فصلنامه علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

سال هشتم، شماره‌های ۲۷ و ۲۸، بهار و تابستان ۱۳۹۵
صص: ۲۱- ۳۶

ارائه روشی مناسب برای دسته‌بندی نامه‌های الکترونیکی تبلیغاتی بر مبنای پروفایل کاربران

محمد فتحیان* رحیم حضرتقلی‌زاده**

* استاد، دانشکده مهندسی صنایع، دانشگاه علم و صنعت ایران
** کارشناسی ارشد، مهندسی فناوری اطلاعان، دانشگاه علم و صنعت ایران

تاریخ دریافت:۹۲/۰۹/۲۵ تاریخ پذیرش: ۹۵/۰۳/۱۸

چکیده

به طور کلی، تعریف هرزنامه در ارتباط با رضایت یا عدم رضایت گیرنده است نه محتوای نامه الکترونیکی. بر طبق این تعریف، مشکلاتی در دسته‌بندی نامه‌های الکترونیکی در بازاریابی و تبلیغات مطرح می‌شود. برای مثال امکان دارد بعضی از نامه‌های الکترونیکی تبلیغاتی، برای عده‌ای از کاربران هرزنامه و برای عده‌ای دیگر هرزنامه نباشد. برای مقابله با این مشکل با توجه به پروفایل و رفتار کاربران، ضد هرزنامه‌های شخصی طراحی می‌شود. به طور عادی برای دسته‌بندی هرزنامه‌ها، روشهای یادگیری ماشینی با دقت خوب به کار می‌رود. اما در هر حال یک روش منحصر به فرد موفق بر مبنای دیدگاه تجارت الکترونیک وجود ندارد. در این مقاله ابتدا پروفایل جدیدی برای شبیه‌سازی بهتر رفتار کاربران، تهیه می‌شود .سپس این پروفایل همراه با نامه‌های الکترونیکی به دانشجویان ارائه شده و پاسخ آنها جمع‌آوری می‌گردد. در ادامه برای دسته‌بندی نامه‌های الکترونیکی، روشهای مشهور به ازای مجموعه داده‌های مختلف آزمایش می‌شود .سرانجام، با مقایسه معیارهای ارزیابی داده کاوی، شبکه عصبی به عنوان بهترین روش با دقت بالا، تعیین می‌گردد.

واژه‌های کلیدی: تجارت‌الکترونیکی، تبلیغات الکترونیکی، دسته‌بندی هرزنامه‌ها، داده‌کاوی، پروفایل

۱. مقدمه

امروزه نامه‌های الکترونیکی[۱] یکی از راههای عمومی، تاثیرگذار و با هزینه پایین در سطح اینترنت می‌باشد که با سرعت زیادی در حال رشد است. یکی از زمینه‌هایی که به وفور از نامه‌های الکترونیکی استفاده می‌شود، حوزه تجارت الکترونیک[۲] مخصوصا بازاریابی و تبلیغات اینترنتی است. این نوع بازاریابی و تبلیغات با عنوان "بازاریابی از طریق نامه‌های الکترونیکی[۳]" نیز مشهور است.

همزمان با رشد استفاده از نامه‌الکترونیکی سوءاستفاده و فریبکاری نیز به تبع آن بالا می‌رود. یکی از نمونه‌های

[۱] Electronic Mail

[۲] Electronic Commerce
[۳] Email marketing

نویسنده عهددار مکاتبات: رحیم حضرتقلی‌زاده h.golizadeh@gmail.com

سوءاستفاده از این روش ارتباطی، ارسال کورکورانه نامه‌های الکترونیکی ناخواسته و بی‌دعوت به نام هرزنامه[۴] می‌باشد[۱،۲،۶،۱۲،۲۰،۲۱،۲۳،۲۴].

تعاریف زیادی برای اسپم یا هرزنامه و چیستی و تفاوت آن با نامه‌های معتبر[۵] وجود دارد. کوتاهترین تعریف متداول از بین تعاریف موجود در مورد هرزنامه، آنرا به عنوان یک نامه الکترونیکی ناخواسته[۶] بیان می‌کند. با این حال تعاریف مشابه زیادی نیز وجود دارد که بیان می‌کند، هرزنامه یک نامه الکترونیکی ناخواسته است که به طور نا مشخص و مستقیم یا غیرمستقیم توسط فردی که نسبتی با گیرنده ندارد، فرستاده شده است. همان‌طور که می‌توان دید نقطه اشتراک برای تعریف هرزنامه ناخواسته بودن آن است. بر طبق تعریف مورد توافق، هرزنامه درباره رضایت یا عدم رضایت است نه محتوا [۱،۴،۱۷،۲۳] . هرزنامه‌ها مشکلات متعددی را به بار می‌آورند که برخی از آنها مستقیما باعث ضررهای اقتصادی می‌شوند، مانند ایجاد ترافیک و اتلاف پهنای باند و برخی دیگر زمان زیادی را تلف می‌کنند تا کاربران نامه‌های زاید را جداسازی کنند. علاوه بر موارد بیان شده، بعضی از هرزنامه‌ها باعث آزار روحی و ایجاد عدم امنیت و اطمینان می‌شوند و سرانجام باعث ایجاد مشکلات قانونی مانند تبلیغات هرمی وکلاهبرداری‌های اقتصادی می‌گردند [۲،۳،۴،۷،۱۲،۲۴].

برای رفع این مشکلات، روشهای زیادی را در مقالات مختلف مطرح کرده‌اند، تا با بالا بردن درصد تخمین و دقت، باری از این هزینه‌ها کم کنند و آرامش و اطمینان را برای کاربران در تمامی حوزه‌ها بوجود بیاورند. با این همه به نظر می‌رسد هنوز هم مشکلاتی در این راه وجود داشته باشد. یکی از این مشکلات وجود خطای زیاد در روشهای مطرح شده می‌باشد، که می‌تواند عامل تاثیرگذار در تجارت الکترونیک باشد. اگر در اینجا دسته هرزنامه‌ها را به عنوان کلاس مثبت و دسته نامه‌های الکترونیکی معتبر را به عنوان کلاس منفی در نظر بگیریم، آنگاه خطای FP [۷] شامل نامه‌های الکترونیکی می‌شود که به اشتباه جزو هرزنامه‌ها دسته‌بندی می‌گردند. خطای FN[۸] هم شامل نامه های الکترونیکی می‌شود که به اشتباه جزو نامه‌های الکترونیکی معتبر دسته‌بندی می‌شود. این خطاها در زمینه بازاریابی و تبلیغات از طریق نامه‌های الکترونیکی مشهودتر است. در مواجهه با این مشکلات بعضی از شرکتها اقدام به طراحی ضد هرزنامه سازگار با زمینه تبلیغات می‌کنند [۵،۲۳].

از آنجاییکه بیشتر هرزنامه‌ها در حوزه بازاریابی و نامه‌های الکترونیکی مطرح می‌شود لذا لازم است، که در طراحی ضد هرزنامه‌ها[۹] دید صحیحی نسبت به حوزه تجارت الکترونیک داشته باشیم. در صورت نداشتن چنین دیدی در طراحی، ضد هرزنامه‌ها با سرویس‌دهنده‌های نامه‌های الکترونیکی تبلیغاتی و بازاریابی هماهنگ نخواهند بود لذا درصد زیادی از نامه‌های الکترونیکی منتشر شده از این سرویس‌دهنده‌ها به جای هرزنامه فیلتر[۱۰] شده و هزینه زیادی را به این سرویس‌دهنده‌ها تحمیل می‌کنند. در صورتی که اگر این ضد هرزنامه‌ها برای دیدگاه خاص مانند تبلیغات از طریق نامه‌های الکترونیکی طراحی شوند و اهداف مشخصی را دنبال کنند بهتر عمل خواهند کرد[۴،۷] .

