پیشنهاد هشتگ در سیستم‌های میکروبلاگ توسط بردار موضوعی: مورد کاربرد توئیتر

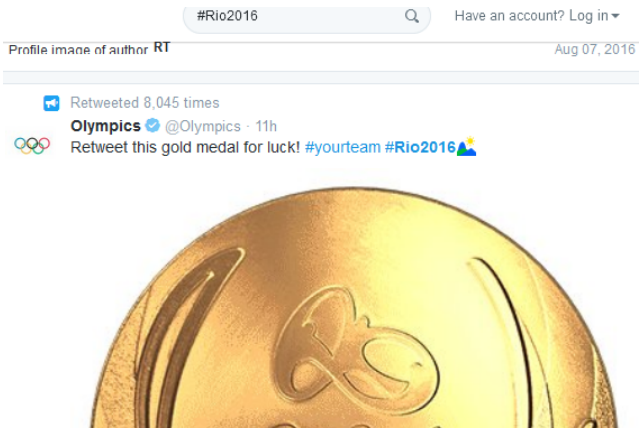
میرسامان تاج‌بخش و جمشید باقرزاده

چکیده: با معرفی وب ۲.۰، داده‌های ایستا که در وب ۱.۰ وجود داشتند، حالت ساخت‌یافته‌تری به خود گرفتند. ویکی‌ها، بلاگ‌ها، شبکه‌های اجتماعی و سیستم‌های بوکمارکینگ اجتماعی مثال‌هایی از آن هستند که کاربران در آنها محتوا تولید می‌کنند. یکی از مشکلات تولید محتوا توسط کاربر، عدم یکپارچگی محتوای تولیدشده می‌باشد که باعث تولید داده‌های ناهمگون شده و اجرای الگوریتم‌ها و تکنیک‌های کامپیوتری را دشوار می‌سازد. راه حل وب ۲.۰ برای کاهش اثر این مشکل، استفاده از هشتگ (تگ) برای مطالب منتشرشده توسط کاربر است که خود کاربر به مطالب منتشرشده خود، تگ می‌زند. این راهکار در میکروبلاگ‌هایی چون توئیتر کماکان رفع نشده است چرا که کاربران با محدودیت کاراکتری (۱۴۰ کاراکتر برای هر توئیت) مواجه هستند و ممکن است تعداد کاراکترهای محتوا باعث شود که برخی کاراکترهای هشتگ در پست نباشد. در این مقاله سعی شده تا با استفاده از روش تخصیص دیریکله نهفته و نمونه‌برداری Gibbs فروریخته، مشکل پیشنهاد هشتگ در محیط ناهمگون توئیتر رفع شود. پیشنهاد هشتگ بر روی ۸۳۹۶۷۴۴ توئیت به زبان انگلیسی پیاده‌سازی و در آزمایش‌های مختلف بین ۱ تا ۵ مرتبط‌ترین هشتگ پیشنهاد شده است. نتایج در حالات مختلف دقت بالای ۲۰٪ و فراخوانی بالای ۴۵٪ را نشان می‌دهد که نشانگر افزایش دقت از ۳٪ به ۲۱٪ و افزایش فراخوانی از ۳۲٪ به ۴۶٪ در مقایسه با دقیق‌ترین روش بررسی‌شده پیشنهاد هشتگ توسط LDA بدون تغییر، توسط نویسندگان است.

کلیدواژه: ***سیستم‌های توصیه‌گر، توصیه هشتگ، بردار موضوعی، تخصیص دیریکله نهفته، نمونه‌برداری Gibbs، میکروبلاگ، توئیتر.***

۱- مقدمه

شبکه اجتماعی توئیتر به عنوان یکی از پرطرفدارترین میکروبلاگ‌های اجتماعی، نقش عمده‌ای در جهان دارد. به طور مثال تأثیر توئیتر در روند انقلاب مصر قابل انکار نمی‌باشد [۱]. یکی از علل موفقیت این شبکه انتقال لحظه‌ای اخبار است که کاربران می‌توانند محتوای مورد نظر خود را تولید کرده و در صفحه خود نشر دهند. پیروان[1] آن کاربر نیز از مطلب منتشرشده مطلع می‌شوند و می‌توانند آن را مجدداً در شبکه تشکیل‌شده از پیروان خود بازنشر دهند. کاربران برای اضافه‌کردن مفهوم و دسته‌بندی هرچه بیشتر مطالب منتشری خود از مفهوم هشتگ استفاده می‌کنند. هشتگ کلماتی در متن توئیت است که با # شروع شده و شامل حروف و اعداد می‌تواند باشد [۲]. استفاده از هشتگ منجر می‌شود تا بر هر مطلب



شکل ۱: مطالب مرتبط با هشتگ #Rio2016.

یک یا چند برچسب زده شود و کلمات مهم آن پست معلوم شوند.

