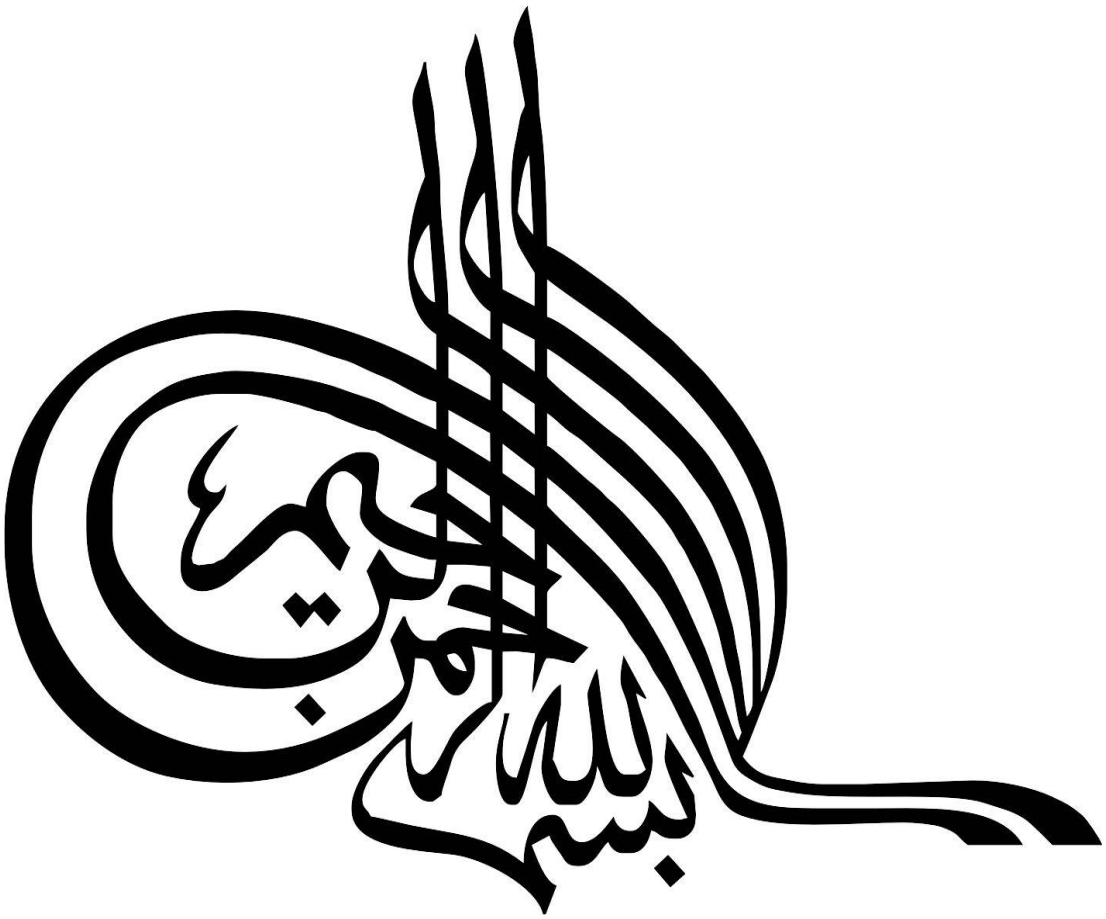
بسم الله الرحمن الرحيم

گروه کامپیوتر

پایان‌نامه برای دریافت درجه کارشناسی ارشد (M.Sc)

گرایش: نرم‌افزار

عنوان:

ارائه روشی جهت تشخیص هرزنامه مبتنی بر الگوریتم‌های جستجوی ممنوعه و ماشین بردار پشتیبان

گروه کامپیوتر

پایان‌نامه برای دریافت درجه کارشناسی ارشد (M.Sc)

گرایش: نرم‌افزار

عنوان:

ارائه روشی جهت تشخیص هرزنامه مبتنی بر الگوریتم‌های جستجوی ممنوعه و ماشین بردار پشتیبان

تاییده پایان نامه توسط هیات داوران (پس از تأیید پایان نامه):

