دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پایان‌نامه کارشناسی

تحلیل شبکه پیچیده‌ای برای شناسایی حساب‌های کاربری جعلی در توییتر

نگارش

سارا اصغری

استاد راهنما

دکتر مصطفی حقیرچهرقانی

اسفند ۱۴۰۰


بسم الله الرحمن الرحيم

به نام خدا

دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)

تعهدنامه اصالت اثر

تاریخ: اسفند ۱۴۰۰

اینجانب سارا اصغری متعهد می‌شوم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آن‌ها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان‌نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

سپاس‌گزاری

شکر و سپاس خداوند عزّوجل را که هر چه دارم از اوست.
از پدر و مادر عزیز و مهربانم که در سختی‌ها و دشواری‌های زندگی همواره یاوری دلسوز و فداکار و پشتیبانی محکم و مطمئن برایم بوده‌اند؛
از استاد گرامی جناب آقای دکتر چهرقانی که در کمال سعه‌ی صدر با حسن خلق و فروتنی هیچ کمکی را در این عرصه بر من دریغ نداشتند؛
و از استاد محترم سرکار خانم دکتر ممتازی که زحمت داوری این پژوهش را متقبل شدند؛
کمال تشکر و قدردانی را دارم.

سارا اصغری
اسفند ۱۴۰۰

چکیده

با گسترش شبکه‌های اجتماعی، روزبه‌روز به تعداد حساب‌های کاربری و اخبار جعلی افزوده می‌شود. به همین دلیل امروزه نیاز به رفع این مشکل و ارائه‌ی روش‌هایی برای شناسایی اطلاعات غیرمعتبر و منابع اخبار جعلی بشدت احساس می‌شود. پیشتر در این زمینه روش‌های بسیاری ارائه شده‌اند که از میان آن‌ها می‌توان به روش‌های مبتنی بر یادگیری ماشین و روش‌های متنوعی (از جمله درخت تصمیم، شبکه‌های عصبی عمیق، رگرسیون لجستیک و SVM) اشاره نمود که برای تشکیل یک مدل کلاس‌بندی و اعطای برچسب حقیقی یا جعلی به هر حساب کاربری به کار رفته‌اند. از نقطه‌نظری متفاوت در هر شبکه‌ی اجتماعی گراف‌هایی وجود دارند که تمامی رخدادها در بستر این گراف‌ها اتفاق می‌افتند؛ به عنوان مثال، در شبکه‌ی اجتماعی توییتر گراف‌های دنبال‌کننده-دنبال‌شونده، کامنت، ری‌توییت، منشن و ... وجود دارند که کاربران حقیقی و جعلی در بستر این گراف‌ها به فعالیت می‌پردازند. هدف این پروژه تحلیل شبکه پیچیده‌ای بر روی حساب‌های کاربری حقیقی و جعلی توییتر می‌باشد تا رفتار کاربران جعلی را در مقایسه با کاربران حقیقی مورد بررسی قرار دهد. در انتها پس از اعمال چندین معیار بر روی مدل‌های گرافی پیاده‌سازی شده، مشاهده شد که برخی از این معیارها در جداسازی کاربران جعلی از حقیقی عملکرد بسزایی دارند و در عمل می‌توان از آن‌ها در شناسایی حساب‌های کاربری جعلی استفاده نمود.

واژه‌های کلیدی:

توییتر، شبکه‌های اجتماعی، شبکه‌های پیچیده، بات، جعلی، حقیقی، حساب کاربری، گراف، همبستگی، مرکزیت، نزدیکی، هارمونیک، رتبه صفحه، دستیابی محلی، درجه

فهرست مطالب

عنوان صفحه

فهرست اشکال

شکل صفحه

فصل اول

مقدمه

۱-۱ معرفی مسئله

در سال‌های اخیر شبکه‌های اجتماعی از جمله توییتر به دلیل قیمت کم، دسترسی آسان و قابلیت انتشار سریع اطلاعات، به یکی از اصلی‌ترین منابع خبری برای میلیون‌ها نفر در سرتاسر جهان تبدیل شده‌اند. با گسترش شبکه‌های اجتماعی، روزبه‌روز به تعداد حساب‌های کاربری و اخبار جعلی (که با هدف گمراه‌کردن خوانندگان و انتشار اطلاعات غلط و مخرب ایجاد می‌شوند) نیز افزوده می‌شود. در سال ۲۰۱۴ شرکت توییتر اعلام کرد بین ۵ الی ۸٫۵ درصد از کاربران توییتر را بات‌ها[۱] تشکیل می‌دهند. همچنین در سال ۲۰۱۷ اونور وارول[۲] و همکاران تعداد کاربران این‌چنینی را عددی مابین ۹ الی ۱۵ درصد تخمین زدند.[۱۸] به کمک این حساب‌ها، سازندگان آن‌ها می‌توانند اطلاعات نادرستی را منتشر نموده و از یک ایده، محصول یا نامزد انتخاباتی پشتیبانی و یا علیه آن اقدام کنند و در نتیجه بر روی تصمیمات میلیون‌ها کاربر حقیقی شبکه تاثیر بگذارند. کاربران در فضای مجازی فاقد سرنخ‌هایی هستند که می‌توانند در دنیای واقعی برای ارزیابی اعتبار اطلاعاتی که در معرض آن‌ها قرار می‌گیرند به کار گیرند. این مشکل برای کاربران بی‌تجربه مشهودتر است، چرا که این افراد به راحتی می‌توانند توسط اطلاعات نامعتبر گمراه شوند. ممکن است بیندیشیم آنچه عقل سلیم و افکار عمومی بیان می‌کند صحیح است، اما به‌طور خاص در موقعیت‌های اضطراری وجود ابزاری برای تشخیص میزان اعتبار اطلاعات آنلاین و شناسایی یک منبع خبری بی‌طرف و قابل اعتماد امری حیاتی است.

۱-۲ معایب کارهای پیشین

پیشتر در این زمینه روش‌های بسیاری ارائه شده‌اند که از میان آن‌ها می‌توان به روش‌های مبتنی بر یادگیری ماشین و روش‌های متنوعی (از جمله درخت تصمیم، شبکه‌های عصبی عمیق، رگرسیون لجستیک[۳] و SVM[۴]) اشاره نمود که برای تشکیل یک مدل کلاس‌بندی و اعطای برچسب حقیقی یا جعلی به هر حساب کاربری به کار رفته‌اند. از نقطه‌نظری متفاوت در هر شبکه‌ی اجتماعی کاربران در بستر چندین گراف به فعالیت می‌پردازند؛ به عنوان مثال، در شبکه‌ی اجتماعی توییتر گراف‌های دنبال‌کننده-دنبال‌شونده،[۵] دیدگاه[۶]، ری‌توییت[۷]، منشن[۸] و ... وجود دارند. بدین ترتیب یک شبکه‌ی پیچیده[۹] تشکیل می‌گردد که می‌توان آن را به کمک مجموعه‌ای

[1] bots
[2] Onur Varol
[3] logistic regression
[4] Support Vector Machines
[5] follower-following
[6] comment
[7] retweet
[8] mention
[9] complex network

از ابزارها و معیارها تحلیل نمود. تا آن‌جا که ما اطلاع داریم، این مطالعه تاکنون بر روی کاربران حقیقی و جعلی شبکه‌های اجتماعی صورت نگرفته است؛ به همین دلیل در این پژوهش سعی شده است این مسئله از دیدگاه مذکور مورد بررسی قرار گیرد.

۱-۳ رویکرد پیش‌رو

هدف این پروژه تحلیل شبکه پیچیده‌ای بر روی حساب‌های کاربری حقیقی و جعلی توییتر می‌باشد تا رفتار کاربران جعلی را در مقایسه با کاربران حقیقی مورد بررسی قرار دهد. پس از ساخت گراف دنبال‌کننده-دوست و گراف دیدگاه، دوازده معیار مرکزیت که در تحلیل شبکه‌های عصبی به کار می‌روند را بر روی هر یک اعمال می‌کنیم. مشاهده خواهد شد که معیارهای نزدیکی، هارمونیک، میانگین کوتاه‌ترین طول مسیر و دستیابی محلی و همچنین همبستگی میان میانگین کوتاه‌ترین طول مسیر و نزدیکی، همبستگی میان میانگین کوتاه‌ترین طول مسیر و هارمونیک، همبستگی میان درجه ورودی و درجه خروجی و همبستگی میان درجه خروجی و رتبه صفحه در جداسازی کاربران جعلی از حقیقی عملکرد بسزایی دارند که این نشان‌دهنده‌ی قابلیت این معیارها در شناسایی حساب‌های کاربری جعلی می‌باشد.

۱-۴ معرفی فصول بعدی

ساختار کلی پایان‌نامه به این صورت است که در فصل دوم جمعاً ۱۲ مقاله مورد مطالعه قرار خواهند گرفت؛ اولین آن‌ها یک مقاله‌ی مروری می‌باشد که به معرفی، مقایسه و دسته‌بندی کارهای پیشین می‌پردازد. در ادامه ۱۱ مقاله آورده می‌شوند که مدل پیاده‌سازی شده‌ی هرکدام با جزئیات مورد بررسی قرار خواهد گرفت. فصل سوم به بیان روش پیشنهادی این پروژه می‌پردازد. در این فصل به تشریح نحوه‌ی پیاده‌سازی، زبان برنامه‌نویسی و کتابخانه‌ی مورد استفاده و معیارهای به کار گرفته شده پرداخته می‌شود. در فصل چهارم مجموعه‌داده‌ی مورد استفاده توصیف می‌شود، نتایج به دست آمده نمایش داده می‌شوند و بطور دقیق مورد تجزیه و تحلیل قرار خواهند گرفت. در نهایت در فصل پنجم به نتیجه‌گیری و جمع‌بندی پرداخته می‌شود و کارهای آتی ذکر خواهند شد.

فصل دوم

مروری بر کارهای مرتبط

در سال ۲۰۲۰ کای شو[1] و همکاران یک مقاله‌ی مروری[2] منتشرکردند که در آن به مرور جامعی بر تاریخچه‌ی اخبار نادرست و ویژگی‌های آن در عصر رسانه‌های اجتماعی پرداخته شده است. همچنین چالش‌های پیش‌رو در کشف اخبار نادرست، انواع مختلف آن که در رسانه‌های اجتماعی رایج‌تر است و رویکردهایی برای شناسایی و جلوگیری از انتشار این اخبار مورد بررسی قرار گرفته است. در انتها درباره‌ی عوامل پشت پرده‌ی انتشار سریع این اخبار و نیز روش‌هایی برای بالابردن سطح دانش عمومی در این حوزه صحبت به میان آمده است.[۱۷]

طبق طبقه‌بندی صورت گرفته، هر داده‌ی جعلی بطور معمول در یکی از دسته‌های زیر قرار می‌گیرد:

۱. عکس‌های جعلی ساخته شده توسط شبکه‌های GAN[3]

۲. ویدیوهای جعلی (به عنوان مثال تغییر چهره‌ی افراد داخل ویدیو و جایگزین‌کردن چهره‌ها بطور هوشمندانه و غیرقابل‌تشخیص)

۳. محتوای چندبُعدی (به عنوان مثال یک عکس جعلی به همراه توضیحات متنی مربوط به آن). این نوع داده در فضای مجازی به وفور یافت می‌شود و معمولا شناسایی آن به کمک روش‌های یادگیری عمیق و شبکه‌های عصبی چالش برانگیز است.

درمورد عواملی که پشت پرده‌ی انتشار اخبار جعلی هستند نیز در این مقاله بطور مفصل توضیح داده شده‌است. برخی از این عوامل به اختصار عبارتند از:

۱. منابع و ناشران

۲. عوامل احساسی نظیر تردید، اضطراب، باورها و غیره

۳. بات‌های کنترل‌کننده‌ی فضای مجازی

مدل‌های شناسایی بات بطور کلی به سه دسته تقسیم می‌شوند:

۱. مدل‌های مبتنی بر گراف: رویکردهای این دسته بر این فرض تکیه می‌کنند که ارتباط میان بات‌ها در شبکه‌های مجازی متفاوت با ارتباط میان کاربران انسانی است. کیانگ کائو[4] و همکاران[۶]، آدام برویر[5]، روئی ایلات[6] و اودی واینسبرگ[7][۵] و توجا خاوند[8] و

[1] Kai Shu
[2] review paper
[3] Generative Adversarial Networks
[4] Qiang Cao
[5] Adam Breuer
[6] Roee Eilat
[7] Udi Weinsberg
[8] Tuja Khaund

همکاران[۱۲] در همین زمینه مطالعاتی انجام داده‌اند که در ادامه روش هریک مورد بررسی قرار گرفته است.

۲. مدل‌های جمع‌سپاری[۹]: در این رویکرد منابع انسانی متخصص برای برچسب زدن به کاربران فضای مجازی (حقیقی - جعلی) استخدام می‌شوند. این رویه قابل اعتماد است و خطای نزدیک به صفر دارد؛ بااین‌حال زمان‌گیر است، مقرون‌به‌صرفه نیست و با توجه به میلیون‌ها کاربرِ رسانه‌های اجتماعی امکان‌پذیر نیست. امروزه رویکردهای جمع‌سپاری و برچسب‌زنی دستی در جمع‌آوری مجموعه‌داده‌های استاندارد برای مدل‌های مبتنی بر ویژگی مورد استفاده قرار می‌گیرند.

۳. مدل‌های مبتنی بر ویژگی: رویکردهای این دسته بر این قاعده استوارند که بات‌ها ویژگی‌های متفاوتی نسبت به کاربران انسانی از خود نشان می‌دهند. برای استفاده از مدل‌های بانظارت[۱۰] مبتنی بر ویژگی، ابتدا باید تفاوت میان کاربران حقیقی و جعلی از منظر ویژگی‌هایی مانند محتوا یا فعالیت در یک مجموعه‌داده‌ی برچسب‌گذاری‌شده مشخص شود. سپس یک دسته‌بند بر روی آن ویژگی‌ها آموزش می‌بیند تا بتواند کاربران جعلی را از کاربران حقیقی در یک مجموعه‌داده‌ی بدون برچسب از یکدیگر جدا کند. به این منظور می‌توان از روش‌های کلاس‌بندی متفاوتی از جمله SVM[۱۶]، جنگل‌های تصادفی[۱۱][۱۴] و شبکه‌های عصبی[۱۳] استفاده نمود. برخی از ویژگی‌های رایج عبارتند از:

(آ) محتوای به‌اشتراک‌گذاشته‌شده توسط کاربر: کلمات، عبارات[۱۸] و موضوعات پُست‌ها[۱۶] در رسانه‌های اجتماعی می‌تواند یک شاخص قوی از فعالیت بات‌ها باشد. همچنین بات‌ها برنامه‌ریزی شده‌اند تا کاربران انسانی را به بازدید از وب‌سایت‌هایی که توسط کنترلگرهایشان اداره می‌شوند، ترغیب کنند؛ از این رو در مقایسه با کاربران انسانی تعداد بیشتری آدرس اینترنتی به اشتراک می‌گذارند.[۷]

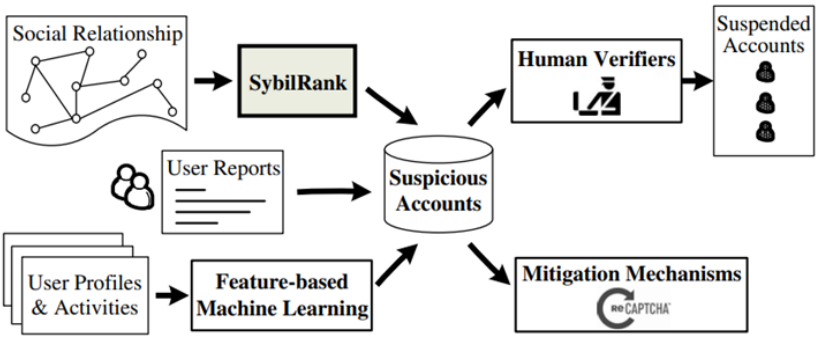
(ب) الگوهای فعالیت: بات‌ها تعداد زیادی توییت را در مدت زمان کوتاهی منتشر می‌کنند و برای مدت طولانی‌تری غیرفعال هستند.[۱۵] علاوه‌براین بات‌ها تمایل دارند الگوهای زمانی بسیار منظم (مانند توییت کردن هر ۱۰ دقیقه یکبار) یا بسیار نامنظمی (وقفه‌ی تصادفی) را دنبال کنند.

(ج) ارتباطات شبکه: بات‌ها تعداد زیادی کاربر انسانی را دنبال[۱۲] می‌کنند به این امید که متقابلاً توسط آن کاربران دنبال شوند. اما این اتفاق معمولاً رخ نمی‌دهد؛ در نتیجه بات‌ها عموماً تعداد دنبال‌شوندگان[۱۳] بسیار بیشتری نسبت به تعداد دنبال‌کنندگانشان

[9] crowdsourcing
[10] supervised
[11] Random Forests
[12] follow
[13] followings

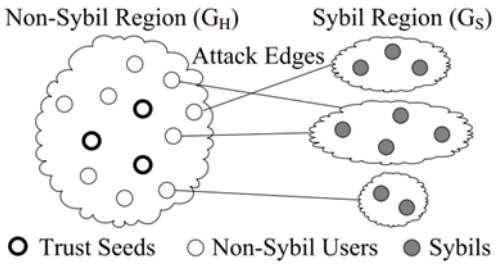
دارند.[۷]

کیانگ کائو و همکاران روشی با عنوان SybilRank معرفی کرده‌اند که از سازوکار مبتنی بر گراف اجتماعی استفاده می‌کند تا کاربران را بر اساس میزان احتمال جعلی (sybil) بودنشان رتبه‌بندی نماید.[۶] پیچیدگی محاسباتی این روش O(n log n) است که در آن n برابر با تعداد گره‌های گراف (تعداد کاربران) می‌باشد. این سیستم می تواند توسط یک ساختار محاسباتی موازی مانند MapReduce نیز پیاده‌سازی شود.



