تابستان ۱۴۰۰، دوره ۶، شماره ۲، پیاپی ۲۱
DOR :20.1001.1.2538161.1400.6.2.2.2

دریافت مقاله: ۱۳۹۹/۱۰/۲۸
پذیرش مقاله: ۱۳۹۹/۱۲/۱۲

نوع مقاله: پژوهشی

شناسایی عناوین محتوای متنی منتشرشده در شبکه اجتماعی توئیتر

فاطمه بهاری‌فرد*
پژوهشکده علوم کامپیوتر، پژوهشگاه دانش‌های بنیادی (IPM)
پست الکترونیکی: f.baharifard@ipm.ir

وحید معتقد
پژوهشکده علوم کامپیوتر، پژوهشگاه دانش‌های بنیادی (IPM)
پست الکترونیکی: vahid.motaghed.2020@gmail.com

چکیده

با رشـــد روز افزون شبکه‌های اجتماعی، میل به تحلیل محتوای منتشرشـــده برای مقاصد گوناگون افزایش یافته اســـت. یک دســـته از عمده فعالیت‌هایی کـــه در این حوزه انجام می‌شود شناسایی و دسته‌بندی محتواهای تولیدشده اســـت. این موضوع به معنی گروه‌بندی مطالب منتشرشده در دســـته‌هایی با موضوعات مشابه و ارائه برچسب‌های پیشنهادی برای هر دسته می‌باشد.

در ایـــن مقالـــه، الگوریتم جدیـــدی برای دســـته‌بندی محتوای متنی شـــبکه اجتماعی توئیتر ارائه شده است. در این الگوریتم ابتدا هر متن، پیش‌پردازش شده و سپس یک گراف ارتباطات جدید مبتنی بر محتوای متن‌های منتشرشده ســـاخته می‌شـــود. این گراف وزن‌دار و بی‌جهت اســـت و روی آن با اســـتفاده از دو روش بدون ناظر، تشـــکل‌های مختلف شناســـایی می‌شوند. برای ارزیابی، داده‌های متنی ارسال‌شده از شهر واشـــینگتن در یک بازه زمانی، با API جمع‌آوری و الگوریتم‌های ارائه‌شـــده روی آن اعمال شده اســـت. برای بررســـی دقت، نتایج حاصل بـــا دو الگوریتم کلاسیک K-means و LDA بر اساس معیار اطلاعات متقابل نرمال‌شده، مقایسه شده است که نشان‌دهنده دقت مناسب الگوریتم پیشنهادی است.

واژه‌های کلیدی: شناســـایی عنوان، گـــراف وزن‌دار، یادگیری بدون ناظر، شبکه‌های اجتماعی، توئیتر

۱- مقدمه

طی ســـالیان اخیر استفاده از شـــبکه‌های اجتماعی در بین عموم جامعه به شـــدت مورد اقبال قرار گرفته اســـت. در این شـــبکه‌ها حجم عظیمی از داده‌ها با سرعت و تنوع زیاد (به‌عنوان مثال تصاویر، متن و فیلم) تولید و منتشـــر می‌شـــود. روزانه ۵۰۰ میلیون توئیت در توئیتر (هر توئیت حداکثـــر ۲۸۰ حـــرف) ارســـال و ۹۰۰ میلیـــون عکس در فیســـبوک و ۰/٤ میلیون ساعت فیلم در یوتیوب بارگذاری می‌شود [۱]. کاربرد این شـــبکه‌ها از برقراری ارتباط بین افراد تا کســـب اطلاعات و اخبار در حوزه‌های وسیعی از مسائل سیاســـی، اجتماعی، فرهنگی و غیره گسترده شده اســـت. با این حال بســـیاری از داده‌های تولیدشده در این بسترها، ساختارمند نیســـتند. پس نمی‌توان به سادگی در مورد چگونگی انتشار موضوعات مختلف و هم‌چنین تعامل

* نویسندهٔ مسئول

کاربران با یکدیگر اطلاعاتی به دست آورد [۲].

در تکنیک‌های خوشه‌بندی اسناد[1]، از مجموعه الگوریتم‌های یادگیری ماشین استفاده می‌شود. اسنادی که در یک خوشه قرار می‌گیرند، نسبت به سند دیگری از خوشه دیگر، مشابه‌تر هستند. در گذشته روش‌های متنوعی برای خوشه‌بندی اسناد ارائه شده است [۳,٤]. برای این منظور نیاز است تا هر سند به ساختار مناسبی برای پردازش تبدیل شود که رایج‌ترین روش‌ها برای این کار عبارتند از کیسه کلمات[2] که در آن هر سند تبدیل به مجموعه‌ای از کلمات می‌شود [۵]، TF-IDF که در آن کلمات مهم هر سند در مقابل کلمات مهم سایر اسناد سنجیده [٦]، و کلمه-به-بردار[3] که در آن هر کلمه با توجه به معنایش به بردار تبدیل می‌شود [۷]. سپس الگوریتم‌های یادگیری ماشین بدون ناظر[4] خوشه‌بندی سلسله مراتبی و تقسیمی مانند K-means روی آن اعمال می‌شود [۸].

مدل‌سازی عناوین[5] شامل روش‌هایی برای کشف الگوهای مناسب از کلمات استفاده شده در اسناد است. از این روش برای تحلیل محتوای متنی موجود در شبکه‌های اجتماعی نیز استفاده می‌شود [۹]. در این حوزه معمولاً عنوان هر سند بر اساس توزیع کلماتی که داخل آن به کار رفته است، استخراج می‌شود. برای مدل‌سازی، روش‌های متنوعی مانند LDA، PLSA، LSA و CTM ارائه شده است [۱۰].

