سال چهاردهم، شماره‌های ۵۱ و ۵۲ ، بهار و تابستان ۱۴۰۱

صص: ۱۲۹_۱۵۴

دو فصلنامه علمی

فناوری اطلاعات و ارتباطات ایران

استفاده از تحلیل احساسات و ترکیب روش‌های یادگیری ماشین برای تشخیص هرزنامه در توییتر

مهدی سالخورده حقیقی* امین الله کرمانی**

*عضو هیئت علمی دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه سجاد، مشهد

**کارشناسی ارشد رایانش امن، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه سجاد، مشهد

تاریخ دریافت: ۱۴۰۰/۰۱/۲۰ تاریخ پذیرش: ۱۴۰۰/۰۹/۰۲

نوع مقاله: پژوهشی

چکیده

محبوبیت شبکه‌های اجتماعی بخصوص توییتر چالش جدیدی را روبروی محققان قرار داده است و آن چیزی نیست جز هرزنامه[۲]. روش‌های گوناگون زیادی برای مقابله با آنها ارائه شده است. بعضی از این روش‌ها اگرچه در ابتدا کارآمد بودند اما به مرور توسط تولیدکنندگان هرزنامه دور زده شدند. در این تحقیق تلاش داریم با استفاده از یکی از جدیدترین روش‌های تشخیص هرزنامه و ترکیب آن با تحلیل احساسات دقت تشخیص هرزنامه را افزایش دهیم. ما با استفاده از روش تعبیه سازی، کلمات متن توییت را به عنوان ورودی به یک معماری شبکه عصبی پیچشی داده و خروجی تشخیص دهنده متن هرزنامه یا متن عادی خواهد بود. هم زمان با استخراج ویژگی‌های مناسب در شبکه توییتر و اعمال روش‌های یادگیری ماشین بر روی آنها تشخیص هرزنامه بودن توییت را بصورت مجزا محاسبه می‌کنیم. در نهایت خروجی هر دو روش را به یک شبکه پیچشی تلفیقی[۳] وارد می‌کنیم تا خروجی آن تشخیص نهایی هرزنامه یا نرمال بودن متن توییت را تعیین کند. ما در این تحقیق از دو مجموعه داده متعادل و نامتعادل استفاده می‌کنیم تا تاثیر مدل پیشنهادی را بر روی دو نوع داده بررسی کنیم. نتایج پژوهش نشان دهنده بهبود کارایی روش پیشنهادی در هر دو مجموعه داده می‌باشد.

واژگان کلیدی: توییتر، هرزنامه، تعبیه لغات، شبکه های عصبی پیچشی، تحلیل احساسات، CNN

۱. مقدمه

با توجه به گسترش روزافزون محبوبیت شبکه‌های اجتماعی، هرزنامه‌ها نیز این بستر را برای گسترش محتوای خود، هدف قرار می‌دهند. توییتر یکی از محبوب‌ترین شبکه‌های اجتماعی است که در آن کاربران با موضوعات مختلف مباحث را مطرح

نویسنده مسئول : مهدی سالخورده حقیقیhaghighi@sadjad.ac.ir

[۲] Spam

[۳] Ensemble

کرده و با هم ارتباط برقرار می‌کنند. اکثر روش‌های فیلتر کردن هرزنامه در توییتر بر شناسایی هرزنامه‌گرها (افرادی که هرزنامه منتشر می‌کنند) و مسدود کردن آنها تمرکز دارند. با این حال، هرزنامه‌گرها می‌توانند یک حساب کاربری جدید ایجاد کرده و دوباره هرزنامه جدید ارسال کنند. در سال ۲۰۱۳، توییتر یکی از ده وبسایت برتر در فهرست محبوب‌ترین وبگاه‌ها اعلام شد و همچنین عنوان پیامک اینترنتی به آن داده شده است[۱] . از سال ۲۰۱۸، توییتر ماهانه بیش از ۳۲۱ میلیون کاربر فعال دارد [۲]. این حجم عظیم کاربر محل جذابی برای تولیدکنندگان هرزنامه می‌باشد تا به شکار قربانیان خود بپردازند. اگرچه تولید و انتشار هرزنامه برای تولیدکنندگان آنها بسیار پر هزینه می‌باشد، اما روش‌های جلوگیری از انتشار آنها برای شرکت‌های میزبان بسیار پر هزینه‌تر است. بر اساس برآورد قوه مقننه آمریکا هزینه هرزنامه در ایالات متحده بالغ بر ۱۳ میلیارد دلار در سال ۲۰۰۷ بوده که شامل پایین آمدن کارایی، اتلاف تجهیزات و نیروی کار لازم بوده است [۳] . تاثیرات مالی مستقیم هرزنامه نیز شامل اضافه بار بر سیستم‌های کامپیوتری و منابع شبکه، اتلاف زمان و منابع انسانی است. به علاوه هرزنامه از چندین بعد دارای هزینه است . این هزینه در مورد شرکتی همچون توییتر با میلیون‌ها کاربر از اهمیت بیشتری برخوردار است.

بنابراین برای شناسایی هرزنامه‌ها در سطح توییت، نیاز به تکنیک‌های تشخیص هرزنامه قوی وجود دارد. این نوع تکنیک ها می‌توانند به صورت بلافاصله از هرزنامه جلوگیری کنند. برای شناسایی هرزنامه در سطح توییت، اغلب ویژگی‌هایی تعریف شده و الگوریتم‌های یادگیری ماشین مناسب بر روی آنها اعمال می‌شود. اما به تازگی، روش‌های یادگیری عمیق[۱] نتایج موثری در کاربرد پردازش زبان طبیعی نشان داده‌اند. ما می‌خواهیم از مزایای بالقوه این روش برای رفع این مشکل استفاده کنیم.

به همین خاطر، در این مقاله یک رویکرد ترکیبی برای تشخیص هرزنامه در سطح توییت ارائه می‌شود و مدل‌های یادگیری عمیق[۱] مختلف را نیز توسعه خواهیم داد. در این نوشتار از مدل مبتنی بر ویژگی به همراه یک مدل مبتنی بر تحلیل احساسات همراه با الگوریتم‌های شبکه‌های عصبی به صورت ترکیبی استفاده شده است. الگوریتم‌های شبکه‌های عصبی پیچشی[۲] با استفاده از روش‌های مختلف تعبیه کلمه[۳] (Word۲vec ،Glove) برای آموزش مدل استفاده می‌شود. مدل مبتنی بر ویژگی، از ویژگی‌های مبتنی بر محتوا، مبتنی بر کاربر و N-gram استفاده می‌کند. همچنین ویژگی‌های احساسی درون توییت‌ها نیز در مدل مبتنی بر تحلیل احساسات بکار گرفته می‌شوند. رویکرد ما ترکیبی از هر دو یادگیری ماشین و مدل‌های مبتنی بر ویژگی‌های سنتی و احساسی با استفاده از یک شبکه عصبی چند لایه است که به عنوان یک ابرطبقه‌بند[۴] عمل می‌کند. شبکه‌های اجتماعی برخط یا به اختصار OSN ابزاری همگانی است که باعث ارتباط میلیون‌ها کاربر اینترنت می‌شود. در این میان بستر توییتر با ارائه سرویس رایگان میکروبلاگین به مشتریان جهت انتشار پیام‌ها، کاربران زیادی را به خود جلب کرده است. هر روزه میلیون‌ها نفر اخبار و احساسات خود را در توییتر به اشتراک می‌گذارند. در این بین حساب‌های کاربری زیادی هستند که جهت منافع شخصی از این بستر سوء استفاده کرده و با ارسال هرزنامه به کاربران دیگر حمله می‌کنند. این حملات می‌تواند با اهداف زیادی همانند بازاریابی و یا شکل خطرناک‌تر آن مانند نفوذ بدافزار باشد. آنچه مسلم است ، هرزنامه موضوعی است که این روزها با زندگی دیجیتال ما گره خورده است. از لحاظ تاریخی، اولین بار هرزنامه از طریق ایمیل وارد حوزه اینترنت شد و با گسترش کاربرد شبکه های اجتماعی به سرعت جای پای خود را در این بخش نیز وارد کرده و امروزه به یک معضل اساسی در حوزه اینترنت بدل شده است. کم هزینه بودن ارسال هرزنامه از یک طرف و نبود قوانین بین المللی مشخص برای محدود کردن آنها از طرف دیگر باعث شده هرزنامه‌ها هر روز به طور وسیعی انتشار یابند. طبق تحقیقات انجام شده در [۴] نرخ کلیک هرزنامه‌های شبکه توییتر به حدود ۰/۱۳٪ رسیده است و این در حالی است که هرزنامه‌های ایمیل تنها به حدود ۰/۰۰۰۳٪ می‌رسد. این شرایط

---

۱ Deep Learning

۲ Convolutional Neural Network(CNN)

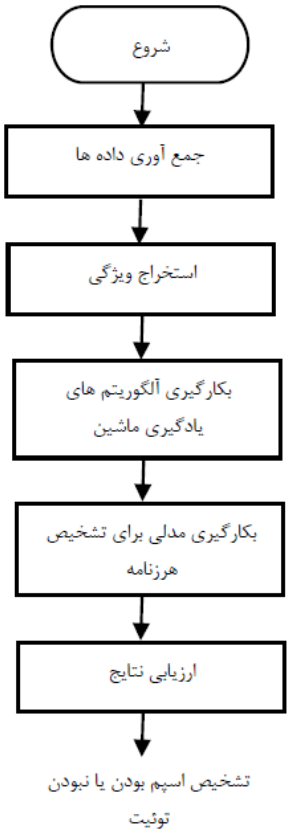
۳ Word Embedding

۴ Meta Classifier

محققان را بر آن داشته تا با ارائه مدل‌هایی جهت تحلیل، شناسایی و مسدودسازی هرزنامه‌ها اقدام کنند. در سال‌های اخیر تحقیقات زیادی بر روی روش‌های تشخیص هرزنامه در بستر توییتر انجام شده است. محققین سعی کردند روش‌هایی را پیدا کنند تا الگوریتم‌های یادگیری ماشین بتوانند خود را با مفاهیم تشخیص هرزنامه وفق دهند. ماشین‌های یادگیر به هر مدل کاربردی امکان یادگیری و پیش‌بینی را می‌دهند. تشخیص هرزنامه یک چهارچوب طبقه‌بندی دودویی می‌باشد. چهارچوبی که تشخیص می‌دهد یک حساب کاربری و یا یک توییت هرزنامه است یا خیر. به دلیل همین ماهیت، محققین توجه خود را معطوف به تکنیک‌های یادگیری ماشین کردند. مدل تشخیص هرزنامه در یادگیری ماشین از دو مرحله تشکیل شده است.

مرحله آموزش : این اولین مرحله‌ای است که مدل یادگیری با نمونه‌های برچسب‌گذاری آموزش می‌بینند.

مرحله آزمون: در این فاز نمونه‌های فاقد برچسب آزمایش شده و بوسیله طبقه‌بندی نمونه‌ها به هرزنامه و یا نرمال تقسیم می‌شوند. مرجع[۵] چهارچوبی کلی برای تشخیص هرزنامه در توییت را ارائه داده است. این چهارچوب که در تمامی تحقیقات تقریبا یکسان است در شکل ۱ نشان داده شده است.



شکل ۱. چهارچوب کلی تشخیص هرزنامه [۵]

اولین مرحله جمع‌آوری داده می‌باشد. این داده‌ها می‌توانند توسط API[۱] های مختص به توییتر جمع‌آوری شوند و یا از مجموعه‌داده‌های عمومی در دسترس، استفاده شود. مرحله بعدی استخراج ویژگی‌ها از مجموعه‌داده می‌باشد. از بین ویژگی های استخراج شده تعدادی از آنها انتخاب می‌شوند. در فاز آموزش تعداد کمتری از کل نمونه‌ها جهت آموزش برچسب-گذاری می‌شوند زیرا برخی توییت‌ها با دلایلی از قبیل استفاده از زبان‌های غیر انگلیسی و نامعتبر بودن برخی ویژگی‌ها قابل استفاده نیستند. این برچسب‌گذاری می‌تواند به صورت دستی یا با استفاده از سرویس‌های فیلترینگ شناسایی هرزنامه انجام شود. پس از آن مدل‌های تشخیص مبتنی بر یادگیری ماشین بوسیله نمونه‌های برچسب‌دار آموزش داده می‌شوند و سپس جهت طبقه بندی داده‌های جدید آزمایش می‌شوند. در نهایت

[۱] Application Programming Interface

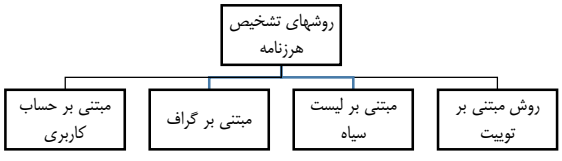
این مدل‌ها توسط پارامترهای دقت، صحت، فراخوانی و غیره ارزیابی می‌شوند.

به دلیل وجود روش‌های ترکیبی و تلفیق روشها با یکدیگر، امکان جداسازی آنها بطور کامل میسر نمی‌باشد. با این وجود، روش‌های تشخیص هرزنامه به چند گروه تقسیم می‌شوند که در بخش بعدی تشریح می‌گردند.

ساختار ادامه مقاله به شرح زیر می‌باشد. در بخش ۲ مروری بر پیشینه تحقیق آورده شده است که به بررسی جنبه‌های مختلف در این حوزه می‌پردازد. در بخش ۳ راهکار پیشنهادی برای تشخیص هرزنامه با جزئیات ارائه گردیده است. در بخش ۴ آزمایشها با استفاده از مجموعه داده‌های انتخابی انجام شده و مقایسه با برخی روش‌ها صورت گرفته و نتایج تحلیل شده است. در بخش ۵ نتیجه‌گیری انجام شده و در نهایت پیشنهادات برای کارهای آینده ارائه گردیده است.

۲. مروری بر پیشینه تحقیق

همانگونه که در انتهای بخش قبل اشاره شد، مرزبندی دقیقی نمی‌توان برای روش‌های تحقیق بکار برد با این وجود بر اساس استخراج ویژگی و استفاده از روش‌های طبقه‌بندی، تکنیک‌های تشخیص هرزنامه در توییتر را می‌توان در چهار گروه کلی بر مبنای شکل ۲ طبقه‌بندی کرد. این روش‌ها عبارتند از تکنیک-های لیست سیاه ، تحلیل ویژگی مبتنی بر گراف، مبتنی بر حساب کاربری و تحلیل مبتنی بر توییت، که این ویژگی آخر خود به دو ویژگی مجزای مبتنی بر خصوصیات توییت و مبتنی بر متن توییت می‌باشد.