مشکل مهم دیگر که بیشتر در حیطه تجارت الکترونیک و تبلیغات از طریق نامه‌های الکترونیکی مطرح می‌شود، در نظر گرفتن مطلق یک نامه الکترونیکی خاص به عنوان هرزنامه یا نامه معتبر اشد. این در حالی است که امکان دارد بعضی از نامه‌های الکترونیکی برای عده‌ای از کاربران هرزنامه و برای عده‌ای دیگر هرزنامه نباشد. برای مثال در تبلیغات از طریق نامه‌های الکترونیکی امکان دارد خرید اتومبیل برای کسی که قصد خرید اتومبیل دارد هرزنامه حساب نشود در صورتیکه برای بعضی دیگر که قصد خرید ندارند هرزنامه حساب شود. پس در این شرایط دسته‌بندی نامه‌های الکترونیکی دچار مشکل می‌شود که این عدم تخمین به صورت مطلق، در نامه‌های الکترونیکی تبلیغاتی فراوان وجود

---

[۴] Spam
[۵] Legitimate , Ham
[۶] Unsolicited E-mail
[۷] False Psitive
[۸] False Negative
[۹] Anti-spam
[۱۰] Filter

دارد، که به آنها نامه‌های الکترونیکی خاکستری[11] نیز می‌گویند. برای مقابله با این مشکل با توجه به رفتار کاربران اقدام به ساخت ضدهرزنامه‌های شخصی[12] می‌کنند که در مقالات متعددی در مورد روشهای مختلف بحث شده است[۲،۴،۵،۲۰،۲۱]

در این مقاله ما ابتدا پروفایل[13] جدیدی را که می‌تواند به شبیه‌سازی بهتر رفتار کاربران منجر شود تهیه می‌کنیم. سپس این پروفایل را همراه با نامه‌های الکترونیکی تبلیغاتی ساختگی در حیطه کتابفروشی بر خط به دانشجویان ارائه کرده و پاسخ آنها را جمع‌آوری می‌کنیم. در ادامه روشهای موجود و مشهور برای دسته‌بندی نامه‌های الکترونیکی را مورد آزمایش و مقایسه قرار می‌دهیم . اجرا و پیاده‌سازی روشهای انتخاب شده در نرم‌افزار کلمنتاین[14] انجام می‌پذیرد. در پایان به تجزیه و تحلیل هر یک از این روشها پرداخته و روش مناسب را برای زمینه بازاریابی و شخصی‌سازی انتخاب می‌کنیم.

سازماندهی بخشهای بعدی به این شکل می‌باشد : در بخش ۲ مقاله، به بررسی کارهای مرتبط و دسته‌بندی آنها می‌پردازیم سپس در بخش ۳ به بیان طراحی پروفایل و تولید داده‌ها و مشخصات جامعه آماری آن می‌پردازیم. در ادامه در بخش ۴ روش پیشنهادی خود را مطرح می‌کنیم. در بخشهای باقیمانده به ارزیابی نتایج و جمع‌بندی و ارائه پشنهادات برای کارهای آتی می‌پردازیم.

۲. مروری بر ادبیات موضوع

به طور کلی برای دسته بندی و پیش بینی هرزنامه ها، تکنیک ها و روشهای زیادی مطرح شده است. دسته بندی نامه‌های الکترونیکی ، یک برنامه کاربردی است که بر اساس تابع (۱) پیاده‌سازی می‌شود :

$$f(m,\theta)=\begin{cases} C_{spam} \\ \\ C_{leg} \end{cases} \Rightarrow \begin{cases} if spam \\ \\ else \end{cases} \qquad (۱)$$

در این تابع m نامه الکترونیکی است که باید دسته‌بندی گردد. بردار پارامتر θ حاصل آموزش دسته بند با استفاده از یک مجموعه داده است که قبلاً جمع‌آوری شده است که می‌توان آنرا به صورت فرمول(۲) بیان کرد :

$$\theta=\Theta(M), \qquad (۲)$$
$$M=\{(m_1,y_1),.....,(m_n,y_n)\}, y_i \in \{C_{spam},C_{leg}\}$$
$$\forall i, i=1,2,.....,n$$

در این فرمول mها نامه‌های الکترونیکی هستند که قبلا جمع‌آوری شده‌اند و yها نیز برچسب متناظر آنها می‌باشد[۱]. عمده کارهای انجام شده در این زمینه را می‌توان به صورت جدول ۱ دسته‌بندی کرد. در این جدول بعضی از روشهای کلی توضیح داده شده است. روشهای بیان شده در این جدول مخصوصاً روشهای یادگیری ماشینی[15] از نظر تخمین و دقت پیش‌بینی نتیجه مطلوبی را در برداشته‌اند. با این همه، برای مقابله با مشکلات اصلی بیان شده لازم است که ضد هرزنامه‌های شخصی در حوزه بازاریابی و تبلیغات و سازگار با این حوزه تولید شود. در زمینه تولید ضد هرزنامه‌های شخصی شده بر مبنای پروفایل و رفتار کاربران، بعضی از کارهای انجام شده را می‌توان به صورت زیر بیان کرد هرچند این تحقیقات نیز به صورت تخصصی به حوزه بازاریابی و تبلیغات از طریق نامه‌های الکترونیکی نپرداخته است.

سوسا و همکاران[16]یک روش تولید ضد هرزنامه شخصی شده همکارانه[17] را بررسی کرده‌اند. در این روش ابتدا پروفایل کاربران دسته‌بندی می‌شود سپس بر مبنای گزاراشات رسیده از هر گروه دسته‌بندی نامه‌های الکترونیکی انجام می‌پذیرد. بدیهی است، در این روش نیاز به انتقال اطلاعات مابین سرویس‌دهنده‌های مختلف می‌باشد که این انتقال از

[11] Gray , Grey
[12] Personalized Anti-spam
[13] Profile
[14] Clementine ۱۲،۰
[15] Machine learning
[16] Sousa and et al.
[17] Collaborative

طریق معماری p2p[۱۸] انجام می‌پذیرد. این ضد هرزنامه به صورت غیر متمرکز بوده که در سرویس دهنده‌های نامه‌های الکترونیکی اجرا می‌شود. از معایب این روش می‌توان به نیاز به امنیت در حین انتقال اطلاعات، نیاز به پهنای باند بیشتر برای انتقال اطلاعات و سختی مدیریت غیر متمرکز را نام برد[۳،۷].

در مقابل راوی و همکاران [۱۹]، روش متمرکز دیگری را بیان کردند که در این روش در دو مرحله با استفاده از روش‌های شبکه عصبی و در سرور[۲۰] نامه‌های الکترونیکی طرف فرستنده اجرا می‌شود. این عمل از اتلاف پهنای باند در ازای پیشگیری ازانتقال هرزنامه‌ها جلوگیری می‌کند و به شکل کاملا منطبق با افکار انسان شکل گرفته است. اما باز هم دارای معایبی از جمله متکی بر رفتار اشخاص خاص به خاطر عدم دستیابی به کل جامعه آماری دارد[۲۲].

از دیگر روشهای اجرا شده استفاده از درخت تصمیم مانند C۴،۵ می‌باشد. در این روش ییح و همکاران [۲۱]، پروفایل کاربران را همراه با نامه‌های الکترونیکی پاسخ داده شده جمع‌آوری کرده‌اند. سپس از طریق روش فرکانس معکوس سند (TF-IDF[۲۲]) به استخراج و انتخاب ویژگیهای مناسب پرداخته است. در حقیقت این مرحله توکن کردن نام دارد که در آن متن نامه‌های الکترونیکی به ریشه کلمات اصلی تبدیل شده و کلمات پرکاربرد به صورت یک مقدار باینری در نظر گرفته می‌شود. در صورت وجود این کلمه در یک متن مقدار آن یک و در غیر اینصورت مقدار آن برابر صفر در نظر گرفته می‌شود. بعد از این مرحله با استفاده از درخت تصمیم به تولید قوانین می‌پردازند. این قوانین تولید شده از جهت دقت [۲۳]پیش‌بینی، مورد بررسی قرار گرفته و قوانین با دقت بالا انتخاب می‌شود[۱۳].

یکی دیگر از روشهای مشابه در این زمینه استفاده از تولید داده‌های ساختگی برای نامه‌های الکترونیکی توسط کیم و همکاران[۲۴]می‌باشد. در این تحقیق به خاطر نیاز به پروفایل کاربران همراه با پاسخ نامه‌های الکترونیکی تبلیغاتی، لازم بود به تولید ساختگی این موارد در قالب پرسشنامه پرداخته شود. در این تحقیق از درخت تصمیم همراه با روشهای معنایی مانند روش قبلی استفاده شده است. تفاوت این تحقیق با روش قبلی در مرحله استخراج و انتخاب ویژگیها می‌باشد. این روش در مقایسه با روش قبلی به خاطر در نظر گرفتن شخصی‌سازی و خصوصیات نامه‌های الکترونیکی خاکستری می‌تواند نمونه مناسبی برای حوزه تجارت الکترونیک باشد. از جمله معایب این روش را می‌توان عدم تطابق بین پروفایل و محتوای نامه‌های الکترونیکی ساختگی با جامعه آماری پاسخ‌دهندگان بیان کرد. در این تحقیق تنها به برچسب زدن نامه‌های الکترونیکی تبلیغاتی در ۱۰ دسته توسط کارشناس انسانی[۲۵] و ارائه آن به دانشجویان برای پاسخگویی پرداخته‌اند. به نظر می‌رسد که این دسته‌بندی جزئی بوده و جامعه آماری پاسخ‌دهنده نتواند جواب دقیق و واقعی را ارائه کند[۵].

