Vol. 3, No. 6, Apr-Jun 2023, (98-111)

Journal of New Achievements in Electrical, Computer and Technology (JNAECT)

مجله دستاوردهای نوین در برق،کامپیوتر و فناوری

دوره ۳، شماره ۶، بهار ۱۴۰۲(۱۱۱-۹۸)

journal homepage: https://jnaect.ir

پیشنهاد روش نوینی برای شناسایی هرزنامه ها در SMS ها

سیما مستخدمین حسینی[۱] ، مریم تعجبیان[۲]

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه آزاد اسلامی، واحد مهدیشهر، سمنان، ایران

Simahosseiniiii1972@gmail.com[1] , taajobian.uni@gmail.com[2]

چکیده

در سال های اخیر اینترنت به بخشی جدایی ناپذیر از زندگی ما انسانها تبدیل شده است. با افزایش استفاده از اینترنت، تعداد کاربران ایمیل روز به روز در حال افزایش است. این استفاده روزافزون از ایمیل مشکلاتی را ایجاد کرده است که ناشی از پیام‌های ایمیل انبوه ناخواسته است که معمولاً به آن اسپم می‌گویند. ایمیل در حال حاضر به یکی از بهترین راه ها برای تبلیغات تبدیل شده است که به واسطه آن ایمیل های اسپم تولید می شود. ایمیل های اسپم، ایمیل هایی هستند که گیرنده تمایلی به دریافت آنها ندارد. تعداد زیادی پیام یکسان برای چندین گیرنده ایمیل ارسال می شود. هرزنامه معمولاً در نتیجه ارائه آدرس ایمیل ما در یک وب سایت غیرمجاز یا غیرقانونی ایجاد می شود. بسیاری از اثرات هرزنامه وجود دارد. صندوق ورودی ما را با تعداد زیادی ایمیل بی فایده پر می کند. سرعت اینترنت ما را تا حد زیادی کاهش می دهد. اطلاعات مفیدی مانند سرقت می کند. لذا در این تحقیق روش نوینی برای شناسایی ایمیل های اسپم نوشته شده به زبان فارسی، پیشنهاد می نماییم. در این روش از ترکیبی از تکنیکهای یادگیری ماشین و یادگیری عمیق برای شناسایی اسپم های ایمیل استفاده نموده ایم. روش پیشنهادی با استفاده از مجموعه داده های آنلاین Kaggle استفاده نمودیم. نتایج بدست آمده نشان داد که روش پیشنهادی از لحاظ معیارهای ارزیابی (دقت، صحت، Recall و F-Measure) کارایی بهتری در مقایسه با روشهای پیشنهاد شده توسط محققان دیگر دارد.

واژه‌های کلیدی: هرزنامه ها، SMSها، مجموعه داده های آنلاین Kaggle، تکنیکهای یادگیری ماشین، تکنیکهای یادگیری عمیق.

.

۱. مقدمه

یادگیری فرایندی است که در آن سیستم با استفاده از تجربیات گذشته، عملکرد خود را بهبود می بخشد. از سال ۲۰۰۶ ، یادگیری عمیق به عنوان یک زیرشاخه جدید از یادگیری ماشین ظاهر شده است، که بر طیف گسترده ای از پردازش سیگنال و اطلاعات در حوزه های سنتی و جدید تأثیر می گذارد. بسیاری از تکنیک های سنتی یادگیری ماشین و پردازش سیگنال از معماری های خاصی بهره می برند که حاوی یک لایه واحد از ویژگی های غیرخطی است.

چند نمونه از یادگیری عمیق در محل کار عبارتند از: یک وسیله نقلیه خودران با نزدیک شدن به گذرگاه عابر پیاده سرعت خود را کم می کند، دستگاه خودپرداز یک اسکناس بانکی تقلبی را رد می کند، یک برنامه تلفن هوشمند ترجمه فوری یک تابلوی نصب شده در خیابان را انجام می دهد. یادگیری عمیق به ویژه برای برنامه های شناسایی هویت مانند شناسایی چهره، ترجمه متن، تشخیص صدا و سیستم های پیشرفته کمک راننده، از جمله و تشخیص علائم بسیار مناسب است [1] .

یادگیری عمیق یکی از زیرشاخه های یادگیری ماشین است. با یادگیری ماشین، به صورت دستی ویژگی های مربوط به یک تصویر را می توان استخراج کرد. با یادگیری عمیق، تصاویر خام را می توان مستقیماً وارد یک شبکه عصبی عمیق کرد که ویژگی ها را به طور خودکار یاد می گیرد. یادگیری عمیق معمولاً به صدها هزار یا میلیون ها تصویر برای دستیابی به بهترین نتیجه نیاز دارد، در حالی که یادگیری ماشین با مجموعه داده های کوچک نتایج خوبی به همراه دارد. یادگیری عمیق همچنین از نظر محاسباتی فشرده است و به CPU با کارایی بالا نیاز دارد. [2].

یادگیری عمیق موثرترین، نظارت شده ترین و هزینه مقرون به صرفه رویکرد یادگیری ماشین است. یادگیری عمیق یک روش یادگیری محدود نیست، اما روش ها و توپوگرافی های مختلفی را که می تواند برای پیش بینی های وسیعی درباره مشکلات پیچیده استفاده شود، رعایت می کند. این تکنیک ویژگی های گویا و افتراقی را به صورت کاملا طبقه بندی شده در بر می گیرد. روشهای یادگیری عمیق با عملکرد قابل ملاحظه در طیف گسترده ای از برنامه ها با ابزارهای امنیتی مفید، دستیابی به موفقیت قابل توجهی داشته اند. یادگیری عمیق در بسیاری از برنامه ها کاربرد دارد از جمله: تجارت، آزمایشات تطبیقی، طبقه بندی تصاویر بیولوژیکی، بینش رایانه ای، تشخیص سرطانها، پردازش زبان طبیعی، تشخیص اشیاء، تشخیص چهره، دست خط، تشخیص گفتار، تحلیل بازار سهام، ایجاد و توسعه شهرهای هوشمند و بسیاری موارد دیگر.[3]

یادگیری ماشینی زیرمجموعه ای از هوش مصنوعی (AI) است که به سیستمها، مزایای یادگیری خودکار مفاهیم و دانش بدون برنامه ریزی صریح را می بخشد. این کار با مشاهداتی مانند تجارب مستقیم برای آماده سازی ویژگی ها و الگوهای موجود در داده ها و تولید نتایج و تصمیمات بهتر در آینده آغاز می شود. یادگیری عمیق به مجموعه الگوریتم های یادگیری ماشین متکی است که انتزاعات سطح بالا را در داده ها با چندین تبدیل غیرخطی مدلسازی می کند. فناوری یادگیری عمیق بر روی سیستم شبکه عصبی مصنوعی (ANN) کار می کند. این شبکه های عصبی به طور مداوم الگوریتم های یادگیری را استفاده نموده و با افزایش مداوم مقدار داده ها، می توان بازدهی فرایندهای آموزش را بهبود بخشید. کارایی الگوریتم های یادگیری عمیق، به حجم داده های حجیم وابسته است. فرآیند آموزش عمیق نامیده می شود زیرا تعداد سطوح شبکه عصبی با گذشت زمان افزایش می یابد [4].

