Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال اول، شماره دوم، پاییز و زمستان ۱۴۰۲، صفحات۵۳-۴۴

مروری بر روش‌های شناسایی سایت‌های فیشینگ از سایت‌های قانونی

فرح مکی الهیجل [۱]، زهرا یعقوبی [۲*]

[۱] دانشجوی کارشناسی ارشد، دانشکده فنی و مهندسی، دانشگاه بین‌المللی امام خمینی(ره)، قزوین، ایران

[۲]استادیار دانشکده فنی و مهندسی، دانشگاه بین‌المللی امام خمینی(ره)، قزوین، ایران

چکیده

محبوبیت روزافزون اینترنت منجر به رشد چشمگیر تجارت الکترونیک شد. بااین‌حال، چنین فعالیت‌هایی دارای چالش‌های امنیتی اساسی است که عمدتاً ناشی از کلاه‌برداری‌های سایبری و سرقت هویت افراد است. ازاین‌رو، بررسی مشروعیت صفحات وب بازدید شده یک کار بسیار مهم برای ایمن کردن هویت مشتریان و جلوگیری از حملات فیشینگ است. استفاده از روش‌های یادگیری ماشین، و یادگیری عمیق به‌طور گسترده به‌عنوان یک‌راه حل امیدوارکننده شناخته‌شده است. تحقیقات، مملو از مطالعاتی است که از روش‌های یادگیری ماشین و یادگیری عمیق برای تشخیص فیشینگ وب‌سایت استفاده می‌کنند. با توجه به این‌که در تحقیقات قبلی، تشخیص سایت-های فیشینگ با استفاده از ویژگی‌های مختلف توسط شبکه‌های عمیق بررسی‌شده است، در تحقیق جاری، ما قصد داریم تشخیص سایت‌های فیشینگ را با استفاده از چندین نوع ویژگی به‌صورت ترکیبی با کمک شبکه‌های عمیق انجام دهیم. بنابراین در این مقاله ابتدا ما به‌مرور روش‌های شناسایی سایت‌های فیشینگ از سایت‌های قانونی می‌پردازیم و در انتها روش پیشنهادی خود را ارائه می‌دهیم.

مقاله پژوهشی

تاریخ دریافت:
۱۴۰۱/۱۱/۲۴

تاریخ پذیرش:
۱۴۰۲/۳/۳۰

کلیدواژه‌ها:
شناسایی سایت‌های فیشینگ، یادگیری ماشین، یادگیری عمیق

نویسنده‌مسئول:
z.yaghoubi@eng.ikiu.ac.ir

: 10.22034/ABMIR.2022.2714

E-ISSN: 2821-2037

۱- مقدمه

فیشینگ[1] به‌عنوان ساده‌ترین و گسترده‌ترین تهدید جرائم سایبری شناخته‌شده است. هکرها نیازی به کرک کردن کد رمز پیچیده و نقض فایروال سخت را ندارند. به‌جای آن، به‌سادگی ایمیل‌های احساسی، انتقادی یا معقول را به‌منظور درخواست از گیرندگان برای معرفی اعتبار شخصی خود با کلیک بر روی یک پیوند ارسال می‌کنند. سپس گیرندگان به صفحات وب جعلی هدایت می‌شوند که بسیار شبیه به وب‌سایت‌های معتبر مورد هدف است. درنتیجه، گیرندگان مانند ماهی در وب‌سایت‌های جعلی گرفتار می‌شوند [۱]. اخیراً، هکرها کار خود را بسیار حرفه‌ای انجام می‌دهند، روند اخیر گزارش‌شده [۲] از فعالیت فیشینگ نشان داد که ۷۸ درصد از تمام وب‌سایت‌های فیشینگ، از محافظت لایه سوکت‌های امن[2] استفاده می‌کنند که منحصراً توسط وب‌سایت‌های معتبر استفاده می‌شود.

۲- مروری بر روش‌های شناسایی

واندرا [۳] در گزارش چشم‌انداز تهدیدات موبایلی خود در سال ۲۰۲۰ اعلام کرد که هر ۲۰ ثانیه یک وب‌سایت جدید فیشینگ راه‌اندازی می‌شود. همه این حقایق از مطالعات تحقیقاتی عمیق در مورد شناسایی و پیشگیری از چنین حملات سایبری حمایت می‌کند. جدول (۱) انواع حملات فیشینگ را نشان می‌دهد.

جدول (۱): انواع حملات فیشینگ

انواع حملات فیشینگ						
ایمیل /اسپم	تحویل مبتنی بر وب	دست‌کاری لینک	فیشینگ تلفن	دزدی داده	فیشینگ مبتنی بر بدافزار	فیشینگ مبتنی بر سیستم نام دامنه

تشخیص حملات فیشینگ از دسته مسائل طبقه‌بندی محسوب می‌شود. بنابراین، روش‌های یادگیری ماشین[3] به‌عنوان راه‌حل‌های امیدوارکننده در نظر گرفته می‌شوند. بااین‌حال، سه جنبه اصلی باید در هنگام پذیرش چنین روش‌هایی در نظر گرفته شود: (۱) انتخاب طبقه‌بندی کننده‌های کارآمد، (۲) استفاده از ویژگی‌های متمایز، و (۳) مجموعه‌ای از نمونه‌های داده نماینده برای آموزش. سیستم‌های مبتنی بر یادگیری ماشین که برای تشخیص فیشینگ توسعه‌یافته‌اند، عمدتاً به دودسته اصلی طبقه‌بندی می‌شود: سیستم‌های مبتنی بر محتوا[4] و سیستم‌های مبتنی بر منبع‌یاب یکنواخت[5]. سیستم‌های مبتنی بر محتوا، فیشینگ توسط بررسی فعال یا غیرفعال بودن محتوای صفحات وب بازدید شده، شناسایی می‌شود و سیستم‌های مبتنی بر منبع‌یاب یکنواخت، فقط URL های صفحات وب بازدید شده موردبررسی قرار می‌گیرند [۱].

