تشخیص وب‌سایت‌های اسپم فارسی با استفاده از پردازش زبان طبیعی

صبا حیدری‌دوست [۱]، امیرحسین کیهانی‌پور [۲]

[۱] کارشناسی مهندسی کامپیوتر، دانشکده مهندسی دانشکدگان فارابی دانشگاه تهران
saba.heydaridoost@ut.ac.ir
[۲] استادیار گروه مهندسی کامپیوتر، دانشکده مهندسی دانشکدگان فارابی دانشگاه تهران
keyhanipour@ut.ac.ir

چکیده

تولید صفحات اسپم به عنوان یکی روش‌های جلب توجه کاربر به محتوای غیر مطلوب، یکی از چالش‌های عمده در حوزه بازیابی اطلاعات به ویژه در محیط وب، بشمار می‌رود و طی سالهای گذشته، الگوریتم‌های مختلفی برای تشخیص آنها مطرح شده است. بر این اساس، روش‌های تولید اسپم نیز همزمان با پیشرفت فناوری، تغییر شکل می‌دهند. امروزه، یکی از روش‌های غیرقانونی افزایش رتبه وب‌سایت، استفاده از وب‌سایت‌های اسپم است. در این مقاله، ابتدا انواع اسپم و روش‌های شناسایی وب‌سایت‌های اسپم مورد بررسی قرار گرفته است. سپس یک مجموعه داده شامل وب‌سایت‌های اسپم و غیر اسپم در وب فارسی، معرفی شده و با استفاده از این مجموعه داده، یک مدل Multinomial Naïve Bayes آموزش دیده است. در این مدل، متون این وب‌سایت‌ها با توجه به تکنیک‌های پردازش زبان طبیعی، مورد بررسی قرار گرفته است و نهایتاً هر وب‌سایت، در یکی از دو دسته اسپم و غیر اسپم، دسته‌بندی می‌شود. نتایج ارزیابی روش پیشنهادی روی مجموعه داده متشکل از حدود هزار وب‌سایت در محیط وب فارسی، حاکی از برتری عملکرد آن نسبت به روش مرجع مورد مقایسه، بر اساس شاخص ارزیابی F-Score و به میزان حدود ٪۲۰٫۲۵ می‌باشد.

کلمات کلیدی: وب‌سایت‌های اسپم، مدل Multinomial Naïve Bayes، پردازش زبان طبیعی.

۱ مقدمه

وب‌سایت‌های اسپم، عمدتاً صفحاتی با محتوای غیر مفید را شامل می‌باشند که قادر به تأمین نیازهای اطلاعاتی کاربران نیستند. از این رو، شناسایی این قبیل وب‌سایت‌ها، یکی از پردازش‌های پایه و کلیدی در فرآیند ایجاد سامانه‌های بازیابی اطلاعات وب و به خصوص جویشگرهای وب، بشمار می‌رود. بر این اساس، در این پژوهش، به مقوله شناسایی وب‌سایت‌های اسپم در محیط وب فارسی با بهره‌گیری از تکنیک‌های پردازش زبان طبیعی، پرداخته شده است.

۱.۱ چیستی اسپم

اسپم به هر گونه ارسال پیام به تعداد زیادی از کاربران در فضای آنلاین، بدون کسب اجازه یا توافق قبلی با آنها اطلاق می‌شود (Internet Society, 2014). اسپم می‌تواند به صورت ایمیل، پیامک، پیام در شبکه‌های اجتماعی و سایر وسایل ارتباطی صورت گیرد. برای مثال، ایمیل‌های تبلیغاتی ارسال شده به تعداد زیادی از افراد بدون موافقت قبلی آنها نوعی اسپم هستند. همچنین، پیامک‌های تبلیغاتی یا پیامک‌هایی که وعده‌ی جایزه رایگان را در بردارند نیز به عنوان اسپم شناخته می‌شوند. البته ممکن است یک محتوای اسپم، تبلیغاتی باشد یا نباشد (encyclopedia by Kaspersky, 2019).

