مقاله پژوهشی

تشخیص اسپم در شبکه اجتماعی توییتر با استفاده از رویکرد یادگیری ترکیبی

مریم فصیحی، محمدجواد شایگان فرد، زهرا سادات حسینی مقدم و زهرا سجده

چکیده: امروزه شبکه‌های اجتماعی، نقش مهمی در گسترش اطلاعات در سراسر جهان دارند. توییتر یکی از محبوب‌ترین شبکه‌های اجتماعی است که در هر روز ۵۰۰ میلیون توییت در این شبکه ارسال می‌شود. محبوبیت این شبکه در میان کاربران منجر شده تا اسپمرها از این شبکه برای انتشار پست‌های هرزنامه استفاده کنند. در این مقاله برای شناسایی اسپم در سطح توییت از ترکیبی از روش‌های یادگیری ماشین استفاده شده است. روش پیشنهادی، چارچوبی مبتنی بر استخراج ویژگی است که در دو مرحله انجام می‌شود. در مرحله اول از Stacked Autoencoder برای استخراج ویژگی‌ها استفاده شده و در مرحله دوم، ویژگی‌های مستخرج از آخرین لایه Stacked Autoencoder به‌عنوان ورودی به لایه softmax داده می‌شوند تا این لایه پیش‌بینی را انجام دهد. روش پیشنهادی با برخی روش‌های مشهور روی پیکره متنی Twitter Spam Detection با معیارهای Accuracy، F۱-Score، Precision و Recall مورد مقایسه و ارزیابی قرار گرفته است. نتایج تحقیق نشان می‌دهند که دقت کشف روش پیشنهادی به ۷۸/۱٪ می‌رسد. در مجموع، این روش با استفاده از رویکرد اکثریت آرا با انتخاب سخت در یادگیری ترکیبی، توییت‌های اسپم را با دقت بالاتری نسبت به روش‌های CNN، LSTM و SCCL تشخیص می‌دهد.

کلیدواژه: **توییتر، شناسایی اسپم، شبکه عصبی، Autoencoder، Softmax.**

۱- مقدمه

شبکه‌های اجتماعی برخط، ابزار همکاری و ارتباطی برای میلیون‌ها کاربر اینترنتی هستند. کاربران در شبکه‌های اجتماعی، نظرهای خود را بیان و در مورد موضوعات مختلف بحث می‌کنند [۱]. توییتر به‌عنوان یکی از اصلی‌ترین شبکه‌های اجتماعی برخط با ارائه خدمات میکروبلاگینگ رایگان برای کاربران، کاربران را به انتشار یا مشاهده پیام‌هایی با طول حداکثر ۲۸۰ کاراکتر، دنبال‌کردن سایر کاربران و ... از طریق دستگاه‌های مختلفی از جمله تلفن‌های همراه و کامپیوترهای رومیزی جذب می‌کند. توییتر ۱/۳ میلیارد کاربر و ۳۶۰ میلیون کاربر فعال ماهانه دارد که ۵۰۰ میلیون توییت در هر روز ارسال می‌کنند. از ویژگی‌های دیگرش می‌توان به پشتیبانی از بیش از ۵۰ زبان اشاره کرد [۱] و [۲]. امکانات و اهمیت بالای این شبکه، عاملی تحریک‌کننده برای برخی افراد سودجوست تا با انجام فعالیت‌های مخرب از طریق ارسال پیام‌های جعلی، مخرب و اسپم[1]، اهداف خود را دنبال کنند. از این رو شبکه اجتماعی توییتر به بستری مناسب و محبوب برای تبلیغ‌کنندگان و اسپمرها[2] برای انتشار پیام‌های آنها تبدیل شده است [۳].

به‌طور کلی روش‌های تشخیص اسپم به سه دسته اصلی تشخیص مبتنی بر تجزیه و تحلیل نحو[3]، تشخیص مبتنی بر تجزیه و تحلیل ویژگی[4] و لیست سیاه[5] تقسیم می‌شوند [۴]:

۱) **تشخیص مبتنی بر تجزیه و تحلیل نحو:** روش‌های مورد استفاده برای تشخیص اسپم با توجه به متن توییت و بخش‌های کلیدی مانند پیوندها[6] انجام می‌شود.

۲) **تشخیص مبتنی بر تجزیه و تحلیل ویژگی:** روش‌های تشخیص در این بخش به دو قسمت تشخیص اسپم با استفاده از اطلاعات آماری و گراف اجتماعی[7] تقسیم می‌شوند. برای تشخیص اسپم با استفاده از روش اطلاعات آماری، ویژگی‌های آماری از پروفایل کاربران و توییت‌های آن‌ها استخراج می‌شود. برای تشخیص اسپم با روش گراف اجتماعی به روابط دنبال‌کنندگان[8] و دنبال‌شوندگان[9] کاربران توجه می‌شود..

۳) **لیست سیاه:** در این روش، مجموعه‌ای از پیوندها که قبلاً در توییت‌های اسپم استفاده شده‌اند و جزء پیوندهای مخرب هستند، جمع‌آوری شده و روش‌های مورد استفاده با توجه به پیوندهای موجود در لیست به تشخیص اسپم می‌پردازند.

اکثر روش‌های ارائه‌شده برای جلوگیری از فعالیت اسپمرها و فیلترکردن هرزنامه در توییتر مبتنی بر شناسایی اسپمرها و مسدودکردن آنهاست. این روش‌ها چندان قدرتمند نبوده و نیازمند به روش‌هایی قوی‌تر برای شناسایی اسپم در سطح توییت هستند. هرچند پژوهش‌های مختلفی در این حوزه انجام شده، ولی امکان بهبود دقت مدل با بهره‌گیری از یادگیری ترکیبی[10] وجود دارد. به همین دلیل در این مقاله برای شناسایی

این مقاله در تاریخ ۱۳ شهریور ماه ۱۴۰۱ دریافت و در تاریخ ۱۵ مرداد ماه ۱۴۰۲ بازنگری شد.

مریم فصیحی، گروه مهندسی کامپیوتر، دانشگاه علم و فرهنگ، تهران، ایران، (email: fasihi.star@gmail.com).

محمدجواد شایگان فرد (نویسنده مسئول)، گروه مهندسی کامپیوتر، دانشگاه علم و فرهنگ، تهران، ایران، (email: shayegan@usc.ac.ir).

