‌ها می‌توانند همبستگی میان محتوای متنی و بصری را شناسایی کرده و تحلیل دقیقی ارائه دهند (Sun et al., 2022). از دیگر تکنیک‌های مؤثر، ادغام داده‌های حسگرهای زیستی (مانند تشخیص واکنش‌های احساسی کاربر) با داده‌های چندرسانه‌ای است که می‌تواند به سیستم‌های شناسایی درک عمیق‌تری از محتوای آزاردهنده ارائه دهد. این ترکیب به خصوص در حوزه‌های حساسی مانند نظارت بر محتوای کودکان و سیستم‌های امنیتی بسیار کاربردی است. تکنیک‌های هوش مصنوعی، به ویژه روش‌هایNLP ، یادگیری عمیق، و سیستم‌های چندمدلی، ابزارهای قدرتمندی برای شناسایی محتوای آزاردهنده هستند. با پیشرفت فناوری‌های محاسباتی و ظهور مدل‌های پیشرفته‌تر، انتظار می‌رود این روش‌ها به طور فزاینده‌ای دقیق‌تر و کارآمدتر شوند، که تأثیر قابل‌توجهی در افزایش امنیت دیجیتال و ارتقاء تجربه کاربران خواهد داشت.

۳.۲.۳. دقت و توانایی سیستم‌های فعلی

تحلیل دقت و توانایی سیستم‌های شناسایی محتوای آزاردهنده در بسترهای دیجیتال نشان‌دهنده پیشرفت چشمگیر الگوریتم‌های هوش مصنوعی در این زمینه است. با این حال، همچنان چالش‌هایی از جمله نرخ بالای مثبت کاذب (False Positives)محدودیت در شناسایی محتوای متنی و چندرسانه‌ای پیچیده وجود دارد. این بخش به بررسی مطالعات موردی، تحلیل آماری، و مقایسه الگوریتم‌ها پرداخته و قابلیت‌های مختلف آنها را ارزیابی می‌کند.

۳.۲.۳.۱.مطالعه نمونه‌های موفق پیاده‌سازی هوش مصنوعی در پلتفرم‌ها

پلتفرم‌های اجتماعی نظیر توییتر، فیسبوک و اینستاگرام از سیستم‌های پیشرفته هوش مصنوعی برای شناسایی و حذف محتوای آزاردهنده استفاده می‌کنند:

-توییتر: توییتر از مدل‌های پیشرفته پردازش زبان طبیعی و شبکه‌های عصبی برای تحلیل متون آزاردهنده بهره می‌برد. طبق گزارش‌های اخیر، سیستم‌های توییتر در سال ۲۰۲۳ موفق به شناسایی و حذف بیش از ۶۵٪ از توییت‌های آزاردهنده پیش از گزارش کاربر شده‌اند .(Twitter Transparency Report, 2023)

-فیسبوک: فیسبوک از معماری چندوجهی برای شناسایی محتوای آزاردهنده در متن، تصویر و ویدئو استفاده می‌کند. این سیستم‌ها توانسته‌اند نرخ شناسایی بیش از ۹۷٪ برای محتوای خشونت‌آمیز و ۹۴٪ برای محتوای نفرت‌پراکن را ثبت کنند .(Meta Community Standards Report, 2023)

-اینستاگرام: الگوریتم‌های اینستاگرام از یادگیری عمیق برای تحلیل تصاویر و ویدئوها استفاده کرده و در ترکیب با مدل‌های NLP برای تحلیل کپشن‌ها و کامنت‌ها، نرخ دقت بالایی در شناسایی محتوای مضر به دست آورده‌اند. این پلتفرم همچنین از بازخورد کاربران برای بهبود سیستم‌های خود بهره می‌برد.