در این تحقیق روش نوینی برای شناسایی هرزنامه های ارسال شده از طریق SMS پیشنهاد خواهیم کرد که در این روش از تکنیکهای یادگیری عمیق برای تشخیص هرزنامه ها و SMS های عادی استفاده خواهیم کرد. یادگیری عمیق (به طور خاص، مدل‌های شبکه عصبی کانولوشن و حافظه کوتاه‌مدت بلندمدت) برای طبقه بندی پیام های متنی هرزنامه و غیر هرزنامه مورد استفاده قرار خواهد گرفت. مدل‌های پیشنهادی فقط بر اساس داده‌های متنی مورد تجزیه و تحلیل قرار خواهند گرفت و مجموعه ویژگی‌ها را در آنها استخراج خواهیم کرد. استفاده از تکنیکهای یادگیری عمیق باعث بالارفتن دقت و سرعت شناسایی هرزنامه ها در بین انبوهی از پیام های SMS ارسالی خواهد شد.

این تحقیق در پنج بخش تنظیم شده است. در بخش دوم پیشینه تحقیقات انجام شده توسط محققان مختلف که مرتبط با این تحقیق هستند را مرور خواهیم کرد. در بخش سوم، روش پیشنهادی را توضیح خواهیم داد. در بخش چهارم پیاده سازی و ارزیابی کارایی روش پیشنهادی را خواهیم داشت و نهایتا در بخش پنجم نتیجه گیری را خواهیم آورد.

۲. مروری بر پیشینه تحقیقات انجام شده توسط محققان دیگر

Gauri Jain و همکارانش [۵] استفاده از یک فناوری یادگیری عمیق به نام شبکه عصبی کانولوشنال (CNN[1]) برای تشخیص هرزنامه با یک لایه معنایی اضافه شده در بالای آن را پیشنهاد دادند مدل حاصل به عنوان یک شبکه عصبی کانولوشن معنایی (SCNN[2]) شناخته می شود. یک لایه معنایی از آموزش بردارهای کلمه تصادفی با کمک Word vec ۲ برای به دست آوردن جاسازی کلمه غنی شده از لحاظ معنایی تشکیل شده است. WordNet و ConceptNet برای یافتن کلمه مشابه در یک جمله داده شده استفاده می شوند. این معماری بر روی دو مجموعه ارزیابی می‌شود: مجموعه داده‌های پیامک هرزنامه (مخزن UCI) و مجموعه داده توییتر (توییت‌هایی که ازمتن توییت‌های زنده عمومی حذف شده‌اند). Sudanagunta و همکارانش [۶] از مفاهیم یادگیری ماشینی پیشرفته برای تشخیص فیلتر اسپم در پیامک ها استفاده نمودند. در سیستم پیشنهادی در این تحقیق، مجموعه داده را از مخزن UCI استفاده کردند و برای تشخیص هرزنامه پیامک، طبقه‌بندی‌کننده‌های یادگیری ماشینی مانند الگوریتم‌های پشتیبان ماشین بردار(SVM) ، نزدیک‌ترین همسایه (KNN) و شبکه‌های عصبی (NN) و با معیارهای آنها مانند دقت پیاده‌سازی نمودند. Xiaoxu Liu و همکارانش [۷] به بررسی امکان مدل ترانسفورماتور در تشخیص پیام‌های سرویس پیام کوتاه هرزنامه (SMS) با پیشنهاد یک مدل تغییر یافته تبدیل‌شده است که برای تشخیص پیام‌های هرزنامه SMS طراحی شده است. ارزیابی ترانسفورماتور هرزنامه پیشنهادی در این تحقیق، بر روی مجموعه داده های SMS Spam Collection v.1 و مجموعه داده های مسابقه تشخیص هرزنامه توییترUtkMl ، با معیار چندین طبقه بندی کننده یادگیری ماشینی ایجاد شده و رویکردهای پیشرفته تشخیص هرزنامه SMS انجام می شود. ارزیابی نشان داد که مدل پیشنهادی عملکرد خوبی را در مجموعه داده‌های توییتر UtkMl به دست می‌آورد که نشان‌دهنده امکان امیدوارکننده‌ای برای انطباق مدل با سایر مشکلات مشابه است. Paras Sethi و همکارانش [۸] نقاط قوت نسبی الگوریتم‌های مختلف یادگیری ماشین را برای شناسایی پیام‌های هرزنامه‌ای که بر روی دستگاه‌های تلفن همراه ارسال می‌شوند، تحلیل و بررسی کردند. آنها داده ها را از مجموعه داده عمومی باز به دست آوردند و دو مجموعه داده را برای اهداف آزمایش و اعتبار سنجی خود آماده کردند. دقت در تشخیص پیام های هرزنامه اولین اولویت در رتبه بندی این الگوریتم ها بود. نتایج آنها به وضوح نشان می‌دهد که الگوریتم‌های مختلف یادگیری ماشین تحت ویژگی‌های مختلف، در طبقه‌بندی پیام‌های هرزنامه عملکرد متفاوتی دارند. Milivoje Popovac و همکارانش [۹] بر این نکته تاکید کردند که پیامک هرزنامه به پیام متنی ناخواسته اشاره دارد. مجموعه داده های مورد استفاده در این تحقیق به عنوان مجموعه داده تیاگو شناخته می شود. مرحله مهم در آزمایش، پیش پردازش داده ها بود، که شامل کاهش متن به حروف کوچک، نشانه گذاری، حذف کلمات توقف بود. شبکه عصبی کانولوشنال روش پیشنهادی برای طبقه بندی بود. دقت کلی مدل ۹۸.۴٪ بود. مدل به دست آمده می تواند به عنوان یک ابزار در بسیاری از برنامه ها استفاده شود. Feng Wei و همکارانش [۱۰] یک مدل عصبی عمیق سبک جدید به نام واحد بازگشتی دروازه‌ای سبک وزن (LGRU) برای تشخیص هرزنامه پیامک پیشنهاد کردند. به‌علاوه، آنها معناشناسی تقویت‌شده بازیابی شده از دانش خارجی (WordNet) را برای تقویت درک ورودی‌های متن پیام کوتاه برای طبقه‌بندی بهتر ترکیب کردند. Dima Suleiman و همکارانش [۱۱] طبقه‌بندی‌کننده جدیدی پیشنهاد کردند که عمدتاً به استفاده از h2o به عنوان پلت فرم برای مقایسه بین الگوریتم‌های مختلف یادگیری ماشین بستگی دارد. علاوه بر این، الگوریتم‌های یادگیری ماشینی که برای مقایسه استفاده شدند، جنگل تصادفی، یادگیری عمیق و خلیج‌های ساده هستند. علاوه بر استفاده از یادگیری عمیق و جنگل تصادفی به عنوان طبقه‌بندی‌کننده، آنها همچنین برای تعیین مهم‌ترین ویژگی‌هایی که می‌توانند به عنوان ورودی برای طبقه‌بندی‌کننده‌های جنگل تصادفی، یادگیری عمیق استفاده شوند، استفاده کردند. نتایج تجربی نشان می‌دهد که مهم‌ترین ویژگی‌هایی که می‌تواند بر تشخیص هرزنامه پیامک تأثیر بگذارد، تعداد ارقام و وجود URL در متن پیامک است.. Nikhil Kumar و همکارانش [۱۲] به این نکته تاکید کردند که، ارسال لینک مخرب از طریق ایمیل های اسپم که می تواند به سیستم ما آسیب برساند و همچنین می تواند به سیستم شما وارد شود. ایجاد یک نمایه و حساب ایمیل