زهرا سادات حسینی مقدم، گروه مهندسی کامپیوتر، دانشگاه علم و فرهنگ، تهران، ایران، (email: hosayni3502@gmail.com).

زهرا سجده، گروه مهندسی کامپیوتر، دانشگاه علم و فرهنگ، تهران، ایران، (email: sajdehzahra@gmail.com).

1. Spam
2. Spammer
3. Detection Based on Syntax Analysis
4. Detection Based on Feature Analysis
5. Blacklist
6. URLs
7. Social Graph
8. Followers
9. Following
10. Ensemble Learning

اسپم در سطح توییت از روش‌های یادگیری ترکیبی استفاده شده است. مدل یادگیری ترکیبی یکی از روش‌های یادگیری ماشین است که در آن، مدل‌ها به‌عنوان اجزای سازنده با یکدیگر ترکیب می‌شوند. در این مقاله، چندین مدل متشکل از ترکیب Autoencoder و لایه softmax برای حل مسئله آموزش داده شده و سپس برای به‌دست‌آوردن نتایج بهتر، این مدل‌ها با استفاده از روش Bagging با یکدیگر ترکیب می‌شوند تا بتوانیم مدل‌های دقیق‌تر و یا پایدارتری به‌دست آوریم. در این روش مدل‌های متفاوت، مستقل از یکدیگر و به‌صورت همزمان آموزش می‌بینند و با رویکرد اکثریت آرا با انتخاب سخت با یکدیگر ترکیب می‌شوند. روش پیشنهادی، چارچوبی مبتنی بر استخراج ویژگی است که در دو مرحله انجام می‌شود. در مرحله اول از Stacked Autoencoder برای استخراج ویژگی‌ها استفاده گردیده است و در مرحله دوم، ویژگی‌های استخراج‌شده از آخرین لایه Stacked Autoencoder به‌عنوان ورودی به لایه softmax داده می‌شود تا این لایه، احتمال تعلق هر نمونه به هر یک از برچسب‌های ممکن را محاسبه کند. در این مقاله از مدل‌های مبتنی بر CNN Autoencoder، Sequence_To_Sequence Autoencoder و Sequential Stacked CNN-LSTM استفاده کرده‌ایم. در واقع مهم‌ترین نوآوری این مقاله، ارائه یک روش یادگیری ترکیبی مبتنی بر تکنیک‌های Autoencoder و لایه softmax برای استخراج ویژگی‌ها و پیش‌بینی توییت‌های اسپم است.

در بخش دوم مقاله، کارهای پیشین شرح داده شده و بخش سوم به توضیح رویکرد پیشنهادی اختصاص دارد. در بخش چهارم ارزیابی کارایی رویکرد پیشنهادی گزارش شده و نهایتاً بخش پنجم، نتیجه‌گیری مقاله را ارائه می‌دهد.

۲- کارهای پیشین

برای رفع مشکل اسپم توییتر در چند سال اخیر، روش‌های تشخیص بسیاری ارائه شده است. تشخیص اسپم یک زمینه پژوهشی است که در آن، پژوهشگران به دنبال ارائه چارچوب‌ها و روش‌هایی برای شناسایی خودکار اسپم با استفاده از تجزیه و تحلیل داده‌های موجود هستند. آنها معمولاً برای تشخیص اسپم از روش‌های یادگیری ماشین مانند ماشین بردار پشتیبان [1](SVM) یا یادگیری عمیق مانند شبکه‌های عصبی استفاده استفاده می‌کنند. مقاله‌های [۵] تا [۸] پژوهش‌هایی هستند که از ماشین بردار پشتیبان و [۱] و [۹] تا [۱۵] از روش‌های یادگیری عمیق برای تشخیص اسپم در شبکه اجتماعی توییتر استفاده کرده‌اند.

در [۷]، [۱۶] و [۱۷] برای تشخیص اسپم در توییتر از ویژگی‌های توییت مانند متن توییت، تاریخ انتشار توییت، تعداد هشتگ‌ها و غیره استفاده گردیده است. در [۱۶] از الگوریتم بدون نظارت نایوبیز، در [۱۷] از اطلاعات کاربری و ساختار شبکه به همراه ویژگی‌های توییت و در [۷] از ترکیب ویژگی‌های مبتنی بر محتوای توییت، ویژگی‌های رفتاری کاربر و ویژگی‌های مبتنی بر حروف که با استفاده از ساختار متن توییت‌ها به‌دست می‌آید و استفاده از روش ترکیبی ماشین بردار پشتیبان به همراه الگوریتم بهینه‌سازی نهنگ برای تشخیص اسپم استفاده شده است.

در [۱۲]، [۱۳] و [۱۸] تا [۲۰]، شبکه‌های عصبی به‌منظور تشخیص اسپم بر روی مجموعه داده‌های مختلف به‌کار برده شده‌اند. مقاله‌های [۱۲] و [۱۳] شبکه‌های عصبی CNN، LSTM و CNN_LSTM را بر روی مجموعه داده توییتر، [۱۸] شبکه‌های عصبی CNN و LSTM را بر روی مجموعه داده‌های Ott و Yelp، [۱۹] شبکه عصبی Autoencoder را بر روی مجموعه داده به زبان چینی و همچنین [۲۰] شبکه عصبی LSTM Autoencoder را بر روی مجموعه داده شامل نظرات ثبت‌شده کاربران در پنج ویدئوی معروف در یوتیوب [۲۱] برای تشخیص اسپم استفاده کرده‌اند.

در [۲۲] ارزیابی احساسات در توییت به‌عنوان معیاری برای شناسایی اسپم‌بودن آن مطرح شده است؛ به همین منظور از الگوریتم‌های مختلف طبقه‌بندی برای تشخیص اسپم و روش‌های یادگیری ماشین و یادگیری عمیق برای تحلیل احساسات توییت استفاده گردیده است. در [۲۳] و [۲۴] روش طبقه‌بندی ترکیبی مطرح شده که در [۲۳] روش طبقه‌بندی ترکیبی هوشمند با پارامترهای بهینه‌سازی‌شده توسط الگوریتم تکاملی و در [۲۴] طبقه‌بندی ترکیبی مبتنی بر روش‌های یادگیری عمیق که هم شامل متن توییت و هم اطلاعات کاربران است، برای شناسایی اسپم معرفی شده است. در [۱]، [۲۵] و [۲۶] روشی ترکیبی برای تشخیص اسپم مطرح شده است. در [۱] از شبکه‌های عصبی چند لایه و الگوریتم جنگل تصادفی، در [۲۵] از روش اکثریت آرا و در [۲۶] از برای شبکه عصبی عمیق DNN برای روش ترکیبی، استفاده شده است.