شکل ۲-۱: SybilRank به عنوان بخشی از یک زنجیره‌ی دفاعی در برابر کاربران جعلی[۶]

یک گراف ساده‌ی همبند به فرم G = (V, E) تعریف می‌شود به گونه‌ای که هر گره در V نمایانگر یک کاربر است و هر یال در E نمایانگر یک رابطه‌ی اجتماعی دوطرفه میان دو کاربر شبکه است. گراف را به دو بخش G_H و G_S تقسیم‌بندی می‌کنیم بطوری که زیرگراف G_H شامل تمام گره‌های non-sybil (کاربران حقیقی، کاربران انسانی) و ارتباطات میان آن‌ها و بطور مشابه زیرگراف G_S شامل تمام گره‌های Sybil (کاربران جعلی) باشد.



شکل ۲-۲: یک نمای سطح بالا از گراف G[۶]

یک پیش‌فرض در اینجا این است که تعداد روابط میان کاربران جعلی و کاربران حقیقی بسیار اندک و محدود است؛ به عبارت دیگر تعداد یال‌هایی که دو زیرگراف G_H و G_S را به هم متصل می‌کنند[۱۴] اندک هستند. یک پیش‌فرض دیگر این است که احتمال رسیدن به یک گره

[14] attack edges

پس از طی تعداد گام‌های کافی در الگوریتم گام‌برداری تصادفی[۱۵] متناسب با درجه‌ی آن گره می‌باشد[۱۶].

هدف این است که گره‌های گراف G با توجه به احتمال نرمال‌سازی شده براساس درجه‌شان[۱۷] - که این احتمالات با اجرای الگوریتم گام‌برداری کوتاه‌شده بر روی گراف و با شروع از یک گره non-sybil به دست آمده‌اند - رتبه‌بندی شوند. بدین منظور از روش تکرار توانی[۱۸] استفاده شده است. همچنین برای شناسایی ساختار چندجامعه‌ای[۱۹] در ناحیه‌ی non-sybil روش لووین[۲۰] مورد استفاده قرار گرفته است تا هیچ اجتماعی از گره‌های non-sybil به اشتباه به عنوان بخشی از ناحیه‌ی sybil در نظر گرفته نشود.

مدل SybilRank از لحاظ محاسباتی کارآمد است و می‌تواند در مقیاس‌های بزرگ بر روی گراف‌هایی با صدها میلیون گره اعمال شود. این سیستم با حداقل ۲۰٪ خطای کاذب مثبت و منفی[۲۱] کمتر نسبت به رویکردهای پیشین خود گره‌های جعلی را تشخیص می‌دهد و نتیجه‌ی اجرای آن بر روی یک مجموعه‌داده از شبکه‌ی اجتماعی اسپانیایی تونتی[۲۲] نشان داد ۹۰٪ از ۲۰۰,۰۰۰ حساب کاربری با پایین‌ترین رتبه، واقعاً جعلی بوده‌اند.

آدام بروير، روئی ایلات و اودی واینسبرگ در سال ۲۰۲۰ رویکردی مبتنی بر گراف به نام SybilEdge ارائه کرده‌اند که قادر است حساب‌های کاربری جعلی را در فاصله‌ی زمانی کوتاهی پس از ایجادشان شناسایی کند.[۵] در این روش جعلی بودن یا نبودن هر حساب کاربری جدید بر اساس درخواست‌های دوستی ارسال شده و عکس‌العمل‌های متقابل دریافتی تعیین می‌شود. SybilEdge در تشخیص این موارد توانمند است: برچسب‌گذاری نویز در داده‌های آموزشی، تشخیص پراکندگی متفاوت حساب‌های کاربری جعلی در شبکه و روش‌های متفاوتی که کاربران جعلی به کار می‌گیرند تا کاربران هدفشان را برای درخواست دوستی انتخاب کنند. با وجود فعالیت‌های اندک کاربران جدید، این الگوریتم عملکرد بالایی (AUC > 0.9) از خود نشان داده است.

توجا خاوند و همکاران به بررسی نقش بات‌های توییتر در جریان بلایای طبیعی سال ۲۰۱۷ و ارزیابی راهبردهای هماهنگ آنان در انتشار اطلاعات پرداخته‌اند.[۱۲] مجموعه داده‌های بررسی شده در طول طوفان‌های هاروی[۲۳]، ایرما[۲۴]، ماریا[۲۵] و زمین‌لرزه‌ی مکزیک -که همگی در سال ۲۰۱۷

[15] random walk
[16] convergence of random walk
[17] node's trust
[18] power iteration
[19] multi-community
[20] Louvein
[21] False Positives and False Negatives
[22] Tuenti
[23] Harvey
[24] Irma
[25] Maria

رخ داده‌اند- از توییتر جمع‌آوری شده‌است. در این مطالعه تمرکز بر روی بات‌هایی است که سعی دارند بر روی رفتار بقیه‌ی کاربران شبکه اثر بگذارند.[۲۶]

مجموعه داده‌ی جمع‌آوری شده شامل تنها کاربرانی است که در ارتباط با هر چهار حادثه‌ی ذکر شده توییت یا ری‌توییت داشته‌اند. به منظور برچسب‌گذاری داده‌های آموزشی، احتمال جعلی بودن هر حساب کاربری به کمک سیستم از پیش پیاده‌سازی شده‌ی BotOrNot[۱۰] به دست آمده‌است که در آن به ازای هر حساب کاربری یک امتیاز[۲۷] در بازه‌ی ۰ تا ۱۰۰ بازگردانده می‌شود و هر چه این عدد بزرگتر باشد احتمال جعلی بودن آن حساب بیشتر است. در مرحله‌ی بعد برای تحلیل قدرتمندتر تنها ۱۰۰ حساب کاربری با بالاترین امتیاز و ۱۰۰ حساب کاربری با پایین‌ترین امتیاز مورد مطالعه قرار گرفته‌اند و باقی داده‌ها از مجموعه‌ی مورد مطالعه حذف شده‌اند. حساب‌های کاربری خصوصی[۲۸] و حساب‌های کاربری که توسط خود توییتر تعلیق شده‌اند نیز از میان این ۲۰۰ حساب کاربری حذف شده‌است. هشتگ[۲۹]های مرتبط با ۴ رخداد ذکر شده نیز از میان توییت‌های این کاربران استخراج شده‌است. سپس الگوریتم شناسایی جوامع ارائه شده توسط وینسنت بلوندل[۳۰] و همکاران[۴] به کار گرفته شده‌است. همچنین یک شبکه از هشتگ‌های همزمانی ساخته می‌شود که به وسیله‌ی آن می‌توان هشتگ‌های مختص یک رویداد و هشتگ‌های مشترک بین رویدادهای گوناگون را شناسایی نمود. سپس یک الگوریتم خوشه‌بندی به این شبکه اعمال می‌شود و هشتگ‌ها را در ۴ خوشه قرار می‌دهد.

مشاهده می‌شود که شبکه‌های متعلق به کاربران انسانی در مقایسه با شبکه‌های متعلق به بات‌ها از تعداد جوامع بیشتری تشکیل شده‌اند، اندازه‌شان کوچکتر است و چگال‌تر هستند (نمایانگر وابستگی و تعلق هر کاربر انسانی به یک خوشه‌ی خاص). از لحاظ ساختارِ جوامع، بات‌ها بیشتر حالت سلسه‌مراتبی دارند. بدین معنا که در هر یک از این جوامع یک هسته‌ی مرکزی از کاربران وجود دارد که اتصالات میانشان قوی است اما اعضای پیرامون آن‌ها اتصالات ضعیفی با هسته و با یکدیگر دارند.

در سال ۲۰۱۶ فرد مورستاتر[۳۱] و همکاران مقاله‌ای منتشر کردند که هدف آن ارائه‌ی یک مدل BoostOR[۳۲] برای شناسایی بات‌های توییتر بود به طوری که مقدار امتیاز F1[۳۳] آن (تمرکز بر روی معیار بازیابی[۳۴] و در عین حال تلاش برای بالا نگه داشتن مقدار دقت) بهینه شود.[۱۶] در این رویکرد دو مجموعه داده به دو روش برچسب‌گذاری کاملا متفاوت جمع‌آوری شده‌اند. روش اول استفاده از فرایند برچسب‌گذاری خود توییتر و تقسیم حساب‌های کاربری به سه دسته‌ی «فعال»،

[26] influence bots
[27] bot likelihood score
[28] private
[29] hashtag
[30] Vincent Blondel
[31] Fred Morstatter
[32] Boosting through Optimizing Recall
[33] F1-score
[34] recall

«تعلیق شده» و «حذف شده» می‌باشد (به کمک streaming API و statuses/user-timeline API endpoint). روش دوم ایجاد یک شبکه شامل ۹ عدد honeypot و جمع‌آوری کاربرانیست که جذب این شبکه شده‌اند. برای جمع‌آوری نمونه‌هایی از کاربران انسانی در روش دوم از تکنیک نمونه‌برداری 1-link snowball استفاده شده‌است.

در این مقاله چهار رویکرد متفاوت اکتشافی[۳۵]، SVM، Adaboost و BoostOR مورد مطالعه و ارزیابی قرار گرفته اند. ویژگی‌های استخراج شده در روش اکتشافی به شرح زیر می‌باشند:

- تعداد ری‌توییت‌ها تقسیم بر تعداد کل توییت‌ها (برای هر کاربر بطور جداگانه)
- میانگین طول توییت‌های منتشر شده (برای هر کاربر بطور جداگانه)
- تعداد توییت‌های شامل آدرس اینترنتی تقسیم بر تعداد کل توییت‌ها (برای هر کاربر بطور جداگانه)
- میانگین فاصله‌ی زمانی میان توییت‌های متوالی هر کاربر

در سه رویکرد دیگر از تکنیک مدل‌سازی بر اساس موضوع بهره گرفته شده است. از آنجا که ویژگی‌های متنی خام، تُنُک و دارای ابعاد بالا می‌باشند، بهره‌گیری از آن‌ها سبب بروز مسئله‌ی نفرین ابعاد[۳۶] می‌شود. به همین دلیل از روش LDA[۳۷] برای استخراج و نمایش موضوعی هر کاربر استفاده شده است. در این روش مجموعه‌ی تمام توییت‌های هر کاربر، یک توزیع حول موضوعات (۲۰۰ موضوع) و در هر موضوع یک توزیع حول تمام کلمات مجزای داخل مجموعه داده درنظر گرفته می‌شود.

در دو روش Adaboost و BoostOR هدف رسیدن به یک کلاس‌بند بهینه از طریق گروه‌بندی کلاس‌بندهای ضعیف است. در الگوریتم Adaboost در هر بار تکرار حلقه، به نمونه‌هایی که در دور قبل اشتباه کلاس‌بندی شده بودند، وزن بیشتری در دور بعد داده می‌شود. بدین ترتیب اگر بات‌ها به درستی کلاس‌بندی شوند، از وزن آنها کاسته می‌شود و تمرکز کلاس‌بندهای دور بعد از روی این بات‌ها برداشته می‌شود. در همین راستا در روش BoostOR (که شباهت بسیاری به روش Adaboost دارد) برای حل این مسئله تمهیداتی اندیشیده شده است به گونه‌ای که تغییر وزن به برچسب کاربر نیز وابسته باشد. ارزیابی روش‌های بررسی شده به کمک معیارهای دقت و بازیابی و امتیاز F1 نشان می‌دهد که روش BoostOR بالاترین میزان کارایی را به خود اختصاص می‌دهد.

در طول ۷ ماه مطالعه‌ی طولانی‌مدت (از ۳۰ دسامبر ۲۰۰۹ تا ۲ آگوست ۲۰۱۰)، کیومین

[35] heuristic

[36] curse of dimensionality

[37] Latent Dirichlet Allocation

لی[۳۸]، برایان دیوید ایوف[۳۹] و جیمز کاورلی[۴۰] توانسته‌اند ۲۳,۸۶۹ حساب کاربری را فریب دهند و آنان را جذب مجموعه‌ای شامل ۶۰ حساب کاربری honeypot کنند.[۱۴] در اینجا هر honeypot یک حساب کاربری توییتر است که هدفش نظارت بر تعاملات سایر کاربران است، بطوری که حساب‌های کاربری که آن را دنبال کنند یا به نوعی با آن تعامل برقرار کنند را پیگیری می‌کند و گزارش می‌دهد؛ چرا که دلیلی ندارد کاربری که در صدد نقض قوانین توییتر نیست با چنین پیام‌هایی وسوسه شود یا چنین حساب‌های کاربری را دنبال کند.

می‌توان در هر honeypot محتوا و نوع توییت‌ها (نرمال، پاسخ به یک honeypot دیگر از طریق منشن، حاوی پیوند[۴۱]، حاوی یکی از داغ‌ترین موضوعات فعلی)، الگوی زمانی ارسال آن‌ها و ساختار شبکه‌ی اجتماعی را به گونه‌ای دلخواه تنظیم کرد. honeypot ها به گونه‌ای طراحی شده‌اند که به هیچ عنوان در فعالیت‌های کاربران حقیقی توییتر اختلالی ایجاد نشود. به همین منظور هر honeypot تنها مجاز است honeypot های دیگر را دنبال کند.

در مرحله‌ی بعد کاربران شناسایی شده توسط الگوریتم EM[۴۲] خوشه‌بندی شده‌اند تا بر اساس ویژگی‌ها و رفتارهای مشابه در ۹ خوشه (۴ گروه اصلی) تقسیم‌بندی شوند. در مرحله‌ی آخر از روش‌های کلاس‌بندی برای جداسازی حساب‌های کاربری جعلی از حقیقی استفاده شده‌است. بدین منظور از جعبه ابزار یادگیری ماشین Weka برای بررسی ۳۰ الگوریتم دسته‌بندی (از جمله بیز ساده[۴۳]، رگرسیون لجستیک، SVM و الگوریتم‌های مبتنی بر درخت) به کمک اعتبارسنجی متقابل ۱۰ لایه[۴۴] بهره گرفته شده است. ویژگی‌های مورد استفاده در این قسمت به ۴ دسته‌ی کلی تقسیم بندی می شوند:

- اطلاعات جمعیتی کاربر[۴۵] : مانند طول عمر حساب کاربری
- شبکه‌های دوستی کاربر : از جمله تعداد دوستان و دنبال‌کنندگان، نسبت تعداد دوستان به تعداد دنبال‌کنندگان و درصد دوستی‌های دوجانبه به تعداد کل دوستان/دنبال‌کنندگان
- توییت‌های منتشر شده : از جمله تعداد کل توییت‌ها، متوسط تعداد توییت‌ها در روز، نرخ تعداد پیوندهای به کار رفته به تعداد کل توییت‌ها، نرخ تعداد منشن‌های به کار رفته به تعداد کل توییت‌ها، میانگین شباهت متنی میان تمام جفت توییت‌های منتشر شده توسط کاربر و میزان فشرده‌سازی متن توییت‌ها
- تاریخچه کاربر : مانند نرخ تغییر تعداد دوستان کاربر در طول زمان

[38] Kyumin Lee
[39] Brian David Eoff
[40] James Caverlee
[41] link
[42] Expectation Maximization
[43] Naive Bayes
[44] 10 fold cross validation
[45] User Demographics

برای ارزیابی کلاس‌بند به کار رفته از معیارهای دقت، بازیابی، امتیاز F1، صحت، AUC[۴۶]، منفی کاذب و مثبت کاذب استفاده شده است. نتایج نشان می‌دهد الگوریتم‌های مبتنی بر درخت و به طور خاص الگوریتم جنگل تصادفی بیشترین میزان صحت را در میان سایر الگوریتم‌های کلاس‌بندی به خود اختصاص می‌دهند. همچنین برای بهبود کارایی این الگوریتم از دو تکنیک standard boosting و bagging استفاده شده‌است.

