

## شناسایی عناوین محتوای متنی منتشر شده در شبکه اجتماعی توئیتر

فاطمه بهاری فرد\*

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

نرمال‌شده، مقایسه شده است که نشان‌دهنده دقت مناسب الگوریتم پیشنهادی است.  
**واژه‌های کلیدی:** شناسایی عنوان، گراف وزن‌دار، یادگیری بدون ناظر، شبکه‌های اجتماعی، توئیتر

### ۱- مقدمه

طی سالیان اخیر استفاده از شبکه‌های اجتماعی در بین عموم جامعه به شدت مورد اقبال قرار گرفته است. در این شبکه‌ها حجم عظیمی از داده‌ها با سرعت و تنوع زیاد (به‌عنوان مثال تصاویر، متن و فیلم) تولید و منتشر می‌شود. روزانه ۵۰۰ میلیون توئیتر در توئیتر (هر توئیتر حداکثر ۲۸۰ حرف) ارسال و ۹۰۰ میلیون عکس در فیس‌بوک و ۴/۰ میلیون ساعت فیلم در یوتیوب بارگذاری می‌شود [۱]. کاربرد این شبکه‌ها از برقراری ارتباط بین افراد تا کسب اطلاعات و اخبار در حوزه‌های وسیعی از مسائل سیاسی، اجتماعی، فرهنگی و غیره گسترده شده است. با این حال بسیاری از داده‌های تولیدشده در این بسترها، ساختارمند نیستند. پس نمی‌توان به سادگی در مورد چگونگی انتشار موضوعات مختلف و همچنین تعامل

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

در این مقاله، الگوریتم جدیدی برای دسته‌بندی محتوای متنی شبکه اجتماعی توئیتر ارائه شده است. در این الگوریتم ابتدا هر متن، پیش‌پردازش شده و سپس یک گراف ارتباطات جدید مبتنی بر محتوای متن‌های منتشرشده ساخته می‌شود. این گراف وزن‌دار و بی‌جهت است و روی آن با استفاده از دو روش بدون ناظر، تشکلهای مختلف شناسایی می‌شوند. برای ارزیابی، داده‌های متنی ارسال‌شده از شهر واشینگتن در یک بازه زمانی، با API جمع‌آوری و الگوریتم‌های ارائه‌شده روی آن اعمال شده است. برای بررسی دقت، نتایج حاصل با دو الگوریتم کلاسیک K-means و LDA بر اساس معیار اطلاعات متقابل

\* نویسنده مسئول

کاربران با یکدیگر اطلاعاتی به دست آورد [۲].

در تکنیک‌های خوشه‌بندی اسناد F0<sup>۱</sup>، 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، ۴]. 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 F1<sup>۲</sup> 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 [۶]، و کلمه-به-بردار F2<sup>۳</sup> که در آن هر کلمه با توجه به معنایش به بردار تبدیل می‌شود [۷]. سپس الگوریتم‌های یادگیری ماشین بدون ناظر<sup>۴</sup> خوشه‌بندی سلسله مراتبی و تقسیمی مانند K-means روی آن اعمال می‌شود [۸].

مدل‌سازی عناوین F3<sup>۵</sup> 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 به عنوان کلمات پیوندی تعریف شده توسط کاربر استفاده شده است. این عبارات می‌توانند به عنوان 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 منتشر شده در شبکه‌های اجتماعی است. این محتواها معمولاً دارای اختلال

F4<sup>۶</sup> (غلط املائی، 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. به عنوان مثال توئیتر فقط مجوز پخش شناسه هر توئیتر را مجاز می‌داند و افراد علاقه‌مند می‌توانند با API به متن هر توئیتر 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. به عنوان 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 انجام شده 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<sup>۷</sup> می‌شود. در ادامه برخی از پژوهش‌های انجام شده در این دو دسته آمده است.