با خوشه‌بندی محتوای متنی شبکه‌های اجتماعی، می‌توان به مدل‌سازی عناوین و دسته‌بندی آن‌ها پرداخت. در شبکه‌های اجتماعی مانند توئیتر، فیسبوک و اینستاگرام از هشتگ به‌عنوان کلمات پیوندی تعریف‌شده توسط کاربر استفاده شده است. این عبارات می‌توانند به‌عنوان ابزاری برای ارزیابی کیفیت خوشه‌بندی متن‌ها و مدل‌سازی عناوین استفاده شود [۱۱]. یکی از چالش‌های استفاده از این الگوریتم‌ها، تفاوت بین اسناد و محتوای متنی منتشرشده در شبکه‌های اجتماعی است. این محتواها معمولاً دارای اختلال[6] (غلط املایی، عبارات عامیانه، متن کوتاه، اشتباهات دستور زبانی و غیره) می‌باشند [۹]. چالش دیگر برای پردازش این نوع داده‌ها این است که داده‌های شبکه‌های اجتماعی، مجوز پخش ندارند و بنابراین مجموعه داده مناسبی برای آموزش و ارزیابی الگوریتم‌های پیشنهادی در این حوزه موجود نیست. به‌عنوان مثال توئیتر فقط مجوز پخش شناسه هر توئیت را مجاز می‌داند و افراد علاقه‌مند می‌توانند با API به متن هر توئیت دسترسی داشته باشند و این مسئله برای هر یک از شبکه‌های اجتماعی تابع مقررات مربوط به خود است [۱۲]. پس پژوهش‌های مختلف، از روش‌های متنوعی برای ارزیابی کیفیت الگوریتم خود استفاده کرده‌اند. به‌عنوان مثال برای ارزیابی خوشه‌بندی داده‌های توئیتر، از معیارهای مبتنی بر برچسب‌گذاری برای هر خوشه تا تفسیر خوشه‌های حاصله توسط عامل انسانی به صورت کیفی، استفاده شده است [۱۳]. بنابراین مقایسه بین پژوهش‌های مختلف، بسیار دشوار خواهد بود.

عمده پژوهش‌های انجام‌شده برای دسته‌بندی و شناسایی عناوین محتوای متنی به دو گروه اصلی تقسیم می‌شوند. در گروه اول پس از پیش‌پردازش هر متن، اقدام به عملیات خوشه‌بندی با روش‌های مختلف می‌شود. در گروه دوم با پردازش هر متن و شناسایی ارتباطات مختلف، گراف ارتباطی-محتوایی تشکیل شده و سپس اقدام به خوشه‌بندی گراف و یا شناسایی تشکل[7] می‌شود. در ادامه برخی از پژوهش‌های انجام شده در این دو دسته آمده است.

در مرجع [۱٤]، نویسندگان برای شناسایی عناوین توئیت‌ها، الگوریتم K-means را روی بردارهای TF-IDF محاسبه‌شده از اسناد اعمال کرده و نشان دادند عملکرد آن‌ها بهتر از LDA خواهد شد. در مقاله [۱۵]، مجموعه داده حوزه خبری موجود در UCI در نظر گرفته شده و با روش کلمه-به-بردار، هر کلمه از هر سند با ۵۰ ویژگی عددی نمایش داده شده است. سپس از SOM برای کاهش بعد و

1- Document Clustering
2- Bag of Words
3-Word2Vec
4- Unsupervised Learning
5- Topic Detection
6- Noise
7- Community Detection

در نهایت از خوشه‌بندی K-means استفاده شده است. در این پژوهش خوشـــه‌بندی روی کلمات اعمال شده و نشان داده شـــده است که کلمات مشـــابه در خوشه‌های یکسان دســـته‌بندی می‌شـــوند. در مقاله [۱۶] برای خوشـــه‌بندی متن‌های منتشرشده در توئیتر و ردیت، چهار روش نمایش ســـند TF-IDF، کلمه–به–بـــردار، کلمه–به–بردار وزن‌دار و هم‌چنین سند–به–بردار به همراه الگوریتم‌های خوشه‌بندی K-median، K-means، خوشه‌بندی سلسله مراتبی و NMF اعمال و نتایج با یکدیگر مقایسه شده است. در [۱۷]، برای خوشه‌بندی مستندات، کلمات هر سند شناسایی و برای هر کدام مطابق الگوریتم کلمه–به–بردار گوگل، یک بردار ۳۰۰ بعدی نمایش داده شده و با میانگین‌گیری از این کلمات، به ازای هر سند یک بردار ۳۰۰ بعدی در نظر گرفته شده است. سپس روی ماتریس خصوصیات مستندات، الگوریتم‌های کلاسیک خوشه‌بندی اعمال شده و نتایج با یکدیگر مقایسه شده‌اند.

در [۱۸] برای شناســـایی عناوین توئیت‌هایی که پس از زلزله ژاپن در سال ۲۰۱۱ منتشر شده است، با پنجره‌بندی توئیت‌ها بر اســـاس زمان ارسال آن‌ها به صورت ساعتی، توئیت‌های هر پنجره طبق کیسه کلمات به فرم بردار دودویی در می‌آید. با استفاده از این ماتریس، گرافی ساخته شده و روی آن خوشه‌بندی انجام می‌شود. نویسندگان این مقاله نشـــان دادند که عملکرد الگوریتم ارایه شده بهتر از LDA و K-means اســـت. در مرجع [۱۹]، کلمات کلیدی هر سند با روش کیسه کلمات شناسایی و با آن گرافی شکل می‌گیرد. در ایـــن گراف اگـــر دو کلمه متفاوت در یک ســـند موجود باشـــند، بین آن‌ها یال در نظر گرفته می‌شود. در نهایت از الگوریتم شناســـایی تشکل Louvain استفاده و تشکل‌هایی که معیار مرکزیت آن‌ها کمتر از حد آســـتانه‌ای باشـــد با یکدیگر ادغام می‌شـــوند. نتایج این پژوهش با K-means و الگوریتم‌های خوشه‌بندی سلسله مراتبی مقایسه و عملکرد خوب آن گزارش شـــده است. هم‌چنین در مقاله [۲۰] برای شناســـایی عناوین توئیت‌ها از الگوریتمی مبتنی بر گراف استفاده شده است. هر توئیت با سه نوع متفاوت از رأس‌ها در یک گراف نمایش داده می‌شود: توئیت، هشتگ و کلمات موجود در هر توئیـــت. بین رأس‌های کلمات و توئیت، بین رأس‌های هشتگ و توئیت و هم‌چنین بین کلمات و هشتگ، یال وزن‌دار ایجاد می‌شـــود. وزن یال‌های هشتگ–توئیت بیشـــینه مقدار یـــک، وزن یال‌های توئیت–کلمـــه مبتنی بر TF-IDF و وزن یال‌های کلمه–هشتگ به میزان دفعاتی که با یکدیگر تکرار شـــده‌اند، در نظر گرفته شده است. در نهایت روی این گراف، خوشـــه‌بندی با الگوریتم RankClus انجام و برخی از خوشه‌ها به‌عنوان اختلال در نظر گرفته شده و حذف شـــده‌اند. این پژوهش با LDA و K-means مقایسه و بهبود عملکرد آن نشان داده شده است.