شکل ۲. روش‌های تشخیص هرزنامه

۲_۱ روش مبتنی بر لیست سیاه

بنابر تحقیقات انجام شده در [۴] حدود ۹۰٪ کلیک‌ها بر روی آدرس‌های URL هرزنامه در همان دو روز اول انجام می‌گیرد در حالیکه به طور متوسط حدود ۴ روز طول می‌کشد تا URL جدید در لیست سیاه قرار گیرد که تاخیر زیادی در بروز رسانی لیست سیاه می‌باشد و در این زمان هرزنامه به سرعت گسترش می‌یابد و این از نقاط ضعف بزرگ این روش می‌باشد. تحقیقات زیادی در این حوزه انجام شده است. به عنوان مثال مولفین مقاله در [۶] از درخت تصمیم و ویژگی‌های آماری جهت تشخیص URL های مخرب استفاده کرده‌اند. برخی ویژگی‌های آنها شامل طول آدرس URL، وجود IP آدرس در Hostname می‌باشد. در مقاله [۷] از سه طریق این ویژگیها را استخراج کردند. ۱- Web Browser و URL نهایی. ۲- DNS و ۳- تحلیل آدرس IP از نظر موقعیت جغرافیایی.

۲_۲ روش های مبتنی بر گراف

این روش، ویژگی‌ها را بر اساس گراف‌های اجتماعی کاربران توییتر بر مبنای روابط بین دنبال‌کنندگان و دنبال‌شوندگان استخراج می‌کند. در این حوزه تحقیقات زیادی در شبکه‌های اجتماعی انجام گرفته است. در روش‌های مبتنی بر گراف که تا حدودی به روش‌های مبتنی بر حساب کاربری شباهت دارند، هر حساب کاربری به عنوان یک گره در نظر گرفته می‌شود و درجه ورودی گره نشانگر تعداد دنبال‌کنندگان و درجه خروجی نمایانگر تعداد دنبال‌شوندگان می‌باشد. همچنین ویژگی‌های مبتنی بر همسایگی نیز در این حوزه قرار می‌گیرند. این ویژگی-ها در طبقه‌بندهای یادگیری ماشین استفاده می‌شوند. به عنوان مثال در مقاله [۸] از سه ویژگی مبتنی بر گراف جهت تشخیص هرزنامه استفاده کردند (چگالی گراف و میانگین کوتاهترین مسیر). آنها همچنین از سه ویژگی قوی نیز استفاده کردند: ویژگی Local Clustering Coefficient ، Betweenness Centrality و Links Bidirectional Ratio. این ویژگی‌ها برای ایجاد گراف اجتماعی کوچک حساب هدف بکار می‌رود. عیب این روش در این است که به عنوان

مثال حساب‌های کاربری افراد مشهور با تعداد زیاد دنبال کننده نیز می‌تواند به عنوان هرزنامه تلقی شود. از طرفی تولید کنندگان هرزنامه نیز می‌توانند خود را با ویژگی‌های جدید مبتنی بر گراف تطبیق داده و باعث گمراهی سیستم تشخیص شوند. ضعف این روش در این است که عملا جمع آوری روابط گراف میلیون‌ها کاربر توییتر غیر ممکن است. همانند روش‌های مبتنی بر حساب کاربر، در این روش هم با بسته شدن حساب تولیدکننده هرزنامه، او مجدد اقدام به ایجاد حساب جدید می-کند.

۲_۳ تشخیص مبتنی بر حساب کاربری

این روش در دیگر شبکه‌های اجتماعی نیز متداول است و به طور موثری حساب‌های کاربری هرزنامه از غیرهرزنامه را تشخیص می‌دهد. تمرکز این روش بر اطلاعات حساب کاربری متمرکز است. به عنوان مثال تعداد دنبال‌کنندگان[۱] و دنبال شوندگان[۲] در حساب‌های عادی بسیار بیشتر از حساب‌های هرزنامه می‌باشد. به عنوان مثالی دیگر، طول عمر[۳] یک حساب هرزنامه به مراتب کمتر از یک حساب عادی می‌باشد. ویژگی مهم دیگر Reputation است که در حساب‌های هرزنامه و غیرهرزنامه متفاوت است. ویژگی Reputation در یک حساب تولیدکننده هرزنامه ۱۰۰٪ یا بسیار کم است، در حالی‌که این مقدار در یک حساب عادی چیزی در حدود ۳۰٪ تا ۹۰٪ می باشد. این فاکتور در تشخیص حساب هرزنامه از غیر هرزنامه بسیار کارآمد است. اگرچه این روش دارای قدرت تشخیص بالایی می‌باشد اما حساب‌های تولیدکننده هرزنامه‌ای نیز وجود دارند که در موارد استثناء دارای تعداد دنبال شوندگان زیادی هستند و به این ترتیب الگوریتم در این موارد دچار اشتباه می شود. معمولا این روش‌ها همراه با دیگر روش‌ها مورد استفاده قرار می‌گیرند. در مقاله [۹] با بررسی ۶ الگوریتم یادگیری ماشین بهترین F-measure را با Random Forest بدست آورده‌اند. آنها از ویژگی‌هایی نظیر تعداد دنبال‌کننده و تعداد دنبال‌شونده و ویژگی طول عمر حساب استفاده کرده‌اند. یکی از نقاط ضعف این روش این است که با بسته شدن حساب کاربری تولیدکننده هرزنامه، او مجددا حساب جدیدی ایجاد می‌کند. همچنین تولیدکنندگان هرزنامه به مرور با دور زدن این ویژگی-ها می‌توانند روش‌های تشخیص را فریب دهند.

۲_۴ روش‌های تشخیص مبتنی بر توییت

تمامی روش‌های مبتنی بر حساب و گراف یک مشکل عمده دارند. پس از مسدود شدن حساب کاربری توسط الگوریتم، تولیدکننده هرزنامه حساب جدیدی ایجاد کرده و به فعالیت خود ادامه می‌دهد. به همین منظور تحقیقات اخیر تمرکز خود را بر روی محتوای خود متن توییت معطوف کرده‌اند. در این روش بدون در نظر گرفتن فرستنده هرزنامه، پس از شناسایی توییت هرزنامه، از انتشار آن جلوگیری می‌شود. با توجه به اینکه هرزنامه‌ها از کلمات و موضوعات مخرب مشابهی استفاده می-کنند، توییت‌های شامل این کلمات و موضوعات می‌توانند هرزنامه باشند. تکنیک‌های تشخیص در این روش مبتنی بر پردازش زبان‌های طبیعی[۴] است.

در تحقیق انجام شده در [۱۰] ، بیشتر از ویژگی‌های ذاتی توییت برای تشخیص هرزنامه در توییتر در سطح توییت استفاده می‌کنند. آنها از روش‌های ترکیبی مبتنی بر کاربر، مبتنی بر محتوای توییت، و N-gram جهت شناسایی توییت‌های هرزنامه استفاده کردند. نویسندگان از دو مجموعه‌داده برای مطالعه استفاده می‌کنند: مجموعه‌داده ۱KS۱۰KN و Social Honeypot. در این تحقیق از روش‌های ترکیبی تشخیص هرزنامه استفاده شده است. روش مبتنی بر کاربر شامل مواردی است همچون تعداد دنبال‌شوندگان و دنبال-کنندگان، طول نام پروفایل کاربر، طول توضیحات پروفایل، عمر اکانت کاربر بر حسب ساعت و غیره. ویژگی‌های مبتنی بر محتوای توییت شامل تعداد کلمات، تعداد کاراکترها، تعداد فاصله ها، تعداد علامت‌های سوال و تعجب و غیره می‌باشد.

---

۱ Follower

۲ Following

۳ Age

۴ Natural language processing

ویژگی‌های مبتنی بر N-gram نیز به سه دسته Uni-gram و Bi-gram و Tri-gram تقسیم‌بندی شده است. ۵ طبقه‌بند بر روی این ویژگی‌ها اعمال می‌شود که عبارتند از الگوریتم‌های Decision Tree, KNN SVM, Naïve Bayes, و Random Forest. بر طبق این تحقیق نتایج بر روی هر دو مجموعه داده با الگوریتم‌های Random Forest و SVM بهترین خروجی را می‌دهد.

ارزیابی عملکرد روش‌های مبتنی بر یادگیری ماشین برای شناسایی هرزنامه در سطح توییت در [۱۱] شرح داده شده است. بررسی روش‌های تشخیص هرزنامه‌های توییتر با تحلیل مقایسه‌ای در [۱۲] شرح داده شده است. داده‌های هرزنامه هشتگ محور توییتر توسط [۱۳] ایجاد شده است. نویسندگان ۱۴ میلیون توییت جمع آوری کرده‌اند و داده‌ها را به عنوان HSpam۱۴ نام‌گذاری کرده‌اند. مولفین در [۱۴] در تحقیق خود یک چارچوب تشخیص هرزنامه را به دست آوردند. آنها از چهار شناسه سبک وزن برای شناسایی هرزنامه در سطح توییت استفاده کرده‌اند.

یک روش مبتنی بر یادگیری عمیق برای شناسایی هرزنامه در [۱۵] ارائه شده است. در این تحقیق از دو روش مبتنی بر شبکه-های عصبی پیچشی به طور همزمان استفاده شده است. یک شبکه عصبی پیچشی وظیفه طبقه‌بندی متن توییت را بر عهده دارد و یک طبقه‌بند از ویژگی‌های فرا داده[۱] استفاده می‌کند. در تحقیق [۱۶] بردار توییت را با ترکیب بردار سند توییت (که با مدل‌سازی بردار پاراگراف بدست می‌آید) ساخته‌اند. این بردارهای ترکیبی به عنوان ویژگی‌های ورودی برای الگوریتم-های یادگیری ماشین عمل می‌کنند (جنگل های تصادفی و شبکه‌های عصبی).

در تحقیق دیگری در مقاله [۱۷] از این روش در حل مشکلات زبان طبیعی پردازش یا NLP استفاده کردند. معماری شبکه-های عصبی پیشنهادی آنها می‌تواند در بسیاری از حوزه‌های NLP قابل استفاده باشد. این روش همانند روش‌های بینایی ماشین پس از تبدیل کلمات جمله به ماتریس اعداد آنها را همانند تصاویر وارد یک معماری شبکه عصبی پیچشی می‌کند. در مقاله [۱۸] همین معماری برای جملات بکار برده شده است. در تحقیق [۱۹] از این معماری برای تشخیص هرزنامه استفاده کردند. در تحقیق آنها از روش تعبیه کلمه جهت تبدیل کلمات جمله به بردار عددی استفاده شده است. آنها علاوه بر استفاده از ویژگی برداری کلمات هر توییت، به طور همزمان از ویژگی‌های مبتنی بر متن و مبتنی بر حساب کاربر و همچنین n-gram و تحقیقات انجام شده در [۱۰] نیز استفاده کرده و این ویژگی‌ها را به طور موازی با معماری شبکه عصبی پیشنهادی به یک طبقه بند مانند SVM می‌دهند. خروجی معماری شبکه عصبی اعمال شده بر روی متن به همراه خروجی طبقه‌بند به صورت تلفیقی وارد یک شبکه عصبی ابرطبقه‌بند می‌شود و در نهایت خروجی این شبکه عصبی تصمیم نهایی را درباره هرزنامه بودن یا نبودن متن توییت بر عهده دارد. در تحقیق انجام شده در [۱۹] همچنین از دو ویژگی n-gram نیز استفاده شده است. با ویژگی Uni-grams و Bi-.grams آنها نتایج مدل پیشنهادی تحقیق خود را با تحقیق [۱۰] مقایسه کرده‌اند. با وجود زمان اجرای بالاتر، روش پیشنهادی به طور قابل ملاحظه‌ای نتایج تشخیص هرزنامه را بهبود داده است.

۲_۵ تحلیل احساسات

یکی از زمینه های جدید در تحقیقات مبتنی بر متن شبکه‌های اجتماعی استفاده از تحلیل احساسات می‌باشد. تجزیه و تحلیل احساسات، که همچنین به عنوان افکار اندیشی یا عقیده کاوی نیز خوانده می‌شود، یکی از مهم‌ترین زیرمجموعه‌های پردازش زبان طبیعی می‌باشد که به طور گسترده‌ای در زمینه داده‌کاوی، وب‌کاوی و متن‌کاوی مورد استفاده قرار می‌گیرد. سیستم‌های تحلیل احساسی تقریباً در هر کسب وکار و حوزه اجتماعی به کار گرفته می‌شوند، زیرا عقاید در تمام فعالیت‌های انسانی نقش اساسی دارد و از تأثیرگذارترین رفتارهای ما می‌باشد. اعتقادات و برداشت ما از واقعیت و انتخاب‌هایی که انجام می‌دهیم ، تا

[۱] Meta Data

حد زیادی مشروط به این است که دیگران چگونه دنیا را می‌بینند و ارزیابی می‌کنند. به همین دلیل، زمانی که ما نیاز به تصمیم‌گیری داریم، اغلب به دنبال عقاید دیگران هستیم. این نه تنها برای افراد بلکه برای سازمان‌ها نیز صادق است. در تحقیق [۲۰] مولفین نشان دادند که سه بعد برجسته معنی شامل شامل ارزشیابی (خوب و بد) ، توانایی (قوی و ضعیف) و فعالیت (فعال ، انفعالی) هستند. ارزشیابی بسیار مشابه خوشایندی /ناخوشی (مثبت و منفی) میباشد.

در تحقیقات دیگری در [۲۱] یک مدل مدور اثر توصیفی با دو بعد را توسعه داد، خوشایندی/ناخوشی و برانگیختگی (میزان واکنش پذیری به محرک). ابعاد قطب ها به درجه مثبت یامنفی احساس ارجاع دارد بطوریکه بعد برانگیختگی به درجه آرامش یا هیجان مرتبط است. محدوده هر دو بعد از ۱ (کاملا منفی یا آرام) تا ۹ (کاملا مثبت یا هیجانی) قرار میگیرد. در نتیجه اکثر تحقیقات در زمینه تحلیل احساسات به عامل خوشایندی/ناخوشی اختصاص یافتند [۲۲].

