

طبقه‌بندی محتوای متنی منتشرشده در شبکه‌های اجتماعی

فاطمه بهاری فرد^۱، وحید معتقد^۲^۱ پژوهشکده علوم کامپیوتر، پژوهشگاه دانش‌های بنیادی (IPM)

f.baharifard@ipm.ir

^۲ پژوهشکده علوم کامپیوتر، پژوهشگاه دانش‌های بنیادی (IPM)

vahid.motaghd2020@gmail.com

چکیده

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

در این مقاله، الگوریتم جدیدی برای دسته‌بندی محتوای متنی شبکه‌های اجتماعی 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 یک روش بدون ناظر، تشکلهای مختلف شناسایی می‌شوند. برای ارزیابی الگوریتم پیشنهادی، بستر شبکه اجتماعی توییتر انتخاب و با API داده‌های متنی ارسال‌شده از شهر واشینگتن در یک بازه زمانی، جمع‌آوری و الگوریتم ارائه‌شده روی آن اعمال شده است. برای بررسی دقت، نتایج حاصل با دو الگوریتم کلاسیک K-means و LDA بر اساس معیار اطلاعات متقابل نرمال‌شده، مقایسه شده است که نشان‌دهنده دقت مناسب الگوریتم پیشنهادی است.

کلمات کلیدی

شناسایی عنوان، گراف وزن‌دار، یادگیری بدون ناظر، شبکه‌های اجتماعی، توییتر

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 (به عنوان مثال تصاویر، متن و فیلم) تولید و منتشر می‌شود. روزانه ۵۰۰ میلیون توییت در توییتر (هر توییت حداکثر ۲۸۰ حرف) ارسال و ۹۰۰ میلیون عکس در فیسبوک و ۰.۴ میلیون ساعت فیلم در یوتیوب بارگذاری می‌شود [1]. 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 کاربران با یکدیگر اطلاعاتی به دست آورد [2].

در تکنیک‌های خوشه‌بندی اسناد^۱، 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 [3,4]. 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^۲ که در آن هر سند تبدیل به مجموعه‌ای از کلمات می‌شود [5]. TF-IDF که در آن کلمات مهم هر سند در مقابل کلمات مهم سایر اسناد سنجیده [6]، و کلمه-به-بردار^۳ که در آن هر کلمه با توجه به معنایش به بردار تبدیل می‌شود [7]. سپس الگوریتم‌های یادگیری ماشین بدون ناظر^۴ خوشه‌بندی سلسله مراتبی و تقسیمی مانند K-means روی آن اعمال می‌شود [8].

مدل‌سازی عناوین^۵ 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 [9]. 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، LDA و CTM ارائه شده است [10].

با خوشه‌بندی محتوای متنی شبکه‌های اجتماعی، می‌توان به مدل‌سازی عناوین و دسته‌بندی آن‌ها پرداخت. در شبکه‌های اجتماعی مانند توییتر، فیسبوک و اینستاگرام از هشتگ به عنوان کلمات پیوندی تعریف‌شده توسط کاربر استفاده شده است. این عبارات می‌توانند به عنوان ابزاری برای ارزیابی کیفیت خوشه‌بندی متن‌ها و مدل‌سازی عناوین استفاده شود [11]. یکی از چالش‌های استفاده از این الگوریتم‌ها، تفاوت بین اسناد و محتوای متنی منتشرشده در شبکه‌های اجتماعی است. این محتواها معمولاً دارای اختلال^۶ (غلط املایی، عبارات عامیانه، متن کوتاه، اشتباهات دستور زبانی و غیره) می‌باشند [9]. 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 توییتر فقط مجوز پخش شناسه هر توییت را مجاز می‌داند و افراد علاقه‌مند می‌توانند با API به متن هر توییت دسترسی داشته باشند و این مساله برای هر یک از شبکه‌های اجتماعی تابع مقررات مربوط به خود است [12]. پس پژوهش‌های مختلف، از روش‌های متنوعی برای ارزیابی کیفیت الگوریتم خود استفاده کرده‌اند. به عنوان مثال برای ارزیابی خوشه‌بندی داده‌های توییتر، از معیارهای مبتنی بر برچسب‌گذاری برای هر خوشه تا تفسیر خوشه‌های حاصله توسط عامل

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 منتشر شده در شبکه‌های اجتماعی 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 تشکلی روی گراف، تلاش می‌شود برای هر دسته عنوانی در نظر گرفته شود. در پایان دقت الگوریتم پیشنهادی با الگوریتم‌های مبنایی در این حوزه با معیار اطلاعات متقابل نرمال شده^۴ مقایسه می‌شود.