[1] Convolutional Neural Network (CNN)
[2] Semantic Convolutional Neural Network (SCNN)

جعلی برای هرزنامه‌نویسان بسیار آسان است، آنها در ایمیل‌های هرزنامه خود مانند یک شخص واقعی وانمود می‌کنند، این هرزنامه‌ها افرادی را هدف قرار می‌دهند که از این کلاهبرداری‌ها آگاه نیستند. بنابراین، برای شناسایی ایمیل‌های هرزنامه که کلاهبرداری هستند، لازم است، این پروژه با استفاده از تکنیک‌های یادگیری ماشینی، آن هرزنامه‌ها را شناسایی می‌کند، این تحقیق، الگوریتم‌های یادگیری ماشینی را مورد بحث قرار می‌دهد و همه این الگوریتم‌ها را بر روی مجموعه داده‌های ما اعمال می‌کند و بهترین الگوریتم برای آن انتخاب می‌شود. تشخیص هرزنامه ایمیل دارای بهترین دقت و صحت است. Andrew Kipkebutو همکارانش [۱۳] بر روی شناسایی هرزنامه‌های پیامکی سمت مشتری در تلفن همراه کنیا با استفاده از یادگیری ماشین تمرکز کردند. الگوریتم Naive's Bayes برای این منظور استفاده شد . Tian Xiaو همکارانش [۱۴] روش جدیدی را بر اساس مدل مخفی مارکوف گسسته (HMM) برای استفاده از اطلاعات ترتیب کلمه و حل مسئله فرکانس پایین در تشخیص هرزنامه پیامک پیشنهاد کردند. مجموعه داده های هرزنامه اس ام اس محبوب از مخزن یادگیری ماشین UCI برای تجزیه و تحلیل عملکرد روش HMM پیشنهادی استفاده شد. آزمایش‌ها نشان می‌دهند که روش HMM پیشنهادی به زبان حساس نیست و می‌تواند هرزنامه‌ها را با دقت بالا در هر دو مجموعه داده شناسایی کند.

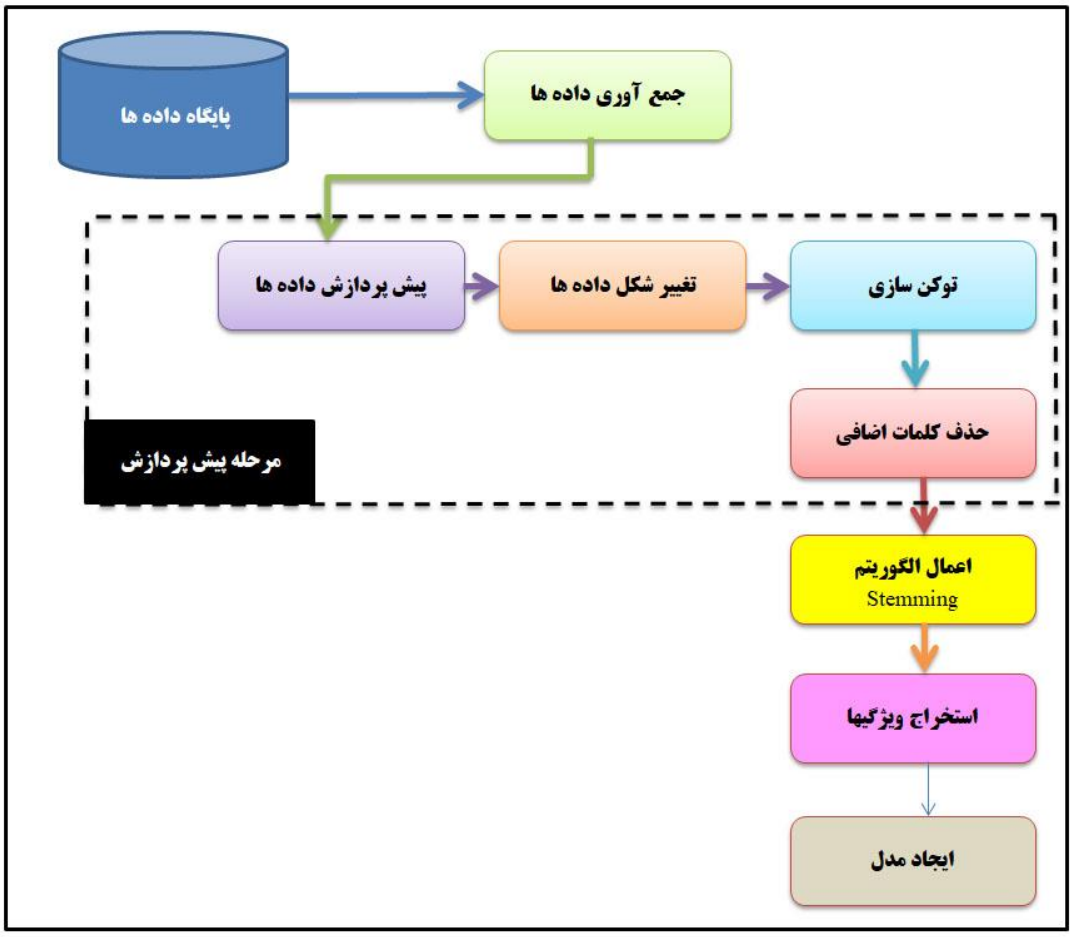
در این تحقیق، روش نوینی برای شناسایی هرزنامه ها در SMS پیشنهاد می کنیم که، در آن از تکنیکهای یادگیری عمیق استفاده خواهیم کرد.

۳. روش پیشنهادی

چارچوب کلی روش پیشنهادی در شکل ۱ نشان داده شده است.

همانطور که در شکل ۱ مشاهده می کنیم، مراحل کلی موجود در روش پیشنهادی عبارتند از:

- **جمع آوری داده ها:** در این تحقیق داده های خام از منبع آنلاین شده Kaggle جمع آوری خواهد شد.
- **پیش پردازش داده ها:** پیش پردازش داده ها در یادگیری ماشین (ML)، عبارت پیش پردازش به سازماندهی و مدیریت داده های خام قبل از استفاده از آن برای آموزش و آزمایش مدل های مختلف یادگیری اشاره دارد. به عبارت ساده‌تر، پیش‌پردازش یک رویکرد داده‌کاوی ML است که داده‌های خام را به ساختاری قابل استفاده و منابع تبدیل می‌کند. اولین گام در ساخت یک مدل ML، پیش پردازش است، که در آن داده های دنیای واقعی، معمولاً ناقص، نادقیق و به دلیل نقص و کمبود، به متغیرها و روندهای ورودی دقیق و قابل استفاده تبدیل می شوند.