۳.۲.۳.۲. تحلیل آماری از دقت سیستم‌ها در شناسایی محتوای آزاردهنده

مطالعات اخیر دقت الگوریتم‌های مختلف را در شناسایی محتوای آزاردهنده تحلیل کرده‌اند:

-دقت در شناسایی متن: مدل‌های NLP نظیر BERT و RoBERTa توانسته‌اند به دقتی در حدود ۹۰-۹۳٪ در شناسایی محتوای متنی توهین‌آمیز دست یابند. با این حال، این مدل‌ها هنوز در شناسایی زبان‌های غیررسمی، اسلنگ‌ها، و کنایه‌ها چالش دارند (Zhou et al., 2022).

-دقت در شناسایی تصاویر: شبکه‌های عصبی پیچشی (CNNs) نرخ دقت ۹۵٪ در شناسایی تصاویر خشونت‌آمیز یا پورنوگرافیک گزارش کرده‌اند، اما در شناسایی تصاویر دستکاری‌شده یا محتوای مبهم عملکرد ضعیف‌تری دارند (Nguyen et al., 2023).

-دقت در تحلیل ویدئوها: مدل‌های پیشرفته نظیر I3D و ViViT توانسته‌اند نرخ دقت بیش از ۹۲٪ در تحلیل ویدئوهای خشونت‌آمیز ارائه دهند، اما کارایی آنها به شدت به کیفیت داده‌های آموزشی وابسته است.

-مثبت‌های کاذب و منفی‌های کاذب: یکی از چالش‌های اصلی سیستم‌ها نرخ مثبت کاذب بالا (تا ۱۵٪) است که می‌تواند محتوای غیرمضر را به اشتباه مسدود کند. از سوی دیگر، در برخی موارد منفی‌های کاذب (تا ۱۰٪) باعث می‌شود محتوای مضر از شناسایی فرار کند (Kim et al., 2021) .

۳.۲.۳.۳.مقایسه الگوریتم‌های مختلف و قابلیت‌های آنها

الگوریتم‌های مختلف هوش مصنوعی نقاط قوت و ضعف خاصی دارند که بسته به کاربرد مورد نظر، عملکرد متفاوتی ارائه می‌دهند:

الگوریتم	مزایا	معایب	کاربردها
BERT and RoBERT	دقت بالا در تحلیل متون	عملکرد ضعیف در مواجهه با کنایه ها	شناسایی متن های آزاردهنده
CNNs	دقت بالا در تحلیل تصاویر	محدودیت در شناسایی محتوای دستکاری شده	تصاویر و محتوای خشونت آمیز
ViViT and I3D	تحلیل موثر ویدیوهای پیچیده	نیاز به داده هایی با کیفیت بالا	شناسایی خشونت در ویدیوها
CLIP	تحلیل چندوجهی (متن و تصویر)	هزینه محاسباتی بالا	سیستم های چندمدلی
شبکه های بازگشتی (RNNs)	مدل سازی توالی زمانی موثر	حساس به تغییرات داده و مشکلات در همگرایی	تحلیل ویدیوهای کوتاه

جدول ۳.۲.۳.۲.۱. مقایسه الگوریتم های مختلف و قابلیت های آنها

دقت و توانایی سیستم‌های فعلی در شناسایی محتوای آزاردهنده به طور قابل‌توجهی بهبود یافته است، اما چالش‌هایی همچنان باقی‌مانده‌اند. مدل‌های پیشرفته‌تر و داده‌های آموزشی متنوع‌تر می‌توانند دقت را افزایش داده و چالش‌های مثبت کاذب و منفی کاذب را کاهش دهند. همچنین، استفاده از ترکیب روش‌ها و سیستم‌های چندمدلی نویدبخش آینده‌ای با امنیت دیجیتال بالاتر است.

۳.۲.۴. چالش‌های فنی

شناسایی محتوای آزاردهنده به کمک هوش مصنوعی همچنان با چالش‌های فنی متعددی مواجه است که مانع از دستیابی به دقت و کارایی کامل سیستم‌ها می‌شود. این چالش‌ها شامل تشخیص زمینه در متون، تحلیل زبان‌های عامیانه و چندگانه، و پردازش تصاویر یا ویدئوهای پیچیده است. در ادامه، هر یک از این چالش‌ها بررسی می‌شوند.