۲- الگوریتم پیشنهادی

در این بخش الگوریتم پیشنهادی برای تعیین عنوان یا موضوع متن منتشر شده در شبکه‌های اجتماعی 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
- (۳) ساخت گراف ارتباطی متناسب با محتوا
- (۴) اجرای الگوریتم شناسایی تشکلی روی گراف حاصل

فرض کنید ورودی الگوریتم یک مجموعه N عضو از متن‌های منتشر شده باشد. متن‌های منتشر شده 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^۵ و سپس علامت‌های نگارشی، کلمات توقف^۶ و لینک‌ها 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 پیش پردازش روی متن n -ام است. در این مقاله، فقط متن‌های منتشر شده 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 از پیش آموزش دیده fastText مبتنی بر دو میلیون کلمه [21]، 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:

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

در ادامه برای هر متن، میانگین بردارهای d بعدی کلمات همان متن، به عنوان بردار ویژگی آن لحاظ می‌شود. یعنی اگر داشته باشیم:

انسانی به صورت کیفی، استفاده شده است [13]. 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 تشکلی^۷ می‌شود. در ادامه برخی از پژوهش‌های انجام شده در این دو دسته آمده است.

در مرجع [14]، نویسندگان برای شناسایی عناوین توییت‌ها، 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 [15]، 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 در نهایت از خوشه‌بندی K-means استفاده شده است. در این پژوهش خوشه بندی روی کلمات اعمال شده و نشان داده شده است که کلمات مشابه در خوشه‌های یکسان دسته‌بندی می‌شوند. در مقاله [16] برای خوشه‌بندی متن‌های منتشر شده در توییتر و ردیت، چهار روش نمایش سند TF-IDF، کلمه-به-بردار، کلمه-به-بردار وزن دار 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 [17]، 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 میانگین گیری 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 [18] برای شناسایی عناوین توییت‌هایی که پس از زلزله ژاپن در سال ۲۰۱۱ منتشر شده است، با پنجره‌بندی توییت‌ها بر اساس زمان ارسال آن‌ها به صورت ساعتی، توییت‌های هر پنجره طبق کیسه کلمات به فرم بردار باینری 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 [19]، 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 تشکلی Louvain استفاده و تشکلی‌هایی 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 [20] برای شناسایی عناوین توییت‌ها از الگوریتمی مبتنی بر گراف استفاده شده است. هر توییت با سه نوع متفاوت از رأس‌ها در یک گراف نمایش داده می‌شود: توییت، هشتگ و کلمات موجود در هر توییت. بین رأس‌های کلمات و توییت، بین رأس‌های هشتگ و توییت و هم‌چنین بین کلمات و هشتگ، یال وزن دار ایجاد می‌شود. وزن یال‌های هشتگ-توییت بیشینه مقدار یک، وزن یال‌های توییت-کلمه مبتنی بر TF-IDF و وزن یال‌های کلمه-هشتگ 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

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

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 قرار دارد. در این مقاله از الگوریتم کلاسیک حریصانه [6] استفاده شده است. این الگوریتم مبتنی بر محاسبه مازولاریتی است و می‌تواند تشکل‌ها را به صورت اکتشافانه 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.

۳- پیاده‌سازی

در این مقاله برای ارزیابی الگوریتم پیشنهادی، به بررسی محتوای متنی تولیدشده در شبکه اجتماعی توییتر می‌پردازیم. برای جمع‌آوری متن توییت‌ها از API توییتر استفاده شده است و مجموعه داده شامل ریتوییت 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 توییت است. پس از فیلتر توییت‌هایی که به زبان‌هایی غیر از انگلیسی ارسال شده‌اند، تعداد توییت‌ها 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 هشتک‌هایی که بیش‌ترین دفعات رخداد را داشته‌اند، برای شناسایی عناوین صحیح توییت‌ها استفاده می‌شود. هم‌چنین هشتک‌ها، ملاک ارزیابی دقت شناسایی عنوان تشکل در نظر گرفته شده است.