نام و نام خانوادگی استاد راهنما:	تاریخ	امضا
نام و نام خانوادگی استاد داور ۱:	تاریخ	امضا
نام و نام خانوادگی استاد داور ۲:	تاریخ	امضا

تعهد نامه اصالت رساله دکتری یا پایان نامه کارشناسی ارشد دانشگاه آزاد اسلامی

اینجانب ..... دانش آموخته مقطع کارشناسی ارشد ناپیوسته در رشته **مهندسی کامپیوتر-نرم‌افزار** که در تاریخ ..... از پایان نامه / رساله خود تحت عنوان**"ارائه روشی جهت تشخیص هرزنامه مبتنی بر الگوریتم‌های جستجوی ممنوعه و ماشین بردار پشتیبان"** با کسب نمره ..... و درجه..... دفاع نموده‌ام بدینوسیله متعهد می شوم :

۱ – این پایان نامه / رساله حاصل تحقیق و پژوهش انجام شده توسط اینجانب بوده و در مواردی که از دستاوردهای علمی و پژوهشی دیگران (اعم از پایان نامه، کتاب، پژوهش و .....) استفاده نموده‌ام، مطابق ضوابط و رویه موجود، نام منبع مورد استفاده و سایر مشخصات آن را در فهرست مربوطه ذکر و درج کرده‌ام.

۲ – این پایان نامه/ رساله قبلاً برای دریافت هیچ مدرک تحصیلی (هم سطح، پائین تر یا بالاتر) در سایر دانشگاه‌ها و موسسات آموزش عالی ارائه نشده است.

۳ – چنانچه بعد از فراغت از تحصیل، قصد استفاده و هرگونه بهره برداری اعم چاپ کتاب، ثبت اختراع و ... از این پایان نامه را داشته باشم، از حوزه معاونت پژوهشی واحد مجوزه‌های مربوطه را اخذ نمایم.

۴ – چنانچه در هر مقطع زمانی خلاف موارد فوق ثابت شود، عواقب ناشی از آن را می پذیرم و واحد دانشگاهی مجاز است با اینجانب مطابق ضوابط و مقررات رفتار نموده و در صورت ابطال مدرک تحصیلی‌ام هیچگونه ادعایی نخواهم داشت.

شیرین پیری

تاریخ و امضاء

معاونت پژوهش و فن آوری

به نام خدا

منشور اخلاق پژوهش

باياری از خداوند سبحان و اعتقاد به اینکه عالم محضر خدا است و همواره ناظر بر اعمال انسان و به منظور پاسداشت مقام بلند دانش و پژوهش و نظر به اهمیت جایگاه دانشگاه در اعتلای فرهنگ و تمدن بشری ما دانشجویان و اعضاء هیئت علمی واحدهای دانشگاه آزاد اسلامی متعهد می گردیم اصول زیر را در انجام فعالیت های پژوهشی مد نظر قرار داده و از آن تخطی نکنیم :

۱-اصل حقیقت جویی: تلاش در راستای پی جویی حقیقت و وفاداری به آن و دوری از هرگونه پنهان سازی حقیقت

۲-اصل رعایت حقوق: التزام به رعایت کامل حقوق پژوهشگران و پژوهیدگان (انسان، حیوان و نبات) و سایر صاحبان حق

۳-اصل مالکیت مادی و معنوی: تعهد به رعایت کامل حقوق مادی و معنوی دانشگاه و کلیه همکاران پژوهش

۴-اصل منافع ملی: تعهد به رعایت مصالح ملی و در نظر داشتن پیشبرد و توسعه کشور در کلیه مراحل پژوهش

۵-اصل رعایت انصاف و امانت: تعهد به اجتناب از هرگونه جانب داری غیر علمی و حفاظت از اموال، تجهیزات و منابع در اختیار

۶-اصل رازداری: تعهد به صیانت از اسرار و اطلاعات محرمانه افراد، سازمان ها و کشور و کلیه افراد و ستادهای مرتبط با تحقیق

۷-اصل احترام: تعهد به رعایت حریم ها و حرمت ها در انجام تحقیقات و رعایت جانب تعدد و خودداری از هرگونه حرمت شکنی

۸-اصل ترویج: تعهد به رواج دانش و اشاعه نتایج تحقیقات و انتقال آن به همکاران علمی و دانشجویان به غیر از مواردی که منع قانونی دارد

۹-اصل برائت: التزام به برائت جویی از هرگونه رفتار غیر حرفه ای و اعلام موضع نسبت به کسانی که حوزه علم و پژوهش را به شائبه های غیر علمی می آلایند.