تغییرات نتایج پیش‌بینی در بین روشهای مرسوم تا حدودی زیاد است. این نتیجه به خاطر انتخاب ویژگیهای متفاوت روشها اتفاق می‌افتد. هر چند عمل مقایسه با دیگر کارهای مشابه به خاطر متفاوت بودن مجموعه داده‌ها و روش پیشنهادی نمی‌تواند مبنایی برای ارزیابی دقیق باشد اما در جدول ۲ به مقایسه اجمالی می‌پردازیم. اگر چه استفاده از تمامی محتوای نامه‌های الکترونیکی می‌تواند دقت نتایج را افزایش دهد اما به خاطر درگیر بودن با مجموعه بزرگی از ویژگیهای استخراج شده، معمولا نیاز به انتخاب ویژگیهای مناسب می‌باشد. این حالت به ازای زمان و فضای مصرفی زیاد می‌تواند دقت پیش‌بینی را تا حدودی بهبود بخشد.

در هر حال روشهای مختلفی برای شخصی‌سازی ضد هرزنامه‌ها وجود دارد که در آن از الگوریتم‌های یادگیری ماشینی و داده‌کاوی مانند شبکه بیزین[۲۶]، شبکه‌های

[۱۸] Peer-to-peer
[۱۹] Ravi and et al.
[۲۰] Server
[۲۱] Yih and et al.
[۲۲] Term Frequency- Inverse Document Frequency
[۲۳] Accuracy
[۲۴] Kim and et al.
[۲۵] Human Expert
[۲۶] Bayesian network

عصبی[۲۷]، درخت تصمیم[۲۸] و SVM[۲۹] استفاده می‌کنند [۱،۵،۶،۸،۱۲،۱۶،۱۷،۱۹،۲۴]. با این همه هیچ یک از این موارد به طور تخصصی به امر تولید ضد هرزنامه‌های شخصی‌سازی شده با نگرش نامه‌های الکترونیکی خاکستری، بازاریابی و تبلیغات نپرداخته است. به خاطر اهمیت این حوزه و اینکه اکثر نامه‌های الکترونیکی امروزه را نامه‌های الکترونیکی خاکستری مخصوصا تبلیغاتی تشکیل می‌دهد نیاز به تولید ضد هرزنامه ساز گار با این حوزه احساس می‌شود تا از این رهگذر نامه‌های الکترونیکی بازاریابی در سرویس‌دهنده‌های پست الکترونیکی هدف به عنوان هرزنامه فیلتر نشوند. از بررسی روشهای بیان شده به این نتیجه می‌رسیم که برای طراحی و تولید ضد هرزنامه‌ها ۴ گام اصلی ضروری می‌باشد که در جدول ۳ می‌توان مشاهده کرد[۱۲]:

۳. استخراج ویژگی‌ها و آماده‌سازی داده‌ها

برای انجام این تحقیق لازم است که در ابتدا نامه‌های الکترونیکی همراه با برچسب (هرز یا معتبر بودن که از طرف کاربران نسبت داده شده است) و پروفایل کاربران جمع‌آوری شود. اگر چه مجموعه نامه‌های الکترونیکی با برچسب قابل اطمینان برای آزمون و آزمایش در شرکتهای تحقیقاتی موجود می‌باشد، اما این مجموعه نامه‌های الکترونیکی یا فاقد پروفایل کاربران می‌باشند و یا در زمینه تبلیغات الکترونیکی نمی‌باشند. لذا از آنجایی که این اطلاعات در اکثر اوقات در دسترس نمی‌باشد باید مانند مقالات موجود در این زمینه به تولید آن بپردازیم.

[۲۷] neural network
[۲۸] decision tree
[۲۹] support vector machines

جدول ۱. انواع روشهای مهم برخورد با هرزنامه‌ها

روشهای غیر فیلتری (پیشگیری از ایجاد و انتقال هرزنامه)	روش اقتصادی			هر شخص برای فرستادن ایمیل بایستی هزینه‌ای را بپردازد که برای هرزنامه‌نویسان این هزینه قابل توجه است [۱].
	روش قانونگذاری			بعضی از قانونگذاران برای حفظ امنیت و آرامش در جامعه مجازی اقدام به وضع قوانینی برای جلوگیری از تولید و انتشار هرزنامه‌ها کرده‌اند[۲۳] .
	روش تغییر پروتوکلها			برای رفع نقص پروتوکلهای موجود ،یک گام برای شناسایی هویت ارسال‌کننده نامه‌های الکترونیکی اضافه می شود[۱۸].
روشهای فیلتری (بعد از انتقال هرزنامه سعی دارند که در سرویس دهنده های پست الکترونیکی به دسته بندی نامه های الکترونیکی به دو دسته هرزنامه و نامه معتبر بپردازند هرچند هنوز بعضی از مشکلات بیان شده همچنان وجود دارد [۱,۶,۳,۱۲])	روشهای غیر محتوایی	لیست سیاه و سفید		در این روش لیستی از آدرسهایی که به عنوان مبدا انتشار هرزنامه شناخته شده‌اند، تهیه می شود. هر نامه الکترونیکی که از این آدرسها فرستاده شود توقیف خواهد شد یا در نسخه دیگر فقط از آدرسهای مشخص شده موجود در لیست سفید، نامه الکترونیکی قبول می‌کند[۹].
		شبکه اجتماعی فرستندگان		در این روش فیلدهای From و To و Cc و Bcc را از سرآیند نامه‌های الکترونیکی کاربران استخراج و بررسی می‌کنند سپس با استفاده از آنها گراف روابط اجتماعی کاربران را می‌سازند در نهایت با استفاده از این گراف روابط اجتماعی اقدام به دسته‌بندی نامه‌های الکترونیکی کاربران می‌کنند.
		رفتار فرستندگان		دانشِ راجع به رفتار که در پشت یک پیغام یا مجموعه‌ای ازپیغامها قرار دارد را از دل ویژگی‌های غیرمتنی استخراج می‌کند و سپس آن را با دانش از پیش تعریف شده‌ی (یا استخراج شده) مربوط به کاربرهای طبیعی و یا خرابکار، مقایسه می کنند
	روشهای محتوایی (در این روشها از محتوای نامه های الکترونیکی مانند متن و عنوان آن استفاده می کنند[۱۰,۱۱])	توکن کردن (در این روش با استفاده از ابزار خاص اقدام به جداسازی کلمات مهم کرده و فقط از مهمترین آنها استفاده می کنند[۱,۱۶])	روشهای اولیه	اولین فیلترها به صورت سطحی فقط وجود یا عدم وجود یک سری توکنهای از پیش تعریف شده را در بدنه پیغام بررسی می کردند و بر مبنای یک سری قواعد ثابت عمل می کردند که به روش کلمات کلیدی و آماری معروف بودند که امروزه با روشهای یادگیری جایگزین شده اند .
			یادگیری ماشینی	در این روش برای پیش بینی از روشهای داده کاوی و یادگیری ماشینی بهره می برند به طوریکه این دسته بیشترین کاربرد را امروزه دارد و از تمامی ابزار داده کاوی که جهت دسته بندی و پیش بینی به کار می رود می توان استفاده کرد.
		آنالیز زبانی		این روش ها بر این فرض استوار هستند که بدنه پیغامها به زبان طبیعی می باشند و روش هایی که مبتنی بر مدلهای مقایسه ای – مانند عمل مقایسه ای مارکف و نیز پیش بینی هستند را استفاده می کنند.

