

پژوهشی

تشخیص قلدری سایبری در شبکه های اجتماعی با یادگیری عمیق مبتنی بر شبکه عصبی CNN و LSTM

محسن اقبالی^۱ | کمال میرزائی^{۲*} | رضا عزیزی^۳

چکیده:

زورگویی سایبری نوعی از خشونت در شبکه های اجتماعی برای تحقیر، تمسخر، باج خواهی و تهدید است. زورگویی سایبری آثار زیانباری دارد و در برخی موارد قربانی آسیب جسمی و روحی شدید می بیند. یکی از رویکردهای امیدوارکننده در تشخیص زورگویی سایبری استفاده از الگوریتم های یادگیری ماشین و یادگیری عمیق است. با این حال، تشخیص آزار سایبری در شبکه های اجتماعی پیچیده است و یک الگوریتم یادگیری ماشین و یادگیری عمیق به تنهایی توانایی زیادی برای تشخیص دقیق زورگویی سایبری ندارند. در این مقاله برای تشخیص زورگویی سایبری در ابتدا با سه روش استخراج ویژگی GloVe، Word2Vec و TF-IDF ویژگی های اولیه متن استخراج می شود. در مرحله دوم انتخاب ویژگی با استفاده از الگوریتم JSO انجام می شود و در نهایت ویژگی های مهم به عنوان ورودی روش IDCNN و LSTM در نظر گرفته می شود. آزمایشات در مجموعه داده توئیتر و فیس بوک برای تشخیص زورگویی سایبری انجام می شود. آزمایشات نشان می دهد دقت، حساسیت و صحت روش پیشنهادی در تشخیص زورگویی سایبری در مجموعه داده توئیتر به ترتیب برابر ۹۸/۲۳ درصد، ۹۷/۸۶ درصد و ۹۷/۷۳ درصد است. نتایج نشان می دهد روش پیشنهادی نسبت به روشهای CNN، LSTM و BERT در تشخیص زورگویی سایبری دارای دقت بیشتری است.

^۱ دانشجوی دکتری مهندسی کامپیوتر، گروه مهندسی کامپیوتر، واحد میبد، دانشگاه آزاد اسلامی، میبد، ایران، m.eghbali@maybodiau.ac.ir

^۲ استادیار گروه مهندسی کامپیوتر، واحد میبد، دانشگاه آزاد اسلامی، میبد، ایران، k.mirzaie@maybodiau.ac.ir

^۳ استادیار گروه مهندسی کامپیوتر، واحد میبد، دانشگاه آزاد اسلامی، میبد، ایران، azizi.reza@maybodiau.ac.ir

نویسنده مسئول

* کمال میرزائی، استادیار گروه کامپیوتر، واحد میبد، دانشگاه آزاد اسلامی، میبد، ایران، k.mirzaie@maybodiau.ac.ir

تاریخ دریافت: ۱۴۰۳/۳/۲۹

تاریخ پذیرش: ۱۴۰۴/۱/۱۶

کلید واژه ها: شبکه اجتماعی، قلدری سایبری، یادگیری عمیق، شبکه عصبی کانولوشن، شبکه عصبی LSTM

۱-مقدمه

ظهور و پذیرش رسانه های اجتماعی به عنوان یک بستر ارتباطی و تعاملی مهم در دنیای امروز باعث شده تا این شبکه ها نقش مهمی در ارتباطات داشته باشند. مطالعات نشان می دهد بیش از ۳ میلیارد کاربر فعال وجود دارد که روزانه از طریق رسانه های اجتماعی با یکدیگر ارتباط و تعامل دارند [۱]. بنابراین، رسانه های اجتماعی برای تعریف و گسترش اطلاعات و محتوا در دنیای مدرن نقشی اساسی دارند. علاوه بر این، پلتفرم های رسانه های اجتماعی، جوامع مجازی مبتنی بر رابطه و علاقه را فعال می کنند که امکان اتصال و شبکه سازی را در بین مردم فراهم می کند. افزایش محبوبیت این پلتفرم های رسانه های اجتماعی (فیس بوک، توئیتر، اینستاگرام، Tinder و غیره) امکان اشتراک گذاری اشکال مختلف پیام های چندرسانه ای را در میان کاربران

خود فراهم می کند. عملیات این پلتفرم های رسانه های اجتماعی در طول همه گیری کووید ۱۹^۱ بسیار مهم بود، زیرا داده های بلادرنگ درباره رویدادها و اکتشافات به راحتی در بین مردم در مکان های مختلف منتشر می شد [۲].

اگرچه آشکار است که پلتفرم های رسانه های اجتماعی مزایای متعددی را برای کاربران خود ارائه می کنند، اما این پلتفرم ها ممکن است برای اهداف بدخواهانه نیز مورد سوء استفاده قرار گیرند. زورگویی سایبری^۲ نمونه بارز و عمیقی از این روندهای بدخواهانه است [۳]. به طور خاص، فراگیر شدن این پلتفرم های رسانه های اجتماعی به طور غیرقابل انکاری جرقه، پیشرفت و تشدید قلندری را در میان کاربران آن ایجاد کرده است. به نظر می رسد قلندری در طول تاریخ تمدن بشری وجود داشته است و مستلزم آزار و اذیت شخصی از طریق شرمساری یا آزار دادن او به هر طریقی است که باعث آسیب عاطفی، روانی یا جسمی شود. هنگامی که این حمله از طریق اینترنت رخ می دهد، به عنوان آزار سایبری یا قربانی سایبری شناخته می شود [۴]. قلندری سایبری را می توان به عنوان قلندری به شخص یا گروهی از افراد که معمولاً به عنوان "قربانی (ها)" شناخته می شوند، با کمک اینترنت، موبایل یا دستگاه الکترونیکی با ارسال محتوای متنی یا غیر متنی نامناسب چند رسانه ای توصیف کرد. به عبارت دیگر، آزار و اذیت سایبری نمایش مستمر اعمال ناخوشایندی است که بر روی «قربانی (ها)» برای ایجاد ترس، آزار، درد یا آسیب از طریق رسانه های الکترونیکی و بسترهای رسانه های اجتماعی انجام می شود. آزار و اذیت سایبری یک مشکل بزرگ در دهه گذشته در فضای مجازی بوده است که بیشتر کودکان و نوجوانان را تحت تأثیر قرار داده است [۵]. به عنوان مثال، یک مطالعه مستقر در ایالات متحده (ایالات متحده) گزارش داد که بیش از ۴۳٪ از نوجوانان در ایالات متحده مورد آزار و اذیت سایبری قرار می گیرند [۶]. بر اساس آمار، حدود ۱۸ درصد از جوانان اروپایی تحت تأثیر فردی قرار گرفته اند که آنها را از طریق اینترنت و تلفن همراه مورد آزار و اذیت قرار می دهد [۷]. همانطور که در گزارش آنلاین کودکان اتحادیه اروپا در سال ۲۰۱۴ بیان شد، بیش از ۲۰ درصد از کودکان (در سنین بین ۱۱ تا ۱۶ سال) آزار سایبری را تجربه می کنند [۸]. در ایران متأسفانه آماری دقیق از زورگویی سایبری در فضای مجازی و شبکه های اجتماعی انتشار پیدا نکرده است اما در سایر کشورها این چالش فضای مجازی گزارش شده است و به عنوان نمونه در سوئد که کشوری توسعه یافته است، شیوع آزار و اذیت سایبری به نقطه اوج رسیده و به تدریج در حال رشد و بدتر شدن است [۹]. این گزارش ها نشان می دهد که توسعه یک راه حل قابل قبول، سریع و آزمایش شده برای این مشکل مبتنی بر اینترنت چقدر حیاتی است. بنابراین، ارزیابی و رسیدگی به آزار و اذیت سایبری از منظرهای مختلف، از جمله شناسایی خودکار و پیشگیری از چنین حوادثی ضروری است. با توجه به توسعه فناوری، امتیازات اجتماعی (ناشناس بودن) که رسانه های اجتماعی ارائه می کنند و دسترسی به مخاطبان گسترده تر، آزار و اذیت سایبری به طور تصاعدی رشد کرده است. این موضوع نیازمند توسعه ابزارها و استراتژی های هوشمندی است که با استفاده از داده های چندرسانه ای اجتماعی موجود، آزار و اذیت سایبری را شناسایی، شناسایی و تجزیه و تحلیل می کنند تا تأثیر مضر آن را کاهش دهند. تشخیص خودکار مزاحمت سایبری یک موضوع طبقه بندی است، با هدف دسته بندی هر نظر/پست/پیام/تصویر توهین آمیز به عنوان قلندری یا غیر آزاردهنده [۱۰].

اگرچه برخی از پلتفرم های رسانه های اجتماعی، مانند یوتیوب و توئیتر، مراکز ایمنی را برای نظارت و کنترل آزار و اذیت سایبری تعبیه کرده اند، اما این مشکل همچنان پابرجاست و نیاز به راه حل های قطعی تری وجود دارد. اخیراً شناسایی خودکار آزار و اذیت سایبری با استفاده از یادگیری عمیق [۱۱] و یادگیری ماشین [۱۲] مورد توجه زیادی قرار گرفته است. روش های یادگیری ماشین و یادگیری عمیق می توانند الگوی متن انتشار یافته برای تشخیص زورگویی سایبری را استخراج نمایند و تئیه های عادی را از تئیه های زورگویی سایبری تشخیص دهند. از جمله روش های یادگیری عمیق که برای تشخیص زورگویی سایبری استفاده می شود می توان به شبکه عصبی کانولوشن [۱۳]، شبکه عصبی بازگشتی [۱۴]، شبکه عصبی مبتنی بر دروازه [۱۵]، شبکه عصبی حافظه کوتاه مدت [۱۶] اشاره نمود. در این روش ها برای تشخیص زورگویی سایبری تلاش شده

^۱ COVID-19^۲ Cyberbullying^۳ Deep learning (DL)^۴ Machine learning (ML)^۵ Convolutional neural network (CNN)^۶ Recurrent neural network (RNN)^۷ Gated recurrent units (GRUs)^۸ Long Short-Term Memory (LSTM)

است تا در ابتدا ویژگی‌های متن استخراج شود و سپس در ادامه ویژگی‌ها در اختیار یک روش طبقه‌بندی قرار داده شود. در این روش‌ها چون انتخاب ویژگی هوشمندانه نیست لذا تشخیص زورگویی سایبری با خطا مواجه است. در روش پیشنهادی برای کاهش دادن خطای تشخیص زورگویی سایبری از یک معماری هوشمند استفاده می‌شود. در روش پیشنهادی برای تشخیص زورگویی در شبکه اجتماعی در ابتدا ویژگی‌ها توسط دو روش GloVe, Word2Vec استخراج شده و در ادامه ویژگی‌های مهم توسط الگوریتم عروس دریایی^۱ [۱۷] انتخاب شده و تحویل طبقه‌بندی کننده ترکیبی CNN و LSTM می‌شود. هدف از این مقاله، ارائه یک رویکرد تشخیص زورگویی سایبری در شبکه اجتماعی با استفاده از ابزارهای هوشمند است. هدف دیگر روش پیشنهادی در این مقاله، کاهش آسیب‌های روحی و روانی به کاربران شبکه‌های اجتماعی با تشخیص زود هنگام زورگویی سایبری است. به طور خاص، این مقاله دارای سهم چندگانه به شرح زیر است:

- ارائه یک روش استخراج ویژگی بر اساس ترکیب سه روش Word2Vec, GloVe و TF-IDF
- ارائه یک نسخه باینری از الگوریتم عروس دریایی
- تلفیق دو طبقه‌بندی کننده CNN و LSTM برای تشخیص زورگویی سایبری
- تلفیق هوش گروهی و یادگیری عمیق در تشخیص زورگویی سایبری

این مقاله یک سیستم تشخیص زورگویی سایبری کارآمد بر اساس معماری یادگیری عمیق و هوش گروهی ارائه می‌دهد. در بخش II کارهای مرتبط در زمینه تشخیص زورگویی سایبری ارائه می‌دهد. در بخش III، سیستم تشخیص زورگویی سایبری بر اساس یادگیری عمیق و الگوریتم بهینه‌سازی عروس دریایی توسعه داده شده است. در بخش IV، روش پیشنهادی پیاده‌سازی و با روشهای مشابه مقایسه می‌شود. در بخش V نتایج تحقیق و یافته‌های تحقیق به همراه پیشنهادات آتی ارائه می‌گردد.

۲-پیشینه تحقیق

روابط اجتماعی در دوران نوجوانی برای رشد اجتماعی ضروری است، زیرا زمانی است که فرد شروع به یادگیری تعامل با جامعه، ایجاد اولین دوستی‌ها و حل اولین تعارضات می‌کند و گاهی اوقات به خشونت تبدیل می‌شوند مانند قلدری. چنین موقعیت‌هایی معمولاً بر زندگی جوانانی که درگیر آن بوده‌اند، بازتاب منفی دارد. با این حال، قلدری نه تنها در مدرسه اتفاق می‌افتد. با توسعه فناوری های جدید اطلاعات و ارتباطات، آزار و اذیت سایبری ظهور کرده است [۱۸].

تأثیر شدید افزایش مداوم و وابستگی به فناوری اطلاعات و ارتباطات، محتوا، عادات و اشکال روابط بین فردی را تغییر داده است. قلدری سایبری هر دو به عنوان یک مشکل شدید بهداشت عمومی شناخته شده‌اند، زیرا تهدیدی برای توسعه سلامت روان و رفاه کودکان و نوجوانان هستند. قلدری سایبری و زورگویی سایبری به نوعی خشونت‌های اشاره دارند که در آن آزار و اذیت سایبری به طور خاص از فناوری اطلاعات و ارتباطات برای آزار و اذیت همسالان استفاده می‌کند. قربانیان قلدری سایبری معمولاً از نظر عاطفی و شدیدتر از قربانیان قلدری سنتی آسیب می‌بینند. در بیشتر موارد، این دو پدیده بر هم منطبق هستند، اما مزاحمت سایبری اغلب به طور جداگانه مورد بحث قرار نمی‌گیرد. برخی از مطالعات نشان می‌دهد رد شیوع آزار و اذیت اینترنتی در اروپا در حدود ۶/۵ درصد و در آمریکا در حدود ۳۵/۴ رشد داشته است [۱۸].

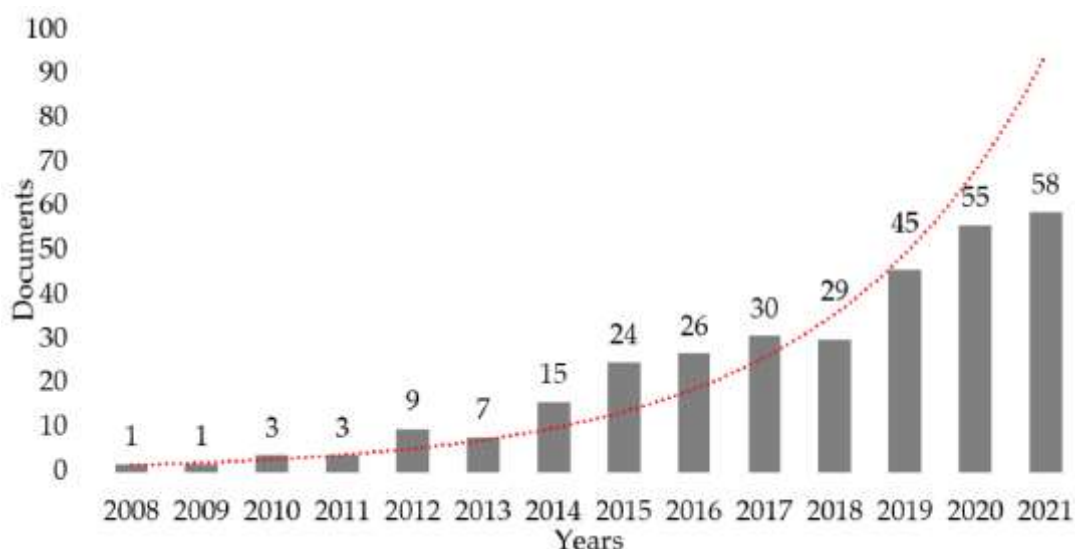
با توجه به ابعاد مختلف روانشناختی، از یک سو، قربانیان سایبری از اثرات منفی مختلفی مانند افزایش علائم افسردگی و اضطراب و استرس عاطفی بیشتر رنج می‌برند. علاوه بر این، ابعاد عاطفی نیز می‌تواند تحت تأثیر قرار گیرد، زیرا آزار و اذیت سایبری باعث کاهش عزت نفس، نارضایتی از زندگی و خودکارآمدی عاطفی پایین می‌شود. تنهایی و خودپنداره پایین برخی از رایج‌ترین اثرات و پیش‌بینی کننده‌ها در میان قربانیان است. معمولاً قربانیان از تنهایی، عدم حمایت خانوادگی و اجتماعی نبود دوستانی که بتوانند در برابر حملات و پیامدهای آن محافظت کنند، رنج می‌برند، بنابراین جنبه‌های اجتماعی-فرهنگی نیز آشکار می‌شود. از سوی دیگر، زورگویان سایبری همچنین استرس، اضطراب و نارضایتی زیادی از زندگی را نشان می‌دهند و این خودکنترلی پایین همراه با سطوح همدلی عاطفی ضعیف‌تر، پیش‌بینی کننده‌ای برای ارتکاب آزار سایبری هستند. علاوه بر این، داشتن قربانی قلدری ممکن است با افزایش تمایل به زورگویی سایبری همراه باشد و انجام قلدری سنتی نیز با ارتکاب

¹ Jellyfish search algorithm

قلندری سایبری در زمان بعدی مرتبط است. تفاوت های اساسی بین قربانیان سایبری و زورگویان اینترنتی و جنسیت در مورد عزت نفس، خودکنترلی، حمایت اجتماعی و خانوادگی و پرخاشگری یافت شده است.