علاوه بر افزودن مفهوم، کاربران جهانی نیز می‌توانند اخبار منتشرشده با یک هشتگ خاص را لحظه به لحظه جویا باشند. به طور مثال شکل ۱ نشانگر اخبار و مطالب مربوط به #Rio2016 می‌باشد که مربوط به المپیک ۲۰۱۶ برزیل است. این اخبار در تاریخ ۷ آگوست ۲۰۱۶ از سیستم جست‌وجوی توئیتر گرفته شده‌اند.

از کاربردهای مهم هشتگ در امر دسته‌بندی و خوشه‌بندی مطالب کاربران است بدین صورت که کاربران می‌توانند با مشخص‌کردن کلمات مهم مطالبشان توسط هشتگ، موضوعات و برچسب‌های کلی آنها را نمایان کنند. البته دو چالش عمده در بحث مفهوم هشتگ (و یافتن مفهوم پست به واسطه آنها) و پیشنهاد آنها موجود است که توضیح داده می‌شود:

۱) ساختار خود هشتگ‌ها ناهمگون است: محدودیت ۱۴۰ کاراکتری در توئیتر، کاربران را مجبور کرده تا بعضی از کلمات و یا هشتگ‌ها را به صورت ناموزون استفاده کنند. مثلاً در مجموعه داده استفاده‌شده در این تحقیق برای هشتگ #followfriday شکل‌های دیگری مانند #ff، #f4f، #follow_Friday، #unfollow_Friday و #followfrida استفاده شده که این گونه عدم توازن، مشکلات عمده‌ای در فرایند تحلیل و داده‌کاوی هشتگ به وجود می‌آورد.

۲) اکثریت پست‌ها هشتگ ندارند: کاربران توئیتر کمتر از هشتگ استفاده می‌کنند. مطالعات مختلف نشان داده که اکثر مطالب منتشرشده توسط کاربران هشتگ ندارند [۳] و [۴].

برای رفع این مشکلات تکنیک‌هایی نیز استفاده گردیده که در بخش کارهای مرتبط توضیح داده شده است. بعضی از محققان با تغییر روش‌های مرسوم پردازش زبان طبیعی چون TF-IDF[2] سعی کرده‌اند که مشکل ۱ و ۲ را حل نمایند. برخی دیگر به روش‌های یافتن شباهت معنایی رو آورده‌اند تا بتوانند مجموعه هشتگ‌هایی چون مجموعه مثال

این مقاله در تاریخ ۲ اردیبهشت ماه ۱۳۹۷ دریافت و در تاریخ ۵ آبان ماه ۱۳۹۷ بازنگری شد.

میرسامان تاج‌بخش، دانشکده مهندسی برق و کامپیوتر، دانشگاه ارومیه، ارومیه، ایران، (email: ms.tajbakhsh@urmia.ac.ir).

جمشید باقرزاده (نویسنده مسئول)، دانشکده مهندسی برق و کامپیوتر، دانشگاه ارومیه، ارومیه، ایران، (email: j.bagherzadeh@urmia.ac.ir).

1. Follower

2. Term Frequency-Inverse Document Frequency

زده شده Follow Friday را در یک دسته بیندازند. عده‌ای دیگر نیز از روش‌های احتمالی (موسوم به روش‌های مبتنی بر LDA) برای این منظور استفاده کرده‌اند.

روش‌های احتمالی (مبتنی بر LDA) بدین صورت کار می‌کنند که متون مشابه از نظر معنا و مفهوم به صورت بدون ناظر[1] در یک دسته قرار می‌گیرد و معمولاً پست‌ها و هشتگ‌های یک دسته، معنی مشابهی داشته و یا در ارتباط با موضوع خاصی صحبت می‌کنند. جهت پیشنهاد هشتگ از روش LDA جهت دسته‌بندی متون استفاده می‌شود. پس از دسته‌بندی متون بدون هشتگ نیز دسته‌بندی شده و از هشتگ‌های آن دسته به آنها پیشنهاد می‌شود. Melis و Saveski با اعمال LDA در توئیتر و پیشنهاد هشتگ به پست‌های بدون هشتگ، یک سیستم پیشنهاد هشتگ را توسعه داده‌اند [۵]. به طور مشابه در [۶] نیز نویسندگان سعی کرده‌اند که با اختصاص برچسب و دسته‌بندی توئیت‌ها، هشتگ پیشنهاد دهند.