در مقاله [۲۳] معتقد بودند که تمامی احساسات در بعد خوشایندی/ناخوشی قرار می‌گیرند و هرگز بعد خنثی ندارند. بعضی از احساسات کاملا منطبق بر یک قطب خوشایندی/ناخوشی هستند مانند شادی که کاملا قطب مثبت خوشایندی/ناخوشی است. اما بعضی لغات می‌توانند مفهوم هر دو قطب را در برداشته باشند مانند کلمه تعجب. درنتیجه تحقیقاتی با تمرکز بر روی تشخیص خودکار برانگیختگی و احساسات انجام شده است، احساساتی شامل عصبانیت، غم و یا مثبت اندیشی. در نتیجه روشهای مبتنی بر طبقه بندی احساسات بر روش مبتنی بر روش خوشایندی/ناخوشی ارجحیت دارند[۲۴].

تحقیقات بسیاری به رابطه میان احساسات مختلف پرداخته‌اند که برای تحقیقات در مورد احساسات به صورت خاص، مطالعه آنها توصیه می‌گردد[۲۵] ,[۲۶]. امروزه تحقیقات زیادی بر روی تحلیل احساسات متن انجام گرفته است. از این رو کتابخانه‌های آماده فراوانی جهت تحلیل احساسات ایجاد شده است، با این وجود در زمینه تحلیل احساسات در حوزه تشخصیص هرزنامه در توییتر تحقیقات کمتری صورت گرفته است. در این تحقیق حاضر از کتابخانه SentiWordNet [۲۷] استفاده شده است.

۳. راه کار پیشنهادی

با توجه به اینکه شبکه‌های اجتماعی از محبوبیت زیادی برخوردار هستند، به دنیای جذاب هرزنامه‌گرها تبدیل شده‌اند. توییتر یکی از محبوب‌ترین شبکه‌های اجتماعی است که در آن کاربران موضوعات مختلفی را به بحث می‌گذارند و با هم ارتباط برقرار می‌کنند. اکثر روش‌های فیلتر کردن هرزنامه در توییتر بر شناسایی هرزنامه‌گرها و مسدود کردن آنها تمرکز دارند. با این حال، هرزنامه‌گرها می‌توانند یک حساب کاربری جدید ایجاد کرده و دوباره توییت‌های هرزنامه ارسال کنند.

بنابراین برای شناسایی هرزنامه در سطح توییت، نیاز به تکنیک‌های تشخیص هرزنامه قوی وجود دارد. این نوع تکنیک-ها می‌توانند با دقت بیشتری از هرزنامه جلوگیری کنند. برای شناسایی هرزنامه در سطح توییت، اغلب ویژگی‌های تعریف شده‌ای وجود دارد و الگوریتم‌های یادگیری ماشین مناسب بکار برده می‌شوند. به تازگی، روش‌های یادگیری ماشین در حال نشان دادن نتایج موثر در چند کاربرد پردازش زبان طبیعی است.

با توجه به دلایل اشاره شده در بخش‌های قبل، تنها استفاده از ویژگی‌های حساب کاربری برای تشخیص می‌تواند ضعف‌هایی داشته باشد. لذا در روش پیشنهادی ویژگی‌های احساسی نیز مورد استفاده قرار گرفته و از یک روش تلفیق به منظور هم افزایی نتایج حاصل از چند روش و دستیابی به نتایج دقیق‌تر استفاده شده است. هدف از ارائه این روش، استفاده از مزایای بالقوه این دو روش برای بهبود کارایی می‌باشد.

به همین منظور، ما یک رویکرد ترکیبی برای تشخیص هرزنامه در سطح توییت ارائه می‌دهیم. در این تحقیق از مدل مبتنی بر ویژگی به همراه یک مدل مبتنی بر تحلیل احساسات در ترکیب با الگوریتم‌های یادگیری ماشین استفاده می‌کنیم. الگوریتم‌های

یادگیری ماشین با استفاده از روش‌های مختلف تعبیه لغات (Word۲vec ،Glove) برای آموزش مدل استفاده می‌شود. مدل مبتنی بر ویژگی، از ویژگی‌های مبتنی بر محتوا، مبتنی بر کاربر و N-gram استفاده می‌کند. همچنین ویژگی‌های احساسی درون توییت‌ها نیز در مدل مبتنی بر تحلیل احساسات بکار گرفته می‌شوند. رویکرد ما ترکیبی از هر دو روش یادگیری ماشین و مدل‌های مبتنی بر ویژگی‌های احساسی با استفاده از یک شبکه عصبی چند لایه است که به عنوان یک ابرطبقه بند عمل می‌کند.

۳_۱ مدل ترکیبی مبتنی بر شبکه عصبی برای تشخیص هرزنامه در توییتر

هدف اصلی از انجام این تحقیق این است که با دادن توییت t سیستم تشخیص دهد که این توییت هرزنامه است یا خیر. در این بخش ، ابتدا معماری شبکه عصبی پیچشی پیشنهادی در مرجع [۱۹] ، با استفاده از روش‌های تعبیه کلمه مختلف و ابعاد متفاوت، برای شناسایی هرزنامه در سطح توییت ارائه می‌شود. در مرحله بعد، مدل مبتنی بر ویژگی را مورد بحث قرار می‌دهیم که از ویژگی‌های مبتنی بر کاربر، مبتنی بر محتوا و ویژگی‌های N-gram به همراه روش پیشنهادی مبتنی بر احساسات استفاده می‌کند.

۳_۲ مدل مبتنی بر شبکه عصبی پیچشی

مدل مبتنی بر شبکه عصبی پیچشی ما از دو بخش تشکیل شده است. یکی انتخاب بازنمایی ویژگی، و دیگری انتخاب معماری شبکه است. در اینجا، ما در مورد این دو جنبه با جزئیات بیشتر توضیح خواهیم داد.

۳_۳ بازنمایی ویژگی

در روش‌های مبتنی بر توییت، ویژگی‌های اصلی یک توییت از کلمات موجود در آن آمده است. کلمات و جملات و اصولا متن، داده‌های غیرساختار یافته می‌باشند که الگوریتم‌های طبقه‌بندی نمی‌توانند آن‌ها را بدون پیش‌پردازش درک کنند. ما باید قادر باشیم متون خود را به داده‌ای قابل محاسبه و سنجش برای مدل درآوریم. در طول سال‌ها روش‌های مختلفی برای این تبدیل ارائه شده است. در واقع هر کلمه در متن به عنوان یک ویژگی عمل می‌کند. تکنیک تعبیه کلمه به عنوان جدیدترین و بهینه‌ترین روش قادر است متون را به بردارهای عددی مبتنی بر روابط معنایی تبدیل کند. مفهوم اصلی تعبیه کلمات این است که تمامی لغات استفاده شده در یک زبان را می‌توان توسط مجموعه‌ای از اعداد اعشاری و در قالب یک بردار بیان کرد. تعبیه کلمات بردارهای n بعدی هستند که تلاش می‌کنند معنای لغات و محتوای آنها را با مقادیر عددی ثبت کنند. هر مجموعه-ای از اعداد یک بردار کلمه به‌حساب می‌آید که به تنهایی برای ما سودمند نیست. آن بخشی از بردار کلمات برای کاربردهای مورد نظر ما مفید هستند که معنای لغات و ارتباط بین آنها را همانطور که بصورت طبیعی مورد استفاده قرار گرفته‌اند، بدست آورده باشند.

به همین خاطر ما از تکنیک‌های تعبیه کلمه Word۲vec و GLOVe جهت تبدیل کلمات متن توییت به بردار عددی استفاده می‌کنیم. این بردارها به عنوان ورودی شبکه عصبی استفاده می‌شوند. مدل Word۲vec توسط ۳ میلیون کلمه Google News با ابعاد ۳۰۰ پیش‌پردازش شده و مدل GLOVe با ۲ میلیارد توییت پیش پردازش شده با ابعاد ۲۵، ۵۰، ۱۰۰و ۲۰۰ تشکیل شده است.

۳_۴ معماری شبکه

شبکه‌های عصبی پیچشی نشان داده‌اند که در بینایی ماشین سودمند هستند. اخیرا از آنها در مسائل مربوط به پردازش زبان طبیعی نیز استفاده می‌شود[۲۸]، [۲۹]. در مرجع [۱۷] یک معماری شبکه عصبی پیشنهاد کردند که می‌توان آن را برای بسیاری از وظایف پردازش زبان طبیعی مانند شناسایی موجودیت‌های نام‌دار، تجزیه، برچسب‌گذاری اجزاء کلام و تکه تکه کردن استفاده کرد. مدل ما از [۱۹] الهام گرفته شده است و در شکل ۳ نشان داده شده است. لایه‌هایی که در معماری

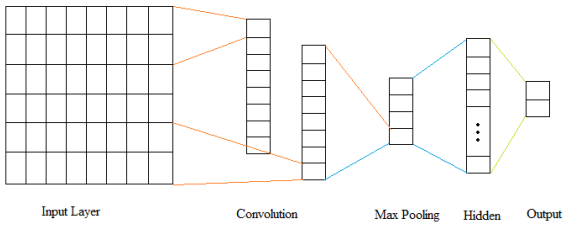
یادگیری ماشین ما حضور دارند عبارتند از: لایه ورودی، لایه پیچشی، لایه جمع کننده[۱]، لایه مخفی و یک لایه خروجی.

هر توییت شامل گروهی از کلمات است، در نتیجه بردار توییت از تلفیق بردارهای کلمات فردی توییت تشکیل شده است. اگر ابعاد بردار کلمه d باشد و طول توییت l باشد آنگاه ابعاد ماتریس توییت $l \times d$ می‌شود. این ماتریس توییت به عنوان لایه ورودی شبکه عصبی است که در شکل ۳ نشان داده شده است. با توجه به [۱۸] اگر هر توییت را به عنوان مجموعه‌ای از کلمات در نظر بگیریم به صورت:

$< term_1 . term_2 . \dots . term_l >$ آنگاه نمایش برداری هر توییت با اندیس v به صورت رابطه (۱) خواهد بود:

$$t_v = x_1 o x_2 o \dots o x_l \qquad (1)$$

که در آن x_i تعبیه کلمه $term_i$ خواهد بود. و o عملگر اتصال می‌باشد.



شکل ۳. معماری شبکه عصبی پیچشی[۱۹]

همانطور که در شکل ۳ نشان داده شده، هر کلمه از توییت به صورت بردار عددی d بعدی تعبیه کلمات نشان داده می‌شود. ماتریس ورودی وارد لایه پیچشی می‌شود. در این لایه عملیات پیچشی شامل یک فیلتر $w \in hk$ به پنجره h کلمه‌ای جملات اعمال می‌شود تا ویژگی جدیدی ساخته شود. به عنوان مثال با توجه به روش [۱۸] ویژگی c_i با یک پنجره کلمات $x_{i:i+h}$ که کلمات آن در اندیس‌های i تا $i+h-1$ قرار دارند به صورت رابطه (۲) تولید می‌شود.

$$c_i = f(w x_{i:i+h-1} + b) \qquad (2)$$

در رابطه (۲) $b \in \mathbb{R}$ بایاس و f تابع غیرخطی مانند Relu می‌باشد. اگر مقدار فیلتر ۱ درنظر گرفته شود، تنها کلمه هدف در جمله در نظر گرفته می‌شود. اگر مقدار فیلتر ۳ باشد کلمه هدف و یک کلمه قبل و بعد از آن در نظر گرفته می‌شود. در صورتیکه این مقدار ۵ باشد کلمه هدف به همراه دو کلمه قبل و دو کلمه بعد از آن به عنوان ویژگی نگاشت می‌شوند. تفاوت حرکت فیلتر در این روش بر خلاف شبکه عصبی بینایی ماشین در این است که حرکت فیلتر بجای پیمایش در فضای دوبعدی به صورت یک بعدی (در واقع هر سطر) ماتریس را پیمایش می‌کند. ضرب فیلتر در ماتریس با طول h باعث ایجاد ماتریس ستونی می‌شود که این مقادیر با b جمع شده و تابع غیرخطی f بر روی آن اعمال می‌شود. در لایه Max Pooling بیشترین مقادیر بدست آمده از مرحله قبلی انتخاب می‌شوند. این کار با هدف کاهش قدرت محاسباتی مورد نیاز برای پردازش داده‌ها از طریق کاهش ابعاد، انجام می‌شود. در واقع در Max Pooling بیشترین فعال سازها انتخاب می‌شوند. اگر فرض کنیم توییت ما شامل ۷ کلمه با ابعاد ۵ باشد، فیلتر با عرض به اندازه ابعاد کلمه و ارتفاع ۲ بر روی تمامی ردیف‌های ماتریس ورودی پیمایش کرده و نتیجه ضرب در ماتریس ستونی بعدی محاسبه می‌شود. پس از جمع بایاس و اعمال تابع فعال ساز نتیجه به لایه Max Pooling وارد شده و بزرگترین عدد از این فیلتر انتخاب می‌شود. این روند تا زمان اعمال تمامی فیلترها با ابعاد گوناگون بر روی ماتریس ورودی ادامه پیدا می‌کند.

لایه‌های تماما متصل[۲]، نگاشت ویژگی‌های ۲ بعدی حاصله از مرحله‌ی Pooling را به بردار ویژگی یک بعدی تبدیل می‌کند. لایه‌های تماما متصل تقریبا ۹۰٪ پارامترهای یک شبکه عصبی پیچشی را شامل می‌شوند. لایه تماما متصل به ما اجازه می‌دهد تا نتیجه شبکه را در قالب یک بردار با اندازه مشخص ارائه کنیم . از این بردار می‌توان برای طبقه بندی استفاده کرد و یا اینکه از آن جهت ادامه پردازش‌های بعدی بهره برد.

[۱] Max Pooling Layer

[۲] Fully connected

در انتها لایه خروجی با تابع فعالساز Sigmoid قرار دارد تا مقادیر خروجی شبکه را بین ۱ و ۰ نگاشت کند.

۳_۵ مدل مبتنی بر ویژگی

به غیر از استفاده از مدل مبتنی بر شبکه عصبی پیچشی که از تعبیه کلمات استفاده می‌کند، ما به طور موازی از مدل مبتنی بر ویژگی نیز استفاده می‌کنیم که از ویژگی‌های مبتنی بر کاربر، مبتنی بر محتوا و روش N-gram استفاده می‌کند. با ظهور شبکه‌های اجتماعی و ورود هرزنامه به این حوزه تحقیقات و پژوهش‌های زیادی درباره انتخاب ویژگی‌های مناسب و تاثیرگذار جهت شناسایی هرزنامه صورت گرفته است. این ویژگی‌ها همانطور که در بخش قبل به آن اشاره شد در بخش-های مختلفی استفاده شده است. ما بر مرجع [۱۰] و همچنین مرجع [۱۹] موثرترین ویژگی‌های تاثیرگذار را در سطح حساب کاربر و محتوای توییت انتخاب کرده‌ایم. ویژگی‌های زیر برای آموزش مدل استفاده می‌شود.