به‌طور کلی پنج نوع مرسوم اسپم وجود دارد؛ از جمله نظر اسپم، اسپم trackback، حمله منفی سئو، حملات DDos با استفاده از spiderها و ربات‌ها و در نهایت ایمیل اسپم (Cojocariu, 2018). در این مقاله منظور از وب‌سایت اسپم، بک لینک‌هایی است که در حملات منفی سئو به کار گرفته می‌شوند. انواع مختلفی وب‌سایت اسپم وجود دارد که در ادامه شرح داده خواهند شد.

۱.۱.۱ انواع وب‌سایت اسپم

انواع وب‌سایت اسپم به سه دسته مبنی بر محتوا، مبنی بر لینک و مبنی بر صفحات پنهان تقسیم می‌شوند.

در وب‌سایت‌های اسپم مبنی بر محتوا، محتوای صفحه طوری تغییر داده شده تا وب‌سایت رتبه بالاتری را دریافت کند. بیش‌تر تکنیک‌های شناسایی اسپم مبنی بر محتوا از پردازش زبان طبیعی استفاده می‌کنند. پرکاربردترین تکنیک پردازش زبان طبیعی برای تشخیص وب‌سایت‌های اسپم شده با این روش، TF-IDF است (Danandeh Oskuie and Razavi, 2014). تکنیک دیگر، کوله کلمات[1] می‌باشد که در آن، وقوع کلمات در یک متن به صورت عددی توصیف می‌شود (Brownlee, 2017). بدنه، عنوان، URL و برچسب متا در یک وب‌سایت مبنی بر محتوا می‌توانند اسپم باشند. همچنین، ممکن است اسپم‌کننده از صفحات دیگر در وب برای صفحه اسپم خود، محتوا کپی کند و در قسمت‌های مختلف آن، اصطلاحات اسپم را به صورت رندوم قرار دهد. علاوه بر آن، در برخی صفحات اسپم تکرار یک یا چند کلمه خاص و استفاده بسیار از اصطلاحات نامرتبط، محتمل است (Danandeh Oskuie and Razavi, 2014).

در راستای به دست آوردن رتبه بالاتر در وب‌سایت‌های اسپم مبنی بر لینک، ساختار لینک دست‌کاری می‌شود. برای این کار، روش‌های دست‌کاری متفاوتی از جمله خرید لینک، استفاده از دامنه‌های منقضی‌شده و مزارع لینک[2] وجود دارد (Danandeh Oskuie and Razavi, 2014). وب‌سایت‌های اسپم مبنی بر صفحات پنهان، از طریق نشان دادن محتوای متفاوت به جویشگرهای وب، رتبه خود را بالاتر می‌برند. در این نوع، دو روش پنهان‌کاری محتوا برای جویشگرهای وب و تغییر مسیر هنگام بارگذاری صفحه به کار برده می‌شوند (Danandeh Oskuie and Razavi, 2014).

[1] Bag-of-Words
[2] Link Farms

۲.۱ اهمیت شناسایی وب‌سایت‌های اسپم

یکی از روش‌های قانونی افزایش رتبه سایت در نتایج جویشگرهای وب، بهبود کیفیت صفحات آن سایت می‌باشد. اما این روش بسیار زمان‌بر و دارای هزینه بالایی است. روش دیگر که غیرقانونی و غیر اخلاقی به شمار می‌رود، فریب دادن جویشگرهای وب با استفاده از بک لینک‌های اسپم است. این وب‌سایت‌ها با هدف تغییر رتبه سایت در نتایج جویشگرهای وب به وجود می‌آیند (Danandeh Oskuie and Razavi, 2014). تشخیص بک لینک‌های اسپم، یکی از چالش برانگیزترین مشکلات برای جویشگرهای وب و کاربران وب به حساب می‌آید.