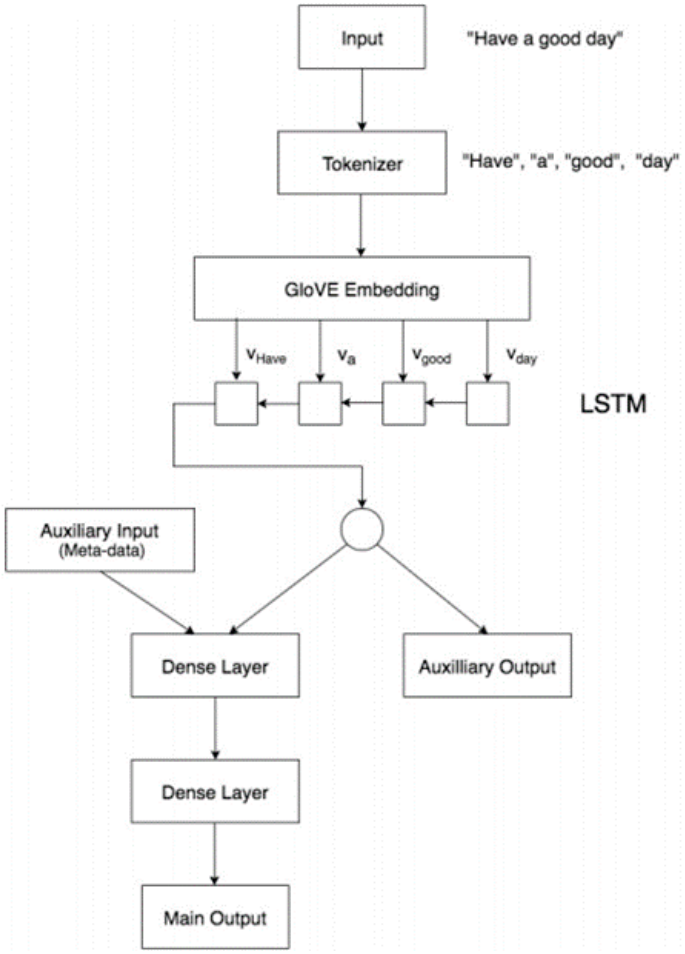
روش‌های ارائه شده در زمینه‌ی شناسایی کاربران جعلی شبکه‌های اجتماعی، عموماً بات‌ها را در سطح حساب کاربری شناسایی کرده‌اند (روش‌هایی از جمله یادگیری بانظارت یا بدون نظارت که از اطلاعات پروفایل کاربران، ساختار شبکه، الگوی زمانی فعالیت‌ها، حالات و احساسات و موارد مشابه استفاده می‌کنند). این روش‌ها گران هستند چرا که به مجموعه داده‌های عظیم برچسب‌گذاری شده برای آموزش مدل و نیز حجم قابل توجهی داده از سوی هر کاربر نیازمندند. اسنها کودوگونتا[۴۷] و امیلیو فرارا[۴۸] از یک معماری LSTM[۴۹] متنی جدید برای شناسایی بات‌ها در سطح توییت استفاده کرده‌اند که این معماری بر پایه‌ی شبکه‌های عصبی عمیق می‌باشد.[۱۳] برای تبدیل متن توییت‌ها به یک فرم سازگار با ورودیِ LSTM از مدل GLOVE[۵۰] بهره گرفته شده‌است که بدین منظور نیاز است توییت‌ها ابتدا یک بار پیش‌پردازش شده و توکن‌هایشان[۵۱] استخراج گردد.

رویکردهای یادگیری عمیق سنتی که با هدف کلاس‌بندی متن به کار می‌روند، تنها بر روی ویژگی‌های متنی مانند حروف یا n-گرام‌ها[۵۲] تکیه می‌کنند، اما نتایج قبلی نشان می‌دهند که متن توییت به تنهایی پیش‌بینی‌کننده‌ی قدرتمندی برای شناسایی بات‌ها نیست. در همین راستا در این مقاله نشان داده می‌شود که اطلاعات فراداده‌ای توییت‌ها (از جمله فراداده‌ی مرتبط با حساب کاربری، اطلاعات ساختار شبکه و الگوی زمانی فعالیت‌ها) - که به وسیله‌ی API توییتر به همراه متن توییت قابل دسترسی هستند و به اقدام فراتری در جمع‌آوری داده‌ها نیاز نیست - اگرچه فی‌نفسه پیشگوی ضعیفی برای شناسایی ماهیت یک حساب کاربری هستند، اما زمانی که در کنار LSTM به کار روند، نرخ خطا را تا ۲۰ درصد کاهش خواهند داد.

علاوه بر این، روش‌هایی براساس تکنیک SMOTE[۵۳] ارائه شده‌اند که یک مجموعه داده‌ی ساختگی عظیم و برچسب‌گذاری شده را از روی حجم کوچکی از داده‌ی حقیقی برچسب‌دار برای فرایند آموزش تولید می‌کنند و صحت شناسایی را به بالاترین حد ممکن نزدیک می‌سازند

[46] Area Under the ROC (receiver operating characteristic) Curve
[47] Sneha Kudugunta
[48] Emilio Ferrara
[49] Long Short-Term Memory
[50] GLObal VEctors for word representation
[51] tokens
[52] n-grams
[53] Synthetic Minority Oversampling TEchnique

(> 99% AUC). این تکنیک با دو تکنیک زیرنمونه‌برداری[۵۴] به نام‌های ENN[۵۵] و Tomek Links ترکیب می‌شود تا هر بایاس ایجاد شده به وسیله‌ی بیش‌نمونه‌برداری[۵۶] را خنثی کند. نتایج نشان می‌دهند استفاده از تکنیک ENN در مقایسه با Tomek Links عملکرد بهتر و موثرتری دارد.



شکل ۲-۳: معماری سیستم LSTM متنی[۱۳]

تمام این روش‌ها از کمترین تعداد ویژگی‌هایی (۱۶ تا) استفاده می‌کنند که بطور مستقیم از خود توییت و اطلاعات فراداده‌ای آن به دست می‌آید و تقریبا همگی آن‌ها بی‌نیاز نسبت به پیش‌پردازش هستند. ۱۰ ویژگی در سطح حساب کاربری استخراج می‌شوند که از میان آن‌ها می‌توان به تعداد دنبال‌کنندگان، تعداد دوستان، استفاده کردن / نکردن از قابلیت تصویر پس‌زمینه و تایید شده / نشده بودن توسط توییتر اشاره کرد. همچنین ۶ ویژگی در سطح توییت استخراج می‌شوند که شامل تعداد ری‌توییت‌ها، پاسخ[۵۷]ها، پسندیدن[۵۸]ها، هشتگ‌ها، آدرس‌های اینترنتی و

[54] undersampling
[55] Edited Nearest Neighbours
[56] oversampling
[57] reply
[58] like

منشن‌ها می‌باشد. کم بودن تعداد ویژگی‌ها موجب می‌شود که سرعت آموزش مدل افزایش یافته و مدل به دست آمده کمتر دچار بیش‌برازش[۵۹] شود. همچنین استفاده از ویژگی‌های محدودی که معنای مشخصی دارند سبب می‌شود مدل به دست آمده قابل تفسیر باشد. در انتها رویکردهای بررسی شده توسط معیارهای دقت، بازیابی، امتیاز F1، صحت و AUC/ROC ارزیابی شده‌اند.

اونور وارول و همکاران ادعا کرده‌اند که حساب‌های کاربری کنترل‌شده توسط نرم‌افزارها رفتارهایی از خود بروز می‌دهند که نمایانگر اهداف و روش عملکرد آن‌هاست، و این رفتارها می‌توانند توسط تکنیک‌های یادگیری ماشین بانظارت شناسایی شوند.[۱۸] بدین منظور یک ساختار برای استخراج یک مجموعه‌ی عظیم از ویژگی‌ها (۱۱۵۰ ویژگی) پیاده‌سازی شده است. این ویژگی‌ها بطورکلی در ۶ دسته قرار می‌گیرند:

۱. ویژگی‌های مبتنی بر کاربر؛ از جمله تعداد دوستان و دنبال‌کنندگان، تعداد توییت‌های منتشر شده توسط کاربر، مشخصات پروفایل و سایر تنظیمات مرتبط به آن

۲. ویژگی‌های مربوط به دوستان؛ شامل استفاده‌ی زبانی، زمان محلی، میزان محبوبیت و غیره

۳. ویژگی‌های شبکه؛ در این سیستم سه نوع شبکه ساخته می‌شود: ری‌توییت، منشن و هشتگ. در شبکه‌های جهت‌دار ری‌توییت و منشن هر گره نمایانگر یک کاربر است و هر یال نمایانگر یک مرتبه انتشار اطلاعات است که جهت آن به سوی کاربریست که ری‌توییت کرده و یا منشن شده است (همان جهت انتشار اطلاعات). در شبکه‌ی هشتگ‌ها هر گره نمایانگر یک هشتگ است و هر یال بدون جهت میان دو گره نشان‌دهنده‌ی این است که آن دو هشتگ بطور همزمان در یک توییت آورده شده‌اند. تمام شبکه‌ها وزن‌دار هستند و وزن در آن‌ها بر اساس تعداد تکرار فعل و انفعالات یا همزمانی‌ها تعریف می‌شود.

۴. ویژگی‌های زمانی؛ از جمله نرخ متوسط تولید محتوا در بازه‌های زمانی مختلف و توزیع فواصل زمانی بین وقایع (فاصله‌ی زمانی میان توییت، ری‌توییت و منشن‌های متوالی)

۵. ویژگی‌های زبانی و متنی؛ عموماً در پیام‌های فریبنده از زبان‌های غیررسمی و جملات کوتاه استفاده می‌شود. در این سیستم ویژگی‌های مرتبط با کیفیت توییت‌ها به کار گرفته نشده‌اند اما اطلاعات آماری مربوط به طول و بی‌نظمی[۶۰] متن توییت‌ها جمع‌آوری شده و مورد استفاده قرار گرفته است. علاوه بر این، برخی ویژگی‌های زبانی با اعمال تکنیک برچسب‌گذاری POS[۶۱] استخراج شده‌اند.

۶. ویژگی‌های مرتبط با حالات و احساسات؛ تحلیل احساسات یک ابزار قوی برای توصیف هیجاناتی است که به وسیله‌ی یک تکه متن منتقل می‌شود، و بطور گسترده‌تر به گرایش

[59] overfitting
[60] entropy
[61] Part-Of-Speech

و حس و حال یک مکالمه می‌پردازد. در این سیستم ویژگی‌هایی نظیر میزان خوشحالی، سطح تحریک و میزان و نوع اثرگذاری (مثبت و منفی) دریافت شده از محتوای هر توییت و همچنین تعداد و بی‌نظمی شکلک[۶۲]های مثبت و منفی به کار رفته در هر توییت استخراج و به کار گرفته می‌شوند.

نتایج نشان می‌دهد ویژگی‌های مبتنی بر کاربر و ویژگی‌های متنی باارزش‌ترین منابع داده در شناسایی بات‌های ساده می‌باشند.

برای آموزش مدل از یک مجموعه حساب‌های کاربری برچسب‌گذاری‌شده به روش honeypot و همچنین یک مجموعه داده‌ی برچسب‌گذاری‌شده به روش دستی[۶۳] به همراه آخرین توییت‌های عمومی منتشر شده توسط آنان استفاده شده است. این مجموعه حاوی بات‌هایی با میزان پیچیدگی متفاوت است. در این سیستم تمرکز تنها بر روی کاربران انگلیسی‌زبان بوده است چرا که این کاربران بزرگترین گروه را در میان کاربران توییتر تشکیل می‌دهند. دسته‌بند مورد استفاده به ازای هر حساب کاربری در بخش تست یک عدد[۶۴] در بازه‌ی ۰ تا ۱ بازمی‌گرداند که نمایانگر احتمال جعلی بودن آن است.

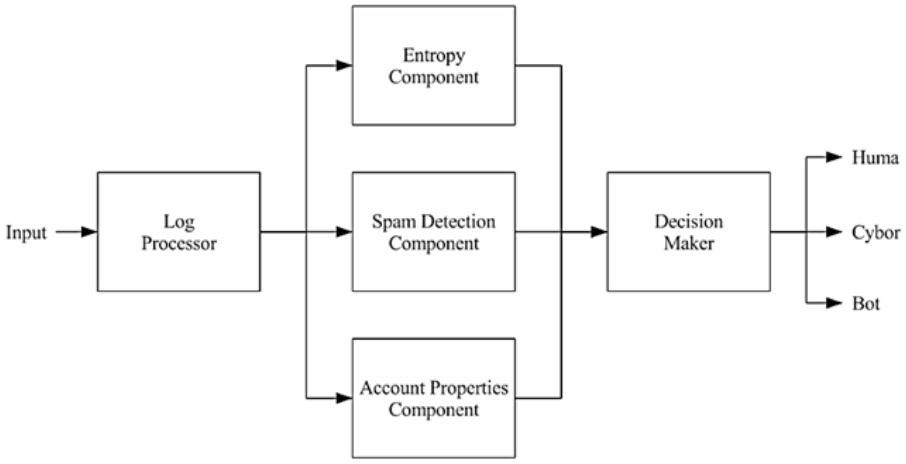
دقت هر مدل با اندازه‌گیری معیار AUC و به کمک روش اعتبارسنجی متقابل ۵ لایه و سپس محاسبه‌ی میانگین مقادیر AUC در هر لایه به دست می‌آید. در میان الگوریتم‌های بررسی شده (جنگل تصادفی، AdaBoost، رگرسیون لجستیک و درخت تصمیم) الگوریتم جنگل تصادفی بالاترین میزان کارایی را داشته و در ادامه نیز همین روش دسته‌بندی به کار گرفته شده است.

برای ارزیابی کارایی سیستم، حساب‌های کاربری بخش تست بر اساس میزان bot-score شان مرتب شده و از هر دهک ۳۰۰ نمونه بطور تصادفی انتخاب شده است. در مرحله‌ی بعد این ۳۰۰۰ نمونه بطور دستی و توسط نیروهای انسانی بررسی و ارزیابی شده‌اند. در انتها با آموزش دسته‌بند از روی ادغام دو مجموعه‌داده‌ی مذکور، دقت بالای 0.94 AUC در تشخیص هم بات‌های ساده و هم پیچیده حاصل شده است.

زی چو[۶۵] و همکاران رفتار کاربران انسانی، بات‌ها و سایبورگ‌ها[۶۶] (انسان با همکاری بات یا بات با همکاری انسان) را مورد مطالعه قرار داده‌اند.[۷] با جستجو در توییتر، داده‌های یک ماه آن به دو روش الگوریتم DFS و استفاده از رابط نرم‌افزاری جدول زمانی[۶۷] ارائه شده توسط خود توییتر جمع‌آوری شده است (بیش از ۵۰۰,۰۰۰ حساب کاربری و ۴۰ میلیون توییت). از میان داده‌های جمع‌آوری شده، یک مجموعه داده‌ی مرجع[۶۸] متشکل از ۲۰۰۰ حساب کاربری از هر دسته (انسان، بات، سایبورگ) بطور تصادفی انتخاب شده و به کمک نیروهای انسانی با روش دستی

[62] emoji
[63] manually annotated
[64] bot-score
[65] Zi Chu
[66] cyborgs
[67] timeline API
[68] ground-truth

برچسب‌گذاری شده‌است. (در این مرحله متون شناسایی شده به عنوان هرزنامه[۶۹] همگی در یک مجموعه ذخیره شده و در الگوریتم کلاس‌بندی از آن‌ها استفاده به عمل خواهد آمد.) بر اساس این داده‌ها ویژگی‌های تفکیک‌کننده‌ی انسان، بات و سایبورگ شناسایی شده‌اند. سپس بر اساس این ویژگی‌ها یک سیستم کلاس‌بندی خودکار طراحی شده‌است که از ۴ بخش تشکیل می‌شود:



شکل ۲-۴: نمای کلی سیستم کلاس‌بندی[۷]

- بخش بی‌نظمی: الگوهای زمانی منظم در ارسال توییت را بررسی می‌کند. به کمک معیار بی‌نظمی این حقیقت کشف شده‌است که کاربران انسانی رفتار زمانی پیچیده‌ای دارند (بی‌نظمی زیاد) درحالی‌که بات‌ها و سایبورگ‌ها عموماً با رفتار زمانی منظم و قاعده‌مند شناخته می‌شوند (بی‌نظمی کم).

- بخش تشخیص هرزنامه: هرزنامه یک شاخص مناسب برای شناسایی حساب‌های کاربری خودکار است؛ اکثر هرزنامه‌ها به وسیله‌ی بات‌ها تولید می‌شوند و تعداد بسیار اندکی بطور دستی توسط کاربران انسانی منتشر می‌شوند. این جزء با کنترل الگوی متنی هر توییت، هرزنامه بودن یا نبودن آن را بررسی می‌کند. بدین منظور از نوعی روش کلاس‌بندی بیز مناسب داده‌های متنی به نام روش OSB[۷۰] استفاده می‌شود.

- بخش ویژگی‌های مربوط به حساب کاربری: به دنبال یافتن مقادیر غیرعادی در ویژگی‌های مرتبط با حساب کاربری است؛ ویژگی‌هایی از جمله نسبت تعداد توییت‌های شامل آدرس اینترنتی به تعداد کل توییت‌های یک کاربر، امنیت پیوندهای به کار رفته در توییت‌ها، تاریخ ثبت‌نام و ایجاد حساب کاربری، نسبت تعداد دنبال‌کنندگان به تعداد دوستان و نسبت تعداد هشتگ‌ها و منشن‌های به کار رفته به تعداد کل توییت‌های کاربر.

- بخش تصمیم‌گیرنده: ویژگی‌های شناسایی‌شده را ترکیب و جمع‌بندی می‌کند و بر اساس

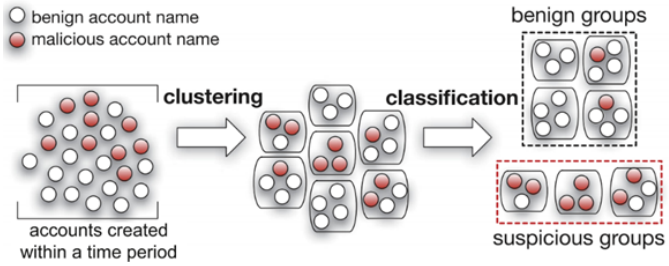
[69] spam

[70] Orthogonal Sparse Bigram

الگوریتم جنگل تصادفی (پیاده‌سازی شده به کمک جعبه‌ابزار Weka) تصمیم می‌گیرد که هر حساب کاربری یک انسان، یک بات و یا یک سایبورگ است.