روش‌های مختلفی برای شناسایی وب‌سایت فیشینگ پیشنهادشده است که عبارت‌اند از: مبتنی بر فهرست[6]، شباهت بصری[7]، اکتشافی[8]، یادگیری ماشین و یادگیری عمیق[9] [٤]. جدول (۲) این روش‌ها را نشان می‌دهد.

جدول (۲): انواع روش‌های شناسایی سایت‌های فیشینگ

روش‌های شناسایی فیشینگ				
اکتشافی	شباهت بصری	مبتنی بر لیست	یادگیری ماشین	یادگیری عمیق

۲-۱ روش مبتنی بر فهرست

مرورگرهایی مانند مایکروسافت آج[10]، فایرفاکس و گوگل کروم از روش‌های مبتنی بر فهرست برای شناسایی وب‌سایت‌های فیشینگ استفاده می‌کند. لیست سفید و لیست سیاه دو نوع از روش‌های مبتنی بر لیست هستند. لیست سفید حاوی فهرستی از URL های معتبر است که مرورگرها می‌توانند به آن‌ها دسترسی داشته باشند، به این معنی که اگر URL در لیست سفید باشد، مرورگر می‌تواند صفحه وب را دانلود کند. در همان زمان، پایگاه داده لیست سیاه شامل فیشینگ یا URL های جعلی است که مرورگرها از دانلود صفحات آن سایت‌ها متوقف‌شده‌اند. عیب عمده این است که یک تغییر جزئی در URL برای دور زدن روش‌های مبتنی بر فهرست کافی است و برای جلوگیری از URL های فیشینگ جدید، این فهرست‌ها باید مرتباً به‌روز شوند [۵].

[1] Phishing
[2] Secure Sockets Layer
[3] Machine Learning
[4] Content-based
[5] (Uniform Resource Locator) URL-based
[6] Lists-Based
[7] Visual Similarity
[8] Heuristic
[9] Deep Learning
[10] Edge

۲-۲ روش شباهت بصری

این رویکرد مظنون را ارزیابی می‌کند و وب‌سایت‌های معتبر را بر اساس ویژگی‌های بصری مختلف احراز هویت می‌کند. زیرابه نظر می‌رسد که صفحه وب فیشینگ بسیار شبیه به صفحه قانونی[1] خودش است، این ابزارها شباهت‌ها را مقایسه می‌کنند: این روش از برگه‌های سبک آبشاری[2]، طرح‌بندی متن، کد منبع، آرم وب‌سایت، تصاویر صفحه وب، و سایر عناصر بصری استفاده می‌کند. از آن جایی که این روش‌ها صفحه وب مشکوک را با صفحات وب قبلاً بازدید شده یا ذخیره‌شده مقایسه می‌کنند آن‌ها نمی‌توانند حملات فیشینگ ساعت صفر[3] را شناسایی کنند [٦].

۲-۳ روش اکتشافی

رویکرد اکتشافی از ویژگی‌های مشتق شده از وب‌سایت فیشینگ استفاده می‌کند. این استراتژی بر چندین ویژگی استوار است که می‌تواند یک وب‌سایت فیشینگ را از یک وب‌سایت واقعی متمایز کند. این روش‌ها داده‌ها را از منابع مختلف، مانند URL ها، محتوای متنی، سیستم نام دامنه، گواهی‌های دیجیتال و ترافیک وب‌سایت‌ها جمع‌آوری می‌کند. مجموعه ویژگی، نمونه‌های آموزشی و الگوریتم‌های طبقه‌بندی، همگی بر موفقیت این روش تأثیر می‌گذارد. یکی از مزایای این روش این است که می‌تواند حملات فیشینگ ساعت صفر را شناسایی کند [٧].

۲-٤ روش یادگیری ماشین

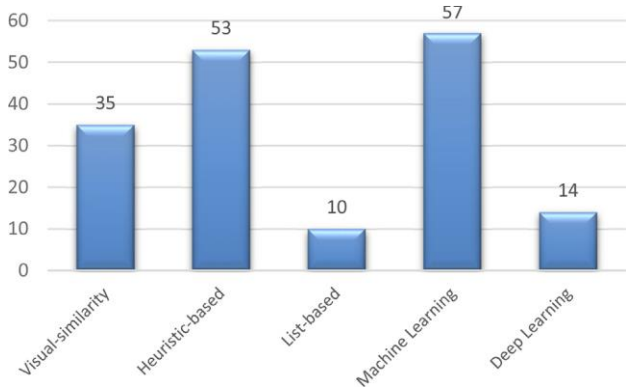
امروزه یادگیری‌ماشین روشی رایج برای شناسایی وب‌سایت‌های فیشینگ است [٨]. ویژگی‌های رایج مانند اطلاعات URL، ساختار وب‌سایت، و ویژگی‌های جاوا اسکریپت برای نشان دادن URL های فیشینگ و وب‌سایت‌های مرتبط جمع‌آوری می‌شوند. سپس، بر اساس آن ویژگی‌ها، مجموعه داده‌های فیشینگ به دست می‌آید. پس‌ازآن، طبقه‌بندی‌کننده‌های یادگیری ماشین برای شناسایی وب‌سایت فیشینگ، بر اساس آن ویژگی‌ها آموزش داده می‌شوند [٩]. این روش‌ها با داده‌های حجیم بسیار خوب کار می‌کند. داده‌های حجیم دارای مجموعه داده‌هایی با سرعت، تنوع، حجم، ارزش و صحت بالا است. طبقه‌بندی‌کننده‌های مبتنی بر یادگیری ماشین بیش از ۹۹ درصد از دقت را به دست آوردند که ثابت کردند موثرترین روش‌ها بودند [١٠].