در مرجع [۱۴]، نویسندگان برای شناسایی عناوین توئیترها، الگوریتم K-means را روی بردارهای TF-IDF محاسبه شده 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 خوشه‌بندی روی کلمات اعمال شده و نشان داده شده است که کلمات مشابه در خوشه‌های یکسان دسته‌بندی می‌شوند. در مقاله [۱۶] برای خوشه‌بندی متن‌های منتشرشده در توئیتر و ردیت، چهار روش نمایش سند TF-IDF، کلمه-به-بردار، کلمه-به-بردار وزن‌دار و همچنین سند-به-بردار به همراه الگوریتم‌های خوشه‌بندی K-means، K-median، 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 شناسایی عناوین توئیتهایی که پس از زلزله ژاپن در سال ۲۰۱۱ منتشر شده است، با پنجره‌بندی توئیتهای بر اساس 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 نشان دادند که عملکرد الگوریتم ارایه شده بهتر از LDA و K-means است. در مرجع [۱۹]، کلمات کلیدی هر سند با روش کیسه کلمات شناسایی و با آن گرافی شکل می‌گیرد. در این گراف اگر دو کلمه متفاوت در یک سند موجود باشند، بین آنها یال در نظر گرفته می‌شود. در نهایت از الگوریتم شناسایی تشکل Louvain استفاده و تشکل‌هایی که معیار مرکزیت آنها کمتر از حد آستانه‌ای باشد با یکدیگر ادغام می‌شوند. نتایج این پژوهش با K-means و الگوریتم‌های خوشه‌بندی سلسله مراتبی مقایسه و عملکرد خوب آن گزارش شده است. همچنین در مقاله [۲۰] برای شناسایی عناوین توئیتهای از الگوریتمی مبتنی بر گراف

استفاده شده است. هر توئیت با سه نوع متفاوت از رأس‌ها در یک گراف نمایش داده می‌شود: توئیت، هشتک و کلمات موجود در هر توئیت. بین رأس‌های کلمات و توئیت، بین رأس‌های هشتک و توئیت و همچنین بین کلمات و هشتک، یال وزن‌دار ایجاد می‌شود. وزن یال‌های هشتک-توئیت بیشینه مقدار یک، وزن یال‌های توئیت-کلمه مبتنی بر TF-IDF و وزن یال‌های کلمه-هشتک به میزان دفعاتی که با یکدیگر تکرار شده‌اند، در نظر گرفته شده است. در نهایت روی این گراف، خوشه‌بندی با الگوریتم RankClus انجام و برخی از خوشه‌ها به عنوان اختلال در نظر گرفته شده و حذف شده‌اند. 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 منتشرشده در شبکه اجتماعی توئیتر ارائه می‌دهیم که در آن متن هر محتوا ابتدا در یک بردار تعبیه شده و در ادامه گراف ارتباطات مربوط به آن ساخته می‌شود. در انتها با شناسایی تشکل روی گراف، تلاش می‌شود برای هر دسته عنوانی در نظر گرفته شود. بر این مبنا دو الگوریتم متفاوت شناسایی تشکل مورد استفاده قرار گرفته است و هر یک از دو الگوریتم پیشنهادی مبتنی بر یکی از الگوریتم‌های شناسایی تشکل می‌باشد. در پایان دقت الگوریتم‌های پیشنهادی با الگوریتم‌های مبنایی در این حوزه با معیار اطلاعات متقابل نرمال‌شده  $F9^A$  مقایسه می‌شود.

## ۲- الگوریتم پیشنهادی

در این بخش الگوریتم پیشنهادی برای تعیین عنوان یا موضوع متن منتشرشده در توئیتر ارائه می‌شود. برای این منظور باید متن‌ها دسته‌بندی و سپس عنوان هر دسته شناسایی شود. 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

(۴) اجرای الگوریتم شناسایی تشکل روی گراف حاصل  
فرض کنید ورودی الگوریتم یک مجموعه  $N$  عضوی از متن‌های منتشرشده باشد. متن‌های منتشرشده در شبکه‌های اجتماعی معمولاً کوتاه، حاوی عبارات عامیانه، اشتباهات نگارشی و املائی می‌باشند. پس باید در ابتدا روی همه متن‌ها عملیات پیش‌پردازش انجام شود.

● **پیش‌پردازش:** در این مرحله، هر یک از اعضای مجموعه متن‌های ورودی، کلمه‌بندی  $F_6$  و سپس علامت‌های نگارشی، کلمات توقف  $F_7$  و پیوندها از بین آن‌ها حذف می‌شوند. پس به ازای متن  $n$ -ام از ورودی، یک بردار کلمات خواهیم داشت:

$$D_n = [T_1^n, \dots, T_{m_n}^n] \quad n \in \{1, \dots, N\} \quad (1)$$

که  $m_n$  تعداد کلمات حاصل از پیش‌پردازش روی متن  $n$ -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-برداری از پیش‌آموزش‌دیده fastText مبتنی بر دو میلیون کلمه [۲۱]، به فضای ۳۰۰ بعدی نگاشت می‌شود. یعنی به ازای کلمه  $i$ -ام از متن  $n$ -ام، مطابق رابطه (۲) برداری محاسبه می‌شود که در آن  $d = 300$  است:

$$fastText(T_i^n) = [x_{i1}^n, \dots, x_{id}^n] \quad i \in \{1, \dots, m_n\} \quad (2)$$

در ادامه برای هر متن، میانگین بردارهای  $d$  بعدی کلمات همان متن، به عنوان بردار ویژگی آن لحاظ می‌شود.