در این مقاله ما الگوریتم جدیدی برای شناسایی عناوین محتوای متنی منتشرشـــده در شبکه اجتماعی توئیتر ارائه می‌دهیـــم که در آن متن هر محتوا ابتدا در یک بردار تعبیه شـــده و در ادامه گـــراف ارتباطات مربوط به آن ســـاخته می‌شـــود. در انتها با شناسایی تشـــکل روی گراف، تلاش می‌شود برای هر دسته عنوانی در نظر گرفته شود. بر این مبنا دو الگوریتم متفاوت شناســـایی تشکل مورد استفاده قرار گرفته اســـت و هر یک از دو الگوریتم پیشنهای مبتنی بر یکی از الگوریتم‌های شناسایی تشکل می‌باشد. در پایان دقت الگوریتم‌های پیشـــنهادی بـــا الگوریتم‌های مبنایی در این حوزه با معیار اطلاعات متقابل نرمال‌شـــده [۸] مقایســـه می‌شود.

۲- الگوریتم پیشنهادی

در ایـــن بخش الگوریتم پیشـــنهادی برای تعیین عنوان یا موضوع متن منتشرشده در توئیتر ارائه می‌شود. برای این منظور باید متن‌ها دسته‌بندی و سپس عنوان هر دسته شناسایی شـــود. بر این مبنا الگوریتم پیشنهادی از چهار قسمت کلی تشکیل شده است:

۱) عملیات پیش پردازش و تهیه مجموعه داده مناسب

۲) محاسبه شباهت مفهومی بین متن‌های مجموعه داده

8-Normalized Mutual Information

۳) ساخت گراف ارتباطی متناسب با محتوا

٤) اجرای الگوریتم شناسایی تشکل روی گراف حاصل

فرض کنید ورودی الگوریتم یک مجموعه $\mathbf{N}$ عضوی از متن‌های منتشرشده باشد. متن‌های منتشرشده در شبکه‌های اجتماعی معمولا کوتاه، حاوی عبارات عامیانه، اشتباهات نگارشی و املایی می‌باشند. پس باید در ابتدا روی همه متن‌ها عملیات پیش‌پردازش انجام شود.

● **پیش‌پردازش**: در این مرحله، هر یک از اعضای مجموعه متن‌های ورودی، کلمه‌بندی [۹] و سپس علامت‌های نگارشی، کلمات توقف [۱۰] و پیوندها از بین آن‌ها حذف می‌شوند. پس به ازای متن $\mathbf{n}$–ام از ورودی، یک بردار کلمات خواهیم داشت:

$$D_n = [T_1^n, ..., T_{m_n}^n] \ , \ n \in \{1, ..., N\}, \qquad (۱)$$

که m_n تعداد کلمات حاصل از پیش پردازش روی متن $\mathbf{n}$–ام است. در این مقاله، فقط متن‌های منتشرشده به زبان انگلیسی بررسی شده و سایر زبان‌ها از مجموعه داده ورودی حذف می‌شوند. برای شناسایی زبان متن، از سرویس گوگل استفاده شده است.

پس از عملیات پیش‌پردازش، برای محاسبه شباهت مفهومی بین متون مختلف، به ازای هر متن، یک بردار ویژگی ساخته شده و سپس ماتریس مربوط به شباهت متن‌ها تشکیل می‌گردد:

● **ساخت بردار ویژگی**: به ازای هر متن ورودی، هر کلمه از آن با استفاده از مدل کلمه–به–بردار ازپیش‌آموزش‌دیده fastText مبتنی بر دو میلیون کلمه [۲۱]، به فضای ۳۰۰ بعدی نگاشت می‌شود. یعنی به ازای کلمه $\mathbf{i}$–ام از متن $\mathbf{n}$–ام، مطابق رابطه (۲) برداری محاسبه می‌شود که در آن $\mathbf{d = 300}$ است:

$$fastText(T_i^n) = [x_{i1}^n, ..., x_{id}^n] \ , \ i \in \{1, ..., m_n\}, \qquad (۲)$$

در ادامه برای هر متن، میانگین بردارهای $\mathbf{d}$ بعدی کلمات همان متن، به‌عنوان بردار ویژگی آن لحاظ می‌شود.

9- Tokenizing
10- Stop Words

یعنی اگر داشته باشیم:

$$F_j^n = \frac{\sum_{i=1}^{m_n} x_j^n}{m_n}, \quad j \in \{1, ..., d\}, \qquad (۳)$$

بردار ویژگی متن $\mathbf{n}$–ام به صورت زیر است:

$$F(D_n) = [F_1^n, ..., F_d^n] \ , \ n \in \{1, ..., N\} \ , \qquad (۴)$$

● **محاسبه ماتریس فاصله**: در این مرحله، از بردار ویژگی‌های به‌دست‌آمده، برای محاسبه ماتریس فاصله استفاده می‌شود. در ماتریس فاصله، $\mathbf{d_{ij}}$ نشانگر فاصله اقلیدسی بین بردار ویژگی متن $\mathbf{i}$–ام و بردار ویژگی متن $\mathbf{j}$–ام است که از رابطه (۵) به دست می‌آید:

$$d_{ij} = \| F(D_i) - F(D_j) \|_2, \ i, j \in \{1, ..., N\}, \qquad (۵)$$

ماتریس $\mathbf{d_{ij}}$ مربعی، متقارن و با اندازه $\mathbf{N \times N}$ می‌باشد که اعضای روی قطر اصلی آن صفر است.

● **ساخت گراف ارتباطات**: برای طبقه‌بندی بهتر متن‌های منتشرشده، از ساختار گراف استفاده می‌کنیم. برای این منظور یک گراف بی‌جهت وزن‌دار به نام $\mathbf{G = (V, E)}$ با مجموعه رئوس $\mathbf{V}$ و مجموعه یال‌های $\mathbf{E}$ ساخته می‌شود. مجموعه رئوس، شامل دو نوع رأس شناسه $(\mathbf{V_1})$ و ویژگی خاص متن $(\mathbf{V_2})$ است. به‌عنوان مثال در فیسبوک، رأس شناسه متن، شناسه هر پست و رأس ویژگی‌های خاص می‌تواند تصویر پست یا هشتگ‌های استخراج‌شده از متن پست باشد.