فهرست مطالب

عنوان **صفحه**

چکیده ۱

فصل اول: کلیات تحقیق

۱-۱- مقدمه ۳

۱-۲- بیان مسئله ٤

۱-۳- ضرورت انجام تحقیق..... ٦

۱-٤- اهداف تحقیق ٦

۱-۵- جنبه نوآوری ۷

۱-٦- ساختار پایان‌نامه ۷

فصل دوم: مبانی تحقیق

۲-۱- تشخیص هرزنامه ۹

۲-۱-۱- دسته‌های مختلف فیلتر کردن هرزنامه ۱۲

۲-۱-۲- تشخیص هرزنامه بر اساس رفتار ۱۳

۲-۱-۳- تشخیص ایمیل‌های عادی از هرزنامه‌ها بر اساس روش‌های رفتار محور..... ۱٤

۲-۱-۳-۱- لیست سیاه..... ۱٤

۲-۱-۳-۲- لیست سفید..... ۱۵

۲-۱-۳-۳- لیست خاکستری ۱۵

۲-۱-٤- تشخیص ایمیل‌های عادی از هرزنامه‌ها بر اساس روش‌های هوشمند ۱٦

۲-۱-٤-۱- الگوریتم درخت تصمیم ۱٦

۲-۱-٤-۲- الگوریتم نایوبیز ۱٦

۲-۱-٤-۳- الگوریتم k نزدیک‌ترین همسایه ۱۷
۲-۱-٤-٤- الگوریتم جنگل تصادفی ۱۷
۲-۱-٤-٥- الگوریتم ماشین بردار پشتیبان ۱۷
۲-۱-٤-٦- الگوریتم جستجوی ممنوعه ۱۹

فصل سوم: پیشینه تحقیق

۳-۱- روش‌های ارائه شده جهت شناسایی هرزنامه ۲۳
۳-۱-۱- روش‌های ارائه شده جهت شناسایی هرزنامه با استفاده از روش‌های یادگیری ماشین ۲۳
۳-۱-۲- روش‌های ارائه شده جهت شناسایی هرزنامه با استفاده از روش‌های بهینه سازی فرامکاشفه‌ای ۳۱
۳-۱-۳- روش‌های ارائه شده جهت شناسایی هرزنامه با استفاده از روش‌های مبتنی بر محتوا ۳۹
۳-۱-٤- روش‌های ارائه شده جهت شناسایی هرزنامه برای مجموعه داده‌های بزرگ ٤۰
۳-۱-٥- روش‌های ارائه شده جهت تشخیص هرزنامه گروهی در اینترنت اشیاء ٤۱

فصل چهارم: روش پیشنهادی

٤-۱- کلیات روش پیشنهادی ٤۷
٤-۲- مراحل روش پیشنهادی ٤۸
٤-۲-۱- فاز اول روش پیشنهادی: انتخاب ویژگی با استفاده از الگوریتم جستجوی ممنوعه ٤۹
٤-۲-۲- فاز دوم روش پیشنهادی: دسته‌بندی با استفاده از ماشین بردار پشتیبان ٥۳
٤-۲-۲-۱- مرحله آموزش ٥۳
٤-۲-۲-۲- مرحله آزمایش ٥۷

فصل پنجم: مقایسه و ارزیابی

٥-۱- بستر شبیه‌سازی ٦۰
٥-۲- پارامترهای ارزیابی ٦۰

۵-۳- مجموعه داده ٦١
۵-٤- تابع تصمیم مورد استفاده در شبیه‌سازی..... ٦٣
۵-۵- نتایج روش پیشنهادی ٦٣
۵-٦- مقایسه نتایج روش پیشنهادی ٦٥