یعنی اگر داشته باشیم:

$$F_j^n = \frac{\sum_{i=1}^{m_n} x_{ij}^n}{m_n}, \quad j \in \{1, \dots, d\} \quad (3)$$

برداری ویژگی متن  $n$ -ام به صورت زیر است:

$$F(D_n) = [F_1^n, \dots, F_d^n] \quad n \in \{1, \dots, N\} \quad (4)$$

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

$$d_{ij} = \|F(D_i) - F(D_j)\|_2, \quad i, j \in \{1, \dots, N\} \quad (5)$$

ماتریس  $d_{ij}$  مربعی، متقارن و با اندازه  $N \times N$  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  $G = (V, E)$  با مجموعه رئوس  $V$  و مجموعه یال‌های  $E$  ساخته می‌شود. مجموعه رئوس، شامل دو نوع رأس شناسه  $(V_1)$  و ویژگی خاص متن  $(V_2)$  است. به عنوان مثال در فیس‌بوک، رأس شناسه متن، شناسه هر پست و رأس ویژگی‌های خاص می‌تواند تصویر پست یا هشتک‌های استخراج‌شده از متن پست باشد.

پس مطابق تعاریف فوق داریم:

$$\begin{aligned} V &= V_1 \cup V_2, \\ V_1 &= \{D_1, \dots, D_N\}, \\ V_2 &= \left\{ \bigcup_{i=1}^N H_i \mid H_i = \{h_{i1}^1, \dots, h_{i n_i}^1\} \right\}, \end{aligned} \quad (5)$$

که  $h_j^i$  ویژگی  $i$ -ام متن  $i$ -ام و  $n_i$  تعداد ویژگی‌های موجود در  $D_i$  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:

$$E = E_1 \cup E_2 \cup E_3. \quad (6)$$

مجموعه یال‌های شناسه-شناسه  $(E_1)$ ، طبق رابطه (۷)

تعریف می‌شود:

$$E_1 = \{uv \mid u, v \in V_1, w(u, v) = w_{uv}\}, \quad (7)$$

$$w_{uv} = \begin{cases} \delta - d_{uv}, & \text{if } d_{uv} < \sigma, \\ 0, & \text{Otherwise.} \end{cases}$$

در این رابطه بین دو پارامتر  $\delta$  و  $\sigma$ ، شرط  $\delta > \sigma$  برقرار است. پارامتر  $\sigma$  را می‌توان حد آستانه‌ای نامید که اگر فاصله بین دو متن بیش‌تر از آن باشد، بین آن‌ها یالی وصل نمی‌شود. با استفاده از پارامتر  $\delta$ ، 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.

$$E_2 = \{uv \mid u \in D_i, v \in H_i, i \in \{1, \dots, N\}, w(u, v) = c_1\}. \quad (8)$$

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.

$$E_3 = \{uv \mid u \in H_i, v \in H_j, u, v \in H_i \cap H_j, i, j \in \{1, \dots, N\}, w(u, v) = c_2\}. \quad (9)$$

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 F8 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  $O(n \log n)$  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 توییتر می‌پردازیم. برای جمع‌آوری متن توییتهای از API توییتر استفاده شده است و مجموعه داده شامل ریتوئیت نیست. توییتهای شهر واشینگتن دی سی (پایتخت ایالات متحده آمریکا) و حوالی آن در بازه زمانی ۱۵ تا ۲۲ اکتبر سال ۲۰۲۰ جمع‌آوری شده است که تقریباً شامل ۱۹۲۵۹۸۹ توییتهای است. پس از فیلتر توییتهایی که به زبان‌هایی غیر از انگلیسی ارسال شده‌اند، تعداد توییتهای به ۱۵۷۳۰۶۹ تقلیل یافت.