پس مطابق تعاریف فوق داریم:

$$V = V_1 \cup V_2,$$
$$V_1 = \{D_1, ..., D_N\},$$
$$V_2 = \{\bigcup_{i=1}^{N} H_i \mid H_i = \{h_1^i, ..., h_{n_i}^i\}\}, \qquad (۵)$$

که h_j^i ویژگی $\mathbf{j}$–ام متن $\mathbf{i}$–ام و $\mathbf{n_i}$ تعداد ویژگی‌های موجود در $\mathbf{D_i}$ است. در این گراف سه نوع یال تعریف می‌شود. یال‌های شناسه–شناسه، یال‌های ویژگی–شناسه و یال‌های ویژگی–ویژگی. پس برای مجموعه یال خواهیم داشت:

$$E = E_1 \cup E_2 \cup E_3. \qquad (۶)$$

مجموعه یال‌های شناسه–شناسه $(\mathbf{E_1})$، طبق رابطه (۷) تعریف می‌شود:

$$E_1 = \{uv \mid u,v \in V_1, w(u,v) = w_{uv}\},$$
$$w_{uv} = \begin{cases} \delta - d_{uv}, & if\ d_{uv} < \sigma, \\ 0, & Otherwise. \end{cases} \quad (۷)$$

در این رابطه بین دو پارامتر δ و σ، شرط $\delta > \sigma$ برقرار است. پارامتر σ را می‌توان حد آستانه‌ای نامید که اگر فاصله بین دو متن بیش‌تر از آن باشد، بین آن‌ها یالی وصل نمی‌شود. با استفاده از پارامتر δ، معیار فاصله به وزن تبدیل می‌شود، یعنی برای هر دو شناسه‌ای که شباهت بیش‌تری به یکدیگر داشته باشند وزن بیش‌تری در نظر گرفته می‌شود.

مجموعه یال‌های شناسه–ویژگی (E_2) طبق رابطه (۸) تعریف و ویژگی‌های خاص هر متن به رأس شناسه آن با یالی با وزن ثابت c_1 متصل می‌شود.

$$E_2 = \{uv \mid u \in D_i, v \in H_i, i \in \{1,...,N\}, w(u,v) = c_1\}. \quad (۸)$$

مطابق رابطه (۹) در مجموعه یال‌های ویژگی–ویژگی (E_3)، یالی با وزن ثابت c_2 بین ویژگی‌هایی که با یکدیگر تشابه معنایی دارند در نظر گرفته می‌شود.

$$E_3 = \{uv \mid u \in H_i, v \in H_j, u,v \in H_i \cap H_j, i,j \in \{1,...,N\}, w(u,v) = c_2\}. \quad (۹)$$

پس از تشکیل گراف ارتباطی–محتوایی، الگوریتم شناسایی تشکل روی آن اعمال می‌شود:

- **شناسایی تشکل مبتنی بر گراف:** برای خوشه‌بندی می‌توان از الگوریتم‌های شناسایی تشکل روی گراف استفاده کرد که در دسته الگوریتم‌های یادگیری بدون ناظر قرار دارد.

الگوریتم اول: در این الگوریتم از الگوریتم کلاسیک حریصانه [۶] استفاده شده است. این الگوریتم مبتنی بر محاسبه پودمانگی است و می‌تواند تشکل‌ها را به صورت اکتشافی روی گراف‌های پیچیده وزن‌دار شناسایی کند. طبق این الگوریتم، برای هر رأس چندین دسته در نظر گرفته شده و سپس به صورت گام به گام تلاش می‌کند تا بهره خروجی کل را با ادغام رأس‌ها در دسته‌های یکسان افزایش دهد و این روند تا جایی ادامه می‌یابد که دیگر بهبودی حاصل نگردد. چون به ازای n رأس، پیچیدگی این الگوریتم از مرتبه $O(nlogn)$ است، این الگوریتم برای اجرا روی گراف‌های بزرگ نیز مناسب است.

الگوریتم دوم: در این الگوریتم از الگوریتم شناسایی تشکل ارائه شده در مقاله [۲۲] استفاده شده است. در این الگوریتم، ابتدا ابرگراف وزن‌دار مبتنی بر موتیف مثلثی از گراف ورودی ساخته می‌شود و سپس تعدادی از بزرگ‌ترین زیرگراف‌ها، به چند زیرگراف دیگر شکسته می‌شوند. هر یک از زیرگراف‌های تولیدی با روش تقویت یال تبدیل به یک زیرگراف کامل شده و یال‌های تقویتی وزن‌دار به گراف اصلی اضافه می‌شوند. در انتها چند روش رایج شناسایی تشکل روی گراف جدید اعمال می‌شود. نتایج ارائه‌شده در این مقاله حاکی از کیفیت بسیار مطلوب این روش روی مجموعه دادگان مختلف می‌باشد. در این روش به دلیل وجود یال‌های تقویتی، گره‌های تکین تولیدشده در مرحله شناسایی موتیف، در تشکل‌های مستقل دسته‌بندی نمی‌شوند.

۳- پیاده‌سازی

در این مقاله برای ارزیابی الگوریتم پیشنهادی، به بررسی محتوای متنی تولیدشده در شبکه اجتماعی توئیتر می‌پردازیم. برای جمع‌آوری متن توئیت‌ها از API توئیتر استفاده شده است و مجموعه داده شامل ریتوئیت نیست. توئیت‌های شهر واشینگتن دی سی (پایتخت ایالات متحده آمریکا) و حوالی آن در بازه زمانی ۱۵ تا ۲۲ اکتبر سال ۲۰۲۰ جمع‌آوری شده است که تقریبا شامل ۱۹۲۵۹۸۹ توئیت است. پس از فیلتر توئیت‌هایی که به زبان‌هایی غیر از انگلیسی ارسال شده‌اند، تعداد توئیت‌ها به ۱۵۷۳۰۶۹ تقلیل یافت.