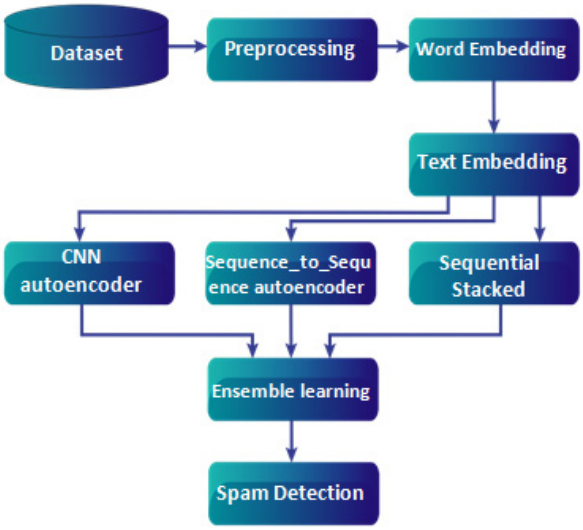
۳- رویکرد ترکیبی پیشنهادی

مدل‌ها در روش‌های یادگیری ماشین ترکیبی به‌عنوان اجزای سازنده با یکدیگر ترکیب می‌شوند و سه روش کلی Bagging، Boosting و Stacking برای ترکیب آنها وجود دارد. در این مقاله از روش Bagging استفاده شده که هدف آن، ایجاد مدلی ترکیبی است که پایدارتر از مدل‌های سازنده خود باشد. در این روش مدل‌های متفاوت به‌صورت مستقل از یکدیگر و هم‌زمان آموزش می‌بینند و با فرایند اکثریت آرا با یکدیگر ترکیب می‌شوند. برای تشخیص توییت‌های اسپم از مدل‌های مختلف یادگیری عمیق مبتنی بر شبکه‌های عصبی Autoencoder مانند CNN Autoencoder، Sequence_to_Sequence Autoencoder و Sequential Stacked CNN-LSTM در این مقاله استفاده شده است.

CNN Autoencoder: شبکه عصبی کانولوشن، نوع خاصی از شبکه عصبی با چندین لایه است. CNN داده‌هایی را که دارای آرایش شبکه‌ای هستند، پردازش می‌کند و سپس به استخراج ویژگی‌های مهم می‌پردازد. یک مزیت بزرگ استفاده از CNN این است که نیازی به انجام بسیاری از مراحل پردازش روی متن نیست. مستقیماً می‌تواند مهم‌ترین ویژگی در فیلترها را آموزش ببیند که این امر باعث صرفه‌جویی بسیاری در زمان آزمایش و خطا می‌شود. در این مقاله از چندین لایه CNN برای انکدر و همچنین چندین لایه CNN برای دیکدر استفاده شده است.

Sequence_to_Sequence Autoencoder: مدل Seq2Seq نوعی Autoencoder است که در این مقاله برای اجزای انکدر و دیکدر آن از معماری LSTM استفاده شده است. برای پیاده‌سازی Autoencoder داده‌هایی که متوالی هستند می‌توان از LSTM Autoencoder استفاده کرد. شبکه‌های حافظه کوتاه‌مدت طولانی (LSTM)، نوعی شبکه عصبی بازگشتی (RNN) هستند که به‌طور خاص برای پشتیبانی از داده‌های متوالی طراحی شده‌اند؛ قادر به یادگیری داده‌های متوالی پیچیده‌اند و از یک حافظه داخلی برای به‌خاطرآوردن یا استفاده از اطلاعات در داده‌های متوالی بزرگ استفاده می‌کنند. در این مقاله از چندین لایه LSTM برای انکدر و همچنین چندین لایه LSTM برای دیکدر استفاده شده است.

1. Support Vector Machine



شکل ۱: فلوچارت رویکرد پیشنهادی.

Sequential Stacked CNN-LSTM: در [۱۲] و [۱۳] یک مدل Sequential طراحی شده است. معماری این مدل به‌ترتیب شامل لایه‌ای از CNN به‌دلیل منحصربه‌فردبودن این شبکه عصبی در استخراج ویژگی و لایه‌ای از LSTM به‌دلیل یادگیری وابستگی‌هاست. سعی بر آن شد تا با قراردادن این دو معماری در کنار هم، میزان معیارهای ارزیابی در پیش‌بینی بالا برده شود. برای پیاده‌سازی انکدر از چندین لایه CNN، Maxpooling و LSTM و نیز از چندین لایه CNN، Maxpooling و LSTM برای دیکدر استفاده شده است.

از یک رویکرد دومرحله‌ای برای تشخیص اسپم استفاده شده است. در مرحله اول از یکی از مدل‌های یادگیری عمیق مبتنی بر شبکه‌های عصبی Autoencoder برای استخراج ویژگی‌ها استفاده می‌شود. در مرحله دوم برای محاسبه احتمال تعلق هر نمونه به هر یک از برچسب‌های اسپم یا اسپم‌نبودن، لازم است که ویژگی‌های استخراج‌شده توئیت‌ها از آخرین لایه Autoencoder به‌عنوان ورودی به لایه softmax داده شود تا این لایه پیش‌بینی را انجام دهد. پس از محاسبه معیارهای ارزیابی برای هر مدل، نتیجه پیش‌بینی هر مدل، ذخیره و سپس با ترکیب مدل‌های طراحی‌شده، پیش‌بینی نهایی انجام شده است. نمودار روند اجرای رویکرد پیشنهادی در شکل ۱ آمده است.