نتیجه‌گیری و پیشنهادات آتی

نتیجه‌گیری ٦٩
پیشنهادات آتی ۷۰
منابع ۷۱
چکیده انگلیسی ۷۳

فهرست جدول‌ها

عنوان **صفحه**

جدول (۱-۳): مروری بر روش‌های پیشین ٤٢

جدول (۱-۵): تعریف مجموعه داده مورد استفاده در شبیه‌سازی ٦٢

جدول (۲-۵): نتایج روش پیشنهادی با توجه به پارامترهای ارزیابی ٦٤

فهرست نمودارها

عنوان **صفحه**

نمودار (۱-۵): نتایج مرحله انتخاب ویژگی روش پیشنهادی..... ٦٤

نمودار (۲-۵): مقایسه روش پیشنهادی با توجه به پارامترهای ارزیابی..... ٦٥

فهرست شکل‌ها

عنوان **صفحه**

شکل (۱-۲): ابر صفحه‌ی موجود و حاشیه‌ها..... ۱۹

شکل (۲-۲): نمودار جریان الگوریتم جستجوی ممنوعه..... ۲۰

شکل (۱-۳): نوع هرزنامه و ترکیب آن‌ها..... ۲۷

شکل (۲-۳): تشخیص هرزنامه ایمیل از طریق الگوریتم انتخاب منفی و تکامل تفاضلی..... ۳۳

شکل (۳-۳): معماری الگوریتم ترکیبی انتخاب منفی و بهینه‌سازی ازدحام..... ۳۶

شکل (٤-۳): الگوریتم انتخاب منفی و مگس میوه برای شناسایی هرزنامه‌های ایمیل..... ۳۸

شکل (٥-۳): تشخیص هرزنامه تصویری با استفاده از ماشین بردار پشتیبان..... ٤۱

شکل (۱-٤): مراحل روش پیشنهادی..... ٤۹

شکل (۲-٤): فلوچارت فاز اول جهت انتخاب ویژگی با استفاده از الگوریتم جستجوی ممنوعه..... ٥۲

چکیده

امروزه، افراد بسیاری به اینترنت دسترسی داشته‌اند و نمی‌توانند بدون تلفن هوشمند و کامپیوترها زندگی کنند. از این رو، استفاده از اینترنت به سرعت رشد یافته است. یکی از تهدیدات این نوع فناوری، هرزنامه می‌باشد. هرزنامه، یک ایمیل/پیام نخواسته می‌باشد. همچنین هرزنامه ایمیل، آن دسته از ایمیل‌های ناخواسته یا تقاضا نشده بوده که برای یک گیرنده خاص نوشته نشده‌اند. ایمیل، رایج‌ترین رسانه مورد استفاده برای دنیای ارتباطات است، به این دلیل که ارزان، قابل‌اعتماد، سریع و به راحتی قابل‌دسترس می‌باشد. ایمیل می‌تواند از نوع هرزنامه و یا نرمال باشد. هرزنامه‌های ایمیل یکی از مشکلات اصلی اینترنت امروزی است که خسارات مالی به شرکت‌ها و کاربران وارد می‌آورد. در این پژوهش، روشی جهت تشخیص هرزنامه‌ها که مبتنی بر الگوریتم‌های جستجوی ممنوعه و ماشین بردار پشتیبان می‌باشد، ارائه گردید. در روش پیشنهادی، جهت تشخیص هرزنامه، دو فاز اصلی وجود داشت. فاز اول روش پیشنهادی، استفاده از الگوریتم جستجوی ممنوعه برای انتخاب ویژگی و فاز دوم روش پیشنهادی، استفاده از روش یادگیری ماشین برای دسته‌بندی بود. پیاده‌سازی مدل پیشنهادی، با استفاده از کدهای نوشته شده در نرم‌افزار متلب انجام شده است. دقت روش پیشنهادی، به ۰,۹۳۸ رسیده است.