۲-۵ روش یادگیری عمیق

با توجه به پیشرفت‌های اخیر در روش‌های یادگیری عمیق، شبکه‌های عصبی عمیق[4]، باید عملکرد بهتری نسبت به روش‌های متداول یادگیری ماشین در تشخیص وب‌سایت‌های فیشینگ داشته باشند. برخی از الگوریتم‌های معروف یادگیری عمیق شامل، شبکه عصبی عمیق، شبکه عصبی مکرر[5]، شبکه عصبی عمیق فید فوروارد[6]، ماشین بولتزمن محدودشده[7]، شبکه عصبی پیچشی[8]، شبکه باور عمیق[9] و رمزگذار خودکار عمیق[10] است [١١].

۳- ۳ - مروری بر کارهای انجام‌شده

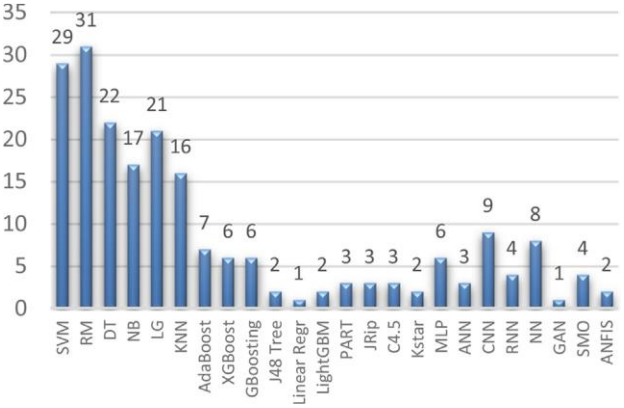
شکل (۳) تعداد مقالات مربوط به انواع روش‌های شناسایی سایت‌های فیشینگ را نشان می‌دهد [٤].



شکل (۳): تعداد مقالات مربوط به انواع روش‌های شناسایی سایت‌های فیشینگ

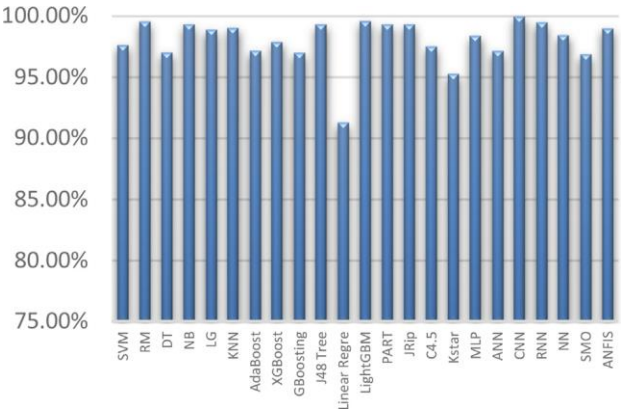
شکل (٤) تعداد مقالات یافت شده برای هر الگوریتم مورداستفاده در شناسایی سایت‌های فیشینگ را نشان می‌دهد [٤].

[1] Legitimate
[2] Cascading Style Sheets
[3] Zero-hour phishing attacks
[4] Deep Neural Network
[5] Recurrent Neural Network
[6] Feed-Forward Deep Neural Network
[7] Limited Boltzmann machine
[8] Convolutional Neural Network
[9] Deep belief network
[10] Deep auto-encoder



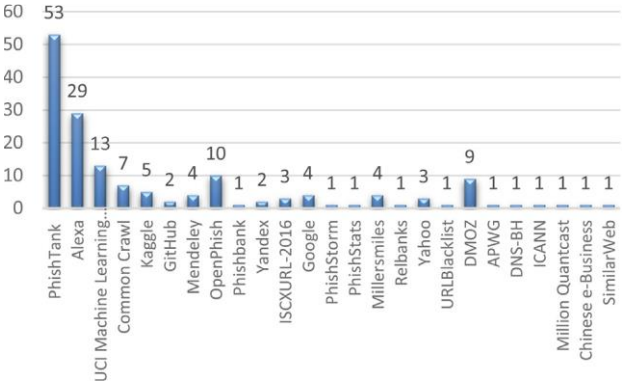
شکل (٤): تعداد مقالات یافت شده برای هر الگوریتم مورداستفاده در شناسایی سایت‌های فیشینگ

شکل (٥) بیشترین دقت به‌دست‌آمده توسط هر الگوریتم را در شناسایی سایت‌های فیشینگ نشان می‌دهد [٤].