مشکل عمده موجود در استفاده از روش‌های مبتنی بر LDA در سیستم‌هایی مثل توئیتر در این است که چون اندازه توئیت‌ها (سندها) کوتاه است، دسته‌بندی متون با دقت کمی انجام می‌شود. همین کمبود دقت منجر می‌شود که هشتگ‌های مناسبی پیشنهاد نشوند و طبعاً دقت سیستم پیشنهاددهنده پایین باشد. به علاوه در این گونه سیستم‌ها، معمولاً هر سند را متعلق به یک موضوع در نظر می‌گیرند در حالی که ممکن است کاربر در ارتباط با بیش از یک موضوع در داخل سند بنویسد. طبعاً برای پیشنهاد هشتگ به توئیت‌هایی که فقط یک موضوع برای آنها در نظر گرفته شود، هشتگ‌های همان موضوع پیشنهاد می‌شود، در صورتی که ممکن است یک متن به موضوعات مختلفی متعلق باشد و طبعاً از هشتگ یک موضوع خاص استفاده نخواهد کرد. به عبارتی در این گونه موارد هشتگ پیشنهادی از یک موضوع خاص، بر اساس ذائقه عمومی کاربران بوده و ذائقه یک کاربر محسوب نمی‌شود، چرا که آن موضوع خاص بر اساس نظرات عمومی کاربران تشکیل شده و ذائقه شخصی کاربر در فرایند پیشنهاد هشتگ دخیل نبوده است.

بنابراین می‌توان از هشتگ‌های موضوعات مختلف برای پیشنهاد هشتگ به توئیت‌های بدون هشتگ استفاده کرد. ایده مورد بررسی در این مقاله چنین است که هر توئیت به صورت برداری از درصد تعلقات به موضوعات مختلف در نظر گرفته می‌شود که خود این نیز منجر به پیشنهاد هشتگ از دسته‌های مختلف می‌گردد. با این روش می‌توان هشتگ‌های مناسب را برای توئیت‌ها پیشنهاد کرد که طبعاً می‌تواند مشکل ۱ را بهبود بخشیده و دقت سیستم‌های پیشنهاددهنده هشتگ را افزایش دهد. به عبارت دیگر هشتگ‌هایی پیشنهاد می‌شوند که بر اساس ذائقه فردی بوده (و نه فقط ذائقه عمومی) و نتایج بهتری ارائه می‌دهد.

در ارتباط با مشکل ۲ که شرایط را سخت‌تر نیز می‌کند چرا که اکثریت پست‌ها هشتگ نداشته و از آنها نمی‌توان جهت یادگیری سیستم و استفاده در LDA بهره برد. لذا پیشنهاد هشتگ باید به صورت هدفمند بوده و برای دستیابی به نتایج قابل قبول بر اساس ذائقه فرد باشد.

راه حل پیشنهادی در دیگر مقالات بررسی‌شده در بخش کارهای پیشین [۵] تا [۱۱] چنین است که ابتدا دسته‌بندی را روی داده‌ها انجام داده و سپس کلمات برجسته هر دسته را پیشنهاد می‌دهند که این کلمات حاصل ذائقه عمومی کاربران است.

نوآوری این کار، افزودن ذائقه فرد در روند پیشنهاد هشتگ بر اساس یافتن توئیت‌هایی با موضوع مشابه است که در بخش سوم توضیح داده شده است. چرا که الگوریتم اصلی LDA بر متون بزرگ نتیجه بهتری نسبت به متون کوتاه (مانند توئیت‌های موجود در توئیتر) دارد. بنابراین سعی شده با تغییر روش پیشنهاد هشتگ دقت روش LDA بدون تغییر در متون کوتاه بیشتر شود. بدین منظور روش انتخاب توئیت‌های مشابه و طبعاً هشتگ برای پیشنهاد به توئیت خاص، نوآوری دوم مقاله است که ذائقه شخصی کاربر را نیز دخالت داده است. کارهای پیشین فرایند یافتن توئیت‌های مشابه و پیشنهاد هشتگ را با در نظر گرفتن ذائقه عمومی و بدون در نظر گرفتن ذائقه شخصی کاربر اعمال کرده‌اند که با مشکل ۱ سازگاری ندارد و طبعاً نمی‌توانند مشکل ۱ را حل کنند. این مقاله چند هدف دارد: ۱) پیشنهاد هشتگ در حین نگارش توئیت، ۲) پیشنهاد هشتگ برای توئیت‌های نگارش‌شده بدون هشتگ و ۳) یافتن شبیه‌ترین توئیت‌ها به ازای هر توئیت.

ساختار مقاله که در ادامه آورده شده به صورت زیر است: در بخش دوم کارهای مرتبط با پیشهاد هشتگ در محیط‌های میکروبلاگ توضیح داده شده است. بخش سوم به توضیح روش تخصیص دیریکله نهفته و نحوه ایجاد بردار موضوعی و نوع استفاده از آن در روش پیشنهادی می‌پردازد. در بخش چهارم نتایج به دست آمده تحلیل‌شده با دیگر روش‌ها مقایسه شده‌اند. بخش پنجم نیز شامل نتیجه‌گیری و کارهای آینده است.