جدول ۲. کارهای مشابه انجام‌شده

توضیحات	دقت	مجموعه داده	روش پیشنهادی	مقاله
در این مقاله استخراج ویژگیها با نظر خبرگان و غیر محتوایی می باشد	۹۱,۷۸٪	خصوصیات کلی نامه ها همراه با فیلد های سرآیند	روش ترکیبی	بینگ و همکاران [۱۱]
دقت به ازای هر قانون متفاوت است ما بیشترین را در نظر گرفته ایم.	٪۸۵,۰	تولید ساختگی نامه های الکترونیکی پرسشنامه ای برای صنایع مختلف	روش تولید قوانین از درخت تصمیم همراه با روش معنایی	کیم و همکاران [۵]
از محتوای نامه ها استفاده کرده است	٪۹۳,۵	تولید ساختگی داده ها همراه با داده واقعی	روش همکارانه	سوسا و همکاران [۳]
از کلیه محتوا استفاده کرده و فیلتر کردن بر مبنای کلمات کلیدی خاص	٪۹۷	داده های واقعی همراه با تکرار چند باره	روش درخت تصمیم همراه با آنتولوژی	یاون و همکاران [۱۵]

جدول ۳. گام‌های طراحی ضد هرزه‌نامه

گامهای طراحی	توضیح	روش کار
گام اول : استخراج ویژگیها	در این روش ویژگیهای و کلمات پرکاربرد از متن نامه ها استخراج می شود. در شبیه سازی(تولید مجموعه داده) دیگر نیازی به مرحله استخراج نیست بلکه به صورت پیش فرض و با نظر خبرگان این ویژگیها مستقیماً پرسش می شود.	ابزار مختلفی مانند متن کاوی و TF-IDF و مشابه آن به کار برده می شود.
گام دوم : انتخاب ویژگیها	در این مرحله از بین ویژگیهای مختلف بعضی از ویژگیهای تاثیر گذار و با اهمیت انتخاب می شود.	برای اینکار از ابزاری مانند IG استفاده می کنند که این ابزار از مهمترین روشها بوده و در نرم افزار کلمنتاین ابزاری بر این مبنا وجود دارد.
گام سوم : چارچوب پیشنهادی	در این مرحله به ارائه روشهای خود با استفاده از روشهای مختلف موجود در داده کاوی و یادگیری ماشینی در قالب چارچوب می پردازند.	طراحان سعی می کنند تا از طریق مقایسه یا ترکیب روشها به دقت بیشتری دست پیدا کنند.
گام چهارم : ارزیابی و اعتبار سنجی	دقت در طراحی وابسته به سه مرحله قبلی می باشد لذا ارزیابی تا حدودی از طریق معیارهای ارزیابی مشهور داده کاوی صورت می گیرد.	ارزیابی از طرق معیارها،ارزیابی از طریق مقایسه مولفه های نوآوری شده،

تولید این مجموعه نامه‌های الکترونیکی از طریق بررسی مجموعه مقالات موجود و نظر خبرگان همراه با بررسی نامه‌های الکترونیکی تبلیغاتی صورت می‌پذیرد. روش کار برای تولید نامه‌های الکترونیکی تبلیغاتی در قالب پرسشنامه به شرح زیر می‌باشد. ابتدا طبق تمامی تحقیقات موجود در این زمینه به این مسئله می‌پردازیم که چه ویژگیهایی می‌تواند از نامه‌های الکترونیکی استخراج شود.

سه دسته ویژگی اساسی، شامل موارد زیر را می‌توان نام برد[۱]:

- ویژگیهایی از سرآیند[۳۰]
- ویژگیهایی از متن[۳۱] نامه که شامل موضوع [۳۲]هم می‌باشد
- و یژگیهایی از کل ساختار[۳۳] نامه الکترونیکی

از آنجاییکه هدف ما پالایش و دسته‌بندی نامه‌های الکترونیکی تبلیغاتی می‌باشد لذا لازم است که محتوای نامه‌های الکترونیکی نیز در همین راستا باشد. در کارهای مشابه انجام شده بدلیل اینکه چنین محدودیتی وجود نداشته، نامه‌های الکترونیکی از صنایع مختلف را برای محتوای نامه‌های الکترونیکی در نظر گرفته‌اند.[۵]

در چنین مواردی بدون در نظر گرفتن ویژگیهای جامعه آماری پاسخ‌دهندگان اقدام به تهیه محتوای نامه‌های الکترونیکی کرده‌اند. ما برای دقت در این تحقیق ابتدا جامعه آماری پاسخ‌دهندگان خود را در نظر گرفته و سپس اقدام به تهیه محتوای نامه‌های الکترونیکی می‌کنیم.

جامعه آماری پاسخ‌دهندگان ما را جامعه دانشگاهی و دانشجویان تشکیل می‌دهند لذا لازم است حوزه‌ای را برگزینیم که افراد آگاهی و تمایل نسبت به این حوزه داشته باشند. در نتیجه ما نامه‌های الکترونیکی تبلیغاتی کتابفروشی برخط را به عنوان نمونه مطالعه موردی انتخاب می‌کنیم.

۳-۱- استخراج ویژگی‌های موثر نامه‌های الکترونیکی

بعد از انتخاب حوزه کاری و جامعه آماری اکنون لازم است که محتوایی برای هر یک از ویژگیهای بیان شده استخراج کنیم. برای اینکار تعدادی از نامه‌های الکترونیکی تبلیغاتی در حوزه کتابفروشی برخط را انتخاب می‌کنیم. سپس با نظر خبرگان در امر تبلیغات و همچنین تعدادی از جامعه آماری اقدام به تهیه محتوا برای هر یک از ویژگی‌های مطرح شده می‌پردازیم. جدول ۴ ویژگیهای استخراج شده نامه‌های الکترونیکی همراه با محتوا برای حوزه کاری مطرح شده را نشان می‌دهد.

در ویژگی سرآیند دو گزینه بیان شده است: اگر فرستنده نامه الکترونیکی تبلیغاتی برای گیرنده آشنا باشد (یعنی در دفترچه آدرس شخص موجود باشد که این نوعی تبلیغات از طریق مشتریان سازمانها می‌باشد که می‌توانند کالا یا خدماتی را به دوستان خود سفارش کنند).

اگر آدرس فرستنده برای گیرنده آشنا نباشد (در حقیقت تبلیغات از طریق شرکتهای موجود و با آدرسهای مختلف انجام شود). در ویژگی عنوان و متن گزینه‌های مختلف و جذابی که در نامه‌های الکترونیکی تبلیغاتی می‌تواند وجود داشته باشد بیان شده است. ویژگی بعدی ویژگی‌های کلی نامه‌های الکترونیکی می‌باشد. محتوای نسبت داده شده به این ویژگیها دارای حالت عمومی هستند و می‌توانند در دیگر حوزه‌ها نیز مطرح شوند. به خاطر وجود ویژگیهای زیاد ما بر مبنای نظر خبرگان و اصول بازاریابی مطرح شده روزیتر و بلمن [۳۴][۱۴] ،دو دسته اصلی برای این ویژگیها را در نظر گرفتیم [۱۶].

جدول ۴ :ویژگی استخراج شده بر مبنای حوزه کاری

ویژگی کلی	ویژگی انتخاب شده بر مبنای حوزه کاری	گزینه های موجود
سرآیند	فرستنده نامه الکترونیکی	۱، از طرف آدرس فرستنده آشنا
		۲، از طرف آدرس فرستنده ناآشنا
موضوع	عنوان نامه الکترونیکی	۱، کتابهای رایگان

[۳۰] Header
[۳۱] Bodey
[۳۲] Subject
[۳۳] General structure
[۳۴] Rossiter and Bellman .(۲۰۰۵)

		۲، تازه ترین کتاب
		۳، پرفروشترین کتاب
متن	زمینه و محتوای نامه الکترونیکی	۱، مهندسی و علوم پایه
		۲، پزشکی
		۳، علوم انسانی
		۴، هنر
		۵، سایر(غیر تخصصی و متفرقه)
ساختار کلی	ویژگیهای دسته بندی شده نامه الکترونیکی	۱، سایز حافظه زیاد (شامل گرافیک یا ضمیمه و...)
		۲، سایزحافظه کم (شامل فقط متن و..)

۲-۳- استخراج پروفایل کاربران

مرحله بعدی استخراج ویژگیها که مهترین مرحله نیز می‌باشد پروفایل کاربران می‌باشد. در اکثر تحقیقات از پروفایل‌های استاندارد موجود در اکثر سایتها مانند شغل، جنسیت، تحصیلات، رشته تحصیلی، سن، علاقه‌مندی و غیره استفاده شده است. از آنجاییکه شخصی‌سازی بر مبنای پروفایل کاربران شکل می‌گیرد، لذا لازم است گزینه‌های دیگری نیز برای بالا بردن دقت در نظر گرفته شود. برای اینکار ما دو گزینه را از مقالات مختلف جمع‌آوری کرده و در پروفایل خود قرار می‌دهیم [۱،۵].