شکل (٥): بیشترین دقت به‌دست‌آمده توسط هر الگوریتم در شناسایی سایت‌های فیشینگ

شکل (٦) تعداد مقالات یافت شده برای انواع مجموعه داده مورداستفاده در شناسایی سایت‌های فیشینگ را نشان می‌دهد [٤]. پس از جمع‌آوری ٥٣٧ مطالعه، فقط ٨٠ مطالعه، مرتبط با حملات فیشینگ بود. از این ٨٠ نشریه، ٣٠ مقاله که ٣٨ درصد از کل را تشکیل می‌دهند، در مجله IEEE، ٢٠ (یا ٢٥٪) در اسپرینگر، ١٧ (یا ٢١٪) در الزویر، ٨ (یا ١٠٪) و ٥ (یا ٦٪) در مجلات نمایه شده Scopus یافت می‌شود. این نشان می‌دهد که IEEE در انتشار این موضوع پیشرو است. هم‌چنین از ٨٠ مقاله منتخب، ٣٤ مقاله در سال ٢٠٢٠، که معادل ٤٣ درصد، و ١٦ مقاله (یا ٢٠ درصد) در سال ٢٠٢١، ١٥ مقاله (یا ١٩٪) در سال ٢٠١٨، ١٣ مقاله تحقیقاتی (یا ١٦٪) در سال ٢٠١٩، و ٢ مقاله (یا ٢٪) در سال ٢٠١٧ منتشر شد [٤].



شکل (٦): تعداد مقالات یافت شده برای انواع مجموعه داده مورداستفاده در شناسایی سایت‌های فیشینگ

٤- روش پیشنهادی

با توجه به اینکه در کارهای قبلی، تشخیص فیشینگ با استفاده از ویژگی‌های مختلف توسط شبکه‌های عمیق بررسی‌شده است، در کار جاری قصد داریم تشخیص فیشینگ را با استفاده از چندین نوع ویژگی در کنار هم با کمک شبکه‌های عمیق انجام دهیم. ویژگی‌های ترکیبی به‌کاررفته عبارت‌اند از:

١) ویژگی‌های مبتنی بر URL مانند طول بخش‌های URL، آی پی، استفاده از کاراکترهای خاص در آدرس URL و غیره.

٢) ویژگی‌های مبتنی بر محتوا مانند تعداد لینک‌های هر صفحه، نسبت لینک‌های داخلی به خارجی، نسبت لینک‌های خراب و غیره.

٣) ویژگی‌های خارجی مانند دامنه ثبت‌شده در WHOIS، دوره اعتبار دامنه، عمر دامنه و غیره.

اطلاعات مربوط به ویژگی‌های فوق برای تعداد ١١٤٨٠ وب‌سایت، در دیتاست phishing_data.csv موجود است. دیتاست ما از دیتا ست‌های آماده موجود در اینترنت است.

در ادامه، به توصیف دقیق ویژگی‌ها می‌پردازیم:

الف) ویژگی‌های مربوط به URL

١- طول URL و بخش‌های آن: به‌منظور مخفی سازی دامنه و زیر دامنه‌های واقعی در سایت‌های فیشینگ از URL های طولانی استفاده می‌شود. به همین خاطر طول کل URL و طول نام

هاست را به‌عنوان ویژگی در نظر گرفتیم (از نوع عدد صحیح).

۲- آدرس IP: برای مخفی‌سازی هویت وب‌سایت‌ها، گاهی در قسمت نام هاست از آدرس آی پی استفاده می‌شود. به همین دلیل، وجود یا عدم وجود IP در نام دامنه را به‌عنوان یک ویژگی در نظر گرفتیم (۱ است اگر وجود داشته باشد و ۰ است اگر وجود نداشته باشد).

۳- کاراکترهای خاص: برای آن‌که کاربر متوجه تفاوت بین دامنه واقعی و دامنه فیشینگ نشود از کاراکترهای خاص مانند نقطه و دیگر علائم نگارشی و غیره (شارپ، دلار و...) استفاده می‌شود، به همین دلیل تعداد این علائم در نام دامنه را به‌عنوان یک ویژگی در نظر گرفتیم (عدد صحیح).

٤- بخش‌های مشترک: اجزای مشترک مانند www و http و com. فقط یک‌بار در نام دامنه ظاهر می‌شوند، در غیر این صورت مشکوک بوده و تعداد آن در این ویژگی قرار می‌گیرد (عدد صحیح).

٥- توکن HTTPS: در مقایسه با وب‌سایت‌های قانونی، بیشتر وب‌سایت‌های فیشینگ، فاقد ابزار امنیتی هستند. به همین دلیل، وجود یا عدم وجود آن را در نظر گرفتیم (۰ و ۱).

٦- نسبت اعداد موجود در دامنه: وجود تعداد زیاد عدد در URL، مشکوک بوده و به‌عنوان نشانه فیشینگ در نظر گرفته‌شده است. نسبت اعداد در کل URL و در قسمت نام هاست را در نظر گرفتیم (اعشاری).

۷- Punycode: پانی کد، به کدهای یونیکد جایگزین شده با کدهای اسکی انگلیسی گفته می‌شود که دلیلی ندارد در سایت‌های قانونی از آن استفاده شود. به همین دلیل با یک متغیر دودویی، URL های دارای پانی کد را مشخص کرده‌ایم (دودویی).

۸- شماره پورت: بندرت از شماره پورت در URL های غیربدخواه استفاده می‌شود. به همین دلیل یک متغیر دودویی برای مشخص کردن این URL ها در نظر گرفتیم (دودویی).