برای پیش‌بینی اسپم در شبکه اجتماعی توییتر نیاز است که پیکره متنی مناسبی جمع‌آوری شود و برای افزایش دقت پیش‌بینی، پیکره متنی باید نرمال‌سازی و پاک‌سازی گردد. ویژگی‌های اصلی توییت از واژه‌های موجود در آن ناشی می‌شود. هر کلمه در پیکره به‌عنوان یک ویژگی عمل می‌کند. روش‌های مختلفی برای نمایش ویژگی‌ها مانند نمایش متراکم[1] وجود دارد. در این روش، ویژگی‌ها در ابعاد کمتری نمایش داده می‌شوند. مزیت نمایش متراکم، قدرت تعمیم آن و سرعت محاسباتی آن به‌دلیل ابعاد کمتر است. نمایش متراکم کلمات را می‌توان با استفاده از روش‌های معمولی مانند Word2vec، FastText و Glove به‌دست آورد که در این مقاله برای نمایش ویژگی‌ها از روش نمایش متراکم استفاده شده است. برای استخراج ویژگی‌های توئیت‌های پیکره، لازم است ابتدا بازنمایی و استخراج بردار کلمات صورت گیرد و سپس به بازنمایی متون کوتاه توییت پرداخته شود. برای بازنمایی و استخراج کلمات از جایگذاری بردارهای کلمات از پیش آموزش داده شده به نام Glove.Twitter.27B در چهار بعد و Word2vec_Twitter_Tokens در یک بعد به ترتیب در ابعاد ۲۵، ۵۰، ۱۰۰ و ۲۰۰ استفاده شده است. برای ساخت بردار متراکم پیکره متنی ۱۴۷۹۵نمونه‌ای، هر متن کوتاه به کلمه‌های تشکیل‌دهنده‌اش تبدیل می‌شود. در ادامه با استفاده از جایگذاری بردار کلمات از پیش آموزش داده شده، بردارهای کلمات هر یک از توییت‌ها استخراج و نهایتاً ماتریس کلمات و بردار متناظرشان ساخته می‌شود. پس از استخراج و ساخت ماتریس کلمات و بردار متناظرشان، به مرحله بازنمایی برداری متون کوتاه توییت و استخراج ویژگی‌های متن کوتاه توییت می‌رسیم. در این مرحله برای استخراج ویژگی‌های متون کوتاه، چندین مدل مبتنی بر شبکه عصبی Autoencoder طراحی شده است. Autoencoder که به‌اختصار AE نیز نامیده می‌شود، نوعی از شبکه‌های عصبی برای یادگیری بدون نظارت است که در اواخر دهه ۸۰ میلادی معرفی شد که این شبکه اغلب برای کاهش ابعاد و استخراج ویژگی استفاده می‌شود [۲۷]. استفاده از AE، یکی از روش‌های قدرتمند برای استخراج ویژگی‌ها از داده‌هاست که به جای ایجاد پیش‌بینی کلاسی، ویژگی‌های اصلی ورودی برای بازسازی شبکه خود را آموزش می‌بینند. AE دارای دو جزء انکدر[2] و دیکدر[3] است. انکدر وظیفه دریافت ورودی با ابعاد بالا و تبدیل آن به برداری با ابعاد کم را دارد. دیکدر، بردار ساخته‌شده توسط انکدر را دریافت و ورودی را دوباره بازسازی می‌کند. برای قسمت بازنمایی متون با استفاده از قسمت انکدر چندلایه‌ای، داده‌های ورودی با یک نگاشت غیرخطی به لایه پنهان برده می‌شوند که یک نمایش غیرخطی فشرده‌شده از ویژگی‌های داده ورودی خواهیم داشت. سپس این ویژگی‌ها به‌عنوان ورودی به دیکدر چندلایه‌ای داده می‌شوند. بدین ترتیب در بخش اول مدل با استفاده از شبکه‌های عصبی، ویژگی‌ها استخراج خواهند شد. با پایان کار شبکه‌های عصبی، بخش اول مدل خاتمه می‌یابد. در بخش دوم ویژگی‌های استخراج‌شده از آخرین لایه به‌عنوان ورودی به لایه softmax داده می‌شوند و محاسبه احتمال تعلق هر نمونه به هر یک از برچسب‌های ممکن را شبکه‌ی عصبی متراکم Dense که یک شبکه عصبی اتصال کامل است، با کمک لایه softmax انجام می‌دهد. بعد از به‌دست‌آمدن نتایج هر یک از مدل‌ها، تمام مدل‌ها با یکدیگر ترکیب می‌شوند تا بتوانند مدل‌های دقیق‌تر و یا پایدارتری به وجود آورند که با ترکیب‌کردن مدل‌ها به رویکرد پیشنهادی بیان‌شده دست می‌یابیم.

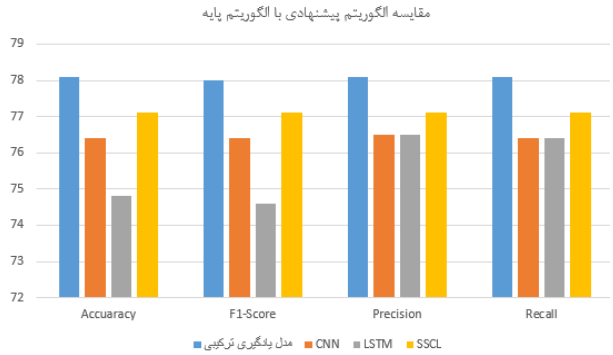
٤- نتایج آزمایش‌ها

٤-١ مجموعه داده

برای پیش‌بینی اسپم در شبکه اجتماعی توییتر نیاز است که پیکره متنی مناسبی جمع‌آوری شود. رویکرد پیشنهادی ارائه‌شده در این مقاله روی پیکره متنی Twitter Spam Detection[4] مورد آزمایش قرار گرفته است. این پیکره متنی شامل ۷ ستون می‌باشد؛ توییت، دنبال‌کننده، دنبال‌شونده، واکنش[5] کاربر، ریتوییت[6]، مکان کاربر و نوع توییت که با دو مقدار هرزنامه و باکیفیت مقداردهی شده است. در این مقاله تنها از متن توییت‌ها استفاده شده است. این پیکره شامل ۱۴۷۹۵ نمونه توییت است که ۷۳۹۷ نمونه متعلق به برچسب اسپم و ۷۳۹۸ نمونه متعلق به برچسب

1. Dense Representation

2. Encoder
3. Decoder
4. https://www.kaggle.com/competitions/twitter-spam/data
5. Actions
6. Retweet



شکل ۲: مقایسه الگوریتم پیشنهادی با الگوریتم‌های پایه.