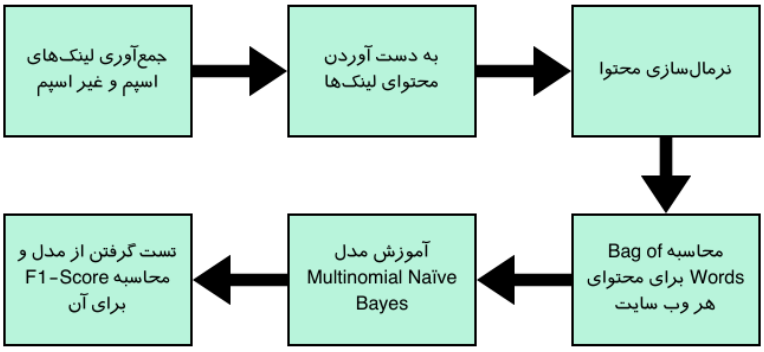
جویشگرهای وب، مانند گوگل، سعی می‌کنند سایت‌ها را بر اساس کیفیت و اعتبار بک لینک‌ها رتبه‌بندی کنند. اما وجود بک لینک‌های اسپم می‌تواند باعث کاهش رتبه وب‌سایت در نتایج جستجو شود (Ntoulas et al., 2006). از طرفی، این لینک‌ها می‌توانند اعتبار یک وب‌سایت را کاهش دهند. وجود بک لینک‌هایی از سایت‌های نامعتبر و مشکوک می‌تواند باعث شود که مخاطبان و جویشگرهای وب، اعتماد کمتری به آن وب‌سایت داشته باشند (Ntoulas et al., 2006). یکی دیگر از آثار مخرب این وب‌سایت‌های اسپم این است که ممکن است باعث کاهش ترافیک وب‌سایت شوند. به این صورت که این بک لینک ها به سایت هدف هدایت شده ولی بازدیدکنندگان به‌سرعت متوجه این موضوع گشته و از سایت خارج می‌شوند. این موضوع باعث کاهش ترافیک و معیارهای مهمی همچون میانگین زمان بازدید و نرخ بازگشت می‌شود.

در نتیجه، برای جلوگیری از خدشه‌دار شدن اعتماد کاربران، هدر رفتن منابع محاسباتی جویشگرهای وب (Najadat & Hmeidi, 2008) و همچنین جلوگیری از کاهش ترافیک وب‌سایت، شناسایی این بک لینک‌ها بسیار حائز اهمیت است و کمک بسیاری به بهبود نتایج جویشگرهای وب کرده و سبب افزایش رضایت کاربران می‌شود.

۳.۱ نوآوری‌ها

- برای ساخت مدل، یک مجموعه داده از لینک وب‌سایت‌های اسپم به همراه محتوای آن‌ها احتیاج بود که برای زبان فارسی، همچین مجموعه داده‌ای وجود نداشت. بنابراین، یک مجموعه داده جدید که شامل لینک وب‌سایت‌های اسپم و همچنین محتوای هر کدام از آنها است، در این مقاله معرفی می‌شود.
- در راستای شناسایی وب‌سایت‌های اسپم، یک مدل یادگیری ماشین که با استفاده از تکنیک‌های پردازش زبان طبیعی این وب‌سایت‌ها را تشخیص می‌دهد، با استفاده از مجموعه داده مذکور، آموزش دیده است.

بر این اساس، ابتدا در بخش دوم این مقاله، مروری بر پژوهش‌های مرتبط صورت خواهد گرفت و سپس در بخش سوم، روش پیشنهادی برای شناسایی وب‌سایت‌های اسپم توضیح داده می‌شود. ارزیابی روش پیشنهادی و نتایج بدست آمده نیز در بخش چهارم ذکر شده است و نهایتاً در بخش پنجم، نتیجه‌گیری این پژوهش، بیان شده است.



شکل ۱: روال به کار رفته در جمع‌آوری داده و آموزش مدل یادگیری ماشین برای شناسایی وب‌سایت‌های اسپم

۲ مروری بر کارهای دیگران

با توجه به اهمیت شناسایی وب‌سایت‌های اسپم، در سال‌های اخیر تلاش‌های بسیاری در این زمینه انجام شده است. در برخی از تلاش‌های انجام‌شده از هوش مصنوعی و یادگیری ماشین برای حل این مسئله استفاده گشته است. (Prieto, Alvarez & Cacheda, 2013) سیستمی به نام SAAD را ارائه دادند که از محتوای وب‌سایت برای تشخیص اسپم بودن یا نبودن آن استفاده می‌کند. (Karmipour, Noroozi & Alizadeh, 2022)، یک روش جدید بر اساس الگوریتم EM با طبقه‌بندی مبنی بر مدل Naïve Bayes برای حل مشکل برچسب‌گذاری پیشنهاد کرده‌اند. این روش، یک مدل طبقه‌بندی را از مجموعه کوچکی از داده‌های برچسب‌دار می‌آموزد تا مجموعه بزرگی از داده‌های بدون برچسب را برچسب‌گذاری کند. (Ntoulas et al., 2006) تعدادی روش مبنی بر محتوا برای شناسایی محتوای اسپم مطرح کرده‌اند. همچنین، (Becchetti et al., 2006) از ویژگی‌های مبنی بر لینک مانند TrustRank و PageRank جهت طبقه‌بندی در دو دسته اسپم و غیر اسپم استفاده کرده‌اند. (Agrawal, 2023) روش TF-IDF را برای ساختاردهی به داده‌ها به کار برده و مدل رگرسیون لجستیک[۳] را برای آموزش داده‌ها انتخاب کرده است.