۳_۶ ویژگی‌های مبتنی بر کاربر

ویژگی‌های مبتنی بر کاربر بر روی مشخصات حساب کاربری تمرکز دارد. پارامترهای استفاده شده به عنوان ویژگی حساب-های یک توییت در جدول ۱ آمده است.

تعداد دنبال‌شوندگان و دنبال‌کننده‌ها یکی از ویژگی‌هایی است که تفاوت‌های میان حساب‌های هرزنامه و عادی را متمایز می-کند. معمولا تعداد دنبال‌کنندگان در یک حساب هرزنامه بسیار بیشتر از یک حساب عادی است. امتیاز reputation همانطور که در جدول ۱ نشان داده شده برای کاربرانی است که دنبال-کنندگان زیادی دارند درحالی که تعداد دنبال شوندگان آنها کم است و نزدیک مقدار یک می‌باشد. برای هرزنامه‌گرها، نسبت به کاربران عادی، این امتیاز کمتر است.

جدول ۱. ویژگی های مبتنی بر کاربر

ویژگی	توضیحات
Follower Count	تعداد دنبال کنندگان کاربر را نشان می‌دهد.
Friends Count	تعداد دوستان کاربر را نشان می‌دهد. در توییتر این ویژگی همان تعداد following های کاربر است.
Reputation Score	$\text{Re}\,putation\ Score = \frac{\#followers}{\#followers + \#friends}$
Registration Age of the User	تعداد روزها از زمان ساخته شدن اکانت کاربر تا آخرین توییتی که کاربر گذاشته است.

۳_۷ویژگی‌های مبتنی بر محتوا

توجه ما در ویژگی‌های مبتنی بر محتوا بر روی متن توییت می‌باشد. ویژگی‌های مورد استفاده در تحقیق ما شامل موارد جدول ۲ می‌باشد.

با توجه به مسدود شدن حساب‌های هرزنامه توسط الگوریتم‌های موجود، حساب‌های هرزنامه عمر کمتری به نسبت حساب‌های عادی دارند. از طرفی تعداد دنبال‌شوندگان حساب‌های هرزنامه بسیار کمتر از حساب‌های عادی است. و بالعکس با توجه به تلاش حساب‌های هرزنامه تعداد دنبال کنندگان این حساب‌ها هم در مقایسه با حساب‌های عادی بیشتر است. با توجه به فرمول شهرت، حساب‌های عادی از امتیاز شهرت بیشتری برخوردارند. از طرفی هرزنامه‌ها به دلیل ماهیت تبلیغاتی خود از آدرس‌های URL بیشتری در متن خود استفاده می‌کنند.

جدول۲. ویژگی‌های مبتنی بر محتوا

ویژگی	توضیحات
Number of Words	تعداد لغاتی که درون توییت قرار دارند
Length of the Tweet	طول یک توییت بر حسب کاراکتر را محاسبه می‌کند.
Number of URL Links	این ویژگی تعداد لینک‌های URL موجود در توییت را محاسبه می کند.

۳_۸ ویژگی های مبتنی بر N-gram

N-gram مدل زبانی است بر پایه احتمالات که پیش‌بینی قلم بعدی را انجام می‌دهد. امروزه این مدل کاربرد فراوانی در زبان های طبیعی پردازشی دارد. به چند روش امکان ارائه این مدل وجود دارد. یکی از این روش‌ها تکنیک TF-IDF می‌باشد. این روش محبوب‌ترین تکنیک استخراج بردار از متن بوده است. TF[۱] به معنی فراوانی وزنی کلمه کلیدی و IDF[۲] به معنیِ برعکسِ تعداد تکرار در متون است. برای به دست آوردن ضریب TF-IDF می‌بایست هر کدام از این دو عبارت را به صورت جداگانه محاسبه کرده و حاصل دو عبارت را در هم ضرب کنیم تا نتیجه حاصله، فراوانی وزنی کلمه کلیدی را به ما نشان دهد. روابط TF-IDF به شرح زیر می‌باشد:

TF عبارت است از تقسیم تعداد تکرار کلمه کلیدی w بر تعداد کل کلمات محتوا N (رابطه (۳)).

$$TF = \frac{w}{N} \qquad (۳)$$

IDF عبارت است از لگاریتم تقسیم تعداد کل محتوا بر محتواهایی که شامل کلمه مورد نظر هستند (رابطه (۴)).

$$IDF = \log(۱ + \frac{c}{d}) \qquad (۴)$$

همچنین c تعداد کل سندها و d تعداد سندهایی است که کلمه کلیدی در آن قرار دارد. در واقع هر چه یک کلمه در یک متن بیشتر تکرار شده باشد TF و در دیگر متون کمتر تکرار شود IDF مقدار TF-IDF آن بیشتر می‌شود و این معیار خوبی جهت تشخیص وزنِ یک کلمه در یک جمله می‌باشد. با این تکنیک می‌توان میزان مهم بودن یک کلمه را سنجید. به این ترتیب ماتریسِ ویژگی با کیفیت بهتری تشکیل می‌شود و معمولاً نتایج بهتری در الگوریتم‌های طبقه‌بندی یا خوشه‌بندی حاصل می‌شود.

Unigram ۹_۳

با توجه به توضیحات مربوط به روش TF-IDF جهت بهبود طبقه‌بندها می‌توان از یک ویژگی مهم دیگری که کاربرد موثری در پردازش متون دارد نیز استفاده کرد. این ویژگی‌ها که به ویژگی‌های N-gram معروف هستند از توالی قرارگیری کلمات پشت سر هم می‌توانند ویژگی جدیدی بوجود آورند. ما در این تحقیق از دو ویژگی Unigram و Bigram استفاده می‌کنیم. در ویژگی Unigram هر کلمه به طور مستقل به عنوان یک ویژگی در نظر گرفته می‌شود. برای محاسبه Unigrams از رابطه (۵) استفاده می‌کنیم.

$$P_{unigram}(w_i) = \frac{count(w_i)}{count(all\ words)} = \frac{N(w_i)}{N_{total}} \qquad (۵)$$

در رابطه بالا $N(w_i)$ تعداد تکرار لغت w_i در توییت می‌باشد. و N_{total} تعداد کل لغات درون توییت می‌باشد.

Bigram ۱۰_۳

Bigram به معنی در نظر گرفتن یک کلمه قبل از هر کلمه می‌باشد و برای محاسبه آن از رابطه (۶) استفاده می‌کنیم.

$$P_{bigram}(w_j|w_i) = \frac{N(w_i w_j)}{N(w_i)} \qquad (۶)$$

که در آن $N(w_i w_j)$ تعداد کلمات دوتایی است که با کلمه w_i شروع می‌شوند و بعد از آن کلمه w_j قرار می‌گیرد و $N(w_i)$ تعداد کل کلمه w_i می‌باشد. مقادیر مختلف j مشخص می‌کند چه کلماتی بعد از کلمه w_i ظاهر می‌شوند.

ما همانند تحقیقات [۱۰] و[۳۰] از روش Unigram و Bigram با محاسبه TF به عنوان یک ویژگی برای عملیات تشخیص هرزنامه استفاده می‌کنیم. در ادامه به بررسی راهکار پیشنهادی می‌پردازیم.

۱۱_۳ بررسی جزئیات راهکار پیشنهادی

همانند تحقیقات گذشته برای تشخیص هرزنامه‌ها در سطح توییت، از ویژگی‌های مختلفی استفاده شده است از جمله لغات بکار رفته در توییت را به عنوان یک ویژگی در نظر می‌گیریم،

[۱] Term Frequency

[۲] Inverse Document Frequency

ویژگی‌های خاص هر کاربر و اطلاعات کاربر و همچنین ویژگی-های مبتنی بر محتوا را نیز در سیستم دخیل می‌کنیم. اما بسیاری از تولیدکنندگان هرزنامه به خاطر اینکه کاربر را ترغیب به کلیک کردن و دنبال کردن لینک‌های درون هرزنامه کنند، از ویژگی‌هایی استفاده می‌کنند که به احساسات کاربر مرتبط می‌باشد. به عبارت دیگر، آنها سعی می‌کنند در توییت، کاربر را از نظر احساسی ترغیب به کلیک کردن روی لینک‌های هرزنامه کنند. در این تحقیق سعی داریم که علاوه بر بعضی ویژگی‌های بکار رفته در کارهای گذشته ویژگیهای احساسی مورد استفاده در تحقیق [۱۵] را نیز به نحو موثری اضافه کنیم تا یک مدل ترکیبی قوی‌تری ایجاد شود.

ما برای تشخیص هرزنامه از دو مدل شبکه عصبی پیچشی استفاده می‌کنیم. مدل آموزش دیده با استفاده از Twitter Glove Word Embeddings و مدل آموزش دیده توسط Google News Corpus Word۲vec Embeddings و نشان می‌دهیم استفاده از ویژگی‌های احساسی و روش انتخاب ویژگی می‌تواند با دقت بیشتری عملیات تشخیص هرزنامه را انجام دهد.

۳_۱۱_۱ تحلیل احساسات

در ادبیات موضوع، واژه‌های تحلیل احساسات، افکار اندیشی و عقیده کاوی معمولا به صورت مترادف استفاده می‌شوند که با پردازش زبان طبیعی مرتبط می‌باشند. در این حوزه تحقیق از روش‌های داده کاوی نیز بطور گسترده استفاده می‌شود. تحلیل احساسات به نوعی با بررسی نظرات افراد که در بخش‌های مختلف از جمله شبکه‌های اجتماعی ارائه می‌شوند مرتبط است. به طور کلی می‌توان نظرات و افکار را در حوزه تحلیل احساسات در سه گروه مثبت، منفی و خنثی تقسیم‌بندی کرد. به عنوان مثال جمله " این کتاب عالی است! " یک جمله با بار مثبت بوده و از جمله " از این غذا متنفرم" می‌توان بار منفی را دریافت کرد. همین برداشت را می‌توان در حوزه نظرات مردم در شبکه های اجتماعی و توییت‌های آن ها نیز استفاده کرد. توییت‌های کاربران می‌تواند دارای بار مثبت، منفی و یا خنثی باشد. با توجه به این نکته که یکی از اهداف تولیدکنندگان هرزنامه ترغیب کاربر به کلیک بر روی یک لینک آدرس می-باشد، لذا بار معنایی یک توییت باید جذاب، جلب کننده و به طور کلی مثبت باشد. شکل ۴ نمونه‌ای از توییت‌های هرزنامه را نشان می‌دهد. در این توییت‌ها کلمات مثبتی همچون Best، Excellent و غیره مشاهده می‌شود. به همین منظور ما با استفاده از ویژگی‌های خاص مرتبط با تحلیل احساسات سعی می‌کنیم بار مفهومی توییت را شناسایی کنیم. امروزه تحقیقات زیادی بر روی تحلیل احساسات متون انجام گرفته است. از این رو کتابخانه‌های آماده فراوانی جهت تحلیل احساسات ایجاد شده است. در این تحقیق کتابخانه SentiWordNet [۲۷] استفاده شده است.

SentiWordNet ابزاری است که در عقیده کاوی به طور گسترده مورد استفاده قرار می‌گیرد و بر اساس یک دیکشنری واژه‌نامه انگلیسی به نام WordNet کار می‌کند. این دیکشنری صفات، اسامی، افعال و سایر کلاس‌های گرامری را به مجموعه-هایی از مترادف‌ها به نام synset طبقه‌بندی می‌کند. SentiWordNet سه امتیاز به synset های دیکشنری WordNet برای نشان دادن احساسات داخل متن نسبت می-دهد: مثبت، منفی و خنثی. این امتیازات که بصورت مقادیر بین صفر و۱ هستند و در رابطه (۷) نشان داده شده با استفاده از یک روش یادگیری ماشین بدون نظارت بدست می‌آیند.

$$PosS.score(term) + Neg.Score(term) + Objective\ Score(term) = ۱ \quad (۷)$$

به عنوان مثال امتیازات یک کلمه در SentiWordNet می-تواند به ترتیب به احساسات مثبت، منفی و خنثی امتیاز [۰,۰.۱۵۰,۰.۸۵۰] باشد. امتیاز هر کلمه در SentiWordNet می‌تواند بر اساس نقش گرامری آن در جمله متفاوت باشد. به همین منظور ما قبل از امتیاز دهی به واژگان در جمله نقش گرامری آنها را استخراج می‌کنیم. برای

این منظور از روش برچسب گذاری اجزاء کلام[1] یا به اختصار POS Tagger استفاده می‌کنیم.

۳_۱۱_۲ برچسب‌گذاری اجزاء کلام

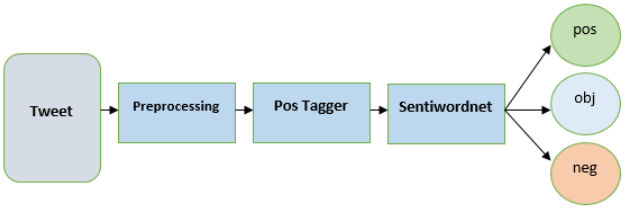
برچسب‌گذاری اجزاء کلام یک نوع برچسب‌گذاری خاص بر روی کلمات یا توکن‌های متن است که بر اساس آن جایگاه گرامری هر کلمه در متن بدست می‌آید. این برچسب‌گذاری در تحقیقات تحلیل متن کاربرد فراوانی دارد. برچسب‌گذاری می‌تواند در سطوح مختلفی اجرا شود، به عنوان مثال از تعیین کلی نقش‌ها مثل فعل، فاعل، صفت و قید تا جزئی‌ترین نقش‌های گرامری کلمه. به منظور امتیازدهی صحیح واژها، برنامه POS بر روی مجموعه داده اعمال می‌شود تا قطب‌بندی (مثبت، منفی، خنثی) انجام شود. تاکنون از روش‌های برچسب‌گذاری گوناگونی در زبان‌های مختلف استفاده شده‌است. این روش‌ها را می‌توان به دو دسته اصلی تقسیم‌بندی کرد. روش‌های آماری و روش-های مبتنی بر الگوریتم‌های یادگیری ماشین.