به دلیل فقدان مجموعه داده استاندارد برای پردازش زبان طبیعی و یا شناسایی عنوان روی داده‌های استخراج‌شده از توییتر، در این مقاله از هشتگ‌هایی که بیش‌ترین دفعات رخداد را داشته‌اند، برای شناسایی عناوین صحیح توییتهای

جدول ۱: عناوین و هشتک‌های مرتبط

| عنوان                  | هشتک‌های نماینده                                        |
|------------------------|---------------------------------------------------------|
| انتخابات آمریکا        | #Elections2020, #BidenHarris2020, #Trump, #vote         |
| سلامتی-کرونا           | #COVID19, #coronavirus, #WearAMask, #WashYourHands      |
| شغل - تحصیل            | #job, #writing, #medium, #books                         |
| سرگرمی-ورزشی           | #game, #NowPlayingm, #Team-Jeneses, #WashingtonFootball |
| محیط زیست              | #cleanenergy, #energyefficiency, #Climate, #solar       |
| فناوری                 | #CyberSecurity, #BigData, #AI, #IoT                     |
| مسائل بین‌المللی       | #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               |

اساس پردازش هشتک‌های پرتکرار، توسط عامل انسانی ارائه شده است. برای هر عنوان، چهار هشتک پرتکرار فهرست شده است.

برخلاف سایر پژوهش‌ها که عمدتاً مبتنی بر جستجو با عبارت و یا هشتک، روی یک موضوع نسبتاً 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 مواجهه با این چالش تلاش شده است که هشتک تکراری در عناوین مختلف وجود نداشته باشد که همین موضوع سبب کاهش دقت خواهد شد. مثلاً می‌دانیم که با وجود شرایط کرونا، عناوین انتخابات آمریکا، شغل و حتی فناوری نیز با مسئله سلامت ترکیب خواهند شد، ولی در جدول (۱)، هشتک کرونا فقط در عنوان سلامتی-کرونا آمده است.

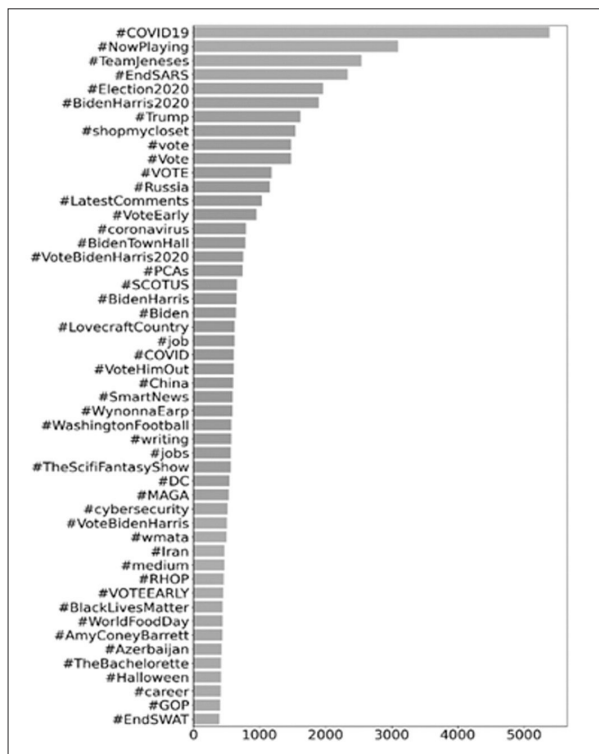
در شکل (۱) نمودار فراوانی ۵۰ هشتک پرتکرار ارائه شده است. در مجموعه داده توئیت‌های هفت روز شهر واشینگتن تعداد کل هشتک‌ها ۴۵۷۱۶۰ است که از این بین تعداد هشتک‌های یکتا ۱۰۰۶۹۳ می‌باشد. با مقایسه جدول (۱) و شکل (۱) مشخص است که عناوین فهرست‌شده در جدول (۱) تقریباً همه موضوعات مهم را در بر می‌گیرد.

مطابق شکل (۱)، هشتک "VOTE" به سه شکل مختلف ("Vote"، "VOTE"، و "vote") ظاهر شده است. اما در جدول (۱) فقط پرتکرارترین شکل از این هشتک ارائه شده است. از سوی دیگر اگر بخواهیم به ازای هر عنوان، لیست طولانی‌تری از هشتک‌ها را داشته باشیم، با مسئله محاسبه دقت مواجه خواهیم شد. چون ممکن است برای هر عنوان تعداد متفاوتی هشتک 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 هشتک‌های تکراری در لیست عناوین می‌شود.

پس از شناسایی عناوین و هشتک‌های مرتبط، اولین گام در پیاده‌سازی، مرحله پیش‌پردازش است. در این مقاله

استفاده می‌شود. همچنین هشتک‌ها، ملاک ارزیابی دقت شناسایی عنوان تشکل در نظر گرفته شده است.