در انتها صحت سیستم کلاس‌بندی به کمک روش اعتبارسنجی متقابل ۱۰ لایه و داده‌های برچسب‌گذاری شده‌ی مرجع ارزیابی شده است. نتایج ارزیابی نشان می‌دهد نرخ مثبت صحیح[۷۱] برای سه کلاس کاربران انسانی، بات‌ها و سایبورگ‌ها به ترتیب برابر با ۹۸٫۶٪، ۹۷٫۶٪ و ۹۱٫۷٪ و میانگین‌شان برابر با ۹۶٪ است.

روش‌های مرسوم، حساب‌های کاربری جعلی را پس از ارسال حداقل یک توییت مخرب از سوی آنان شناسایی می‌کنند؛ به عبارت دیگر در این روش‌ها پیش از اجرای الگوریتم‌های شناسایی، زمان قابل توجهی صرف جمع‌آوری اطلاعات می‌شود. سنگو لی[۷۲] و جانگ کیم[۷۳] یک رویکرد جدید معرفی کرده‌اند که حساب‌های کاربری مخرب بالقوه را در زمان ایجادشان شناسایی می‌کند و منتظر شروع رفتارهای مخرب نمی‌ماند.[۱۵] در این رویکرد حساب‌های کاربری بر اساس میزان شباهت نام کاربری‌شان (ویژگی‌های مشابه مبتنی بر نام کاربری) خوشه‌بندی شده و سپس بر اساس ویژگی‌های هر خوشه در دو کلاس بی‌خطر و مشکوک دسته‌بندی می‌شوند. این روش می‌تواند به عنوان یک سیستم زنگ خطر اولیه برای نظارت بر روی حساب‌های کاربری مخرب بالقوه مورد استفاده قرار گیرد.



شکل ۲-۵: نمای کلی سیستم[۱۵]

این رویکرد از این نکته بهره گرفته است که میان نام‌های کاربری تولید شده بصورت الگوریتمی و تولید شده به دست انسان عموماً تفاوت‌های چشمگیری وجود دارد؛ به این دلیل که تولید نام‌های کاربری بصورت الگوریتمی بطوری که به نام‌های انتخابی توسط انسان‌ها شبیه باشند و پیش‌تر توسط حساب کاربری دیگری انتخاب و اشغال نشده باشند کار ساده‌ای نیست.

برای خوشه‌بندی حساب‌های کاربری، از یک الگوریتم خوشه‌بندی سلسله‌مراتبی جمع‌کننده[۷۴] استفاده می‌شود. این الگوریتم در ابتدا n خوشه می‌سازد که هر کدام حاوی تنها یک موجودیت (نام کاربری) هستند و سپس بطور پیوسته خوشه‌های مشابه را با هم ادغام می‌کند تا زمانی که

[71] True Positive rate
[72] Sangho Lee
[73] Jong Kim
[74] Agglomerative hierarchical clustering algorithm

شرط خاتمه فراهم گردد. بدین منظور تابع $d_{max}(C_i, C_j) = max_{n_i \in C_i, n_j \in C_j} dist(n_i, n_j)$ برای محاسبه‌ی فاصله‌ی میان دو خوشه تعریف می‌گردد. در هر مرحله از اجرای الگوریتم، اگر مقدار $d_{max}(C_i, C_j)$ از یک حد آستانه‌ی تعریف شده کمتر باشد، دو خوشه‌ی C_i و C_j با هم ادغام خواهند شد. الگوریتم زمانی خاتمه می‌یابد که فاصله‌ی میان تمام خوشه‌ها بزرگتر یا مساوی مقدار آستانه شود. برای محاسبه‌ی $dist(n_i, n_j)$ احتمال تولید هر نام کاربری (n_i, n_j) به کمک یک زنجیره‌ی مارکوف[۷۵] محاسبه می‌گردد که این احتمال با نماد $\ell(\ell(n_i|m), \ell(n_j|m))$ نمایش داده می‌شود. زنجیره‌ی مارکوف با استفاده از رشته‌های دوحرفی استخراج شده از نام‌های کاربری معتبر ساخته می‌شود. برای کلاس‌بندی خوشه‌ها، یک کلاس‌بند SVM به کمک کتابخانه‌ی LIBSVM بر اساس ویژگی‌های مبتنی بر اسم حساب‌های کاربری تعلیق شده آموزش داده می‌شود.

مجموعه داده‌ی مورد استفاده شامل تمام حساب‌های کاربری ایجاد شده بین آوریل ۲۰۱۱ و آگوست همان سال می‌باشد. که در این مدت حدود ۷.۴ میلیون حساب کاربری ایجاد شده‌اند. برای برچسب‌گذاری این داده‌ها وضعیت فعال / تعلیق شده آن‌ها در مارچ ۲۰۱۲ بررسی شده‌است. همچنین نام حساب‌های کاربری مورد تایید توییتر به عنوان مرجعی برای نام‌های کاربری تولید شده توسط انسان در نظر گرفته شده‌اند.

برای ارزیابی سیستم از دو معیار نرخ منفی کاذب و نرخ مثبت کاذب و روش اعتبارسنجی متقابل ۵ لایه استفاده شده‌است. هدف کاهش نرخ منفی کاذب با حفظ نرخ مثبت کاذب نسبتاً پایین است که با تنظیم نسبت بین نمونه‌های مثبت و منفی داده‌های آموزشی حاصل می‌شود. در نهایت باید به این نکته توجه داشت که این رویکرد نمی‌تواند یک سیستم شناسایی کامل به حساب آید، چرا که این روش اطلاعات کافی (جزئیات پروفایل، اطلاعات مربوط به روابط میان حساب‌های کاربری، پیام‌های انتشار یافته) برای قضاوت صحیح را در اختیار ندارد. از سوی دیگر دقت نسبتاً پایین این سیستم قابل قبول است؛ چرا که هدف در این‌جا شناسایی دقیق کاربران مخرب نیست، بلکه هدف فیلتر کردن حساب‌های کاربری مشکوک برای بررسی‌های بیشتر بعدی است. به همین دلیل نیاز است تا بررسی‌های فراتری (تحقیقات انسانی، کدهای کپچا[۷۶]، ...) بر روی نتایج به دست آمده از این روش صورت گیرد.

در سال ۲۰۲۱ کوسم کوماری بهارتی[۷۷] و شیوانجلی پاندی[۷۸]یک مدل دو فازی (فاز انتخاب ویژگی[۷۹] و فاز طبقه‌بندی) را برای تشخیص حساب‌های کاربری جعلی پیشنهاد دادند.[۳] ابتدا از تکنیک‌های انتخاب ویژگیِ بهره‌ی اطلاعاتی[۸۰]، همبستگی[۸۱] و حداکثر ارتباط - حداقل افزونگی[۸۲]

[75] Markov chain
[76] captcha
[77] Kusum Kumari Bharti
[78] Shivanjali Pandey
[79] feature selection
[80] information gain
[81] correlation
[82] Maximum Relevance — Minimum Redundancy (MRMR)

برای استخراج زیرمجموعه‌ی مفیدی از ویژگی‌ها که خصوصیات پروفایل کاربران را تعیین می‌کنند استفاده شد. سپس از آن‌جا که کارایی الگوریتم طبقه‌بندی تا حد زیادی متکی بر چگونگی انتخاب پارامترها می‌باشد، الگوریتم رگرسیون لجستیک همراه با بهینه‌سازی ازدحام ذرات[۸۳] برای طبقه‌بندی مؤثر حساب‌های حقیقی و جعلی به کار گرفته شد. علاوه‌براین، مقداردهی اولیه مبتنی بر مخالفت[۸۴] با بهینه‌سازی ازدحام ذرات ترکیب شد تا کاوش فضای جستجو با مجموعه‌ی مناسبی از راه‌حل‌ها آغاز شود. در انتها نتایج به‌دست‌آمده در قیاس با رویکردهای پیشین از نظر آماری معنادارتر است و دقت مدل ۹۶٫۲٪ ثبت شده است.

در سال ۲۰۲۱ احمد حمصی[۸۵] و همکاران با هدف بررسی تأثیر همبستگی در کنار الگوریتم‌های طبقه‌بندی برای شناسایی حساب‌های کاربری جعلی، چهار الگوریتم یادگیری ماشین J48، جنگل تصادفی، بیز ساده و k-نزدیک‌ترین همسایه[۸۶] و دو تکنیک کاهش داده‌ی تحلیل مؤلفه‌های اصلی[۸۷] و همبستگی را بر روی مجموعه داده‌ی MIB توییتر [۱] به‌کارگرفتند.[۱۱] بدین منظور از جعبه‌ابزار Weka استفاده شده است. نتایج این پژوهش نشان می‌دهد ترکیب همبستگی با الگوریتم جنگل تصادفی بالاترین میزان دقت (حدود ۹۸٫۶٪) و ترکیب آن با الگوریتم بیز ساده پایین‌ترین میزان دقت (۸۲٫۱٪) دارد.

[83] Particle Swarm Optimization (PSO)

[84] Opposition-Based Initialization

[85] Ahmad Homsi

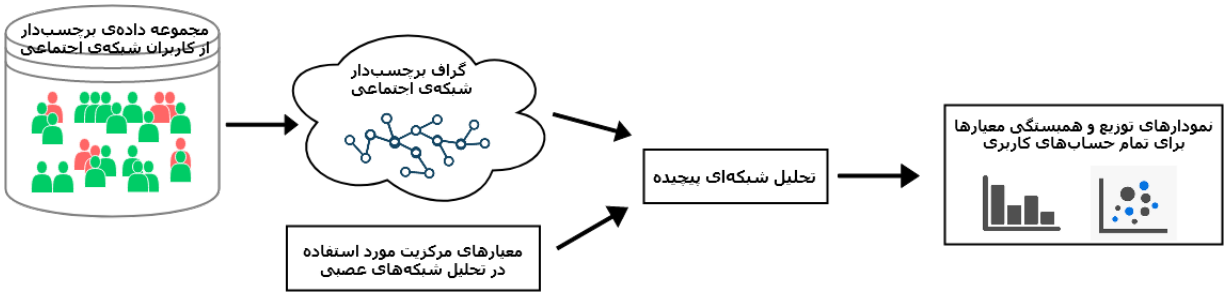
[86] k-nearest neighbors (KNN)

[87] Principal Component Analysis (PCA)

فصل سوم

توصیف کار پیشنهادی

در هر شبکه‌ی اجتماعی کاربران در بستر چندین گراف به فعالیت می‌پردازند. در هر یک از این گراف‌ها، هر گره نمایانگر یک حساب کاربری است و هر یال بیانگر یک رابطه بین دو حساب کاربری می‌باشد. تعدادی از گره‌های این گراف را کاربران جعلی و بقیه را کاربران حقیقی تشکیل می‌دهند. به عنوان مثال، در شبکه‌های اجتماعی توییتر، اینستاگرام[۱]، لینکدین[۲] و یوتیوب[۳] گراف‌های دنبال‌کننده-دنبال‌شونده، پسندیدن، دیدگاه و منشن وجود دارند. به طور مجزا در توییتر گراف‌های پاسخ‌دهی و ری‌توییت، در اینستاگرام گراف‌های به‌اشتراک‌گذاری مجدد[۴] و برچسب‌زنی[۵]، در لینکدین گراف به‌اشتراک‌گذاری مجدد و در یوتیوب گراف نپسندیدن[۶] نیز وجود دارند. بدین ترتیب چندین شبکه‌ی پیچیده تشکیل می‌گردند که می‌توان آن‌ها را به کمک مجموعه‌ای از ابزارها و معیارها تحلیل نمود. در این پروژه یک رویکرد تحلیل شبکه پیچیده‌ای بر روی کاربران حقیقی و جعلی شبکه‌ی اجتماعی توییتر به کار گرفته شده است. پس از تشکیل گراف، بررسی خواهد شد که گره‌های حقیقی در برابر معیارهایی که در تحلیل شبکه‌های عصبی مورد استفاده قرار می‌گیرند، چه رفتاری از خود نشان می‌دهند و بطور مشابه گره‌های جعلی در برابر آن معیارها چه رفتاری از خود نشان می‌دهند.



شکل ۳-۱: نمودار متنی سطح صفر

۳-۱ معیارهای مورد استفاده در تحلیل شبکه‌های عصبی

معیارهای مرکزیت[۷] که در تحلیل شبکه‌های عصبی به کار می‌روند عبارتند از:

۱. درجه ورودی[۸]: تعداد یال‌های جهت‌داری که به یک گره از گراف وارد می‌شوند. به بیان

[1] Instagram
[2] Linkedin
[3] Youtube
[4] resharing
[5] tagging
[6] dislike
[7] centrality
[8] in-degree

دقیق، درجه ورودی برای هر گره‌ی i برابر است با:

$$C(i) = |\{e_{ji} | i, j \in V, e_{ji} \in E\}|$$

۲. درجه خروجی[9]: تعداد یال‌های جهت‌داری که از یک گره از گراف خارج می‌شوند. به بیان دقیق، درجه خروجی برای هر گره‌ی i برابر است با:

$$C(i) = |\{e_{ij} | i, j \in V, e_{ij} \in E\}|$$

۳. میانگی[10]: یک روش برای تشخیص میزان تأثیری است که یک گره بر جریان اطلاعاتی در گراف دارد و معمولاً برای پیدا کردن گره‌هایی که به عنوان پل ارتباطی از یک قسمت از گراف به قسمت دیگر عمل می‌کنند، استفاده می‌شود. در یک شبکه‌ی اجتماعی، یک گره با میانگی بالاتر، کنترل بیشتری بر روی شبکه خواهد داشت، زیرا اطلاعات بیشتری از آن عبور می‌کنند. به بیان دقیق، مقدار میانگی برای هر گره‌ی v برابر است با:

$$C_B(v) = \sum_{s,t \in V} \frac{\sigma(s, t|v)}{\sigma(s, t)}$$

به‌طوری‌که V مجموعه‌ی گره‌ها، $\sigma(s, t)$ تعداد کل کوتاه‌ترین مسیرها از گره‌ی s به t و $\sigma(s, t|v)$ تعداد آن کوتاه‌ترین مسیرهایی از گره‌ی s به t است که از گره‌ی v نیز عبور می‌کنند. در یک گراف وزن‌دار، کوتاه‌ترین مسیر میان دو گره مسیری است که مجموع وزن یال‌های تشکیل‌دهنده‌اش کمینه باشد.

۴. نزدیکی[11]: در یک گراف همبند، از طریق معکوس کردن حاصل‌جمع کوتاه‌ترین مسیر میان گره‌ی مورد نظر و تمام گره‌های دیگر گراف محاسبه می‌شود. بطور کلی در هر گراف با بیش از یک مؤلفه‌ی همبندی، مقدار نزدیکی برای هر گره‌ی u برابر است با:

$$C(u) = \frac{n-1}{N-1} \frac{n-1}{\sum_{v=1}^{n-1} d(v, u)}$$

به‌طوری‌که n تعداد گره‌هایی است که از طریق حداقل یک مسیر به گره‌ی u متصل هستند (تعداد گره‌ها در مؤلفه‌ی همبندی u)، N تعداد کل گره‌های گراف است و $d(v, u)$ طول

[9] out-degree
[10] betweenness
[11] closeness

کوتاه‌ترین مسیر از گره‌ی v به u می‌باشد.

۵. هارمونیک[۱۲]: این معیار یک نسخه‌ی تغییریافته از مرکزیت نزدیکی می باشد که برای حل مشکلی که فرمول اولیه در رابطه با گراف‌های ناهمبند دارد ، ابداع شده است. به بیان دقیق، در این الگوریتم برای هر گره مجموع معکوس کوتاه‌ترین مسیر از تمام گره‌های دیگر به آن گره محاسبه می‌شود. در صورتی که یک گره از طریق هیچ مسیری به گره‌ی موردنظر متصل نشده باشد، فاصله‌اش تا آن گره برابر با بی‌نهایت و معکوس فاصله‌اش برابر با صفر می‌گردد و بدین‌ترتیب بطور خودکار از فرمول حذف خواهد شد. به بیان دقیق، مقدار هارمونیک برای هر گره‌ی u برابر است با:

$$C(u) = \sum_{v \neq u} \frac{۱}{d(v,u)}$$

به‌طوری‌که $d(v,u)$ طول کوتاه‌ترین مسیر از گره‌ی v به u می‌باشد.

۶. بردار ویژه[۱۳]: مرکزیت یک گره را براساس مرکزیت گره‌های همسایه‌اش محاسبه می‌کند. پیوندهایی که از گره‌های با امتیاز بالا سرچشمه می‌گیرند، بیشتر به امتیاز یک گره کمک می‌کنند تا پیوندهای نشأت گرفته از گره‌های با امتیاز پایین. به بیان دقیق، مقدار مرکزیت بردار ویژه برای گره‌ی i از گراف G برابر است با i اُمین درایه از بردار x که به صورت زیر تعریف می‌شود:

$$Ax = \lambda x$$

به‌طوری‌که A ماتریس مجاورت گراف G و λ بزرگترین مقدار ویژه‌ی آن است و تنها یک بردار x وجود دارد که در معادله صدق می‌کند و تمام درایه‌های آن مثبت است.