۱-۲- اهمیت تشخیص زورگویی سایبری

در شکل (۱)، تعداد مطالعات در زمینه زورگویی سایبری بین سالهای ۲۰۰۸ تا ۲۰۲۱ را نشان می دهد. با توجه به نمودار می توان گفت که موضوع زورگویی سایبری در سالهای اخیر بسیار مورد توجه پژوهشگران قرار گرفته شده است و این موضوع اهمیت تشخیص زورگویی سایبری را نشان می دهد [۱۹].

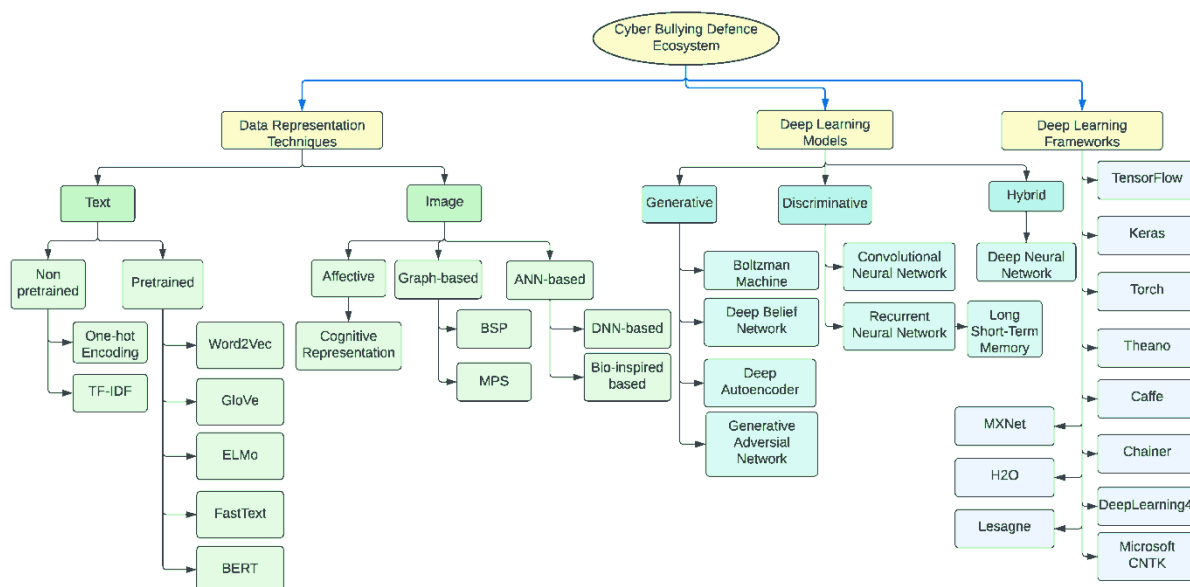


شکل ۱: افزایش مطالعات زورگویی سایبری در شبکه های اجتماعی در سالهای اخیر [۱۹]

شناسایی آزار و اذیت سایبری برای جلوگیری از مشکل تهدید کننده در شبکه های اجتماعی مهم است. تشخیص آزار و اذیت سایبری به دلیل فقدان پارامترهای قابل شناسایی و عدم وجود استاندارد قابل سنجش، کار دشواری است.

۲-۲- روشهای تشخیص زورگویی سایبری

در شکل (۲)، یک دسته بندی اصلی برای تشخیص زورگویی سایبری نمایش داده شده است [۲۰].



شکل ۲: دسته بندی روشهای تشخیص زورگویی سایبری در شبکه های اجتماعی [۲۰]

با توجه به شکل (۲)، محققان از الگوریتم های یادگیری ماشین سنتی برای شناسایی آزار سایبری (به عنوان مثال، قالب متن و تصویر) استفاده می کنند، در حالی که اکثر راه حل های موجود مبتنی بر روش های یادگیری نظارت شده هستند. به دلیل ماهیت ذهنی عبارات قلدر، مدل های سنتی یادگیری ماشین در تشخیص آزار سایبری نسبت به رویکردهای مبتنی بر یادگیری عمیق عملکرد کمتری دارند. گزارشات بر اساس [۲۰] نشان می دهد، که مدل های یادگیری عمیق از الگوریتم های سنتی یادگیری ماشین در مورد شناسایی آزار سایبری بهتر عمل می کنند. شبکه های عصبی عمیق مانند شبکه عصبی بازگشتی، واحد بازگشتی دروازه ای، حافظه کوتاه مدت بلند مدت و چندین مدل یادگیری عمیق دیگر را می توان برای تشخیص زورگویی سایبری استفاده کرد.

۲-۳- کارهای مرتبط

در [۲۱]، در سال ۲۰۲۳، راه حل های تشخیص آزار سایبری بر اساس معماری های یادگیری عمیق را بررسی کردند. مطالعات آنها نشان می دهد آزار و اذیت سایبری، سوء رفتار آنلاین آزاردهنده و نگران کننده است. به اشکال مختلف ظاهر می شود و معمولاً در اکثر شبکه های اجتماعی به صورت متنی است. سیستم های هوشمند برای تشخیص خودکار این حوادث ضروری است. برخی از آزمایش های اخیر این مسئله را با مدل های یادگیری ماشین سنتی حل کرده اند. اکثر مدل ها در یک زمان در یک شبکه اجتماعی اعمال شده اند. آخرین تحقیقات نشان داده است که مدل های مختلف مبتنی بر الگوریتم های یادگیری عمیق بر تشخیص آزار سایبری تأثیر می گذارند. این پژوهش یک تحلیل تجربی برای تعیین اثربخشی و عملکرد الگوریتم های یادگیری عمیق در تشخیص توهین در تفسیر اجتماعی انجام می دهد. چهار مدل یادگیری عمیق زیر برای نتایج تجربی استفاده شده است که عبارتند از، حافظه کوتاه مدت دو جهته (BiLSTM)، واحدهای بازگشتی دردار (GRU)، حافظه کوتاه مدت طولانی (LSTM)، و شبکه عصبی بازگشتی (RNN). مراحل پیش پردازش داده ها دنبال شد که شامل پاکسازی متن، نشانه گذاری، ریشه یابی، Lemmatization و حذف کلمات توقف می شود. پس از انجام پیش پردازش داده ها، داده های متنی تمیز برای پیش بینی به الگوریتم های یادگیری عمیق منتقل می شوند. نتایج نشان می دهد که مدل BLSTM در مقایسه با RNN، LSTM و GRU دارای دقت بیشتری در تشخیص زورگویی سایبری یافته است. در [۲۲]، در سال ۲۰۲۳، یک مدل گروهی یادگیری عمیق برای شناسایی آزار و اذیت سایبری در دوران شیوع بیماری کووید ۱۹ ارائه شده است. آنها برای تشخیص زورگویی سایبری در این پژوهش از الگوریتم های یادگیری ماشین مختلف و یادگیری عمیق مانند شبکه عصبی کانولوشن، BiLSTM، BERT را برای تشخیص آزار سایبری انگلیسی- هندی استفاده نمودند. ارزیابی آنها نشان می دهد دقت مدل های یادگیری عمیق در تشخیص زورگویی سایبری بیشتر از روش های یادگیری ماشین است. در [۲۳]، در سال ۲۰۲۳، تشخیص آزار و اذیت سایبری در رسانه های اجتماعی با استفاده از مدل سازی یادگیری عمیق ارائه شده است. رویکرد پیشنهادی آنها مدل سازی موضوعی مبتنی بر خوشه بندی بهینه سازی تعادل تطبیقی فازی (FAEO) و شبکه عصبی پیچیده (ECNN) را برای افزایش دقت فرایند تشخیص زورگویی سایبری را ادغام می کند. در ابتدا، پیش پردازش به منظور پاکسازی مجموعه داده انجام می شود. در مرحله بعد، ویژگی ها با استفاده از چندین مدل استخراج می شوند. بهینه سازی تعادل تطبیقی فازی بدون نظارت برای کشف موضوعات پنهان از داده های ورودی از پیش پردازش شده استفاده می شود، که به طور خودکار داده های متن را بررسی می کند و خوشه هایی از کلمات را ایجاد می کند. در نهایت، طبقه بندی آزار سایبری از الگوریتم ECNN و الگوریتم بهینه سازی باران برای شناسایی زورگویی سایبری از پست ها/متون استفاده می کند. مدل پیشنهادی آنها در تشخیص آزار سایبری از مدل های مانند شبکه عصبی کانولوشن بهتر عمل کرد. در [۲۴]، در سال ۲۰۲۳، یک شبکه عصبی کانولوشن ترکیبی با انتخاب ویژگی N-gram برای تشخیص آزار سایبری در شبکه های اجتماعی آنلاین ارائه دادند. در این پژوهش مکانیسم یادگیری عمیق را با مدل شبکه عصبی کانولوشنال هفت لایه با انتخاب ویژگی N-gram، به نام CNN عمیق ترکیبی اتخاذ شده است. با استفاده از این مدل در چهار معماری مرحله ای، تشخیص آزار سایبری سطح کلمه و سطح کاراکتر به طور موثری انجام می شود. مهاجمان با استفاده از انواع کلمات مبهم و توهین آمیز تأثیر منفی بر شبکه های اجتماعی دارند و لذا الگوی تشخیص توهین ها پیچیده است و نیاز به ابزارهای مارآمد مانند یادگیری عمیق دارد. این تحقیق عمدتاً بر روی

تشخیص آزار سایبری در سطح کاراکتر مترادف تمرکز دارد که در اینجا به مدل CyberNet اشاره می شود. تجزیه و تحلیل نتایج کیفی ثابت می کند که رویکرد پیشنهادی عملکرد بسیار بهتری را در مقایسه با رویکردهای مرسوم ارائه می دهد.

در [۲۵]، در سال ۲۰۲۳، تحلیل عملکرد تکنیک های تشخیص حاشیه نویسی برای پیام های قلندری سایبری با استفاده از شبکه های عصبی عمیق تعبیه شده در کلمه را پیشنهاد نمودند. پیام های رسانه های اجتماعی ساختاری ندارند زیرا شامل متن، پیوند URL، شکلک ها، اختصارات و غیره است. بیشتر کارهای قبلی برای شناسایی پیام های قلندری تنها با در نظر گرفتن کلمات مهم در متن انجام شده اند، و از سایر ویژگی های پیام مانند لینک های URL، ایموجی ها غفلت می کنند. در این پژوهش، یک تکنیک پیش پردازش پیشرفته با در نظر گرفتن برخی از ویژگی های موجود در پیام ها مانند URL، مخفف، شماره، ایموجی ها و غیره برای شناسایی پیام های قلندری پیشنهاد شده است. در این کار، شش مدل، یعنی سه مدل یادگیری عمیق ترکیب شده با دو مدل مختلف جاسازی کلمه برای تشخیص حاشیه نویسی به کار گرفته شده اند. عملکرد هر یک از این شش مدل دو بار با استفاده از پیش پردازش سنتی و پیش پردازش پیشرفته پیشنهادی اندازه گیری می شود. ارزیابی آنها نشان می دهد روش پیشنهادی بکار رفته آنها از روش های یادگیری عمیق مانند شبکه عصبی کانولوشن دقت بیشتری دارد. در [۲۶]، در سال ۲۰۲۳، تشخیص آزار سایبری بر اساس احساسات را پیشنهاد دادند. انگیزه تحقیق ارائه شده در این پژوهش این است که احساسات منفی می تواند توسط آزار و اذیت سایبری ایجاد شود. این پژوهش مدل های تشخیص آزار سایبری را پیشنهاد می کند که بر اساس ویژگی های زمینه ای، احساسات و احساسات آموزش داده می شوند. یک مدل تشخیص احساسات (EDM) با استفاده از مجموعه داده های توییت که از نظر حاشیه نویسی بهبود یافته اند، ساخته شد. نتایج نشان می دهد که خشم، ترس و احساس گناه اصلی ترین احساسات مرتبط با آزار و اذیت اینترنتی بودند. متعاقباً، احساسات استخراج شده به عنوان ویژگی علاوه بر ویژگی های زمینه ای و احساسی برای آموزش مدل هایی برای تشخیص آزار سایبری استفاده شد. در [۲۷]، در سال ۲۰۲۳، یک روش ترکیبی مبتنی بر یادگیری فدرال، جاسازی کلمات، و ویژگی های احساسی برای تشخیص آزار سایبری ارائه دادند. در این مطالعه، آنها یک تحقیق عمیق در مورد ادغام جاسازی های کلمه، ویژگی های احساسی و یادگیری فدرال انجام دادند که با استفاده از BERT، شبکه های عصبی کانولوشنال (CNN)، شبکه های عصبی عمیق (DNN) و مدل های حافظه (LSTM) ارائه شده است. فرآیندها و معماری عصبی برای یافتن پیکربندی های بهینه مورد بررسی قرار می گیرند که منجر به تولید نتایج برتر می شود. این تکنیک ها در زمینه تشخیص آزار سایبری، با استفاده از مجموعه داده های آزار سایبری چند پلتفرمی (رسانه های اجتماعی) در دسترس عموم استفاده می شوند. نتایج نشان دهنده توانایی افزایش یافته برای شناسایی و مبارزه موثر با حوادث مزاحم سایبری است که به ایجاد محیط های آنلاین امن تر کمک می کند. به ویژه، مدل BERT به طور مداوم از سایر مدل های یادگیری عمیق (CNN، DNN، LSTM) در تشخیص آزار سایبری بهتر عمل می کند و در عین حال حریم خصوصی مجموعه داده های محلی را برای هر پلتفرم اجتماعی از طریق تنظیمات یادگیری فدرال بهبود یافته ما حفظ می کند. در [۲۸]، در سال ۲۰۲۳، تشخیص قلندری سایبری در رسانه های اجتماعی با استفاده از آموزش گروهی و BERT را پیشنهاد دادند. این پژوهش یک رویکرد یادگیری انباشته برای تشخیص آزار سایبری در توییت با استفاده از ترکیبی از روش های شبکه عصبی عمیق ارائه می کند. همچنین این پژوهش یک مدل BERT اصلاح شده را معرفی می کند. مجموعه داده مورد استفاده در این مطالعه از توییت جمع آوری شده و برای حذف اطلاعات نامربوط پیش پردازش شده است. فرآیند استخراج ویژگی شامل استفاده از word2vec به همراه CBOW برای تشکیل وزن ها در لایه جاسازی بود. این ویژگی ها سپس به یک مکانیسم کانولوشنی و ادغام تبدیل شدند، و به طور موثر ابعاد آنها را کاهش دادند و ویژگی های تغییرناپذیر موقعیت کلمات توهین آمیز را به دست آوردند. اعتبار سنجی مدل انباشته پیشنهادی و BERT-M با استفاده از معیارهای ارزیابی مدل شناخته شده انجام شد. نتایج آزمایشات نشان داد که رویکرد یادگیری گروه انباشته به دقت ۹۷.۴ درصد در تشخیص آزار سایبری در مجموعه داده توییت و ۹۰.۹۷ درصد در مجموعه داده های ترکیبی توییت و فیس بوک دست یافت. در [۲۹]، در سال ۲۰۲۴، رویکرد مبتنی بر یادگیری ماشین برای شناسایی زورگویی سایبری ارائه شده است. در این پژوهش درخت های تصمیم گیری، جنگل تصادفی و XGBoost، بکارگرفته شده است. نتایج آزمایش ها نشان داد که این چارچوب می تواند شش نوع مختلف آزار و اذیت سایبری را با دقت ۰.۹۰۷۱ به طور مؤثرتری شناسایی کند. در [۳۶]، سال ۲۰۲۴، یک الگوریتم بهینه سازی حشره شبتاب برای تشخیص آزار سایبری در رسانه های اجتماعی با آموزش گروهی ارائه شده است. این مطالعه بر

