Journal of New Technologies in Distributed Systems and Algorithmic Computing
Islamic Azad University of Sabzevar
E-ISSN: 3115-705X
https://sanad.iau.ir/journal/ntds

Reaserch Article

Detection of Cyberbullying in Social Networks with Deep Learning based on CNN and LSTM Neural Network

Mohsen Eghbali [1] | Kamal Mirzaie *[2] | Reza Azizi [3]

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

[2]Department 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

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

Main Subjects: Cyberbullying in Social Networks
Received: 18 June 2024
Revised: 26 Septembr 2024
Accepted: 26 March 2025

https://doi.org/10.82195/ntds.2025.1123148

Abstract

One of the promising approaches in cyberbullying detection is machine learning and deep learning algorithms. Nevertheless, 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 including 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. The 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 demonstrate 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.

فناوری های نوین در سیستم های توزیع شده و محاسبات الگوریتمی

دانشگاه آزاد اسلامی واحد سبزوار

E-ISSN: 3115-705X
https://sanad.iau.ir/journal/ntds

پژوهشی

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

محسن اقبالی[۱] | کمال میرزائی*[۲] | رضا عزیزی[۳]

[۱]گروه مهندسی کامپیوتر، واحد میبد، دانشگاه آزاد اسلامی، میبد، ایران،
M.eghbali@maybodiau.ac.ir

[۲]گروه مهندسی کامپیوتر، واحد میبد، دانشگاه آزاد اسلامی، میبد، ایران،
K.mirzaie@maybodiau.ac.ir

[۳]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

موضوع اصلی: قلدری سایبری در شبکه‌های اجتماعی
تاریخ دریافت : ۱۴۰۳/۳/۲۹
تاریخ بازنگری: ۱۴۰۳/۷/۵
تاریخ پذیرش: ۱۴۰۴/۱/۰۶

چکیده:

زورگویی سایبری نوعی از خشونت در شبکه‌های اجتماعی برای تحقیر، تمسخر، باج‌خواهی و تهدید است. زورگویی سایبری آثار زیان‌باری دارد و در برخی موارد قربانی آسیب جسمی و روحی شدید می‌بینند. یکی از رویکردهای امیدوارکننده در تشخیص زورگویی سایبری استفاده از الگوریتم‌های یادگیری ماشین و یادگیری عمیق است. بااین‌حال، تشخیص آزار سایبری در شبکه‌های اجتماعی پیچیده است و یک الگوریتم یادگیری ماشین و یادگیری عمیق به‌تنهایی 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 ویژگی‌های اولیه متن استخراج می‌شود. در مرحله دوم انتخاب ویژگی با استفاده از الگوریتم JSO انجام می‌شود و در نهایت ویژگی‌های مهم به‌عنوان ورودی روش 1DCNN و LSTM در نظر گرفته می‌شود. آزمایش‌ها در مجموعه‌داده توییتر و فیس‌بوک برای تشخیص زورگویی سایبری انجام می‌شود. آزمایش‌ها نشان می‌دهد دقت، حساسیت و صحت روش پیشنهادی در تشخیص زورگویی سایبری در مجموعه‌داده توییتر به ترتیب برابر ۹۸/۲۳ درصد، ۹۷/۸۶ درصد و ۹۷/۷۳ درصد است. نتایج نشان می‌دهد روش پیشنهادی نسبت به روش‌های CNN، LSTM و BERT در تشخیص زورگویی سایبری دارای دقت بیشتری است.

https://doi.org/10.82195/ntds.2025.1123148

کلیدواژه‌ها: شبکه اجتماعی، قلدری سایبری، یادگیری عمیق، شبکه عصبی کانولوشن، شبکه عصبی LSTM

۱-مقدمه

ظهور و پذیرش رسانه‌های اجتماعی به‌عنوان یک بستر ارتباطی و تعاملی مهم در دنیای امروز باعث شده تا این شبکه‌ها نقش مهمی در ارتباطات داشته باشند. مطالعات نشان می‌دهد بیش از ۳ میلیارد کاربر فعال وجود دارد که روزانه از طریق رسانه‌های 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، پلتفرم‌های رسانه‌های اجتماعی، جوامع مجازی مبتنی بر رابطه و علاقه را فعال می‌کنند که امکان اتصال و شبکه‌سازی را در بین مردم فراهم می‌کند. افزایش محبوبیت این پلتفرم‌های رسانه‌های اجتماعی (فیس‌بوک، توییتر، اینستاگرام، Tinder و غیره) امکان اشتراک‌گذاری اشکال مختلف پیام‌های چندرسانه‌ای را در میان کاربران خود فراهم می‌کند. عملیات این پلتفرم‌های رسانه‌های اجتماعی در طول همه‌گیری کووید۱۹[1] بسیار مهم بود، زیرا داده‌های بلادرنگ درباره رویدادها و اکتشافات به راحتی در بین مردم در مکان‌های مختلف منتشر می‌شد[۲].

اگرچه آشکار است که پلتفرم‌های رسانه‌های اجتماعی مزایای متعددی را برای کاربران خود ارائه می‌کنند، اما این پلتفرم‌ها ممکن است برای اهداف بدخواهانه نیز مورد سوءاستفاده قرار گیرند. زورگویی سایبری[2] نمونه بارز و عمیقی از این روندهای بدخواهانه است[۳]. به طور خاص، فراگیرشدن این پلتفرم‌های رسانه‌های اجتماعی به طور غیرقابل‌انکاری جرقه، پیشرفت و تشدید قلدری را در میان کاربران آن ایجاد کرده است. به نظر می‌رسد قلدری در طول تاریخ‌تمدن بشری وجود داشته است و مستلزم آزار و اذیت شخصی از طریق شرمساری یا آزاردادن او به هر طریقی است که باعث آسیب عاطفی، روانی یا جسمی شود. هنگامی که این حمله از طریق اینترنت رخ می‌دهد، به‌عنوان آزار سایبری یا قربانی سایبری شناخته می‌شود[۴]. قلدری سایبری را می‌توان به‌عنوان قلدری به شخص یا گروهی از افراد که معمولاً به‌عنوان قربانی(ها) شناخته می‌شوند، با کمک اینترنت، موبایل یا دستگاه الکترونیکی با ارسال محتوای متنی یا غیرمتنی نامناسب چندرسانه‌ای توصیف کرد. به‌عبارت‌دیگر، آزار و اذیت سایبری نمایش مستمر اعمال ناخوشایندی است که بر روی قربانی(ها) برای ایجاد ترس، آزار، درد یا آسیب از طریق رسانه‌های الکترونیکی و بسترهای رسانه‌های اجتماعی انجام می‌شود. آزار و اذیت سایبری یک مشکل بزرگ در دهه گذشته در فضای مجازی بوده است که بیشتر کودکان و نوجوانان را تحت‌تأثیر قرار داده است[۵]. به‌عنوان‌مثال، 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 می‌گیرند [۶]. بر اساس آمار، حدود ۱۸ درصد از جوانان اروپایی تحت‌تأثیر فردی قرار گرفته‌اند که آنها را از طریق اینترنت و تلفن همراه مورد آزار و اذیت قرار می‌دهد [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 می‌کنند [۸]. در ایران متأسفانه آماری دقیق از زورگویی سایبری در فضای مجازی و شبکه‌های اجتماعی انتشار پیدا نکرده است؛ اما در سایر کشورها این چالش فضای مجازی گزارش شده است و به‌عنوان نمونه در سوئد که کشوری توسعه‌یافته است، شیوع آزار و اذیت سایبری به نقطه اوج رسیده و به‌تدریج درحال‌رشد و بدتر شدن است [۹]. این گزارش‌ها نشان می‌دهد که توسعه یک راه‌حل قابل‌قبول، 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. باتوجه‌به توسعه فناوری، امتیازات اجتماعی (ناشناس بودن) که رسانه‌های اجتماعی ارائه می‌کنند و دسترسی به مخاطبان گسترده‌تر، آزار و اذیت سایبری به طور تصاعدی رشد کرده است. این موضوع نیازمند توسعه ابزارها و استراتژی‌های هوشمندی است که با استفاده از داده‌های چندرسانه‌ای اجتماعی موجود، آزار و اذیت سایبری را شناسایی، شناسایی و تجزیه‌وتحلیل می‌کنند تا تأثیر مضر آن را کاهش دهند. تشخیص خودکار مزاحمت سایبری یک موضوع طبقه‌بندی است، باهدف دسته‌بندی هر نظر/پست/پیام/تصویر توهین‌آمیز به‌عنوان قلدری یا غیر آزاردهنده [۱۰].

اگرچه برخی از پلتفرم‌های رسانه‌های اجتماعی، مانند یوتیوب و توییتر، مراکز ایمنی را برای نظارت و کنترل آزار و اذیت سایبری تعبیه کرده‌اند، اما این مشکل همچنان پابرجاست و نیاز به راه‌حل‌های قطعی‌تری وجود دارد. اخیراً شناسایی خودکار آزار و اذیت سایبری با استفاده از یادگیری عمیق[3]۱۱و یادگیری ماشین[4][۱۲]مورد توجه زیادی قرار گرفته است. روشهای

[1] COVID-19
[2] Cyberbullying
[3] Deep learning (DL)
[4] Machine learning (ML)

یادگیری ماشین و یادگیری عمیق می‌توانند الگوی متن انتشار یافته برای تشخیص زورگویی سایبری را استخراج نمایند و توئیتهای عادی را از توئیتهای زورگویی سایبری تشخیص دهند. از جمله روشهای یادگیری عمیق که برای تشخیص زورگویی سایبری استفاده می‌شود می‌توان به شبکه عصبی کانولوشن[۱][۱۳] شبکه عصبی بازگشتی[۲][۱۴] شبکه عصبی مبتنی بر دروازه[۳][۱۵] شبکه عصبی حافظه کوتاه مدت[۴][۱۶]اشاره نمود. در این روشها برای تشخیص زورگویی سایبری تلاش شده 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[۵]۱۷[نتخاب شده و تحویل طبقه‌بندی کننده ترکیبی CNN و LSTM می‌شود. هدف از این مقاله، ارایه 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 GloVe، Word2Vec و TF-IDF
- ارائه یک نسخه باینری از الگوریتم عروس دریایی
- تلفیق دو طبقه‌بندی‌کننده 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 پیشنهادهای 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. بااین‌حال، قلدری نه‌تنها در مدرسه اتفاق می‌افتد. با توسعه فناوری‌های 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 بین‌فردی را تغییر داده است. قلدری سایبری هر دو به‌عنوان یک مشکل شدید بهداشت عمومی شناخته شده‌اند، زیرا تهدیدی برای توسعه سلامت روان و رفاه کودکان و نوجوانان هستند. قلدری سایبری و زورگویی سایبری به‌نوعی 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 موردبحث 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].