گزینه اول تعداد دفعاتی است که یک شخص بعد از دریافت یک نامه الکترونیکی آن را هرزنامه اعلام می‌کند. این گزینه برای افراد مختلف متفاوت است به صورتیکه امکان دارد یک شخص در مرحله اول یک نامه الکترونیکی را هرزنامه اعلام کند در صورتیکه امکان دارد شخص دیگری در دفعات تکرار زیاد نامه الکترونیکی مذکور را هرزنامه اعلام کند. گزینه‌ای مشابه این گزینه با نام قدرت مورد انتظار برای ضد هرزنامه شخصی شده وجود دارد[۵]. علت اصلی قرار دادن چنین گزینه‌ای در پروفایل اشخاص به خاطر نامه‌های الکترونیکی خاکستری می‌باشد.

گزینه دومی که در پروفایل اشخاص قرار داده شده و مورد پرسش قرار می‌گیرد، نسبت خطاهای مورد تحمل شخص در فیلتر کردن است که می‌تواند قبول کند. در حقیقت افراد متفاوت ویژگیهای رفتاری مختلفی دارند. بعضی از افراد عنوان می‌کنند که باید هیچ یک از نامه‌های الکترونیکی معتبر آنها به اشتباه فیلتر نشود و در مقابل دریافت چند هرزنامه روزانه را قبول می‌کنند، مخصوصاً وقتی که این هرزنامه‌ها، نامه‌های الکترونیکی تبلیغاتی باشد. در مقابل بعضی از افراد راضی به دریافت هیچ هرزنامه‌ای نیستند، هر چند بعضی از نامه‌های الکترونیکی معتبر آنها به اشتباه فیلتر شود. در حقیقت با این گزینه دو گروه افراد متفاوت را از لحاظ رفتاری می‌توان تشخیص داد.

این بخش مطابق با گام اول یعنی گام استخراج ویژگیها از نامه‌های الکترونیکی می‌باشد. در اینجا ما نیاز به ابزاری خاص برای توکن کردن و غیرساختاری کردن متن نیاز نداریم. در حقیقت با استخراج ویژگیها و محتوا با روش مطالعه مقالات مشابه ، مطالعه نمونه موردی ، نظر خبرگان و رتبه‌بندی آنها توانستیم به ویژگیهای مورد نیاز خود دست یابیم.

همانطور که ملاحظه می‌شود یک نمونه نامه الکترونیکی از ضرب دکارتی موارد بیان شده از جدول ۴ حاصل می‌شود که تعداد ۲*۳*۵*۲=۶۰ قالب نامه الکترونیکی بدست می‌آید. تعداد سوالات پروفایل نیز برابر با ۱۰ عدد می‌باشد که هریک از آنها مقادیر مختلفی می‌تواند داشته باشد. در نتیجه هر پرسشنامه شامل ۱۰ سوال برای پروفایل کاربران و تعداد ۶۰ نامه الکترونیکی ساختگی و پرچسب پاسخگویی (هرزنامه یا معتبر) می‌باشد.

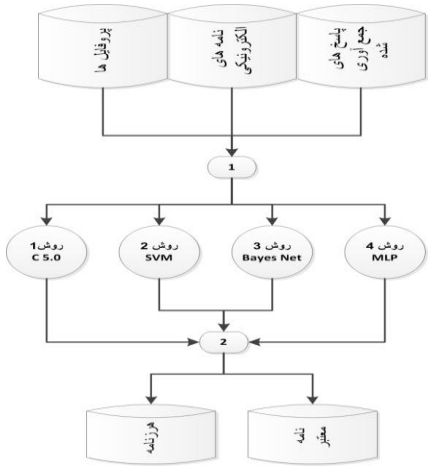
این پرسشنامه بعد از طراحی از طریق وب و در برخی موارد به صورت رودرو توسط ۷۰ نفر از دانشگاهیان پاسخ داده شد، که پس از اعمال پاکسازی تعداد ۶۶ عدد از آنها مورد استفاده قرار گرفت که از این تعداد ۳۰ نفر را زن و ۳۶ را مرد تشکیل می‌دهد. در ادامه ما داده‌های جمع‌آوری شده را به صورت تصادفی درهم کردیم سپس این داده‌ها را به دو قسمت مساوی تقسیم کردیم. در ادامه برای ارزیابی چارچوب از دو نوع مجموعه داده زیر استفاده کردیم. (هر رکورد شامل ۱۰ فیلد پروفایل و ۴ فیلد ویژگی نامه الکترونیکی و یک برچسب پاسخگویی می‌باشد.)

- مجموعه داده نوع اول: تعداد ۱۸۴۳ رکورد که شامل ۱۱۷۲ رکورد هرزنامه می‌باشد.

- مجموعه داده نوع دوم: تعداد ۱۸۴۳ رکورد که شامل ۹۵۹ رکورد هرزنامه می‌باشد.

۴. انتخاب ویژگی‌ها و چارچوب پیشنهادی

بعد از گام اول نوبت به گام دوم می‌رسد. در این گام لازم است که از بین ویژگیهای موجود در پروفایل و ویژگیهای نامه‌های الکترونیکی ساختگی بهترین آنها را برای چارچوب پیشنهادی خود انتخاب کنیم. بدیهی است که تمامی ویژگیهای بیان شده نمی‌تواند برای چارچوب مفید واقع شود. در بعضی موارد حتی مشاهده شده است که وجود بعضی از ویژگیها باعث کاهش دقت شده است. ما برای انتخاب ویژگیهای مناسب از ابزار انتخاب ویژگی موجود در نرم‌افزار کلمنتاین که از روش IG بهره می‌برد، استفاده می‌کنیم. این روش در اکثر مقالات موجود در این زمینه به کار گرفته شده و نتیجه مطلوبی را به همراه داشته است [۵]. در گام سوم روش پیشنهادی خود را در قالب یک چارچوب ارائه می‌دهیم. در این چارچوب سعی می‌کنیم از اکثر روشهای مشهور و زیاد استفاده شده از حوزه داده کاوی، یادگیری ماشینی و آمار استفاده کنیم. شکل ۱ چارچوب پیشنهادی ما را نمایش می‌دهد.



شکل ۱. چارچوب ارائه شده

بخش اول پایگاه داده‌ها می‌تواند هم پایگاه داده مربوط به آموزش و هم آزمایش چارچوب را شامل شود. برای راحتی کار ما در این چارچوب فقط یکی از این دو را نمایش دادیم. اما بدیهی است که بعد از تقسیم مجموعه داده‌های موجود به دو بخش آموزش (۷۰٪ داده ها بر اساس اکثر تحقیقات) و آزمایش می‌توانیم هر کدام از آنها را به چارچوب اعمال کنیم که در بخش آموزش چارچوب ایجاد می‌شود و در بخش آزمایش به ارزیابی چارچوب می‌پردازیم.

بخش اول پایگاه داده‌ها در کل شامل موارد زیر می‌باشد:

پروفایل کاربران: این پروفایلها از پاسخ‌دهندگان به نامه‌های الکترونیکی ساختگی جمع‌آوری شده است.

نامه‌های الکترونیکی: این پایگاه داده نتیجه مطالعات، نظر خبرگان، بررسی نامه‌های الکترونیکی مختلف در حیطه تبلیغات می‌باشد.

پاسخهای جمع‌آوری شده: این بخش برچسب نسبت داده شده از طرف پاسخ‌دهندگان به نامه‌های الکترونیکی شبیه‌سازی شده می‌باشد. در حقیقت شامل دو گزینه نامه معتبر یا هرزنامه می‌باشد.

بخش دوم پایگاه داده‌ها شامل پیش‌بینی‌های چارچوب می‌باشد این قسمت شامل دو پایگاه داده به صورت زیر می‌باشد:

پایگاه داده معتبر: این پایگاه داده در حقیقت شامل پیش‌بینی نامه‌های معتبر چارچوب می‌باشد.

پایگاه داده هرزنامه: این پایگاه داده شامل نامه‌های الکترونیکی است که چارچوب آنها را به عنوان هرزنامه شناخته و برای استفاده آتی در اینجا ذخیره کرده است.