اسپم‌نبودن است. ۶۰٪ از توییت‌های پیکره معادل ۸۸۷۷ نمونه توییت برای آموزش، ۲۰٪ آن معادل ۲۹۵۹ نمونه توییت برای اعتبارسنجی و ۲۰٪ دیگر آن معادل ۲۹۵۹ نمونه برای آزمایش روش پیشنهادی اختصاص داده شده است. همچنین توییت‌های این پیکره شامل یک تا ۴۰ کلمه است که در این پژوهش حداکثر طول کلمات ۴۰ در نظر گرفته شده است. برای افزایش دقت پیش‌بینی نیاز است پیکره متنی نرمال‌سازی و پاک‌سازی شود. برای پاک‌سازی پیکره، تمام حروف موجود به حروف کوچک تبدیل و اعداد، علائم نقطه‌گذاری، فضاهای خالی و کلمات بی‌اثر نیز پاک شده است. سپس عمل Tokenization برای تبدیل جمله به اجزای سازنده‌اش انجام شده است. از تابع Tokenizer Word برای توکنایزکردن و از دو کتابخانه re و NLTK برای انجام تمام کارهای ذکرشده در بالا استفاده شده است.

٤-٢ معیارهای ارزیابی

برای بررسی و مقایسه عملکرد روش پیشنهادی با روش‌های پایه از چهار معیار Accuracy، Precision، Recall و F۱−Score استفاده شده است. معیار Recall نشان می‌دهد که مقدار اندازه‌گیری‌شده تا چه اندازه به مقدار واقعی نزدیک است. منظور از معیار Precision، نسبت تعداد جملاتی است که در کلاس‌های درستی قرار داده شده‌اند به مجموع تعداد کل جملاتی که به‌درستی به کلاس‌ها تخصیص داده شده‌اند و جملاتی که به کلاس درستی تخصیص داده نشده‌اند. Recall کسری است از جملاتی که در کلاس‌های درستی قرار داده شده‌اند، نسبت به تعداد کل جملات. F۱−Score ترکیب Recall و Precision با یکدیگر است. معیارهای ذکرشده به‌صورت (۱) تا (۴) تعریف می‌شوند

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (۱)$$

$$Precision = \frac{TP}{TP + FP} \quad (۲)$$

$$Recall = \frac{TP}{TP + FN} \quad (۳)$$

$$F۱ - Score = ۲ \times \frac{Precision \times Recall}{Precision + Recall} \quad (۴)$$

٤-٣ نتیجه آزمایش‌ها

٤-٣-١ آموزش مدل‌ها

در رویکرد پیشنهادی، پیکره متنی Twitter Spam Detection روی مدل‌های مبتنی بر CNN Autoencoder، Sequence_to_Sequence Autoencoder و Sequential Stacked CNN-LSTM آموزش داده شده است. برای هر یک از مدل‌ها به‌منظور بازنمایی بردار کلمات از جایگذاری بردار کلمات Glove در چهار بعد و جایگذاری بردار کلمات Word2vec در یک بعد استفاده شده است. در اینجا برای هر پنج بعد، دو حالت تک‌لایه و چندلایه طراحی گردیده که در مجموع ۱۰ مدل آموزش داده شده است. منظور از تک‌لایه، یک انکدر، لایه پنهان و دیکدر و منظور از چندلایه، چندین انکدر، لایه پنهان و دیکدر است. در هر دو گروه تک‌لایه‌ای و چندلایه‌ای، شبکه دو ورودی را دریافت می‌کند که ورودی اول Embedding است که در آن ماتریس کلمات و بردار متناظرشان فراخوانی می‌شود. ورودی دوم فایل‌های متنی به نام‌های Train، Valid و Test می‌باشد که به‌ترتیب ۶۰٪، ۲۰٪ و ۲۰٪ از متن‌های توییت را به خود اختصاص داده‌اند. نتایج به‌دست‌آمده برای مدل‌های مبتنی بر CNN Autoencoder، Sequence_To_Sequence Autoencoder و Sequential Stacked CNN-LSTM به‌ترتیب در جداول ۱ تا ۳ آورده شده است.

٤-٣-٢ پیش‌بینی اسپم در توییتر مبتنی بر یادگیری ترکیبی

در این مقاله برای پیش‌بینی اسپم در توییتر با روش یادگیری ترکیبی، تقریباً ۳۰ مدل طراحی شده که در بالا به‌طور مختصر بیان گردید. خروجی هر یک از مدل‌ها به‌عنوان یک رأی در نظر گرفته می‌شود که در اینجا دو کلاس اسپم و اسپم‌نبودن است. کلاسی که اکثریت آرا را کسب کند، خروجی مدل ترکیبی خواهد بود. خروجی هر مدل در زمان اجرا ذخیره شده است. خروجی هر مدل به صورت یک لیست است که بعد از تمام شدن اجرای هر مدل ذخیره می‌شود. پس از اجرای تمام مدل‌ها ، حداکثر آرا این مدل‌ها محاسبه می‌شود که نتیجه حداکثر آن نیز به صورت لیست است. سپس مقدار واقعی نیز به‌صورت یک لیست فراخوانی گردیده است. با داشتن دو لیست پیش‌بینی، مقدار واقعی معیارهای تعریف‌شده برای ارزیابی محاسبه گردیدند. نتیجه حاصل از این روش پیشنهادی در جدول ۴ قابل مشاهده است.

٤-٣-٣ مقایسه روش پیشنهادی با الگوریتم‌های پایه

در این مقاله از پیکره متنی استفاده شده که در مقاله‌های پیشین آموزش داده نشده است؛ بنابراین ابتدا الگوریتم‌های پایه‌ای طراحی گردیده و سپس پیکره متنی روی این الگوریتم‌های طراحی‌شده آموزش داده شده است. بیسواز و همکاران [۱۵] با تحقیق روی مجموعه داده‌های Ott و Yelp با شبکه‌های عصبی عمیق که برای پیش‌بینی اسپم مناسب است، اقدام به پیش‌بینی اسپم کردند. آنها بعد از جمع‌آوری داده به پاکسازی و نرمال‌سازی آن پرداختند و سپس شبکه عصبی عمیق را برای آموزش مدل کردند. در این شبکه عمیق از لایه‌های CNN و LSTM استفاده شده است. شارما و همکاران [۱۲] و [۱۳] با طراحی شبکه‌های عصبی SCNN، SLSTM و SSCL مجموعه داده Twitter را برای پیش‌بینی اسپم آموزش داده‌اند. در این مقالات ابتدا با کمک لایه معنایی، بازنمایی بردار کلمات ارائه گردید و سپس، شبکه‌های عصبی CNN، LSTM و CNN_LSTM برای دسته‌بندی اسپم مورد آموزش قرار گرفتند. در لایه معنایی از جایگذاری بردار کلمات Word2vec برای بازنمایی بردار کلمات استفاده شده است. معماری‌هایی که مطرح شد در این مقاله به‌عنوان الگوریتم‌های پایه در نظر گرفته شده و معیارهای بیان‌گردیده برای این الگوریتم‌ها نیز ارزیابی شده است. همان طور که مشاهده می‌شود رویکرد پیشنهادی برای چهار معیار Accuracy، F۱−Score، Precision و Recall توانسته در مقایسه با روش CNN نتایج را تقریباً تا ۱/۷٪، روش LSTM بیش از ۳/۵٪ و روش SSCL نیز تا ۰/۹٪ بهبود دهد. نتایج

جدول ۱: نتایج به‌دست‌آمده برای پیش‌بینی اسپم در توییتر مبتنی بر مدل CNN Autoencoder.