طراحی و توسعه یادگیری عمیق گروهی با الگوریتم بهینه‌سازی ازدحام کرم شب تاب برای تشخیص و طبقه‌بندی آزار سایبری در داده‌های توییتر متمرکز است. هدف این مطالعه بررسی داده‌های رسانه‌های اجتماعی از طریق استفاده از پردازش زبان طبیعی (NLP) و فرآیند یادگیری گروهی است. این تکنیک EDL-TSGSO توییت‌های خام را از قبل پردازش می‌کند و سپس از تکنیک جاسازی کلمه Glove استفاده می‌کند. علاوه بر این، تکنیک ارائه‌شده از حافظه کوتاه‌مدت مجموعه با مدل Adaboost برای تشخیص و طبقه‌بندی موثر آزار سایبری استفاده می‌کند. در [۳۷]، سال ۲۰۲۴، یک روش بهینه‌سازی ماشین بردار پشتیبانی با عملکرد هسته مکعبی برای تشخیص آزار سایبری در شبکه‌های اجتماعی ارائه دادند. در این مطالعه تشخیص آزار و اذیت سایبری در برخورد با داده‌های سیاست دولتی مانند "cipta kerja" با استفاده از روش SVM که با استفاده از تابع هسته مکعبی بهینه شده است، انجام می‌شود. مقدار دقت به دست آمده در روش آنها برابر ۹۲/۳ درصد است و این در حالی است که دقت ماشین بردار پشتیبان برابر ۹۰ درصد است. در [۳۸]، سال ۲۰۲۴، یک مدل یادگیری عمیق ترکیبی برای تشخیص پیشگیرانه آزار سایبری در رسانه‌های اجتماعی ارائه دادند. این مطالعه از یک مدل ترکیبی تصادفی تصادفی مبتنی بر جنگل CNN برای طبقه‌بندی متن استفاده کرده است که نقاط قوت هر دو رویکرد را ترکیب می‌کند. مجموعه داده‌های بلادرنگ از توییتر و اینستاگرام جمع‌آوری و حاشیه‌نویسی شد تا اثربخشی تکنیک پیشنهادی را نشان دهد. عملکرد الگوریتم‌های مختلف ML و DL مقایسه شد و مدل CNN مبتنی بر RF از نظر دقت و سرعت اجرا بهتر از آنها بود. این مدل به دقت ۹۶ درصد دست یافت و نتایج را ۳.۴ ثانیه سریعتر از مدل‌های استاندارد CNN ارائه کرد. در [۳۹]، سال ۲۰۲۴، یک روش تشخیص خودکار آزار و اذیت سایبری با استفاده از رویکرد ترکیبی SVM و NLP ارائه دادند. این مقاله یک رویکرد نوآورانه برای شناسایی خودکار آزار و اذیت سایبری در متن آنلاین با استفاده از ترکیبی از پردازش زبان طبیعی (NLP)، طبقه‌بندی‌کننده‌های ماشین بردار پشتیبانی (SVM)، فرکانس مدت-فرکانس بار معکوس سند (TF-IDF)، و تحقیق زبانی و ابزار شمارش کلمات (LIWC2) ارائه دادند. سیستم پیشنهادی از قدرت NLP برای پیش‌پردازش و تجزیه و تحلیل داده‌های متنی استفاده می‌کند و امکان استخراج ویژگی‌های ضروری را که نشان‌دهنده آزار سایبری است، می‌دهد. طبقه‌بندی‌کننده‌های SVM سپس برای طبقه‌بندی نمونه‌های متنی به دسته‌های مزاحم سایبری یا غیر آزاردهنده استفاده می‌شوند و دقت پیش‌بینی مدل را افزایش می‌دهند. برای دریافت جنبه‌های معنایی و متنی متن، از TF-IDF برای سنجش اهمیت کلمات در مجموعه سند استفاده می‌شود. این به تمایز بین زبان رایج و کلمات خاص در موارد زورگویی سایبری کمک می‌کند. علاوه بر این، LIWC2 برای استخراج بینش‌های زبانی و روان‌شناختی استفاده می‌شود و به شناسایی الگوهای عاطفی و روان‌شناختی مرتبط با آزار سایبری کمک می‌کند. آزمایش‌های انجام‌شده بر روی مجموعه داده‌های دنیای واقعی، توانایی سیستم را در شناسایی دقیق موارد زورگویی سایبری، ارائه بینش‌های ارزشمند برای محققان، سیاست‌گذاران و پلتفرم‌های آنلاین در نبرد مداوم علیه آزار و اذیت آنلاین نشان می‌دهد. این تحقیق نشان دهنده گام مهمی در توسعه ابزارهای خودکار برای مبارزه با آزار سایبری و محافظت از رفاه ذهنی و عاطفی کاربران آنلاین است. ارزیابی‌ها نشان داد دقت روش آنها برابر ۹۳.۱۵ درصد است. در جدول (۱)، خلاصه‌ای از مطالعات و مزیت و معایب آنها و حوزه هر یک ارائه شده است.

جدول ۱: خلاصه کارهای مرتبط

پژوهش	رویکرد	مزیت	چالش	حوزه
در [۲۱]، در سال ۲۰۲۳،	مرور روشهای تشخیص آزار سایبری بر اساس معماری های یادگیری عمیق	اثبات کارایی روشهای مبتنی بر شبکه عصبی LSTM	عدم بررسی تعادل و بالانس مجموعه داده ها و تاثیر آن بر دقت	مقاله مروری در حوزه یادگیری عمیق
در [۲۲]، در سال ۲۰۲۳،	مدل گروهی یادگیری عمیق	دقت بیشتر از شبکه عصبی کانولوشن، BERT، BiLSTM	زمان یادگیری زیاد و پیچیدگی بالا و عدم انتخاب ویژگی	یادگیری عمیق-یادگیری گروهی
در [۲۳]، در سال ۲۰۲۳،	تطبیقی فازای (FAEO) و شبکه عصبی پیچیده (ECNN)	از مدل های مانند شبکه عصبی کانولوشن بهتر عمل کرد.	پیچیدگی بالا و عدم انتخاب ویژگی هوشمندانه	یادگیری عمیق-منطق فازای
در [۲۴]، در سال ۲۰۲۳،	شبکه عصبی کانولوشن ترکیبی با انتخاب ویژگی N-gram	دقت بیشتر از CNN	عدم انتخاب ویژگی هوشمندانه	یادگیری عمیق
در [۲۵]، در سال ۲۰۲۳،	شبکه های عصبی عمیق تعبیه شده در کلمه	شامل ویژگی های بصری و متن مانند متن پیام، پیوند URL، شکلک ها، اختصارات و غیره است.	عدم انتخاب ویژگی هوشمندانه و عدم تعادل مجموعه داده	یادگیری عمیق
در [۲۶]، در سال ۲۰۲۳،	تشخیص آزار سایبری بر اساس تحلیل احساسات	دقت بالا	فقط احساسات ملاک در نظر گرفته شده است.	یادگیری عمیق
در [۲۷]، در سال ۲۰۲۳،	روش ترکیبی مبتنی بر یادگیری فدرال، جاسازی کلمات	دقت بیشتر از CNN، DNN، LSTM	پیچیدگی بالای BERT	یادگیری عمیق و پردازش زبان طبیعی
در [۲۸]، در سال ۲۰۲۳،	آموزش گروهی و شبکه BERT	دقت ۹۷.۴ درصد در تشخیص آزار سایبری	پیچیدگی بالای BERT عدم تعادل مجموعه داده	یادگیری گروهی و پردازش زبان طبیعی
در [۲۹]، در سال ۲۰۲۴،	درخت های تصمیم گیری، جنگل تصادفی و XGBoost	عدم پیچیدگی	دقت اندک	یادگیری ماشین
در [۳۶]، سال ۲۰۲۴،	الگوریتم بهینه سازی حشره شب تاب با پردازش زبان طبیعی	دقت بالا	عدم متعادل سازی مجموعه داده	هوش گروهی
در [۳۷]، سال ۲۰۲۴،	ماشین بردار پشتیبان با کرنل مکعبی	دقت بیشتر از ماشین بردار پشتیبان	دقت اندک	یادگیری ماشین
در [۳۸]، سال ۲۰۲۴،	مدل ترکیبی تصادفی مبتنی بر جنگل CNN	دقت بیشتر از CNN و ۳.۴ ثانیه سریعتر از مدل های استاندارد	پیچیدگی بیشتر از CNN	یادگیری عمیق
در [۳۹]، سال ۲۰۲۴،	رویکرد ترکیبی SVM و NLP	استخراج بینش های زبانی و روان شناختی	دقت نسبتاً اندک	یادگیری ماشین و پردازش زبان طبیعی

بررسی کارهای مرتبط نشان می دهد که هر کدام از آنها دارای مزایا و معایبی می باشند. بررسی کارها نشان می دهد روش های یادگیری عمیق برای تشخیص زورگویی سایبری یک روش موثر و کارآمد است اما با این وجود چالشهای برای آن وجود دارد. عدم تعادل یا بالانس در مجموعه داده آموزشی یک چالش مهم است که دقت روشهای یادگیری عمیق را کاهش می دهد. در روش پیشنهادی برای رفع این چالش از تئوری بازی و شبکه عصبی GAN برای متعادل سازی مجموعه داده استفاده می شود تا تعداد نمونه های کلاس اقلیت افزایش داده شود. یکی از چالشهای دیگر روشهای تشخیص حملات سایبری عدم استفاده از تکنیکهای هوش گروهی جدید برای تشخیص ویژگی های مهم و کاهش ورودی روشهای یادگیری عمیق است که در روش پیشنهادی به کمک الگوریتم عروس دریایی این چالش بر طرف شده است. مزیت اصلی کار و روش پیشنهادی نسبت به روشهای موجود را می توان در موارد ذیل خلاصه نمود:

- استخراج ویژگی با سه ترکیب GloVe، Word2Vec و TF-IDF و استفاده از توانایی سه روش در استخراج ویژگی
- انتخاب هوشمندانه ویژگی های مرتبط با زورگویی سایبری با الگوریتم عروس دریایی
- تلفیق معماری LSTM و CNN در تشخیص زورگویی سایبری
- متعادل سازی مجموعه داده برای افزایش دقت مدل یادگیری با تئوری بازی و شبکه عصبی GAN

۳- روش پیشنهادی

روش پیشنهادی در این بخش برای تشخیص زورگویی سایبری ارائه می شود. روش پیشنهادی برای تشخیص زورگویی سایبری دارای نوآوری های ذیل است که در ادامه بیشتر توضیح داده می شود و در شکل (۳)، مراحل آن ارائه گردیده است:

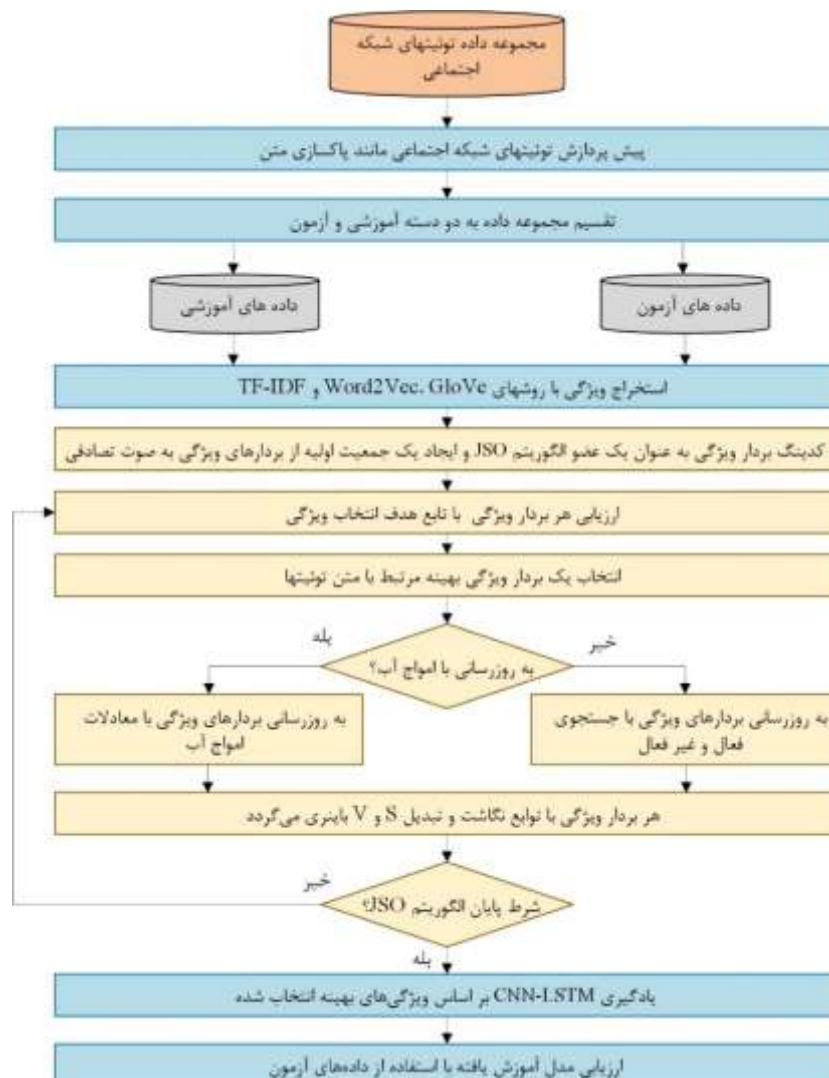
- پیش پردازش مجموعه داده شامل پاکسازی مجموعه داده
- متعادل سازی مجموعه داده با روشهای نظیر GAN یا شبکه عصبی متخاصم
- استخراج ویژگی با استفاده از تلفیق GloVe، Word2Vec و TF-IDF

- انتخاب ویژگی با الگوریتم عروس دریایی در تشخیص ویژگی های مهم زورگویی سایبری
- طبقه بندی توئیتها با استفاده از شبکه عصبی CNN و LSTM بهبود یافته با الگوریتم عروس دریایی



شکل ۳: مراحل سیستم تشخیص توئیت های زورگویی سایبری

در شکل (۴)، چارچوب روش پیشنهادی برای تشخیص زورگویی سایبری در شبکه های اجتماعی ارایه شده است.



شکل ۴: چارچوب سیستم تشخیص توئیت های زورگویی سایبری

در روش پیشنهادی چند مرحله اصلی برای تشخیص زورگویی سایبری وجود دارد که به شرح ذیل است:

- در ابتدا مجموعه داده که مجموعه ای از توئیت های شبکه اجتماعی نظیر توئیتر است گردآوری می شود و در ادامه مورد پیش پردازش مانند پاکسازی متن، توکن گذاری و غیره قرار گرفته می شود.
- در مرحله بعدی مجموعه داده به دو دسته آموزشی و آزمون تقسیم می شود و از نمونه های آموزشی برای ایجاد مدل طبقه بندی و از نمونه های آزمون برای ارزیابی مدل ایجاد شده استفاده می شود.
- در مرحله بعدی فاز استخراج ویژگی به سه روش ترکیبی GloVe، Word2Vec و TF-IDF انجام می شود به گونه ای هر روش یک مجموعه ویژگی را استخراج کرده و در نهایت این سه مجموعه ویژگی با هم ترکیب می شود.
- انتخاب ویژگی ها در مرحله بعدی با استفاده از الگوریتم عروس دریایی انجام می شود و هر بردار ویژگی یک عروس دریایی است که با جستجو بر اساس امواج آب و رفتار گروهی (جستجوی فعال و غیرفعال) به روزرسانی می شوند.
- طبقه بندی توئیتها به دو دسته عادی و زورگویی سایبری در فاز آخر و توسط طبقه بندی کننده CNN-LSTM انجام می شود.
- در نهایت مدل آموزش یافته به کمک داده های آزمون مورد ارزیابی قرار گرفته می شود و با شاخص های مانند دقت و حساسیت با روشهای مشابه مورد مقایسه قرار گرفته می شود.