۹- محل TLD: در URL های خوش‌فرم، TLD یا دامنه سطح بالا (مثل .com)، قبل از مسیر ظاهر می‌شود. وقتی TLD در مسیر یا در بخش زیردامنه قرار دارد، URL مشکوک به فیشینگ است و این ویژگی را با متغیر دودویی در نظر گرفتیم (دودویی).

۱۰- زیر دامنه‌های غیرعادی: ممکن است URL فیشینگ، به‌جای www از ترکیب w و اعداد و حالات مختلف آن استفاده کند. پس چنین URL هایی مشکوک به فیشینگ هستند و با متغیر دودویی مشخص می‌شوند.

۱۱- تعداد زیردامنه ها: URL های فیشینگ، معمولاً دارای تعداد زیردامنه بیشتری از سایت‌های قانونی هستند و تعداد آن‌ها را به‌عنوان یکی از ویژگی‌های فیشینگ در نظر گرفتیم (عدد صحیح).

۱۲- پیشوند-پسوند: پیشوندها و پسوندها که با علامت – از هم جدا می‌شوند، برای وانمود کردن ظاهری قانونی و بی‌خطر به کاربران در URL استفاده می‌شوند. به همین دلیل، وجود این علامت را (-) را با متغیر دودویی به‌عنوان ویژگی فیشینگ در نظر گرفتیم (دودویی).

۱۳- دامنه‌های تصادفی: URL های فیشینگ، از کلمات مرکب شامل کاراکترهای تصادفی تشکیل‌شده‌اند و این مسئله را با متغیر دودویی در نظر گرفته‌ایم (دودویی).

۱٤- سرویس‌های کوتاه‌سازی URL: نام‌های کوتاه شده‌ای که به دیگر URL های پیچیده و طولانی هدایت می‌شوند و در فیشینگ برای مخفی سازی نام هاست های واقعی قابل‌استفاده است. در صورت استفاده از این سرویس‌ها آن را با متغیر دودویی مشخص کرده‌ایم (دودویی).

۱۵- پسوند مسیر: اسکریپت‌های بدخواه را می‌توان به صفحات سالم اضافه کرد. برخی پسوند فایل‌های به‌کاررفته در URL ها به این منظور استفاده می‌شوند که در این کار، پسوندهای txt و exe و js را در نظر گرفتیم (دودویی).

۱٦- هدایت مجدد (redirect): از هدایت مجدد یا تغییر مسیر URL، برای بازکردن صفحاتی غیر از صفحه انتخابی کاربر استفاده می‌شود که برای جلوگیری از دستیابی به لینک‌های خراب (پس از حذف شدن صفحه اصلی) از آن استفاده می‌شود. هدایت مجدد می‌تواند به صفحات داخلی سایت یا

کلاً مسیرهای خارجی استفاده شود که تعداد هر یک از این موارد را در یک ویژگی مشخص کرده‌ایم (عدد صحیح).

۱۷- ویژگی‌های NLP: برای تشخیص فیشینگ از ویژگی‌های NLP (پردازش زبان طبیعی) استفاده می‌شود که در این کار تعدادی از موارد زیر را به صورت ویژگی‌های جداگانه در نظر گرفتیم: تعداد کلمات، تکرار کاراکتر، کلمات کوتاه شده در URL و نام هاست و مسیر، طولانی‌ترین کلمه در URL و نام هاست و مسیر، میانگین طول کلمات در URL ها، نام هاست ها و مسیرها (عدد صحیح).

۱۸- Phish hints (سرنخ‌های فیشینگ): کلمات حساسی هستند که برای جلب اعتماد به صفحات بازدید شده به کار می‌روند. تعداد چنین کلماتی در URL، به‌عنوان یک ویژگی در نظر گرفته شد (عدد صحیح).

۱۹- دامنه‌های برند: URL های فیشینگ از نام دامنه‌های برند در بخش‌های مختلف استفاده می‌کنند. استفاده از نام برند در بخش دامنه، مجاز و استفاده از آن در زیردامنه یا مسیرها را غیرمجاز در نظر گرفتیم و این موارد را با متغیرهای دودویی نشان دادیم (دودویی).

۲۰- TLD های مشکوک: TLD ها بررسی‌شده‌اند و فهرستی از TLD های مشکوک در تحقیق جاری استفاده‌شده است که این ویژگی با متغیر دودویی نشان داده‌شده است (دودویی).

۲۱- گزارش آماری: دامنه‌های URL بررسی شدند و IP آن‌ها با دامنه‌های مهم فیشینگ، تطبیق داده‌شده است (دودویی).

ب) ویژگی‌های محتوایی

۱- تعداد ابرلینک‌ها: وب‌سایت‌های غیرفیشینگ، معمولاً تعداد صفحات بیشتری نسبت به سایت‌های فیشینگ دارند ازاین‌رو، تعداد لینک‌های موجود در محتوای صفحه را به‌عنوان یک ویژگی در نظر گرفتیم (عدد صحیح).

۲- نسبت لینک‌های داخلی/خارجی: وب‌سایت‌های غیرفیشینگ، بیشتر دارای لینک‌های داخلی هستند و وب‌سایت‌های فیشینگ برعکس. ازاین‌رو، نسبت لینک‌های داخلی به خارجی را به‌عنوان یک ویژگی در نظر گرفتیم (عدد اعشاری).