زمانی که بخواهیم از هشتک برای تخمین برچسب صحیح توئیت استفاده کنیم با چالش‌هایی نظیر وجود توئیت‌های بدون هشتک، وجود توئیت‌هایی با بیش از یک هشتک، وجود توئیت‌هایی با هشتک‌های نامربوط و وجود هشتک‌های متفاوت برای یک موضوع مواجه هستیم. در این‌جا همه توئیت‌های بدون هشتک را به دلیل فقدان معیاری برای ارزیابی نتیجه، از مجموعه داده حذف کرده و در نهایت مجموعه داده ورودی به ۲۰۸۷۶۳ توئیت تقلیل پیدا می‌کند. برای حذف هشتک‌های بی‌ربط و یا موضوعات کم‌اهمیت، با سعی و خطا تعدادی از تشکل‌های شناسایی‌شده کم‌جمعیت، از مجموعه نهایی حذف می‌شوند. در جدول (۱)، ده موضوع پر اهمیت بر



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

در الگوریتم اول، با الگوریتم پیشینه‌سازی حریصانه ماژولاریتی و در الگوریتم دوم، با الگوریتم EdMot روی گراف ساخته‌شده، تشکلهای شناسایی می‌شود. در این الگوریتم‌ها، ۹۰ درصد تشکلهای پرجمعیت به‌عنوان خروجی صحیح و ۱۰ درصد به‌عنوان اختلال دسته‌بندی می‌شوند.

#### ۴- ارزیابی و مقایسه

در این بخش برای مقایسه کیفیت الگوریتم‌های پیشنهادی، الگوریتم‌های کلاسیک K-means و LDA نیز روی مجموعه داده ورودی اجرا شده و نتایج به دست آمده با نتیجه حاصل از الگوریتم‌های پیشنهادی مقایسه می‌شوند.

برای ارزیابی کیفیت الگوریتم‌ها از معیار اطلاعات متقابل F9 استفاده شده است. اطلاعات متقابل معیاری است که تقابل بین دو دسته متغیر تصادفی گسسته را ارزیابی می‌کند و برای دو متغیر تصادفی  $X$  و  $Y$  با توزیع احتمال

مجموعه داده به صورت پنجره‌بندی‌شده روزانه پردازش می‌شود و به ازای هر روز، دقت الگوریتم پیشنهادی و سایرین محاسبه و گزارش می‌شود.

توئیت‌ها به صورت روزانه وارد مرحله پیش‌پردازش شده و با محاسبه بردار ویژگی برای همه توئیت‌ها  $(F(D_i) \quad i = 1, \dots, 208763)$ ، ماتریس فاصله با فرمول (۵) محاسبه می‌شود. در این مقاله ویژگی خاص هر متن (توئیت)، هشتگ‌هایی است که در همان توئیت استفاده شده است. پس با در نظر گرفتن موارد فوق گراف ارتباطات با مجموعه رئوس  $V_1$  که شامل شناسه توئیت‌ها است و مجموعه رئوس  $V_2$  که شامل هشتگ‌های موجود در مجموعه داده است، ساخته می‌شود.

مجموعه یال‌های  $E_1$  بر اساس ماتریس فاصله و پارامترهای  $\delta$  و  $\sigma$  ایجاد می‌شوند. به ازای داده‌های هر روز، تابع توزیع ماتریس فاصله محاسبه و پارامتر حد آستانه‌ای بر اساس آن تنظیم می‌گردد. مقدار  $\sigma_t$  به‌عنوان پارامتر حد آستانه‌ای داده‌های روز  $t$  -ام برابر با  $\sigma_t = \frac{M_t}{r}$  قرار داده می‌شود که  $M_t$  میانگین تابع توزیع روی ماتریس فاصله روز  $t$  -ام است. با در نظر گرفتن مقدار  $r$ ، فقط بین متنی‌هایی که شباهت چشم‌گیری وجود داشته باشد، یال وزن‌دار اضافه می‌شود. اگر مقدار پارامتر  $r$  زیاد باشد، تعداد یال‌های گراف کم و اگر این مقدار کم باشد، تعداد بیش‌تری یال در گراف حضور خواهند داشت. همچنین به ازای تمام اعضای مجموعه داده  $\delta = 2$  است.