مسیر ۱ از هر سه پایگاه داده موجود در چارچوب استفاده می‌کند. این مسیر خود به تنهایی به ۴ مسیر فرعی منشعب می‌شود. در هر مسیر فرعی چارچوب از روشهای داده کاوی و یادگیری ماشینی برای پیش‌بینی استفاده شده است. هر روش شامل دو قسمت انتخاب ویژگیها از میان ویژگیهای استخراج شده و انجام اعمال پیش‌بینی را شامل می‌شود. در مسیر ۲ عمل مقایسه ما بین نتایج بدست آمده صورت می‌گیرد. چهار روش یا الگوریتم استفاده شده در این مسیر به شرح زیر می‌باشد[۶،۱۰،۱۲،۱۶] :

مسیر ۱ از هر سه پایگاه داده موجود در چارچوب استفاده می کند. این مسیر خود به تنهایی به ۴ مسیر فرعی منشعب می‌شود. در هر مسیر فرعی چارچوب از روشهای داده‌کاوی و یادگیری ماشینی برای پیش‌بینی استفاده شده است. هر روش شامل دو قسمت انتخاب ویژگیها از میان ویژگیهای استخراج شده و انجام اعمال پیش‌بینی را شامل می‌شود. در مسیر ۲ عمل مقایسه ما بین نتایج بدست آمده صورت

می‌گیرد. چهار روش یا الگوریتم استفاده شده در این مسیر به شرح زیر می‌باشد[۶،۱۰،۱۲،۱۶] :

C۵,۰: روش یا الگوریتم اول که همان درخت تصمیم نیز می‌باشد به وفور و در منابع مختلف برای اعمال پیش‌بینی استفاده می‌شود. برای ایجاد این درخت روشهای زیادی وجود دارد که امروزه با نرم‌افزارهای موجود و در دسترس به سادگی می‌توان از C۴,۵ و یا C۵,۰ استفاده کرد. در اینجا ما از C۵,۰ که در نرم‌افزار کلمنتاین وجود دارد استفاده کردیم.

SVM: روش یا الگوریتم دوم که در اینجا استفاده می‌شود SVMمی‌باشد. این روش برای پیش‌بینی بعضی از ویژگیها مخصوصا در حیطه تصاویر کاربرد دارد.

BN: روش یا الگوریتم سوم همان شبکه بیزین می‌باشد. این الگوریتم در روشهای محتوایی به فراوان و کرات مورد استفاده قرار گرفته است. این روش بیشتر مطابق با روشهای آماری و یادگیری ماشینی می‌باشد.

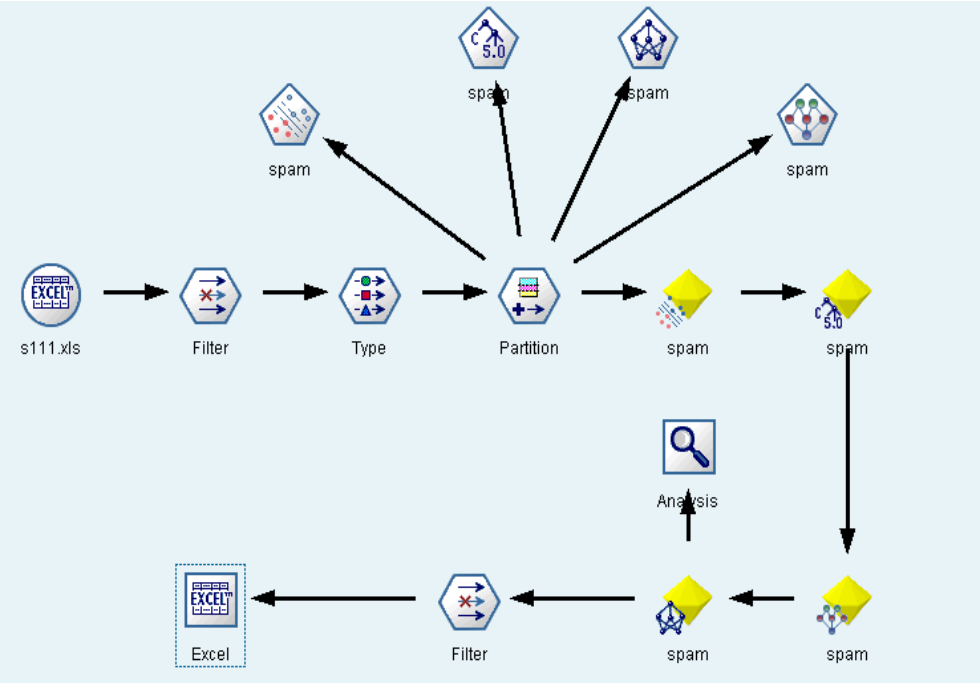
MLP: این روش یا الگوریتم همان روش شبکه‌های عصبی برای پیش‌بینی می‌باشد. انواع روشهای مختلفی برای پیش‌بینی با استفاده از شبکه‌های عصبی موجود می‌باشد. اما از میان روشهای مختلف موجود و ابزار موجود در نرم‌افزار کلمنتاین روشی را برمی‌گزینیم که نسبت به روش‌های دیگر از نظر زمانی همخوانی داشته باشد. ما بیشتر از روش شبکه‌های عصبی چند لایه استفاده می‌کنیم که MLP[۳۵] می‌تواند بهترین مورد هم از نظر زمان و هم از نظر پیش‌بینی باشد.

۵. ارزیابی نتایج و تفسیر آنها

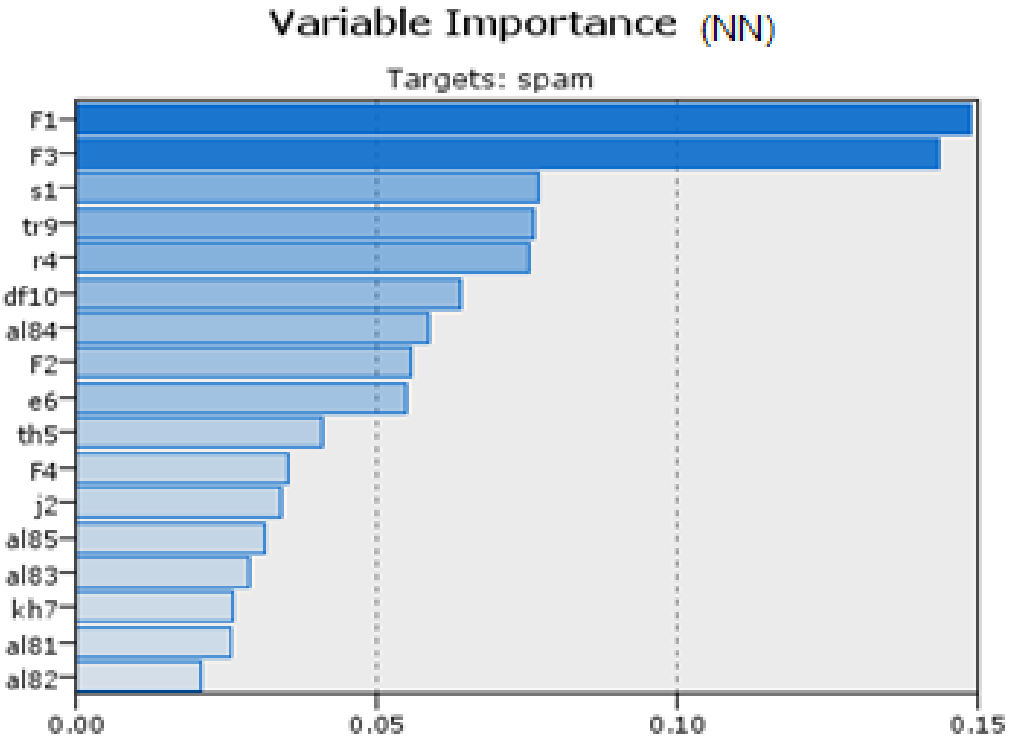
بعد از اینکه داده‌های مورد نیاز خود را جمع‌آوری و پردازش کردیم اکنون نوبت به اجرا و پیاده‌سازی چارچوب ارائه شده مـی‌رسد. ما اطـلاعات مورد نیاز برای ارزیابی، نتیجه‌گیری و تفسیر نتایج را از اجرا و پیاده‌سازی چارچوب در نرم‌افزار کلمنتاین بدست می‌آوریم. نمودار کلی برای اجرای این مرحله را، می‌توان به صورت شکل ۲ در نرم‌افزار کلمنتاین نمایش داد. این ساختار برای هر دو مجموعه داده موجود یکسان بوده و مبنایی برای اجرا و پیاده‌سازی این روشها می‌باشد. در این نمودار ابتدا مجموعه داده وارد نرم‌افزار می‌شود. سپس اگر نیازی به فیلتر کردن بعضی از ویژگیها باشد اقدام به فیلتر کردن ویژگی مورد نظر می‌کنیم. در گره بعدی نوع داده‌های ویژگیها را برای نرم‌افزار مشخص می‌کنیم. مهمترین ویژگی که نوع داده آن باید به درستی مشخص شود، ویژگی مورد پیش‌بینی می‌باشد. گره بعدی گره پارتیشن می‌باشد. این گره وظیفه انتخاب تصادفی مجموعه آموزش (در اینجا ۷۰٪ مجموعه) و مجموعه آزمایش را دارد. گره‌های بعدی چهار روش عنوان شده در چارچوب می‌باشد. بعد از این مرحله پیاده‌سازی خود را در این نرم‌افزار اجرا می‌کنیم. هر یک از روشها بعد از اجرا دارای نتایجی می‌باشند که ما فقط بعضی از نتایج که برای ارزیابی چارچوب لازم است را ارائه می‌کنیم. چنانچه قبلاً نیز بیان شد در هر یک از روشها قبل از اجرا به انتخاب ویژگیهای با اهمیت می‌پردازیم، نمودار شکل ۳ نمونه‌ای از این انتخاب ویژگیهای با اهمیت را، برای شبکه عصبی نشان می‌دهد. در مقایسه این نمودارها برای هر چهار روش می‌توان موارد زیر را بیان کرد :