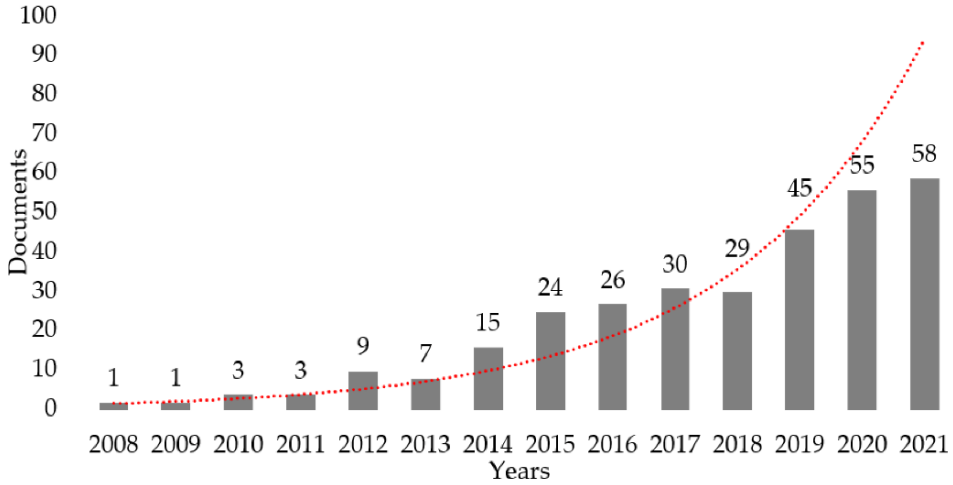
باتوجه‌به ابعاد مختلف روان‌شناختی، از یک سو، قربانیان سایبری از اثرات منفی مختلفی مانند افزایش علائم افسردگی و اضطراب و استرس عاطفی بیشتر رنج می‌برند. علاوه بر این، ابعاد عاطفی نیز می‌تواند تحت‌تأثیر قرار گیرد، زیرا آزار و اذیت سایبری باعث کاهش عزت‌نفس، نارضایتی از زندگی و خودکارآمدی عاطفی پایین می‌شود. تنهایی و خودپنداره پایین برخی از رایج‌ترین اثرات

[1] Convolutional neural network (CNN)
[2] Recurrent neural network (RNN)
[3] Gated recurrent units (GRUs)
[4] Long Short-Term Memory (LSTM)
[5] Jellyfish search algorithm

و پیش‌بینی‌کننده‌ها در میان قربانیان است. معمولاً قربانیان از تنهایی، عدم حمایت خانوادگی و اجتماعی و 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، پیش‌بینی‌کننده‌ای 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 قلدری سایبری در زمان بعدی مرتبط است. تفاوت‌های اساسی بین قربانیان سایبری و زورگویان اینترنتی و جنسیت در مورد عزت‌نفس، خودکنترلی، حمایت اجتماعی و خانوادگی و پرخاشگری یافت شده است.

۲-۱- اهمیت تشخیص زورگویی سایبری

در شکل (۱)، تعداد مطالعات در زمینه زورگویی سایبری بین سالهای۲۰۰۸ تا ۲۰۲۱ را نشان می‌دهد. با توجه به نمودار می‌توان 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 می‌دهد[۱۹].



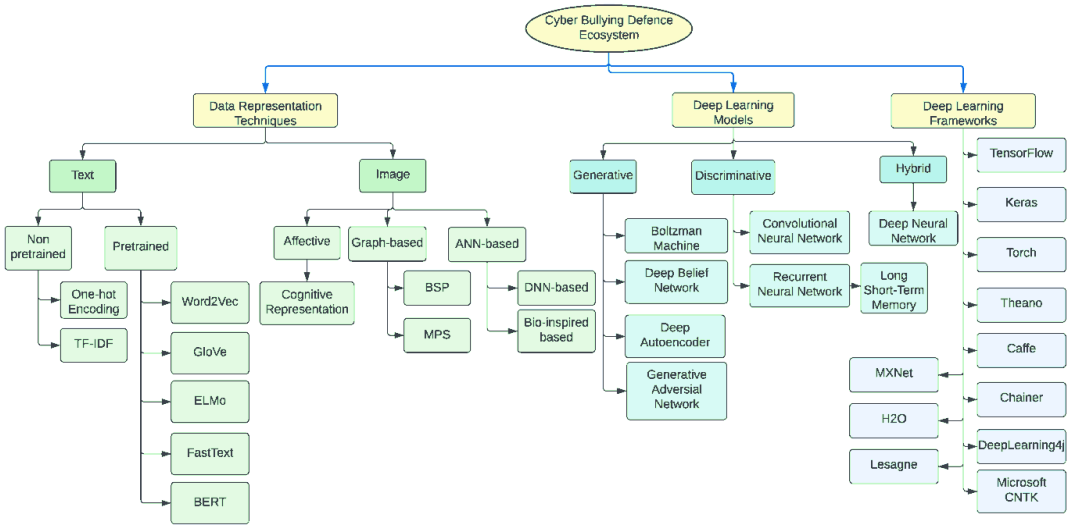
شکل ۱: افزایش مطالعات زورگویی سایبری در شبکه‌های اجتماعی در سالهای اخیر[۱۹]

Figure 1: An increase in studies on cyberbullying on social media in recent years

شناسایی آزار و اذیت سایبری برای جلوگیری از مشکل تهدیدکننده در شبکه‌های اجتماعی مهم است. تشخیص آزار و اذیت سایبری به دلیل فقدان پارامترهای قابل‌شناسایی و عدم وجود استاندارد قابل‌سنجش، کار دشواری است.

۲-۲- روشهای تشخیص زورگویی سایبری

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



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

Figure 2: Classification of methods for detecting cyberbullying on social media

باتوجه‌به شکل (۲)، محققان از الگوریتم‌های یادگیری ماشین سنتی برای شناسایی آزار سایبری (به عنوان مثال، قالب متن و تصویر) استفاده می‌کنند، در حالی که اکثر راه‌حل‌های موجود مبتنی بر روش‌های یادگیری نظارت شده هستند.

به دلیل ماهیت ذهنی عبارات قلدر، مدل‌های سنتی یادگیری ماشین در تشخیص آزار سایبری نسبت به رویکردهای مبتنی بر یادگیری عمیق عملکرد کمتری دارند. گزارش‌ها بر اساس[۲۰] نشان می‌دهد که مدل‌های یادگیری عمیق از الگوریتم‌های سنتی یادگیری ماشین در مورد شناسایی آزار سایبری بهتر عمل می‌کنند. شبکه‌های عصبی عمیق مانند شبکه عصبی بازگشتی، واحد بازگشتی دروازه‌ای، حافظه کوتاه‌مدت بلندمدت و چندین مدل یادگیری عمیق دیگر را می‌توان برای تشخیص زورگویی سایبری استفاده کرد.

۲-۳-کارهای مرتبط

در[۲۱]، در سال ۲۰۲۳، راه‌حل‌های تشخیص آزار سایبری بر اساس معماری‌های یادگیری عمیق را بررسی کردند. مطالعات آنها نشان می‌دهد آزار و اذیت سایبری، سوءرفتار آنلاین آزاردهنده و نگران‌کننده است. به اشکال مختلف ظاهر می‌شود و معمولاً در اکثر شبکه‌های اجتماعی به‌صورت متنی است. سیستم‌های هوشمند برای تشخیص خودکار این حوادث ضروری است. برخی از آزمایش‌های اخیر این مسئله را با مدل‌های یادگیری ماشین سنتی حل کرده‌اند. اکثر مدل‌ها در یک‌زمان در یک شبکه اجتماعی اعمال شده‌اند. آخرین تحقیقات نشان داده است که مدل‌های مختلف مبتنی بر الگوریتم‌های یادگیری عمیق بر تشخیص آزار سایبری تأثیر می‌گذارند. این پژوهش یک تحلیل تجربی برای تعیین اثربخشی و عملکرد الگوریتم‌های یادگیری عمیق در تشخیص توهین در تفسیر اجتماعی انجام می‌دهد. چهار مدل یادگیری عمیق زیر برای نتایج تجربی استفاده شده است که عبارت‌اند از حافظه کوتاه‌مدت دو جهته (BiLSTM)، واحدهای بازگشتی دردار (GRU)، حافظه کوتاه‌مدت طولانی (LSTM)، و شبکه عصبی بازگشتی (RNN). مراحل پیش‌پردازش داده‌ها دنبال شد که شامل پاک‌سازی متن، نشانه‌گذاری، ریشه‌یابی، Lemmatization و حذف کلمات توقف می‌شود. پس از انجام پیش‌پردازش داده‌ها، داده‌های متنی تمیز برای پیش‌بینی به الگوریتم‌های یادگیری عمیق منتقل می‌شوند. نتایج نشان می‌دهد 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 الگوریتم‌های یادگیری ماشین مختلف و یادگیری عمیق مانند شبکه عصبی کانولوشن، BiLSTM، BERT را برای تشخیص آزار سایبری انگلیسی - هندی استفاده نمودند. ارزیابی آنها نشان می‌دهد دقت مدل‌های یادگیری عمیق در تشخیص زورگویی سایبری بیشتر از روش‌های یادگیری ماشین است. در[۲۳]، در سال ۲۰۲۳، تشخیص آزار و اذیت سایبری در رسانه‌های

اجتماعی با استفاده از مدل‌سازی یادگیری عمیق ارائه شده است. رویکرد پیشنهادی آنها مدل‌سازی موضوعی مبتنی بر خوشه‌بندی بهینه‌سازی تعادل تطبیقی فازی (FAEO) و شبکه عصبی پیچیده (ECNN) را برای افزایش دقت فرایند تشخیص زورگویی سایبری را ادغام می‌کند. در ابتدا، پیش‌پردازش به‌منظور پاک‌سازی مجموعه‌داده انجام می‌شود. در مرحله بعد، ویژگی‌ها با استفاده از چندین مدل استخراج می‌شوند. بهینه‌سازی تعادل تطبیقی فازی بدون نظارت برای کشف موضوعات پنهان از داده‌های ورودی از پیش‌پردازش شده استفاده می‌شود که به طور خودکار داده‌های متن را بررسی می‌کند و خوشه‌هایی از کلمات را ایجاد می‌کند. در نهایت، طبقه‌بندی آزار سایبری از الگوریتم ECNN و الگوریتم بهینه‌سازی باران برای شناسایی زورگویی سایبری از پست‌ها/متون استفاده می‌کند. مدل پیشنهادی آنها در تشخیص آزار سایبری از مدل‌های مانند شبکه عصبی کانولوشن بهتر عملکرد. در[۲۴]، در سال ۲۰۲۳، یک شبکه عصبی کانولوشن ترکیبی با انتخاب ویژگی N-gram برای تشخیص آزار سایبری در شبکه‌های اجتماعی آنلاین ارائه دادند. در این پژوهش مکانیسم یادگیری عمیق را با مدل شبکه عصبی کانولوشنال هفت‌لایه با انتخاب ویژگی N-gram، به نام CNN عمیق ترکیبی اتخاذ شده است. با استفاده از این مدل در چهار معماری مرحله‌ای، تشخیص آزار سایبری سطح کلمه و سطح کاراکتر به طور مؤثری انجام می‌شود. مهاجمان با استفاده از انواع کلمات مبهم و توهین‌آمیز تأثیر منفی بر شبکه‌های اجتماعی دارند و لذا الگوی تشخیص توهین‌ها پیچیده است و نیاز به ابزارهای کارآمد 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 می‌شود. تجزیه‌وتحلیل نتایج کیفی ثابت می‌کند که رویکرد پیشنهادی عملکرد بسیار بهتری را در مقایسه با رویکردهای مرسوم ارائه می‌دهد.