در این مقاله نیز با به کار گرفتن مدل یادگیری ماشین Multinomial Naïve Bayes و تکنیک‌های پردازش زبان طبیعی، یک مدل برای شناسایی و طبقه‌بندی وب‌سایت‌ها در دو دسته اسپم و غیر اسپم با توجه به محتوای آنها ارائه شده است.

۳ روش پیشنهادی

در پیاده‌سازی مدل شناسایی‌کننده وب‌سایت‌های اسپم، از روال شکل ۱ استفاده شده است.

سه مرحله اول از این روال، مربوط به بخش جمع‌آوری داده بوده و سه مرحله دوم، در بخش مدل یادگیری ماشین، انجام شده‌اند.

[3] Logistic Regression

جدول ۱: آمار مجموعه داده

	اسپم	غیر اسپم	کل
تعداد لینک‌های به دست آمده	۷۳۸	۱۶۹۸	۲۴۳۶
تعداد لینک‌های در دسترس	۶۰۴	۱۳۵۱	۱۹۵۵
تعداد لینک‌ها با محتوای استخراج شده	۳۶۵	۷۵۳	۱۱۱۸
تعداد لینک‌ها پس از تمیز شدن دادگان	۳۶۵	۵۶۰	۹۲۵

۱.۳ جمع‌آوری داده

هدف از آموزش مدل، تشخیص وب‌سایت‌های اسپم و غیر اسپم در محیط وب فارسی است. برای زبان فارسی، چنین مجموعه داده‌ای وجود نداشت. بنابراین، در این مقاله یک مجموعه داده از وب‌سایت‌های اسپم و غیر اسپم فارسی معرفی می‌شود.

در فرآیند جمع‌آوری داده، احتیاج به دو گروه وب‌سایت وجود داشت. گروه اول، شامل وب‌سایت‌های اسپم بوده و گروه دوم از وب‌سایت‌های غیر اسپم تشکیل شده است.

برای گروه اول، ۷۳۸ بک لینک اسپم جمع‌آوری شد. از آن جایی که این لینک‌ها در گذر زمان از دسترس خارج می‌شوند، تمامی این لینک‌ها بررسی شده و در نهایت ۳۶۵ لینک اسپم در دسترس بوده و محتوای موجود در آن‌ها استخراج شده است.

در رابطه با گروه دوم، ابتدا چندین وب‌سایت مانند فارس نیوز، ایسنا، مجله دیجی‌کالا و ... در نظر گرفته شده و سپس تمامی لینک‌های موجود در آن‌ها استخراج گشته و لینک‌هایی که به صفحات دیگر مربوط بودند از میان تمامی لینک‌ها جدا شده‌اند. علت این کار این است که برخی از لینک‌ها، لینک وب‌سایت نبوده و به طور مثال، لینک مربوط به تصاویر بوده‌اند. در ادامه، محتوای لینک‌های جمع‌آوری شده نیز استخراج شد. در نهایت، تعداد ۵۶۰ وب‌سایت به همراه محتوای موجود در آن‌ها، به عنوان وب‌سایت غیر اسپم جمع‌آوری شد.