کلمات کلیدی: تشخیص هرزنامه، هرزنامه ایمیل، اینترنت، الگوریتم جستجوی ممنوعه، الگوریتم ماشین بردار پشتیبان

فصل اول

کلیات تحقیق

۱-۱- مقدمه

هرزنامه، نامه‌های الکترونیکی است که امکان دارد با اهداف متفاوتی به صندوق پستی کاربران مختلف، ارسال گردند. هرزنامه‌ها، برای افراد بسیاری آزاردهنده می‌باشند و منجر به اتلاف زمان و افزایش تأخیرهای ارتباطی می‌شوند. درصورتی‌که ارسال هرزنامه می‌تواند بدون محدودیت و با حداقل هزینه انجام گیرد، هزینه ناشی از ترافیک دریافت هرزنامه امکان دارد برای فراهم‌کننده‌های سرویس پستی و فراهم‌کننده‌های سرویس اینترنت بسیار زیاد باشد (فردین پور، ۱۳۹۵).

از اصلی‌ترین شاخه‌های هوش مصنوعی، یادگیری ماشین است. به درک کردن ماهیت یک فرآیند با بررسی ساختار آن حاوی ورودی، خروجی و رفتار فرآیند، یادگیری ماشین گفته می‌شود. بر مبنای یادگیری ماشین، پیش‌بینی خروجی سیستم به ازای هر ورودی امکان‌پذیر است. الگوریتم‌های یادگیری ماشین، جهت انتخاب ویژگی و دسته‌بندی هرزنامه‌ها مورد استفاده قرار می‌گیرند. تاکنون، الگوریتم‌های دسته‌بندی بسیاری ارائه گشته‌اند که می‌توان به روش شبکه عصبی مصنوعی اشاره نمود. شبکه عصبی مصنوعی، یک سیستم پردازش داده‌ها می‌باشد که از مغز انسان ایده می‌گیرد و پردازش داده‌ها را به عهده تعداد زیادی پردازنده‌های کوچک می‌گذارد که به‌طور شبکه‌ای به هم پیوسته و موازی با هم رفتار می‌نمایند تا یک مشکل را حل کنند (دالملون[1]، ۲۰۱۰).

یادگیری، از مسائلی است که به زمینه هوش بسیار نزدیک می‌باشد. در حقیقت بدون وجود یادگیری، هوش نمی‌تواند موجود باشد؛ چون یادگیری، ابزار دریافت دانش‌های نوین محسوب می‌گردد. یادگیری، انسان را قادر

[1] Dalmolen

می‌سازد که تجارب و حوادثی که برای خودشان پدید آمده است را به کار گیرند. قابلیت یادگیری، ابزاری قدرتمند به شمار می‌رود؛ به همین دلیل هدف بسیاری از برنامه نویسان، ساختن برنامه‌ای است که بتواند این ابزار را به همان ترتیبی که انسان‌ها به کار می‌گیرند، استفاده کنند. برنامه‌ای که قادر به انجام این کار است، از نظر تئوری آخرین برنامه‌ای می‌باشد که نوشته می‌شود؛ چون این برنامه، قادر به یادگیری آسان وظایف گوناگون خود است. فن‌آوری نوین اطلاعات که سالانه به تولید بسیاری کامپیوتر قدرتمندتر و تازه‌تر، منجر می‌گردد؛ در حال حاضر، گردآوری، انتقال، ادغام و ذخیره حجم بسیاری از اطلاعات را با هزینه بالا، امکان‌پذیر نموده است.