۳.۲.۴.۱.تشخیص زمینه (Context) در متون

تشخیص دقیق محتوای آزاردهنده در متون نیازمند درک صحیح زمینه است.

۱ .کنایه و طنز: محتوای کنایه‌آمیز یا طنزگونه، به ویژه زمانی که کلمات استفاده‌شده معنای چندگانه‌ای دارند، می‌تواند سیستم‌ها را به اشتباه بیندازد. برای مثال، جمله‌ای که ظاهراً خنثی است، بسته به زمینه می‌تواند به عنوان آزاردهنده تعبیر شود.

۲ .وابستگی به اطلاعات پس‌زمینه: بسیاری از پیام‌ها معنای کامل خود را تنها در صورتی منتقل می‌کنند که اطلاعات پس‌زمینه‌ای خاصی در دسترس باشد. سیستم‌های فعلی اغلب قادر به استفاده از این اطلاعات نیستند.

۳ .مدل‌سازی روابط بلندمدت در متن: تحلیل پیام‌هایی که در مکالمات چندمرحله‌ای یا طولانی رخ می‌دهند، نیازمند توانایی مدل‌سازی وابستگی‌های بلندمدت میان جملات است. مدل‌های فعلی مانند BERT یا GPT-3 در این زمینه بهبودهایی داشته‌اند، اما هنوز کامل نیستند (Zhou et al., 2022).

راهکارهای پیشنهادی: استفاده از مدل‌های پیشرفته‌تر مانند Longformer که برای تحلیل متن‌های بلند بهینه‌سازی شده‌اند، می‌تواند عملکرد را بهبود بخشد. همچنین، تقویت سیستم‌ها با دانش خارجی (external knowledge graphs) می‌تواند در تشخیص بهتر زمینه مؤثر باشد.

۳.۲.۴.۲. مقابله با زبان‌های عامیانه، ایموجی‌ها و زبان‌های چندگانه

محیط‌های دیجیتال به طور فزاینده‌ای از زبان‌های عامیانه، اسلنگ‌ها، ایموجی‌ها و زبان‌های ترکیبی (code-switching) استفاده می‌کنند.

۱ .زبان‌های عامیانه و اسلنگ‌ها: زبان‌های غیررسمی اغلب قواعد گرامری مشخصی ندارند و شامل کلمات اختصاری، تغییرات املایی عمدی، و اصطلاحات جدید می‌شوند که در مدل‌های زبانی مرسوم پوشش داده نشده‌اند(Nguyen et al., 2023) .

۲ .ایموجی‌ها و نمادها: ایموجی‌ها می‌توانند به تنهایی یا در ترکیب با متن معنای خاصی ایجاد کنند. مدل‌های فعلی در تحلیل دقیق معنای ایموجی‌ها و تفسیر آنها در زمینه مناسب محدودیت دارند.

۳ .زبان‌های چندگانه و تغییر کد زبانی: کاربران در محیط‌های چندفرهنگی اغلب از چند زبان در یک جمله استفاده می‌کنند (مانند ترکیب انگلیسی و عربی). مدل‌های فعلی در تشخیص و تحلیل صحیح این نوع زبان‌ها دقت کمتری دارند.

راهکارهای پیشنهادی :

- استفاده از مدل‌های زبانی چندزبانه مانند mBERT و XLM-RoBERTa که برای تحلیل زبان‌های مختلف آموزش دیده‌اند.