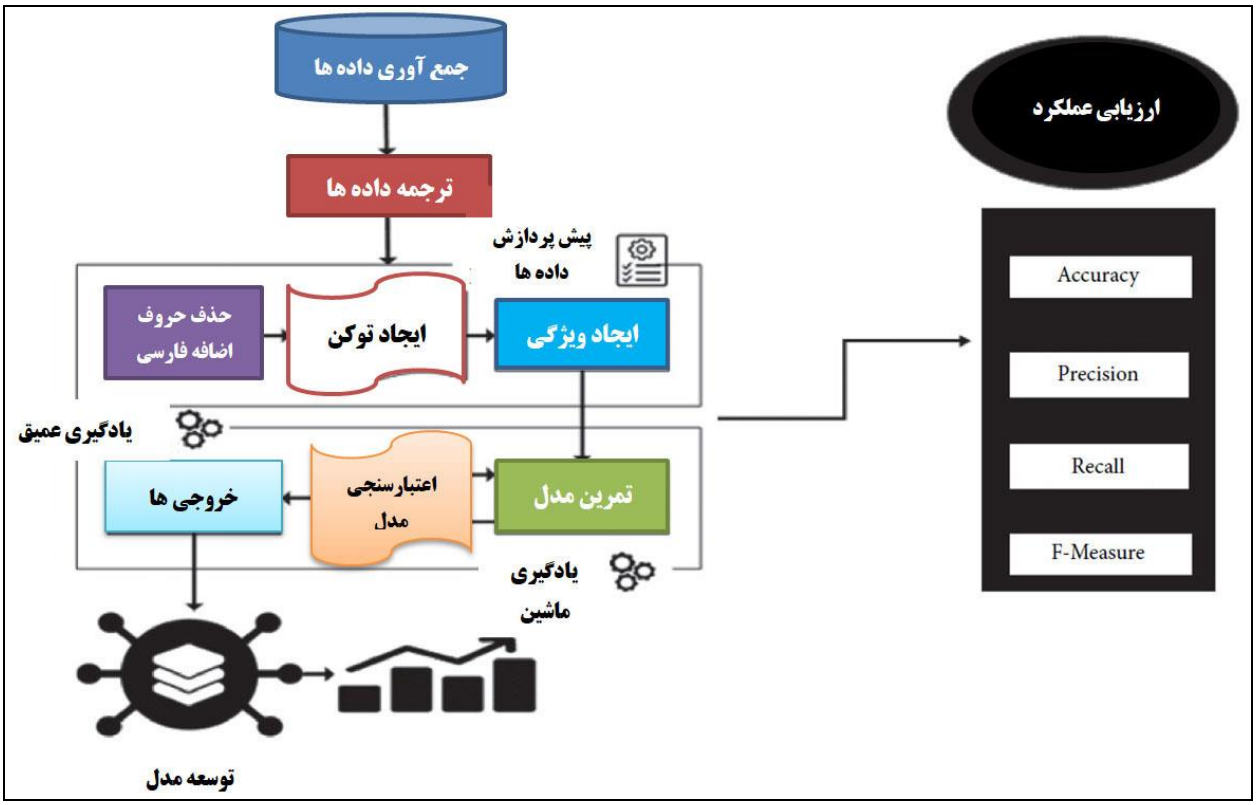
شکل ۱-چارچوب کلی روش پیشنهادی در این تحقیق

- **تبدیل داده ها به فرمت دلخواه :** بر حسب نیاز و نرم افزار مورد استفاده، داده ها به فرمت دلخواه ما تبدیل خواهند شد.
- **توکن سازی به عنوان یک مرحله حیاتی از پیش پردازش:** در این مرحله تمام کلمات ایمیل ها جمع آوری شده و تعداد دفعات ظاهر شدن هر کلمه و مکان ظاهر شدن شمارش می شود [۱۵]. با کمک Count Vectorizer، ما توانستیم تکرار کلمات را در مجموعه داده خود پیدا کنیم. به هر کلمه یک عدد منحصر به فرد داده می شود، و از این رو، آنها را نشانه نامیده می شود، همچنین وقوع آنها و تعداد وقوع آنها را نشان می دهد. توکن شامل یکی از انواع مقادیر ویژگی است که بعداً در ایجاد بردارهای ویژگی کمک خواهد کرد. در مرحله توکن سازی، به هر کلمه یک نشانه منحصر به فرد اختصاص داده می شود.
- **حذف کلمات اضافی و بدون استفاده یا غیرضروری**
- **اعمال الگوریتم Stemming :** برای تبدیل عبارات مشتق شده مجموعه داده به شکل اصلی خود[۱۶]. ابتدا، اصطلاحات پایه از پیشوندها و پسوندها حذف می شوند. در مرحله بعد، هر دو کلمه اصلاح شده یا غلط املایی با استفاده از الگوریتم Stemming به کلمات پایه یا ریشه خود تبدیل می شوند.
- **استخراج و انتخاب ویژگی:** استخراج ویژگی فرآیند تبدیل یک مجموعه داده خام بزرگ به یک قالب قابل مدیریت است. هر متغیر، ویژگی یا کلاس را می توان در طول این مرحله، بسته به مجموعه داده اصلی از مجموعه داده استخراج کرد. استخراج ویژگی یک مرحله مهم در آموزش مدل است که به تولید نتایج قابل اعتمادتر و دقیق تر

کمک می کند. در طی فرآیند استخراج ویژگی، از بین بسیاری از ویژگی های ممکن، روش انتخاب برخی از متغیرهای کلیدی که به درستی داده ها را مشخص می کنند، انتخاب ویژگی نامیده می شود[۱۶] .

- **ایجاد مدل:** مدل با استفاده از ویژگی ها یا متغیرهای انتخاب شده ساخته می شود [۱۷, ۱۸]. اگر انتخاب ویژگی به درستی انجام شود، ساخت مدل به درستی انجام خواهد شد.

می توان کل عملیات انجام شده در روش پیشنهادی برای تشخیص ایمیل های اسپم و غیراسپم را در شکل ۲ نشان داده شده است.



شکل ۲-شمای کلی روش پیشنهادی در این پایان نامه

۴. پیاده سازی و ارزیابی کارایی روش پیشنهادی

در این پژوهش، داده های خام جمع آوری شده از منبع آنلاین Kaggle به دست آمده است که برای آموزش مدل های یادگیری ماشین استفاده خواهد شد. داده ها در ابتدا به زبان انگلیسی در دسترس بودند و در قالب مقادیر جدا شده با کاما (CSV)به دست آمدند [۱۲]. علاوه بر این، مجموعه داده به‌دست‌آمده با استفاده از کتابخانه Googletrans python در زبان فارسی، که از Google Translate Ajax API استفاده می‌کند، ترجمه شد. پس از این، تصحیح دستی داده های ترجمه شده را انجام دادیم. در ادامه، از مجموعه داده‌های اسکریپتی زبان فارسی ترجمه شده خودمان استفاده کرده‌ایم که شامل ۵۰۰۰ ایمیل هرزنامه و غیرهرزنامه است .

در ادامه مجموعه داده خود را ایجاد کردیم، زیرا زبان فارسی با استفاده از الفبای انگلیسی نوشته می شود و خط فارسی بر اساس الفبای عربی است. در پایان این فرآیند، در مجموع ۵۰۰۰ ایمیل به دست آوردیم که در دو ستون فهرست شده بودند.

- ستون اول، با برچسب "نوع" دارای دو مقدار ممکن به عنوان هرزنامه یا غیرهرزنامه، قرار بود برای طبقه بندی ایمیل ها استفاده شود .