	Model	Dimension	Measure			
			Accuracy	F۱-Score	Precision	Recall
Single Layla	Glove	۲۵	۷۳٫۹	۷۳٫۷	۷۴٫۴	۷۳٫۹
		۵۰	۷۴٫۷	۷۴٫۷	۷۴٫۷	۷۴٫۷
		۱۰۰	۷۷٫۵	۷۷٫۵	۷۷٫۵	۷۷٫۵
		۲۰۰	۷۷٫۰	۷۷٫۰	۷۷٫۰	۷۷٫۰
	Word2vec	۱۰۰	۷۶٫۴	۷۶٫۴	۷۶٫۵	۷۶٫۴
Multi-layer	Glove	۲۵	۷۴٫۳	۷۴٫۲	۷۴٫۶	۷۴٫۳
		۵۰	۷۵٫۷	۷۵٫۷	۷۵٫۸	۷۵٫۷
		۱۰۰	۷۱٫۷	۷۱٫۰	۷۴٫۲	۷۱٫۷
		۲۰۰	۷۶٫۶	۷۶٫۶	۷۶٫۸	۷۶٫۶
	Word2vec	۱۰۰	۷۲٫۵	۷۲٫۰	۷۴٫۳	۷۲٫۰

جدول ۲: نتایج به‌دست‌آمده برای پیش‌بینی اسپم در توییتر مبتنی بر مدل Sequence_to_Sequence Autoencoder.

	Model	Dimension	Measure			
			Accuracy	F۱-Score	Precision	Recall
Single layer	Glove	۲۵	۷۴٫۳	۷۴٫۳	۴٫۷۴	۷۴٫۳
		۵۰	۷۶٫۱	۷۶٫۱	۷۶٫۱	۷۶٫۱
		۱۰۰	۷۵٫۱	۷۵٫۱	۷۵٫۱	۷۵٫۱
		۲۰۰	۷۴٫۹	۷۴٫۹	۷۵٫۰	۷۴٫۹
	Word2vec	۱۰۰	۷۴٫۸	۷۴٫۶	۷۶٫۵	۷۶٫۴
Multi-layer	Glove	۲۵	۷۵٫۴	۷۵٫۴	۷۵٫۴	۷۵٫۴
		۵۰	۷۶٫۳	۷۶٫۳	۷۶٫۳	۷۶٫۳
		۱۰۰	۷۴٫۵	۷۴٫۳	۷۵٫۵	۷۴٫۵
		۲۰۰	۷۴٫۸	۷۴٫۷	۷۵٫۲	۷۴٫۸
	Word2vec	۱۰۰	۷۴٫۹	۷۴٫۹	۷۵٫۰	۷۴٫۹

جدول ۳: نتایج به‌دست‌آمده برای پیش‌بینی اسپم در توییتر مبتنی بر مدل Sequential Stacked CNN-LSTM.

	Model	Dimension	Measure			
			Accuracy	F۱-Score	Precision	Recall
Single layer	Glove	۲۵	۷۴٫۳	۷۴٫۳	۴٫۷۴	۷۴٫۳
		۵۰	۷۶٫۱	۷۶٫۱	۷۶٫۱	۷۶٫۱
		۱۰۰	۷۵٫۱	۷۵٫۱	۷۵٫۱	۷۵٫۱
		۲۰۰	۷۴٫۹	۷۴٫۹	۷۵٫۰	۷۴٫۹
	Word2vec	۱۰۰	۷۴٫۸	۷۴٫۶	۷۶٫۵	۷۶٫۴
Multi-layer	Glove	۲۵	۷۵٫۴	۷۵٫۴	۷۵٫۴	۷۵٫۴
		۵۰	۷۶٫۳	۷۶٫۳	۷۶٫۳	۷۶٫۳
		۱۰۰	۷۴٫۵	۷۴٫۳	۷۵٫۵	۷۴٫۵
		۲۰۰	۷۴٫۸	۷۴٫۷	۷۵٫۲	۷۴٫۸
	Word2vec	۱۰۰	۷۴٫۹	۷۴٫۹	۷۵٫۰	۷۴٫۹

جدول ۴: نتایج به‌دست‌آمده برای پیش‌بینی اسپم در توییتر مبتنی بر یادگیری ترکیبی.

مدل	نام معیار			
	Accuracy	F۱-Score	Precision	Recall
رویکرد ترکیبی	۷۸٫۱	۷۸٫۰۱	۷۸٫۱	۷۸٫۱

به‌دست‌آمده در جدول ۵ و شکل ۲ قابل مشاهده است.

برای مقایسه نتایج این تحقیق با روش‌های ترکیبی دیگر، در بررسی سوابق مشخص گردید که مجموعه داده مورد استفاده این تحقیق در کارهای پیشین استفاده نشده است. در این میان [۲۶] از رویکردی کلی مشابه این تحقیق استفاده کرده؛ اما روی یک مجموعه داده دیگر مدل‌سازی انجام داده است. در [۲۶] روش ترکیبی پیشنهادشده شامل ۲ جزء اصلی ماژول پایه و ماژول ترکیبی است. در ماژول پایه از شش الگوریتم یادگیری ماشین برای استخراج ویژگی‌ها از داده اصلی استفاده شده است. داده به‌دست‌آمده از ماژول پایه در ترکیب با متا داده‌ای با

جدول ۵: نتایج به‌دست‌آمده برای روش پیشنهادی و الگوریتم‌های پایه.