۳-۱- پیش پردازش

موفقیت یک مدل یادگیری ماشین به شدت به کیفیت داده های آموزشی متکی است، که بر اهمیت پیش پردازش داده ها در فرآیند توسعه تأکید می کند. تمیز کردن موثر داده ها برای اطمینان از وضوح و جلوگیری از کاهش دقت هنگام وارد کردن داده ها به مدل ضروری است. با استفاده از کتابخانه NLTK در پایتون ابزاری که به طور گسترده برای پیش پردازش داده ها مورد استفاده قرار می گیرد، می توان داده ها را به دقت پاکسازی کرد. از طریق NLTK، عناصر نامطلوب مانند برچسب ها، هشتگ ها، لینک ها، موارد تکراری، علائم نگارشی و اعداد را می توان حذف کرد. علاوه بر این می توان همه توئیت ها به طور یکنواخت به حروف کوچک تبدیل نمود تا متن یک دست حاصل شود. تعدادی از پیش پردازش های مهم بکار رفته در روش پیشنهادی به شرح ذیل است:

- پاکسازی: با عبارات منظم تمام علائم نگارشی، پیوندها، هشتگ ها، نمادها، متن های تکراری و الفبای غیر انگلیسی را می توان حذف کرد.
- حذف کلمات توقف^۱
- یافتن بن و ریشه کلمات^۲: برای این منظور می توان از کتابخانه WordNet Lemmatizer برای انجام فرآیند Lemmatization Word جهت یافتن ریشه کلمات استفاده نمود.

۳-۲- متعادل سازی داده ها با شبکه عصبی متخاصم

در بیشتر موارد توئیت های بکار رفته در مورد زورگویی سایبری تعداد کمتری نسبت به سایر توئیت های عادی دارند و در واقع کلاس اقلیت محسوب می شوند. آموزش روش های یادگیری ماشین و یادگیری عمیق روی داده های نامتعادل باعث می شود که دقت تشخیص زورگویی سایبری توسط این روشها کاهش داده شود. یک روش برای متعادل سازی مجموعه داده ها آن است که تعدادی نمونه مصنوعی به کلاس اقلیت (کلاس زورگویی سایبری اجتماعی) اضافه شود. شبکه های متخاصم مولد^۳ به موضوع تحقیقاتی اخیر و به سرعت در حال توسعه در یادگیری ماشین تبدیل شده اند. هدف اصلی GAN ها تولید خودکار داده ها است. تفاوت اصلی آن با سایر مدل های تولیدی این است که مستقیماً از توزیع داده های واقعی استفاده نمی کند. در عوض، از طریق یک طبقه بندی کننده عمل می کند. مدل مولد^۴، تصادفی است و خود را گام به گام بر اساس پاسخ طبقه بندی کننده تنظیم می کند و به طور مداوم خروجی را اصلاح می کند تا زمانی که تمایز کننده^۵ نتواند بین داده های واقعی و مصنوعی تمایز قائل

¹ Stop Words Removal

² Lemmatization

³ Generative adversarial networks (GANs)

⁴ Generator(G)

⁵ Discriminator(D)

شود. برای دستیابی به این هدف، از دو شبکه عصبی - یعنی مولد و تمایزکننده - استفاده می شود که در یک فرآیند رقابتی شرکت می کنند. مولد داده های مصنوعی شبیه داده های واقعی تولید می کند، در حالی که تمایزکننده تلاش می کند بین داده های واقعی و داده های ارائه شده توسط مولد تمایز قائل شود. در روش پیشنهادی از اجازه دهید یک معرفی فنی از مدل GAN ارائه شود و برای این منظور فرضیات ذیل در نظر گرفته شده است [۳۵]:

- $G(z)$: خروجی ژنراتور از نویز z است و این داده از نوع، داده مصنوعی است.
- $D(x)$: خروجی تفکیک کننده است که یک نمونه واقعی x را پردازش می کند.
- $D(G(z))$: پیش بینی تمایز دهنده بر روی داده های مصنوعی است.
- P_x و P_z به ترتیب توزیع داده های واقعی و نویز هستند.
- E_x و $E_{G(z)}$ به ترتیب احتمالات گزارش مورد انتظار از خروجی های مختلف داده های واقعی و تولید شده هستند.
- θ^D و θ^G : به ترتیب وزن های مدل تفکیک کننده و مولد هستند.

عبارتی که برای شبکه کامل متشکل از متمایز کننده و مولد در نظر گرفته می شود با V نشان داده می شود و به صورت رابطه ۱، تعریف می شود:

$$V(\theta^D, \theta^G) = E_{x \sim P_x} [\log D(x)] + E_{z \sim P_z} [\log (1 - D(G(z)))] \quad (1)$$

این تابع مقدار با هدف به حداکثر رساندن تلفات تفکیک کننده و به حداقل رساندن تلفات مولد به یک استراتژی حداقل-حداکثر ارسال می شود که در رابطه ۲، ارایه می شود:

$$\min_{\theta^G} \max_{\theta^D} V(\theta^D, \theta^G) \quad (2)$$

در واقع بین مولد و متمایزگر یک بازی ماکزیمم و کمینه اجراء می شود و هدف مولد فریب دادن متمایزگر است به گونه ای که داده های با کیفیت و مصنوعی شبیه داده های واقعی تولید کند و متمایزگر فرض کند این داده ها واقعی است.

۳-۳- استخراج ویژگی

در روش پیشنهادی از سه تکنیک Word2Vec, GloVe و TF-IDF به طور همزمان استفاده می شود و ویژگی های هر کدام از این روشها در مجموعه F1, F2 و F3 قرار داده می شود و در نهایت اجتماع این مجموعه ها به عنوان ورودی فاز انتخاب ویژگی در نظر گرفته می شود. TF-IDF یک روش مهندسی ویژگی است که برای استخراج ویژگی ها از داده های متنی استفاده می - شود. این روش در زمینه تحلیل متن بسیار محبوب است. در TF-IDF، به هر عبارت در سند یک نمایش عددی اختصاص داده می شود که بر اساس وزن هر دو ویژگی فرکانس عبارت (TF) و فرکانس سند معکوس (IDF) تعیین می شود. کلمات با وزن بالاتر در سند در مقایسه با کلمات دارای وزن کمتر اهمیت بیشتری دارند. برای محاسبه TF-IDF، TF و IDF باید جداگانه بدست آیند. فرکانس مدت (TF) اغلب برای تعیین وزن یک اصطلاح استفاده می شود. رابطه ۳ نحوه محاسبه TF را نشان می دهد [۲۹].

$$TF_{t,d} = \frac{n_{t,d}}{\sum_k TF_{t,d}} \quad (3)$$

فرکانس اصطلاح یا $TF_{t,d}$ با پخش کردن تعداد کل یک عبارت t خاص در یک سند $d(n_{t,d})$ با تعداد کل عبارت های سند $\sum_k TF_{t,d}$ تعیین می شود. این رابطه فراوانی عبارت را نسبت به تعداد کلی ترم در سند نشان می دهد. رابطه بالا روند تعیین TF را روشن می کند، جایی که به هر عبارت در سند یک نمایش عددی اختصاص داده می شود. در مقابل، IDF مقدار بازنمایی را برای عباراتی که در مجموعه غیرمعمول هستند، محاسبه می کند. این بدان معناست که وقتی کلمات نادر یا غیر معمول در یک یا چند سند ظاهر می شوند، این کلمات حاوی اطلاعات مهم و معنی داری هستند. رابطه ۴ محاسبه وزن IDF را نشان می دهد [۲۹].

$$IDF(t) = \log\left(\frac{N}{df_t}\right) \quad (4)$$

فرکانس معکوس سند (IDF) برای شناسایی امتیاز تعداد کل اسناد (N) و (ft) محاسبه می شود که با تعداد اسناد حاوی عبارت نشان داده می شود. روش تعبیه کلمات Word2Vec نیز یکی از روشهای استخراج ویژگی است که در روش پیشنهادی استفاده می شود. با توجه به مطالب ارایه شده، فرکانس واژه فرکانس معکوس سند یا TF-IDF برای تعیین میزان مرتبط بودن یک اصطلاح در یک سند، با ارتباط کلمه به مقدار اطلاعات ارائه شده در مورد زمینه اصطلاح استفاده می شود. فراوانی ترم (TF) معیاری است که تعداد دفعات ظاهر شدن یک عبارت در یک سند را کمیت می کند. اگر اصطلاحی بیشتر از سایر اصطلاحات در یک متن ظاهر شود، نسبت به سایر اصطلاحات با محتوا ارتباط بیشتری دارد. علاوه بر این، امتیاز معکوس فراوانی اسناد (IDF) با تقسیم تعداد کل اسناد بر تعداد کل اسناد موجود در مجموعه حاوی آنها محاسبه می شود. این رویکرد به کاهش وزن عباراتی که اغلب در مجموعه مقالات ظاهر می شوند کمک می کند. به طور کلی، TF-IDF، که اساساً ضرب امتیازهای TF و IDF است، برای شناسایی نیازهای مربوط به یک متن استفاده می شود تا مهم ترین و آموزنده ترین کلمات به راحتی پیدا شوند [۲۹]. Word2Vec روشی برای بازآفرینی زمینه های زبانی کلمه است. این روش دارای یک شبکه عصبی با دو لایه است. مجموعه وسیعی از کلمات به عنوان ورودی استفاده می شود و نتیجه یک فضای برداری با صدها بعد است. یک فضای برداری منطبق به هر کلمه منحصر به فرد در پیکره اختصاص داده می شود. بردارهای کلمه در پیکره به گونه ای چیده شده اند که کلمات با زمینه های مشابه یا معانی تقریباً یکسان در کنار هم در فضا قرار می گیرند. Word2Vec یک روش محاسباتی سریع برای یادگیری جاسازی کلمات از متن خام است. Word2vec از دو روش جداگانه استفاده می کند [۲۰]:

- مدل Continuous Bag-of-Words (CBOW)
- مدل Skip-Gram

GloVe^۱ یک تکنیک یادگیری بدون نظارت است که می تواند در استخراج ویژگی استفاده شود. استنفورد GloVe را ایجاد کرد تا با تجمیع ماتریس همروی کلمه به کلمه، جاسازی کلمه را بسازد. نتیجه جاسازی در فضای برداری، زیرساخت های خطی جذاب کلمه را نشان می دهد. مدل GloVe روشی موثر برای استخراج ویژگی از پیکره سراسری متن است. هدف اصلی GloVe بردار سازی کلمات و خروجی بردارهای کلمه از طریق پیکره ورودی متن است. روش پیاده سازی آن به این صورت است که در ابتدا یک ماتریس همزمانی کلمه بر اساس کل پیکره ایجاد می شود و در مرحله بعد، بردار کلمه یادگیری با توجه به ماتریس همزمانی و مدل GloVe پردازش می شود. مدل GloVe را می توان با رابطه ۴، توصیف کرد [۳۰]:

$$J = \sum_{i,j}^N (X_{ij}) (V_i^T V_j + b_i + b_j - \ln(X_{ij}))^2 \quad (4)$$

۳-۴- انتخاب ویژگی

در فاز استخراج ویژگی ۴۲ ویژگی مختلف برای تشخیص زورگویی سایبری استخراج شده و تحویل مرحله انتخاب می شود. برخی از این ویژگی ها، ویژگی های زبانی و الفاظ زشت، برخی از ویژگی ها مرتبط با جنسیت، برخی از ویژگی ها مرتبط با نژاد، برخی از ویژگی ها مرتبط با لینک و تصاویر ارسال شده است. در روش پیشنهادی بعد از فاز استخراج ویژگی توسط سه روش استخراج ویژگی در نهایت ویژگی ها در اختیار الگوریتم عروس دریایی قرار گرفته می شوند. هر عروس دریایی یک بردار ویژگی است. در الگوریتم پیشنهادی هر بردار ویژگی دارای dim مولفه یا ویژگی است و یک جمعیت اولیه و تصادفی از بردارهای ویژگی به عنوان یک ماتریس مانند رابطه ۵، ایجاد می شود:

$$X = \begin{bmatrix} X_{1,1} & X_{2,1} & \dots & \dots & X_{1,dim} \\ X_{2,1} & X_{2,2} & \dots & \dots & X_{2,dim} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{N,1} & X_{N,2} & \dots & \dots & X_{N,dim} \end{bmatrix} \quad (5)$$

در اینجا، X_{ij} نشان دهنده ویژگی j از بردار ویژگی یا عروس دریایی i ام است. N تعداد جمعیت اولیه از بردارهای ویژگی برای تشخیص زورگویی در شبکه اجتماعی است. بردار ویژگی در تشخیص زورگویی در شبکه اجتماعی دارای الگوی صفر و یک

^۱ Global Vector for Word Representation

است که مطابق رابطه ۶، اگر هر ویژگی برابر صفر باشد آنگاه ویژگی مورد نظر در یادگیری استفاده نمود و اگر برابر یک باشد آن ویژگی در یادگیری استفاده می شود:

$$F_i = \begin{cases} 1 & \text{Selected feature} \\ 0 & \text{Unselected feature} \end{cases} \quad (۶)$$

در الگوریتم عروس دریایی، عروس های دریایی به عنوان راه حل در نظر گرفته می شوند و دارای دو رفتار و جستجوی اصلی به شرح ذیل می باشند:

- رفتار جستجو بر اساس حرکت در جهت امواج آب
- رفتار جستجوی گروهی که در این حالت دو جستجوی فعال و غیرفعال وجود دارد.

برای تعیین نوع رفتار عروس های دریایی می توان از متغیر $C(t)$ استفاده نمود که در رابطه ۷ فرموله شده است [۱۷].

$$c(t) = \left| \left(1 - \frac{t}{Maxt} \right) \times (2.rand - 1) \right| \quad (۷)$$

در این رابطه، t شمارنده تکرار الگوریتم عروس دریایی است و T بیشترین شمارنده تکرار الگوریتم است. $rand$ یک عدد تصادفی بین صفر و یک است. برای به روزرسانی هر عروس دریایی یا بردار ویژگی در ابتدا $C(t)$ محاسبه می شود و اگر بیشتر از ۰/۵ باشد آنگاه نوع جستجو طبق امواج آب و بر اساس رابطه ۸، انجام می شود:

$$X_i(t+1) = X_i(t) + rand(0,1) \times (X^* - \beta \times rand(0,1) \times \mu) \quad (۸)$$