نمونه خروجی جمله "Manchester United isn't looking to sign any forward." بعد از جداسازی و برچسب گذاری اجزاء کلام در کتابخانه spaCY پایتون به صورت زیر می‌باشد.

Manchester PROPN compound
United PROPN nsubj
is VERB aux
n't ADV neg
looking VERB ROOT
to PART aux
sign VERB xcomp
any DET advmod
forward ADV advmod
. PUNCT punct

همانطور که نتایج نشان می‌دهد هر کلمه بر اساس نقش خود در جمله برچسب گذاری شده است. نام، صفت، قید و افعال به صورت مثبت یا منفی برچسب گذاری می‌شوند.

SentiWordNet با یک دیکشنری برچسب دار ارزیابی می-شود. در شکل ۴ مراحل امتیاز دهی کلمات متن توییت نشان داده شده است. به دلیل آنکه مقادیر امتیاز یک کلمه با توجه به نقش آن در جمله بر حسب صفت، قید و یا فعل متفاوت است، ابتدا نقش کلمات توسط POS تعیین شده و سپس توسط SentiWordNet امتیاز دهی می‌شود. شکل ۴ مراحل پردازش ویژگی‌های احساسی روش پیشنهادی را نشان می‌دهد.



شکل ۴. مراحل پردازش ویژگی‌های احساسی روش پیشنهادی

۳_۱۱_۳ ویژگی‌های مبتنی بر تحلیل احساسات

تمرکز در تحقیق [۱۹] تنها محدود به ویژگی‌های تکراری و محدودی بوده است. در این تحقیق برای اینکه بتوانیم از احساسات بکار رفته در توییت نیز بهره ببریم از یک زیرسیستم تحلیل احساسات بکار رفته در تحقیق [۱۵] استفاده می‌کنیم و تلاش می‌کنیم ویژگی‌های احساسی بکار رفته در توییت را استخراج کنیم. پس از استخراج این ویژگی‌ها، از یک الگوریتم شبکه عصبی برای آموزش روی تمامی این ویژگی‌ها استفاده می‌کنیم.

ما ویژگی‌های احساسی (همراه با محتوا و ویژگی‌های مبتنی بر کاربر که در جدول ۱و۲ ذکر شدند) را به عنوان بخشی از رویکرد شناسایی هرزنامه برای توییتر پیشنهاد می‌کنیم. این ویژگی‌ها در جدول ۳ شرح داده می‌شوند.

[1] Part of Speech Tagging

جدول ۳. ویژگی‌های پیشنهادی مبتنی بر احساسات [۱۵]

ویژگی	توضیحات
تعداد کلمات منفی	تعداد کل لغات منفی در یک توییت.
نرخ کلمات منفی	این ویژگی از طریق فرمول زیر محاسبه می‌شود. $\frac{TotalNegativeWords}{TweetLength}\times 100$
امتیاز منفی	مقدار این متغیر از طریق جمع کردن همه negative words scores های یک توییت بدست می‌آید.
تعداد کلمات مثبت	تعداد کل لغات مثبت در یک توییت.
نرخ کلمات مثبت	این ویژگی از طریق فرمول زیر محاسبه می‌شود. $\frac{TotalPositiveWords}{TweetLength}\times 100$
امتیاز مثبت	مقدار این متغیر از طریق جمع کردن همه positive words scores های یک توییت بدست می‌آید.
امتیاز خنثی	مجموع امتیاز خنثی کلمات توییت محاسبه می‌شود
صفت‌های مثبت	مقدار این ویژگی برابر است با همه صفت‌های درون یک توییت به صورتی که مقدار احساسی مثبت آن از یک آستانه ثابت بیشتر باشد.
افعال مثبت	مقدار این ویژگی برابر است با همه فعل‌های درون یک توییت به صورتی که مقدار احساسی مثبت آن از یک آستانه ثابت بیشتر باشد.
قیود مثبت	مقدار این ویژگی برابر است با همه قیود درون یک توییت به صورتی که مقدار احساسی مثبت آن از یک آستانه ثابت بیشتر باشد.

در این نوشتار، از ویژگی‌های احساسی ذکر شده در جدول ۳ همراه با ویژگی‌های ذکر شده در جدول‌های ۱ و ۲ استفاده می‌کنیم. برای شناسایی کلمات با بار مثبت امتیاز Positive کلماتی که بیشتر از ۰/۵ باشند را به عنوان کلمات مثبت در نظر می‌گیریم. به منظور شناسایی جملاتی که بار مثبت آنها در فعل، صفت و یا قید جمله قرار دارند، امتیاز این کلمات را نیز بطور مجزا محاسبه می‌کنیم.

۳_۱۱_۴ استفاده از طبقه‌بند برای آموزش ویژگی‌ها

حال باید ویژگی‌های یاد شده بخش قبل را توسط یک الگوریتم یادگیری ماشین آموزش دهیم. در تحقیقات Wang و همکاران [۱۰] طبقه بندهای Random Forest, SVM, KNN و Decision Tree بر روی ویژگی‌های مبتنی بر محتوا و مبتنی بر کاربر اعمال شده است. نتایج تحقیقات [۱۰] و [۱۹] نشان می‌دهد بهترین نتیجه برای مجموعه داده‌های متعادل و نامعتادل هرزنامه توسط دو طبقه‌بند Random Forest و SVM بدست می‌آید. به همین منظور ما تمامی ویژگی‌های انتخابی را به عنوان ورودی به این دو الگوریتم یادگیری ماشین وارد می‌کنیم. Random Forest یک الگوریتم یادگیری ماشین با قابلیت استفاده آسان است که اغلب اوقات نتایج بسیار خوبی را حتی بدون تنظیم فراپارامترهای آن، فراهم می‌کند. این الگوریتم به دلیل سادگی و قابلیت استفاده، هم برای طبقه-بندی و هم رگرسیون[۱]، یکی از پرکاربردترین الگوریتم‌های یادگیری ماشین محسوب می‌شود. در این تحقیق برای بهبود نتایج، پارامتر number_of_trees را برابر ۵۰ و min_samples_split را برابر ۱۰ در نظر می‌گیریم. از این طبقه بند بر روی هر دو مجموعه داده استفاده میکنیم و نتایج حاصل با طبقه‌بند SVM مقایسه می‌شود. بهترین نتیجه بر اساس معیار F-Measure به عنوان طبقه‌بند نهایی در معماری نهایی قرار می‌گیرد. طبقه‌بند SVM بر اساس طبقه بندی خطی داده‌ها عمل می‌کند. از این طبقه‌بند در هر مسئله تشخیص الگو و طبقه‌بندی می‌توان استفاده کرد. برای عملکرد بهتر این طبقه‌بند پارامتر C = ۰,۸ ، kernel = linear و penaltyرا معادل ۱۲ در نظر می‌گیریم. این مقادیر بر مبنای مقادیر استفاده شده در مقالات و انجام چند آزمایش انتخاب شده‌اند. این طبقه‌بند هم بر روی هر دو مجموعه داده اعمال می‌شود و نتایج با طبقه‌بند Random Forest مقایسه می-شود.

۳_۱۱_۵ انتخاب طبقه‌بندهای ساختار تلفیقی

[۱] Regression

ساختار روش پیشنهادی در شکل ۵ نشان داده شده است. ما برای روش پیشنهادی خود از تلفیق مدل مبتنی بر شبکه عصبی پیچشی و مدل مبتنی بر یادگیری ماشین استفاده می‌کنیم.

روش‌های مبتنی بر شبکه عصبی پیچشی با تعبیه کلمات با ابعاد مختلف آموزش داده می‌شوند. اولین شبکه عصبی پیچشی با روش تعبیه کلمه GLOVe با ابعاد ۲۵ و ۵۰ آموزش داده می شود. بهترین نتیجه خروجی بر مبنای پارامتر F۱ از بین این دو شبکه عصبی پیچشی به عنوان شبکه عصبی پیچشی برگزیده انتخاب می‌شود. دومین شبکه عصبی پیچشی با تعبیه کلمه Word۲vecبا ابعاد ۳۰۰ آموزش داده می‌شود. این شبکه عصبی پیچشی نیز به عنوان شبکه عصبی پیچشی انتخابی در معماری تلفیقی قرار می‌گیرد. با توجه به تابع Sigmoid استفاده شده در این شبکه‌های عصبی، خروجی مقدار پیوسته بین صفر تا ۱ می‌دهد.

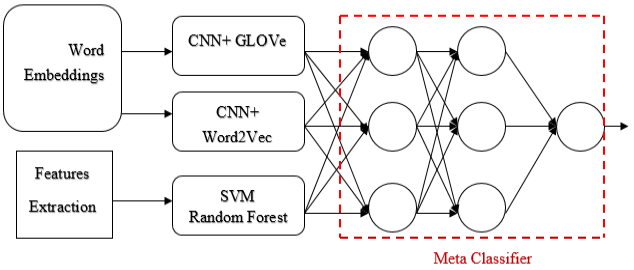
در ادامه مدل مبتنی بر ویژگی که شامل ویژگی‌های مبتنی بر کاربر، مبتنی بر محتوا، ویژگی‌های N-gram و ویژگی‌های احساسی هستند به صورت مجزا توسط طبقه بند Random Forest و SVM آموزش می‌بینند. خروجی این دو طبقه‌بند دودویی صفر برای ورودی غیرهرزنامه و ۱ برای ورودی هرزنامه می‌باشد. از بین این طبقه‌بند نیز بهترین نتیجه ارزیابی F۱ به عنوان طبقه‌بند منتخب در معماری تلفیقی ما جای می‌گیرد. در کل ما از ۳ طبقه‌بند نهایی برای ساخت معماری تلفیقی خود بهره می‌گیریم.

۳_۱۱_۶ استفاده از ابرطبقه‌بند[۱] برای تلفیق خروجی‌ها

در مرحله آخر برای طراحی مدل نهایی ما سه طبقه‌بند منتخب داریم. دو طبقه‌بند مبتنی بر شبکه عصبی پیچشی که خروجی بین صفر و ۱ دارند و طبقه‌بند منتخب مبتنی بر ویژگی که خروجی صفر و یا ۱ نهایی خود را ارائه می‌دهد. در مقالات مختلف روش‌های مختلفی برای تلفیق‌کردن طبقه‌بندها و استخراج نتیجه بهینه‌تر استفاده می‌شود. روش‌هایی مانند Bootstrap Aggregation، Boosting، Majority Voting، Weighted Voting برخی از آنها می‌باشند. بیشتر برندگان رقابت چالش داده از روش‌های تلفیقی استفاده می‌کنند [۳۱], [۳۲]. در یادگیری تلفیقی، کارایی گروهی اغلب بهتر از کارایی روش‌هایی است که به تنهایی استفاده می‌شوند.

در روش پیشنهادی، یک مجموعه داده جدید بوسیله خروجی های ۲ مدل شبکه عصبی پیچشی و یک مدل مبتنی بر ویژگی ساخته می‌شود. یک ابر طبقه‌بند مبتنی بر شبکه عصبی روی مجموعه داده ساخته شده اعمال می‌شود تا توییت داده شده را طبقه‌بندی کرده و تشخیص دهد هرزنامه است یا خیر.

مقادیر ورودی این مدل بازه عددی صفر تا ۱ خروجی طبقه‌بند-های قبلی می‌باشد. این شبکه عصبی دارای دو لایه مخفی می-باشد که هر کدام ۳ گره داخلی دارد. تابع فعال سازی Relu در لایه‌های مخفی استفاده می‌شود و تابع Sigmoid در لایه خروجی استفاده می‌شود. در این تحقیق از تابع فعالساز Sigmoid در لایه خروجی استفاده می‌شود تا اطمینان حاصل شود که خروجی نهایی در بین مقادیر صفر و ۱ قرار می‌گیرند. اگر مقادیر خروجی از حدآستانه کمتر باشند به عنوان توییت نرمال و در غیر اینصورت به عنوان هرزنامه شناسایی می‌شوند. شکل ۵ معماری این ابرطبقه بند را نشان می‌دهد.

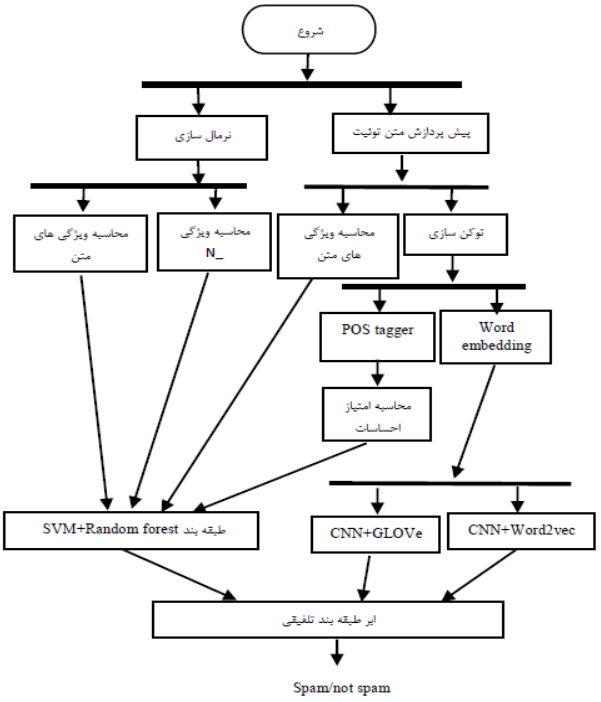


شکل ۵. معماری تلفیقی پیشنهادی

همانطور که از شکل ۵ مشخص است ابتدا متن توییت توسط دو روش تعبیه لغت GLOVe و Word۲vec به بردارهای عددی تبدیل شده و هر کدام به صورت مجزا به معماری شبکه عصبی پیچشی توضیح داده شده در بخش قبل وارد می‌شوند.