- ستون دوم دارای برچسب "متن ایمیل" است و حاوی انواع محتوای ایمیل است .
- تصمیم گرفته شد که حداکثر ۸۰ درصد از ایمیل ها برای آموزش مدل ها (تقریباً ۴۰۰۰ ایمیل) استفاده شود، در حالی که ۲۰ درصد باقی مانده برای آزمایش مدل ها به صورت جداگانه (۱۰۰۰ ایمیل) استفاده شود.

۱-۴-یش پردازش داده ها

- **پیش پردازش داده ها در یادگیری ماشین (ML)**

عبارت پیش پردازش به سازماندهی و مدیریت داده های خام قبل از استفاده از آن برای آموزش و آزمایش مدل های مختلف یادگیری اشاره دارد. به عبارت ساده‌تر، پیش‌پردازش یک رویکرد داده‌کاوی ML است که داده‌های خام را به ساختاری قابل استفاده و منابع تبدیل می‌کند.

اولین گام در ساخت یک مدلML، پیش پردازش است، که در آن داده های دنیای واقعی، معمولاً ناقص، نادقیق و به دلیل نقص و کمبود، به متغیرها و روندهای ورودی دقیق، دقیق و قابل استفاده تبدیل می شوند.

بخش ذکر شده، هر مرحله را که در مرحله پیش پردازش داده دخیل است، برجسته می کند، که همچنین به زیبایی در شکل ۲-۳ به عنوان معماری کلی پیشنهادی نشان داده شده است.

- **وارد کردن داده ها**

بعد از جمع آوری و تبدیل داده ها به فرمت CSV. آنها را در نرم افزار مورد نظر خودمان (در اینجا متلب) وارد می کنیم.

- **توکن سازی**

به عنوان یک مرحله حساس از پیش پردازش، در این مرحله، تمام کلمات ایمیل ها جمع آوری می شوند و تعداد دفعات ظاهر شدن هر کلمه و مکان ظاهر شدن شمارش می شود. با کمکCount Vectorizer، ما توانستیم تکرار کلمات را در مجموعه داده خود پیدا کنیم. به هر کلمه یک عدد منحصر به فرد داده می شود، و از این رو، آنها را نشانه نامیده می شود، همچنین وقوع آنها و تعداد وقوع آنها را نشان می دهد.

توکن شامل یکی از انواع مقادیر ویژگی است که بعداً در ایجاد بردارهای ویژگی کمک خواهد کرد. در مرحله توکن سازی، به هر کلمه یک نشانه منحصر به فرد اختصاص داده می شود.

- **حذف کلمات توقف و حروف اضافه:**

هنگامی که مجموعه داده به نشانه های منحصر به فرد تبدیل شد، گام بعدی حذف هر کلمه غیر ضروری (به عنوان مثال، فاصله سفید، کاما، نقطه نقطه، دونقطه، نقطه ویرگول، و علائم نگارشی بی معنا) بدون اهمیت است. پایتون دارای کتابخانه داخلی است که به عنوان جعبه ابزار زبان طبیعی (NLTK[3]) شناخته می شود که در پردازش زبان بسیار مورد استفاده قرار گرفته است. در اینجا از جعبه ابزار NLTK برای فرآیند حذف کلمات توقف برای حذف کلمات و فاصله های غیر ضروری استفاده کردیم .

- **Stemming**

پس از ایجاد توکن ها، گام بعدی این است که آنها را ریشه یابی کنیم. Stemming روشی برای تبدیل عبارات مشتق شده مجموعه داده به شکل اصلی خود است. ابتدا، اصطلاحات پایه از پیشوندها و پسوندها حذف می شوند. در مرحله بعد، هر دو کلمه اصلاح شده یا غلط املایی با استفاده از الگوریتم stemming به کلمات پایه یا ریشه خود تبدیل می شوند. برای این مرحله نیز، از کتابخانه پایتون NLTK برای انجام یک فرآیند بنیادی کامل استفاده کردیم. پس از ارسال ایمیل های محتوایی، کلمات هرزنامه را می توان به راحتی شناسایی کرد.

[3] Natural Language Toolkit (NLTK)

- **استخراج و انتخاب ویژگی ها:** استخراج ویژگی یک مرحله مهم در آموزش مدل است که به تولید نتایج قابل اعتمادتر و دقیق تر کمک می کند. در طی فرآیند استخراج ویژگی، از بین بسیاری از ویژگی های ممکن، روش انتخاب برخی از متغیرهای کلیدی که به درستی داده ها را مشخص می کنند، انتخاب ویژگی نامیده می شود. سپس مدل با استفاده از این ویژگی ها یا متغیرهای انتخاب شده ساخته می شود [۱۸]. اگر انتخاب ویژگی به درستی انجام شود، در عوض، ساخت مدل زمان کمتری می برد.

جدول ۱ ویژگیهای مجموعه داده های استخراج شده را نشان می دهد.

جدول ۱–ویژگیهای مجموعه داده های استخراج شده

ویژگیهای مجموعه داده	مقدار
تعداد متغیرها	۲
تعداد مشاهدات	۵۰۰۰
تعداد داده های از دست رفته	۰
درصد داده های از دست رفته	۰٪
داده های تکراری	۲۳۸
درصد داده های تکراری	۴.۸٪
اندازه کلی حافظه مورد نیاز	۷۸.۲ کیلوبایت
متوسط اندازه رکوردهای حافظه	۱۶ بایت

۴-۲-آموزش و تست داده ها

آموزش و تست داده ها در این بخش ارائه شده است.

در این پایان نامه، ۵۰۰۰ ایمیل از منبع آنلاین «kaggle» جمع‌آوری کرده ایم و با استفاده از کتابخانه پایتون Googletrans که از API Ajax Google Translate استفاده می‌کند، آنها را به فارسی ترجمه کردیم. چهار هزار ایمیل برای آموزش مدل های مختلف ML و DL استفاده شد. هزار ایمیل برای آزمایش به منظور تعیین کمیت معیارهای دقت و ارزیابی استفاده شد. پیاده سازی روش پیشنهادی با استفاده از متلب و RMiner انجام می شوند.

۴-۳-نتایج ارزیابی روش پیشنهادی

معیارهای ارزیابی شامل دقت، دقت، یادآوری و اندازه‌گیری‌های f را ارزیابی کرده‌ایم که معیارهای ارزیابی با استفاده از SVM و Naive Bayes اندازه‌گیری می‌شوند.

CNN و LSTM برای اندازه‌گیری ROC-AUC و مدل‌سازی مقادیر تلفات استفاده می‌شوند. در نهایت با استفاده از نمودارهای مختلف، مقایسه مدل ها در زیر ارائه شده است.