در[۲۵]، در سال ۲۰۲۳، تحلیل عملکرد تکنیک‌های تشخیص حاشیه‌نویسی برای پیام‌های قلدری سایبری با استفاده از شبکه‌های عصبی عمیق تعبیه‌شده در کلمه را پیشنهاد نمودند. پیام‌های رسانه‌های اجتماعی ساختاری ندارند؛ زیرا شامل متن، پیوند URL، شکلک‌ها، اختصارات و غیره است. بیشتر کارهای قبلی برای شناسایی پیام‌های قلدری تنها با درنظرگرفتن کلمات مهم در متن انجام شده‌اند، و از سایر ویژگی‌های پیام مانند لینک‌های URL، ایموجی‌ها غفلت می‌کنند. در این پژوهش، یک تکنیک پیش‌پردازش پیشرفته با درنظرگرفتن برخی از ویژگی‌های موجود در پیام‌ها مانند URL، مخفف، شماره، ایموجی‌ها و غیره برای شناسایی پیام‌های قلدری پیشنهاد شده است. در این کار، شش مدل، یعنی سه مدل یادگیری عمیق ترکیب شده با دو مدل مختلف جاسازی کلمه برای تشخیص حاشیه‌نویسی به کار گرفته شده‌اند. عملکرد هر یک از این شش مدل دو بار با استفاده از پیش‌پردازش سنتی و پیش‌پردازش پیشرفته پیشنهادی اندازه‌گیری می‌شود. ارزیابی آنها نشان می‌دهد روش پیشنهادی بکار رفته آنها از روش‌های 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 می‌تواند توسط آزار و اذیت سایبری ایجاد شود. این پژوهش مدل‌های تشخیص آزار سایبری را پیشنهاد می‌کند که بر اساس ویژگی‌های زمینه‌ای، احساسات و احساسات آموزش داده می‌شوند. یک مدل تشخیص احساسات (EDM با استفاده از مجموعه‌داده‌های توییتر که از نظر حاشیه‌نویسی بهبود یافته‌اند، ساخته شد. نتایج نشان می‌دهد که خشم، ترس و احساس گناه اصلی‌ترین احساسات مرتبط با آزار و اذیت اینترنتی بودند. متعاقباً، احساسات استخراج‌شده به‌عنوان ویژگی علاوه بر ویژگی‌های زمینه‌ای و احساسی برای آموزش مدل‌هایی برای تشخیص آزار سایبری استفاده شد. در[۲۷]، در سال ۲۰۲۳، یک روش ترکیبی مبتنی بر یادگیری فدرال، جاسازی کلمات، و ویژگی‌های احساسی برای تشخیص آزار سایبری ارائه دادند. در این مطالعه، آنها یک تحقیق عمیق در مورد ادغام جاسازی‌های کلمه، ویژگی‌های احساسی و یادگیری فدرال انجام دادند که با استفاده از BERT، شبکه‌های عصبی کانولوشنال (CNN)، شبکه‌های عصبی عمیق (DNN) و مدل‌های حافظه (LSTM) ارائه شده است. فراپارامترها و معماری عصبی برای یافتن پیکربندی‌های بهینه موردبررسی قرار می‌گیرند که منجر به تولید نتایج برتر می‌شود. این تکنیک‌ها در زمینه تشخیص آزار سایبری، با استفاده از مجموعه‌داده‌های آزار سایبری چند پلتفرمی (رسانه‌های اجتماعی) در دسترس عموم استفاده می‌شوند. نتایج نشان‌دهنده توانایی افزایش‌یافته برای شناسایی و مبارزه مؤثر با حوادث مزاحم سایبری است که به ایجاد محیط‌های آنلاین امن‌تر کمک می‌کند. به‌ویژه، مدل BERT به طور مداوم از سایر مدل‌های یادگیری عمیق (CNN، DNN، LSTM) در تشخیص آزار سایبری بهتر عمل می‌کند و درعین‌حال حریم خصوصی مجموعه‌داده‌های محلی را برای هر پلتفرم اجتماعی از طریق تنظیمات یادگیری فدرال بهبودیافته ما حفظ می‌کند. در[۲۸]، در سال ۲۰۲۳، تشخیص قلدری سایبری در

رسانه‌های اجتماعی با استفاده از آموزش گروهی و BERT را پیشنهاد دادند. این پژوهش یک رویکرد یادگیری انباشته برای تشخیص آزار سایبری در توییتر با استفاده از ترکیبی از روش‌های شبکه عصبی عمیق ارائه می‌کند. همچنین این پژوهش یک مدل BERT اصلاح شده را معرفی می‌کند. مجموعه‌داده مورداستفاده در این مطالعه از توییتر جمع‌آوری‌شده و برای حذف اطلاعات نامربوط پیش‌پردازش شده است. فرایند استخراج ویژگی شامل استفاده از word2vec به همراه CBOW برای تشکیل وزن‌ها در لایه جاسازی بود. این ویژگی‌ها سپس به یک مکانیسم کانولوشنی و ادغام تبدیل شدند، و به طور مؤثر ابعاد آنها را کاهش دادند و ویژگی‌های تغییرناپذیر موقعیت کلمات توهین‌آمیز را به دست آوردند. اعتبارسنجی مدل انباشته پیشنهادی و BERT-M با استفاده از معیارهای ارزیابی مدل شناخته شده انجام شد. نتایج آزمایش‌ها نشان داد که رویکرد یادگیری گروه انباشته به‌دقت ۹۷.۴ درصد در تشخیص آزار سایبری در مجموعه‌داده توییتر و ۹۰.۹۷ درصد در مجموعه‌داده‌های ترکیبی توییتر و فیس‌بوک دست‌یافت. در[۲۹]، در سال ۲۰۲۴، رویکرد مبتنی بر یادگیری ماشین برای شناسایی زورگویی سایبری ارائه شده است. در این پژوهش درخت‌های تصمیم‌گیری، جنگل تصادفی و XGBoost، بکار گرفته شده است. نتایج آزمایش‌ها نشان داد که این چارچوب می‌تواند شش نوع مختلف آزار و اذیت سایبری را بادقت ۰.۹۰۷۱ به طور مؤثرتری شناسایی کند. در[۳۶]، سال ۲۰۲۴، یک الگوریتم بهینه‌سازی حشره شب‌تاب برای تشخیص آزار سایبری در رسانه‌های اجتماعی با آموزش گروهی ارائه شده است. این مطالعه بر طراحی و توسعه یادگیری عمیق گروهی با الگوریتم بهینه‌سازی ازدحام کرم شب‌تاب 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 فرایند 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 مؤثر 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 به‌دست‌آمده 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 مبتنی بر جنگل CNN برای طبقه‌بندی متن استفاده کرده است که نقاط قوت هر دو رویکرد را ترکیب می‌کند. مجموعه‌داده‌های 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 به‌دقت ۹۶ درصد دست‌یافت 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 تجزیه‌وتحلیل 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 مجموعه‌داده‌های 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 نشان‌دهنده 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
Table 1: Summary of Related Works

پژوهش	رویکرد	مزیت	چالش	حوزه
در[۲۱]، در سال ۲۰۲۳،	مرور روش‌های تشخیص آزار سایبری بر اساس معماری‌های یادگیری عمیق	اثبات کارایی روش‌های مبتنی بر شبکه عصبی LSTM	عدم بررسی تعادل و بالانس مجموعه‌داده‌ها و تأثیر آن بر دقت	مقاله مروری در حوزه یادگیری عمیق
در[۲۲]، در سال ۲۰۲۳،	مدل گروهی یادگیری عمیق	دقت بیشتر از شبکه عصبی کانولوشن، BiLSTM، BERT	زمان یادگیری زیاد و پیچیدگی بالا و عدم انتخاب ویژگی	یادگیری عمیق - یادگیری گروهی
در[۲۳]، در سال ۲۰۲۳،	تطبیقی فازی (FAEO) و شبکه عصبی پیچیده (ECNN)	از مدل‌های مانند شبکه عصبی کانولوشن بهتر عملکرد.	پیچیدگی بالا و عدم انتخاب ویژگی هوشمندانه	یادگیری عمیق - منطق فازی
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،	شبکه‌های عصبی عمیق تعبیه‌شده در کلمه	شامل ویژگی‌های بصری و متن مانند متن پیام، پیوند URL، شکلک‌ها، اختصارات و غیره است.	عدم انتخاب ویژگی هوشمندانه و عدم تعادل مجموعه‌داده	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 مجموعه‌داده	یادگیری گروهی و پردازش زبان طبیعی
در[۲۹]، در سال ۲۰۲۴،	درخت‌های تصمیم‌گیری، جنگل تصادفی و XGBoost	عدم پیچیدگی	دقت اندک	یادگیری ماشین
در[۳۶]، سال ۲۰۲۴،	الگوریتم بهینه‌سازی حشره شب‌تاب با پردازش زبان طبیعی	دقت بالا	عدم متعادل‌سازی مجموعه‌داده	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 بینش‌های زبانی و روان‌شناختی	دقت نسبتاً اندک	یادگیری ماشین و پردازش زبان طبیعی

بررسی کارهای مرتبط نشان می‌دهد که هر کدام از آنها دارای مزایا و معایبی هستند. بررسی کارها نشان می‌دهد روش‌های یادگیری عمیق برای تشخیص زورگویی سایبری یک روش مؤثر و کارآمد است؛ اما بااین‌وجود چالش‌های برای آن وجود دارد. عدم تعادل یا بالانس در مجموعه‌داده آموزشی یک چالش مهم است که دقت روش‌های یادگیری عمیق را کاهش می‌دهد. در روش پیشنهادی برای رفع این چالش از تئوری بازی و شبکه عصبی GAN برای متعادل‌سازی مجموعه‌داده استفاده می‌شود تا تعداد نمونه‌های کلاس اقلیت افزایش داده شود. یکی از چالش‌های دیگر روش‌های تشخیص حملات سایبری عدم استفاده از تکنیک‌های هوش گروهی جدید برای تشخیص ویژگی‌های 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 می‌توان 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 ویژگی‌های مرتبط با زورگویی سایبری با الگوریتم عروس دریایی
- تلفیق معماری LSTM و CNN در تشخیص زورگویی سایبری
- متعادل‌سازی مجموعه‌داده برای افزایش دقت مدل یادگیری با تئوری بازی و شبکه عصبیGAN

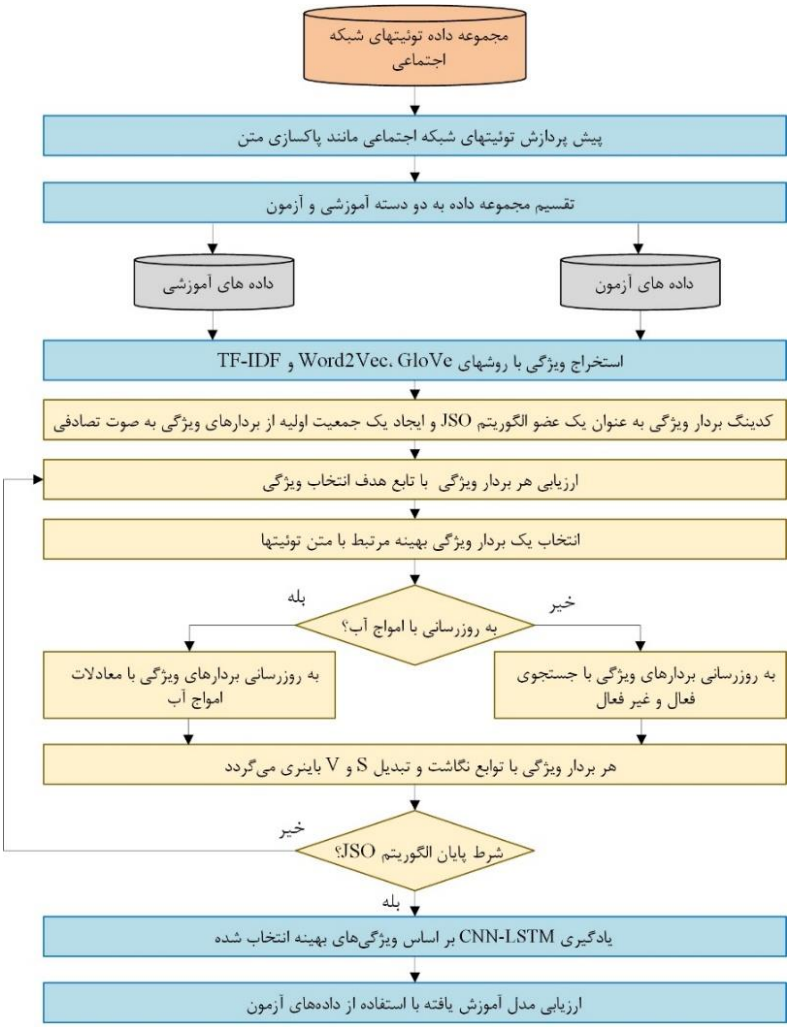
۳- روش پیشنهادی

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

- پیش‌پردازش مجموعه‌داده شامل پاک‌سازی مجموعه‌داده
- متعادل‌سازی مجموعه‌داده 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 ویژگی‌های مهم زورگویی سایبری
- طبقه‌بندی توییت‌ها با استفاده از شبکه عصبی CNN و LSTM بهبودیافته با الگوریتم عروس دریایی

شکل ۳: مراحل سیستم تشخیص توئیتهای زورگویی سایبری
Figure 2: Stages of a cyberbullying detection system for tweets

در شکل ۴، چارچوب روش پیشنهادی برای تشخیص زورگویی سایبری در شبکه‌های اجتماعی ارایه شده است.