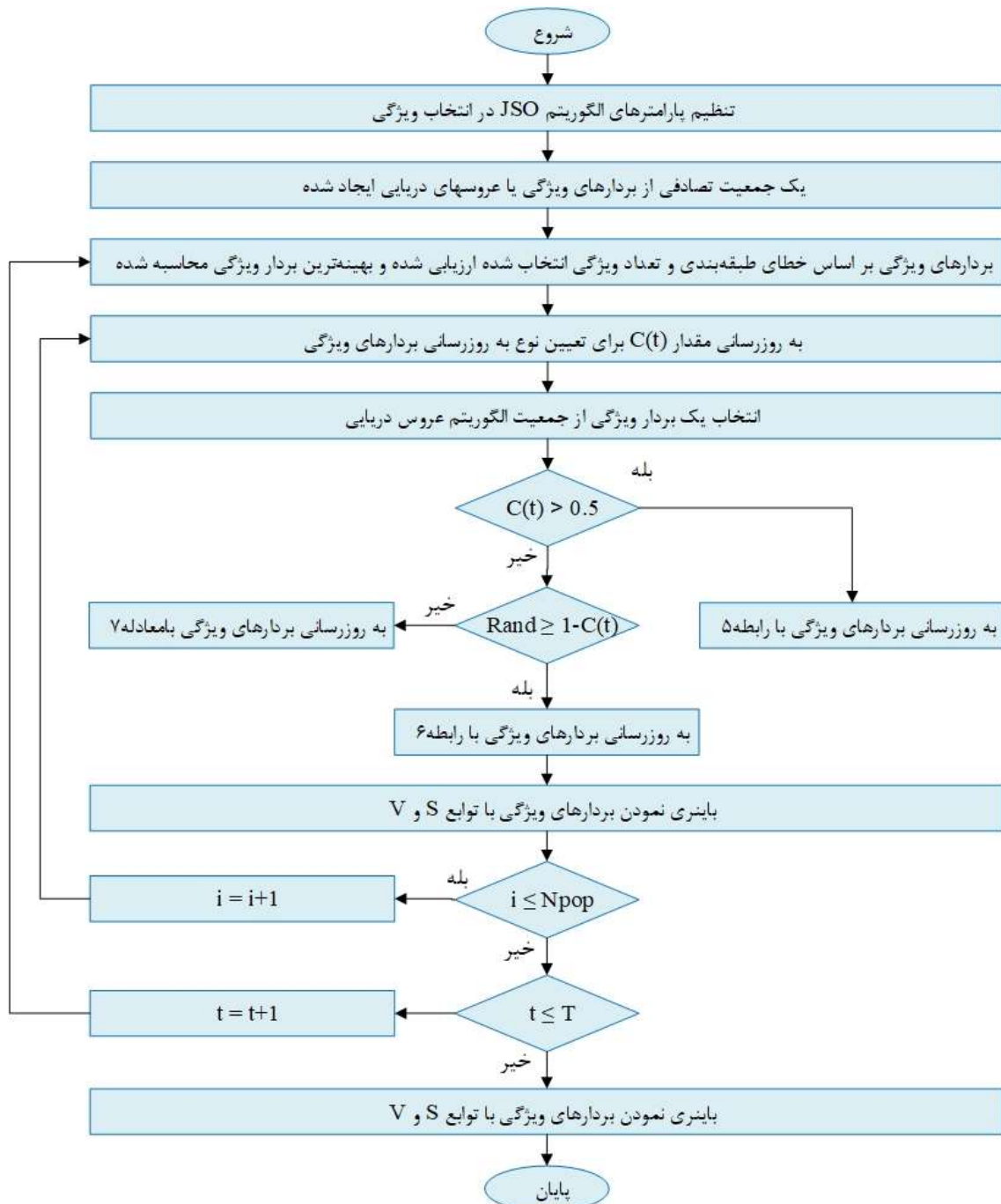
در این رابطه، $X_i(t)$ به بردار ویژگی شماره i و در تکرار t اشاره دارد و μ میانگین بردارهای ویژگی در یک بعد خاص آن است. مقدار β برای تنظیم حرکت در راستای امواج آب در نظر گرفته می شود. بهینه ترین راه حل یا بردار ویژگی بهینه با X^* نمایش داده می شود. اما در مقابل اگر $C(t)$ از ۰/۵ بیشتر نباشد آنگاه رفتارهای هوش گروهی باعث به روزرسانی بردارهای ویژگی می شود. برای این منظور یک عدد تصادفی بین صفر و یک برای هر بردار ویژگی تولید می شود که اگر عدد تصادفی مورد نظر از $(1-c(t))$ بیشتر باشد آنگاه جستجوی غیرفعال برای به روزرسانی بردارهای ویژگی مطابق رابطه ۹ انجام می شود و در غیر این صورت جستجوی فعال مطابق رابطه ۱۰، بکار گرفته می شود.

$$X_i(t+1) = X_i(t) + \gamma.rand(0,1) \times (U_b - L_b) \quad (۹)$$

$$X_i(t+1) = X_i(t) + rand.(X_j(t) - X_i(t)) \quad (۱۰)$$

در این رابطه، U_b و L_b محدوده بالا و پایین بردارهای ویژگی است و γ ضریب همگرایی در جستجوی غیرفعال عروس های دریایی است.

در شکل (۵)، فلوچارت روش پیشنهادی در انتخاب ویژگی توسط الگوریتم عروس دریایی نشان داده شده است.



شکل ۵: فلوچارت انتخاب ویژگی توسط الگوریتم عروس دریایی

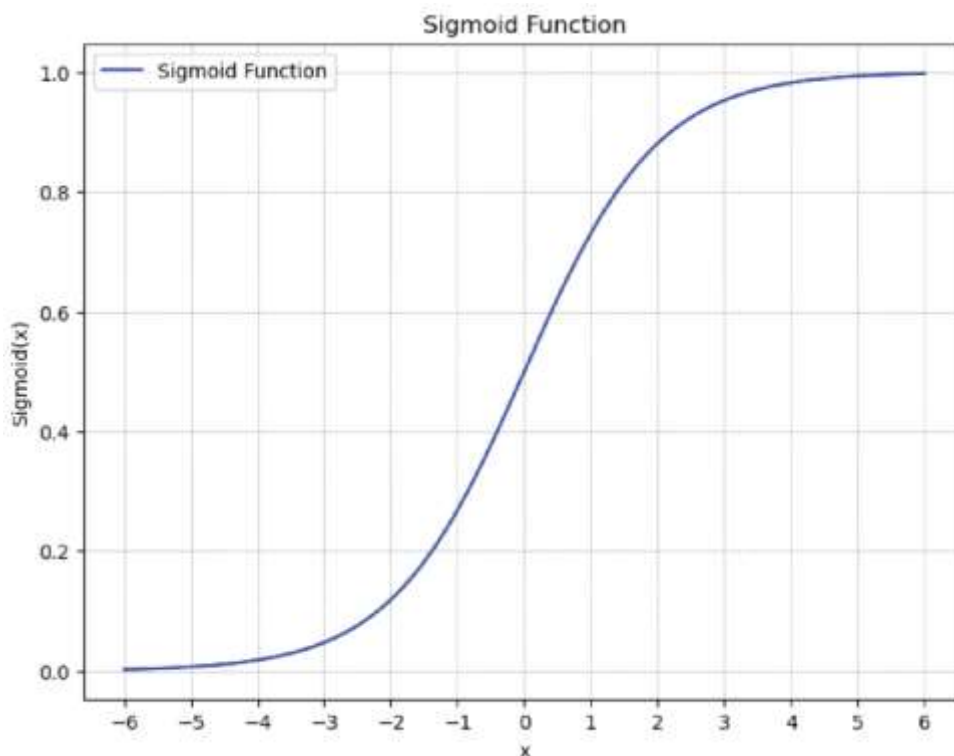
تابع هدف در فاز انتخاب برای ارزیابی بردارهای ویژگی در رابطه ۱۱، فرموله شده است:

(۱۱)

$$Cost(X_i) = w_1 \cdot E + w_2 \frac{size(X_i)}{D}$$

در این رابطه، E خطای تشخیص طبقه بندی توئیتها به دو دسته زورگویی سایبری و عادی به ازای بردار ویژگی i -ام یا X_i است. در اینجا از یک شبکه عصبی چند لایه برای ارزیابی بردارهای ویژگی استفاده می شود. $size(X_i)$ تعداد ویژگی انتخاب شده در بردار ویژگی X_i در زورگویی سایبری است. ابعاد بردارهای ویژگی برابر D نمایش است. w_1 و w_2 دو وزن خطا و کاهش ابعاد است که بین صفر و یک انتخاب می شوند و مجموع آنها برابر یک است. کمینه نمودن این تابع هدف نشان دهنده شایستگی بیشتر یک بردار ویژگی است.

مقادیر بردارهای ویژگی باید صفر و یک باشند اما در حین فرآیند به روزرسانی بردارهای ویژگی توسط الگوریتم عروس دریایی این مقادیر حالت صفر و یک خود را از دست داده و می توانند اعشاری شوند. بعد از هر به روزرسانی بردارهای ویژگی می توان آنها را مجدد باینری نمود. برای این منظور ابتدا توسط تابع S یا V ، مقادیر بردارهای ویژگی بین صفر و یک نرمال می شود. در شکل (۶)، تابع سیگموئید یا S برای نرمالسازی مقادیر ویژگی های مجموعه داده نمایش داده می شود. مشاهده می شود که برد تابع S بین صفر و یک است که برای نرمالسازی مقادیر بردارهای ویژگی بین صفر و یک در نظر گرفته شده است.



شکل ۶: تابع S یا سیگموئید برای نرمالسازی مقادیر ویژگی ها بین صفر و یک

در روابط ۱۲، ۱۳، ۱۴، ۱۵، ۱۶، ۱۷، ۱۸ و ۱۹ به ترتیب چهار نوع تابع سیگموئید یا S و چهار تابع تانژانت سیگموئید یا V نمایش داده شده است [۳۴].

$$S_1 = \frac{1}{1 + e^{-2x}} \quad (12)$$

$$S_2 = \frac{1}{1 + e^{-x}} \quad (13)$$

$$S_3 = \frac{1}{1 + e^{-x/2}} \quad (14)$$

$$S_4 = \frac{1}{1 + e^{-x/3}} \quad (15)$$

$$V1 = |\tanh(x)| \quad (16)$$

$$V2 = \left| \operatorname{erf} \left(\frac{\sqrt{\pi}}{2} x \right) \right| = \left| \frac{\sqrt{\pi}}{2} \int_0^x e^{-t^2} dt \right| \quad (17)$$

$$V3 = \left| \frac{x}{\sqrt{1+x^2}} \right| \quad (18)$$

$$V3 = \left| \frac{2}{\pi} \arctan \frac{\pi}{2} x \right| \quad (19)$$

بعد از مرحله نرمالسازی بردارهای ویژگی توسط توابع S و V می توان توسط رابطه های ۲۰ و ۲۱ مقدار آنها به صفر و یک تبدیل نمود [۳۴].

$$x_i^j(t+1) = \begin{cases} 0, & \text{if } random < Sfunction(x_i^j(t+1)) \\ 1, & \text{if } random \geq Sfunction(x_i^j(t+1)) \end{cases} \quad (20)$$

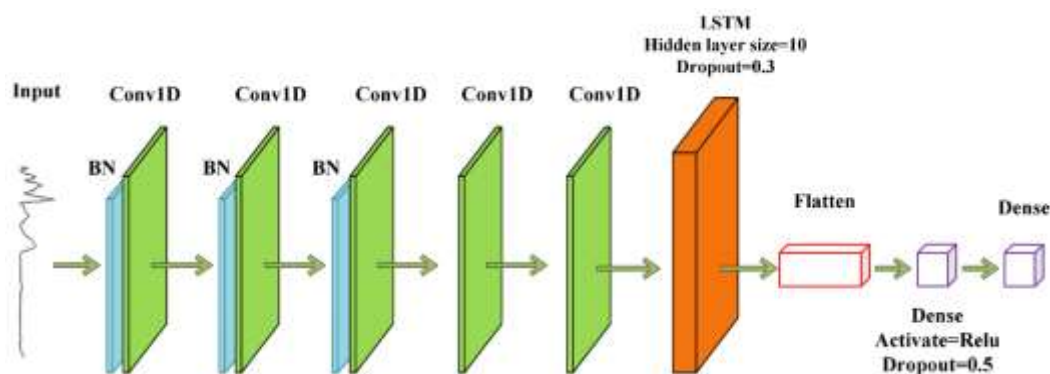
$$x_i^j(t+1) = \begin{cases} x_i^j(t), & \text{if } random < Vf inction(x_i^j(t+1)) \\ x_i^j(t), & \text{if } random \geq Vf inction(x_i^j(t+1)) \end{cases} \quad (21)$$

در این پژوهش از مجموعه داده بکار رفته در [۳۲] و [۳۳] مورد استفاده قرار گرفته شده است و ورودی های فاز انتخاب ویژگی از این دو مجموعه داده تغذیه می شود. ویژگی های مهمی که روش پیشنهادی آنها را انتخاب می کند عبارتند از توهین جنسی، توهین نژادی، باج خواهی، الفاظ زشت، انتساب اسم حیوان به اشخاص، تهدید با سلاح سرد و گرم، ارسال فایل های تحقیر آمیز، ارسال اطلاعات شخصی برای دیگران و غیره است.

یکی از چالش های یادگیری ماشین بیش برازش^۱ است که می تواند ناشی از پیچیدگی زیاد مدل و آموزش بیش از حد آن باشد. عواملی که باعث بیش برازش می شود می تواند عدم انتخاب ویژگی، حجم نمونه های اندک و عدم پیش پردازش مجموعه داده باشد. یکی از راهکارهای ما برای مقابله با این چالش بکارگیری انتخاب ویژگی با الگوریتم عروس دریایی و از طرفی متعادل سازی مجموعه داده است و همچنین روش ما برای رفع این چالش پیش پردازش و نرمال سازی داده ها را انجام می دهد.

۳-۵- طبقه بندی با معماری LSTM و CNN

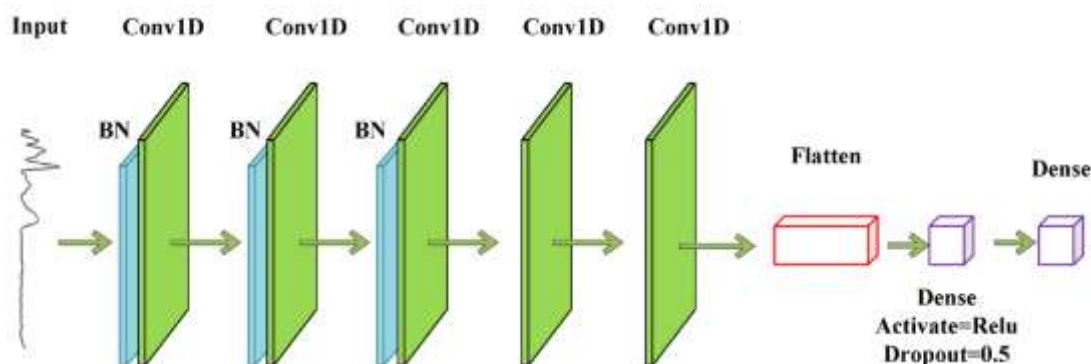
ترکیب CNN و شبکه های عصبی مکرر کاربردهای تحقیقاتی متعددی داشته است و اخیراً نتایج قابل توجهی در زمینه های زیر به همراه داشته است که از جمله آنها می توان به زمینه پردازش زبان طبیعی، مانند تجزیه و تحلیل احساسات و تشخیص گفتار، در پیش بینی داده های بلادرنگ، مانند پیش بینی سهام و پیش بینی آلودگی هوا و غیره اشاره نمود [۳۱]. در مدل ترکیبی CNN پارامترها را از طریق به اشتراک گذاری وزن کاهش می دهد تا کارایی یادگیری مدل را بهبود بخشد. علاوه بر این، LSTM یک مدل شبکه تکراری است که مشکلات گرایان طولانی مدت و ناپدید شدن گرایان را در RNN حل می کند. چارچوب شبکه یادگیری عمیق CNN-LSTM در شکل (۷)، نشان داده شده است. در این مطالعه، دو یادگیرنده بر اساس 1D-CNN و LSTM ساخته شدند. این دو یادگیرنده برای استخراج انتزاع محلی و اطلاعات موقعیت توالی از طیف ها ساخته شدند. به طور خاص، ویژگی های داده درون طیفی که توسط CNN آموخته شد به LSTM وارد شد و LSTM برای استخراج اطلاعات موقعیت در داده های طیفی استفاده شد. مدل CNN-LSTM می تواند به طور کامل ویژگی های ذاتی داده ها را استخراج کند، و همچنین مدل را قادر می سازد تا توانایی خاصی برای استخراج اطلاعات داده های توالی متوسط و طولانی داشته باشد. این مدل می تواند ویژگی های پنهان نمونه ها را بیاموزد و این اطلاعات را برای شناسایی بهتر نمونه های دسته های مختلف به دست آورد.



¹ Overfitting

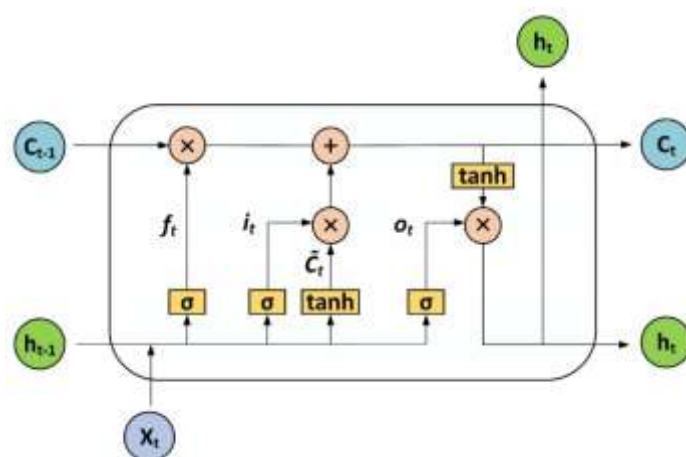
شکل ۷: معماری ترکیبی CNN و LSTM

1D-CNN ساخته شده در این مقاله از چارچوب Alexnet استفاده می‌کند، و ساختار مدل 1D-CNN در شکل (۸)، نشان داده شده است. Alexnet یک مدل یادگیری عمیق کلاسیک است که شامل یک تبدیل ۸ لایه با ۵ لایه کانولوشنال و ۲ لایه کاملاً متصل است. و یک لایه خروجی کاملاً متصل Alexnet تابع فعال‌سازی واحد خطی اصلاح‌شده (ReLU) را بعد از هر لایه کانولوشن اضافه می‌کند، که باعث می‌شود Alexnet از تکنیک حذف تصادفی برای نادیده گرفتن انتخابی نورون‌ها در لایه کاملاً متصل در طول آموزش استفاده کند تا از برازش بیش از حد مدل جلوگیری کند.