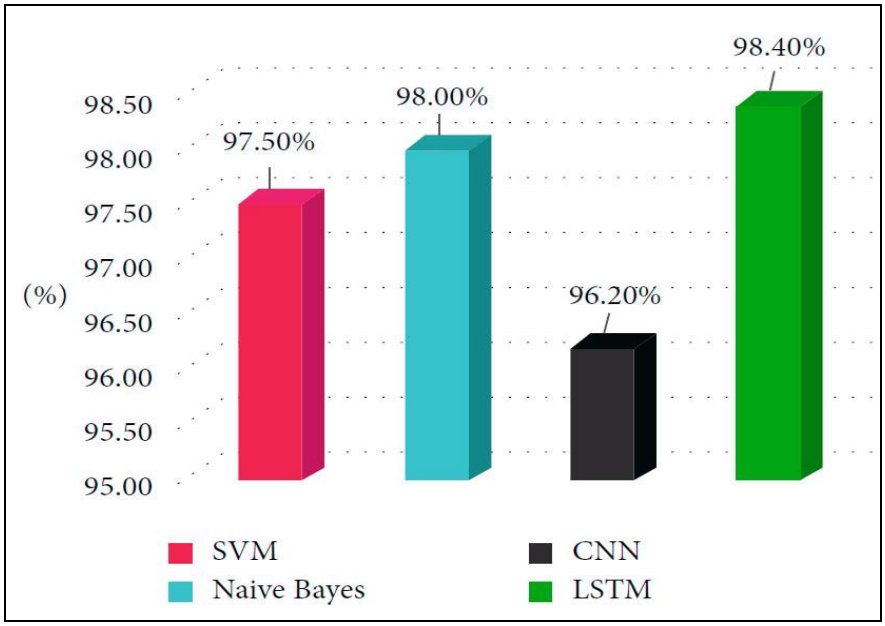
یافته های موجود در جدول ۲ نشان می دهد که الگوریتم یادگیری عمیق (LSTM) روش قوی تری برای شناسایی ایمیل های هرزنامه اردو با دقت بالای ۹۸.۴ درصد است.

جدول ۲–دقت مدلهای استفاده شده در این تحقیق

مدلها	دقت
LSTM	98.4%
CNN	96.2%
Naïve Bayes	98%
SVM	97.5%

در جدول ۲ مشاهده می کنیم که، دقت چهار مدل مختلف ML و DL را با هم مقایسه کرده ایم. می‌توانیم ببینیم که مدل LSTM در بین همه مدل‌ها دقیق‌ترین است، اما آموزش آن به زمان زیادی نیاز دارد. مدل‌های ML مانند SVM و Naive Bayes تقریباً درصد دقت کمتری نسبت به LSTM/CNN دارند که همچنین یک مدل DL است و کمترین درصد دقت را دارد.

شکل ۳ مقایسه دقت مدل های ML و DL را نشان می دهد. مدل‌های ML (یعنی SVM و Naive Bayes) برای محاسبه پارامترهای ارزیابی مانند دقت، فراخوان، و اندازه‌های f که در جدول ۳ توضیح داده شده‌اند استفاده می‌شوند و موفق بوده و نتایج بهتری به همراه دارند. با این حال، SVM بالاترین درصد دقت را در مقایسه با Naive Bayes ایجاد می کند.



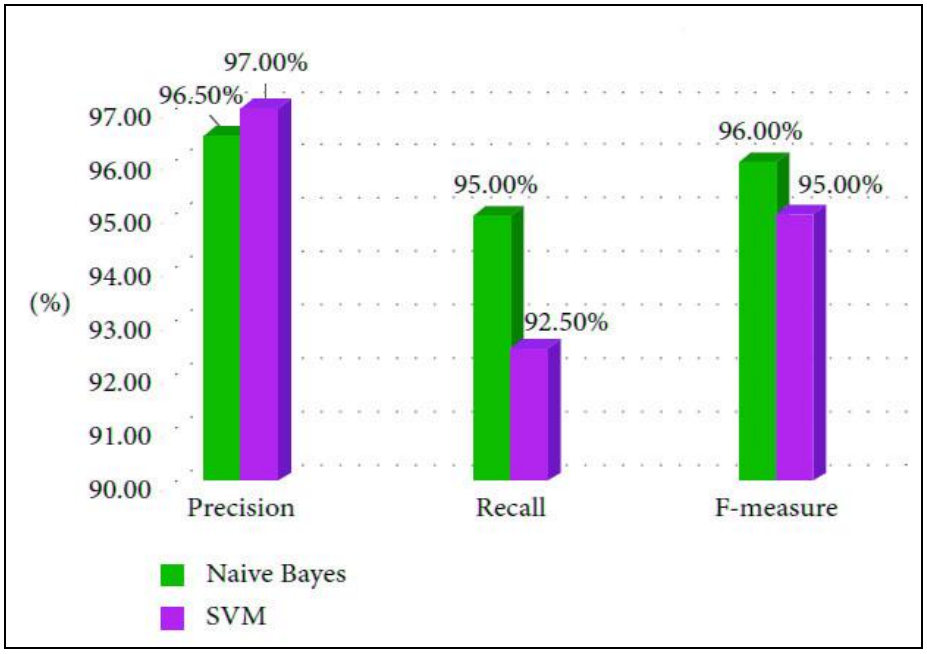
شکل ۳-مقایسه دقت مدلهای ML و DL

نتایج تجزیه و تحلیل مقایسه ای ارائه شده در جدول ۳ نشان می دهد که Naive Bayes از نظر فراخوانی و اندازه گیری f به نتایج بهتری دست می یابد، در حالی که SVM با توجه به دقت به نتایج بهتری دست می یابد.

جدول ۳-مقادیر پارامترهای ارزیابی مدلهای **ML**

مدلهای یادگیری ماشین	صحت%	Recall %	F-Measure %
	۹۶.۵٪	۹۵٪	۹۶٪
	۹۷٪	۹۲٪	۹۵٪

مقایسه نتایج بدست آمده توسط Naive Bayes و SVM به صورت بصری در شکل ۴ نشان داده شده است.



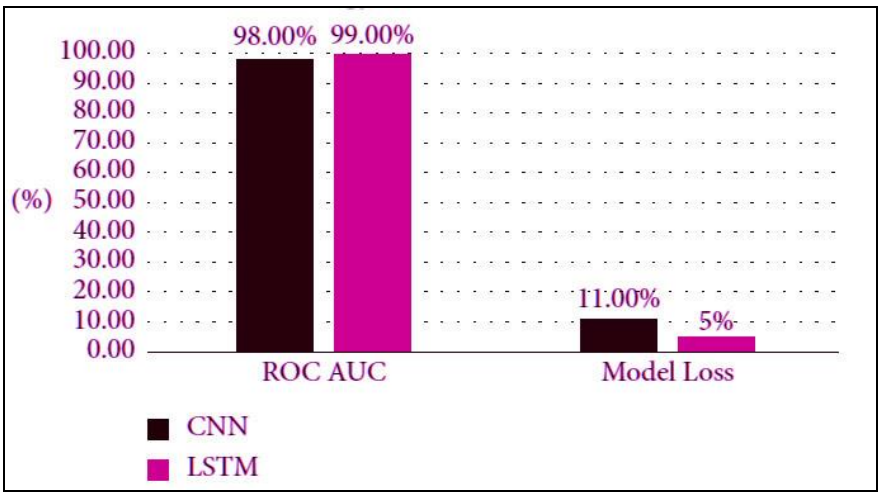
شکل ۴- مقایسه نتایج بدست آمده توسط Naive Bayes و SVM به صورت بصری

همانطور که در جدول ۴ نشان داده شده است، در مقایسه با CNN، LSTM دارای درصد بیشتری از ROC-AUC و نرخ تلفات مدل کمتری است. در نهایت، کل یافته ها با هم مقایسه شدند. ما دریافتیم که LSTM دارای دقت و مقدار ROC-AUC بیشتر و نرخ تلفات مدل بسیار پایین است. در مقایسه با CNN، یافته های مطالعه مقایسه ای نشان می دهد که مدل LSTM نتایج ROC-AUC و درصد از دست دادن داده های بهتری را تولید می کند. در مدل LSTM درصد از دست دادن داده ها کمتر از ۵٪ و ROC-AUC بالای ۹۹٪ دارد.