نام پژوهش	روش پیشنهادی	نام معیار			
		Accuracy	F۱−Score	Precision	Recall
الگوریتم پیشنهادی	یادگیری ترکیبی	۷۸٫۱	۷۸	۷۸٫۱	۷۸٫۱
بیسواز	CNN	۷۶٫۴	۷۶٫۴	۷۶٫۵	۷۶٫۴
	LSTM	۷۴٫۸	۷۴٫۶	۷۶٫۵	۷۶٫۴
شارما	SSCL	۷۷٫۱	۷۷٫۱	۷۷٫۱	۷۷٫۱

ویژگی‌های جدید به‌عنوان داده ورودی در ماژول ترکیبی استفاده شده است. در ماژول ترکیبی، یک شبکه عصبی عمیق (DNN)[1] به‌عنوان یک متا طبقه‌بندی‌کننده برای گرفتن اطلاعات عمیق پنهان به‌کار گرفته شده است. در این تحقیق مقدار Accuracy و F۱−Score برابر ۷۰٪ گزارش شده که از مقدار روش پیشنهادی این تحقیق کمتر است. بنابراین با توجه به نتایج به‌دست‌آمده می‌توان نتیجه گرفت که روش پیشنهادی در مقایسه با روش‌های پایه قدرتمندی چون LSTM، CNN و SSCL و همین طور روش ترکیبی ارائه‌شده در [۲۶] نتایج بهتری را کسب نموده است.

۵- نتیجه‌گیری

در سال‌های اخیر، رفتارهای مخرب اسپمرها تهدیدی جدی برای امنیت اطلاعات کاربران بوده است. توییتر یکی از محبوب‌ترین شبکه‌های اجتماعی است و به بستری مناسب برای تبلیغ‌کنندگان و اسپمرها به‌منظور انتشار پیام‌های آنها تبدیل شده است. در سال‌های اخیر برای رفع مشکل اسپم توییتر، روش‌های تشخیص بسیاری ارائه گردیده که اکثر آنها برای جلوگیری از فعالیت اسپمرها و فیلترکردن هرزنامه در توییتر مبتنی بر شناسایی اسپمرها و مسدودکردن آنهاست. محققان، پیوسته روش‌هایی جدیدتر برای شناسایی اسپم با دقت بالاتر و کارایی بیشتر ارائه داده‌اند؛ به‌ویژه اینکه شناسایی اسپم در توییت‌ها جای کار و بهبود دارد. در این مقاله برای شناسایی اسپم در سطح توییت از ترکیبی از روش‌های یادگیری ماشین استفاده شده است. روش پیشنهادی، چارچوبی مبتنی بر استخراج ویژگی است که در دو مرحله انجام می‌شود. در مرحله اول از Stacked Autoencoder برای استخراج ویژگی‌ها استفاده شده و در مرحله دوم ویژگی‌های مستخرج از آخرین لایه Stacked Autoencoder به‌عنوان ورودی به لایه softmax داده می‌شود تا این لایه پیش‌بینی را انجام دهد. نتیجه آزمایش روش پیشنهادی با نتایج روش‌های پایه روی پیکره متنی Twitter Spam Detection برای معیارهای Accuracy، F۱−Score، Precision و Recall مورد مقایسه و ارزیابی قرار گرفته است. رویکرد پیشنهادی در مقایسه با روش‌های پایه CNN، LSTM و SSCL توانسته که دقت را به‌ترتیب ۱٫۷٪، ۳٫۳٪ و ۱٪ بهبود بخشد. با توجه به نتایج به‌دست‌آمده می‌توان گفت که روش پیشنهادی توانسته با استفاده از رویکرد اکثریت آرا با انتخاب سخت در یادگیری ترکیبی، پیش‌بینی توییت‌های اسپم را با دقت بالاتری نسبت به روش‌های پایه تشخیص دهد. رویکرد یادگیری ترکیبی با روش‌های متفاوتی محاسبه می‌شود که می‌توان آنها را در آینده مورد بررسی قرار داد. از جمله این روش‌ها می‌توان به روش انتخاب نرم، Boosting و Stacking اشاره کرد. این نکته نیز حائز اهمیت است که برای هر روش، انتخاب الگوریتم نقشی اساسی دارد.

1. Deep Neural Network

مراجع

[1] S. Madisetty and M. S. Desarkar, "A neural network-based ensemble approach for spam detection in Twitter," *IEEE Trans. Comput. Soc. Syst.*, vol. 5, no. 4, pp. 973-984, Dec. 2018.

[2] M. McCord and M. Chuah, "Spam detection on twitter using traditional classifiers," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. LNCS6906, pp. 175-186, Sept. 2011.

[3] X. Zhang, S. Zhu, and W. Liang, "Detecting spam and promoting campaigns in the Twitter social network," in *Proc. IEEE International Conf. on Data Mining, ICDM*, pp. 1194-1199, Brussels, Belgium , 10-13 Dec. 2012.

[4] A. T. Kabakus and R. Kara, "A survey of spam detection methods on Twitter," *International J. of Advanced Computer Science and Applications*, vol. 8, no. 3, pp. 29-38, 2017.

[5] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing*, vol. 159, no. 1, pp. 27-34, Jul. 2015.

[6] J. Martinez-Romo and L. Araujo, "Detecting malicious tweets in trending topics using a statistical analysis of language," *Expert Syst. Appl.*, vol. 40, no. 8, pp. 2992-3000, Jun. 2013.

[7] A. M. Al-Zoubi, H. Faris, J. Alqatawna, and M. A. Hassonah, "Evolving support vector machines using whale optimization algorithm for spam profiles detection on online social networks in different lingual contexts," *Knowledge-Based Syst.*, vol. 153, pp. 91-104, Aug. 2018.

[8] S. B. S. Ahmad, M. Rafie, and S. M. Ghorabie, "Spam detection on Twitter using a support vector machine and users' features by identifying their interactions," *Multimed. Tools Appl.*, vol. 80, no. 8, pp. 11583-11605, Mar. 2021.

[9] Z. Alom, B. Carminati, and E. Ferrari, "A deep learning model for Twitter spam detection," *Online Soc. Networks Media*, vol. 18, Article ID: 100079, Jul. 2020.

[10] X. Ban, C. Chen, S. Liu, Y. Wang, and J. Zhang, "Deep-learnt features for Twitter spam detection," in *Proc. Int. Symp. Secur. Priv. Soc. Networks Big Data*, pp. 22-26, Santa Clara, CA, USA, 10-11 Dec. 2018.

[11] Y. Liu, L. Wang, T. Shi, and J. Li, "Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and Bi-LSTM," *Inf. Syst.*, vol. 103, Article ID: 101865, Jan. 2022.