[۱] Meta Classifier

در همین حال ویژگی‌های استخراج شده که شامل ویژگی‌های مبتنی بر متن توییت، مبتنی بر کاربر و ویژگی‌های احساسی هستند وارد طبقه‌بند انتخابی Random Forest و یا SVM می‌شوند. دو مدل مبتنی بر شبکه عصبی پیچشی با توجه به تشخیص توییت، امتیازی بین صفر و یک به متن توییت می-دهند. هرچه این عدد به مقدار عددی ۱ نزدیک‌تر باشد نشان-دهنده احتمال بالای هرزنامه بودن متن توییت می‌باشد. و بالعکس هرچه این عدد به سمت صفر نزدیک‌تر باشد سالم بودن توییت بیشتر است. هر دو روش مبتنی بر شبکه عصبی پیچشی مقادیر خروجی خود را به عنوان ورودی به شبکه عصبی تلفیقی وارد می‌کنند. از طرف دیگر یکی از دو طبقه‌بند SVM و یا Random Forest با توجه به عملکرد بهتر خود در مواجهه با مجموعه‌دادها انتخاب شده و بر اساس ویژگی‌های تعریف شده، عمل تشخیص هرزنامه یا عادی بودن توییت را انجام می‌دهد. این طبقه‌بند خروجی ۱ را برای هرزنامه و خروجی صفر را برای ویژگی‌های غیر هرزنامه در نظر می‌گیرد. خروجی این طبقه‌بند مبتنی بر ویژگی نیز به همراه دو مدل مبتنی بر شبکه عصبی پیچشی وارد شبکه عصبی تلفیقی می‌شود. معماری شبکه عصبی تلفیقی ما دارای دو لایه مخفی می‌باشد که هر کدام شامل ۳ گره هستند. تابع فعال ساز Relu بر روی این لایه اعمال شده و Sigmoid به عنوان لایه خروجی در نظر گرفته شده است. با توجه به خروجی عددی پیوسته بین صفر و ۱ لایه خروجی، از یک حدآستانه برای تشخیص هرزنامه یا سالم بودن توییت استفاده می‌کنیم. اگر مقدار خروجی بیشتر از حدآستانه انتخابی باشد خروجی هرزنامه در نظر گرفته می‌شود. در صورتیکه خروجی مقداری مساوی و یا کمتر از حدآستانه داشته باشد، به عنوان توییت سالم شناسایی می‌شود. نمودار فعالیت روش پیشنهادی در شکل ۶ داده شده است.



شکل ۶. نمودار روش پیشنهادی

شبکه عصبی پیشنهادی دارای پنج لایه است که لایه اول ورودی است. ابعاد این لایه $l \times d$ است که l طول توییت‌ها و d طول بردار کلمات است. لایه دوم convolution است که بر روی پنجره h کلمه‌ای اعمال می‌شود. برای $h = ۳$ کلمه هدف و یک کلمه قبل و بعد از آن درنظر گرفته می‌شود. در لایه خروجی تابع فعالساز سیگموئید استفاده شده است تا مقادیر خروجی شبکه را بین صفر و ۱ نگاشت کند. در این مدل جهت جلوگیری از بیش برازش از رگرسیون L۲ استفاده شده است. تعداد فیلترهای مورد استفاده ۲۵۰ بوده و پارامترهای

$Dropout = ۰.۲ . Batchsize = ۳۲ . Optimizer = adam$

و تابع هزینه Binary cross entropy تنظیم شده است. در Random Forest پارامتر number_of_trees برابر ۵۰ و min_samples_split برابر ۱۰ تنظیم شده است. پارامترهای طبقه بند SVM عبارتند از $C = ۰.۸$ و $kernel = linear$. طبقه بند شبکه عصبی پیچشی با روش تعبیه کلمه GLOVe با ابعاد ۲۵ و ۵۰ آموزش داده می شود. دومین شبکه عصبی پیچشی با تعبیه کلمه Word۲vec با ابعاد ۳۰۰ آموزش داده می‌شود.

۴. آزمایشها و تحلیل نتایج

در این بخش، رویکرد پیشنهادی برای شناسایی پست‌های هرزنامه در رسانه‌های اجتماعی را ارزیابی می‌کنیم. قبل از بحث در مورد نتایج تجربی، ما شرح مختصری از مجموعه داده‌ها و معیارهای ارزیابی استفاده شده برای آزمایش‌های خود و همچنین روش‌هایی که برای مقایسه استفاده شده را ارائه می-دهیم.

۴_۱ مجموعه داده

ما در این مقاله از دو مجموعه‌داده برای آزمایشات خود استفاده کرده‌ایم. اولین مجموعه زیر مجموعه HSpam۱۴ [۱۳] است که از آن به عنوان مجموعه‌داده HSpam یاد می‌کنیم. مجموعه داده‌های اصلی شامل ۱۴ میلیون توییت است و روند جمع آوری مجموعه داده‌ها به مدت دو ماه انجام گرفته است.

ما ۲۰۰۰ مورد از این توییت‌ها را در نظر می‌گیریم. این مجموعه‌داده شامل ۱۰۰۰ توییت هرزنامه و ۱۰۰۰ توییت غیر هرزنامه است. برای مجموعه‌داده‌های HSpam۱۴ ، هر نمونه یک توییت است و برچسب کلاس مرتبط برای توییت (هرزنامه یا غیر هرزنامه) از قبل موجود می‌باشد. در مجموعه داده ۱KS۱۰KN [۳۳] مجموعه اصلی شامل ۱۰۰۰ توییت هرزنامه و ۱۰۰۰۰ توییت غیرهرزنامه می‌باشد. ما برای تحقیقات خود از یک مجموعه نامتعادل انتخابی استفاده می‌کنیم. لیست انتخابی ما شامل ۲۰۰ توییت هرزنامه و ۲۰۰۰ توییت غیر هرزنامه است. ما از این مجموعه‌داده به عنوان مجموعه داده ۱KS۱۰KN یاد خواهیم کرد. بنابراین در آزمایشات ما ، دو مجموعه‌داده داریم که یکی از آنها متعادل است حاوی تعداد تقریباً مساوی از هرزنامه و غیر هرزنامه در حالی که داده‌های دیگر با عدم تعادل کلاس به صورتی که نمونه‌های کلاس غیر هرزنامه دارای تعداد قابل توجهی بیشتر هستند مواجه هستیم. برای هر دو مجموعه‌داده، ما داده‌ها را به دو مجموعه آموزشی و آزمایشی تقسیم می‌کنیم.

۴_۲ ارزیابی داده‌ها

جهت ارزیابی مدل پیشنهادی باید داده‌ها به دو دسته آموزشی و آزمایشی تقسیم شوند. بطور معمول ۷۰% مجموعه داده به عنوان داده‌های آموزشی و۳۰% باقیمانده به عنوان مجموعه آزمایشی مدل پیشنهادی در نظر گرفته می‌شود. ۷۰٪ مجموعه آموزش برای آموزش هر کدام از طبقه‌بندها مورد استفاده قرار می‌گیرد. در صورت استفاده از روش اعتبارسنجی متقابل جهت آموزش و ارزیابی هر طبقه‌بند، داده‌ها به k زیرمجموعه تقسیم می‌شوند. در هر مرحله تعداد k-۱ مجموعه به عنوان داده‌های آموزشی استفاده شده و یک مجموعه باقی مانده به عنوان داده آزمایشی جهت ارزیابی استفاده می‌شود. این روال k بار انجام می‌گیرد و به این ترتیب همه داده‌ها هم در آموزش مدل و هم در آزمایش مدل مورد استفاده قرار می‌گیرند. میانگین نتایج آزمایش به عنوان ارزیابی نهایی مورد استفاده قرار می‌گیرد.در این آزمایشها مقدار k برابر ۵ انتخاب شده است که در نتیجه باعث می‌شود تقسیم بندی داده‌ها بجای ۷۰٪-۳۰٪ مقدار ۸۰٪-۲۰٪ باشد.

مجموعه داده‌ها شامل اطلاعات کاربر به همراه متن توییت می باشد. توییت‌های هر کاربر در دو گروه هرزنامه و غیر هرزنامه برچسب‌گذاری شده است.

۴_۳ معیارهای ارزیابی

معیارهای مورد استفاده برای ارزیابی روش پیشنهادی ما عبارتند از: دقت [۱]، فراخوانی[۲] ، F-measure که به ترتیب در رابطه‌های (۸) و (۹) و (۱۰) نشان داده شده‌اند.

کلاس هرزنامه را کلاس مثبت و کلاس غیر هرزنامه را کلاس منفی درنظر می‌گیریم.

[۱] Precision

[۲] Recall

True positive (TP) به تعداد توییت‌های هرزنامه که به طور صحیح به عنوان هرزنامه طبقه‌بندی می‌شوند ، اشاره می-کند.

False negative (FN) تعداد توییت‌های هرزنامه را که به اشتباه به عنوان غیر هرزنامه طبقه‌بندی شده‌اند نشان می‌دهد.

False positive (FP) به تعداد توییت‌های غیر هرزنامه اشاره دارد که به اشتباه به عنوان هرزنامه طبقه‌بندی شده است.

True negative (TN) تعداد توییت‌های غیر هرزنامه را که به درستی به عنوان غیر هرزنامه طبقه‌بندی شده‌اند اشاره دارد.

در ادامه فرمول‌ها و نحوه محاسبه معیارهای ارزیابی را نشان می‌دهیم:

$$precision = \frac{TP}{TP + FP} \qquad (۸)$$

$$recall = \frac{TP}{TP + FN} \qquad (۹)$$

$$F1 = 2 * \frac{precision * recall}{precision + recall} \qquad (۱۰)$$

با توجه به اینکه هزینه تشخیص اشتباه توییت سالم به عنوان توییت هرزنامه بیشتر است ، در نتیجه سیستم ارزیابی ما باید با انتخاب مناسب حدآستانه نرخ FPR را تا حد امکان کاهش دهد. رابطه نرخ تشخیص صحیح TPR و نرخ تشخیص اشتباه FPR با رابطه‌های (۱۱) و (۱۲) نشان داده شده است.

$$TPR = \frac{TP}{TP + FN} \qquad (۱۱)$$

$$FPR = \frac{FP}{FP + TN} \qquad (۱۲)$$

۴-۴ روشهای استفاده شده برای مقایسه

در این تحقیق سعی شده است نتایج روش پیشنهادی با تحقیق [۱۹] مقایسه شود. به دلیل محدودیت‌های پردازشی معماری مدل فراطبقه‌بند ما از دو شبکه عصبی پیچشی انتخابی به همراه یک طبقه‌بند مبتنی بر ویژگی تشکیل شده است. به همین منظور پس از ساخت مدل نهایی، این مدل را با ویژگی‌های تحقیق [۱۹] و ویژگی‌های پیشنهادی جهت ارزیابی مقایسه می‌کنیم. برای طراحی مدل تلفیقی نهایی ابتدا هر مدل بر اساس مجموعه داده Hspam و ۱KS۱۰KN آموزش داده شده و بهترین طبقه‌بندها از نظر معیار F-measure به عنوان مدل انتخابی در مجموعه نهایی قرار داده می‌شود. ما در این تحقیق مدل شبکه عصبی پیچشی مبتنی بر تعبیه کلمه GLOVe را با دو بُعد ۲۵ و ۵۰ مقایسه می‌کنیم. بهترین مدل از بین این دو بُعد بر اساس مجموعه‌داده‌های متعادل و نامتعادل در مدل نهایی قرار داده می‌شود. از معماری شبکه عصبی پیچشی مبتنی بر تعبیه کلمه Word۲vec نیز به عنوان یکی دیگر از مدل‌های انتخابی در مدل نهایی استفاده می‌کنیم. از بین طبقه‌بندهای مبتنی بر ویژگی Random Forest و SVM نیز بهترین گزینه انتخاب و در مدل نهایی تلفیقی قرار می‌گیرد.

۴-۵ تحلیل نتایج آزمایش‌ها

نتایج ارزیابی الگوریتم‌ها روی دو مجموعه داده ۱KS۱۰KN و HSpam در ادامه نشان داده شده است. در ادامه روش پیشنهادی و روش مقاله [۱۹] با توجه به ویژگی‌های متفاوت با هم مقایسه شده‌اند. با توجه به جداول خروجی، نتایج در دو مجموعه متعادل و نامتعادل متفاوت هستند. جهت بررسی تاثیر ویژگی‌های پیشنهادی عملکرد هر طبقه‌بند نیز به تنهایی با ویژگی‌های مورد استفاده در مقاله مذکور مقایسه شده است. به این ترتیب می‌توان میزان تاثیر ویژگی‌های پیشنهادی را بر روی هر طبقه‌بند به طور مجزا و بر روی کل مدل مشاهده کرد.

۴_۵_۱ نتایج ارزیابی مجموعه‌داده ۱KS۱۰KN

نتایج جدول ۴ از دو جنبه باید ارزیابی گردند. یکی اینکه نتایج روشهای قبلی ذکر شده در این جدول اگر با ویژگی‌های احساسی همراه شوند بهبودی در نتایج حاصل می‌شود. دوم

اینکه روش تلفیقی پیشنهادی نسبت به بقیه روش‌ها چگونه عمل کرده است. این نکته نیز قابل ذکر است که آزمایشهای روش پیشنهادی و آزمایشهای مربوط به تاثیر ویژگی‌های احساسی با درصدی از کل داده‌ها انجام شده‌اند در حالی که روشهای دیگر کل داده‌ها را مورد استفاده قرار داده‌اند. با این وجود روش پیشنهادی کاهشی را در نتایج نشان نمی‌دهد بلکه افزایشی هم به دنبال داشته است. درنتیجه علیرغم افزودن ویژگی‌های احساسی که انتظار می‌رود باعث کاهش سرعت محاسبات گردد، ولی با دستیابی به نتایجی قابل قبول و حتی کمی بیش از روشهای مورد مقایسه، این مورد جبران می‌گردد.

جهت بررسی تاثیر داده‌های نامتعادل بر روش پیشنهادی از مجموعه‌داده ۱KS۱۰KN استفاده کرده‌ایم. همانطور که جدول ۴ نشان می‌دهد بهترین نتیجه در بین روش‌های مبتنی بر شبکه عصبی پیچشی با تعبیه کلمه GLOVe با ابعاد ۵۰ بوده است. از طرف دیگر مشاهده می‌شود که با افزایش دقت، فراخوانی کاهش می‌یابد و این دو عکس یکدیگر عمل می‌کنند. این نتیجه نشان می‌دهد با افزایش ابعاد تعبیه کلمه دقت مدل افزایش می‌یابد. همچنین هر دو روش تعبیه کلمه GLOVe از روش Word۲vec بهتر عمل کرده‌اند.

همانند نتایج موجود در مقاله مرجع [۱۹]، با ویژگی‌های اولیه و ویژگی‌های روش پیشنهادی برای دقت و معیار F-measure، الگوریتم جنگل تصادفی عملکرد بهتری دارد. در حالی که برای فراخوانی، عملکرد الگوریتم SVM بهتر است. با این وجود عملکرد هر دو طبقه‌بند با ویژگی‌های پیشنهادی در مقایسه با ویژگی‌های اولیه بهتر عمل کرده‌اند. این نتایج نشان می‌دهد انتخاب ویژگی‌های احساسی پیشنهادی نسبت به ویژگی‌های کلاسیک مبتنی بر کاربر و یا متن توییت در مقاله پایه از کارایی بالاتری در داده‌های نامتعادل برخوردار است.