۳- نسبت ابرلینک‌های تهی: برای تقلید از سایت‌های قانونی، لینک صفحات غیرفیشینگ به صورت تهی در سایت‌های فیشینگ قرار داده می‌شود (یعنی عملاً به‌جایی اشاره نمی‌کنند). نسبت لینک‌های تهی به سالم به‌عنوان ویژگی محتوایی دیگر در نظر گرفته‌شده است (اعشاری).

٤- تعداد فایل CSS خارجی: در وب‌سایت‌های قانونی، از یک سبک داخلی یا بیش از یک فایل CSS خالی استفاده می‌شود. اما در وب‌سایت‌های فیشینگ، فقط از یک فایل CSS خارجی استفاده‌شده که حاوی لینک‌هایی به فایل‌های CSS سایت هدف است. ازاین‌رو تعداد فایل‌های CSS خارجی را به‌عنوان ویژگی دیگر در نظر گرفتیم (عدد صحیح).

۵- تعداد تغییر مسیر (redirect) های داخلی/خارجی: لینک‌های صفحات فیشینگ، ممکن است به دیگر صفحات جعلی یا سالم، اشاره کنند. نسبت تغییر مسیرهای داخلی و خارجی برای تشخیص صفحات وب فیشینگ در نظر گرفته‌شده است (عدد صحیح).

٦- نسبت خطاهای داخلی/خارجی: معمولاً در صفحات فیشینگ، از لینک‌های جعلی استفاده می‌شود. بنابراین، همه لینک‌های صفحات وب، بازرسی شد و نسبت خطاهای اتصال داخلی و خارجی به‌عنوان ویژگی در نظر گرفته‌شده است (اعشاری).

۷- فرم‌های ورود (لاگین): فرم‌های لاگین، ابزار پرکاربرد دیگری برای ربودن اطلاعات کاربران وب‌سایت‌ها محسوب می‌شوند. وجود فرم‌های لاگین با لینک اکشن خارجی یا اکشن تهی به‌عنوان دیگر ویژگی فیشینگ در نظر گرفته‌شده است که با ۰ یا ۱ مشخص می‌شود (دودویی).

۸- فاو آیکن: برای تقلید از وب‌سایت‌های قانونی، در سایت‌های فیشینگ، از فاو آیکن‌های مشابه با سایت هدف استفاده می‌شود. به همین جهت، استفاده از فاو آیکن‌های خارجی به‌عنوان ویژگی در نظر گرفته‌شده است (دودویی).

۹- لینک در تگ: در وب‌سایت‌های قانونی، انتظار می‌رود تگ‌های <links>، از لینک‌هایی استفاده کنند که به صفحات

وب با دامنه مشابه با URL اشاره کنند. ازاین‌رو، نسبت لینک‌های داخلی در تگ‌های <link> به لینک‌های خارجی نیز در نظر گرفته‌شده است (اعشاری).

۱۰- ارسال به ایمیل: وب‌سایت‌های فیشینگ، ورودی‌های کاربران در وب فرم‌ها را به ایمیل‌های خاص ارسال می‌کنند. فرم‌های حاوی :mailto یا ()mail به‌عنوان ویژگی در نظر گرفته شدند (دودویی).

۱۱- نسبت رسانه‌های داخلی/خارجی: وب‌سایت‌های قانونی، بیشتر از رسانه‌هایی که از عکس، صوت و فیلم استفاده می‌کنند در دامنه مشابه با URL ذخیره‌شده‌اند. به همین دلیل نسبت لینک‌های فایل رسانه‌ای داخلی و خارجی را نیز به‌عنوان ویژگی در نظر گرفتیم (عدد صحیح).

۱۲- SFH: معمولاً، پس از دریافت اطلاعات از کاربر، باید اقدامی (اکشن) اجرا شود. از همین رو فرم‌های با رشته تهی یا about:blank را به‌عنوان بخشی از سایت فیشینگ در نظر گرفتیم (دودویی).

۱۳- iFrame های غیرقابل مشاهده: از تگ‌های iFrame برای گنجاندن دیگر صفحات وب در یک صفحه، استفاده می‌شود. وب‌سایت‌های فیشینگ معمولاً از این تگ‌ها با مرزهای (border) نامرئی استفاده می‌کنند تا کاربر فکر کند که در بخشی از صفحه واقعی است. این مسئله را نیز به‌عنوان ویژگی در نظر گرفتیم (دودویی).

۱۴- پنجره‌های بازشو (pop up): در سایت‌های قانونی، برای هشدار دادن به کاربران استفاده می‌شود و به‌ندرت برای جمع‌آوری و ارسال داده استفاده می‌شوند. وجود چنین پنجره‌هایی با فیلدهای متنی به‌عنوان نشانه فیشینگ استفاده‌شده است (دودویی).

۱۵- لنگر امن: از تگ <a>، برای لینک یک صفحه به صفحه دیگر استفاده می‌شود. تگ‌هایی با لینک‌های '#', ' ', 'javascript', 'mailto'' را به‌عنوان ناامن در نظر گرفته و تعداد آن‌ها را در یک ویژگی قراردادیم (عدد صحیح).

۱۶- کلیک راست: اسکریپت‌هایی برای غیرفعال سازی راست کلیک وجود دارد که سایت‌های فیشینگ برای جلوگیری از مشاهده سورس کد، از این قابلیت استفاده می‌کنند و آن را به‌عنوان یک ویژگی در نظر گرفتیم (دودویی).