شکل ۴: چارچوب سیستم تشخیص توئیتهای زورگویی سایبری
Figure 4: Framework of a system for detecting cyberbullying tweets

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

- در ابتدا مجموعه‌داده که مجموعه‌ای از توییت‌های شبکه اجتماعی نظیر توییتر است گردآوری می‌شود و در ادامه مورد پیش‌پردازش مانند پاک‌سازی متن، توکن‌گذاری و غیره قرار گرفته می‌شود.
- در مرحله بعدی مجموعه‌داده به دودسته آموزشی و آزمون تقسیم می‌شود و از نمونه‌های آموزشی برای ایجاد مدل طبقه‌بندی و از نمونه‌های آزمون برای ارزیابی مدل ایجاد شده استفاده می‌شود.
- در مرحله بعدی فاز استخراج ویژگی به سه روش ترکیبی GloVe، Word2Vec و TF-IDF انجام می‌شود به‌گونه‌ای هر روش یک مجموعه ویژگی را استخراج کرده و در نهایت این سه مجموعه ویژگی با هم ترکیب می‌شود.
- انتخاب ویژگی‌ها در مرحله بعدی با استفاده از الگوریتم عروس دریایی انجام می‌شود و هر بردار ویژگی یک عروس دریایی است که با جستجو بر اساس امواج آب و رفتار گروهی (جستجوی فعال و غیرفعال) به‌روزرسانی می‌شوند.
- طبقه‌بندی توییت‌ها به دودسته عادی و زورگویی سایبری در فاز آخر و توسط طبقه‌بندی‌کننده CNN-LSTM انجام می‌شود.
- در نهایت مدل آموزش‌یافته به کمک داده‌های آزمون مورد ارزیابی قرار گرفته می‌شود و با شاخص‌های مانند دقت و حساسیت با روش‌های مشابه مورد مقایسه قرار گرفته می‌شود.

۳- ۱-پیش پردازش

موفقیت یک مدل یادگیری ماشین به‌شدت به کیفیت داده‌های آموزشی متکی است که بر اهمیت پیش‌پردازش داده‌ها در فرایند توسعه تأکید می‌کند. تمیزکردن مؤثر داده‌ها برای اطمینان از وضوح و جلوگیری از کاهش دقت هنگام واردکردن داده‌ها به مدل ضروری است. با استفاده از کتابخانه NLTK در پایتون ابزاری که به طور گسترده برای پیش‌پردازش داده‌ها مورداستفاده قرار می‌گیرد، می‌توان داده‌ها را به‌دقت پاک‌سازی کرد. از طریق NLTK، عناصر نامطلوب مانند برچسب‌ها، هشتگ‌ها، لینک‌ها، موارد تکراری، علائم نگارشی و اعداد را می‌توان حذف کرد. علاوه بر این می‌توان همه توییت‌ها به طور یکنواخت به حروف کوچک تبدیل نمود تا متن یک‌دست حاصل شود. تعدادی از پیش‌پردازش‌های مهم بکار رفته در روش پیشنهادی به شرح ذیل است:

- پاک‌سازی: با عبارات منظم تمام علائم نگارشی، پیوندها، هشتگ‌ها، نمادها، متن‌های تکراری و الفبای غیرانگلیسی را می‌توان حذف کرد.
- حذف کلمات توقف[1]
- یافتن بن و ریشه کلمات[2]: برای این منظور می‌توان از کتابخانه WordNet Lemmatizer برای انجام فرایند Lemmatization Word جهت یافتن ریشه کلمات استفاده نمود.

۳-۲- متعادلسازی داده ها با شبکه عصبی متخاصم

در بیشتر موارد توییت‌های بکار رفته در مورد زورگویی سایبری تعداد کمتری نسبت به سایر توییت‌های عادی دارند و در واقع کلاس اقلیت محسوب می‌شوند. آموزش روش‌های یادگیری ماشین و یادگیری عمیق روی داده‌های نامتعادل باعث می‌شود که دقت تشخیص زورگویی سایبری توسط این روش‌ها کاهش داده شود. یک روش برای متعادل‌سازی مجموعه‌داده‌ها آن است که تعدادی نمونه مصنوعی به کلاس اقلیت (کلاس زورگویی سایبری اجتماعی) اضافه شود. شبکه‌های متخاصم مولد[3] به موضوع تحقیقاتی اخیر و به‌سرعت درحال‌توسعه در یادگیری ماشین تبدیل شده‌اند. هدف اصلیGAN ها تولید خودکار داده‌ها است. تفاوت اصلی آن با سایر مدل‌های تولیدی این است که مستقیماً از توزیع داده‌های واقعی استفاده نمی‌کند. در عوض، از طریق یک‌طبقه بندی کننده عمل می‌کند. مدل مولد[4]، تصادفی است و خود را گام‌به‌گام بر اساس پاسخ طبقه‌بندی‌کننده تنظیم می‌کند و به طور مداوم خروجی را اصلاح می‌کند تا زمانی که متمایزکننده[5] نتواند بین داده‌های واقعی و مصنوعی تمایز قائل شود. برای دستیابی به این هدف، از دو شبکه عصبی – یعنی مولد و متمایزکننده – استفاده می‌شود که در یک فرایند رقابتی شرکت می‌کنند. مولد داده‌های مصنوعی شبیه داده‌های واقعی تولید می‌کند، درحالی‌که تمیزکننده تلاش می‌کند بین داده‌های واقعی و داده‌های ارائه‌شده 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 تفکیک‌کننده است که یک نمونه واقعی x را پردازش می‌کند.
- D(G(z)): پیش‌بینی تمایزدهنده بر روی داده‌های مصنوعی است.
- Px و Pz به ترتیب توزیع داده‌های واقعی و نویز هستند.
- Ex و EG(z) به ترتیب احتمالات گزارش موردانتظار از خروجی‌های مختلف داده‌های واقعی و تولید شده هستند.
- θ^D و θ^G: به ترتیب وزن های مدل تفکیک‌کننده و مولد هستند.

عبارتی که برای شبکه کامل متشکل از متمایزکننده و مولد در نظر گرفته می‌شود با V نشان داده می‌شود و به‌صورت رابطه ۱، تعریف می‌شود:

$$V\left(\theta^D,\theta^G\right)=E_{x\sim Px}\left[\log D(x)\right]+E_{z\sim P_z}\left[\log\left(1-D\left(G(z)\right)\right)\right] \quad (۱)$$

[1] Stop Words Removal
[2] Lemmatization
[3] Generative adversarial networks (GANs)
[4] Generator(G)
[5] Discriminator(D)

این تابع مقدار باهدف به حداکثر رساندن تلفات تفکیک‌کننده و به‌حداقل‌رساندن تلفات مولد به یک استراتژی حداقل - حداکثر ارسال می‌شود که در رابطه ۲، ارایه می‌شود:

$$\min_{\theta^G} \max_{\theta^D} V\left(\theta^D, \theta^G\right) \quad (۲)$$

در واقع‌بین مولد و متمایز گر یک بازی ماکزیمم و کمینه اجرا می‌شود و هدف مولد فریب‌دادن متمایز گر است به‌گونه‌ای که داده‌های باکیفیت و مصنوعی شبیه داده‌های واقعی تولید کند و متمایز گر فرض کند این داده‌ها واقعی است.

۳-۳- استخراج ویژگی

در روش پیشنهادی از سه تکنیک GloVe، Word2Vec و TF-IDF به طور هم‌زمان استفاده می‌شود و ویژگی‌های هر کدام از این روش‌ها در مجموعه F1، F2 و F3 قرار داده می‌شود و در نهایت اجتماع این مجموعه‌ها به‌عنوان ورودی فاز انتخاب ویژگی در نظر گرفته می‌شود. TF-IDF یک روش مهندسی ویژگی است که برای استخراج ویژگی‌ها از داده‌های متنی استفاده می‌شود. این روش در زمینه تحلیل متن بسیار محبوب است. در TF-IDF، به هر عبارت در سند یک نمایش عددی اختصاص داده می‌شود که بر اساس وزن بر اساس هر دو ویژگی فرکانس عبارت (TF) و فرکانس سند معکوس (IDF) تعیین می‌شود. 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 به دست آیند. فرکانس مدت (TF) اغلب برای تعیین وزن یک اصطلاح استفاده می‌شود. رابطه ۳ نحوه محاسبه TF را نشان می‌دهد[۲۹].

$$TF_{t,d} = \frac{n_{t,d}}{\sum_k TF_{t,d}} \quad (۳)$$

فرکانس اصطلاح یا $TF_{t,d}$ با پخش‌کردن تعداد کل یک عبارت t خاص در یک سند $d(n_{t,d})$ با تعداد کل عبارت‌های سند k $\sum TF_{t,d}$ تعیین می‌شود. این رابطه فراوانی عبارت را نسبت به تعداد کلی‌ترم در سند نشان می‌دهد. رابطه بالا روند تعیین TF را روشن می‌کند، جایی که به هر عبارت در سند یک نمایش عددی اختصاص داده می‌شود. در مقابل، IDF مقدار بازنمایی را برای عباراتی که در مجموعه غیرمعمول هستند، محاسبه می‌کند. این بدان معناست که وقتی کلمات نادر یا غیرمعمول در یک یا چند سند ظاهر می‌شوند، این کلمات حاوی اطلاعات مهم و معنی‌داری هستند. رابطه ۴ محاسبه وزن IDF را نشان می‌دهد[۲۹].

$$IDF\left(t\right) = \log\left(\frac{N}{dft}\right) \quad (۴)$$

فرکانس معکوس سند (IDF) برای شناسایی امتیاز تعداد کل اسناد (N) و (ft) محاسبه می‌شود که با تعداد اسناد حاوی عبارت نشان داده می‌شود. روش تعبیه کلمات Word2Vec نیز یکی از روشهای استخراج ویژگی است که در روش پیشنهادی استفاده می‌شود. 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 می‌شود. فراوانی ترم (TF) معیاری است که تعداد دفعات ظاهر شدن یک عبارت در یک سند را کمیت می‌کند. 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 می‌شود. این رویکرد به کاهش وزن عباراتی که اغلب در مجموعه مقالات ظاهر می‌شوند کمک می‌کند. به طور کلی، TF-IDF، که اساساً ضرب امتیازهای TF و IDF است، برای شناسایی نیازهای مربوط به یک متن استفاده می‌شود تا مهم‌ترین و آموزنده‌ترین کلمات به راحتی پیدا شوند[۲۹].
Word2Vecروشی برای بازآفرینی زمینه‌های زبانی کلمه است. این روش دارای یک شبکه عصبی با دو لایه است. مجموعه وسیعی از کلمات به عنوان ورودی استفاده می‌شود و نتیجه یک فضای برداری با صدها بعد است. یک فضای برداری منطبق به هر کلمه منحصر به فرد در پیکره اختصاص داده می‌شود. 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 می‌گیرند. Word2Vec یک روش محاسباتی سریع برای یادگیری جاسازی کلمات از متن خام است. Word2vec از دو روش جداگانه استفاده می‌کند[۲۰]:

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

GloVe[1] یک تکنیک یادگیری بدون نظارت است که می‌تواند در استخراج ویژگی استفاده شود. استنفورد GloVe را ایجاد کرد تا با تجمیع ماتریس هم‌روی کلمه‌به‌کلمه، جاسازی کلمه را بسازد. نتیجه جاسازی در فضای برداری، زیرساخت‌های خطی جذاب کلمه را نشان می‌دهد. مدل GloVe روشی مؤثر برای استخراج ویژگی از پیکره سراسری متن است. هدف اصلی GloVe بردارسازی کلمات و خروجی بردارهای کلمه از طریق پیکره ورودی متن است. روش پیاده‌سازی آن به این صورت است که در ابتدا یک ماتریس هم‌زمانی کلمه بر اساس کل پیکره ایجاد می‌شود و در مرحله بعد، بردار کلمه یادگیری باتوجه‌به ماتریس هم‌زمان و مدل GloVe پردازش می‌شود. مدل GloVe را می‌توان با رابطه ۵، توصیف کرد[۳۰]:

$$J = \sum_{i,j}^{N} f\left(X_{ij}\right)\left(V_i^T V_j + b_i + b_j - \ln\left(X_{ij}\right)\right)^2 \quad (۵)$$