هر یک از روشها مجموعه‌ای از انتخاب ویژگیهای منحصر به خود را دارد. در این نمودارها سوال ۹ (tr۹) و سوال ۱۰ (df۱۰)که همان موارد اضافه شده در پروفایل کاربران در این تحقیق می‌باشد دارای جایگاه خوبی می‌باشند. دقت و نتیجه به دست آمده از هر روشی تا حدودی وابسته به ویژگیهای مورد استفاده در روش می‌باشد.

[۳۵] Multilayer Perceptron



شکل ۲. اجرای چهار روش چارچوب در نرم‌افزار کلمنتاین



شکل ۳. متغیرهای با اهمیت در انتخاب ویژگی برای شبکه عصبی

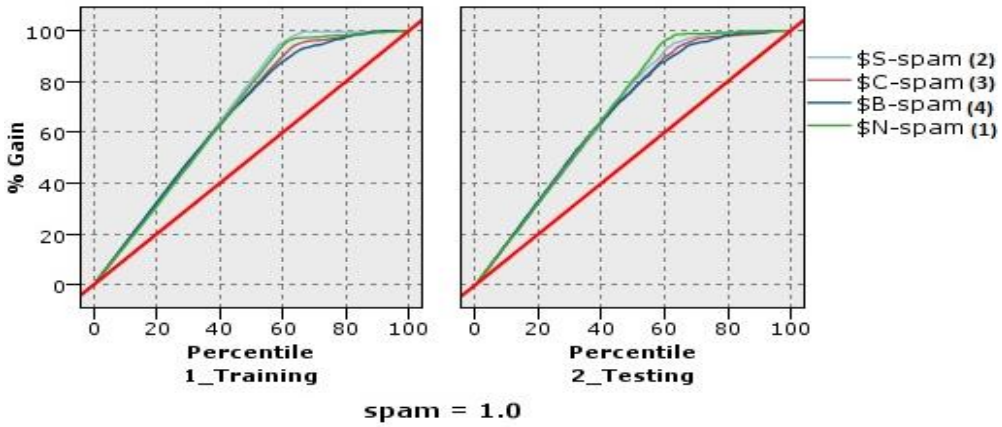
حال نتایج به دست آمده برای هر دو مجموعه داده اول و دوم را ارائه می‌کنیم. در جدول ۵ معیارهای ارزیابی [۱،۸،۱۲،۱۶،۱۹] مورد مقایسه برای هر دو مجموعه بر مبنای چهار روش اجرا شده را می‌توان مشاهده کرد.

در هر یک از مجموعه داده‌ها به ازای هر روش تقریبا نتایج یکسانی حاصل می‌شود. در نتایج هر مجموعه داده شبکه عصبی بهترین نتیجه و پیش‌بینی را نسبت به دیگر روشها دارا می‌باشد. در نتیجه شبکه عصبی به عنوان پایدارترین و بهترین روش می تواند مورد توجه قرار گیرد .

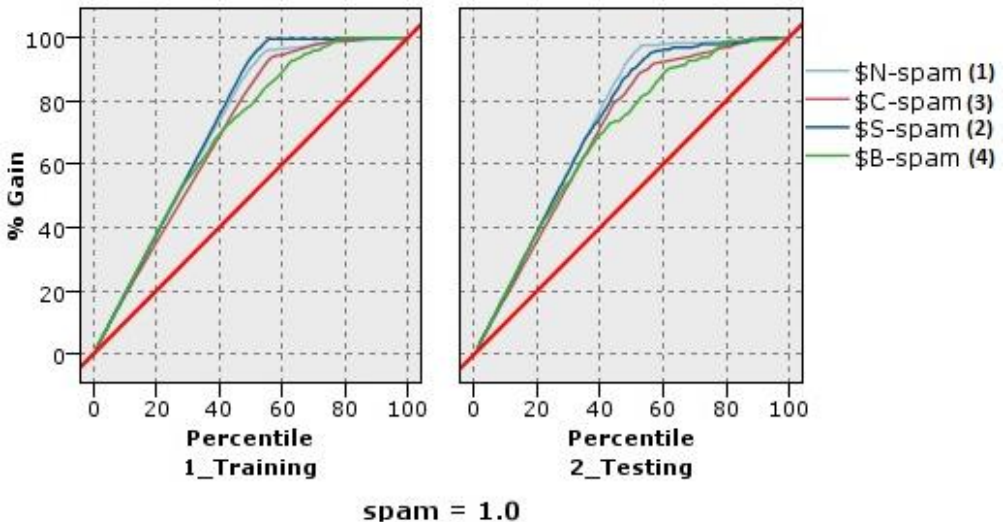
در پایان نمودار بهره (Gain) برای مجموعه داده اول و مجموعه داده دوم به ترتیب در قالب نمودارهای شکل ۴ و ۵ برای هر چهار روش نمایش داده می‌شود. این نمودارها در حقیقت یک روش بصری و آماری برای کمک به درک کارآیی روشهای مطرح شده می‌باشد .در این نمودارها ابتدا مجموعه داده‌ها به صد قسمت تقسیم شده و به صورت درصدی در محور افقی نمایش داده می‌شود، سپس به صورت تجمعی تعداد کل پیش‌بینی‌های درست هرزنامه به ازای کل پیش‌بینی‌ها در قالب نمودار در محور عمودی به صورت درصد نمایش داده می‌شود. این نمودارها گویای آن است که روش شبکه عصبی بهترین بهره را دارا می‌باشد. هر چند می‌توان این نمودار را به صورت نقطه‌ای و غیرتجمعی و با تقسیمات مختلف داده نمایش داد که از آن صرف‌نظر می‌کنیم.

جدول۵ : معیارهای ارزیابی برای هر دو مجموعه داده:(a :هرزنامه‌ای که به عنوان هرزنامه پیش‌بینی شده، d: نامه معتبری که به عنوان نامه معتبر پیش‌بینی شده، b:هرزنامه که به عنوان نامه معتبر پیش‌بینی شده(FN)، c:نامه معتبر که به عنوان هرزنامه پیش‌بینی شده(FP))

روش | معیار / مجموعه داده			مجموعه اول	مجموعه دوم
Accuracy	$\frac{a+d}{a+d+b+c}$	C۵,۰	۸۷,۸۷	۸۷,۱۷
		SVM	۹۱,۳۹	۹۰,۵۱
		BN	۸۵,۵۹	۷۷,۸۶
		NN	۹۵,۴۳	۹۵,۰۸
Error Rate	$1-Accturacy$	C۵,۰	۱۲,۱۳	۱۲,۸۳
		SVM	۸,۶۱	۹,۴۹
		BN	۱۴,۴۱	۲۲,۱۴
		NN	۴,۵۷	۴,۹۲
FP Rate	$\frac{c}{d+c}$	C۵,۰	۲۱,۶	۱۴,۵
		SVM	۱۴,۶	۸,۸
		BN	۲۲,۵	۲۰,۹۹
		NN	۷,۰	۴,۶
Spam Recall	$\frac{a}{b+a}$	C۵,۰	۹۴,۱	۸۸,۸۸
		SVM	۹۵,۳	۸۹,۹۳
		BN	۹۰,۹	۷۶,۷۳
		NN	۹۷,۰	۹۴,۷۹
Spam Precision	$\frac{a}{c+a}$	C۵,۰	۸۶,۸۲	۸۶,۱۹
		SVM	۹۰,۰۸	۹۱,۱۹
		BN	۸۵,۹۵	۷۸,۹۲
		NN	۹۵,۲۵	۹۵,۴۵



شکل ۴. نمودار بهره Gain برای داده اول



شکل ۵. نمودار بهره Gain برای داده دوم

۶. نتیجه‌گیری و پیشنهاد کارهای آتی

به طور کلی در ایران در زمینه دسته‌بندی و پالایش نامه‌های الکترونیکی در امر بازایابی و تبلیغات پژوهش زیادی انجام نشده است، لذا در این طرح سعی بر ایجاد یک ضد هرزنامه شخصی شده برای تخمین اهمیت و دسته‌بندی نامه‌های الکترونیکی تبلیغاتی کاربران با توجه به رفتار. پروفایل آنها شده است. در واقع ما از سه منبع مقالات، مطالعه و بررسی نامه‌های الکترونیکی تبلیغاتی و نظر خبرگان برای انجام این تحقیق استفاده کرده‌ایم.