به دلیل فقدان مجموعه داده استاندارد برای پردازش زبان طبیعی و یا شناسایی عنوان روی داده‌های استخراج‌شده از توئیتر، در این مقاله از هشتگ‌هایی که بیش‌ترین دفعات رخداد را داشته‌اند، برای شناسایی عناوین صحیح توئیت‌ها استفاده می‌شود. هم‌چنین هشتگ‌ها، ملاک ارزیابی دقت

جدول ۱: عناوین و هشتگ‌های مرتبط

عنوان	هشتگ‌های نماینده
انتخابات آمریکا	#Elections2020, #BidenHarris2020, #Trump, #vote
سلامتی-کرونا	#COVID19, #coronavirus, #WearAMask, #WashYourHands
شغل – تحصیل	#job, #writing, #medium, #books
سرگرمی-ورزشی	#game, #NowPlayingm, #TeamJeneses, #WashingtonFootball
محیط زیست	#cleanenergy, #energyefficiency, #Climate, #solar
فناوری	#CyberSecurity, #BigData, #AI, #IoT
مسائل بین المللی	#Russia, #China, #Azerbaijan, #Iran
مسئله نیجریه	#EnSARS, #EndSWAT, #Nigeria, #BuhariMustGo
انتساب قاضی دیوان عالی	#SCOTUS, #AmyConeyBarrett, #ACB, #BlockBarrett
مسئله جنبش سیاه پوستان	#BlackLivesMatter, #BLM, #racism, #Antifa

شناسایی عنوان تشکل در نظر گرفته شده است.

زمانی‌که بخواهیم از هشـــتگ برای تخمین برچســـب صحیح توئیت اســـتفاده کنیم با چالش‌هایـــی نظیر وجود توئیت‌هـــای بدون هشـــتگ، وجود توئیت‌هایـــی با بیش از یک هشـــتگ، وجـــود توئیت‌هایی با هشـــتگ‌های نامربوط و وجـــود هشـــتگ‌های متفاوت برای یـــک موضوع مواجه هســـتیم. در این‌جـــا همـــه توئیت‌های بدون هشـــتگ را به دلیل فقـــدان معیاری بـــرای ارزیابی نتیجـــه، از مجموعه داده حـــذف کـــرده و در نهایت مجموعـــه داده ورودی به ۲۰۸۷۶۳ توئیت تقلیل پیدا می‌کند. برای حذف هشـــتگ‌های بی‌ربط و یا موضوعات کم‌اهمیت، با ســـعی و خطا تعدادی از تشکل‌های شناسایی‌شده کم‌جمعیت، از مجموعه نهایی حذف می‌شـــوند. در جـــدول (۱)، ده موضوع پر اهمیت بر اســـاس پردازش هشتگ‌های پرتکرار، توسط عامل انسانی ارائه شـــده اســـت. برای هر عنوان، چهار هشتگ پرتکرار فهرست شده است.

برخلاف سایر پژوهش‌ها که عمدتاً مبتنی بر جستجو با عبارت و یا هشـــتگ، روی یک موضوع نسبتا بزرگ است، در این مقاله ماهیت مجموعه متن‌های توئیت‌شده، مربوط به توئیت‌های ارسال شده از یک شهر است. در این وضعیت بدیهی است که موضوعات متنوعی وجود داشته و برخی از موضوعات با یکدیگر هم‌پوشانی نیز داشته باشند. برای مواجه با این چالش تلاش شده است که هشتگ تکراری در عناوین مختلف وجود نداشته باشد که همین موضوع سبب کاهش دقت خواهد شـــد. مثلا می‌دانیم که با وجود شرایط کرونا، عناوین انتخابات آمریکا، شـــغل و حتی فناوری نیز با مسئله سلامت ترکیب خواهند شـــد، ولی در جدول (۱)، هشتگ کرونا فقط در عنوان سلامتی-کرونا آمده است.

در شـــکل (۱) نمودار فراوانی ۵۰ هشتگ پرتکرار ارائه شـــده اســـت. در مجموعه داده توئیت‌های هفت روز شهر واشینگتن تعداد کل هشتگ‌ها ٤٥٧١٦٠ است که از این بین تعداد هشتگ‌های یکتا ۱۰۰۶۹۳ می‌باشد. با مقایسه جدول (۱) و شـــکل (۱) مشخص است که عناوین فهرست‌شده در جدول (۱) تقریبا همه موضوعات مهم را در بر می‌گیرد.

مطابق شـــکل (۱)، هشتگ "VOTE" به سه شکل مختلف ("Vote"، "VOTE" و "vote") ظاهـــر شـــده اســـت. اما در جدول (۱) فقط پرتکرارترین شکل از این هشتگ ارائه شده است. از سوی دیگر اگر بخواهیم به ازای هر عنوان، لیست طولانی‌تری از هشتگ‌ها را داشته باشیم، با مسئله محاسبه دقت مواجه خواهیم شد. چون ممکن است برای هر عنوان تعداد متفاوتی هشـــتگ وجود داشته باشد. از این رو سعی شده است این لیست با تعداد اعضای برابر برای هر عنوان در نظر گرفته شود و این موضوع منجر به نبود هشتگ‌های تکراری در لیست عناوین می‌شود.

پس از شناســـایی عناوین و هشتگ‌های مرتبط، اولین گام در پیاده‌سازی، مرحلهٔ پیش‌پردازش است. در این مقاله مجموعه داده به صورت پنجره‌بندی‌شده روزانه پردازش

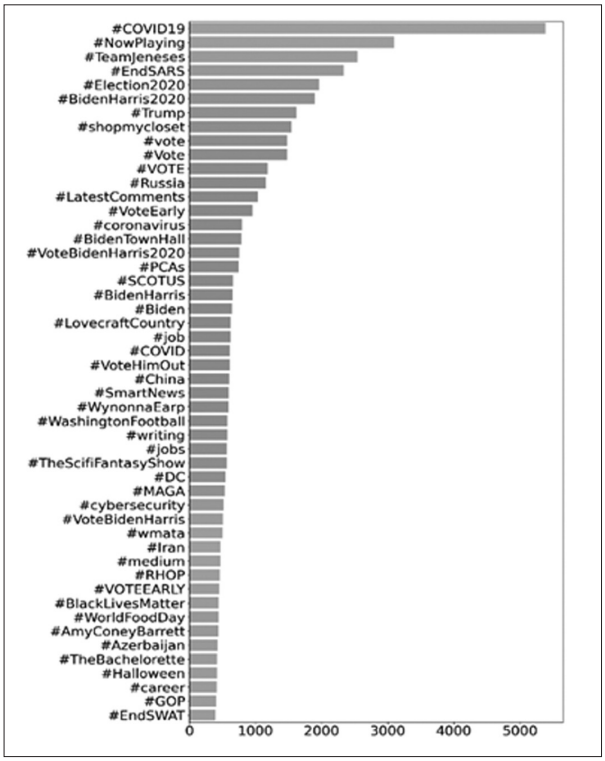
می‌شـــود و به ازای هر روز، دقت الگوریتم پیشـــنهادی و سایرین محاسبه و گزارش می‌شود.