۱۷- عنوان تهی: بیشتر وب‌سایت‌های قانونی، عنوان صفحه خود را در تگ <title> قرار می‌دهند. عدم وجود عنوان را در یک ویژگی در نظر گرفتیم (دودویی).

۱۸- وجود نام دامنه در عنوان: وب‌سایت‌های قانونی، از نام دامنه به‌عنوان بخشی از عنوان صفحه استفاده می‌کنند. وب‌سایت‌های فیشینگ، از دامنه‌های قانونی، در عنوان خود برای فریب دادن کاربران استفاده می‌کنند. ازاین‌رو، وجود دامنه URL به‌عنوان بخشی از عنوان صفحه را به‌عنوان ویژگی سایت‌های قانونی در نظر گرفتیم (دودویی).

۱۹- وجود دامنه در کپی‌رایت: وب‌سایت‌های قانونی، نام دامنه خود را در لوگوی کپی‌رایتشان نمایش می‌دهند که این امر در مورد وب‌سایت‌های فیشینگ درست نیست. این مسئله را به‌عنوان یک ویژگی در نظر گرفتیم (دودویی).

ج) ویژگی‌های به‌دست‌آمده از منابع خارجی:

۱- ثبت در whois: برخلاف سایت‌های قانونی، دامنه وب‌سایت‌های فیشینگ، معمولاً در WHOIS ثبت‌نشده است. این مسئله را به‌عنوان یک ویژگی در نظر گرفتیم (دودویی).

۲- طول عمر ثبت دامنه: سایت‌های فیشینگ معمولاً، عمر کوتاهی دارند (برخلاف سایت‌های قانونی). به همین دلیل تعداد سال تمدید دامنه را به‌عنوان یک ویژگی در نظر گرفتیم (عدد صحیح).

۳- عمر دامنه: عمر دامنه (فاصله زمانی تا انقضا)، نیز به‌عنوان ویژگی در نظر گرفته شد (زیرا سایت‌های فیشینگ، عمر کوتاهی دارند) (عدد صحیح).

۴- ترافیک وب: وب‌سایت‌های فیشینگ، معمولاً بازدیدکننده کمتری نسبت به سایت‌های قانونی دارند و به همین دلیل این مسئله را با استفاده از سایت الکسا، به‌عنوان یک ویژگی در نظر گرفتیم (عدد صحیح).

۵- رکورد DNS:سیستم نام دامنه، برای بازیابی آدرس IP و دستیابی به URL، ضروری است. ازاین‌رو، دامنه‌های URL

باید در DNS ثبت شوند. عدم وجود رکورد DNS، به‌عنوان نشانه فیشینگ در نظر گرفته‌شده است (دودویی).

۶- نمایه گوگل: وب‌سایت‌های فیشینگ، عمر کوتاهی داشته و غالباً از طریق لینک‌های مستقیم در ایمیل به قربانی ارسال می‌شوند و نیازی به نمایه گوگل ندارند. ازاین‌رو، نمایه بودن یا نبودن در گوگل را به‌عنوان ویژگی در نظر گرفتیم (دودویی).

۷- Page rank: وب‌سایت‌های فیشینگ، محبوبیت چندانی ندارند، ازاین‌رو، فرض می‌شود که نسبت به سایت‌های قانونی، رتبه صفحه پائینی داشته باشند. بنابراین از Openpagerank برای به دست آوردن این ویژگی استفاده کردیم (عدد صحیح).

پیاده‌سازی کار به این صورت است که پس از آماده‌سازی داده‌ها (تمیز سازی داده‌ها، جداسازی برچسب‌ها از داده‌ها و نرمال‌سازی در صورت لزوم)، مدل‌های مختلف شبکه عمیق شامل، شبکه Dense، شبکه LSTM و شبکه RNN ساخته می‌شود. بر اساس آزمایش و خطا، تعداد بهینه لایه‌ها و نرون‌ها را برای هر مدل تعیین می‌کنیم. سپس با استفاده از نمودارهایی، میزان loss، و تعداد اپک بهینه را برای هر مدل تعیین می‌کنیم و مدل را آموزش می‌دهیم. سپس مدل را روی‌داده‌های تست (۲۰ درصد از کل داده‌ها) ارزیابی می‌کنیم. در صورت عدم کسب دقت رضایت‌بخش، با استفاده از cross-validation بازه‌های مختلف داده‌ها را به‌عنوان داده آموزش یا تست ارزیابی می‌کنیم. درنهایت، بهترین دقت به‌دست‌آمده به‌عنوان مبنای مقایسه باکارهای قبلی استفاده‌شده و نتیجه‌گیری مربوطه انجام می‌شود. به دلیل حجم بالای کار پیاده‌سازی روش پیشنهادی، ما قصد داریم در مقاله تحقیقی دیگری فرایند پیاده‌سازی ایده خود را به‌طور کامل شرح دهیم و در این مقاله به‌مرور تحقیقات گذشتگان به‌طور خلاصه و مفید پرداختیم و نیز ایده پیشنهادی خود را به‌طور خلاصه شرح دادیم.