زمانی که بخواهیم از هشتک برای تخمین برجسب صحیح توییت استفاده کنیم با چالش‌هایی نظیر وجود توییت‌های بدون هشتک، وجود توییت‌هایی با بیش از یک هشتک، وجود توییت‌هایی با هشتک‌های نامربوط و وجود هشتک‌های متفاوت برای یک موضوع مواجه هستیم. در این‌جا همه توییت‌های بدون هشتک را به دلیل فقدان معیاری برای ارزیابی نتیجه، از مجموعه داده حذف کرده و در نهایت مجموعه داده ورودی به ۲۰۸۷۶۳ توییت تقلیل پیدا می‌کند. برای حذف هشتک‌های بی‌ربط و یا موضوعات کم‌اهمیت، با سعی و خطا تعدادی از تشکل‌های شناسایی شده کم‌جمعیت، از مجموعه نهایی حذف می‌شوند. در

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

بردار 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:

$$V = \{V_1 \cup V_2\}, \\ V_1 = \{D_1, \dots, D_N\}, \quad (6) \\ V_2 = \left\{ \bigcup_{i=1}^N H_i \mid H_i = \{h_1^i, \dots, h_{n_i}^i\} \right\},$$

که h_j^i ویژگی j -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:

$$E = \{E_1 \cup E_2 \cup E_3\}. \quad (7)$$

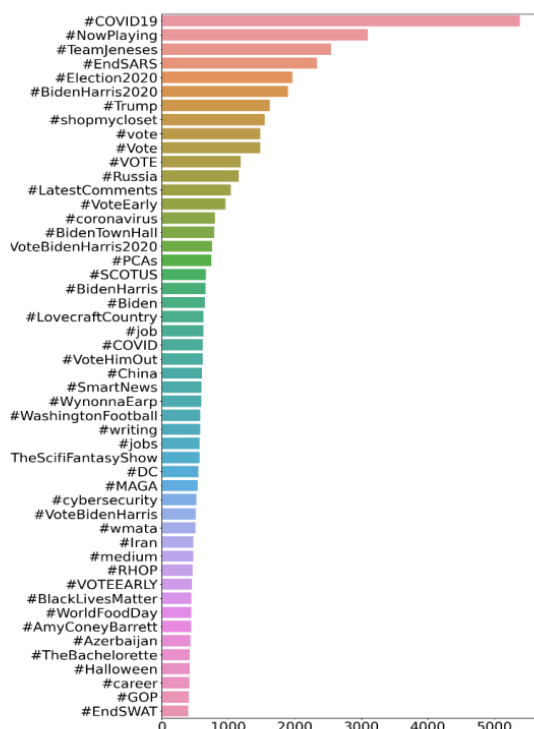
مجموعه یال‌های شناسه-شناسه (E_1), طبق رابطه (۷) تعریف می‌شود:

$$E_1 = \{uv \mid u, v \in V_1, w(u, v) = w_{uv}\}, \\ w_{uv} = \begin{cases} \delta - d_{uv}, & \text{if } d_{uv} < \sigma, \\ 0, & \text{Otherwise.} \end{cases} \quad (8)$$

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.

پردازش می‌شود و به ازای هر روز، دقت الگوریتم پیشنهادی و سایرین محاسبه و گزارش می‌شود.

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



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

مجموعه یال‌های E_1 بر اساس ماتریس فاصله و پارامترهای δ و σ ایجاد می‌شوند. به ازای داده‌های هر روز، تابع توزیع ماتریس فاصله محاسبه و پارامتر حد آستانه‌ای بر اساس آن تنظیم می‌گردد. مقدار σ_t به عنوان پارامتر حد آستانه‌ای داده‌های روز t برابر با $\sigma_t = \frac{M_t}{4}$ قرار داده می‌شود که M_t میانۀ تابع توزیع روی ماتریس فاصله روز t است. با این مقدار فقط بین متنی‌هایی که شباهت چشم‌گیری وجود داشته باشد، یال وزن‌دار اضافه می‌شود. همچنین به ازای تمام اعضای مجموعه داده $\delta = 2$ است.