جدول ٤-**ROC-AUC** و مقادیر از دست دادن داده ها در مدل‌های **ML**

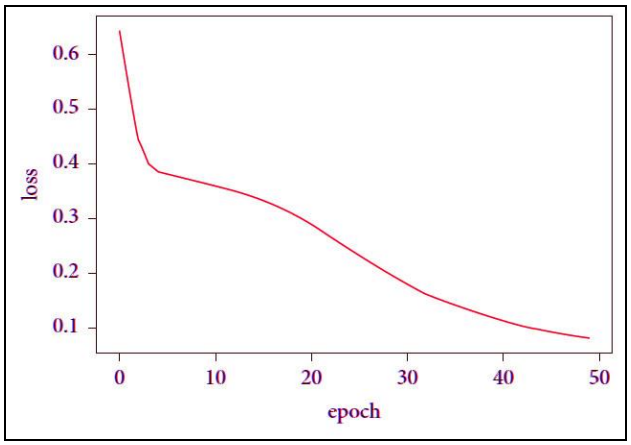
مدلهای یادگیری ماشین	ROC-AUC%	% از دست دادن داده ها
	99%	5
	98%	11

شکل ۵ یک نمایش گرافیکی از نتایج بدست آمده توسط CNN و LSTM را نشان می دهد.

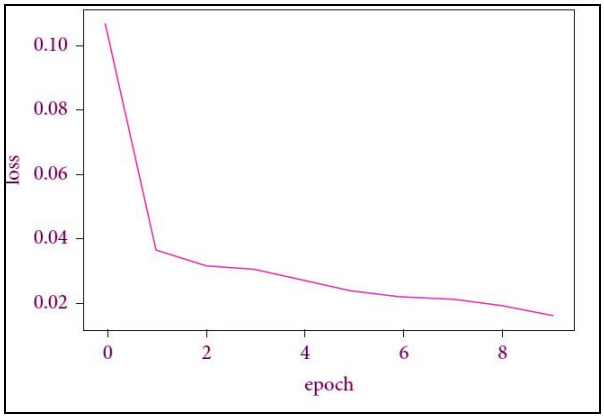


شکل۵-نمودار مقایسه ROC-AUC و تلفات داده ها

نمودارهای شکل ۶ و ۷ میزان از دست دادن داده ها را برای مدل CNN و LSTM را برای هر دوره نشان می دهد.



شکل ۶- نمودار از دست دادن داده ها برای شبکه های عصبی کانولوشن با استفاده از روش پیشنهادی



شکل ۷- نمودار از دست دادن داده ها برای LSTM با استفاده از روش پیشنهادی

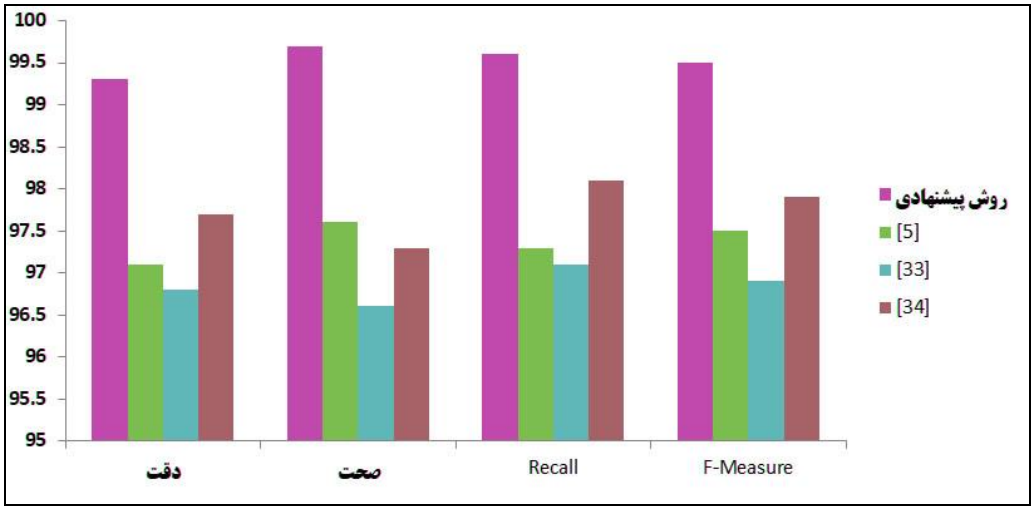
همانطور که مشاهده می شود، خط نمودار آشکارا با افزایش دوره ها کاهش می یابد. هنگامی که تعداد دوره ها افزایش می یابد، نرخ تلفات داده ها هم کاهش می یابد. این اشکال نشان می‌دهد که مدل‌های DL به‌ویژه در شناسایی و طبقه‌بندی ایمیل‌های هرزنامه های به زبان فارسی خوب هستند، زیرا تشخیص دقیق‌تری دارند. در این مطالعه از مدل‌های موجود برای

شناسایی ایمیل‌های هرزنامه فارسی استفاده کردیم و آموزش‌های بیشتر و تشخیص بهتر نیز برای SVM، Naive Bayes، CNN و LSTM توضیح داده شد.

علاوه بر این، دقت هر مدل محاسبه شد و معیارهای ارزیابی مانند دقت، فراخوان و اندازه‌گیری f برای SVM و Naive Bayes و برای CNN و LSTM و همچنین اندازه‌گیری‌های ROC-AUC و از دست دادن مدل برای ارزیابی مقایسه‌ای استفاده شد. بر اساس یافته‌ها، مدل LSTM با امتیاز ۹۸.۴ درصد دقت بالاتری نسبت به سایر مدل‌ها کسب کرد.

۴-4-مقایسه روش پیشنهادی با روشهای دیگر

شکل ۸ دقت، صحت، Recall و F-Measure بدست آمده از روش پیشنهادی را با سه روش دیگر پیشنهادی توسط محققان دیگر مقایسه می کند [۵, 19, 20] .



شکل ۸-مقایسه کارایی روش پیشنهادی با روشهای دیگر

همانطور که در شکل ۸ مشاهده می شود از لحاظ دقت، صحت، Recall, F-Measure روش پیشنهادی در مقایسه با روشهای محققان دیگر بسیار بهتر می باشد. لذا می توان این روش را برای شناسایی اسپم ها در ایمیلهای نوشته شده به زبان فارسی به کار برد.