توئیت‌ها بـــه صورت روزانه وارد مرحله پیش‌پردازش شـــده و با محاســـبه بردار ویژگـــی برای همـــه توئیت‌ها $(F(D_i), i = 1,...,208763)$، ماتریـــس فاصله با فرمول (۵) محاســـبه می‌شـــود. در این مقاله ویژگی خـــاص هر متن (توئیت)، هشتگ‌هایی است که در همان توئیت استفاده شده اســـت. پس با در نظر گرفتن موارد فـــوق گراف ارتباطات با مجموعه رئوس V_1 که شـــامل شناســـه توئیت‌ها است و مجموعه رئوس V_2 که شـــامل هشـــتگ‌های موجود در مجموعه داده است، ساخته می‌شود.

مجموعـــه یال‌های E_1 بر اســـاس ماتریـــس فاصله و پارامترهـــای δ و σ، ایجاد می‌شـــوند. بـــه ازای داده‌های هـــر روز، تابع توزیع ماتریس فاصله محاســـبه و پارامتر حد آســـتانه‌ای بر اســـاس آن تنظیم می‌گـــردد. مقدار σ_t به‌عنوان پارامتر حد آســـتانه‌ای داده‌های روز t-ام برابر با $\sigma_t = \frac{M_t}{r}$ قرار داده می‌شـــود که M_t میانه تابع توزیع روی ماتریس فاصله روز t-ام اســـت. بـــا در نظر گرفتن مقدار r، فقط بین متن‌هایی که شـــباهت چشـــم‌گیری وجود داشته باشـــد، یال وزن‌دار اضافه می‌شـــود. اگر مقدار پارامتر r زیاد باشـــد، تعداد یال‌های گراف کم و اگـــر این مقدار کم باشد، تعداد بیش‌تری یال در گراف حضور خواهند داشت. هم‌چنین به ازای تمام اعضای مجموعه داده $\delta = 2$ است.

مجموعـــه یال‌های E_2 مربوط به یال‌های بین هر توئیت و هشتگ‌هایی است که در همان توئیت ظاهر شده است که مقدار وزن آن‌هـــا را ثابت $c_1 = 1$ قرار می‌دهیم. با توجه به مقادیر δ و c_1، عمدتـــاً اهمیت یال‌های بین توئیت و توئیت بیش‌تر از یال‌های بین توئیت و هشـــتگ است. زیرا بیشینه مقدار وزن یال بین توئیت و توئیت تقریبا دو برابر وزن یال بین توئیت و هشتگ است. علاوه بر این، در این مقاله برای جلوگیـــری از پیچیدگی گراف، مجموعـــه یال‌های E_3 را در نظر نگرفته و در واقع $c_2 = 0$ را مقداردهی می‌کنیم.

در الگوریتم اول، با الگوریتم بیشینه‌ســـازی حریصانه



شکل ۱: نمودار فراوانی ۵۰ هشتگ پرتکرار شهر واشینگتن در هفت روز متوالی

ماژولاریتـــی و در الگوریتـــم دوم، بـــا الگوریتـــم EdMot روی گراف ساخته‌شده، تشـــکل‌ها شناسایی می‌شود. در این الگوریتم‌ها، ۹۰ درصد تشـــکل‌های پرجمعیت به‌عنوان خروجی صحیح و ۱۰ درصد به‌عنوان اختلال دســـته‌بندی می‌شوند.

۴- ارزیابی و مقایسه

در ایـــن بخش بـــرای مقایســـه کیفیـــت الگوریتم‌های پیشـــنهادی، الگوریتم‌های کلاســـیک K-means و LDA نیز روی مجموعه داده ورودی اجرا شـــده و نتایج به دســـت آمده با نتیجه حاصل از الگوریتم‌های پیشـــنهادی مقایسه می‌شوند.

بـــرای ارزیابی کیفیـــت الگوریتم‌هـــا از معیار اطلاعات متقابل استفاده شـــده است. اطلاعات متقابل معیاری است که تقابل بین دو دســـته متغیر تصادفی گسسته را ارزیابی می‌کند و بـــرای دو متغیر تصادفی X و Y با توزیع احتمال ادغامی $p(X,Y)$، $MI(X,Y)$ به صورت زیر محاسبه می‌شود:

$$MI(X,Y)=\sum_{x\in X}\sum_{y\in Y}\log\left(\frac{p(x,y)}{p(x)p(y)}\right). \quad (۹)$$

برای نگاشت مقدار اطلاعات متقابل به بازه ۰ تا ۱، معمولاً از معیار متقابل نرمال‌شده مطابق رابطه زیر استفاده می‌شود:

$$NMI(X,Y)=\frac{MI(X,Y)}{\sqrt{H(X)H(Y)}}, \quad (۱۰)$$

که در آن $H(X)$ و $H(Y)$ تعداد اعضای هر یک از مجموعه‌های X و Y می‌باشد.

در این مقاله، با توجه به معیارهای معرفی‌شده، به ازای روز t–ام، مقدار $MNMI(Date_t)$ مطابق رابطه (۱۱) محاسبه می‌شود:

$$MNMI(Date_t)=\frac{1}{|C_t|}\sum_{i=1}^{|C_t|}\max_{Y_j\in Y}(NMI(X_i^t,Y_j)). \quad (۱۱)$$

از این رابطه به‌عنوان معیاری برای محاسبه دقت الگوریتم‌ها استفاده می‌شود که در آن، Y مجموعه عناوین موجود در جدول (۱)، Y_j مجموعه هشتگ‌های عنوان j–ام و X_i^t مجموعه هشتگ‌های نماینده دسته i–ام از داده‌های مربوط به روز t–ام است. هم‌چنین اندازه C_t تعداد دسته‌های روز t–ام را نشان می‌دهد.