۳-۴- انتخاب ویژگی

در فاز استخراج ویژگی ۴۲ ویژگی مختلف برای تشخیص زورگویی سایبری استخراج شده و تحویل مرحله انتخاب می‌شود. برخی از این ویژگی‌ها، ویژگی‌های زبانی و الفاظ زشت، برخی از ویژگی‌ها مرتبط با جنسیت، برخی از ویژگی‌ها مرتبط با نژاد، برخی از ویژگی‌ها مرتبط با لینک و تصاویر ارسال شده است. در روش پیشنهادی بعد از فاز استخراج ویژگی توسط سه روش استخراج ویژگی در نهایت ویژگی‌ها در اختیار الگوریتم عروس دریایی قرار گرفته می‌شوند. 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 رابطه۶، ایجاد می‌شود:

$$X = \begin{bmatrix} X_{1,1} & X_{2,1} & \cdots & \cdots & X_{1,\dim} \\ X_{2,1} & X_{2,2} & \cdots & \cdots & X_{2,\dim} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{N,1} & X_{N,2} & \cdots & \cdots & X_{N,\dim} \end{bmatrix} \quad (۶)$$

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 است که مطابق رابطه۷، اگر هر ویژگی برابر صفر باشد آنگاه ویژگی مورد نظر در یادگیری استفاده نمود و اگر برابر یک باشد آن ویژگی در یادگیری استفاده می‌شود:

$$F_i = \begin{cases} 1 \;\; Selected \;\; feature \\ 0 Unselected \;\; feature \end{cases} \quad (۷)$$

در الگوریتم عروس دریایی، عروس‌های دریایی به‌عنوان راه‌حل در نظر گرفته می‌شوند و دارای دو رفتار و جستجوی اصلی به شرح ذیل هستند:

- 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 عروس‌های دریایی می‌توان از متغیر C(t) استفاده نمود که در رابطه ۸ فرموله شده است[۱۷].

$$c(t) = \left|(1 - \frac{t}{Maxt}) \times (2.rand - 1)\right| \quad (۸)$$

دراین‌رابطه، t شمارنده تکرار الگوریتم عروس دریایی است و T بیشترین شمارنده تکرار الگوریتم است. rand یک عدد تصادفی بین صفر و یک است. برای به‌روزرسانی هر عروس دریایی یا بردار ویژگی در ابتدا C(t) محاسبه می‌شود و اگر بیشتر از ۰/۵ باشد آنگاه نوع جستجو طبق امواج آب و بر اساس رابطه۹، انجام می‌شود:

$$X_i(t+1) = X_i(t) + rand(0,1) \times (X^* - \beta \times rand(0,1) \times \mu) \quad (۹)$$

دراین‌رابطه، $X_i(t)$ 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 می‌شود. بهینه‌ترین راه‌حل یا بردار ویژگی بهینه با X^* نمایش

[1] Global Vector for Word Representation

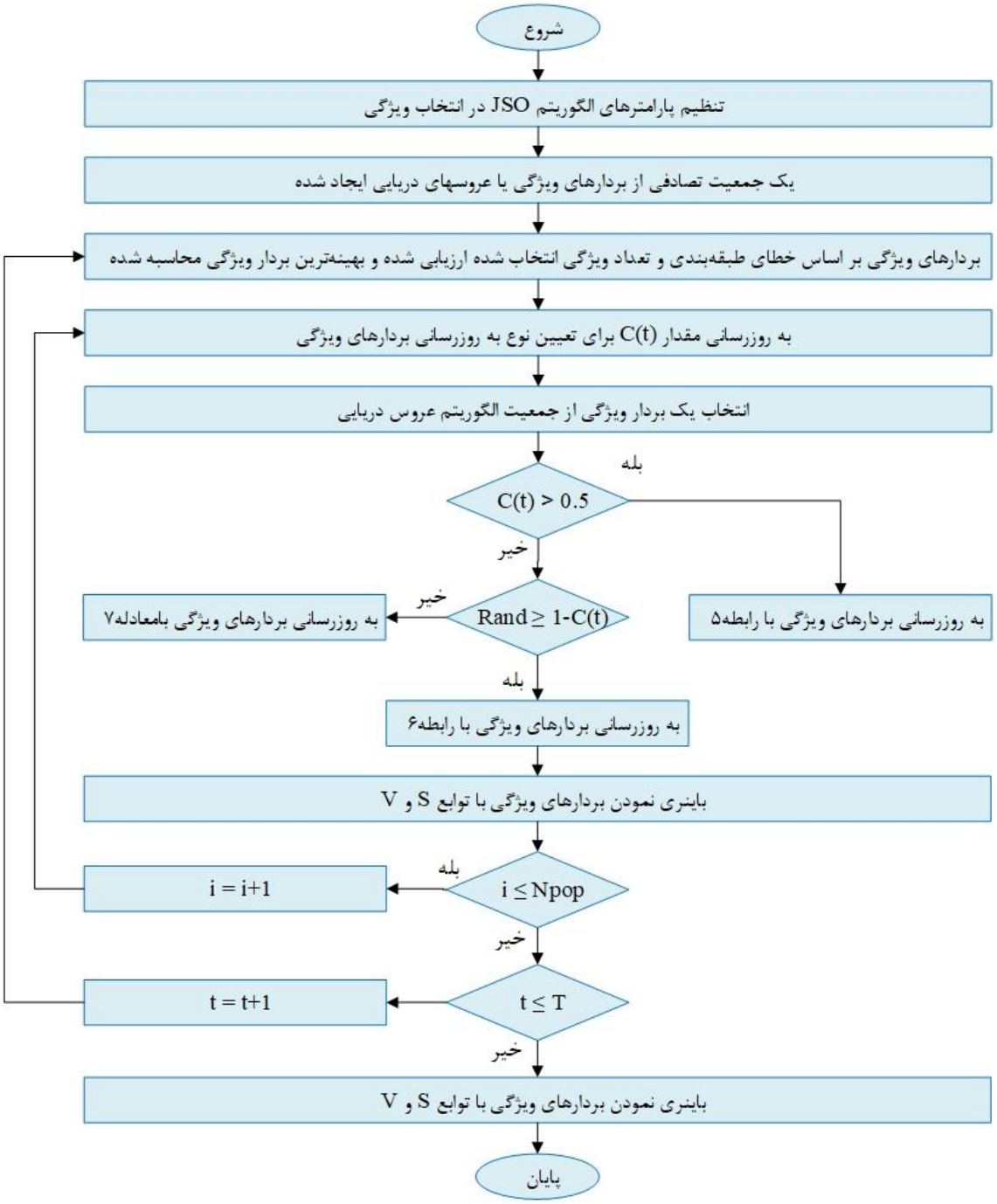
داده می‌شود. اما در مقابل اگر C(t) از ۰/۵ بیشتر نباشد آنگاه رفتارهای هوش گروهی باعث به‌روزرسانی بردارهای ویژگی می‌شود. برای این منظور یک عدد تصادفی بین صفر و یک برای هر بردار ویژگی تولید می‌شود که اگر عدد تصادفی موردنظر از ((1-c(t)) بیشتر باشد آنگاه جستجوی غیرفعال برای به‌روزرسانی بردارهای ویژگی مطابق رابطه ۱۰ انجام می‌شود و در غیر این صورت جستجوی فعال مطابق رابطه ۱۱، بکار گرفته می‌شود.

$$X_i(t+1) = X_i(t) + \gamma.rand(0,1) \times (U_b - L_b) \quad (۱۰)$$

$$X_i(t+1) = X_i(t) + rand.(X_j(t) - X_i(t)) \quad (۱۱)$$

دراین‌رابطه، U_b و L_b محدوده بالا و پایین بردارهای ویژگی است و γ ضریب همگرایی در جستجوی غیرفعال عروس‌های دریایی است.

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



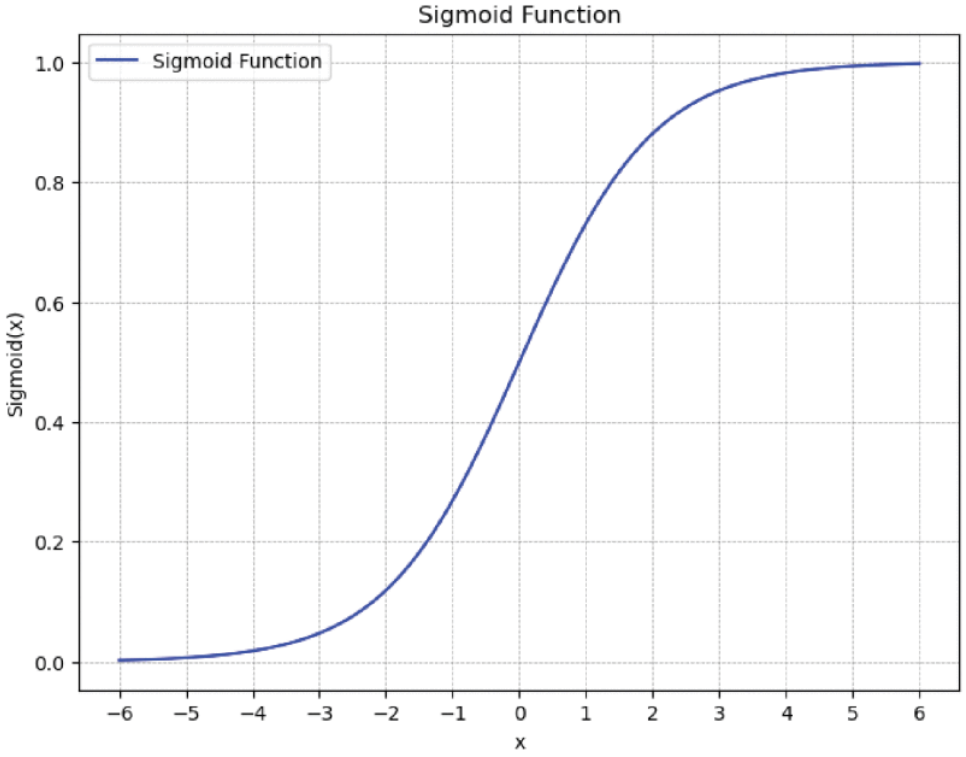
شکل ۵: فلوچارت انتخاب ویژگی توسط الگوریتم عروس دریایی
Figure 5: Flowchart of feature selection using the Jellyfish algorithm

تابع هدف در فاز انتخاب برای ارزیابی بردارهای ویژگی در رابطه ۱۲، فرموله شده است:

$$Cost(X_i) = w_1.E + w_2 \frac{size(X_i)}{D} \quad (۱۲)$$

دراین‌رابطه، E خطای تشخیص طبقه‌بندی توییت‌ها به دودسته زورگویی سایبری و عادی به‌ازای بردار ویژگی i-ام یا X_i است. در اینجا از یک شبکه عصبی چندلایه برای ارزیابی بردارهای ویژگی استفاده می‌شود. 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 می‌شوند و مجموع آنها برابر یک است. کمینه نمودن این تابع هدف نشان‌دهنده شایستگی بیشتر یک بردار ویژگی است.