۷. کتز[۱۴]: معیاری است که برای سنجش درجه‌ی نسبی تأثیرگذاری یک گره در شبکه به کار می‌رود. کتز تعمیمی از مرکزیت درجه است؛ به‌طوری‌که مرکزیت درجه تعداد همسایگان مستقیم هر گره را درنظر می‌گیرد، درحالی‌که کتز تعداد کل گره‌هایی که بطور مستقیم یا غیرمستقیم (از طریق یک مسیر) با گره‌ی مورد نظر در ارتباط هستند را درنظر می‌گیرد و البته برای گره‌های دورتر جریمه‌ای لحاظ می‌کند. به بیان دقیق، مقدار مرکزیت کتز برای هر گره‌ی i برابر است با:

$$x_i = \alpha \sum_j A_{ij} x_j + \beta$$

به‌طوری‌که A ماتریس مجاورت گراف G با مقادیر ویژه‌ی λ است، β مرکزیت اولیه را

[12] harmonic

[13] eigenvector

[14] katz

کنترل می‌کند و قادر است به همسایگان مستقیم i وزن بیشتری اختصاص دهد و α عامل تضعیف[15] تأثیر گره‌های دورتر است که باید در شرط $\alpha < \frac{1}{\lambda_{max}}$ صدق کند. در صورتی که $\alpha = \frac{1}{\lambda_{max}}$ و $\beta = 0$ باشد، مقدار مرکزیت کتز با مقدار مرکزیت بردار ویژه برابر خواهد شد.

۸. رتبه صفحه[16]: الگوریتمی است که رتبه‌ی گره‌های گراف را بر اساس ساختار پیوند (یال) های ورودی محاسبه می‌نماید. در این الگوریتم به هر گره امتیازی تعلق می‌گیرد که برابر با مجموع امتیازات یال‌های ورودی به آن گره است. از سوی دیگر امتیاز هر گره بطور مساوی میان تمام یال‌های خروجی از آن گره تقسیم می‌شود. به بیان دقیق مقدار رتبه صفحه برای هر گره‌ی i برابر است با:

$$x_i = \alpha \sum_j A_{ji} \frac{x_j}{L(j)} + \frac{1-\alpha}{N}$$

به‌طوری‌که A ماتریس مجاورت گراف G و $L(j)$ برابر با درجه خروجی گره‌ی i می‌باشد ($L(j) = \sum_j A_{ji}$). این الگوریتم در ابتدا برای رتبه‌بندی صفحات وب در نتایج جست‌وجو طراحی شده بود. رتبه صفحه یک نسخه‌ی تغییریافته از مرکزیت بردار ویژه می‌باشد.

۹. ضریب خوشه‌بندی[17]: این معیار مشخص می‌کند که در همسایگی هر گره چگالی محلی به چه صورت می‌باشد. به عبارت دیگر، این معیار به هر گره مقداری نسبت می‌دهد که متناسب است با تعداد گره‌هایی که تمایل به قرار گرفتن در یک خوشه در کنار آن گره دارند. ضریب خوشه‌بندی میزان مشابهت همسایگان هر گره به یک زیرگراف کامل را می‌سنجد. به بیان دقیق، در هر گراف جهت‌دار بدون وزن، مقدار ضریب خوشه‌بندی برای هر گره‌ی i برابر است با:

$$C(i) = \frac{|\{e_{jk} | j,k \in N_i, e_{jk} \in E\}|}{deg(i)(deg(i)-1)}$$

به‌طوری‌که e_{jk} بیانگر یالی است که گره‌ی j را به k متصل می‌کند، N_i مجموعه‌ی همسایگان گره‌ی i است و $deg(i)$ مجموع درجه ورودی و خروجی گره‌ی i می‌باشد. همچنین در هر گراف جهت‌دار وزن‌دار، مقدار ضریب خوشه‌بندی برای هر گره‌ی i برابر است با:

$$C(i) = \frac{\sum_{j,k \in N_i} (\hat{w}_{ij} \hat{w}_{ik} \hat{w}_{jk})^{\frac{1}{3}}}{deg(i)(deg(i)-1)}$$

[15] attenuation factor

[16] pagerank

[17] clustering coefficient

به‌طوری‌که $\hat{w}_{ij} = \frac{w_{ij}}{max(w)}$.

۱۰. میانگین کوتاه‌ترین طول مسیر[۱۸]: میانگین فاصله‌ی هر گره از گره های دیگر شبکه را تعیین می‌کند. به بیان دقیق، میانگین کوتاه‌ترین طول مسیر برای هر گره‌ی u برابر است با:

$$C(u) = \frac{\sum_{v=1}^{n-1} d(v,u)}{N-1}$$

به‌طوری‌که n تعداد گره‌ها در مؤلفه‌ی همبندی گره‌ی u، N تعداد کل گره‌های گراف و $d(v,u)$ طول کوتاه‌ترین مسیر میان دو گره‌ی v و u می‌باشد.

۱۱. میانگین درجات همسایگی[۱۹]: سنجشی از میزان وابستگی میان درجات گره‌های همسایه نسبت به هم می‌باشد. به بیان دقیق، میانگین درجات همسایگی برای هر گره‌ی i در هر گراف وزن‌دار برابر است با:

$$C(i) = \frac{1}{deg(i)} \sum_{j \in N_i} w_{ij} deg(j)$$

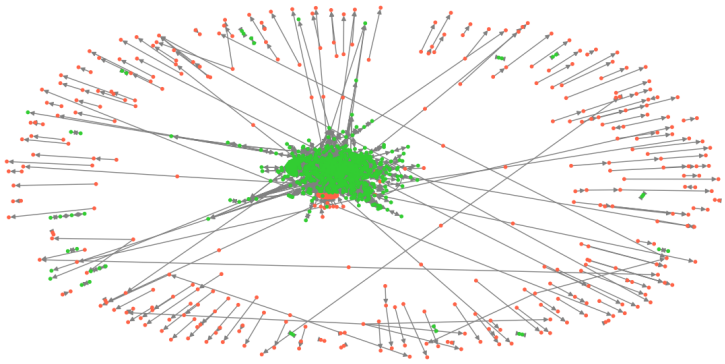
۱۲. دستیابی محلی[۲۰]: در یک گراف جهت‌دار، نسبتی از قابل‌دسترسی بودن گره‌های دیگر از گره‌ی مربوطه می‌باشد.

۳-۲ نحوه‌ی پیاده‌سازی

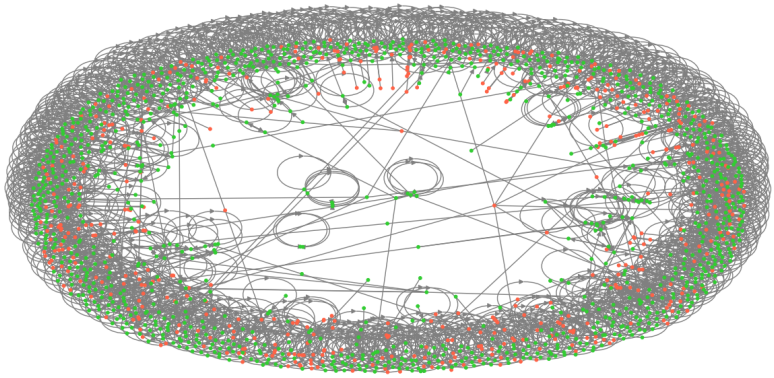
پیاده‌سازی این پروژه به زبان پایتون انجام شده و از کتابخانه‌ی NetworkX[۲] برای ساخت گراف و محاسبه‌ی برخی معیارها (میانگی، رتبه صفحه، کتز، ضریب خوشه‌بندی و دستیابی محلی) استفاده شده است. ابتدا به کمک اطلاعات پروفایل و توییت‌های حساب‌های کاربری موجود در مجموعه داده، گراف شبکه‌های دنبال‌کننده-دوست و دیدگاه را تشکیل داده‌ایم؛

- در گراف دنبال‌کننده-دوست : هر یال از A به B نشان می‌دهد که حساب کاربری A حساب کاربری B را دنبال می‌کند. (گراف بدون وزن)
- در گراف دیدگاه : هر یال از A به B نشان می‌دهد که حساب کاربری B حداقل یکبار زیر یکی از توییت‌های حساب کاربری A دیدگاه گذاشته است. (تعداد دفعات تکرار، وزن یال را مشخص می‌کند و جهت هر یال در جهت انتشار اطلاعات است.)

[18] average shortest path length
[19] average neighbor degree
[20] local reaching



شکل ۳-۲: گراف دنبال‌کننده-دوست. نقاط سبز نمایانگر گره‌های حقیقی و نقاط قرمز نمایانگر گره‌های جعلی می‌باشند. گره‌های با درجه‌ی صفر حذف شده‌اند.



شکل ۳-۳: گراف دیدگاه. نقاط سبز نمایانگر گره‌های حقیقی و نقاط قرمز نمایانگر گره‌های جعلی می‌باشند. گره‌های با درجه‌ی صفر حذف شده‌اند.

از آن‌جا که حجم داده‌ی مورد استفاده به نسبت زیاد است، برای جلوگیری از تکرار فرایند ساخت گراف‌ها در هر بار اجرای برنامه، از تابع write_gexf موجود در کتابخانه‌ی networkx برای ذخیره‌ی گراف‌ها داخل فایل به فرمت GEXF استفاده شده است. در بعضی از معیارهای مورد استفاده (میانگی، نزدیکی، هارمونیک و میانگین کوتاه‌ترین طول مسیر) وزن یال به معنای فاصله‌ی میان دو گره است، درحالیکه همانطور که پیش‌تر گفته شد، نحوه‌ی محاسبه‌ی وزن یال در گراف دیدگاه به گونه‌ای است که بیانگر استحکام و قدرت اتصال میان دو گره است. در نتیجه در محاسبه‌ی معیارهای مذکور باید از معکوس وزن یال‌ها استفاده شود؛ به همین دلیل در ساخت گراف دیدگاه از همان ابتدا یک گراف با وزن‌های معکوس نیز ایجاد شده است.

کد کامل پیاده‌سازی پروژه در آدرس زیر قابل مشاهده است:

https://github.com/S-Asghari/Complex-Network-Analysis-of-Twitter-Accounts

۳-۲-۱ چگونگی حذف داده‌های پرت

برای بالا بردن وضوح هر نمودار و نمایش دقیق‌تر مقادیر، داده‌های پرت حذف شده‌اند. بدین منظور الگوریتم removing_outliers برای حذف داده‌های پرت در نمودارهای همبستگی، پیاده‌سازی شده است که عملکرد آن به شرح زیر است:

می‌دانیم در یک مجموعه‌داده‌ی تک‌بُعدی (مانند نمودار توزیع یک معیار)، اولین داده‌ای که به عنوان داده‌ی پرت حذف می‌شود، داده‌ای‌ست که بیشترین فاصله را از نقطه‌ی مرکزی (میانه) داشته باشد. اما از آن‌جا که نمودار همبستگی یک نمودار دوبُعدی (حاوی دو مؤلفه‌ی x و y) است، موقعیت نقطه‌ی مرکزی باید به طریق دیگری محاسبه شود و فرمول فاصله نیز متناسب با آن تغییر کند. در این الگوریتم ابتدا میانه‌ی مقادیر x (x_m) و میانه‌ی مقادیر y (y_m) محاسبه شده است. سپس از فرمول فاصله‌ی اقلیدسی[۲۱] استفاده شده است:

$$d = \sqrt{(x - x_m)^۲ + (y - y_m)^۲}$$

درعین‌حال باید به این نکته توجه داشت که مقادیر دو معیار x و y لزوماً دامنه تغییرات یکسانی ندارند. به همین دلیل باید یک مرحله نرمال‌سازی داخل فرمول فاصله لحاظ گردد. از آن‌جا که مقدار میانه در بسیاری از معیارها برابر با صفر است، نمی‌توان از این شاخص برای نرمال‌سازی استفاده نمود. بدین منظور میانگین چارک اول و سوم برای هر دو معیار x و y محاسبه می‌شود که به ترتیب با نمادهای x_normalizer و y_normalizer نمایش داده می‌شوند. فلذا فرمول محاسبه‌ی d به فرم زیر است:

$$d = \sqrt{((x - x_m)/x\_normalizer)^۲ + ((y - y_m)/y\_normalizer)^۲}$$

به ازای هر داده داخل نمودار، این مقدار محاسبه می‌شود و داده‌ای که بیشترین مقدار d را داشته باشد، اولین داده‌ای‌ست که باید از نمودار حذف گردد.

[21] Euclidean distance

فصل چهارم

نتایج تجربی

از میان دوازده معیار مورد استفاده، الگوریتم محاسبه‌ی معیار درجه ورودی، درجه خروجی، میانگین درجات همسایگی، میانگین کوتاه‌ترین طول مسیر، نزدیکی، هارمونیک و بردار ویژه پیاده‌سازی شده‌اند. در ادامه مجموعه‌داده‌ی مورد استفاده توصیف می‌شود، نتایج تجربی الگوریتم‌های به‌کاررفته نمایش داده می‌شوند و بطور دقیق مورد تجزیه و تحلیل قرار خواهند گرفت.

۴-۱ مجموعه داده‌ی مورد استفاده

در این پروژه از مجموعه داده‌ی MIB استفاده شده‌است که شامل دو سری مجزای Cresci-2015[۸] و Cresci-2017[۹] می‌باشد.[۱]

مجموعه‌ی Cresci-2015 شامل:

- جزئیات حساب‌های کاربری (شامل: شناسه، نام، تعداد دنبال‌کنندگان، تعداد دوستان، زمان ایجاد پروفایل، زبان، موقعیت مکانی، عکس پروفایل و غیره)
- توییت‌ها (شامل: شناسه، شناسه‌ی فرستنده، زمان ارسال، محتوا، شناسه‌ی پیامی که این توییت در پاسخ به آن نوشته شده است (در صورت وجود) به همراه شناسه‌ی فرستنده‌ی آن، شناسه‌ی پیامی که این توییت آن را تکرار (ری‌توییت) کرده است (در صورت وجود)، تعداد ری‌توییت‌ها، تعداد پاسخ‌ها، تعداد لایک‌ها، تعداد هشتگ‌های بکار برده شده و غیره)
- شناسه‌ی دنبال‌کنندگان و دوستان

برای بیش از ۵۰۰۰ حساب کاربری توییتر می‌باشد (۱۹۵۰ کاربر حقیقی و ۳۳۵۱ کاربر جعلی). بخشی از این مجموعه در گراف دنبال‌کننده-دوست به کار برده شده است.

مجموعه‌ی Cresci-2017 حاوی همان اطلاعات برای بیش از ۱۴,۰۰۰ حساب کاربری توییتر می‌باشد (۳۴۷۴ کاربر حقیقی و ۱۰,۹۲۴ کاربر جعلی) با این تفاوت که شناسه‌ی دنبال‌کنندگان و دوستان را شامل نمی‌شود. از این مجموعه در گراف دیدگاه استفاده شده است. اسنها کودوگونتا و امیلیو فرارا [۱۳] نیز پیش‌تر از مجموعه‌ی Cresci-2017 برای آموزش مدلشان بهره گرفته بودند.