۱-۲- بیان مسئله

جهت تنظیم معیارهای شناسایی هرزنامه، هر روش، به مجموعه آموزشی بستگی دارد. کاربران، قادرند کلمات را از دستگاه‌های شخصی یا حساب‌های خود به مجموعه آموزش بیفزایند تا برنامه آموزش جهانی را توسعه بخشند و نتایج مطلوبی استخراج کنند. ارائه‌دهنده خدمات، به کاربر حق انتخاب می‌دهد؛ به خاطر اینکه مجموعه آموزش جهانی را انتخاب نماید یا خیر. این مجموعه جهانی، قادر به نظم دادن ایمیل‌ها با توجه به فرکانس است. هرزنامه ایمیل که برای بسیاری از مشتریان فرستاده می‌شود، حاوی این مجموعه آموزشی می‌باشد؛ درصورتی‌که هرزنامه که دارای نرخ فراوانی (فرکانس) پایینی است، فقط به تعداد اندکی از مشتریان ارسال می‌گردد و قادرند برای مجموعه آموزشی اختیاری باشند. ارائه‌دهنده خدمات، قادر به افزودن کنترل‌کننده‌ای با حساسیت به این مجموعه آموزشی است تا با حساسیت زیاد همه هرزنامه ایمیل‌ها را برای آموزش مجموعه داده جهانی بدون مدنظر قرار دادن فرکانس آن، داشته باشد. حساسیت کم، تنها هرزنامه ایمیل با فرکانس بالا در مجموعه آموزش خواهد داشت، بدین‌سان باعث سهولت کار کاربران جهت طبقه‌بندی پست الکترونیکی در مجموعه آموزشی می‌گردد. نخست، بایستی همه شیوه‌های طبقه‌بندی در جداسازی هرزنامه ایمیل از ایمیل‌های دیگر، پیش از استفاده آن‌ها، آموزش داده شود. جهت آموزش این شیوه‌ها، مجموعه داده‌ای به نام مجموعه آموزش به کار رفته است. هزاران نمونه در این مجموعه آموزشی به کار می‌رود و طبقه‌بندی کننده را قادر به جداسازی هرزنامه ایمیل

می‌سازد؛ ولی حتی بعد از این عمل، بسیاری از هرزنامه ایمیل همان‌طور برقرار می‌مانند؛ چون هر روز یک نوع جدیدی از هرزنامه ایمیل معرفی می‌گردد؛ لذا، چنانچه پست الکترونیکی هرزنامه قدیمی، مرتب‌سازی و نشانه‌گذاری شود، یک نمونه تازه همچنان وارد می‌گردد (سینگ[1]، ۲۰۱۸). برای اینکه بتوان پیغام‌های هرز را از پیغام‌های سالم جدا کرد، مراحل زیر ضروری می‌باشد:

- به وجود آوردن یک بانک اطلاعاتی از کلمات مشابه که در هرزنامه‌های مختلف تکرار گشته‌اند. در هر پیغام هرز، کلماتی وجود دارند که بسیار تکرار شده‌اند، بعلاوه می‌توان به‌کارگیری آن‌ها را در سایر پیغام‌ها مشاهده نمود. بعد از پژوهش‌های گوناگون یک بانک اطلاعاتی کوچک از این کلمات را می‌توان در فایل یافت.
- هنگامی که یک پیغام دریافت گردید، نرم‌افزار باید متن داخل آن را بخواند و کلمات آن را با بانک اطلاعاتی مورد مقایسه قرار دهد. چنانچه در این پیغام، کلمات مشابه به کار رفته باشد، کدی منحصربه‌فرد جایگزین آن، خواهد شد. در غیر این صورت پیغام، سالم است.

در این تحقیق، یک روش برای شناسایی هرزنامه‌ها که مبتنی بر الگوریتم‌های جستجوی ممنوعه و ماشین بردار پشتیبان است، ارائه خواهد شد. برای شناسایی هرزنامه، دو فاز مهم در نظر گرفته شده است. فاز اول روش پیشنهادی، به‌کارگیری روش‌های جستجوی پیشرفته به منظور انتخاب ویژگی می‌باشد. برای انتخاب ویژگی‌های بهینه، الگوریتم جستجوی ممنوعه به کار می‌رود. در فاز دوم روش پیشنهادی، جهت دسته‌بندی، روش یادگیری ماشین مورد استفاده می‌باشد. برای دسته‌بندی پیغام‌ها به دو دسته نرمال و هرزنامه، از روش بگینگ و یا بوستینگ استفاده خواهد شد. روش مدنظر، از بهترین روش‌های موجود در این زمینه به نام ماشین بردار پشتیبان می‌باشد. هدف از روش پیشنهادی، ایجاد ترکیب جدید به منظور انتخاب ویژگی و دسته‌بندی است تا دقت تشخیص هرزنامه افزایش پیدا کند.