- ایجاد مجموعه داده‌های متنوع و شامل نمونه‌های واقعی از زبان‌های عامیانه و ایموجی‌ها برای بهبود عملکرد مدل‌ها

- تحلیل ایموجی‌ها با استفاده از مدل‌های گرافی یا تکنیک‌های چندمدلی برای ترکیب اطلاعات متنی و تصویری

۳.۲.۴.۳. تشخیص تصاویر یا محتوای ویدئویی پیچیده

شناسایی محتوای آزاردهنده در تصاویر و ویدئوها با چالش‌های متعددی همراه است:

۱ .تغییرات زمینه‌ای و پیچیدگی محتوای بصری: تصاویر یا ویدئوها می‌توانند شامل محتوایی باشند که در ظاهر بی‌ضرر به نظر می‌رسد، اما در یک زمینه خاص آزاردهنده است. به عنوان مثال، نمادهای خاص یا ژست‌های مبهم ممکن است معنای خاصی در یک فرهنگ یا گروه داشته باشند.

۲ .تشخیص تصاویر دستکاری‌شده یا فیلترگذاری‌شده: تکنیک‌های ویرایشی مانند اعمال فیلترها یا تغییرات جزئی در تصاویر می‌توانند باعث فرار از شناسایی شوند.

۳ .پیچیدگی ویدئوها: تحلیل ویدئوها به دلیل وجود هزاران فریم، محاسبات سنگینی را می‌طلبد. علاوه بر این، ترکیب متن، تصویر و صوت در ویدئوها پیچیدگی بیشتری ایجاد می‌کند.

راهکارهای پیشنهادی:

- استفاده از معماری‌های پیشرفته مانند Vision Transformers (ViTs) و مدل‌های Multimodal Transformers که قادر به ترکیب اطلاعات از مختلف هستند(Dosovitskiy et al., 2021) .

- به‌کارگیری تکنیک‌های پیشرفته مانند adversarial training برای شناسایی تصاویر دستکاری‌شده

- استفاده از روش‌های نمونه‌برداری هوشمند برای کاهش هزینه محاسباتی در تحلیل ویدئوها، مانند انتخاب فریم‌های کلیدی (Key Frame Selection).

سیستم‌های شناسایی محتوای آزاردهنده با چالش‌های مهمی نظیر درک زمینه در متون، تحلیل زبان‌های غیررسمی و چندگانه، و شناسایی پیچیدگی‌های محتوای بصری مواجه هستند. حل این چالش‌ها نیازمند توسعه مدل‌های پیشرفته‌تر، استفاده از داده‌های متنوع‌تر، و بهره‌گیری از تکنیک‌های چندمدلی است. پیشرفت در این حوزه می‌تواند گام مهمی در بهبود ایمنی دیجیتال و تجربه کاربران باشد.

۴. نتیجه‌گیری

شناسایی و مدیریت محتوای آزاردهنده در فضای دیجیتال یکی از بزرگ‌ترین چالش‌های عصر حاضر است. با افزایش استفاده از شبکه‌های اجتماعی، پلتفرم‌های آنلاین، و محیط‌های تعاملی، نیاز به ابزارهایی که بتوانند به طور مؤثر و دقیق این نوع محتوا را شناسایی کرده و حذف کنند، بیش از پیش احساس می‌شود. این مقاله به بررسی تکنیک‌ها، دقت سیستم‌ها و چالش‌های فنی در شناسایی محتوای آزاردهنده پرداخته و راهکارهایی برای بهبود این فناوری‌ها ارائه داده است. در ادامه، یافته‌های کلیدی و اهمیت بهبود این سیستم‌ها مورد تأکید قرار گرفته و به ضرورت تحقیقات و همکاری‌های بیشتر در این حوزه پرداخته می‌شود.

۵. خلاصه‌ای از یافته‌های کلیدی

۵.۱. پیشرفت‌های تکنیکی در هوش مصنوعی :

الگوریتم‌های پردازش زبان طبیعی (NLP) مانند BERT و RoBERTa و معماری‌های یادگیری عمیق مانند CNNs و I3D به‌طور موفقیت‌آمیزی برای شناسایی محتوای متنی، تصویری، و ویدئویی آزاردهنده به کار گرفته شده‌اند. این روش‌ها با بهره‌گیری از داده‌های بزرگ و مدل‌های پیش‌آموزش‌دیده توانسته‌اند دقت قابل‌توجهی در محیط‌های مختلف دیجیتال ارائه دهند .