مقادیر بردارهای ویژگی باید صفر و یک باشند؛ اما در حین فرایند به‌روزرسانی بردارهای ویژگی توسط الگوریتم عروس دریایی این مقادیر حالت صفر و یک خود را ازدست‌داده و می‌توانند اعشاری شوند. بعد از هر به‌روزرسانی بردارهای ویژگی می‌توان آنها را مجدد باینری نمود. برای این منظور ابتدا توسط تابع S یا V ، مقادیر بردارهای ویژگی بین صفر و یک نرمال می‌شود. در شکل (۶)، تابع سیگموئید یا S برای نرمالسازی مقادیر ویژگی‌های مجموعه داده نمایش داده می‌شود. مشاهده می‌شود 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 سیگموییدد برای نرمالسازی مقادیر ویژگی‌ها بین صفر و یک
Figure 6: The sigmoid (S) function for normalizing feature values between zero and one

در روابط ۱۳ الی ۲۰ به ترتیب چهار نوع تابع سیگموید یا S و چهار تابع تانژانت سیگموید یا V نمایش داده شده است[۳۴].

$$S_1 = \frac{1}{1+e^{-2x}} \quad (۱۳)$$

$$S_2 = \frac{1}{1+e^{-x}} \quad (۱۴)$$

$$S_3 = \frac{1}{1+e^{-x/2}} \quad (۱۵)$$

$$S_4 = \frac{1}{1+e^{-x/3}} \quad (۱۶)$$

$$V1 = \left|\tanh(x)\right| \quad (۱۷)$$

$$V2 = \left|crf\left(\frac{\sqrt{\pi}}{2}x\right)\right| = \left|\frac{\sqrt{\pi}}{2}\int_0^{\frac{\sqrt{\pi}}{2}} e^{-t2}dt\right| \quad (۱۸)$$

$$V3 = \left|\frac{x}{\sqrt{1+x^2}}\right| \quad (۱۹)$$

$$V3 = \left|\frac{2}{\pi}\arctan\frac{\pi}{2}x\right| \quad (۲۰)$$

بعد از مرحله نرمال‌سازی بردارهای ویژگی توسط توابع S و V می‌توان توسط رابطه‌های ۲۱ و ۲۲ مقدار آنها به صفر و یک تبدیل نمود[۳۴].

$$x_i^j(t+1) = \begin{cases} 0, if\ random < Sfunction\left(x_i^j(t+1)\right) \\ 1, if\ random \geq Sfunction\left(x_i^j(t+1)\right) \end{cases} \quad (۲۱)$$

$$x_i^j(t+1) = \begin{cases} x_i^j(t), \ if\ random < Vf\ unction\left(x_i^j(t+1)\right) \\ x_i^j(t), if\ random \geq Vf\ iction\left(x_i^j(t+1)\right) \end{cases} \quad (۲۲)$$

در این پژوهش از مجموعه‌داده بکار رفته در [۳۲] و [۳۳] مورداستفاده قرار گرفته شده است و ورودی‌های فاز انتخاب ویژگی از این دو مجموعه‌داده تغذیه می‌شود. ویژگی‌های مهمی که روش پیشنهادی آنها را انتخاب می‌کند عبارت‌اند از توهین جنسی، توهین نژادی، باج‌خواهی، الفاظ زشت، انتساب اسم حیوان به اشخاص، تهدید با سلاح سردوگرم، ارسال فایل‌های تحقیرآمیز، ارسال اطلاعات شخصی برای دیگران و غیره است.

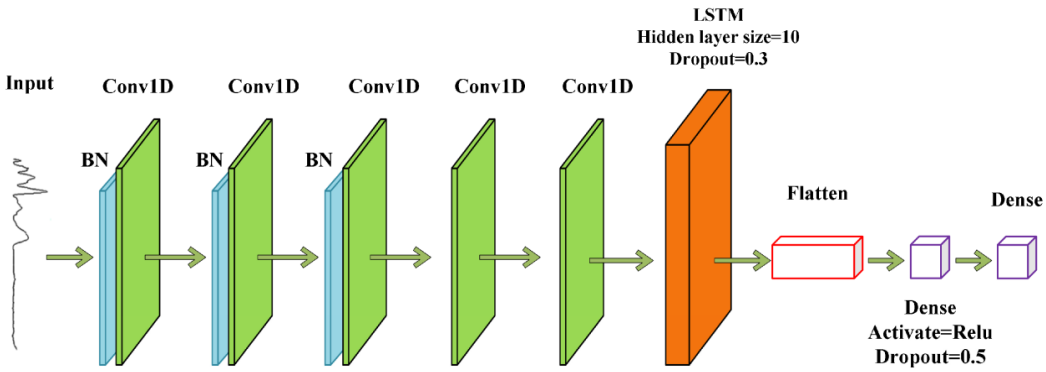
یکی از چالش‌های یادگیری ماشین بیش برازش[1] است که می‌تواند ناشی از پیچیدگی زیاد مدل و آموزش بیش از حد آن باشد. عواملی که باعث بیش برازش می‌شود می‌تواند عدم انتخاب ویژگی، حجم نمونه‌های اندک و عدم پیش‌پردازش مجموعه‌داده باشد. یکی از راهکارهای ما برای مقابله با این چالش به‌کارگیری انتخاب ویژگی با الگوریتم عروس دریایی و از طرفی متعادل‌سازی مجموعه‌داده است و همچنین روش ما برای رفع این چالش پیش‌پردازش و نرمال‌سازی داده‌ها را انجام می‌دهد.

۳- ۵- طبقه بندی با معماری CNN و LSTM

ترکیب CNN و شبکه‌های عصبی مکرر کاربردهای تحقیقاتی متعددی داشته است و اخیراً نتایج قابل‌توجهی در زمینه‌های زیر به همراه داشته است که از جمله آنها می‌توان به زمینه‌پردازش زبان طبیعی، مانند تجزیه‌وتحلیل احساسات و تشخیص گفتار، در پیش‌بینی داده‌های بلادرنگ، مانند پیش‌بینی سهام و پیش‌بینی آلودگی هوا و غیره اشاره نمود[۳۱]. در مدل ترکیبی CNN پارامترها را از طریق به‌اشتراک‌گذاری وزن کاهش می‌دهد تا کارایی یادگیری مدل را بهبود بخشد. علاوه بر این، LSTM یک مدل شبکه تکراری است که مشکلات گرادیان طولانی‌مدت و ناپدیدشدن گرادیان را در RNN حل می‌کند. چارچوب شبکه یادگیری عمیق CNN-LSTM در شکل (۷)، نشان‌داده‌شده 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 طیف‌ها ساخته شدند. به طور خاص، ویژگی‌های داده درون طیفی که توسط CNN آموخته شد به LSTM وارد شد و LSTM برای استخراج اطلاعات موقعیت در داده‌های طیفی استفاده شد. مدل CNN-LSTM می‌تواند به طور کامل ویژگی‌های ذاتی داده‌ها را استخراج کند، و همچنین

[1] Overfitting

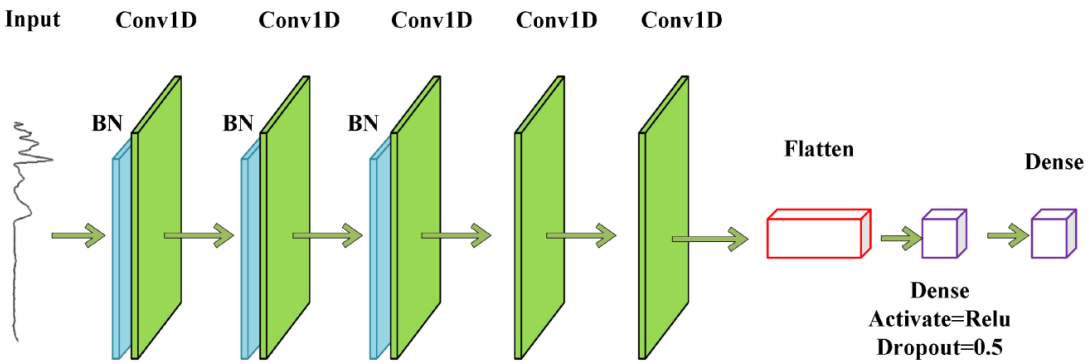
مدل را قادر می‌سازد تا توانایی خاصی برای استخراج اطلاعات داده‌های توالی متوسط و طولانی داشته باشد. این مدل می‌تواند ویژگی‌های پنهان نمونه‌ها را بیاموزد و این اطلاعات را برای شناسایی بهتر نمونه‌های دسته‌های مختلف به دست آورد.



شکل ۷: معماری ترکیبی CNN و LSTM

Figure 7: Hybrid CNN-LSTM architecture

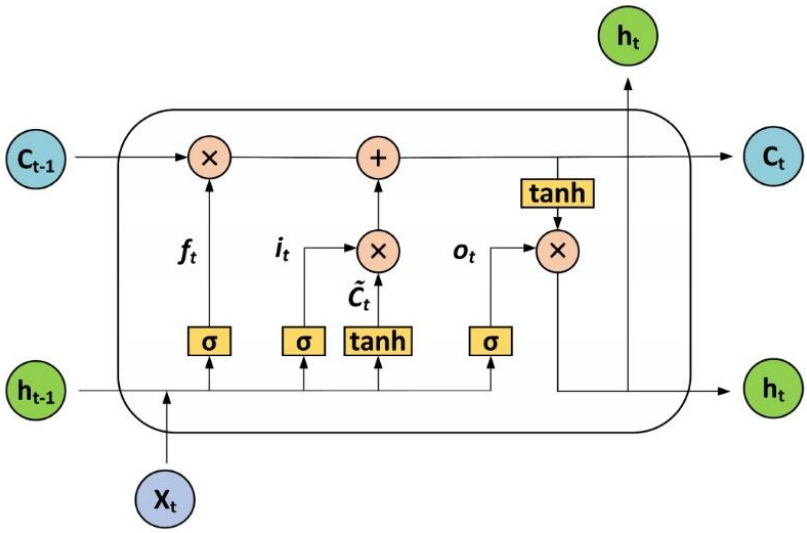
1D-CNN ساخته شده در این مقاله از چارچوب Alexnet استفاده می‌کند، و ساختار مدل 1D-CNN در شکل (۸)، نشان داده شده است. Alexnet یک مدل یادگیری عمیق کلاسیک است که شامل یک تبدیل ۸ لایه با ۵ لایه کانولوشنال و ۲ لایه کاملاً متصل است. و یک لایه خروجی کاملا متصل Alexnet تابع فعال‌سازی واحد خطی اصلاح‌شده (ReLU) 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

Figure 8: One-dimensional convolutional neural network (1D-CNN) architecture

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



شکل ۹: یک واحد از شبکه
Figure 9: One unit of the network

برای هر ورودی xt از داده‌های طیفی، مدل LSTM یک فعال‌سازی پنهان ht ایجاد می‌کند. در مرحله بعد، هر قطعه از داده‌های طیفی وارد شده و بیشتر برای پیش‌بینی استفاده می‌شود. مدل LSTM رابطه تبدیل ht نمایش پنهان را از طریق واحد LSTM تعریف می‌کند که ورودی xt را در حالت فعلی و اطلاعات به‌دست‌آمده ht-1 را می‌پذیرد.

۴-نتایج تجربی

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

۴-۱-پارامترهای پیاده‌سازی

برای پیاده‌سازی یادگیری عمیق از پایتون و کتابخانه Keras و Tensorfolw استفاده می‌شود. در مدل 1D-CNN، اندازه دسته ۳۲، بهینه ساز از نوع Adam انتخاب می‌شود، نرخ یادگیری روی ۰/۰۰۱ تنظیم شده است. اندازه جمعیت بردارهای ویژگی برابر ۱۰ و تعداد تکرار الگوریتم انتخاب ویژگی نیز برابر ۳۰ تنظیم می‌شود. نوع اعتبارسنجی روش پیشنهادی از نوع متقاطع است و ۷۰ رصد از داده‌ها آموزشی و مابقی نیز به نسبت ۱۵ درصد و ۱۵درصد به ترتیب داده آزمون و اعتبارسنجی است. هر آزمایش نیز ۲۵ مرتبه تکرار می‌شود و متوسط شاخص‌ها در ارزیابی استفاده می‌شود. در روش پیشنهادی برای نرمالسازی از روش خطی ماکزیمم و کمینه استفاده می‌شود و محدوده نرمالسازی بین [0,1] است. در پیاده‌سازی‌ها دو مجموعه داده [۳۳و۳۲] به عنوان ورودی روش پیشنهادی در نظر گرفته شده و پیش پردازش مجموعه داده‌ها با کتابخانه‌های پایتون مانند spacy انجام می‌شود. در ادامه مجموعه داده‌ها فازهای مانند متعادلسازی با شبکه GAN و استخراج ویژگی انجام می‌شود و سپس انتخاب ویژگی با الگوریتم عروس دریایی انجام می‌شود. طبقه‌بندی نهایی مدل روی ویژگی‌های خروجی الگوریتم عروس دریایی توسط CNN و LSTM انجام می‌شود. هر آزمایش در روش پیشنهادی ۲۵ مرتبه تکرار شده و متوسط شاخص‌های ارزیابی مانند دقت محاسبه و با روشهای مشابه مقایسه می‌شود.