الگوریتم ارائه‌شده در زبان پایتون و روی سخت‌افزاری با ۱۶ گیگابایت حافظه پیاده‌سازی شده است. برای اندازه‌گیری کیفیت هر دسته، چهار هشتگ با بیش‌ترین درجه، به‌عنوان نماینده هر دسته انتخاب شده‌اند.

برای اجرای الگوریتم K-means از کتابخانه scikit-learn در پایتون استفاده شده و تعداد خوشه‌ها به‌عنوان دانش اولیه برای هر مجموعه داده، در آن اعمال شده است. به منظور رعایت شرایط یکسان، بردار ویژگی متن توئیت‌ها به‌عنوان ورودی الگوریتم K-means در نظر گرفته شده است. برای اندازه گیری کیفیت K-means به ازای هر خوشه، چهار هشتگی که بیش‌ترین دفعات تکرار را داشته‌اند، به‌عنوان نماینده هر دسته انتخاب می‌شود.

برای اجرای الگوریتم LDA از کتابخانه gensim در زبان پایتون استفاده شده است. در این الگوریتم تعداد دسته‌ها مشابه تعداد خوشه‌ها در K-means مقداردهی می‌شود. در الگوریتم LDA چهار کلمه با بیش‌ترین آمارگان به‌عنوان نماینده هر دسته انتخاب و توسط عامل انسانی رابطه (۱۱) برای آن محاسبه می‌شود.

جدول ۲: مقادیر MNMI به‌دست‌آمده برای الگوریتم‌های پیشنهادی اول و دوم به ازای $r=4$، الگوریتم K-means و الگوریتم LDA به ازای هفت روز مجموعه داده مربوط به متن توئیت‌های ارسالی از شهر واشینگتن

روز	الگوریتم پیشنهادی دوم	الگوریتم پیشنهادی اول	K-means	LDA
2020-10-15	0.95	0.754	0.466	0.500
2020-10-16	0.92	0.704	0.616	0.440
2020-10-17	0.90	0.609	0.583	0.360
2020-10-18	0.89	0.791	0.433	0.400
2020-10-19	0.91	0.462	0.433	0.300
2020-10-20	0.89	0.625	0.516	0.280
2020-10-21	0.93	0.634	0.400	0.240

در جدول (۲) مقادیر MNMI به‌دست‌آمده برای الگوریتم‌های پیشنهادی به ازای پارامتر $r=4$، الگوریتم K-means و الگوریتم LDA به ازای هفت روز مجموعه داده مربوط به متن توئیت‌های ارسال‌شده از شهر واشینگتن گزارش شده است.

مطابق جدول (۲)، الگوریتم پیشنهادی دوم نسبت به سایرین از دقت مناسب‌تری برخوردار است. الگوریتم پیشنهادی اول هم با وجود استفاده از الگوریتم سنتی شناسایی تشکل روی گراف، نسبت به الگوریتم‌های K-means و LDA از دقت بیش‌تری برخوردار است. یکی از دلایل تفاوت زیاد بین نتایج الگوریتم K-means و LDA، استفاده از بردار ویژگی متن هر توئیت است که با این کار عملاً توئیت‌ها در فضای ۳۰۰ بعدی بر حسب ویژگی‌های آن‌ها نگاشت شده و روی آن الگوریتم K-means اجرا شده است.

در جدول (۳) تاثیر پارامتر r در الگوریتم پیشنهادی دوم مورد بررسی قرار گرفته است. در این جدول معیار MNMI به ازای سه مقدار $\sigma_t=\frac{M_t}{4}$، $\sigma_t=\frac{M_t}{3}$ و $\sigma_t=\frac{M_t}{2}$ برای هفت روز داده‌های شهر واشنگتن محاسبه شده است.

مطابق جدول (۳)، با افزایش مقدار پارامتر r دقت افزایش

جدول ۳: مقادیر MNMI به‌دست‌آمده برای الگوریتم‌های پیشنهادی دوم به ازای $r = 2$، $r = 3$ و $r = 4$، به ازای هفت روز مجموعه داده مربوط به متن توئیت‌های ارسالی از شهر واشینگتن

روز	$M_{t}/_{4}$	$M_{t}/_{3}$	$M_{t}/_{2}$
2020-10-15	0.95	0.96	0.86
2020-10-16	0.92	0.9	0.74
2020-10-17	0.90	0.8	0.58
2020-10-18	0.89	0.74	0.47
2020-10-19	0.91	0.9	0.64
2020-10-20	0.89	0.86	0.62
2020-10-21	0.93	0.82	0.68

یافته است. در روش ارائه‌شده در این مقاله، برای هر دو رأس از نوع توئیت، یالی وزن‌دار در نظر گرفته شده و مقدار این وزن بر اساس تعداد کلمات مشابه استفاده‌شده در دو رأس موردنظر محاسبه می‌شود. می‌دانیم کلماتی هستند که از نوع کلمات رایج و پرکاربرد حوزه‌های مختلف به شمار می‌روند و در بسیاری از توئیت‌ها استفاده شده‌اند پس اگر تعداد یال‌ها با کاهش r، افزایش یابد این حوزه‌های مختلف با یکدیگر ارتباط ضعیفی خواهند داشت. از سوی دیگر در الگوریتم دوم، زمانی‌که ابرگراف مبتنی بر موتیف ساخته می‌شود به یال‌ها وزن اختصاص داده می‌شود و عملا وزن یال‌های کم‌اهمیت افزایش می‌یابد و زمانی‌که تقسیم‌بندی نهایی انجام می‌شود، عملاً تشکل‌هایی که با یکدیگر ارتباط معنایی اندکی دارند هم در یک تشکل دسته‌بندی می‌شوند و همین موضوع سبب کاهش دقت می‌شود.

۵- نتیجه‌گیری

در این مقاله الگوریتم جدیدی برای شناسایی عناوین محتوای متنی تولیدشده در شبکه‌های اجتماعی ارائه شد. الگوریتم ارائه‌شده مبتنی بر شناسایی تشکل‌های روی گراف است. برای ساخت گراف، محتوای متن‌های تولیدشده (رئوس نوع اول) به‌عنوان معیار وزن یال‌های نوع اول و سایر ویژگی‌ها به‌عنوان رئوسی از نوع دوم در نظر گرفته شد. در نهایت از دو الگوریتم برای شناسایی تشکل استفاده شد.