۵- نتیجه‌گیری

کار انجام‌شده در این مقاله شامل، بررسی آن دسته از مطالعاتی است که عملکرد روش‌های شناسایی وب‌سایت فیشینگ را تحلیل کردند. این مقاله مجموعه داده‌های استفاده‌شده و الگوریتم‌های مورداستفاده محققین در پنج سال گذشته در تشخیص وب‌سایت فیشینگ را گزارش می‌کند. پنج روش تشخیص فیشینگ به‌طور عمده توسط جامعه تحقیقاتی، مورداستفاده قرار گرفت که در میان آن‌ها، یادگیری ماشین بیشترین استفاده از رویکردهای موردمطالعه بوده است. از ۸۰ مورد تحقیقاتی، ۵۷ مقاله، یا ۷۱٫۲۵٪ از مطالعات، از رویکردهای یادگیری ماشین در کار خود استفاده کردند. همچنین نظرسنجی‌ها نشان داد که عمدتاً از دو منبع برای تحلیل فیشینگ مجموعه داده‌ها، استفاده شد. ۵۳ یا ۶۶٫۲۵٪ از مطالعات، از وب‌سایت PhishTank استفاده کردند، درحالی‌که، برای مجموعه داده‌های قانونی، ۲۹ یا ۳۶٫۲۵٪ از مطالعات، از وب‌سایت الکسا استفاده کردند. بیشتر نویسندگان از طبقه‌بندی جنگل تصادفی استفاده کردند، که ۳۸٫۷۵٪ از ۸۰ مقاله است. اگرچه الگوریتم جنگل تصادفی در میان الگوریتم‌های سنتی یادگیری ماشین، بیشتر مورداستفاده قرار می‌گیرد، اما دقت الگوریتم تکاملی شبکه عصبی پیچشی، ۹۹٫۹۸٪ است و در میان تمام مطالعاتی که در این نظرسنجی گنجانده‌شده است بهترین است.

References

[1] Hannousse, A., & Yahiouche, S. (2021). Towards benchmark datasets for machine learning based website phishing detection: An experimental study. Engineering Applications of Artificial Intelligence, 104, 104347.

[2] Anti-Phishing Working Group. (2006). Phishing Activity Trends Report-May, 2006. http://www.anti-phishing.org/reports/apwg_report_May2006. pdf.

[3] Wandera (2020). Mobile Threat Landscape 2020: Understanding the key trends in mobile enterprise security in 2020. Technical Report. https://www.wandera.com/mobile-threat-landscape/.

[4] Safi, A., & Singh, S. (2023). A Systematic Literature Review on Phishing Website Detection Techniques. Journal of King Saud University-Computer and Information Sciences.

[5] Yang, L., Zhang, J., Wang, X., Li, Z., Li, Z., & He, Y. (2021). An improved ELM-based and data preprocessing integrated approach for phishing detection considering comprehensive features. Expert Systems with Applications, 165, 113863.

[6] Jain, A. K., & Gupta, B. B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. In Cyber Security: Proceedings of CSI 2015 (pp. 467-474). Springer Singapore.

[7] Jain, A. K., & Gupta, B. B. (2018). Two-level authentication approach to protect from phishing attacks in real time. Journal of Ambient Intelligence and Humanized Computing, 9, 1783-1796.

[8] Sindhu, S., Patil, S. P., Sreevalsan, A., Rahman, F., & AN, M. S. (2020, October). Phishing detection using random forest, SVM and neural network with backpropagation. In 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE) (pp. 391-394). IEEE.

[9] Zhu, E., Ju, Y., Chen, Z., Liu, F., Fang, X. (2020). DTOF-ANN: an artificial neural network phishing detection model based on decision tree and optimal features. Appl. Soft Comput. J. 95,. https://doi.org/10.1016/j.asoc.2020.106505 106505.

[10] Alkawaz, M. H., Steven, S. J., Hajamydeen, A. I., & Ramli, R. (2021, April). A comprehensive survey on identification and analysis of phishing website based on machine learning methods. In 2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE) (pp. 82-87). IEEE.

[11] Basit, A., Zafar, M., Javed, A. R., & Jalil, Z. (2020, November). A novel ensemble machine learning method to detect phishing attack. In 2020 IEEE 23rd International Multitopic Conference (INMIC) (pp. 1-5). IEEE

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 1, No. 2, Autumn & Winter 2023, pp. 44-53

Copy – Move Detection using genetic algorithm

Farah Maki Alhijel[1], Zahra Yaghoubi[1*]

[1]Engineering Department, Imam Khomeini International University, Qazvin, Iran

Article Information

Original Research Paper

Received:
2023 February 13

Accepted:
2023 June 20

Keywords:
Phishing website detection, Machine learning, Deep learning

Corresponding Author*:
z.yaghoubi@eng.ikiu.ac.ir

Abstract

The increasing popularity of the Internet has led to the dramatic growth of e-commerce. However, such activities have significant security challenges, mainly due to cyber fraud and identity theft. Therefore, checking the legitimacy of visited web pages is a very important task to secure the identity of customers and prevent phishing attacks. The use of machine learning methods, and deep learning is widely recognized as a promising solution. Research is full of studies that use machine learning and deep learning methods to detect website phishing. However, their findings depend on the data set and are far from generalizable. The two main reasons for the lack of generalization are impractical replication and lack of appropriate benchmark data sets for fair evaluation of systems. Furthermore, phishing methods are constantly evolving and the proposed models do not keep up with the rapid changes. In this article, we review the methods of identifying phishing sites from legal sites and finally reach the final conclusion.

:10.22034/ABMIR.2023.19703.1020:

E-ISSN: 2821-2037