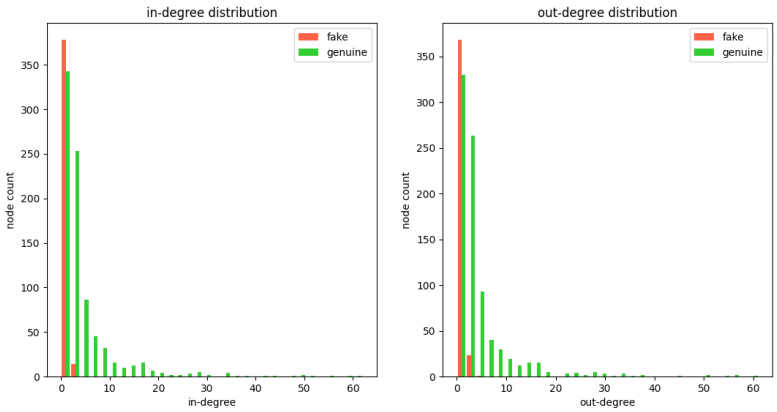
۴-۲ نتایج به دست آمده

۴-۲-۱ نمودارهای توزیع

در زیر نمودارهای توزیع میله‌ای هر معیار برای گراف دنبال‌کننده-دوست را پس از حذف داده‌های پرت[1] مشاهده می‌کنید:

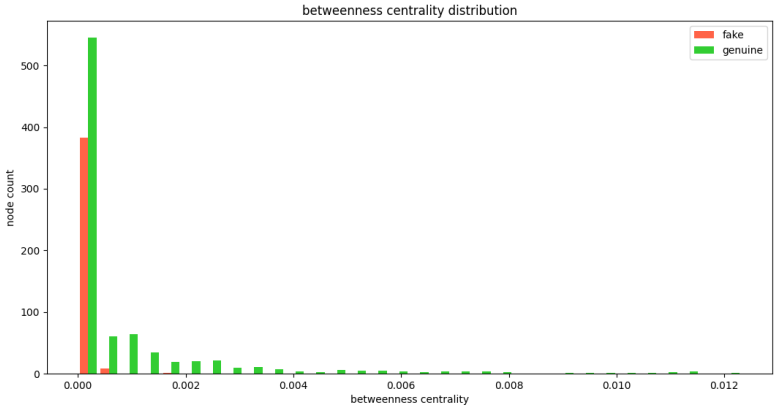
[1] outliers

- توزیع معیار درجه ورودی و درجه خروجی:



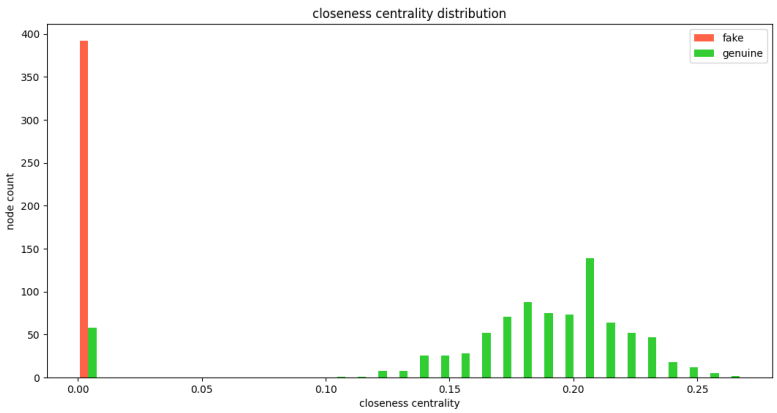
شکل ۴-۱: توزیع معیار درجه ورودی و خروجی

- توزیع معیار مرکزیت میانگی:



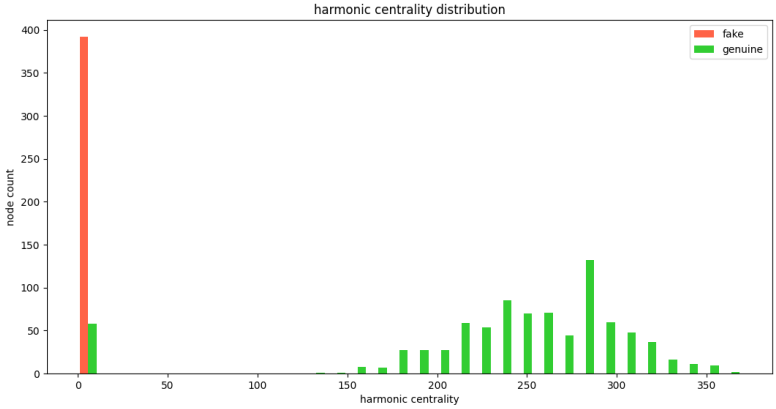
شکل ۴-۲: توزیع معیار مرکزیت میانگی

- توزیع معیار مرکزیت نزدیکی:



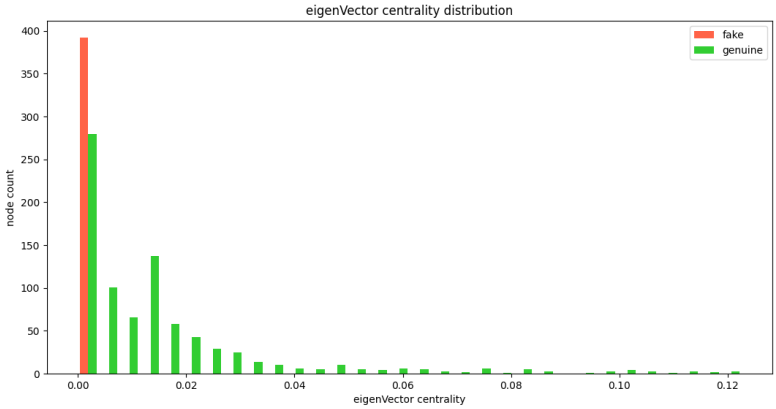
شکل ۴-۳: توزیع معیار مرکزیت نزدیکی

- توزیع معیار مرکزیت هارمونیک:



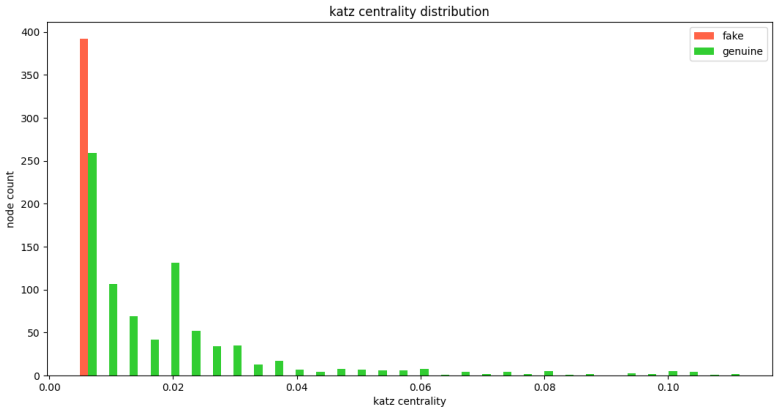
شکل ۴-۴: توزیع معیار مرکزیت هارمونیک

- توزیع معیار مرکزیت بردار ویژه:



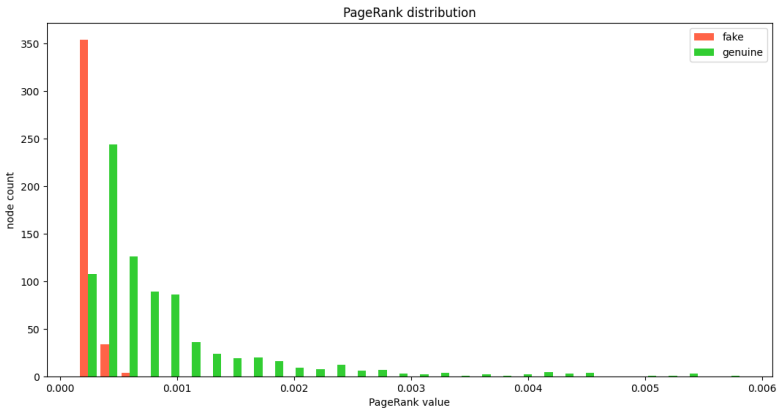
شکل ۴-۵: توزیع معیار مرکزیت بردار ویژه

- توزیع معیار مرکزیت کتز:



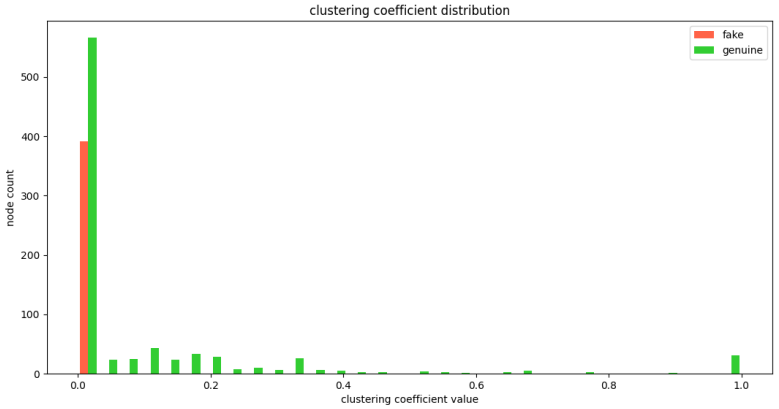
شکل ۴-۶: توزیع معیار مرکزیت کتز

- توزیع معیار رتبه صفحه:



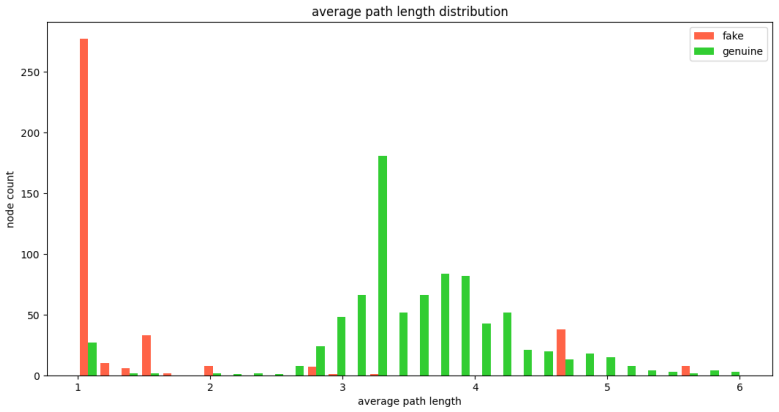
شکل ۴-۷: توزیع معیار رتبه صفحه

- توزیع معیار ضریب خوشه‌بندی:



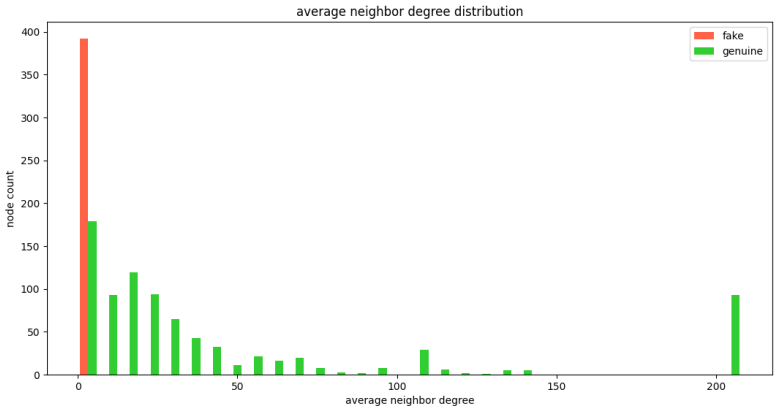
شکل ۴-۸: توزیع معیار ضریب خوشه‌بندی

- توزیع معیار میانگین کوتاه‌ترین طول مسیر:



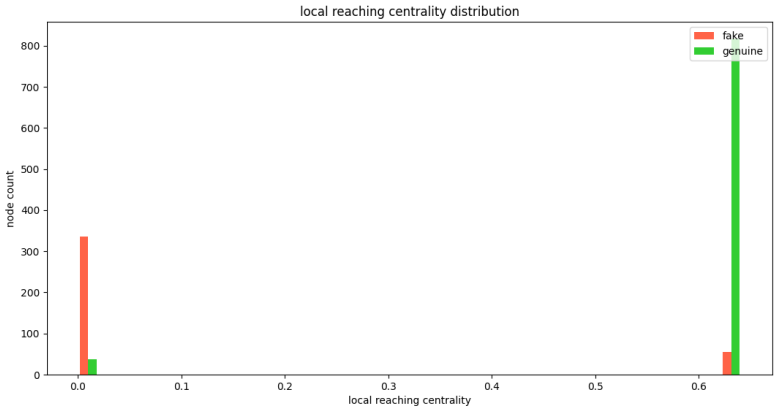
شکل ۴-۹: توزیع معیار میانگین کوتاه‌ترین طول مسیر

- توزیع معیار میانگین درجات همسایگی:



شکل ۴-۱۰: توزیع معیار میانگین درجات همسایگی

- توزیع معیار مرکزیت دستیابی محلی:

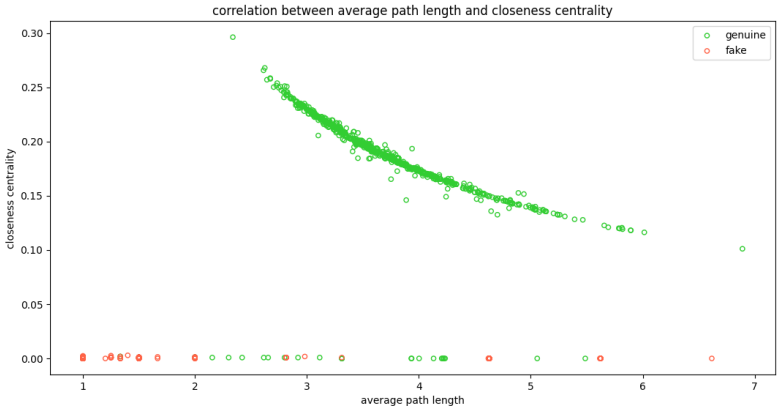


شکل ۴-۱۱: توزیع معیار مرکزیت دستیابی محلی

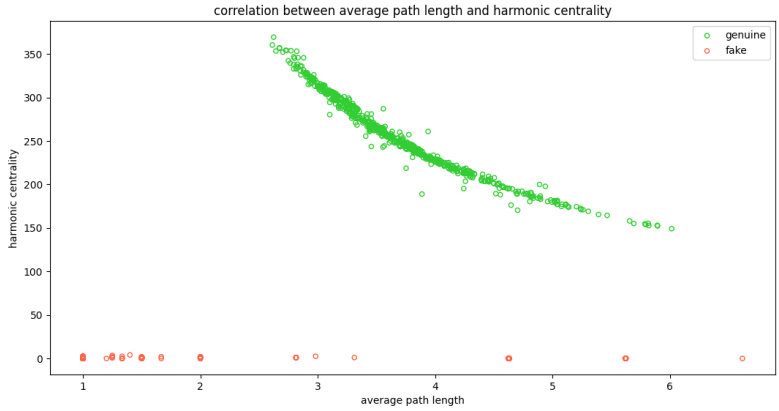
۴-۲-۲ نمودارهای همبستگی

در گام بعد همبستگی بین هر دو معیار را بررسی کردیم. نمودارهایی که در ادامه آورده شده‌اند، آن‌هایی هستند که دربردارنده‌ی اطلاعات معنادار و مفیدی می‌باشند (داده‌های پرت حذف شده‌اند).

در دو نمودار زیر، در گره‌های جعلی با افزایش میانگین کوتاه‌ترین طول مسیر مقدار مرکزیت نزدیکی / هارمونیک تقریباً ثابت و برابر با صفر باقی می‌ماند، درحالیکه در گره‌های حقیقی با افزایش میانگین کوتاه‌ترین طول مسیر الگوی یک منحنی با شیب منفی (مشابه تابع $y = 1/x$) مشاهده می‌شود (که کران پایین آن از صفر بالاتر است) :

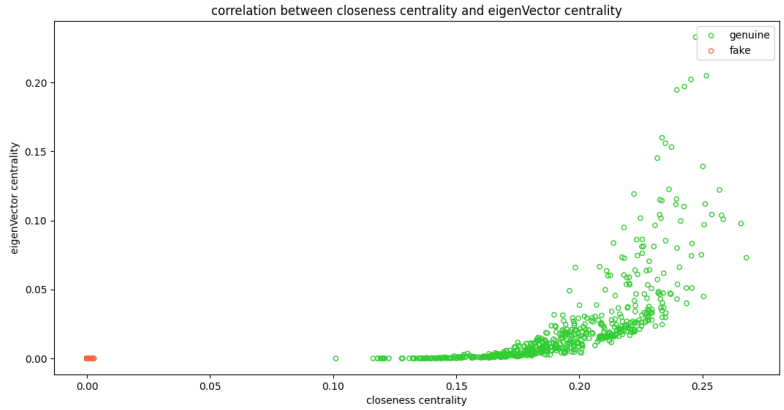


شکل ۴-۱۲: همبستگی میان میانگین کوتاه‌ترین طول مسیر و مرکزیت نزدیکی



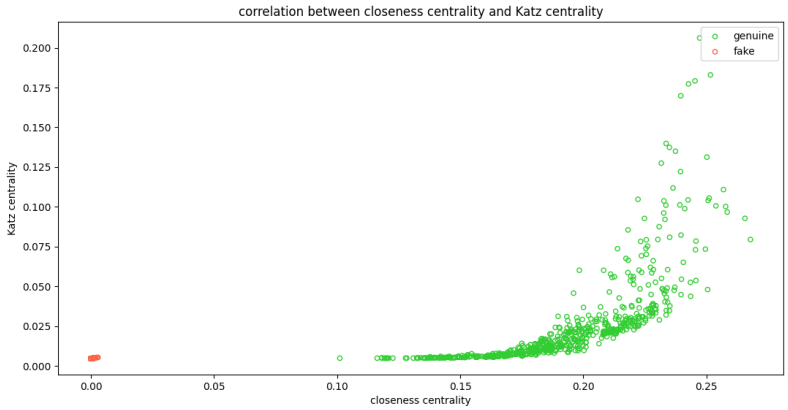
شکل ۴-۱۳: همبستگی میان میانگین کوتاه‌ترین طول مسیر و مرکزیت هارمونیک

در گره‌های جعلی با افزایش مقدار نزدیکی مقدار بردار ویژه ثابت و صفر باقی می‌ماند، اما در گره‌های حقیقی مقدار نزدیکی از حدود ۰٫۱۰ شروع می‌شود و با افزایش آن مقدار بردار ویژه مشابه یک تابع نمایی افزایش می‌یابد:



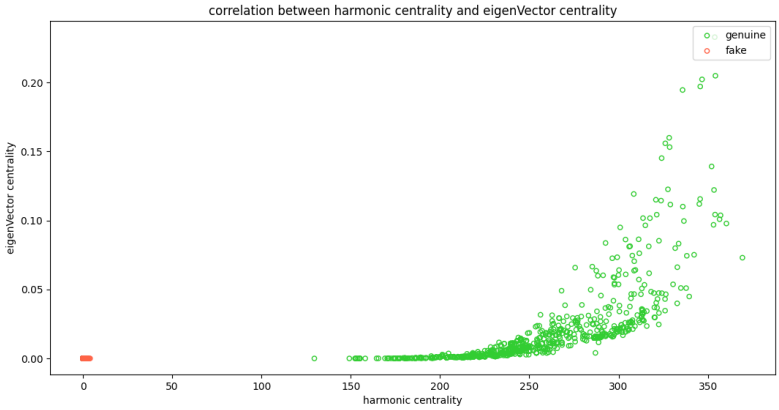
شکل ۴-۱۴: همبستگی میان مرکزیت نزدیکی و مرکزیت بردار ویژه