شبکه اجتماعی توئیتر برای اجرای الگوریتم پیشنهادی انتخاب و ویژگی خاص متن هر توئیت، هشتگ‌های متن‌ها قرار داده شد. هم‌چنین برای محاسبه وزن یال بین توئیت‌ها از الگوریتم ازپیش‌آموزش‌دیده fastText استفاده گردید. توئیت‌هایی که مورد بررسی قرار گرفتند، مربوط به هفت روز محتوای منتشرشده ساکنان شهر واشینگتن بوده و فقط محتواهای با زبان انگلیسی در این مقاله مورد بررسی قرار گرفت.

با مقایسه الگوریتم‌های پیشنهادی با الگوریتم‌های K-means و LDA، نشان داده شد که با وجود استفاده از الگوریتم کلاسیک شناسایی تشکل حریصانه در الگوریتم پیشنهادی اول، دقت این روش بهتر از الگوریتم‌های K-means و LDA است. هم‌چنین الگوریتم پیشنهادی دوم نسبت به الگوریتم پیشنهادی اول و الگوریتم‌های کلاسیک از دقت بسیار بهتری برخوردار است که این موضوع نشان‌دهنده تأثیر بسیار زیاد الگوریتم شناسایی تشکل در این الگوریتم است.

به دلیل استفاده از هشتگ به‌عنوان برچسب هر دسته، در این مقاله صرفاً از توئیت‌هایی که دارای هشتگ باشند استفاده شد. با توجه به دقت بسیار مناسب الگوریتم دوم، به ازای هر تشکل می‌توان چند هشتگ پر درجه را به‌عنوان موضوع آن تشکل در نظر گرفت و این‌گونه شناسایی عناوین را انجام داد.

در آینده سعی داریم با ترکیب روش ارائه‌شده و الگوریتم‌های موجود در حوزه یادگیری شبه‌نظارتی [11]، توئیت‌هایی که در آن از هشتگ استفاده نشده باشد را نیز با دقت مناسبی دسته‌بندی کنیم. هم‌چنین با توجه به اهمیت داده، می‌توان از روش‌های موجود در حوزه یادگیری فعال برای تعیین داده‌های ورودی نیز بهره گرفت.

مراجع

1. BigData. Available online: https://www.whishworks.com/blog/big-data/understanding-the-3-vs-of-big-data-volume-velocity-and-variety
2. Guille, A., Hacid, H., Favre, C., Zighed, D.A., "Information diffusion in online social networks: A survey" ACM

11- Semi-Supervised Learning

Sigmod Record, 42(2), 17-28, 2013.
3. Bisht, S. Paul, A., "Document clustering: a review", International Journal of Computer Applications, 73(11), 2013.
4. Naik, M.P., Prajapati, H.B., Dabhi, V.K., "A survey on semantic document clustering", IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 1-10, 2015.
5. Tsur, O., Littman, A., Rappoport, A., "Efficient clustering of short messages into general domains". In Icwsm, 2013.
6. Ramos, J.,"Using tf-idf to determine word relevance in document queries" Proceedings of the first instructional conference on machine learning, 242, 2003.
7. Mikolov, T., Chen, K., Corrado, G., Dean, J., "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781, 2013.
8. Irfan, R., King, C.K., Grages, D., Ewen, S., Khan, S.U., Madani, S.A., Kolodziej, J., Wang, L., Chen, D., Rayes, A., Tziritas, N., "A survey on text mining in social networks", The Knowledge Engineering Review, 30(2), 157-170, 2015.
9. Chinnov, A., Kerschke, P., Meske, C., Stieglitz, S., Trautmann, H., "An Overview of Topic Discovery in Twitter Communication through Social Media Analytics", 2015.
10. Alghamdi, R., Alfalqi, K., "A survey of topic modeling in text mining", Int. J. Adv. Comput. Sci. Appl. (IJACSA), 6(1), 2015.
11. Steinskog, A., Therkelsen, J., Gambäck, B., "Twitter topic modeling by tweet aggregation", In Proceedings of the 21st nordic conference on computational linguistics, 77-86, 2017.
12. Stieglitz, S., Mirbabaie, M., Ross, B. and Neuberger, C.,"Social media analytics–Challenges in topic discovery, data collection, and data preparation", International journal of information management, 39, 156-168, 2018.
13. Alnajran, N., Crockett, K., McLean, D., Latham, A., "Cluster analysis of twitter data: A review of algorithms", Proceedings of the 9th international conference on agents and artificial intelligence – volume 2: ICAART, INSTICC. SciTePress, 239–249, 2017.
14. Rosa, K.D., Shah, R., Lin, B., Gershman, A., Frederking, R. "Topical clustering of tweets", In Proceedings of the ACM SIGIR: SWSM, Beijing, China, 2011.
15. Miranda D., Raiol G., Pasti, R., Nunes de Castro, L., "Detecting Topics in Documents by Clustering Word Vectors", International Symposium on Distributed Computing and Artificial Intelligence, 2019.
16. Curiskis, S., Drake, B., Osborn, T. R., Kennedy, P. J., "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit", Information Processing & Management, 57.2, 2020.
17. Mendonça, I., Trouvé, A., Fukuda, A., Murakami, KJ., Tsai, CF., Hu, YH., Wang, MC., Liu, KE., Yu, X., Yuan, Y., Chung, YM, "On Clustering Algorithms: Applications in Word-Embedding Documents", JCP, 14.2, 88-92, 2019.
18. Hashimoto, T., Shepard, D.L., Kuboyama, T., Shin, K., Kobayashi, R. Uno, T., "Analyzing temporal patterns of topic diversity using graph clustering", The Journal of Supercomputing, 1-14, 2020.
19. Beniwal, A., Roy, G., Bhavani, S.D., "Text Document Clustering Using Community Discovery Approach", International Conference on Distributed Computing and Internet Technology, 336-346, 2020.
20. Majdabadi, Z., Sabeti, B., Golazizian, P., Asli, S., Momenzadeh, A.A., "Twitter Trend Extraction: A Graph-based Approach for Tweet and Hashtag Ranking, Utilizing No-Hashtag Tweets", Proceedings of the 12th Language Resources and Evaluation Conference. 2020.
21. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H. and Mikolov, T., "fastText. zip: Compressing text classification models", arXiv preprint arXiv:1612.03651, 2016.
22. Li, P.Z., Huang, L., Wang, C.D. and Lai, J.H., "Edmot: An edge enhancement approach for motif-aware community detection", In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 479-487. 2019.