مجموعه یال‌های E_2 مربوط به یال‌های بین هر توییت و هشتگ‌هایی است که در همان توییت ظاهر شده است که مقدار وزن آن‌ها را ثابت $c_1 = 1$ قرار می‌دهیم. با توجه به مقادیر δ و c_1 ، عمدتاً اهمیت یال‌های بین توییت و توییت

جدول (۱)، ده موضوع پر اهمیت بر اساس پردازش هشتگ‌های پرتکرار، توسط عامل انسانی ارائه شده است. برای هر عنوان، چهار هشتگ پرتکرار فهرست شده است.

برخلاف سایر پژوهش‌ها که عمدتاً مبتنی بر جستجو با عبارت و یا هشتگ، روی یک موضوع نسبتاً بزرگ است، در این مقاله ماهیت مجموعه متن‌های توییت‌شده، مربوط به توییت‌های 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 هشتگ 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 تکنولوژی نیز با مساله سلامت ترکیب خواهند شد، ولی در جدول (۱)، هشتگ کرونا فقط در عنوان سلامتی-کرونا آمده است.

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
تکنولوژی	#CyberSecurity, #BigData, #AI, #IoT
مسائل بین المللی	#Russia, #China, #Azerbaijan, #Iran
مساله نیجریه	#EnSARS, #EndSWAT, #Nigeria, #BuhariMustGo
انتساب قاضی دیوان عالی	#SCOTUS, #AmyConeyBarrett, #ACB, #BlockBarrett
مساله جنبش سیاه پوستان	#BlackLivesMatter, #BLM, #racism, #Antifa

در شکل (۱) نمودار فراوانی ۵۰ هشتگ پرتکرار ارائه شده است. در مجموعه داده توییت‌های هفت روز شهر واشینگتن تعداد کل هشتگ‌ها ۴۵۷۱۶۰ است که از این بین تعداد هشتگ‌های 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)، هشتگ "VOTE" به سه شکل مختلف ("VOTE", "Vote" و "vote") ظاهر شده است. 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 مساله محاسبه دقت مواجه خواهیم شد. چون ممکن است برای هر عنوان تعداد متفاوتی هشتگ 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 شناسایی عناوین و هشتگ‌های مرتبط، اولین گام در پیاده‌سازی، فاز پیش‌پردازش است. در این مقاله مجموعه داده به صورت پنجره‌بندی‌شده روزانه

مقداردهی می‌شود. در الگوریتم LDA چهار کلمه با بیش‌ترین آمارگان به عنوان نماینده هر دسته انتخاب و توسط عامل انسانی رابطه (۱۱) برای آن محاسبه می‌شود.

در جدول (۲) مقادیر $MNMI$ به دست آمده برای الگوریتم پیشنهادی، الگوریتم K-means و الگوریتم LDA به ازای هفت روز مجموعه داده مربوط به متن توییت‌های ارسال شده از شهر واشینگتن گزارش شده است.

جدول (۲): مقادیر $MNMI$ به دست آمده برای الگوریتم پیشنهادی، الگوریتم K-means و الگوریتم LDA به ازای هفت روز مجموعه داده مربوط به متن توییت‌های ارسال از شهر واشینگتن

روز	الگوریتم پیشنهادی	K-means	LDA
۲۰۲۰-۱۰-۱۵	۰,۷۵۴	۰,۴۶۶	۰,۵۰۰
۲۰۲۰-۱۰-۱۶	۰,۷۰۴	۰,۶۱۶	۰,۴۴۰
۲۰۲۰-۱۰-۱۷	۰,۶۰۹	۰,۵۸۳	۰,۳۶۰
۲۰۲۰-۱۰-۱۸	۰,۷۹۱	۰,۴۳۳	۰,۳۰۰
۲۰۲۰-۱۰-۱۹	۰,۴۶۲	۰,۴۳۳	۰,۳۰۰
۲۰۲۰-۱۰-۲۰	۰,۶۲۵	۰,۵۱۶	۰,۲۸۰
۲۰۲۰-۱۰-۲۱	۰,۶۳۴	۰,۴۰۰	۰,۲۴۰

مطابق جدول (۲)، الگوریتم ارائه شده علی‌رغم استفاده از الگوریتم سستی شناسایی شکل 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 توییت است که با این کار عملاً توییت‌ها در فضای ۳۰۰ بعدی بر حسب ویژگی‌های آن‌ها نگاشت شده و روی آن الگوریتم K-means اجرا شده است.