[12] G. Jain, M. Sharma, and B. Agarwal, "Optimizing semantic LSTM for spam detection," *Int. J. Inf. Technol.*, vol. 11, no. 2, pp. 239-250, Jun. 2019.

[13] G. Jain, M. Sharma, and B. Agarwal, "Spam detection in social media using convolutional and long short term memory neural network," *Annals of Mathematics and Artificial Intelligence,* vol. 85, no. 1, pp. 21-44, 2019.

[14] T. Wu, S. Liu, J. Zhang, and Y. Xiang, "Twitter spam detection based on deep learning," in *Proc. ACM Int. Conf. Proc. Ser.*, 8 pp., Geelong, Australia, 30 Jan.-3 Feb 2017.

[15] G. M. Shahariar, S. Biswas, F. Omar, F. M. Shah, and S. Binte Hassan, "Spam review detection using deep learning," in *Proc. IEEE 10th Annu. Inf. Technol. Electron. Mob. Commun. Conf., IEMCON'19*, pp. 27-33, Vancouver, Canada, 17-19 Oct. 2019.

[16] A. T. Kabakus and R. Kara, "'TwitterSpamDetector': a spam detection framework for twitter," *International J. of Knowledge and Systems Science*, vol. 10, no. 3, pp. 1-14, Jul. 2019.

[17] H. Shen, *et al.*, "Discovering social spammers from multiple views," *Neurocomputing*, vol. 225, pp. 49-57, Feb. 2017.

[18] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots + machine learning," in *Proc. SIGIR Proc.-33rd Annual International ACM SIGIR Conf. on Research and*

Development in Information Retrieval, pp. 435-442, Geneva, Switzerland, 19-23 Jul. 2010.

[19] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@spam: the underground on 140 characters or less," in *Proc. of the ACM Conf. on Computer and Communications Security*, pp. 27-37, Chicago, IL, USA, 4-8 Oct. 2010.

[20] S. Saumya and J. P. Singh, "Spam review detection using LSTM autoencoder: an unsupervised approach," *Electron. Commer. Res.*, vol. 22, no. 1, pp. 113-133, Mar. 2022.

[21] J. V. Lochter, T. A. Almeida, and T. C. Alberto, "TubeSpam: comment spam filtering on YouTube," in *Proc. IEEE 14th Int, Conf. on Machine Learning and Applications*, pp. 138-143, Miami, FL, USA, 9-11 Dec. 2015.

[22] M. M. Abdulhasan, H. Alchilibi, M. A. Mohammed, and R. Nair, "Real-time sentiment analysis and spam detection using machine learning and deep learning," *in Proc. 3rd Int. Conf. on Data Science and Big Data Analytics*, pp. 507-533, Indore, India, 16-17 Jun. 2023.

[23] A. Ahraminezhad, M. Mojarad, and H. Arfaeinia, "An intelligent ensemble classification method for spam diagnosis in social networks," *International J. of Intelligent Systems and Applications*, vol. 14, no. 1, pp. 24-31, Feb. 2022.

[24] Z. Alom, B. Carminati, and E. Ferrari, "A deep learning model for Twitter spam detection," *Online Social Networks and Media*, Article ID: 100079, Jul. 2020.

[25] S. Liu, Y. Wang, J. Zhang, C. Chen, and Y. Xiang, "Addressing the class imbalance problem in twitter spam detection using ensemble learning," *Computers & Security*, vol. 69, pp. 35-49, Aug. 2017.

[26] C. Zhao, Y. Xin, X. Li, Y. Yang, and Y. Chen, "A heterogeneous ensemble learning framework for spam detection in social networks with imbalanced data," *Applied Sciences*, vol. 10, no. 3, Article ID" 936, Jan. 2020.

[27] M. Usama, *et al.*, "Unsupervised machine learning for networking: techniques, applications and research challenges," *IEEE Access*, vol. 7, pp. 65579-65615, 2019.

مریم فصیحی تحصیلات خود در مقطع کارشناسی را در رشته مهندسی فناوری اطلاعات گرایش طراحی صفحات وب در دانشگاه علمی کاربردی به پایان رساند و هم‌اکنون فارغ‌التحصیل رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه علم و فرهنگ است. زمینه تحقیقاتی مورد علاقه ایشان تحلیل شبکه‌های اجتماعی و هوش مصنوعی است.

محمدجواد شایگان فرد تحصیلات خود در مقاطع کارشناسی و کارشناسی ارشد را در دانشگاه علوم و تحقیقات در گرایش‌های مهندسی نرم‌افزار و معماری کامپیوتر در سال ۱۳۷۷ به پایان رساند. سپس در سال ۱۳۸۸ در رشته دکتری فناوری اطلاعات و سیستم‌های چندرسانه‌ای از دانشگاه یوپی‌ام مالزی فارغ‌التحصیل شد. وی هم‌اکنون عضو هیأت علمی و دانشیار گروه مهندسی کامپیوتر دانشگاه علم و فرهنگ است. دکتر شایگان فرد پایه‌گذار موضوع وب پژوهی در کشور است و در این راستا کنفرانس بین‌المللی وب پژوهی را از سال ۱۳۹۴ در کشور بنا نهاد که تا کنون بطور پیاپی برگزار و در IEEE نمایه شده است. در ادامه مجله وب پژوهی توسط وی از سال ۱۳۹۸ بنا نهاده شد که هم‌اکنون از نشریات معتبر وزارت علوم، تحقیقات و فناوری است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: داده‌کاوی و یادگیری ماشین، علوم وب، تجارت و بازاریابی الکترونیکی و سیستم‌های توزیعی.

زهرا سادات حسینی تحصیلات خود در مقطع کارشناسی را در رشته مهندسی کامپیوتر نرم‌افزار در دانشگاه پیام نور قم به پایان رساند و هم‌اکنون فارغ التحصیل رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه علم و فرهنگ است. زمینه تحقیقاتی مورد علاقه ایشان پردازش زبان طبیعی، یادگیری عمیق و هوش مصنوعی است.

زهرا سجده تحصیلات خود در مقطع کارشناسی را در رشته مهندسی کامپیوتر- نرم‌افزار در دانشگاه ابرار به پایان رساند و هم اکنون فارغ التحصیل رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه علم و فرهنگ است. زمینه تحقیقاتی مورد علاقه ایشان تحلیل شبکه‌های اجتماعی، تحلیل متن و هوش مصنوعی است.