در گره‌های جعلی با افزایش مقدار نزدیکی مقدار کتز بطور خطی با شیب کم افزایش می‌یابد، اما در گره‌های حقیقی مقدار نزدیکی از حدود ۰٫۱۰ شروع می‌شود و با افزایش آن مقدار کتز مشابه یک تابع نمایی افزایش می‌یابد:



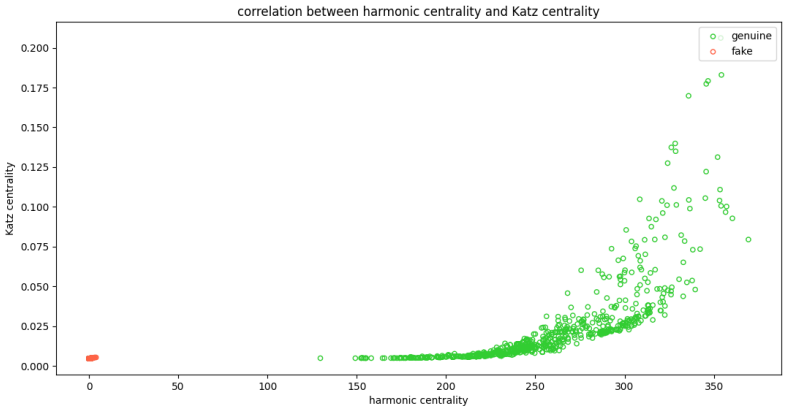
شکل ۴-۱۵: همبستگی میان مرکزیت نزدیکی و مرکزیت کتز

در گره‌های جعلی با افزایش مقدار هارمونیک مقدار بردار ویژه ثابت و صفر باقی می‌ماند، اما در گره‌های حقیقی مقدار هارمونیک از حدود ۱۳۵ شروع می‌شود و با افزایش آن مقدار بردار ویژه مشابه یک تابع نمایی افزایش می‌یابد:



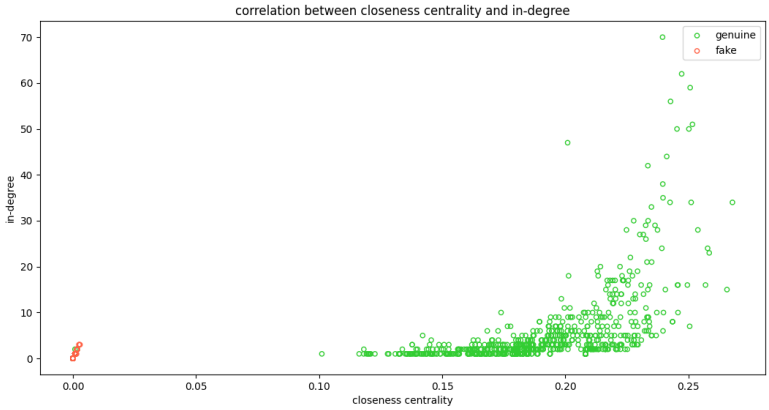
شکل ۴-۱۶: همبستگی میان مرکزیت هارمونیک و مرکزیت بردار ویژه

در گره‌های جعلی با افزایش مقدار هارمونیک مقدار کتز بطور خطی با شیب کم افزایش می‌یابد، اما در گره‌های حقیقی مقدار هارمونیک از حدود ۱۳۵ شروع می‌شود و با افزایش آن مقدار کتز مشابه یک تابع نمایی افزایش می‌یابد:



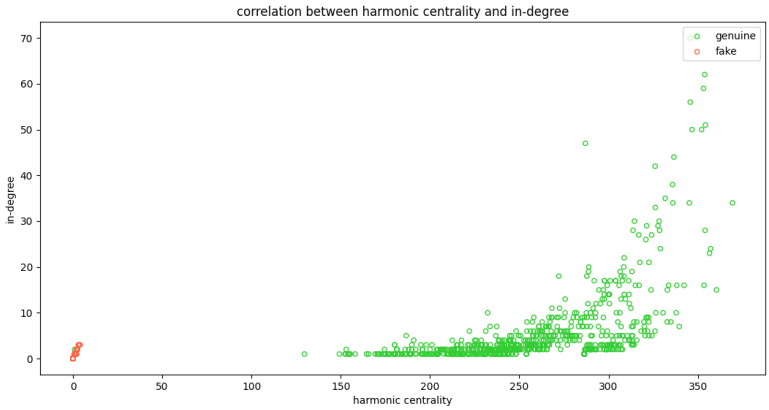
شکل ۴-۱۷: همبستگی میان مرکزیت هارمونیک و مرکزیت کتز

در گره‌های جعلی با افزایش مقدار نزدیکی مقدار درجه ورودی بطور خطی با شیب تند افزایش می‌یابد، اما در گره‌های حقیقی مقدار نزدیکی از حدود ۰٫۱۰ شروع می‌شود و با افزایش آن مقدار درجه ورودی مشابه یک تابع نمایی افزایش می‌یابد:



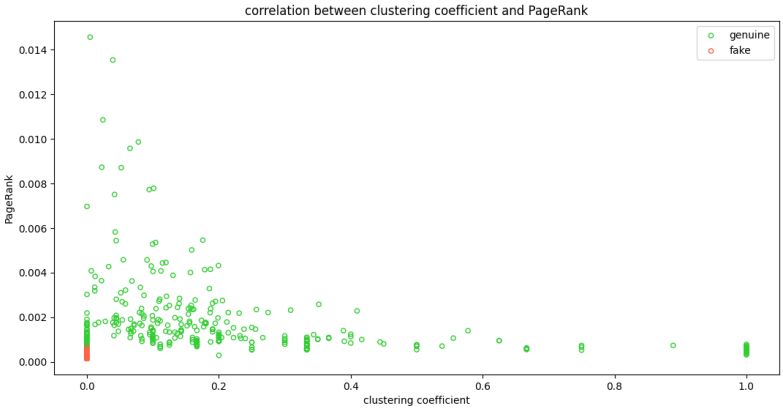
شکل ۴-۱۸: همبستگی میان مرکزیت نزدیکی و درجه ورودی

در گره‌های جعلی با افزایش مقدار هارمونیک مقدار درجه ورودی بطور خطی با شیب تند افزایش می‌یابد، اما در گره‌های حقیقی مقدار هارمونیک از حدود ۱۳۵ شروع می‌شود و با افزایش آن مقدار درجه ورودی مشابه یک تابع نمایی افزایش می‌یابد:



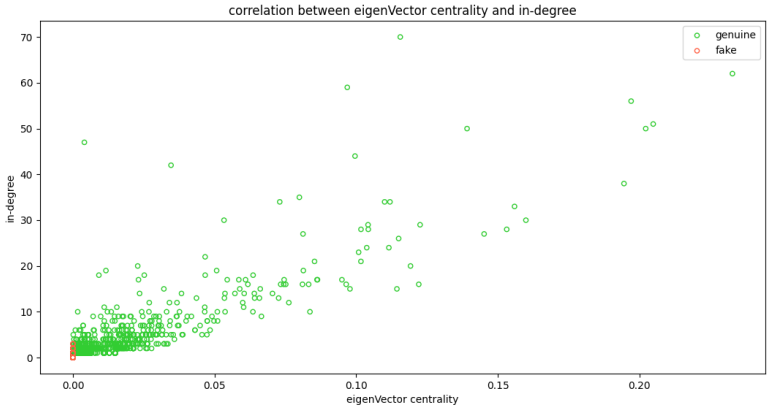
شکل ۴-۱۹: همبستگی میان مرکزیت هارمونیک و درجه ورودی

در گره‌های جعلی مقدار ضریب خوشه‌بندی ثابت و صفر است و با افزایش رتبه صفحه ثابت می‌ماند، اما در گره‌های حقیقی این الگو مشاهده نمی‌شود.

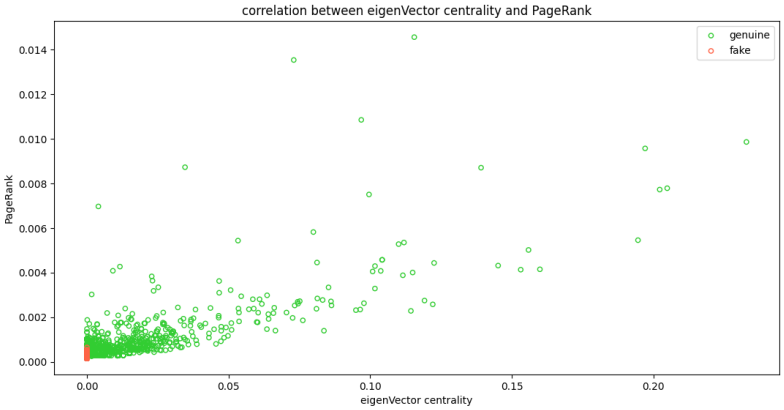


شکل ۴-۲۰: همبستگی میان ضریب خوشه‌بندی و رتبه صفحه

در دو نمودار زیر، در گره‌های حقیقی یک الگوی تقریبی به فرم یک خط با شیب مثبت کم را میان دو معیار بردار ویژه و درجه ورودی / رتبه صفحه مشاهده می‌کنیم، اما در گره‌های جعلی این الگو مشاهده نمی‌شود؛ بطوری که با افزایش درجه ورودی / رتبه صفحه مقدار بردار ویژه ثابت و صفر باقی می‌ماند:

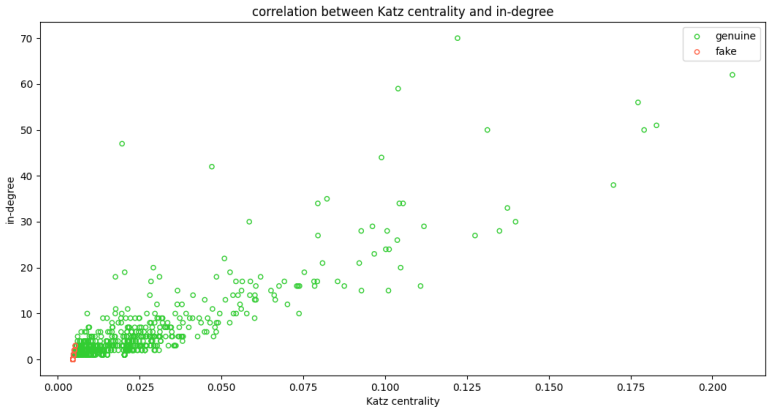


شکل ۴-۲۱: همبستگی میان مرکزیت بردار ویژه و درجه ورودی

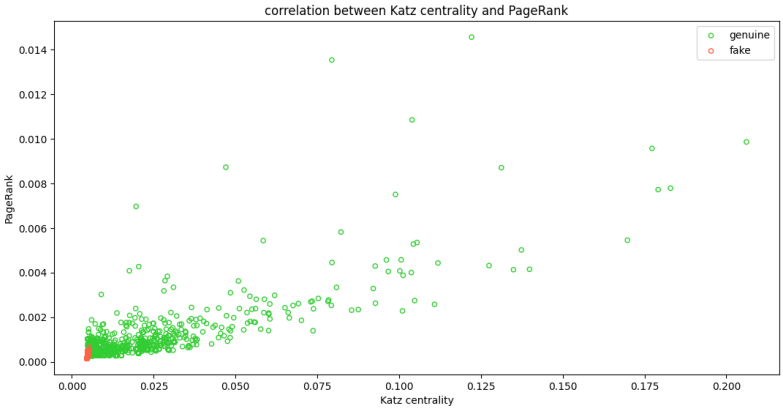


شکل ۴-۲۲: همبستگی میان مرکزیت بردار ویژه و رتبه صفحه

در دو نمودار زیر، در گره‌های حقیقی یک الگوی تقریبی به فرم یک خط با شیب مثبت کم را میان دو معیار کتز و درجه ورودی / رتبه صفحه مشاهده می‌کنیم، اما در گره‌های جعلی با افزایش مقدار کتز درجه ورودی / رتبه صفحه به فرم یک خط با شیب تند افزایش می‌یابد:

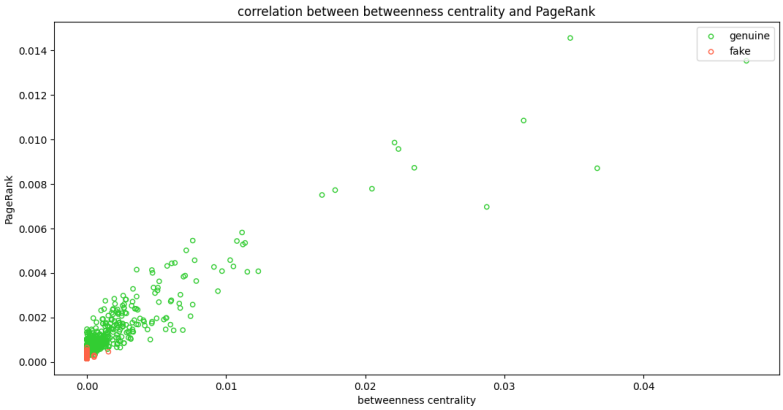


شکل ۴-۲۳: همبستگی میان مرکزیت کتز و درجه ورودی



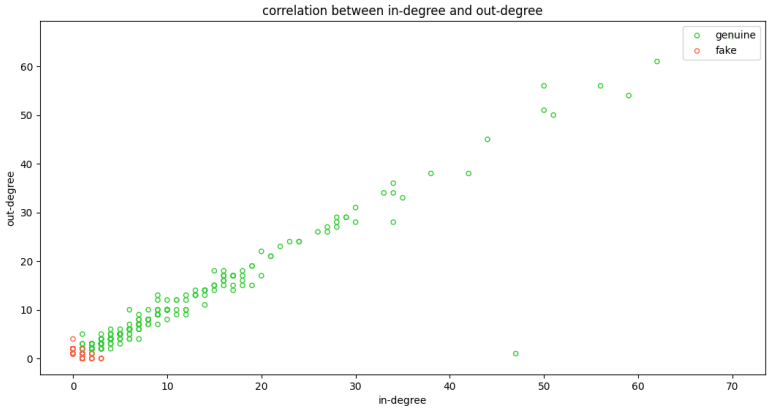
شکل ۴-۲۴: همبستگی میان مرکزیت کتز و رتبه صفحه

در گره‌های حقیقی یک الگوی تقریبی به فرم یک خط با شیب مثبت تند را میان دو معیار میانگی و رتبه صفحه مشاهده می‌کنیم، اما در گره‌های جعلی با افزایش رتبه صفحه مقدار میانگی تقریباً ثابت و صفر باقی می‌ماند:



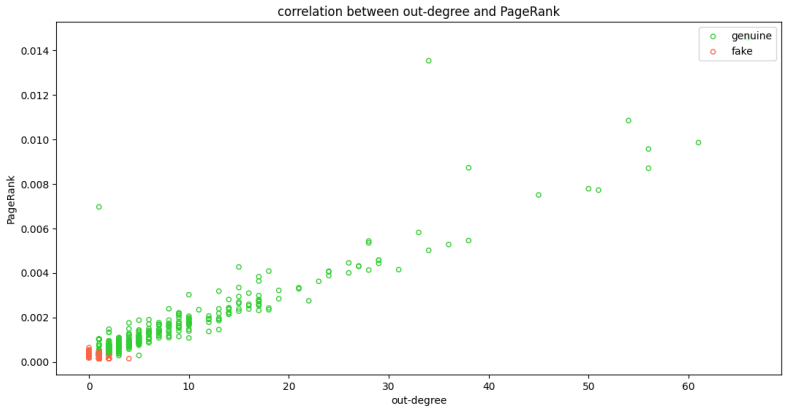
شکل ۴-۲۵: همبستگی میان مرکزیت میانگی و رتبه صفحه

در گره‌های حقیقی با افزایش درجه ورودی درجه خروجی نیز بطور خطی افزایش می‌یابد، درحالیکه در گره‌های جعلی بطور کلی مقدار درجه ورودی و خروجی بسیار کم است و با افزایش درجه ورودی درجه خروجی نسبتاً کاهش می‌یابد:



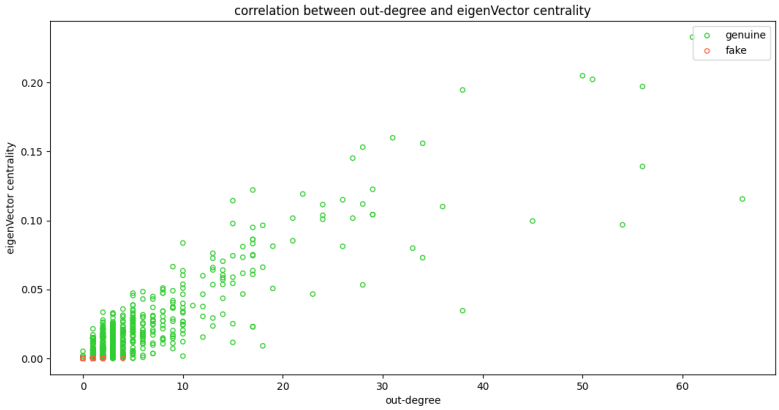
شکل ۴-۲۶: همبستگی میان درجه ورودی و درجه خروجی

در گره‌های حقیقی با افزایش درجه خروجی رتبه صفحه نیز بطور خطی افزایش می‌یابد، درحالیکه در گره‌های جعلی بطور کلی مقدار درجه خروجی و رتبه صفحه بسیار کم است و با افزایش درجه خروجی رتبه صفحه لزوماً افزایش نمی‌یابد (نسبتاً کاهش می‌یابد):