برای طراحی بهتر لازم بود که حوزه کاری و بعدی از تجارت الکترونیک که طراحی ضد هرزنامه برای آن صورت می‌پذیرد، مشخص شود. طراحی ضد هرزنامه برای نامه‌های الکترونیکی تبلیغاتی که بیشتر در حوزه نامه‌های الکترونیکی خاکستری قرار می‌گیرد، صورت پذیرفته است. در بعد تجارت الکترونیکی، برای بازاریابی از طریق نامه‌های الکترونیکی تبلیغاتی و بیشتر برای بازاریابی B۲C سازگار شده است. بعد از طراحی پرسشنامه، جمع آوری پاسخها، مقایسه روشها بر مبنای معیارهای ارزیابی مطرح شده و دو مجموعه داده مجزا با یکدیگر مشخص شد که شبکه عصبی دارای دقت می‌باشد. این دقت و ثبات در نتایج بدست آمده از شبکه عصبی به خاطر نوع خاص روش شبکه عصبی می‌باشد که بر مبنای افکار انسان عمل می‌کند.

به طور کلی می‌توان در هر چهار مرحله بیان شده طراحی نوآوری‌هایی دیگری در نظر گرفت. در مرحله اول می‌توان دسته‌بندی‌های دیگر و بر مبنای جامعه آماری دیگر شبیه‌سازی کرد. در قسمت چارچوب می‌توان از ترکیب روشهای موجود مانند رای‌گیری برای پیش‌بینی استفاده کرد. هر روش در حین انتخاب ویژگیها به صورت منحصر به فرد عمل می‌کند لذا هریک بسته به ویژگیهای انتخابی دقت محدودی را در همان بازه کسب می‌کند. با ترکیب کردن نتایج روشهای مختلف می‌توان از حداکثر ویژگیهای استخراج شده بهره برد. برای دقت بیشتر می‌توان از روشهای معنایی، آنتولوژی همراه با روش‌های همکارانه بهره برد. برای مثال می‌توان به خوشه‌بندی پروفایل‌های کاربران پرداخت. سپس با استفاده از این خوشه‌بندی در مراحل مختلف اقدام به بهبود دقت چارچوب کرد.

منابع

١. Blanzieri E., Bryl A. ,(٢٠٠٨) A survey of learning-based techniques of email spam filtering, Artif Intell Rev, vol.٢٩,pp.٦٣–٩٢.

٢. Cukier W. L., Cody S., Nesselroth E. J., (٢٠٠٦)Genres of Spam: Expectations and Deceptions, Proceedings of the ٣٩th Hawaii International Conference on System Sciences, .

٣. Sousa p., et al,(٢٠١٠) A Collaborative Approach for Spam Detection ,Second International Conference on Evolving Internet, IEEE.

٤. Raad M.,et al,(٢٠١٠)Impact of spam advertisement through e-mail: A study to assess the influence of the anti-spam on the e-mail marketing, African Journal of Business Management, Vol. ٤(١١), pp. ٢٣٦٢-٢٣٦٧.

٥. Kim J., Dou D., Liu H., Kwak D., (٢٠٠٧)Constructing a User Preference Ontology for Anti-spam Mail Systems, Canadian AI ٢٠٠٧, LNAI ٤٥٠٩, pp. ٢٧٢ – ٢٨٣.

٦. Kakade A.G., Kharat P.K., Gupta A.K,(٢٠١٣),Survey of Spam Filtering Techniques and Tools, and Map Reduce with SVM, IJCSMC, Vol. ٢, Issue. ١١, November ٢٠١٣, pg.٩١ – ٩٨.

٧. Wenxuan S.,Maoqiang X.,(٢٠١٣) A Reputation-based Collaborative Approach for Spam Filtering, ٢٠١٣ AASRI Conference on Parallel and Distributed Computing and Systems, Volume ٥, ٢٠١٣, Pages ٢٢٠–٢٢٧

٨. Almeida, T. A., Yamakami, A.,(٢٠١٢) Facing the spammers: A very effective approach to avoid junk e-mails, Expert Systems with Applications,vol. ٣٩ ,pp.٦٥٥٧–٦٥٦١.

٩. Cook D., Hartnett J., Manderson K., scanlan J., (٢٠٠٦)catching Spam Before it Arrives: Domain Secific Dynamic Blacklists , ACM Inrenational Conference Proceeding Series; Vol.١٦٧,pp.١٩٣-٢٠٢.

١٠. Almeida T.A., Yamakami A.,(٢٠١٠) Content-Based Spam Filtering, The ٢٠١٠ International Joint Conference on Neural Networks (IJCNN), IEEE.

١١. Ying K.C., et al,(٢٠١٠)An ensemble approach applied to classify spam e-mails, Expert Systems with Applications.vol ٣٧,pp. ٢١٩٧–٢٢٠١.

١٢. Saad O., Darwish A., Faraj R.,(٢٠١٢) A survey of machine learning techniques for Spam filtering., International Journal of Computer Science and Network ecurity, VOL.١٢ No.٢, February.

١٣. Yih W., McCann R., Kołcz A.,(٢٠٠٧) Improving Spam Filtering by Detecting Gray Mail, In Proceedings of the ٤rd Conference on Email and Anti-Spam.

١٤. Rossiter J. R., Bellman S.,(٢٠٠٥) "Marketing Communications" Prentice Hall, English.

١٥.Youn S., McLeod D.,(٢٠٠٩) Spam Decisions on Gray E-mail using Personalized Ontologies, Proceedings of the ٢٠٠٩ ACM Symposium on Applied Computing (SAC), Honolulu, Hawaii, USA, pp. ١٢٦٢-١٢٦٦.

١٦. Guzella T.S., Caminhas W.M.,(٢٠٠٩) A review of machine learning approaches to Spam filtering , Expert Systems with Applications,vol. ٣٦,pp.١٠٢٠٦–١٠٢٢.

١٧.Saad O.,Darwish A.,faraj R.,(٢٠١٢)A survey of machine learning techniques for Spam filtering., International Journal of Computer Science and Network security, VOL.١٢ No.٢, February.

١٨. Dwork C., Naor M.,(١٩٩٢) Pricing via processing or combatting junk mail, In Advances in Cryptology - Crypto ٩٢ Proceedings, Springer Verlag, pp ١٣٩–١٤٧.

١٩. SHI L., WANG Q. , MA X. , WENG M. , QIAO H.,(٢٠١٢) Spam Email Classi_cation Using Decision Tree Ensemble, Journal of Computational Information Systems,vol. ٨: ٣,pp. ٩٤٩–٩٥٦.

٢٠. Spam definition.(٢٠١٢)Availabe at http://en.wikipedia.org/wiki/Spam_(electronic) .

٢١. GrayEmail definition,(٢٠١٢) Availabe at http://en.wikipedia.org/wiki/Graymail_(email).

٢٢. Ravi J., Shi W., Xu C., (٢٠٠٥)Personalized Email Management at Network Edges, IEEE Internet Computing, Vol.٩(٢) ,pp.٥٤-٦٠.

٢٣. Nicola L.,(٢٠٠٤) European union vs. spam: A legal response, In Proceedings of the First Conference on Email and Anti-Spam, CEAS'٢٠٠٤.

٢٤.Rafiqul I., Jemal A.,(٢٠١٣) A multi-tier phishing detection and filtering approach, Journal of Network and Computer Applications, Volume 36, Issue 1, January 2013, Pages 324–335.