شکل ۸: معماری شبکه عصبی CNN یک بعدی

واحدهای LSTM در شکل (۹)، نشان داده شده‌اند و کار هر واحد LSTM در ادامه توضیح داده می‌شود.



شکل ۹: یک واحد از شبکه

برای هر ورودی x_t از داده‌های طیفی، مدل LSTM یک فعال‌سازی پنهان h_t ایجاد می‌کند. در مرحله بعد، هر قطعه از داده های طیفی وارد شده و بیشتر برای پیش بینی استفاده می‌شود. مدل LSTM رابطه تبدیل h_t نمایش پنهان را از طریق واحد LSTM تعریف می‌کند، که ورودی x_t را در حالت فعلی و اطلاعات به دست آمده h_{t-1} را می‌پذیرد.

۴- نتایج تجربی

در این بخش روش پیشنهادی برای تشخیص توثیهای زورگویی سایبری در شبکه اجتماعی مورد پیاده‌سازی و تحلیل قرار گرفته شده است. در ابتدا پارامترهای پیاده‌سازی و سپس مجموعه داده شرح داده می‌شود. در ادامه نیز متریکهای ارزیابی بیان شده و روش پیشنهادی در نهایت با روشهای دیگر مقایسه می‌شود.

۴-۱- پارامترهای پیاده سازی

برای پیاده سازی یادگیری عمیق از پایتون و کتابخانه Keras و Tensorflow استفاده می شود. در مدل 1D-CNN، اندازه دسته ۳۲، بهینه ساز از نوع Adam انتخاب می شود، نرخ یادگیری روی ۰/۰۰۱ تنظیم شده است. اندازه جمعیت بردارهای ویژگی برابر ۱۰ و تعداد تکرار الگوریتم انتخاب ویژگی نیز برابر ۳۰ تنظیم می شود. نوع اعتبارسنجی روش پیشنهادی از نوع متقاطع است و ۷۰ درصد از داده ها آموزشی و مابقی نیز به نسبت ۱۵ درصد و ۱۵ درصد به ترتیب داده آزمون و اعتبارسنجی است. هر آزمایش نیز ۲۵ مرتبه تکرار می شود و متوسط شاخص ها در ارزیابی استفاده می شود. در روش پیشنهادی برای نرمال سازی از روش خطی ماکزیمم و کمینه استفاده می شود و محدوده نرمال سازی بین [0,1] است. در پیاده سازی ها دو مجموعه داده [۳۲ و ۳۳] به عنوان ورودی روش پیشنهادی در نظر گرفته شده و پیش پردازش مجموعه داده ها با کتابخانه های پایتون مانند spacy انجام می شود. در ادامه مجموعه داده ها فازهای مانند متعادل سازی با شبکه GAN و استخراج ویژگی انجام می شود و سپس انتخاب ویژگی با الگوریتم عروس دریایی انجام می شود. طبقه بندی نهایی مدل روی ویژگی های خروجی الگوریتم عروس دریایی توسط CNN و LSTM انجام می شود. هر آزمایش در روش پیشنهادی ۲۵ مرتبه تکرار شده و متوسط شاخص های ارزیابی مانند دقت محاسبه و با روش های مشابه مقایسه می شود.

جدول ۲: پارامترهای مدل یادگیری عمیق

Operator	Filters	Kernel Size	Strides
Conv1D	64	11	4
	64	5	1
	92	3	1
	92	3	1
	64	3	1
LSTM Hidden layer size	50		
Dropout	0.3		
Dense Activate	Relu		
batch size	32		

۴-۲- مجموعه داده

این مطالعه از دو مجموعه داده برای آزمایش و اعتبارسنجی عملکرد مدل های پیشنهادی استفاده می کند. اولین مجموعه داده معیار توییت (که در حال حاضر به عنوان X شناخته می شود) در [۳۲] برای شناسایی آزار و اذیت سایبری با شناسایی توییت های توهین آمیز و غیر توهین آمیز استفاده شده است و این مجموعه داده با اندازه ۳۷۳۷۳ توییت است. داده ها به صورت عددی با ۱ یا ۰ برچسب گذاری شدند که در آن ۱ نشان دهنده توییت توهین آمیز و ۰ نشان دهنده توییت خنثی است، به این معنی که توییت به دسته توهین آمیز تعلق ندارد. جدول (۳)، نمونه ای از نمونه های توییت های آزار و اذیت سایبری در مجموعه داده های توییت را نشان می دهد.

جدول ۳: چند نمونه از توییت های عادی و زورگویی سایبری

d	Cyberbullying Tweet Samples	Pred	Label
1	Fat people are dump	Offensive (cyberbullying)	1
2	WTF are you talking about Men? No men thats not a menage that's just gay.	Offensive (cyberbullying)	1
3	Fake friends are no different than shadows, they stick around during your brightest moments, but disappear during your darkest.	Non-offensive (non-bullying)	0
4	You are big black s**t.	Offensive (cyberbullying)	1
5	Today something is dope. Tomorrow that same thing is trash. Next month it is irrelevant. Next year it's classic.	Non-offensive (non-bullying)	0

مجموعه داده دوم مورد استفاده در این مطالعه [۳۳] شامل داده های جمع آوری شده از گروه های توییت و فیس بوک است که بر فعالیت های مشکوک مانند نژادپرستی، تبعیض، زبان توهین آمیز و تهدید، که عمدتاً با حوادث آزار و اذیت سایبری مرتبط است، تمرکز دارد. داده های موجود در مجموعه داده بر اساس وجود کلمات مشکوک مورد استفاده در توییت ها و نظرات حاشیه نویسی شدند. نقاط داده مشکوک به صورت دستی با برچسب ۱ برچسب گذاری شدند، در حالی که نقاط داده غیر مشکوک با ۰ برچسب گذاری شدند. در مجموع، مجموعه داده شامل تقریباً ۲۰ هزار ردیف احساسات است. در این میان، حدود ۱۲ هزار نقطه داده با احساسات منفی برچسب گذاری شدند که نشان دهنده وجود ویژگی هایی مانند نژادپرستی،

تبعیض و سوء استفاده است. برعکس، هشت هزار نداده با احساسات مثبت یا خنثی برچسب گذاری شدند، که نشان می دهد داده ها ویژگی های غیر مشکوک را نشان می دهند. اطلاعات ورودی هر دو مجموعه داده اساساً بر اساس توییت ها و نظرات به زبان انگلیسی است، با برخی از روش های پیش پردازش و پاکسازی داده ها در زیر بخش زیر توضیح داده شده است.

۳-۴- متریک های ارزیابی

برای ارزیابی روش پیشنهادی در تشخیص زورگویی سایبری از متریک های نظیر دقت^۱، حساسیت^۲ و صحت^۳ استفاده می شود که ضابطه آنها به ترتیب در رابطه های ۹، ۱۰، ۱۱ ارایه شده است.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (9)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \quad (10)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (11)$$

هر کدام از شاخص های ارزیابی TP، TN، FP و FN را برای تحلیل زورگویی سایبری می توان به صورت ذیل تعریف نمود:

- نمونه های صحیح مثبت^۴ (TP): توثیت مورد نظر از نوع زورگویی سایبری است و روش پیشنهادی آن را در دسته زورگویی سایبری قرار داده است.
- نمونه های غلط مثبت^۵ (FP): توثیت مورد نظر از نوع عادی است و روش پیشنهادی آن را در دسته زورگویی سایبری قرار داده است.
- نمونه های صحیح منفی^۶ (TN): توثیت مورد نظر از نوع عادی است و روش پیشنهادی آن را در دسته عادی قرار داده است.
- نمونه های غلط منفی^۷ (FN): توثیت مورد نظر از نوع زورگویی سایبری است و روش پیشنهادی آن را در دسته عادی قرار داده است.

۳-۴- تحلیل و ارزیابی

در این بخش روش پیشنهادی برای تشخیص زورگویی سایبری مورد پیاده سازی قرار گرفته شده است. در نمودار شکل (۱۰) و (۱۱) شاخص دقت، صحت و حساسیت روش پیشنهادی با حالت های مختلف LSTM، CNN، CNN-LSTM مورد مقایسه قرار گرفته شده است. روش پیشنهادی در اینجا با JSO-CNN-LSTM تعریف می شود که به اختصار JCL در نظر گرفته می شود. هدف از این مقایسه ها آن است که نشان داده شود که روش پیشنهادی نسبت به هر کدام از اجزای سازنده آن چقدر دقت دارد.

¹ Accuracy

² Sensitivity

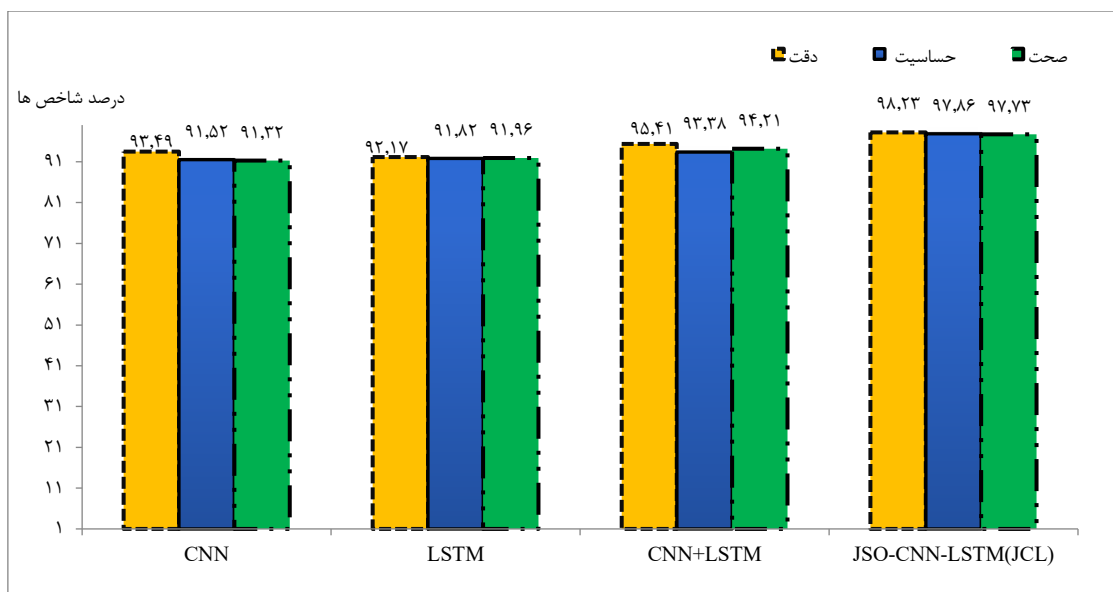
³ Precision

⁴ True positive (TP)

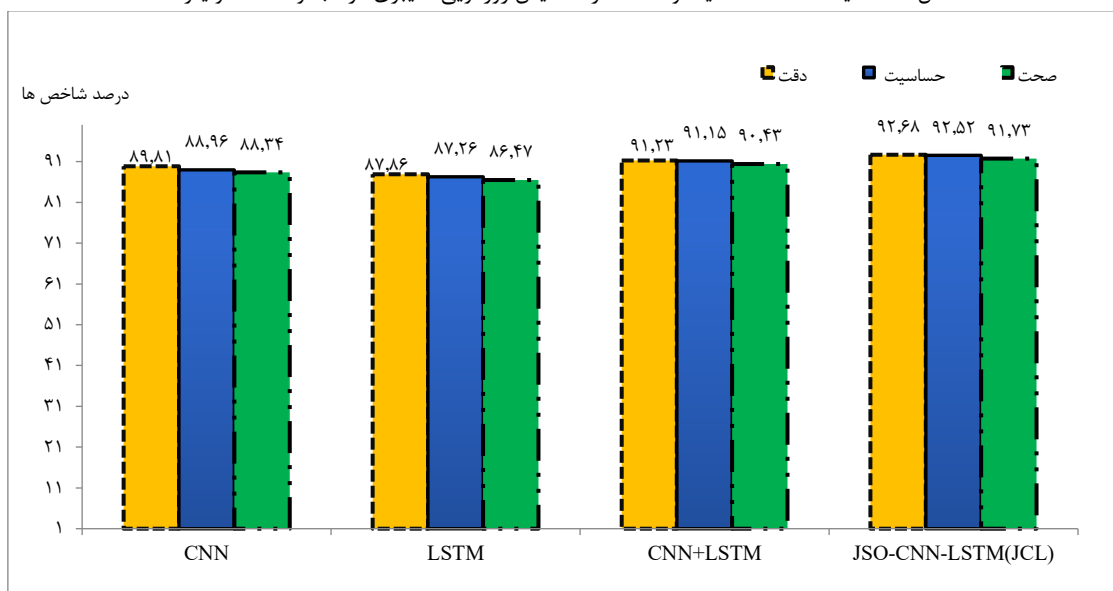
⁵ False positive (FP)

⁶ True negative (TN)

⁷ False negative (FN)



شکل ۱۰: مقایسه دقت، حساسیت و صحت در تشخیص زورگویی سایبری در مجموعه داده توئیت

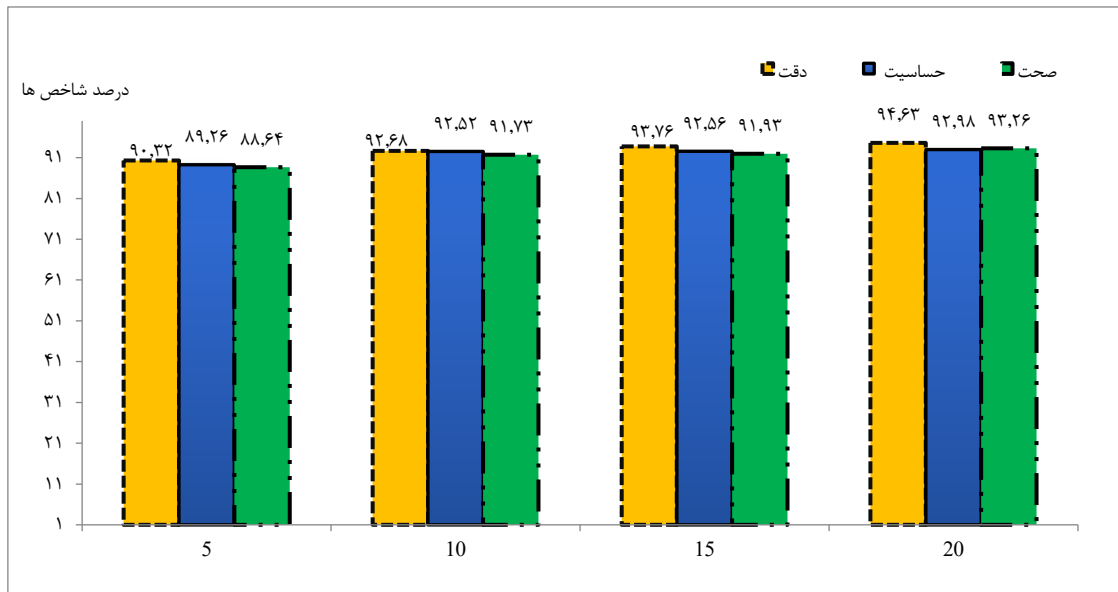


شکل ۱۱: مقایسه دقت، حساسیت و صحت در تشخیص زورگویی سایبری در مجموعه داده فیس بوک

آزمایشات نشان می‌دهد که اگر از مجموعه داده توئیت برای ارزیابی استفاده شود آنگاه دقت، حساسیت و صحت روش پیشنهادی در تشخیص زورگویی سایبری به ترتیب برابر ۹۸/۲۳ درصد، ۹۷/۸۶ درصد و ۹۷/۷۳ درصد است. روش پیشنهادی به دلیل انتخاب ویژگی با الگوریتم JSO دارای دقت بیشتری نسبت به معماری CNN-LSTM است و از طرفی به دلیل ترکیب و دو معماری CNN و LSTM دارای دقت بیشتری نسبت به هر کدام از این دو روش یادگیری عمیق است. در مجموعه داده فیس بوک دقت، حساسیت و صحت روش پیشنهادی برای تشخیص زورگویی سایبری به ترتیب برابر ۹۲/۶۸ درصد، ۹۲/۵۲ درصد و ۹۱/۷۳ درصد است و روش پیشنهادی دارای دقت بیشتری از روشهای نظیر CNN، LSTM و CNN-LSTM است. آزمایشات نشان می‌دهد روش پیشنهادی در مجموعه داده توئیت دارای دقت بیشتری نسبت به مجموعه داده فیس بوک در تشخیص زورگویی سایبری است.