۵. نتیجه گیری

با افزایش استفاده از اینترنت و شبکه های اجتماعی، ارتباطات آنلاین به بخشی ضروری از زندگی روزمره ما تبدیل شده است. یکی از رسانه های پرکاربرد برای ارتباطات رسمی و تجاری، پست الکترونیکی است که به دلیل دسترسی آزاد، سرعت و قابلیت اطمینان آن است. با افزایش محبوبیت ایمیل، ایمیل‌های هرزنامه به سرعت در حال افزایش بودند، هرزنامه یک یا چند پیام ناخواسته است که به شکل تبلیغات یا مواد تبلیغاتی مانند طرحهای کاهش بدهی، طرحهای سریع ثروتمند شدن، دوستیابی آنلاین و محصولات مرتبط با سلامتی و غیره است. ارائه روشهای خودکار برای تشخیص هرزنامه موضوع جدیدی نیست. کسب و کار و شرکت ها همیشه به دنبال بهبود تجربه کاربران خود هستند. برای محافظت از سرورهای خود در برابر آسیب های احتمالی؛ در صورتی که ویروس ها در داخل ایمیل های اسپم پیچیده شده باشند، بسیار علاقمند هستند. علاوه بر این شناسایی هرزنامه ها به صرفه جویی در منابع شبکه (پهنای باند) و زمان از هدر رفتن کمک می کند. بسته به نوع داده ها و وظیفه ای که به آنها داده می شود، رویکردهای طبقه بندی مختلف ممکن است بدتر یا بهتر عمل کنند. با کمک یادگیری ماشین و یادگیری عمیق برای درک زبان طبیعی می توان به شناسایی زودهنگام هرزنامه ها می توان کمک کرد. در روش پیشنهادی از شبکه عصبی کانولوشنال (CNN) برای طبقه بندی جملات استفاده می کنیم و LSTM دو طرفه مبتنی بر توجه برای طبقه‌بندی رابطه‌ای و تحلیل احساسات مبتنی بر موضوع می باشد. روش پیشنهادی با استفاده از داده های آنلاین Kaggle پیاده سازی شده و مورد ارزیابی قرار گرفت و نتایج بدست آمده عملکرد بالای آن را در شناسایی صحیح و بادقت هرزنامه ها در مقایسه با کارهای دیگر انجام شده توسط محققان دیگر، نشان داد.

۱۲. منابع و مراجع

۱. Pouyanfar, S., et al., *A survey on deep learning: Algorithms, techniques, and applications.* ACM Computing Surveys (CSUR), ۲۰۱۸. ۵۱(۵): p. ۱-۳۶.

۲. Alom, M.Z., et al., *A state-of-the-art survey on deep learning theory and architectures.* Electronics, ۲۰۱۹. ۸(۳): p. ۲۹۲.

۳. Dong, S., P. Wang, and K. Abbas, *A survey on deep learning and its applications.* Computer Science Review, ۲۰۲۱. ۴۰: p. ۱۰۰۳۷۹.

۴. Zhang, Q., et al., *A survey on deep learning for big data.* Information Fusion, ۲۰۱۸. ۴۲: p. ۱۴۶-۱۵۷.

۵. Jain, G., M. Sharma, and B. Agarwal, *Spam detection on social media using semantic convolutional neural network.* International Journal of Knowledge Discovery in Bioinformatics (IJKDB), ۲۰۱۸. ۸(۱): p. ۱۲-۲۶.

۶. Sudanagunta, S., *SMS Spam Detection Framework Using Machine Learning Algorithms*. ۲۰۲۰, Southeast Missouri State University.

۷. Liu, X., H. Lu, and A. Nayak, *A spam transformer model for SMS spam detection.* IEEE Access, ۲۰۲۱. ۹: p. ۸۰۲۶۳-۸۰۲۵۳.

۸. Sethi, P., V. Bhandari, and B. Kohli. *SMS spam detection and comparison of various machine learning algorithms*. in ۲۰۱۷*international conference on computing and communication technologies for smart nation (IC۳TSN)*. ۲۰۱۷. IEEE.

۹. Popovac, M., et al. *Convolutional neural network based SMS spam detection*. in ۲۶ ۲۰۱۸*th Telecommunications Forum (TELFOR)*. ۲۰۱۸. IEEE.

10. Wei, F .and T. Nguyen. *A lightweight deep neural model for sms spam detection*. in ۲۰۲۰ *International Symposium on Networks, Computers and Communications (ISNCC)*. ۲۰۲۰. IEEE.

۱۱. Suleiman, D. and G. Al-Naymat, *SMS spam detection using H۲O framework.* Procedia computer science, ۲۰۱۷. ۱۱۳: p. ۱۵۴-۱۶۱.

۱۲. Kumar, N. and S. Sonowal. *Email spam detection using machine learning algorithms*. in ۲۰۲۰ *Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. ۲۰۲۰. IEEE.

۱۳. Kipkebut, A., M .Thiga, and E. Okumu, *Machine Learning Sms Spam Detection Model*. ۲۰۱۹.

۱۴. Xia, T. and X. Chen, *A discrete hidden Markov model for SMS spam detection.* Applied Sciences, ۲۰۲۰. ۱۰(۱۴): p. ۵۰۱۱.

۱۵. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, ۲۰۱۲. ۲۵.

۱۶. Chen, X.-l., et al. *A method of spam filtering based on weighted support vector machines*. in ۲۰۰۹*IEEE International Symposium on IT in Medicine & Education*. ۲۰۰۹. IEEE.

۱۷. Sharma, P. and U. Bhardwaj, *Machine learning based spam e-mail detection.* International Journal of Intelligent Engineering and Systems, ۲۰۱۸. ۱۱(۳): p. ۱-۱۰.

۱۸. Jameel, N.G.M. and L.E. George, *Detection of phishing emails using feed forward neural network.* International Journal of Computer Applications, ۲۰۱۳. (۷)۷۷.

19. Yaseen, Q., *Spam email detection using deep learning techniques.* Procedia Computer Science, ۲۰۲۱. ۱۸۴: p. ۸۵۳-۸۵۸.

۲۰. Rathod, S.B. and T.M. Pattewar. *Content based spam detection in email using Bayesian classifier*. in ۲۰۱۵*International Conference on Communications and Signal Processing (ICCSP)*. ۲۰۱۵. IEEE.

Proposing a New Method to Identify Spams in SMS

Author1, Author 2, Author 3, Author 4 - Font :Times New Roman 12
Title and address of the first author , Email
Title and address of the second author, Email
Title and address of the third author , Email

Abstract— In recent years, the Internet has become an integral part of our human lives. With the increasing use of the Internet, the number of email users is increasing day by day. This increasing use of e-mail has created problems caused by unsolicited bulk e-mail messages, commonly referred to as spam. Email has now become one of the best ways to advertise, which is why spam emails are generated. Spam emails are emails that the recipient does not want to receive. Too many identical messages are sent to multiple email recipients. Spam is usually generated as a result of providing our email address on an unauthorized or illegal website. There are many effects of spam. It fills our inbox with tons of useless emails. It slows down our internet speed to a great extent. It steals useful information. Therefore, in this research, we propose a new method to identify spam emails written in Persian language. In this method, we have used a combination of machine learning and deep learning techniques to identify email spam. We used the proposed method using Kaggle online data set. The obtained results showed that the proposed method has a better performance in terms of evaluation criteria (accuracy, accuracy, recall and F-Measure) compared to the methods proposed by other researchers.

Keywords: spam, SMS, Kaggle online dataset, machine learning techniques, deep learning techniques.