۵- نتیجه‌گیری

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

شبکه اجتماعی تویتر برای اجرای الگوریتم پیشنهادی انتخاب و ویژگی خاص متن هر توییت، هشتک‌های متن‌ها قرار داده شد. همچنین برای محاسبه وزن یال بین توییت‌ها از الگوریتم افزایش آموزش دیده fastText استفاده گردید. توییت‌هایی 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 الگوریتم پیشنهادی با الگوریتم‌های K-means و LDA، نشان داده شد که علی‌رغم استفاده از الگوریتم کلاسیک شناسایی تشکلهای حریصانه در الگوریتم پیشنهادی، دقت این روش بهتر از الگوریتم‌های K-means و LDA است.

به دلیل استفاده از هشتک به عنوان برجسب هر دسته، در این مقاله صرفاً از توییت‌هایی که دارای هشتک باشند استفاده شد. با این وجود در آینده سعی

بیشتر از یال‌های بین توییت و هشتک است. زیرا بیشینه مقدار وزن یال بین توییت و توییت تقریباً دو برابر وزن یال بین توییت و هشتک 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.

در ادامه با الگوریتم پیشنهادی حریصانه ماژولاریتی، روی گراف ساخته شده، تشکلهای شناسایی و ۹۰ درصد آن‌ها به عنوان خروجی صحیح و ۱۰ درصد به عنوان اختلال دسته‌بندی می‌شوند.

۴- ارزیابی و مقایسه

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

برای ارزیابی کیفیت الگوریتم‌ها از معیار اطلاعات متقابل 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 ادغامی $MI(X, Y)$ و $p(X, Y)$ به صورت زیر محاسبه می‌شود:

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

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

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

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 معرفی شده، 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 به عنوان معیاری برای محاسبه دقت الگوریتم‌ها استفاده می‌شود که در آن، Y مجموعه عناوین موجود در جدول (۱)، مجموعه هشتک‌های عنوان t -ام و X_i^t مجموعه هشتک‌های 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 ارائه شده در زبان پایتون و روی سخت‌افزاری با ۱۶ گیگابایت حافظه پیاده‌سازی شده است. برای اندازه‌گیری کیفیت هر دسته، چهار هشتک با بیش‌ترین درجه، به عنوان نماینده هر دسته انتخاب شده‌اند.

برای اجرای الگوریتم K-means از کتابخانه scikit-learn در پایتون استفاده شده و تعداد خوشه‌ها به عنوان دانش اولیه برای هر مجموعه داده، در آن اعمال شده است. به منظور رعایت شرایط یکسان، بردار ویژگی متن توییت‌ها به عنوان ورودی الگوریتم K-means در نظر گرفته شده است. برای اندازه‌گیری کیفیت K-means به ازای هر خوشه، چهار هشتکی که بیش‌ترین دفعات تکرار را داشته‌اند، به عنوان 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

- 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 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.
- داریم با ترکیب روش ارائه‌شده و الگوریتم‌های موجود در حوزه یادگیری شبه‌نظارتی^{۱۱}، توییت‌هایی که در آن از هشتگ استفاده نشده باشد را نیز با دقت مناسبی دسته‌بندی کنیم. همچنین برای افزایش دقت الگوریتم ارائه‌شده می‌توان از سایر الگوریتم‌های موجود برای شناسایی تشکل روی گراف استفاده کرد.

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 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

پانویس‌ها

- ^{*} Community Detection
[^] Normalized Mutual Information
[^] Tokenizing
[^] Stop Words
[^] Semi-Supervised Learning

- [^] Document Clustering
[^] Bag of Words
[^] Word2Vec
[^] Unsupervised Learning
[^] Topic Detection
[^] Noise