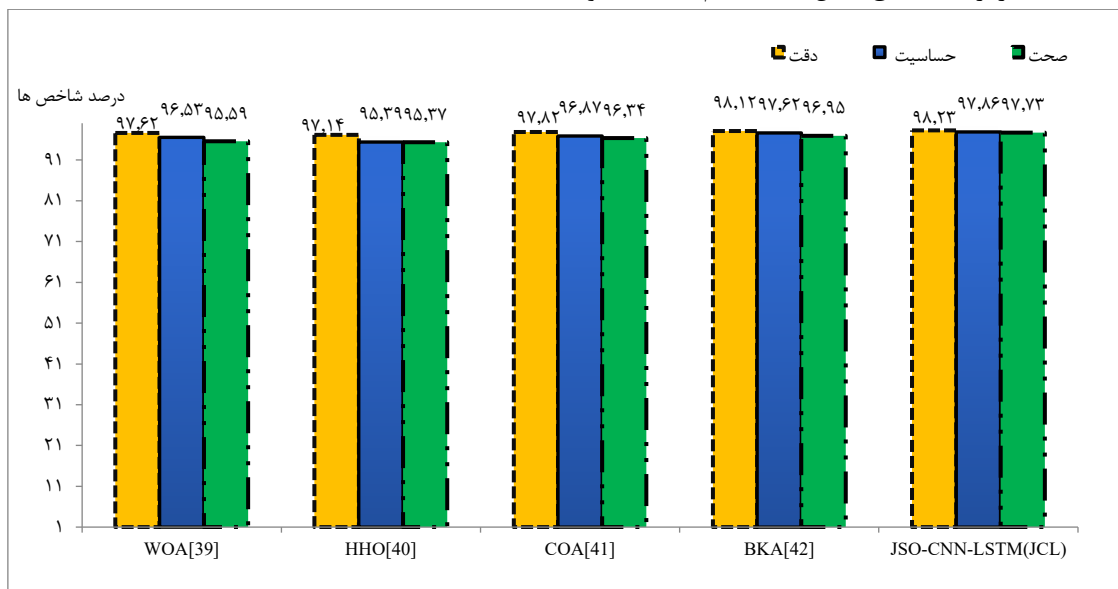
یکی از شاخص‌ها و پارامترهای مهم برای ارزیابی روش پیشنهادی اندازه جمعیت اولیه بردارهای ویژگی است که افزایش آن را می‌توان باعث افزایش توانایی جستجوی روش پیشنهادی در یافتن بردار ویژگی در نظر گرفت. در نمودار شکل (۱۲)، تعداد

بردارهای ویژگی در الگوریتم عروس دریایی به عنوان مهمترین پارامتر به عنوان متغیر در نظر گرفته شده است و مقدار آن ۵، ۱۰، ۱۵ و ۲۰ تنظیم شده است.



شکل ۱۲: مقایسه دقت، حساسیت و صحت در تشخیص زورگویی سایبری با افزایش تعداد بردارهای ویژگی

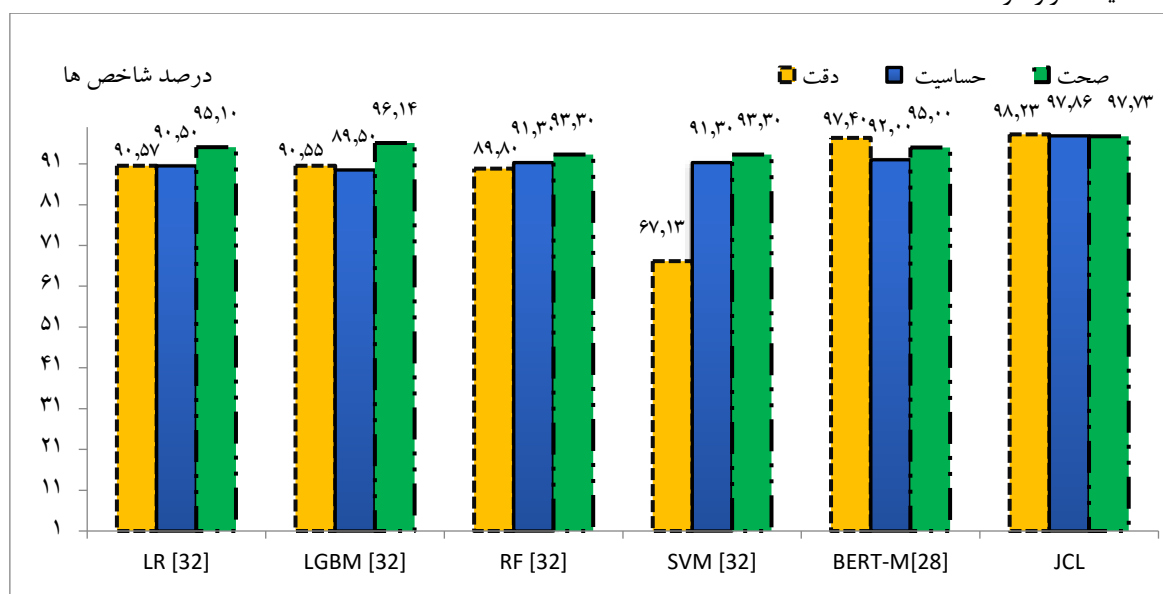
ارزیابی‌ها نشان می‌دهد با افزایش تعداد بردارهای ویژگی توانایی الگوریتم عروس دریایی برای یافتن بردار ویژگی افزایش پیدا می‌کند و دقت، حساسیت و صحت افزایش می‌یابد. اگر تعداد بردارهای ویژگی برابر ۵ باشد دقت، حساسیت و صحت روش پیشنهادی برابر ۹۰/۳۲ درصد، ۸۹/۲۶ درصد و ۸۸/۶۴ درصد است و با افزایش تعداد بردار ویژگی به ۲۰ عدد مقدار دقت، حساسیت و صحت در حدود ۴/۳۱ درصد، ۳/۷۲ درصد و ۴/۶۲ درصد افزایش خواهد یافت که تاثیر جمعیت الگوریتم فراابتکاری در فاز انتخاب ویژگی را نشان می‌دهد اما با افزایش تعداد بردارهای ویژگی از ۵ به ۲۰ زمان اجرای فاز انتخاب ویژگی از ۳/۶۸ به ۷/۹۲ افزایش خواهد یافت. می‌توان دقت، حساسیت و صحت روش پیشنهادی را با الگوریتم‌های فراابتکاری به ازای جمعیت برابر ۱۰ مطابق شکل (۱۳)، با هم مقایسه نمود.



شکل ۱۳: مقایسه دقت، حساسیت و صحت در تشخیص زورگویی سایبری در مجموعه داده فیس بوک با روشهای انتخاب ویژگی

در این آزمایش الگوریتم عروس دریایی در تشخیص زورگویی سایبری با روشهای فراابتکاری در فاز انتخاب ویژگی از جمله الگوریتم بهینه‌سازی وال، الگوریتم بهینه‌سازی شاهین، الگوریتم بهینه‌سازی کوانتی و الگوریتم بهینه‌سازی عقاب بال سیاه مقایسه و ارزیابی شده است. آزمایشات نشان داد روش الگوریتم عروس دریایی بیشترین دقت، حساسیت و صحت را در

تشخیص زورگویی سایبری دارد و بدترین عملکرد نیز مرتبط با الگوریتم شاهین هریس است. زمان اجرای روش پیشنهادی به ازای تعداد بردار ویژگی برابر ۱۰ در حدود ۴.۸۷ ثانیه است و زمان اجرای الگوریتم بهینه سازی وال، الگوریتم بهینه سازی شاهین، الگوریتم بهینه سازی کوآتی و الگوریتم بهینه سازی عقاب بال سیاه به ترتیب برابر ۴.۳۵، ۵.۸۲، ۴.۹۳ و ۴.۵۱ است. به عبارت بهتر زمان اجرای الگوریتم پیشنهادی فقط نسبت به الگوریتم وال در فاز انتخاب ویژگی بیشتر است و در حالت کلی زمان اجرای الگوریتم پیشنهادی از الگوریتم بهینه سازی شاهین، الگوریتم بهینه سازی کوآتی و الگوریتم بهینه سازی عقاب بال سیاه کمتر است. در نمودار شکل (۱۴)، روش پیشنهادی در شاخص های سه گانه با چند روش یادگیری در تشخیص زورگویی مورد مقایسه قرار گرفته شده است.



شکل ۱۴: مقایسه دقت، حساسیت و صحت در تشخیص زورگویی سایبری در مجموعه داده توئیت با مطالعات مرتبط

آزمایشات و ارزیابی نشان می دهد که روش پیشنهادی نسبت به رگرسیون خطی، گرادیان LGBM، جنگل تصادفی یا RF، ماشین بردار پشتیبان یا SVM و شبکه پردازش زبان BRET دارای دقت بیشتری است. آزمایشات و مقایسه ها نشان می دهد که روش پردازش زبان BERT رقیب اصلی روش پیشنهادی در تشخیص زورگویی اجتماعی است. در جدول (۴)، روش پیشنهادی با چند روش یادگیری عمیق در تشخیص زورگویی در شبکه اجتماعی توئیت نیز مورد مقایسه قرار گرفته شده است.

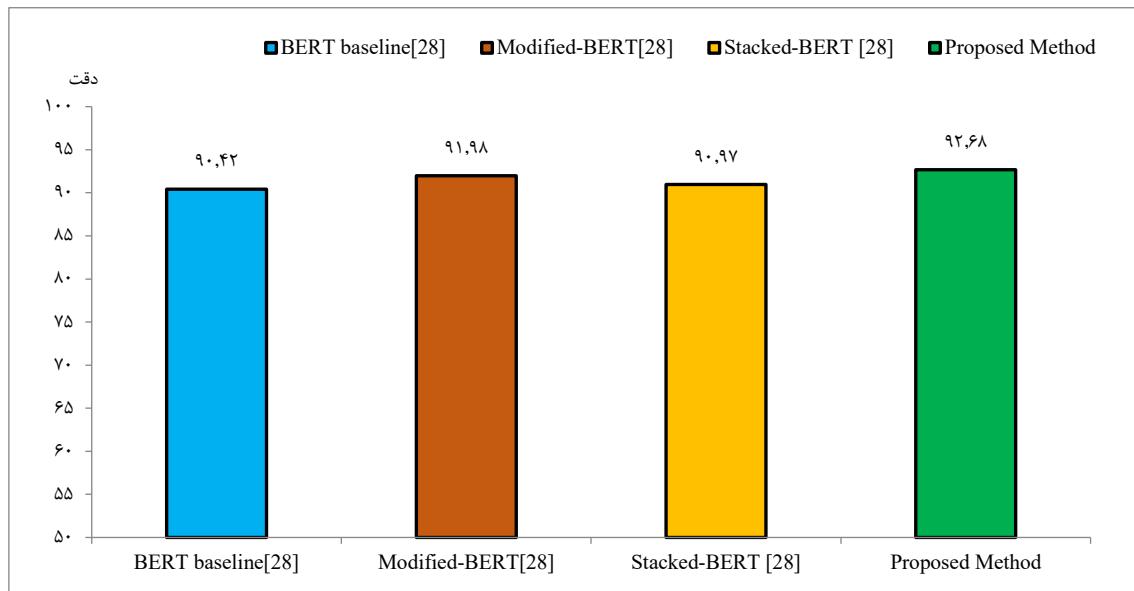
جدول ۴: مقایسه روش پیشنهادی در تشخیص توئیت های عادی و زورگویی سایبری در مجموعه داده توئیت

روش	دقت	حساسیت	صحت
LSTM[28]	0.8011	0.7281	0.8142
Conv1DLSTM[28]	0.8649	0.8919	0.8146
CNN[28]	0.8496	0.7908	0.8836
BiLSTM[28]	0.7795	0.8130	0.8373
BERT[28]	0.921	0.915	0.915
Tuned-BERT[28]	0.9384	0.91	0.92
Stacked[28]	0.974	0.92	0.950
CNN-LSTM[43]	0.9752	0.9896	0.9828
JCL(Proposed Method)	98.23	97.86	97.73

شاخص دقت، حساسیت و صحت روش پیشنهادی در تشخیص زورگویی اجتماعی از روش های یادگیری عمیق در مجموعه داده توئیت از LSTM، Conv1DLSTM، CNN، BiLSTM، BERT، Tuned-BERT و Stacked بیشتر است. ارزیابی ها نشان می دهد که دقت روش پیشنهادی از مدل ارایه شده در [۴۳] بیشتر است اما حساسیت و صحت روش پیشنهادی نسبت به روش ارایه شده در این مطالعه کمتر است.

دلیل اصلی آنکه روش پیشنهادی نسبت به بیشتر روش های یادگیری عمیق دقت بیشتری در تشخیص زورگویی اجتماعی دارد در ذیل خلاصه شده است:

- استخراج ویژگی بر خلاف این روش ها توسط سه روش انجام شده و نتایج با هم تلفیق شده
 - انتخاب ویژگی با الگوریتم هوش گروهبی در روش پیشنهادی انجام شده است.
 - روش پیشنهادی از ترکیب دو معماری IDCNN و LSTM برای طبقه بندی دقیق توئیتها به دو دسته عادی و زورگویی سایبری استفاده می کند.
- در نمودار شکل (۱۵)، نیز دقت روش پیشنهادی در تشخیص زورگویی سایبری در مجموعه داده فیس بوک با روشهای پیشرفته مانند BERT baseline، Modified-BERT و Stacked-BERT مقایسه شده است.



شکل ۱۵: مقایسه دقت، روش پیشنهادی در تشخیص زورگویی سایبری در مجموعه داده فیس بوک

دقت روش BERT baseline، Modified-BERT و Stacked-BERT به ترتیب برابر ۹۰/۴۲ درصد، ۹۱/۹۸ درصد و ۹۰/۹۷ درصد در تشخیص زورگویی است حال آنکه دقت روش پیشنهادی در مجموعه داده فیس بوک بیشتر از این روشها و برابر ۹۲/۸ درصد است.