جدول ۲: پارامترهای مدل یادگیری عمیق
Table 2: Deep Learning Model Parameters

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 مجموعه‌داده برای آزمایش و اعتبارسنجی عملکرد مدل‌های پیشنهادی استفاده می‌کند. اولین مجموعه‌داده معیار توییتر (که در حال حاضر به‌عنوان X شناخته می‌شود) در [۳۲] برای شناسایی آزار و اذیت سایبری با شناسایی توییت‌های توهین‌آمیز و غیرتوهین‌آمیز استفاده شده است و این مجموعه‌داده با اندازه ۳۷۳۷۳ توییت است. داده‌ها به‌صورت عددی با ۱ یا ۰ برچسب‌گذاری شدند که در آن ۱ نشان‌دهنده توییت توهین‌آمیز و ۰ نشان‌دهنده توییت خنثی است، به این معنی که توییت به دسته توهین‌آمیز تعلق ندارد. جدول (۳)، نمونه‌ای از نمونه‌های توییت‌های آزار و اذیت سایبری در مجموعه‌داده‌های توییتر را نشان می‌دهد.

جدول ۳: چند نمونه از توئیتهای عادی و زورگویی سایبری

Table 3: Several Examples of Normal and Cyberbullying Tweets

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

مجموعه‌داده دوم مورداستفاده در این مطالعه [۳۳] شامل داده‌های جمع‌آوری‌شده از گروه‌های توییتر و فیس‌بوک است که بر فعالیت‌های مشکوک مانند نژادپرستی، تبعیض، زبان توهین‌آمیز و تهدید، که عمدتاً با حوادث آزار و اذیت سایبری مرتبط است، تمرکز دارد. داده‌های موجود در مجموعه داده بر اساس وجود کلمات مشکوک مورد استفاده در توییت ها و نظرات حاشیه نویسی شدند. نقاط داده مشکوک به‌صورت 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 ویژگی‌هایی مانند نژادپرستی، تبعیض و سوء استفاده است. برعکس، هشت هزار نداده با احساسات مثبت یا خنثی برچسب‌گذاری شدند، که نشان می‌دهد داده‌ها ویژگی‌های غیر مشکوک را نشان می‌دهند. اطلاعات ورودی هر دو مجموعه داده اساساً بر اساس توییت‌ها و نظرات به زبان انگلیسی است، با برخی از روش‌های پیش پردازش و پاکسازی داده‌ها در زیر بخش زیر توضیح داده شده است.

۴-۳-متریکهای ارزیابی

برای ارزیابی روش پیشنهادی در تشخیص زورگویی سایبری از متریک‌های نظیر دقت[1]، حساسیت[2] و صحت[3] استفاده می‌شود که ضابطه آنها به ترتیب در رابطه‌های ۲۳، ۲۴، ۲۵ ارایه شده است.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (۲۳)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \quad (۲۴)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (۲۵)$$

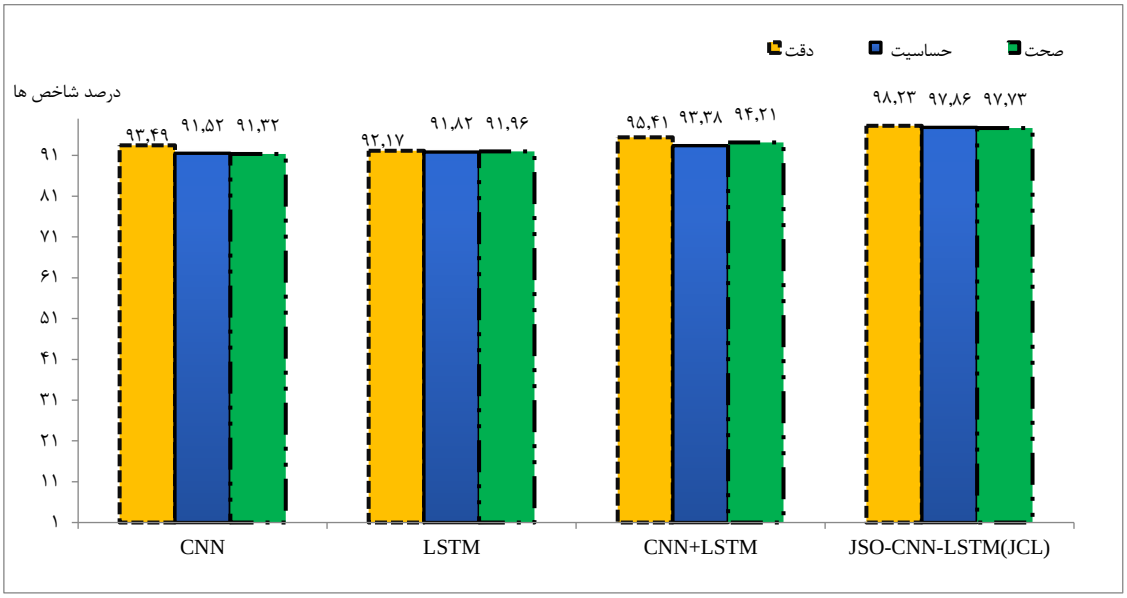
هر کدام از شاخص‌های ارزیابی TP، TN، FP و FN را برای تحلیل زورگویی سایبری می‌توان به‌صورت ذیل تعریف نمود:

[1] Accuracy
[2] Sensitivity
[3] Precision

- نمونه‌های صحیح مثبت[1](TP): توییت موردنظر از نوع زورگویی سایبری است و روش پیشنهادی آن را در دسته زورگویی سایبری قرار داده است.
- نمونه‌های غلط مثبت[2](FP): توییت موردنظر از نوع عادی است و روش پیشنهادی آن را در دسته زورگویی سایبری قرار داده است.
- نمونه‌های صحیح منفی[3](TN): توییت موردنظر از نوع عادی است و روش پیشنهادی آن را در دسته عادی قرار داده است.
- نمونه‌های غلط منفی[4](FN): توییت موردنظر از نوع زورگویی سایبری است و روش پیشنهادی آن را در دسته عادی قرار داده است.

۴-۳-تحلیل و ارزیابی

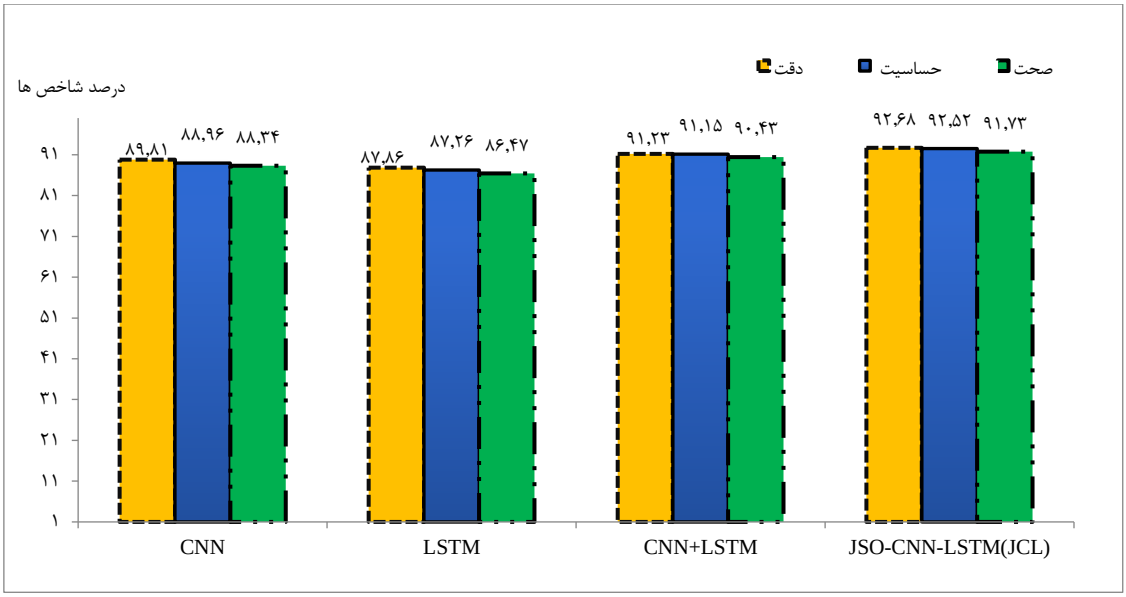
در این بخش روش پیشنهادی برای تشخیص زورگویی سایبری مورد پیاده‌سازی قرار گرفته شده است. در نمودار شکل (۱۰) و (۱۱) شاخص دقت، صحت و حساسیت روش پیشنهادی با حالتهای مختلف LSTM، CNN، CNN-LSTM مورد مقایسه قرار گرفته شده است. روش پیشنهادی در اینجا با JSO-CNN-LSTM تعریف می‌شود که به اختصار JCL در نظر گرفته می‌شود.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



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

Figure 10: Comparison of accuracy, sensitivity, and specificity in cyberbullying detection on a Twitter dataset

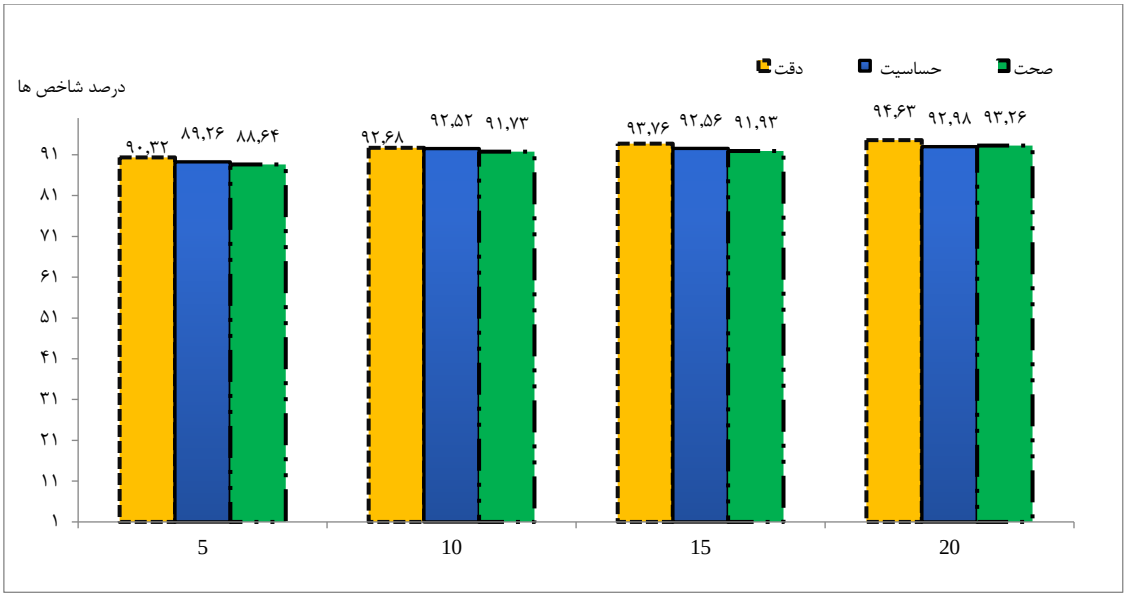
[1] True positive(TP)
[2] False positive(FP)
[3] True negative (TN)
[4] Flase negative (FN)



شکل ۱۱: مقایسه دقت، حساسیت و صحت در تشخیص زورگویی سایبری در مجموعه داده فیس بوک
Figure 11: Comparison of Accuracy, Sensitivity, and Specificity in Cyberbullying Detection on a Facebook Dataset

آزمایش‌ها نشان می‌دهد که اگر از مجموعه‌داده توییتر 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 توئیتر دارای دقت بیشتری نسبت به مجموعه داده فیس بوک در تشخیص زورگویی سایبری است.