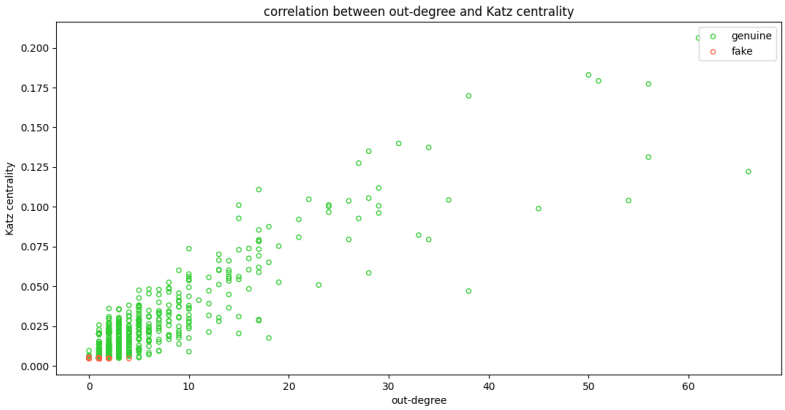
شکل ۴-۲۷: همبستگی میان درجه خروجی و رتبه صفحه

در گره‌های حقیقی با افزایش درجه خروجی مقدار بردار ویژه نیز افزایش می‌یابد، درحالیکه در گره‌های جعلی با افزایش درجه خروجی مقدار بردار ویژه ثابت و صفر باقی می‌ماند:



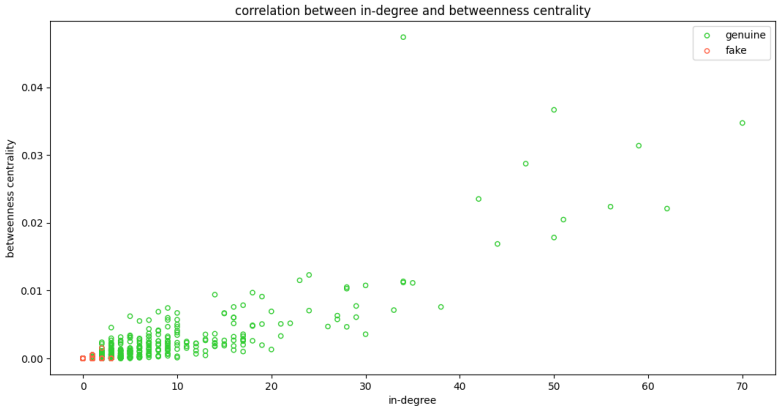
شکل ۴-۲۸: همبستگی میان درجه خروجی و مرکزیت بردار ویژه

در گره‌های حقیقی با افزایش درجه خروجی مقدار کتز نیز افزایش می‌یابد، درحالیکه در گره‌های جعلی با افزایش درجه خروجی مقدار کتز ثابت می‌ماند:

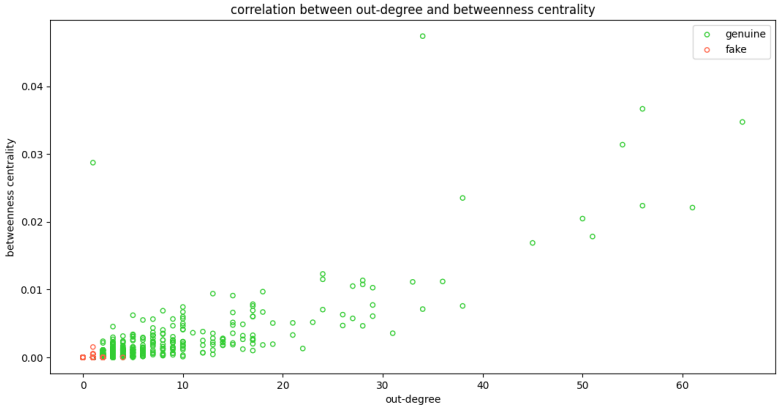


شکل ۴-۲۹: همبستگی میان درجه خروجی و مرکزیت کتز

در دو نمودار زیر، در گره‌های حقیقی با افزایش درجه ورودی / خروجی مقدار میانگی نیز بطور خطی با شیب کم افزایش می‌یابد، درحالیکه در گره‌های جعلی با افزایش درجه ورودی / خروجی مقدار میانگی تقریباً ثابت و صفر باقی می‌ماند:



شکل ۴-۳۰: همبستگی میان درجه ورودی و مرکزیت میانگی



شکل ۴-۳۱: همبستگی میان درجه خروجی و مرکزیت میانگی

۴-۳ تجزیه و تحلیل نتایج

همان‌طور که در شکل ۳-۲ مشاهده شد، در مجموعه داده‌ی پیش‌رو گره‌های حقیقی چگال‌تر و متمرکزتر هستند و هر گره‌ی حقیقی بطور میانگین تعداد دنبال‌کنندگان و دنبال‌شوندگان بیشتری نسبت به یک گره‌ی جعلی دارد. به همین سبب همان‌طور که انتظار می‌رود، در تمامی معیارهای بررسی شده گره‌های حقیقی مرکزیت بالاتری نسبت به گره‌های جعلی داشته‌اند.

همان‌گونه که در آزمایشات تجربی ما مشاهده می‌شود، از میان نمودارهای توزیع چهار معیار نزدیکی، هارمونیک، میانگین کوتاه‌ترین طول مسیر و دستیابی محلی بهترین عملکرد را در تفکیک گره‌های حقیقی از گره‌های جعلی داشته‌اند. به بیان دقیق، در نمودار توزیع نزدیکی تمام گره‌های جعلی مقداری بین ۰ و ۰٫۱ دارند، درحالی‌که اکثر گره‌های حقیقی مقداری بین ۰٫۱۲ و ۰٫۲۶ دارند. در نمودار توزیع هارمونیک تمام گره‌های جعلی مقداری بین ۰ و ۱۰ دارند، درحالی‌که اکثر گره‌های حقیقی مقداری بین ۱۵۰ و ۳۶۰ دارند. در نمودار توزیع میانگین کوتاه‌ترین طول

مسیر، اکثر گره‌های جعلی مقداری بین ۱ و ۲ دارند، در حالی‌که اکثر گره‌های حقیقی مقداری بین ۲٫۶ و ۶ دارند. در نمودار توزیع دستیابی محلی اکثر گره‌های جعلی مقداری بین ۰ و ۰٫۲ دارند، درحالی‌که اکثر گره‌های حقیقی مقداری بین ۰٫۶۲ و ۰٫۶۴ دارند. در هر سه معیار نزدیکی، هارمونیک و میانگین کوتاه‌ترین طول مسیر، فاصله‌ی یک گره تا تمام گره‌های دیگر محاسبه می‌شود و از آن‌جا که اکثر گره‌های جعلی در زیرگراف‌های ناهمبند کوچکی قرار گرفته‌اند، مقدار این سه معیار برایشان در مقایسه با گره‌های حقیقی بطور قابل توجهی کمتر است. همچنین طبق تعریف، دستیابی محلی به ازای هر گره تمام گره‌هایی را که بطور مستقیم یا غیرمستقیم با گره‌ی مورد نظر در ارتباط هستند، در نظر می‌گیرد. بدین ترتیب می‌توان نتیجه گرفت معیارهایی که به نوعی تمام گره‌های قابل دسترسی از گره‌ی مورد نظر را در محاسبات خود لحاظ می‌کنند، عملکرد بهتری در جداسازی گره‌های حقیقی از جعلی دارند.

از میان نمودارهای همبستگی، همبستگی میان میانگین کوتاه‌ترین طول مسیر و نزدیکی، همبستگی میان میانگین کوتاه‌ترین طول مسیر و هارمونیک، همبستگی میان بردار ویژه و درجه ورودی، همبستگی میان بردار ویژه و رتبه صفحه، همبستگی میان درجه ورودی و درجه خروجی و همبستگی میان درجه خروجی و رتبه صفحه به خوبی رفتار متفاوت گره‌های حقیقی را در مقابل گره‌های جعلی به نمایش گذاشته‌اند. به عنوان نمونه در نمودار همبستگی میان درجه ورودی و درجه خروجی، دلیل عملکرد مشاهده شده را می‌توان این‌طور تفسیر نمود که در میان کاربران حقیقی، کاربری که تعداد بیشتری کاربر را دنبال کند (افزایش درجه خروجی)، بطور معمول توسط تعداد کاربران بیشتری نیز دنبال خواهد شد (افزایش درجه ورودی)، اما در میان کاربران جعلی، چه بسا کاربری که تعداد بیشتری کاربر را دنبال کند (افزایش درجه خروجی)، توسط تعداد کاربران کمتری دنبال شود (کاهش درجه ورودی). در نمودار همبستگی میان درجه خروجی و رتبه صفحه نیز عملکرد مشابهی مشاهده می‌شود. دلیلش آن است که رتبه‌صفحه مانند درجه ورودی امتیاز هر گره را بر اساس امتیاز یال‌های ورودی‌اش محاسبه می‌کند. در ارتباط با دو نمودار همبستگی میان میانگین کوتاه‌ترین طول مسیر و نزدیکی و همبستگی میان میانگین کوتاه‌ترین طول مسیر و هارمونیک، ذکر این نکته خالی از لطف نیست که در گره‌های حقیقی یک رفتار طبیعی و منطقی به چشم می‌خورد، اما در گره‌های جعلی یک رفتار غیرطبیعی دیده می‌شود. به بیان دقیق، افزایش میانگین کوتاه‌ترین طول مسیر به معنای فاصله گرفتن از مرکزیت است و طبیعتاً مقدار دو معیار نزدیکی و هارمونیک که با مرکزیت نسبت مستقیم دارند، باید کاهش یابد. بااین‌وجود در گره‌های جعلی مقدار نزدیکی و هارمونیک تقریباً صفر است که با افزایش میانگین کوتاه‌ترین طول مسیر هم تقریباً ثابت باقی می‌ماند.

ضمناً تمامی نمودارهای توزیع و همبستگی از گراف دیدگاه هم گرفته شد، اما به دلیل داشتن عملکرد نسبتاً ضعیف در تفکیک گره‌های حقیقی از گره‌های جعلی، از آوردن نمودارهای آن در این پایان‌نامه صرف‌نظر شد.

فصل پنجم

جمع‌بندی و نتیجه‌گیری

۵-۱ جمع‌بندی و نتیجه‌گیری

با گسترش شبکه‌های اجتماعی، روزبه‌روز به تعداد حساب‌های کاربری و اخبار جعلی افزوده می‌شود. به همین دلیل امروزه نیاز به رفع این مشکل و ارائه‌ی روش‌هایی برای شناسایی اطلاعات غیرمعتبر و منابع اخبار جعلی بشدت احساس می‌شود. در این پروژه یک رویکرد تحلیل شبکه پیچیده‌ای بر روی کاربران حقیقی و جعلی شبکه‌ی اجتماعی توییتر به کار گرفته شده است. بدین منظور گرافی ساخته شده که بخشی از گره‌های آن را کاربران جعلی و بقیه را کاربران حقیقی تشکیل داده‌اند. سپس معیارهایی که در تحلیل شبکه‌های عصبی مورد استفاده قرار می‌گیرند، برای تحلیل رفتار کاربران جعلی و حقیقی بر روی این گراف مورد بررسی قرار گرفته‌اند. از میان معیارهای مورد مطالعه، چهار معیار نزدیکی، هارمونیک، میانگین کوتاه‌ترین طول مسیر و دستیابی محلی بهترین عملکرد را در تفکیک گره‌های حقیقی از گره‌های جعلی داشته‌اند. می‌توان نتیجه گرفت معیارهایی که به نوعی تمام گره‌های قابل دسترسی از هر گره را در محاسبات خود لحاظ می‌کنند، قادرند به خوبی رفتار متفاوت کاربران حقیقی را در مقابل کاربران جعلی به نمایش بگذارند. همچنین از میان نمودارهای همبستگی، همبستگی میان میانگین کوتاه‌ترین طول مسیر و نزدیکی، همبستگی میان میانگین کوتاه‌ترین طول مسیر و هارمونیک، همبستگی میان درجه ورودی و درجه خروجی و همبستگی میان درجه خروجی و رتبه صفحه بهتر می‌توانند تفاوت رفتار گره‌های حقیقی و جعلی را آشکار سازند. چگونگی عملکرد دو نمودار اخیر را می‌توان این‌گونه استدلال نمود که عموماً میان تعداد دنبال‌کنندگان و تعداد دنبال‌شوندگان گره‌های حقیقی رابطه‌ی مستقیمی وجود دارد، درحالی‌که روند مشخصی میان تعداد دنبال‌کنندگان و تعداد دنبال‌شوندگان گره‌های جعلی به چشم نمی‌خورد.

۵-۲ کارهای آتی

می‌توان از نتایج این پروژه و تفاوت‌های آشکار شده میان رفتار کاربران حقیقی و رفتار کاربران جعلی شبکه‌های اجتماعی در طراحی روش‌های پیشرفته مبتنی بر یادگیری ماشین برای شناسایی هرچه دقیق‌تر حساب‌های کاربری جعلی استفاده نمود. به بیان دقیق، در بسیاری از روش‌های یادگیری ماشین یک بردار ویژگی[1] از مشخصه‌های قابل اندازه‌گیری (در موضوع مورد بحث، مشخصه‌های استخراج شده از پروفایل و فعالیت‌های کاربران) ساخته می‌شود. می‌توان به منظور تقویت این بردار، هریک از معیارهای بررسی شده را به عنوان یک ویژگی به خانه‌های این بردار اضافه نمود و در نتیجه به یک الگوریتم پیش‌بینی‌کننده‌ی قوی‌تر دست یافت.

[1] feature vector

منابع و مراجع

[1] MIB Datasets. http://mib.projects.iit.cnr.it/dataset.html.

[2] Networkx, a python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. https://networkx.org/.

[3] Bharti, K. K., and Pandey, S. Fake account detection in twitter using logistic regression with particle swarm optimization. Soft Computing 25, 16 (2021), 11333–11345.

[4] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008, 10 (2008), P10008.

[5] Breuer, A., Eilat, R., and Weinsberg, U. Friend or faux: graph-based early detection of fake accounts on social networks. in Proceedings of The Web Conference 2020 (2020), pp. 1287–1297.

[6] Cao, Q., Sirivianos, M., Yang, X., and Pregueiro, T. Aiding the detection of fake accounts in large scale social online services. in 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12) (2012), pp. 197–210.

[7] Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE Transactions on dependable and secure computing 9, 6 (2012), 811–824.

[8] Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. Fame for sale: Efficient detection of fake twitter followers. Decision Support Systems 80 (2015), 56–71.

[9] Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. in Proceedings of the 26th international conference on world wide web companion (2017), pp. 963–972.

[10] Davis, C. A., Varol, O., Ferrara, E., Flammini, A., and Menczer, F. Botornot: A system to evaluate social bots. in Proceedings of the 25th international conference companion on world wide web (2016), pp. 273–274.

[11] Homsi, A., Al Nemri, J., Naimat, N., Kareem, H. A., Al-Fayoumi, M., and Snober, M. A. Detecting twitter fake accounts using machine learning and data reduction techniques.

[12] Khaund, T., Al-Khateeb, S., Tokdemir, S., and Agarwal, N. Analyzing social bots and their coordination during natural disasters. in International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (2018), Springer, pp. 207–212.

[13] Kudugunta, S., and Ferrara, E. Deep neural networks for bot detection. Information Sciences 467 (2018), 312–322.

[14] Lee, K., Eoff, B., and Caverlee, J. Seven months with the devils: A long-term study of content polluters on twitter. in Proceedings of the international AAAI conference on web and social media (2011), volume 5, pp. 185–192.

[15] Lee, S., and Kim, J. Early filtering of ephemeral malicious accounts on twitter. Computer Communications 54 (2014), 48–57.

[16] Morstatter, F., Wu, L., Nazer, T. H., Carley, K. M., and Liu, H. A new approach to bot detection: striking the balance between precision and recall. in 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (2016), IEEE, pp. 533–540.

[17] Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., and Liu, H. Combating disinformation in a social media age. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10, 6 (2020), e1385.

[18] Varol, O., Ferrara, E., Davis, C., Menczer, F., and Flammini, A. Online human-bot interactions: Detection, estimation, and characterization. in Proceedings of the international AAAI conference on web and social media (2017), volume 11.

Abstract

With the growing rate of social networks, the number of fake accounts and fake news is multiplying day by day. To detect bots, researchers have proposed several approaches, among which we can mention Machine Learning techniques and classification models (e.g., Decision Trees, Deep Neural Networks, Logistic Regression, and Support Vector Machines) which label each user as human or bot. From a different point of view, in every social network users interact with each other in the context of graphs. For example, in Twitter there are graphs of follower-following, comment, retweet, mention, and so on. Thus, a Complex Network is formed that can be analyzed using a set of tools and criteria. The aim of this project is to analyze Complex Networks on real and fake Twitter accounts, in order to examine the behavior of bots compared to humans. After applying several criteria to the implemented graphs, we found some of them effective in separating fake users from real ones, and consequently, we propose they be used to identify fake user accounts.

Keywords:

Twitter, social media, complex networks, bot, fake, genuine, account, graph, correlation, centrality, closeness, harmonic, pagerank, local reaching, degree

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Computer Engineering

BSc Thesis

Complex Network Analysis for Detecting Fake Accounts in Twitter

By:

Sara Asghari

Advisor:

Dr. Mostafa HaghirChehreghani

February 2022