۵- نتیجه گیری

شیوع آزار و اذیت سایبری در پلتفرم های رسانه های اجتماعی به یک نگرانی قابل توجه برای افراد، سازمان ها و جامعه در کل تبدیل شده است. تشخیص زودهنگام و مداخله زورگویی سایبری در رسانه های اجتماعی برای کاهش اثرات مضر آن بسیار مهم است. در سال های اخیر، یادگیری عمیق نتایج امیدوارکننده ای برای تشخیص آزار سایبری در رسانه های اجتماعی نشان داده است. روش پیشنهادی برای تشخیص توئیت های زورگویی سایبری از یک روش ترکیبی سه مرحله ای استفاده می کند. در مرحله اول روش پیشنهادی با سه روش ترکیبی GloVe، Word2Vec و TF-IDF ویژگی های متن را استخراج نموده و در مرحله دوم با استفاده از الگوریتم عروس دریایی ویژگی های مهم را انتخاب و تحویل طبقه بندی کننده یادگیری عمیق می نماید. در مرحله سوم از طبقه بندی کننده IDCNN+LSTM در تشخیص توئیت های زورگویی سایبری استفاده می شود. آزمایشات نشان داد روش پیشنهادی در تشخیص زورگویی سایبری از روشهای مانند IDCNN، LSTM، CNN، BERT baseline، Modified-BERT و Stacked-BERT دارای دقت بیشتری در تشخیص زورگویی سایبری است. روش پیشنهادی دارای مزایای مختلفی است که از جمله آنها دقت بیشتر نسبت به روشهای یادگیری عمیق، ترکیب هوش گروهبی و معماری های مختلف یادگیری عمیق و استخراج ویژگی با استفاده از سه روش موازی است و این موضوع باعث می شود تا ویژگی های مختلف متن در نظر گرفته شود. از محدودیتهای روش پیشنهادی، پیچیدگی بیشتر آن نسبت به روشهای مانند CNN و LSTM است. دلایل ذیل را می توان برای برتری مدل پیشنهادی برای نسبت به روشهای مشابه ارایه داد:

- متعادلسازی مجموعه داده با روشهای هوشمند و بر پایه یادگیری عمیق و تئوری بازی مانند GAN باعث می‌شود تا تعداد نمونه‌های آموزشی کلاس اقلیت (نمونه‌های زورگویی سایبری) افزایش یابد و یادگیری فقط به کلاس اکثریت (نمونه‌های غیر زورگویی سایبری) محدود نشوند و مدل یادگیری خطای کمتری ارائه دهد.
 - استخراج ویژگی بر اساس اجتماع ویژگی‌های استخراج شده توسط سه روش GloVe, Word2Vec و TF-IDF باعث می‌شود تا کمتر ویژگی در تشخیص زورگویی سایبری از قلم بیفتد.
 - انتخاب ویژگی با الگوریتم عروس دریایی به دلیل رفتار جستجوی اکتشافی و محلی و توازن بین این دو جستجو توسط پارامتر C(t) باعث هوشمندی این الگوریتم برای جستجو در فضاهای چند بعدی شده است.
 - تلفیق توانایی شبکه عصبی CNN و LSTM باعث افزایش دقت در طبقه بندی توئیتها می‌شود.
- در پژوهش آتی تلاش می‌شود تا مجموعه داده با روشهای نظیر GAN متعادلسازی شود و به جای استفاده از LSTM از روشهای نظیر BiLSTM استفاده شود.

مراجع

- [1] H. Parlak Sert, & H. Başkale, "Students' increased time spent on social media, and their level of coronavirus anxiety during the pandemic, predict increased social media addiction," *Health Information & Libraries Journal*, vol. 40, no. 3, pp. 262-274, 7 Jul 2023, doi: 10.1111/hir.12448.
- [2] J. W. Patchin, & S. Hinduja, "Cyberbullying among Asian American youth before and during the COVID-19 pandemic," *Journal of school health*, vol. 93, no. 1, pp. 82-87, 2023, doi: 10.1111/josh.13249.
- [3] D. M. H. Kee, A. Anwar, & I. Vranjes, "Cyberbullying victimization and suicide ideation: The mediating role of psychological distress among Malaysian youth," *Computers in Human Behavior*, vol. 150, 108000, January 2024, doi: 10.1016/j.chb.2023.108000.
- [4] P. J. Macaulay, O. L. Steer, & L. R. Betts, "Bystander intervention to cyberbullying on social media," In *Handbook of Social Media Use Online Relationships, Security, Privacy and Society*, vol. 2, pp. 73-99, Academic Press, 2024, doi: 10.1016/B978-0-443-28804-3.00001-6.
- [5] E. Mahajan, H. Mahajan, & S. Kumar, "EnsMulHateCyb: Multilingual hate speech and cyberbully detection in online social media," *Expert Systems with Applications*, vol. 236, 121228, Feb 2024, doi: 10.1016/j.eswa.2023.121228
- [6] S. M. Fati, A. Muneer, A. Alwadain, & A. O. Balogun, "Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction," *Mathematics*, vol. 11, no. 16, 3567, 15 August 2023, doi: 10.3390/math11163567.
- [7] C. Iwendi, G. Srivastava, S. Khan, & P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimedia Systems*, vol. 29, no. 3, pp. 1839-1852, June 2023, doi: 10.1007/s00530-020-00701-5.
- [8] M. Dadvar, & K. Eckert, "Cyberbullying detection in social networks using deep learning based models. In *Big Data Analytics and Knowledge Discovery: 22nd International Conference, DaWaK 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings 22*, Springer International Publishing, pp. 245-255, Sep 2020, doi: 10.1007/978-3-030-59065-9_20.
- [9] A. Bozyiğit, S. Utku, & E. Nasibov, "Cyberbullying detection: Utilizing social media features," *Expert Systems with Applications*, vol. 179, 115001, 1 October 2021, doi: 10.1016/j.eswa.2021.115001.
- [10] T. Mahmud, M. Ptaszynski, J. Eronen, & F. Masui, "Cyberbullying detection for low-resource languages and dialects: Review of the state of the art," *Information Processing & Management*, vol. 60, no. 5, 103454, 27 June 2023, doi: 10.1016/j.ipm.2023.103454.

- [11] C. Iwendi, G. Srivastava, S. Khan, & P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimedia Systems*, vol. 29, no. 3, pp. 1839-1852, June 2023, doi: 10.1007/s00530-020-00701-5.
- [12] A. Akhter, U. K. Acharjee, M. A. Talukder, M. M. Islam, & M. A. Uddin, "A robust hybrid machine learning model for Bengali cyber bullying detection in social media," *Natural Language Processing Journal*, vol. 4, 100027, September 2023, doi: 10.1016/j.nlp.2023.100027.
- [13] H. Saini, H. Mehra, R. Rani, G. Jaiswal, A. Sharma, & A. Dev, "Enhancing cyberbullying detection: a comparative study of ensemble CNN-SVM and BERT models," *Social Network Analysis and Mining*, vol. 14, no. 1, 2 December 2023, doi: 10.1007/s13278-023-01158-w.
- [14] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, & H. D. E. Al-Ariki, "DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform," *IEEE Access*, vol. 10, pp. 25857-25871, 23 February 2022, doi: 10.1109/ACCESS.2022.3153675.
- [15] A. Kumar, & N. Sachdeva, "A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media," *World Wide Web*, vol. 25, no. 4, pp. 1537-1550, 01 July 2022, doi: 10.1007/s11280-021-00920-4.
- [16] A. Dass, & D. K. Daniel, "Cyberbullying Detection on Social Networks using LSTM Model," In *2022 International Conference on Innovations in Science and Technology for Sustainable Development (ICISTSD)*, pp. 293-296, IEEE, August 2020, doi: 10.22214/ijraset.2024.60420.
- [17] A. Alam, P. Verma, M. Tariq, A. Sarwar, B. Alamri, N. Zahra, & S. Urooj, "Jellyfish search optimization algorithm for mpp tracking of pv system," *Sustainability*, vol. 13, no. 21, 11736, 24 October 2021, doi: 10.3390/su132111736.
- [18] A. B. Barragán Martín, M. D. M. Molero Jurado, M. D. C. Pérez-Fuentes, M. D. M. Simon Marquez, Á. Martos Martínez, M. Sisto, & J. J. Gazquez Linares, "Study of cyberbullying among adolescents in recent years: A bibliometric analysis," *International journal of environmental research and public health*, vol. 18, no. 6, 3016, 15 March 2021, doi: 10.3390/ijerph18063016.
- [19] Á. Denche-Zamorano, S. Barrios-Fernandez, C. Galán-Arroyo, S. Sánchez-González, F. Montalva-Valenzuela, A. Castillo-Paredes, ... & P. R. Olivares, "Science mapping: a bibliometric analysis on cyberbullying and the psychological dimensions of the self," *International journal of environmental research and public health*, vol. 20, no. 1, 209, 23 December 2022, doi: 10.3390/ijerph20010209.
- [20] M. T. Hasan, M. A. E. Hossain, M. S. H. Mukta, A. Akter, M. Ahmed, & S. Islam, "A Review on Deep-Learning-Based Cyberbullying Detection," *Future Internet*, vol. 15, no. 5, 179, 11 May 2023, doi: 10.3390/fi15050179.
- [21] C. Iwendi, G. Srivastava, S. Khan, & P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimedia Systems*, vol. 29, no. 3, pp. 1839-1852, June 2023, doi: 10.1007/s00530-020-00701-5.
- [22] S. Paul, S. Saha, & J. P. Singh, "COVID-19 and cyberbullying: deep ensemble model to identify cyberbullying from code-switched languages during the pandemic," *Multimedia tools and applications*, vol. 82, no. 6, pp. 8773-8789, March 2023, doi: 10.1007/s11042-021-11601-9.
- [23] B. A. H. Murshed, Suresha, J. Abawajy, M. A. N. Saif, H. M. Abdulwahab, & F. A. Ghanem, "FAEO-ECNN: cyberbullying detection in social media platforms using topic modelling and deep learning," *Multimedia Tools and Applications*, vol. 82, no. 30, pp. 46611-46650, December 2023, doi: 10.1007/s11042-023-15372-3.
- [24] V. L. Paruchuri, & P. Rajesh, "CyberNet: a hybrid deep CNN with N-gram feature selection for cyberbullying detection in online social networks," *Evolutionary Intelligence*, vol. 16, no. 6, pp. 1935-1949, December 2023, doi: 10.1007/s12065-022-00774-3.
- [25] S. Giri, & S. Banerjee, "Performance analysis of annotation detection techniques for cyber-bullying

- messages using word-embedded deep neural networks,” *Social Network Analysis and Mining*, vol. 13, no. 23, 14 January 2023, doi: 10.1007/s13278-022-01023-2.
- [26] M. Al-Hashedi, L. K. Soon, H. N. Goh, A. H. L. Lim, & E. G. Siew, “Cyberbullying Detection Based on Emotion,” *IEEE Access*, vol. 11, no. 12, pp. 53907-53918, 29 May 2023, doi: 10.1109/ACCESS.2023.3280556.
- [27] N. A. Samee, U. Khan, S. Khan, M. M. Jamjoom, M. Sharif, & D. H. Kim, “Safeguarding Online Spaces: A Powerful Fusion of Federated Learning, Word Embeddings, and Emotional Features for Cyberbullying Detection,” *IEEE Access*, vol. 11, 2 November 2023, doi: 10.1109/ACCESS.2023.3329347.
- [28] A. Muneer, A. Alwadain, M. G. Ragab, & A. Alqushaibi, “Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT,” *Information*, vol. 14, no. 8, 467, 18 August 2023, doi: g/10.3390/info14080467.
- [29] A. F. Alqahtani, & M. Ilyas, “An Ensemble-Based Multi-Classification Machine Learning Classifiers Approach to Detect Multiple Classes of Cyberbullying,” *Machine Learning and Knowledge Extraction*, vol. 6, no. 1, pp.156-170, 12 January 2024, doi: 10.3390/make6010009.
- [30] L. Xiaoyan, R. C. Raga, & S. Xuemei, “GloVe-CNN-BiLSTM model for sentiment analysis on text reviews,” *Journal of Sensors*, vol. 2022, no. 1, 22 October 2022, doi: 10.1155/2022/7212366.
- [31] P. Sun, J. Wang, & Z. Dong, “CNN–LSTM Neural Network for Identification of Pre-Cooked Pasta Products in Different Physical States Using Infrared Spectroscopy,” *Sensors*, vol. 23, no. 10, 4815, 17 May 2023, doi: 10.3390/s23104815.
- [32] A. Muneer, & S. M. Fati, “A comparative analysis of machine learning techniques for cyberbullying detection on twitter,” *Future Internet*, vol. 12, no. 11, 187, 29 October 2020, doi: 10.3390/fi12110187.
- [33] S.A.R. Zaidi, Suspicious Communication on Social Platforms. [Online]. Available: <https://www.kaggle.com/datasets/syedabbasraza/suspicious-communication-on-social-platforms> [Accessed on 20 November 2022].
- [34] D. A. Kristiyanti, I. S. Sitanggang, & S. Nurdianti, (2023). “Feature selection using new version of v-shaped transfer function for salp swarm algorithm in sentiment analysis, ” *Computation*, vol. 11, no. 3, 56, 23 January 2023, doi: 10.3390/computation11030056.
- [35] A. Ruiz-Gándara, & L. Gonzalez-Abril, “Generative Adversarial Networks in Business and Social Science”, *Applied Sciences*, vol. 14, no. 17, pp.1-23, 7438, 20 August 2024, doi: 10.3390/app14177438.
- [36] R. Daniel, T. S. Murthy, C. D. Kumari, E. L. Lydia, M. K. Ishak, M. Hadjouni, & S. M. Mostafa, “Ensemble Learning With Tournament Selected Glowworm Swarm Optimization Algorithm for Cyberbullying Detection on Social Media”, *IEEE Access*, vol.11, pp. 123392-123400, January 2023, doi: 10.1109/access.2023.3326948
- [37] A. K. Al-Khowarizmi, I. P. Sari, & H. Maulana, “Optimization of support vector machine with cubic kernel function to detect cyberbullying in social networks”, *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 22, no. 2, pp. 329-339. 12 Jan 2024, doi: 10.12928/telkomnika.v22i2.25437
- [38] T. Nitya Harshitha, M. Prabu, E. Suganya, S. Sountharajan, D. P. Bavirisetti, N. Gadde, & L. S. Uppu, “ProTect: a hybrid deep learning model for proactive detection of cyberbullying on social media,” *Frontiers in artificial intelligence*, vol. 7, 1269366, 6 March 2024, doi: 10.3389/frai.2024.1269366.
- [39] S. Mirjalili, & A. Lewis, “The whale optimization algorithm,” *Advances in engineering software*, vol. 95, pp. 51-67, May 2016, doi: 10.1016/j.advengsoft.2016.01.008.
- [40] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, & H. Chen, “Harris hawks optimization: Algorithm and applications,” *Future generation computer systems*, vol. 97, pp. 849-872, 18 February 2019, doi: 10.1016/j.future.2019.02.028.

- [41] M. Dehghani, Z. Montazeri, E. Trojovská, & P. Trojovský, "Coati Optimization Algorithm: A new bio-inspired metaheuristic algorithm for solving optimization problems," *Knowledge-Based Systems*, vol. 259, 110011, 10 January 2023, doi: 10.1016/j.knosys.2022.110011.
- [42] J. Wang, W. C.Wang, X. X. Hu, L. Qiu, & H. F. Zang, "Black-winged kite algorithm: a nature-inspired meta-heuristic for solving benchmark functions and engineering problems," *Artificial Intelligence Review*, vol. 57, no. 4, 98, 4 February 2024, doi: 10.1007/s10462-024-10723-4.
- [43] D.Sultan, M.Mendes, A.Kassenkhan, & O. Akylbekov, "Hybrid CNN-LSTM Network for Cyberbullying Detection on Social Networks using Textual Contents, " *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, January 2023, doi : 10.14569/IJACSA.2023.0140978.

Detection of Cyberbullying in social networks with deep learning based on CNN and LSTM neural network

Mohsen Eghbali, PhD Student ¹ | Kamal Mirzaie, Assistant Professor ² | Reza Azizi, Assistant Professor ³

¹Department of Computer Engineering, Maybod Branch, Islamic Azad University, Maybod, Iran , m.eghbali@maybofiau.ac.ir

²Department of Computer Engineering, Maybod Branch, Islamic Azad University, Maybod, Iran, k.mirzaie@maybodiau.ac.ir

³ Department of Computer Engineering, Maybod Branch, Islamic Azad University, Maybod, Iran, aziz.reza@maybodiau.ac.ir

Correspondence

Kamal Mirzaie, Associate Professor, Department of Computer Engineering, Maybod Branch, Islamic Azad University, Maybod, Iran, k.mirzaie@maybodiau.ac.ir

Abstract

Cyberbullying is a type of violence in social networks for humiliation, ridicule, blackmail and threats. Cyberbullying has harmful effects and in some cases the victim suffers severe physical and mental damage. One of the promising approaches in cyberbullying detection is machine learning and deep learning algorithms. Detecting cyberbullying in social networks(SN) is complicated. Machine and deep learning methods for detecting cyberbullying in social networks can have a lot of errors. In this manuscript, to detect cyberbullying, the primary features of the text are extracted using three feature extraction methods: GloVe, Word2Vec, and TF-IDF methods. In the second phase, feature selection is done using the JSO algorithm, and finally, the essential features are considered as input to the 1DCNN and LSTM methods. Experiments are conducted on Twitter and Facebook datasets to detect cyberbullying. Results show that the accuracy, sensitivity(recall), and precision of the proposed method(JSO-CNN-LSTM) in detecting cyberbullying on Twitter are 98.23%, 97.86%, and 97.73%. The results show

that the proposed method is more accurate than CNN, LSTM, and BERT methods in detecting cyberbullying.

Keywords: Social network, Cyberbullying, Deep learning, Convolutional neural network, LSTM neural network.