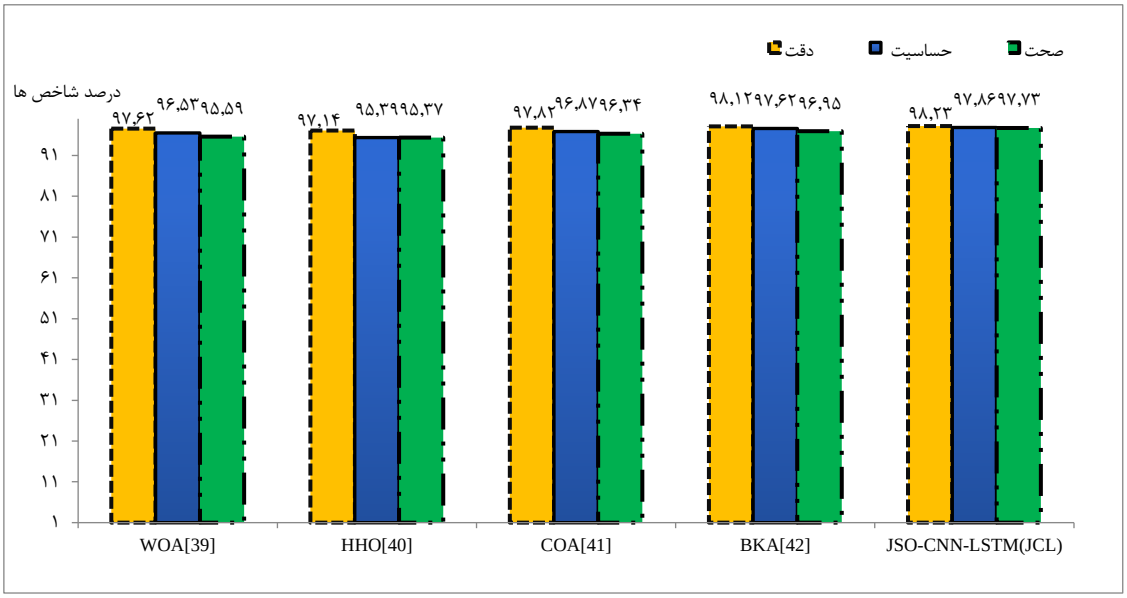
یکی از شاخص‌ها و پارامترهای مهم برای ارزیابی روش پیشنهادی اندازه جمعیت اولیه بردارهای ویژگی است که افزایش آن را می‌توان باعث افزایش توانایی جستجوی روش پیشنهادی در یافتن بردار ویژگی در نظر گرفت. در نمودار شکل (۱۲)، تعداد بردارهای ویژگی در الگوریتم عروس دریایی به عنوان مهمترین پارامتر به عنوان متغیر در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

Figure 12: Comparison of Accuracy, Sensitivity, and Specificity in Cyberbullying Detection with Increasing Feature Vector Size

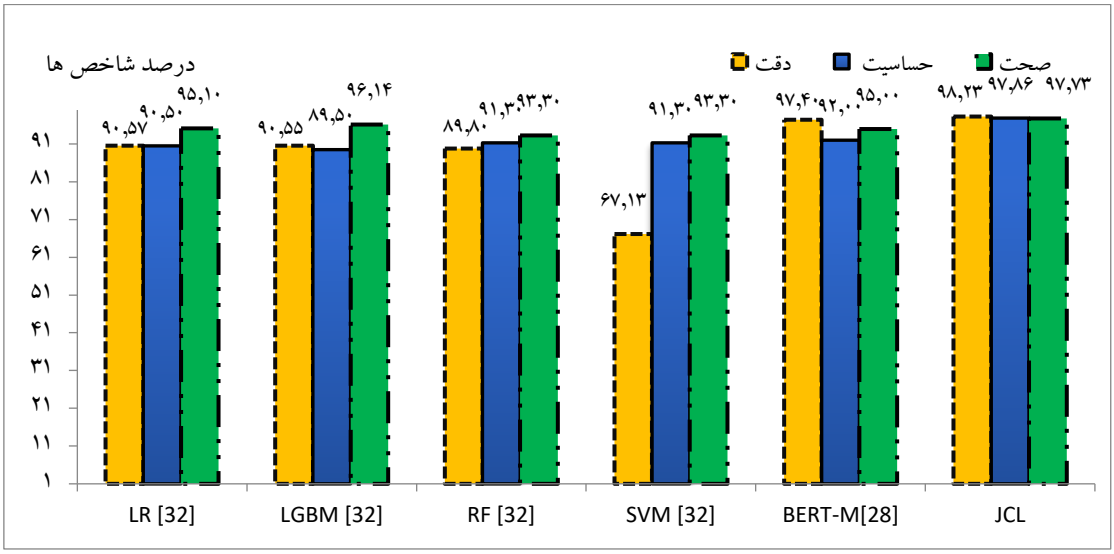
ارزیابی‌ها نشان می‌دهد با افزایش تعداد بردارهای ویژگی توانایی الگوریتم عروس دریایی برای یافتن بردار ویژگی افزایش پیدا می‌کند و دقت، حساسیت و صحت افزایش می‌یابد. اگر تعداد بردارهای ویژگی برابر ۵ باشد دقت، حساسیت و صحت روش پیشنهادی برابر ۹۰/۳۲ درصد، ۸۹/۲۶ درصد و ۸۸/۶۴ درصد است و با افزایش تعداد بردار ویژگی 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

Figure 13: Comparison of Accuracy, Sensitivity, and Specificity in Cyberbullying Detection on a Facebook Dataset Using Feature Selection Methods

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 کواتی الگوریتم بهینه‌سازی عقاب بال‌سیاه مقایسه و ارزیابی شده است. آزمایش‌ها 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 به‌ازای تعداد بردار ویژگی برابر ۱۰ در حدود ۴.۸۷ ثانیه است و زمان اجرای الگوریتم بهینه‌سازی وال، الگوریتم بهینه‌سازی شاهین، الگوریتم بهینه‌سازی کواتی و الگوریتم بهینه‌سازی عقاب بال‌سیاه 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 بهینه‌سازی شاهین، الگوریتم بهینه‌سازی کواتی و الگوریتم بهینه‌سازی عقاب بال‌سیاه کمتر است. در نمودار شکل (۱۴)، روش پیشنهادی در شاخص‌های سه‌گانه 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 توئیتر با مطالعات مرتبط

Figure 14: Comparison of Accuracy, Sensitivity, and Specificity in Cyberbullying Detection on a Twitter Dataset with Related Studies

آزمایش‌ها و ارزیابی نشان می‌دهد 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. آزمایش‌ها و مقایسه‌ها نشان می‌دهد 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 توییتر نیز مورد مقایسه قرار گرفته شده است.

جدول ۴: مقایسه روش پیشنهادی در تشخیص توئیتهای عادی و زورگویی سایبری در مجموعه داده توئیتر

Table 4: “Comparison of the Proposed Method in Detecting Normal and Cyberbullying Tweets in the Twitter Dataset”

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(Propsed Method)	98.23	97.86	97.73

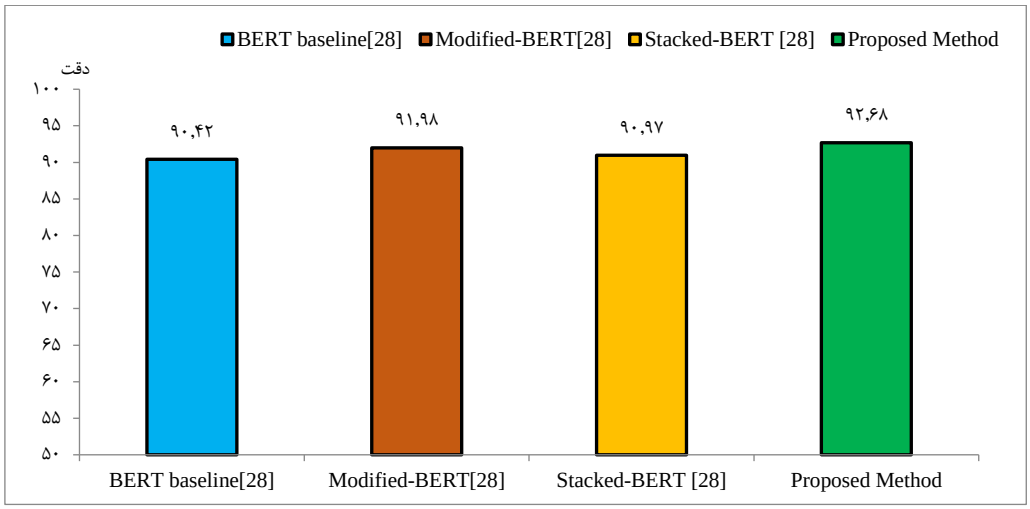
شاخص دقت، حساسیت و صحت روش پیشنهادی در تشخیص زورگویی اجتماعی از روش‌های یادگیری عمیق در مجموعه‌داده توییتر از LSTM، Conv1DLSTM، CNN، BiLSTM، BERT، Tuned-BERT و Stacked بیشتر است. ارزیابی‌ها نشان می‌دهد

که دقت روش پیشنهادی از مدل ارائه شده در [۴۳] بیشتر است؛ اما حساسیت و صحت روش پیشنهادی نسبت به روش ارائه شده در این مطالعه کمتر است.

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

- استخراج ویژگی بر خلاف این روش‌ها توسط سه روش انجام شده و نتایج با هم تلفیق شده
- انتخاب ویژگی با الگوریتم هوش گروهی در روش پیشنهادی انجام شده است.
- روش پیشنهادی از ترکیب دو معماری 1DCNN و LSTM برای طبقه‌بندی دقیق توییت‌ها به دودسته عادی و زورگویی سایبری استفاده می‌کند.

در نمودار شکل (۱۵)، نیز دقت روش پیشنهادی در تشخیص زورگویی سایبری در مجموعه داده فیس‌بوک 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

Figure 15: Comparison of Accuracy: Proposed Method in Cyberbullying Detection on the Facebook Dataset

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 فیس‌بوک بیشتر از این روشها و برابر ۹۲/۸ درصد است.

۵-نتیجه گیری

شیوع آزار و اذیت سایبری در پلتفرم‌های رسانه‌های اجتماعی به یک نگرانی قابل‌توجه برای افراد، سازمان‌ها و جامعه در کل تبدیل شده است. تشخیص زودهنگام و مداخله زورگویی سایبری در رسانه‌های اجتماعی برای کاهش اثرات مضر آن بسیار مهم است. در سال‌های اخیر، یادگیری عمیق نتایج امیدوارکننده‌ای برای تشخیص آزار سایبری در رسانه‌های اجتماعی نشان داده است. روش پیشنهادی برای تشخیص توییت‌های زورگویی سایبری از یک روش ترکیبی سه‌مرحله‌ای استفاده می‌کند. در مرحله اول روش پیشنهادی با سه روش ترکیبی GloVe، Word2Vec و TF-IDF ویژگی‌های متن را استخراج نموده و در مرحله دوم با استفاده از الگوریتم عروس دریایی ویژگی‌های مهم را انتخاب و تحویل طبقه‌بندی‌کننده یادگیری عمیق می‌نماید. در مرحله سوم از طبقه‌بندی‌کننده 1DCNN+LSTM در تشخیص توییت‌های زورگویی سایبری استفاده می‌شود. آزمایش‌ها نشان داد روش پیشنهادی در تشخیص زورگویی سایبری از روش‌های مانند 1DCNN، CNN، LSTM، 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 گروهی و معماری‌های مختلف یادگیری

عمیق و استخراج ویژگی با استفاده از سه روش موازی است و این موضوع باعث می‌شود تا ویژگی‌های مختلف متن در نظر گرفته شود. از محدودیت‌های روش پیشنهادی، پیچیدگی بیشتر آن نسبت به روش‌های مانند CNN و LSTM است. دلایل ذیل را می‌توان برای برتری مدل پیشنهادی برای نسبت به روش‌های مشابه ارائه داد:

- متعادل‌سازی مجموعه‌داده 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 چندبعدی شده است.
- تلفیق توانایی شبکه عصبی CNN و LSTM باعث افزایش دقت در طبقه‌بندی توییت‌ها می‌شود.

در پژوهش آتی تلاش می‌شود تا مجموعه‌داده با روش‌های نظیر GAN متعادل‌سازی شود و به‌جای 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. Nurdiati, (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. Sountharrajan, 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.