متون استخراج شده، احتیاج به تمیز شدن به صورت دستی داشتند؛ زیرا برخی از آنها شامل کاراکترهای نامفهوم بودند یا هنگام بارگذاری صفحه تغییر مسیر می‌دادند. تغییر مسیر باعث می‌شود هنگام استخراج متن از وب‌سایت، به جای متن اصلی صفحه، متن موجود در زمان تغییر مسیر استخراج شود. در این فرآیند، وب‌سایت‌هایی که متن استخراج شده از آنها شامل کاراکترهای نامفهوم یا متن غیر از متن اصلی بود، به‌صورت دستی از مجموعه داده حذف شده‌اند. در مجموع ۱۱۱۸ متن اسپم و غیر اسپم استخراج شده بود؛ که این مقدار پس از تمیز کردن دادگان به ۹۲۵ متن رسید.

تمامی محتواهای به دست آمده، با استفاده از کتابخانه Parsivar (Mohtaj et al., 2018)، نرمال‌سازی شده‌اند. علت این کار این است که تمام متون استخراج شده از یک استاندارد یکسان پیروی کنند و یکپارچه شوند. این کار بخشی از فرآیند پیش‌پردازش دادگان قبل از آموزش مدل می‌باشد.

در جدول ۱، آمار مربوط به مجموعه داده قابل مشاهده است.

۲.۳ مدل یادگیری ماشین

یادگیری نظارت‌شده یکی از انواع روش‌های یادگیری ماشین است که خروجی آن واضح یا برچسب‌زده‌شده می‌باشد. یکی از دسته‌های این روش، طبقه‌بندی[۴] نام دارد. طبقه‌بندی در مسائلی که خروجی آنها مقادیر محدودی دارد، قابل استفاده است. در این مقاله، با توجه به این‌که خروجی مدل اسپم بودن یا نبودن یک وب سایت است، تعداد خروجی‌ها محدود به ۲ بوده و مدل مورد نیاز آن در دسته الگوریتم‌های طبقه‌بندی قرار می‌گیرد.

مدل Naïve Bayes یک روش طبقه‌بندی بر اساس تئوری بیز میباشد. تئوری بیز احتمال رخ دادن یک پیشامد را هنگامی که پیشامد دیگر اتفاق افتاده باشد، به دست می‌آورد. مدل Bayes Na ve برمبنای احتمال شرطی است. در این مدل، فرض بر این است که ویژگی‌ها نسبت به یکدیگر مستقل‌اند (Gandhi, 2018).

به‌صورت کلی، سه نوع مدل Naïve Bayes وجود دارد (Saxena, 2021):

- Gaussian Naïve Bayes (GaussianNB): این مدل برای داده‌های پیوسته که از توزیع گاوسی (نرمال) پیروی می‌کنند، بهترین انتخاب است.
- Multinomial Naïve Bayes (MultinomialNB): این مدل هنگام برخورد با داده‌های گسسته، مانند شمارش فرکانس، مفید است. معمولا در موارد استفاده پردازش زبان طبیعی مانند طبقه‌بندی اسپم به کار برده می‌شود.
- Bernoulli Naïve Bayes (BernoulliNB): این مدل با متغیرهای بولی، یعنی متغیرهایی با دو مقدار، مانند True و False یا ۰ و ۱ استفاده می‌شود.

با توجه به موارد بررسی شده، نوع Multinomial Naïve Bayes برای استفاده در تشخیص وب‌سایت‌های اسپم، انتخاب مناسب‌تری است.

برای آموزش مدل، مجموعه داده به دو قسمت آموزش و تست تقسیم شده و سپس برای هر قسمت مقدار کوله کلمات، محاسبه شده است. پس از آن، مدل Multinomial Naïve Bayes با استفاده از مجموعه داده آموزشی، آموزش دیده است.

۴ ارزیابی

برای ارزیابی و مقایسه، احتیاج به مدلی مرتبط که روی زبان فارسی آموزش داده شده باشد، وجود داشت. از مدل‌های ذکر شده در بخش ۲، تنها مدل معرفی شده توسط (Agrawal, 2023) که روی زبان انگلیسی آموزش داده شده، در دسترس بود. در راستای انجام ارزیابی، این مدل بار دیگر روی زبان فارسی با استفاده از مجموعه داده معرفی شده در این مقاله، آموزش داده شد. مجموعه داده تست که شامل ۲۳۲ وب‌سایت

[4] Classification

جدول ۲: ارزیابی مدل با استفاده از مجموعه داده تست

سیستم ارزیابی	مدل (Agrawal, 2023)	مدل پیشنهادی این مقاله
(macro) F۱-score	۶۸۱۷.۰	**۹۶۲۴.۰**
(micro) F۱-score	۸۰۲۹.۰	**۹۶۵۵.۰**
(weighted) F۱-score	۷۶۷۲.۰	**۹۶۵۲.۰**

می‌شود، برای محاسبه F1-score برای هر دو مدل به کار رفته است. نتایج در جدول ۲ قابل بررسی هستند. بهترین نتیجه در هر سیستم، پررنگ شده است.