طبقه‌بند Random Forest به علت استفاده از درخت‌های تصمیم مختلف، ویژگی‌های بیشتری را برای یادگیری در نظر می‌گیرد و این امر باعث افزایش دقت و F-measure شده است. SVM نسبت به Random Forest از ویژگی‌های کمتری استفاده می‌کند و تنها نمونه‌های مرزی[۱]را در نظر می‌گیرد. لذا فراخوانی بالاتری را بدست می‌آورد. با توجه به عملکرد بهتر Random Forest در معیار F-measure از این طبقه‌بند به عنوان طبقه‌بند برگزیده برای داده‌های نامتعادل استفاده می‌کنیم.

از طرف دیگر فرآیند آموزش تحت تأثیر این مجموعه داده نامتعادل قرار می‌گیرد زیرا تعداد بیشتری از نمونه‌های آموزش غیرهرزنامه بوده و تعداد نمونه‌های کمتری هرزنامه هستند. علیرغم نتایج ضعیف طبقه‌بندهای مبتنی بر ویژگی، مدل‌های شبکه عصبی پیچشی با تعبیه کلمه عملکرد خوبی در برابر داده‌های نامتعادل داشته‌اند.

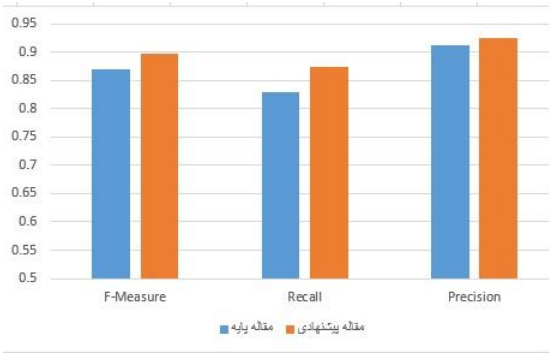
جدول ۴. نتایج ارزیابی برای دیتاست ۱KS۱۰KN

Method	Precision	Recall	F-Measure
CNN + Glove۲۵d	۰,۹۲۱	۰,۸۰۸	۰,۸۶۱
CNN + Glove۵۰d	۰,۹۳۶	۰,۸۱۵	۰,۸۷۱
CNN + Google۳۰۰d	۰,۸۰۸	۰,۸۳۶	۰,۸۲۲
Random Forest ویژگی مقاله پایه	۰,۷۳۰	۰,۹۱۶	۰,۸۱۲
SVM ویژگی مقاله پایه	۰,۶۸۶	۰,۹۳۲	۰,۷۹۰
Random Forest ویژگی‌های پیشنهادی	۰,۷۴۲	۰,۹۲۵	۰,۸۲۳
SVM ویژگی‌های پیشنهادی	۰,۶۹۷	۰,۹۴۵	۰,۸۰۲
مقاله پایه	۰,۹۱۲	۰,۸۳۰	۰,۸۶۹
روش پیشنهادی	**۰,۹۲۵**	**۰,۸۷۳**	**۰,۸۹۸**

با توجه به مقادیر بدست آمده، هر دو طبقه‌بند Random Forest و SVM با ویژگی‌های پیشنهادی عملکرد بهتری نسبت به ویژگی‌های مقاله پایه دارند. این بهبود در عملکرد کلی روش پیشنهادی نسبت به روش تحقیق [۱۹] نیز اثرگذار بوده است. بهبود عملکرد روش پیشنهادی به دلیل استفاده از ویژگی‌های احساسی در روش پیشنهادی می‌باشد. این نتیجه نشان

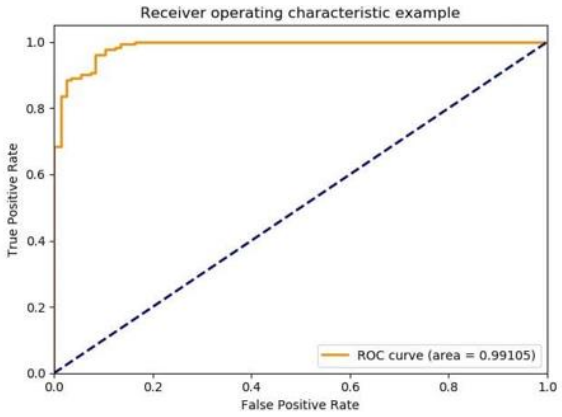
[۱] marginal

می‌دهد ویژگی‌های مورد استفاده در مقاله پایه تاثیرگذاری کمتری نسبت به ترکیب ویژگی‌های احساسی با ویژگی‌های منتخب مبتنی بر کاربر، مبتنی بر متن و N-gram ما دارد. با این حال، وجود ویژگی‌های احساسی باعث افزایش احتمالی زمان اجرای روش پیشنهادی نسبت به مقاله پایه می‌شود. به طور کلی کمترین زمان اجرا متعلق به ویژگی‌های مبتنی بر حساب کاربر می‌باشد. ویژگی‌های مبتنی بر متن و N-gram در رتبه بعدی قرار داشته و بیشترین زمان اجرا به محاسبه ویژگی‌های احساسی اختصاص دارد. لازم به ذکر است در این تحقیق تمرکز اصلی بر استفاده از تحلیل احساسات به همراه دیگر ویژگی‌های توییت‌ها و تحلیل اثر گذاری آنها بوده و افزایش احتمالی زمان اجرا به دلیل تحلیل احساسات در تحقیقات آتی قابل بهبود خواهد بود. اگرچه استفاده از درصد کمتری از داده‌ها برای آموزش و تعداد طبقه‌بندهای کمتر توانسته است تا حدی افزایش زمان را جبران نماید. شکل ۷ مقایسه عملکرد کلی معماری روش پیشنهادی و روش مقاله پایه را نشان می‌دهد.



شکل ۷. مقایسه دو مدل در مجموعه داده ۱KS۱۰KN

نمودار ROC[۱] مجموعه داده ۱KS۱۰KN با AUC[۲] ۰,۹۹۱۰۵ در شکل ۸ نشان داده شده است. محور عمودی نشان دهنده نرخ مثبت درست و محور افقی نشان دهنده نرخ مثبت اشتباه می‌باشد که این نقاط بر اساس حد آستانه‌های مختلف ترسیم شده است. برای هر مورد ورودی، خروجی نهایی طبقه‌بند امتیازی بین صفر تا ۱ را محاسبه می‌کند. امتیاز بالاتر خروجی، شانس هرزنامه بودن را بالا می‌برد. به همین منظور ما با درنظر گرفتن یک حدآستانه بین صفر و۱ کلاس خروجی را تعیین می‌کنیم. اگر مقدار نهایی خروجی عددی بزرگتر از حدآستانه باشد، خروجی هرزنامه شناسایی می‌شود و اگر این مقدار مساوی و یا کوچکتر از حد آستانه باشد خروجی کلاس نرمال تشخیص داده می‌شود. مقدار AUC ۰,۹۹۱۰۵ نشان می‌دهد برای هر جفت توییت هرزنامه و غیر هرزنامه در ۹۹,۱۰۵٪ موارد توییت هرزنامه امتیاز بیشتری از توییت غیر هرزنامه دارد.



شکل ۸. نمودار ROC مجموعه داده ۱KS۱۰KN

۴-۵-۲ نتایج حاصل روی مجموعه‌داده HSpam

همانگونه که در مورد مجموعه داده قبلی اشاره شد، با بررسی نتایج جدول ۵ از دو جنبه‌ی تاثیر ویژگی‌های احساسی بر روی روشهای قبلی و عملکرد روش پیشنهادی با بکارگیری حجم کمتری از داده‌ها، مشاهدی می‌گردد نه تنها ویژگی‌های احساسی تاثیر مثبت در روشهای قبلی دارد بلکه روش پیشنهادی نیز علیرغم استفاده از حجم کمتر داده‌ها نتایج قابل توجهی دارد. نتایج حاصل از رویکرد شبکه عصبی پیچشی ما با روش تعبیه کلمه،Google News ، Twitter Glove Word۲vec، برای مجموعه داده‌های HSpam در جدول ۵ ارائه شده است. Hspam یک مجموعه‌داده متعادل است.

با توجه به متعادل بودن مجموعه‌داده عملکرد هر دو روش مقاله پایه و روش پیشنهادی نسبت به مجموعه‌داده نامتعادل ۱KS۱۰KN عملکرد بهتری دارد. با این حال مقایسه دو روش

[۱] Receiver Operating Characteristic

[۲] Area Under Curve

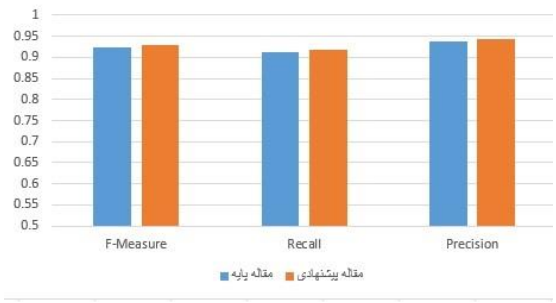
نشان از عملکرد بهتر روش پیشنهادی دارد. در بین روش‌های تعبیه کلمه، شبکه عصبی پیچشی با روش GLOVe با ابعاد ۲۵ بیشترین دقت را دارد. با این وجود شبکه عصبی پیچشی با روش GLOVe با ابعاد ۵۰ معیار F-measure بهتری ارائه می‌دهد. به همین دلیل از بین ابعاد ۲۵ و ۵۰ ما روش GLOVe با ابعاد ۵۰ را برای مدل نهایی انتخاب می‌کنیم. در این مجموعه‌داده بر خلاف مجموعه نامتعادل قبلی، طبقه‌بند SVM دارای معیار F-measure بهتری نسبت به طبقه‌بند Random Forest می‌باشد. لذا در مدل نهایی داده‌های متعادل از این طبقه‌بند استفاده خواهیم کرد. کارایی روش‌های انفرادی در مجموعه‌داده‌های HSpam بسیار بهتر است. از این رو، وقتی ما از روش تلفیقی استفاده می‌کنیم، افزایش قابل توجهی در عملکرد نهایی وجود ندارد زیرا روش‌های فردی به تنهایی دارای عملکرد مناسب هستند. با این حال، برای مجموعه‌داده ۱KS۱۰KN، روش‌های فردی از اندازه و عدم تعادل داده‌ها رنج می‌برند و عملکرد آنها متوسط است. فراطبقه-بند در روش تلفیقی می‌تواند عملکرد را با حاشیه قابل توجهی برای این مجموعه‌داده بالا ببرد.

تولیدکنندگان هرزنامه اغلب با تغییر استراتژی‌های هرزنامه سعی در فریب تکنیک‌های شناسایی هرزنامه دارند. به همین دلیل، الگوریتم‌های شناسایی هرزنامه نیاز به بروزرسانی داشته و یا حداقل به صورت دوره‌ای باید دوباره آموزش داده شوند. همچنین، با گذشت زمان افراد توییت‌هایی درباره رویدادها یا موضوعات جدیدتر ارسال می‌کنند و در نتیجه بسیاری از کلمات و هشتگ‌های جدید به واژگان افزوده می‌شوند. هرزنامه‌گرها سعی می‌کنند از ویژگی‌های سیستم تشخیص مطلع شده و می-توانند نوع توییت‌های هرزنامه را تغییر دهند به گونه‌ای که ویژگی‌های قدیمی مشخص شده در توییت‌های آنها وجود نداشته باشد. سیستم‌هایی که فقط از روش‌های مبتنی بر ویژگی استفاده می‌کنند، شناسایی این نوع توییت‌های هرزنامه برایشان دشوار است. با این حال، الگوریتم ما هر دو روش مبتنی بر ویژگی و مبتنی بر یادگیری عمیق را ترکیب می‌کند. شناخت ویژگی‌های روش‌های یادگیری عمیق بسیار دشوار است زیرا آنها از ویژگی‌های تعبیه کلمات استفاده می‌کنند. بنابراین حتی اگر تولیدکنندگان هرزنامه سعی کنند سیستم تشخیص را فریب دهند، روش ما به اندازه کافی قوی است که بتواند توییت‌های هرزنامه را تشخیص دهد.

جدول۵. نتایج ارزیابی برای دیتاست HSpam

Method	Precision	Recall	F-Measure
CNN + Glove۲۵d	۰,۹۴۳	۰,۸۶۲	۰,۹۰۱
CNN + Glove۵۰d	۰,۹۳۲	۰,۸۹۴	۰,۹۱۲
CNN + Google۳۰۰d	۰,۹۲۹	۰,۸۹۵	۰,۹۱۱
Random Forest ویژگی مقاله پایه	۰,۹۵۲	۰,۷۹۲	۰,۸۶۴
SVM ویژگی مقاله پایه	۰,۹۲۹	۰,۸۵۷	۰,۸۹۱
Random Forest ویژگی‌های پیشنهادی	۰,۹۵۶	۰,۸۰۲	۰,۸۷۱
SVM ویژگی‌های پیشنهادی	۰,۹۳۴	۰,۸۶۴	۰,۸۹۶
مقاله پایه	۰,۹۳۸	۰,۹۱۳	۰,۹۲۵
روش پیشنهادی	**۰,۹۴۲**	**۰,۹۱۸**	**۰,۹۲۹**

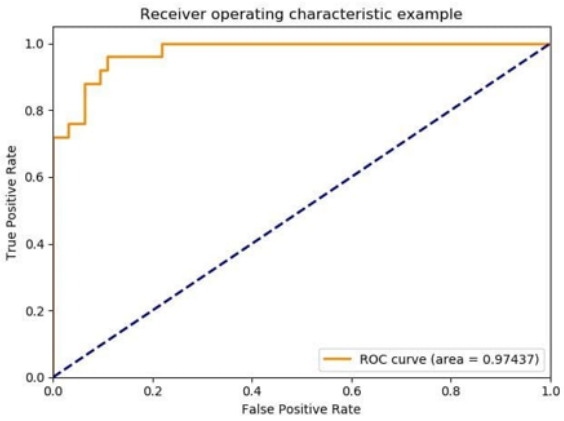
در شکل ۹ مقایسه نتایج نهایی مدل تحقیق [۱۹] و روش پیشنهادی نشان داده شده است. همانطور که در نمودار نشان داده شده روش پیشنهادی در مجموعه‌داده‌های متعادل نیز عملکرد بهتری در مقایسه با روش مقاله پایه دارد.