[1] Singh

۱-۳- ضرورت انجام تحقیق

افرادی که هرزنامه می‌نویسند، قادرند روزانه هزاران هرزنامه را با حداقل هزینه بفرستند و حتی امکان دارد این کار موجب کسب درآمدی نیز برای آن‌ها، شود. توسعه سیستم‌های فیلترینگ، هرزنامه نویس‌ها را وادار می‌سازد به منظور ادامه بقا، ساختارهای خود را برای دور زدن سیستم‌های فیلترینگ تغییر دهند. این کار، منجر به ایجاد تقابل پویایی بین سیستم‌های فیلترینگ و هرزنامه نویس‌ها می‌گردد. در نتیجه، سیستم‌های فیلترینگ باید همیشه به تشخیص این ساختارها و رویارویی با آن‌ها بپردازند. در سالیان اخیر، الگوریتم‌های دسته‌بندی یادگیری ماشین جهت فیلتر نمودن خودکار هرزنامه‌ها، نظر بسیاری از محققان را به خود معطوف داشته‌اند. در این زمینه، الگوریتم‌های بسیاری نظیر الگوریتم‌های ماشین بردار پشتیبان، تکامل تفاضلی، درخت تصمیم، جنگل تصادفی و نزدیک‌ترین همسایه مورد استفاده قرار گرفته‌اند. در همه پژوهش‌های تحقیقاتی در زمینه تشخیص هرزنامه، هدف جداسازی هرزنامه‌ها از معتبرها است. برای این منظور، ابتدا، استخراج و انتخاب ویژگی‌ها را صورت داده‌اند و در مرحله دوم، دسته‌بندی این ویژگی‌ها به دو دسته معتبر و غیرمعتبر صورت گرفته است. هرزنامه‌ها، به یک مشکل همه‌گیر مبدل گشته‌اند که قادرند تأثیر منفی روی توانایی به‌کارگیری پست الکترونیکی داشته باشند. افزون بر اتلاف زمان کاربران و تلاش جهت اسکن و حذف تعداد بسیاری از هرزنامه‌های دریافتی، پهنای باند شبکه و فضای ذخیره‌سازی از طریق این نوع ایمیل‌ها مصرف شده، سرعت سرورهای ایمیل را نیز کاهش می‌دهند و علاوه بر این، یک رسانه جهت توزیع مطالب توهین‌آمیز و زیان‌آور مهیا می‌گردد.

۱-٤- اهداف تحقیق

هدف اصلی روش پیشنهادی، افزایش دقت و کارایی سیستم تشخیص هرزنامه است. هدف دیگر، به‌کارگیری الگوریتم‌های جستجوی ممنوعه و ماشین بردار پشتیبان برای به وجود آوردن یک سیستم تشخیص هرزنامه می‌باشد. هدف مهم دیگر، افزایش کارایی فیلتر نمودن هرزنامه‌های الکترونیکی با حذف ویژگی‌های زائد در مرحله انتخاب ویژگی است.

۱-۵- جنبه نوآوری

هرزنامه‌ها، از بزرگ‌ترین مسائل کاربران پست الکترونیکی می‌باشند که منجر به اتلاف زمان، کاهش امنیت و کاهش کارایی کامپیوتر می‌گردند. جهت فایق آمدن بر این مشکل، روش‌های گوناگونی ارائه گردیده است. در این تحقیق، تشخیص هرزنامه با به‌کارگیری الگوریتم‌های جستجوی ممنوعه و ماشین بردار پشتیبان صورت می‌پذیرد. هدف از روش پیشنهادی، به وجود آوردن ترکیب جدید به منظور انتخاب ویژگی و دسته‌بندی می‌باشد تا دقت تشخیص هرزنامه افزایش پیدا کند که نوآوری این تحقیق محسوب می‌گردد. روش‌های پیشین که روی تشخیص هرزنامه کار کرده‌اند، غالباً خطای بسیاری داشته‌اند و نمی‌توانند از نظر کاربردی قابل استناد باشند؛ لذا، دستیابی به یک سیستم با کمترین خطا و کارایی و دقت بالاتر در این زمینه، ضروری می‌باشد. با به‌کارگیری الگوریتم جستجوی ممنوعه، در زمان محاسباتی معقول، جواب‌هایی نزدیک جواب بهینه مطلق حاصل می‌گردد. ازآنجایی‌که الگوریتم جستجوی ممنوعه یک الگوریتم فرامکاشفه‌ای می‌باشد، می‌توان آن را با روش‌های یادگیری ماشین نظیر ماشین بردار پشتیبان ادغام نمود تا الگوریتم جستجو ممنوعه با استخراج جواب‌های مناسب برای مرحله انتخاب ویژگی، جواب‌های دسته‌بندی هرزنامه‌ها را نیز بهبود دهد.