۵ نتیجه‌گیری

در این مقاله، ابتدا انواع اسپم بررسی شد و سپس یک مجموعه داده، شامل لینک وب‌سایت‌های اسپم و غیر اسپم به همراه برچسب مربوطه و محتوای هر لینک، جهت وب فارسی، معرفی گردید. در نهایت، با استفاده از مجموعه داده، مدل Multinomial Naïve Bayes برای طبقه‌بندی وب‌سایت‌ها به دو دسته اسپم و غیر اسپم، آموزش دیده است. در ارزیابی این مدل، مقدار ۰/۹۶۵۵ برای معیار F1-score ، به دست آمده است؛ که این نتیجه بسیار قابل قبول بوده و مدل می‌تواند در راستای شناسایی وب‌سایت‌های اسپم با استفاده از محتوای آن‌ها و در نتیجه آن، افزایش اعتماد کاربران و جلوگیری از گمراه شدن جویشگرهای وب، به کار رود.

کد منبع این مقاله و همچنین مجموعه داده معرفی‌شده، در گیت‌هاب[5] برای استفاده توسط دیگر محققان در کارهای آینده، منتشر شده است.

سپاس‌گزاری

با تشکر از آقایان محسن محمدی، شایان داودی و عماد نعمتی که در جمع‌آوری بک لینک‌های اسپم ما را یاری کردند.

مراجع

[1] Agrawal, I. (2023, March 29). Spam-Detection-Model. Github. https://github.com/ishitvaagrawal

[2] A. Ntoulas, M. Najork, M. Manasse and D. Fetterly, “Detecting Spam Web Pages through Content Analysis”, The Web Conference, 2006.

[3] Brownlee, J. (2019, August 7). A Gentle Introduction to the Bag-of-Words Model. Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-bag-words-model/

[5] https://github.com/Saba-Heydaridoost/spam-detection

[4] Cojocariu, A. (n d). 5 Common Types of Spam & How You Can Protect Yourself Against Them. COGNITIVESEO. https://cognitiveseo.com/blog/18718/5-common-types-spam-can-protect/

[5] encyclopedia by Kaspersky. (n d). What is Spam?. encyclopedia by Kaspersky. https://encyclopedia.kaspersky.com/knowledge/what-is-spam/

[6] Gandhi, R. (2018, May 5). Naïve Bayes Classifier. Towards Data Science. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

[7] H. Najadat and I. Hmeidi, "Web Spam Detection Using Machine Learning in Specific Domain Features", Journal of Information Assurance and Security, vol. 3, pp. 220–229, 2008.

[8] Internet Society. (2014, July 27). What Is Spam. Internet Society. https://www.internetsociety.org/resources/doc/2014/what-is-spam/

[9] J. Karimpour, A. Noroozi, S. Alizadeh, "Web Spam Detection by Learning from Small Labeled Samples", International Journal of Computer Applications, vol. 50–No. 2, 2022.

[10] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, R. Baeza-Yates, "Using Rank Propagation and Probabilistic Counting for Link-Based Spam Detection", 2006.

[11] M. Danandeh Oskuie and N. Rasavi, "A Survey of Web Spam Detection Techniques", International Journal of Computer Applications Technology and Research, vol. 3, pp. 180–185, 2014.

[12] Saxena, S. (2021, April 6). Inroduction to Naïve Bayes Algorithm. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/03/introduction-to-naive-bayes-algorithm/

[13] S. Mohtaj, B. Roshanfekr, A. Zafarian, H. Asghari, "Parsivar: A Language Processing Toolkit for Persian", LREC, pp. 1112–1118, 2018.

[14] V. Prieto, M. Alvarez, F. Cacheda, "SAAD, a content based Web Spam Analyzer and Detector", Journal of Systems and Software, vol. 86, pp. 2906–2918, 2013.