۵.۲. دقت سیستم‌ها در پلتفرم‌های اجتماعی :

گزارش‌های مربوط به پلتفرم‌های بزرگ مانند توییتر، فیسبوک، و اینستاگرام نشان می‌دهد که این سیستم‌ها توانسته‌اند بیش از ۹۰٪ از محتوای مضر را شناسایی کنند. برای مثال، فیسبوک در سال ۲۰۲۳ نرخ دقت ۹۷٪ را در شناسایی محتوای خشونت‌آمیز گزارش کرده است(Meta Community Standards Report, 2023) . با این حال، نرخ مثبت‌های کاذب (false positives) همچنان یک چالش است و باعث حذف اشتباه محتوای بی‌ضرر می‌شود.

۵.۳. چالش‌های باقی‌مانده :

شناسایی کنایه و طنز، تحلیل زبان‌های غیررسمی و چندگانه، و پردازش محتوای پیچیده بصری و ویدئویی از مهم‌ترین موانع فنی هستند که هنوز به راه‌حل‌های کامل دست نیافته‌اند. علاوه بر این، تحلیل محتوای آزاردهنده در فرهنگ‌ها و زبان‌های مختلف نیازمند توسعه مدل‌هایی است که بتوانند با داده‌های متنوع و محیط‌های چندفرهنگی سازگار باشند .

۶. اهمیت بهبود سیستم‌های هوش مصنوعی

بهبود سیستم‌های هوش مصنوعی برای شناسایی محتوای آزاردهنده از چند جهت اهمیت دارد :

۱ .حفظ امنیت روانی کاربران :

محتوای آزاردهنده تأثیرات مخربی بر سلامت روان کاربران، به‌ویژه کودکان و نوجوانان، دارد. سیستم‌های مؤثر هوش مصنوعی می‌توانند نقش مهمی در کاهش مواجهه با چنین محتوایی ایفا کنند .

۲ .ایجاد محیط دیجیتال سالم‌تر:

جلوگیری از گسترش نفرت‌پراکنی، خشونت آنلاین، و آزار و اذیت سایبری باعث ایجاد محیط‌های دیجیتال امن‌تر و جذاب‌تر برای کاربران می‌شود. این امر همچنین می‌تواند اعتماد کاربران به پلتفرم‌های آنلاین را افزایش دهد .

۳ .پاسخگویی به قوانین و مقررات :

دولت‌ها و نهادهای نظارتی در سراسر جهان قوانین سخت‌گیرانه‌ای برای کنترل محتوای مضر وضع کرده‌اند. سیستم‌های پیشرفته هوش مصنوعی می‌توانند به پلتفرم‌ها در تطابق با این قوانین کمک کنند و از تحریم‌های احتمالی جلوگیری کنند .

۴ .بهبود تجربه کاربری :

حذف محتوای آزاردهنده بدون تأثیر منفی بر محتوای سالم می‌تواند تجربه کاربری را بهبود بخشد و تعامل کاربران را با پلتفرم‌ها افزایش دهد .

۷. پیشنهادات

- تحقیقات میان‌رشته‌ای: توسعه سیستم‌های شناسایی محتوای آزاردهنده نیازمند تحقیقات میان‌رشته‌ای است که شامل دانش زبان‌شناسی، روان‌شناسی، علوم اجتماعی، و هوش مصنوعی می‌شود. این همکاری‌ها می‌توانند به درک بهتر محتوای آزاردهنده و توسعه راه‌حل‌های جامع‌تر کمک کنند.

-ایجادمجموعه داده‌های استاندارد : یکی از موانع بزرگ در توسعه سیستم‌های دقیق، کمبود مجموعه داده‌های استاندارد و متنوع است. ایجاد و به اشتراک‌گذاری مجموعه داده‌هایی که زبان‌ها، فرهنگ‌ها و محتوای چندرسانه‌ای مختلف را پوشش دهند، می‌تواند گامی بزرگ در این زمینه باشد.