۱-٦- ساختار پایان‌نامه

در ادامه، بخش‌بندی این تحقیق، بدین شرح است که در فصل دوم، مفاهیم مربوط به تشخیص هرزنامه مورد بررسی قرار می‌گیرد. در فصل سوم، به بررسی روش‌های ارائه شده برای تشخیص هرزنامه که مبتنی بر الگوریتم‌های یادگیری ماشین هستند، پرداخته می‌شود. در فصل چهارم، جزییات روش پیشنهادی بیان می‌گردد. در فصل پنجم، نتایج به دست آمده از شبیه‌سازی روش پیشنهادی با الگوریتم‌های دیگر و روش‌های متداول در این زمینه، مورد مقایسه قرار می‌گیرد. در پایان، نتیجه‌گیری از این تحقیق و پیشنهادهایی برای تحقیقات آینده مطرح می‌گردد.

فصل دوم

مبانی تحقیق

در این فصل، به بررسی مفاهیم تشخیص هرزنامه پرداخته می‌شود.

۲-۱- تشخیص هرزنامه

هرزنامه، از مواردی می‌باشد که اثربخشی به‌کارگیری پست الکترونیک را ضعیف می‌نماید. از آنجایی که هرزنامه، ویران‌کننده ارتباطات الکترونیکی می‌باشد؛ با افزایش تعداد هرزنامه‌ها، مشکلات بسیاری ایجاد می‌شود که می‌توان به مواردی همچون پر شدن صندوق پستی کاربر، غرق شدن پیام‌های شخصی کاربر در انبوهی از پیغام‌ها، از دست دادن فضای ذخیره و پهنای باند ارتباطی و وقت کاربر در حذف پیغام‌های هرزنامه اشاره کرد. روش‌های گوناگون و زیادی برای رویارویی با این پدیده ارائه شده است که هر یک دارای مزیت‌ها و معایبی هستند، شناخت درست کارکرد هر یک از این الگوریتم‌ها و میزان دقت در تشخیص و فیلتر نمودن نامه‌های الکترونیکی نامعتبر، مهم تلقی می‌گردد. هرزنامه‌هایی که به شکل پست الکترونیکی می‌باشند، تنها به‌عنوان زباله در نظر گرفته نمی‌شوند و از آنجا که فایل پیوست ویروس و عوامل نرم‌افزارهای جاسوسی را در بر می‌گیرند، قادرند برای یک سیستم و دریافت‌کنندگان آن، خطرناک باشند و منجر به از دست دادن اطلاعات شوند؛ لذا، به ابزارهایی برای تشخیص هرزنامه احتیاج می‌باشد. تعداد زیادی از راه‌های تشخیص هرزنامه بر مبنای روش‌های یادگیری ماشین پیشنهاد گردید. همان‌گونه که میزان هرزنامه به‌طور قابل‌ملاحظه‌ای با به‌کارگیری ابزارهای پستی افزایش پیدا کرده است، روش‌هایی برای تشخیص هرزنامه و بعلاوه رویارویی با آن، لازم می‌باشد (هاشمی، ۱۳۹۳).

هرزنامه، سوءاستفاده از سیستم‌های پیام‌دهی الکترونیکی (حاوی رسانه‌های داده پراکنی و سیستم‌های تحویل دیجیتالی اطلاعات) جهت ارسال پیام‌های بسیاری برای افراد نامعلوم است. درصورتی‌که هرزنامه‌های پست