مجموعه یال‌های  $E_2$  مربوط به یال‌های بین هر توئیت و هشتگ‌هایی است که در همان توئیت ظاهر شده است که مقدار وزن آن‌ها را ثابت  $c_1 = 1$  قرار می‌دهیم. با توجه به مقادیر  $\delta$  و  $c_1$ ، عمدتاً اهمیت یال‌های بین توئیت و توئیت بیش‌تر از یال‌های بین توئیت و هشتگ است. زیرا پیشینه مقدار وزن یال بین توئیت و توئیت تقریباً دو برابر وزن یال بین توئیت و هشتگ است. علاوه بر این، در این مقاله برای جلوگیری از پیچیدگی گراف، مجموعه یال‌های  $E_3$  را در نظر نگرفته و در واقع  $c_2 = 0$  را مقداردهی می‌کنیم.

ادغامی  $p(X, Y)$ ،  $MI(X, Y)$  به صورت زیر محاسبه می‌شود:

$$M(X, Y) = \sum_{x \in X} \sum_{y \in Y} \log \left( \frac{p(x, y)}{p(x)p(y)} \right). \quad (9)$$

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

$$NMI(X, Y) = \frac{M(X, Y)}{\sqrt{H(X)H(Y)}}, \quad (10)$$

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:

$$MNMI(Date_t) = \frac{1}{|C_t|} \sum_{i=1}^{|C_t|} \max_{Y_j \in Y} (NMI(X_i^t, Y_j)). \quad (11)$$

از این رابطه به‌عنوان معیاری برای محاسبه دقت الگوریتم‌ها استفاده می‌شود که در آن،  $Y$  مجموعه عناوین موجود در جدول (۱)،  $Y_i$  مجموعه هشتگ‌های عنوان  $t$ -ام و  $X_i^t$  مجموعه هشتگ‌های نماینده دسته  $t$ -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 اندازه‌گیری 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

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 |

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.

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 ارسالی 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  $\sigma_t = \frac{M_t}{2}$ ،  $\sigma_t = \frac{M_t}{3}$ ،  $\sigma_t = \frac{M_t}{4}$  برای هفت روز داده‌های شهر واشینگتن محاسبه شده است.

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

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

شبکه اجتماعی توئیتر برای اجرای الگوریتم پیشنهادی انتخاب و ویژگی خاص متن هر توئیت، هشتک‌های متن‌ها قرار داده شد. همچنین برای محاسبه وزن یال بین توئیت‌ها از الگوریتم از پیش آموزش دیده fastText استفاده گردید. توئیت‌هایی که مورد بررسی قرار گرفتند، مربوط به هفت روز محتوای منتشر شده 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 هشتک به عنوان برچسب هر دسته، در این مقاله صرفاً از توئیت‌هایی که دارای هشتک باشند استفاده شد. با توجه به دقت بسیار مناسب الگوریتم دوم، به ازای هر تشکل می‌توان چند هشتک پر درجه را به عنوان موضوع آن تشکل در نظر گرفت و این‌گونه شناسایی عناوین را انجام داد.

در آینده سعی داریم با ترکیب روش ارائه شده و الگوریتم‌های موجود در حوزه یادگیری شبه نظارتی F10<sup>11</sup>، توئیت‌هایی که در آن از هشتک 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>

مطابق جدول (۳)، با افزایش مقدار پارامتر  $r$  دقت افزایش یافته است. در روش ارائه شده در این مقاله، برای هر دو رأس از نوع توئیت، یالی وزن دار در نظر گرفته شده و مقدار این وزن بر اساس تعداد کلمات مشابه استفاده شده در دو رأس مورد نظر 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، زمانی که ابرگراف مبتنی بر موتیف ساخته می‌شود به یال‌ها وزن اختصاص داده می‌شود و عملاً وزن یال‌های کم اهمیت افزایش می‌یابد و زمانی که 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 تولید شده در شبکه‌های اجتماعی ارائه شد. الگوریتم ارائه شده مبتنی بر شناسایی تشکل‌های روی گراف است. برای ساخت گراف، محتوای متن‌های تولید شده (رئوس نوع اول) به عنوان معیار وزن یال‌های نوع اول و سایر ویژگی‌ها به عنوان رئوسی از نوع دوم در نظر گرفته شد. در نهایت از دو الگوریتم برای شناسایی تشکل استفاده شد.

- 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.
  2. Guille, A., Hacid, H., Favre, C., Zighed, D.A., “Information diffusion in online social networks: A survey” ACM 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, K.J., Tsai, C.F., Hu, Y.H., Wang, M.C., Liu, K.E., Yu, X., Yuan, Y., Chung, Y.M., “On Clustering Algorithms: Applications