-همکاری بین‌المللی : محتوای آزاردهنده یک چالش جهانی است که نیازمند همکاری‌های بین‌المللی میان دولت‌ها، شرکت‌های فناوری، و دانشگاه‌ها است. ایجاد چارچوب‌های مشترک برای تحلیل و مدیریت این محتوا می‌تواند بهبود عملکرد سیستم‌ها را تسریع کند.

- تمرکز بر مسائل اخلاقی: هرگونه پیشرفت در سیستم‌های هوش مصنوعی باید با در نظر گرفتن اصول اخلاقی، مانند حفظ حریم خصوصی کاربران، شفافیت در عملکرد سیستم‌ها، و جلوگیری از سوگیری‌های ناعادلانه صورت گیرد.

بهبود سیستم‌های هوش مصنوعی برای شناسایی محتوای آزاردهنده گامی ضروری در جهت ارتقای امنیت دیجیتال و سلامت اجتماعی است. با وجود پیشرفت‌های چشمگیر، چالش‌های موجود نشان‌دهنده نیاز به تحقیقات بیشتر، داده‌های بهتر، و همکاری‌های گسترده‌تر است. سرمایه‌گذاری در این حوزه نه‌تنها به کاهش محتوای مضر در محیط‌های آنلاین کمک می‌کند، بلکه باعث تقویت اعتماد کاربران به فناوری و ایجاد محیطی امن‌تر برای نسل‌های آینده خواهد شد.

منابع:

فارسی:

۱ .تبریزی، مصطفی.، استکی، مهناز. و ملاعلی حسینی، سمانه. (۱۳۹۳). مقایسه ی اختلالات رفتاری دانش آموزان سایبربولینگ و عادی. فصلنامه ی روانشناسی مدرسه،۴(۴)، ۶-۲۰.

۲. حاجی وند، امین.، خوش منظر، علی. و سیاری زهان، صابر. (۱۴۰۳). تاریخچه مختصری از هوش مصنوعی: گذشته، حال و آینده هوش مصنوعی. ویژه نامه هوش مصنوعی. ۶(۱۸). ۷۳-۹۰.

۳. حسینی، آمنه.، زندی، آمنه. و تدین، عباس. (۱۴۰۲). بررسی جرم قلدری مجازی در حقوق کیفری ایران و آمریکا. *پژوهش های حقوق تطبیقی.۲۷* (۳). ۷۶-۱۱۲.

لاتین:

1. Dosovitskiy, A., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition“, 2021, 10023-10033.

2. Kim, J., Kim, Y., & Lee, H. (2021). Sentiment Analysis Using Bidirectional LSTM and Attention Mechanism for Online Harassment Detection. “IEEE Access, 9“, 36784-36796.

3. Franco, L., & Ghanayim, K. (2019). The criminalization of Cyberbullying among children and Youth, Sanata Clara. *Journal of International Law*,Vol.17, PP. 1-49.

4. Meta Community Standards Report. (2023). Content Moderation Performance Metrics. [Online Report].

5. Nguyen, H., Tran, Q., & Vo, T. (2023). Deep Learning Models for Violent Content Detection in Video Streams. “Pattern Recognition Letters, 174“, 78-87.

6. Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. “International Conference on Machine Learning (ICML)“, 2021.

7. Sun, C., Myers, A., & Singh, V. (2022). VideoBERT: A Model for Video-Language Understanding. “Transactions on Multimedia“, 2022.

8. Twitter Transparency Report. (2023). Addressing Abuse and Harassment on Twitter. [Online Report].

9. Zhou, J., Li, C., & Yang, Z. (2022). Advances in Textual Abuse Detection Using Natural Language Processing Techniques. “Journal of Artificial Intelligence Research, 65“, 213-238.