شکل ۹. مقایسه دو مدل در مجموعه داده Hspam

نمودار ROC برای مجموعه داده Hspam نیز در شکل ۱۰ نشان داده شده است. همانطور که از نمودار مشخص است مقدار

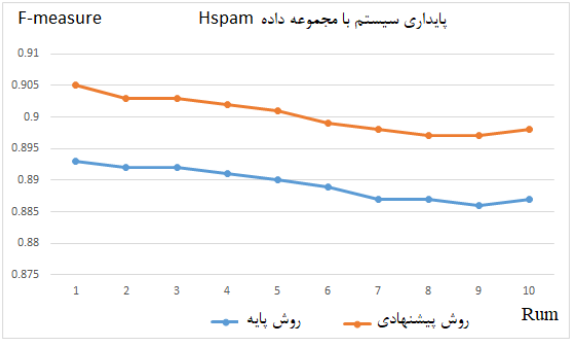
AUC برای نمودار ۰,۹۷۴۳۷ می‌باشد که نشان می‌دهد با دادن جفت توییت ورودی هرزنامه و غیر هرزنامه، در ۹۹,۴۳٪ موارد توییت هرزنامه امتیاز بیشتری نسبت به توییت غیر هرزنامه دارد.



شکل ۱۰. نمودار ROC مجموعه داده Hspam

با توجه به نتایج بدست آمده از آزمایشهای انجام شده بر روی این مجموعه داده‌ها، علیرغم عدم استفاده از کل داده‌ها برای آموزش طبقه‌بندهای پایه و استفاده از تعداد کمتر طبقه‌بندها در روش پیشنهادی، نتایج بیانگر میزان موثر بودن ویژگی‌های احساسی در تشخیص هرزنامه می‌باشد. همچنین می‌توان نتیجه گرفت که از روش پیشنهادی با استفاده از ویژگی‌های احساسی می‌توان در یادگیری انتقالی به شکل موثری بهره برد.

تحلیل دیگری که بر روی روش پیشنهادی و مقاله پایه صورت گرفته است نشان دهنده میزان پایداری خروجی نسبت به مقداردهی اولیه طبقه‌بندهای پایه است. این آزمایش با تعداد کمتری از داده‌های ورودی که ۱۰۰۰ توییت از مجموعه داده Hspam است انجام شده تا با تعداد داده‌های کم میزان پایداری سنجیده شود. در هر اجرا مقادیر اولیه که به صورت تصادفی داده می‌شوند تغییر نموده و با این داده‌ها مقدار f-measure اندازه‌گیری شده که در نمودار شکل ۱۱ نتایج برای ۱۰ اجرای متوالی مشاهده می‌گردد. با بررسی این نمودار ملاحظه می‌گردد که حتی با تعداد داده‌های ورودی کم نیز هر دو سیستم پایداری قابل قبولی نسبت به مقداردهی‌های اولیه دارند، اگرچه روش پیشنهادی مقدار بیشتری را برای این معیار نشان می‌دهد.



شکل ۱۱. مقایسه میزان پایداری سیستم پسشنهادی و روش پایه

۵. نتیجه گیری

یکی از بزرگترین چالش‌های شبکه‌های اجتماعی انتشار هرزنامه در سطح وسیع است که هزینه بسیار زیادی را به این شرکت‌ها تحمیل می‌کند. مشکل بیشتر روش‌های جلوگیری از هرزنامه در شبکه توییتر حذف حساب‌های کاربری انتشار دهنده هرزنامه می‌باشد. در این روش‌ها با حذف حساب، انتشار دهندگان هرزنامه به راحتی حساب‌های جدیدی ایجاد می‌کنند. یکی از روش‌های کارامد، شناسایی و حذف خود توییت هرزنامه بجای حساب کاربری است. ما در این تحقیق با بهره‌گیری از شبکه‌های عصبی پیچشی و ترکیب آن با ویژگی‌های موثر تحقیقات قبلی و ویژگی جدید تحلیل احساسات تلاش کردیم تا دقت تشخیص هرزنامه را افزایش دهیم. استدلال ما این است که هرزنامه‌ها تلاش دارند با ترغیب کاربر نظر وی را جهت کلیک بر روی یک لینک خاص جلب کنند. در نتیجه بار معنایی متن هرزنامه جنبه مثبت دارد. مقایسه روش پیشنهادی با روش‌های قبلی نشان-دهنده این موضوع است که بسیاری از ویژگی‌های مبتنی بر کاربر و متن مانند تعداد کاراکترهای هشتگ و یا علامت سوال تاثیر چندانی در بهبود نتایج روش تشخیص ندارند. از طرفی تولیدکنندگان هرزنامه نیز با بروز کردن روش‌های خود سعی در فریب الگوریتم‌های تشخیص دارند و این موضوع در آینده بر دیگر ویژگی‌های استفاده شده نیز تاثیر خواهد داشت. به همین

دلیل روش‌های تشخیص باید بیشتر بر روی مفاهیم متن توییت از جمله ویژگی‌های احساسی و عقیده کاوی تمرکز داشته باشند.

۶. پیشنهادهایی برای کارهای آینده

با توجه به کارهای گذشته حجم بسیار کمی از تحقیقات تمرکز خود را بر روی متن توییت قرار داده‌اند. در حالیکه این روش دارای پتانسیل بالایی جهت جلوگیری کارآمد انتشار هرزنامه‌ها می‌باشد. اگرچه استفاده از ویژگی‌های مبتنی بر متن مانند تعبیه کلمه و تحلیل احساسات ممکن است باعث افزایش زمان پردازش گردد، می‌توان تمرکز تحقیقات بعدی را بر روی ارائه روش‌هایی به منظور بهبود زمان پردازش قرار داد. آنچه مسلم است، این تحقیق مثبت بودن تاثیر ویژگی‌های احساسی را نشان می‌دهد و کاهش زمان در اولویت بعدی در تحقیقات آتی قرار دارد.

از طرفی تمامی تحقیقات انجام شده تا کنون بر روی توییت‌های زبان‌های غیر فارسی بوده است. با گسترش روزافزون فعالیت کاربران فارسی زبان در توییتر، تمرکز بیشتر بر روی روش‌های تشخیصی هرزنامه در متن توییت‌های فارسی بیش از پیش احساس می‌شود. اگرچه این امر منوط به گسترش کتابخانه‌های تعبیه کلمه و تحلیل احساسات فارسی در زبان‌های برنامه نویسی است.

در تحقیقات آینده می‌توان موارد متعددی را در روش ارائه شده در نظر گرفت اگر از مجموعه داده‌هایی با داشتن ویژگی‌های مرتبط استفاده شود از جمله تاثیر ایموجی‌ها و غیره. در نتیجه این روش قابل توسعه است بخصوص با وجود روش تلفیق استفاده شده.

مراجع

[۱] Top Sites. Alexa Internet. Archived from the original on ۲۳ August ۲۰۱۹. Retrieved May ۱۳, ۲۰۱۳

[۲] Twitter overcounted active users since ۲۰۱٤, shares surge on profit hopes, USA Today, Archived from the original on ۱ January ۲۰۲۰. Retrieved ٤ November ۲۰۱۹

[۳] “California business and professions code". Spamlaws. Retrieved ۲۰۱۳-۰۹-۰۳.

[٤] Grier, C., Thomas, K., Paxson, V., & Zhang, M., Spam: the underground on ۱٤۰ characters or less. In Proceedings of the ۱۷th ACM conference on Computer and communications security, ۲۰۱۰, pp. ۲۷-۳۷.

[٥] Gheewala, S., & Patel, R. Machine learning based Twitter Spam account detection: a review. Second International Conference on Computing Methodologies and Communication (ICCMC), ۲۰۱۸, pp. ۷۹-۸٤.

[٦] Patil, D. R., & Patil, J. B., Malicious URLs detection using decision tree classifiers and majority voting technique. Cybernetics and Information Technologies, ۱۸(۱), ۲۰۱۸, pp. ۱۱-۲۹.

[۷] Thomas K, Grier C, Ma J, Paxson V, Song D. Design and evaluation of a real-time url spam filtering service, in IEEE Symposium on Security and Privacy, IEEE, ۲۰۱۱, pp. ۴۴۷–۶۲.

[۸] Yang C, Harkreader R, Gu G. Empirical evaluation and new design for fighting

[۱۵] Alom, Z., Carminati, B., & Ferrari, E., A deep learning model for Twitter spam detection. Online Social Networks and Media, ۲۰۲۰.

[۱۶] Le, Q., & Mikolov, T., Distributed representations of sentences and documents. In International conference on machine learning, ۲۰۱۴, pp. ۱۱۸۸-۱۱۹۶.

[۱۷] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P., Natural language processing (almost) from scratch. Journal of machine learning research, ۲۰۱۱, pp. ۲۴۹۳-۲۵۳۷.

[۱۸] Kim, Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv:۱۴۰۸,۵۸۸۲, ۲۰۱۴.

[۱۹] Madisetty, S., & Desarkar, M. S. A neural network-based ensemble approach for spam detection in Twitter. IEEE Transactions on Computational Social Systems, ۲۰۱۸, Vol ۵(۴), pp. ۹۷۳-۹۸۴.

[۲۰] Osgood, Charles Egerton, George J. Suci, and Percy H. Tannenbaum, The measurement of meaning. No. ۴۷. University of Illinois press, ۱۹۵۷.

[۲۱] Russell, James A, 'A circumplex model of affect', Journal of personality and social psychology, ۱۹۸۰, Vol ۳۹, pp. ۱۱۶۱.

[۲۲] Russell, James A, and Lisa Feldman Barrett. 'Core affect, prototypical emotional episodes, and other things called emotion: evolving twitter spammers. IEEE Trans InfForensics Secur ۲۰۱۳, Vol ۸(۸), pp ۱۲۸۰–۹۳.

[۹] Chen, C., Zhang, J., Xie, Y., Xiang, Y., Zhou, W., Hassan, M. M. Alrubaian, M., A performance evaluation of machine learning-based streaming spam tweets detection. IEEE Transactions on Computational social systems, ۲۰۱۵, Vol ۲(۳), pp. ۶۵-۷۶.

[۱۰] Wang, B., Zubiaga, A., Liakata, M., & Procter, R., Making the most of tweet-inherent features for social spam detection on Twitter. arXiv preprint arXiv:۱۵۰۳,۰۷۴۰۵, ۲۰۱۵.

[۱۱] X. Zhang, Y. Wang, N. Mou, and W. Liang, "Propagating both trust and distrust with target differentiation for combating link-based Web spam," ACM Trans. Web, vol. ۸, no. ۳, ۲۰۱۴, Art. no. ۱۵.

[۱۲] Wu, T., Wen, S., Xiang, Y., & Zhou, W., Twitter spam detection: Survey of new approaches and comparative study. Computers & Security, ۲۰۱۸, Vol *۷۶*, pp. ۲۶۵-۲۸۴.

[۱۳] Sedhai, S., & Sun, A., Hspam۱۴: A collection of ۱۴ million tweets for hashtag-oriented spam research. In Proceedings of the ۳۸th International ACM SIGIR Conference on Research and Development in Information Retrieval, *۲۰۱۵*, pp. ۲۲۳-۲۳۲.

[۱۴] Sedhai, S., & Sun, A. (۲۰۱۷). Semi-supervised spam detection in Twitter stream. IEEE Transactions on Computational Social Systems, ۲۰۱۵, Vol *۵*(۱), pp.۱۶۹-۱۷۵.

[۲۹] Perveen, N., Missen, M. M. S., Rasool, Q., & Akhtar, N. Sentiment based twitter spam detection. International Journal of Advanced Computer Science and Applications (IJACSA), ۲۰۱۶, ۷(۷), ۵۶۸-۵۷۳.

[۳۰] Martinez-Romo, J., & Araujo, L., Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications, ۲۰۱۳, Vol ۴۰(۸), pp. ۲۹۹۲-۳۰۰۰.

[۳۱] Töscher, A., Jahrer, M., & Bell, R. M., The bigchaos solution to the netflix grand prize. Netflix prize documentation, ۲۰۰۹, pp. ۱-۵۲.

[۳۲] Niculescu-Mizil, A., Perlich, C., Swirszcz, G., Sindhwani, V., Liu, Y., Melville, P., ... & Shang, W. X. Winning the KDD cup orange challenge with ensemble selection. In KDD-Cup ۲۰۰۹ Competition, pp. ۲۳-۳۴.

[۳۳] C. Yang, R. C. Harkreader, and G. Gu. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In Proceedings of RAID, RAID'۱۱, Berlin, Heidelberg, Springer-Verlag, ۲۰۱۱, pp. ۳۱۸-۳۳۷.

dissecting the elephant', Journal of personality and social psychology, ۱۹۹۹, pp. ۷۶: ۸۰۵.

[۲۳] Andrew Ortony, Terence J. Turner, What's Basic About Basic Emotions, Psychological Review, ۱۹۹۰, Vol ۹۷(۳), pp. ۳۱۵-۳۱.

[۲۴] Mohammad, Saif M., Sentiment analysis: Detecting valence, emotions, and other affectual states from text, In Emotion measurement, Woodhead Publishing, ۲۰۱۶, pp. ۲۰۱-۲۳۷.

[۲۵] Kuppens, P., Tuerlinckx, F., Russell, J.A. and Barrett, L.F, The relation between valence and arousal in subjective experience, Psychological Bulletin, ۲۰۱۳, Vol ۱۳۹(۴), pp. ۹۱۷.

[۲۶] Kuppens, P., Tuerlinckx, F., Yik, M., Koval, P., Coosemans, J., Zeng, K.J. and Russell, J.A, the relation between valence and arousal in subjective experience varies with personality and culture, Journal of personality, ۲۰۱۷, Vol ۸۵(۴), pp. ۵۳۰-۵۴۲.

[۲۷] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet ۳,۰: an enhanced lexical resource for sentiment analysis and opinion mining, In Lrec, ۲۰۱۰, vol. ۱۰, pp. ۲۲۰۰-۲۲۰٤.

[۲۸] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P., Gradient-based learning applied to document recognition, Proceedings of the IEEE, ۱۹۹۸, Vol ۸۶(۱۱), pp. ۲۲۷۸-۲۳۲۴.