۲- کارهای پیشین

در این بخش در ارتباط با کارهای مرتبط با پیشنهاد هشتگ در شبکه‌های اجتماعی و روش‌های مختلف بر اساس مدل‌سازی موضوعی صحبت شده است. بحث اصلی پیشنهاد هشتگ مرتبط با سیستم‌های پیشنهاددهنده است [۱۲] تا [۱۴]. در کاری که توسط نویسندگان این مقاله صورت گرفته است از روش Semantic Vector که گسترش‌یافته روش TF-IDF به صورت معنایی است، جهت یافتن توئیت‌های مشابه و پیشنهاد هشتگ استفاده شده است [۱۵].

از طرف دیگر بحث مدل‌سازی موضوعی در زمینه‌های مختلف کاربرد دارد [۱۶] و همچنین در میکروبلاگ‌ها نیز مورد بررسی فراوان قرار گرفته است. Melis و Saveski با مطرح‌کردن مشکلات LDA[2] در فضای محدود توئیتر، با جمع‌بندی توئیت‌ها، مکالمات، کاربران و هشتگ‌ها، سعی در تشکیل سندهای بزرگ‌تر[3] و رفع مشکل ۱۴۰ کاراکتری هر سند در توئیتر کرده‌اند [۵].

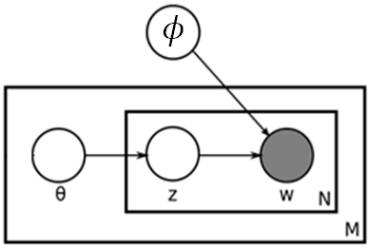
در کار دیگری، Mehrotra و همکاران نیز با دیدگاه مشابهی سعی در تجمیع توئیت‌ها کرده‌اند ولی در این کار با محاسبه شباهت کسینوسی هر توئیت هشتگ‌دار با توئیت‌های بدون هشتگ، مجموعه سند خود را بزرگ‌تر کرده‌اند. جهت وزن‌دهی بردار هر توئیت از راهکار TF-IDF استفاده کرده‌اند [۶]. در کار دیگر نیز روش‌های LDA، AT و ART در توئیتر به صورت اعمال هر یک از روش‌ها بر یک توئیت و یا اعمال بر روی تجمیع چندین توئیت بررسی شده‌اند. در روش ایشان نیز از هشتگ جهت دسته‌بندی توئیت‌ها استفاده شده است [۷].

کارهای توضیح داده شده در ادامه در حوزه پیشنهاد هشتگ بر اساس روش LDA بدون تغییر می‌باشند. در این کارها، تمرکز بر یافتن توئیت‌های مشابه با هر توئیت و پیشنهاد هشتگ بر اساس آن شباهت بوده است. کارهای جدیدتر، بر اساس دسته‌بندی خود توئیت‌ها و پیشنهاد هشتگ‌های آن دسته، متمرکز هستند. در یکی از این کارها، گودین و

1. Unsupervised

2. Latent Dirichlet Allocation

3. Tweet Polling



شکل ۲: گراف احتمالی فرایند دیریکله.

همکارانش از LDA جهت دسته‌بندی و موضوع‌بندی توئیت‌ها استفاده کرده‌اند. بعد از اتمام دسته‌بندی به ازای هر توئیت جدید، دسته و در نتیجه موضوع آن توئیت پیدا شده و محتمل‌ترین هشتگ‌های همان موضوع به آن توئیت پیشنهاد داده می‌شود [۸]. در این کار کلمات محتمل هر موضوع به هر توئیت پیشنهاد داده شده است. در [۹] تا [۱۱] هشتگ در مدل احتمالی وارد شده و پس از اعمال مدل دیریکله نهفته، احتمال اختصاص هشتگ‌ها به یک پست حساب می‌شوند.

مشکل موجود در کارهایی که از روش‌های مبتنی بر LDA استفاده می‌کنند، تمرکز بر یک دسته است. به طور مثال هر توئیت الزاماً هنگام پیشنهاد هشتگ، بایستی متعلق به یک دسته باشد و اصل اساسی روش LDA که تعلق احتمالی به هر دسته است در نظر گرفته نمی‌شود. نوآوری این تحقیق در استفاده از تعلق توئیت به موضوعات مختلف جهت پیشنهاد هشتگ می‌باشد که جزئیات روش پیشنهادی در بخش ۳ آمده است.

۳- ادبیات موضوع

در این بخش جزئیات روش تخصیص دیریکله نهفته و نحوه استفاده از آن در تشکیل بردار موضوعی شرح داده می‌شود.

۳-۱ توزیع دیریکله

توزیع دیریکله، یک توزیع پیوسته ریاضی است که حالت گسترده توزیع بتا می‌باشد. یکی از کاربردهای این توزیع در مقداردهی اولیه فرایندهای احتمالی بیزی است. عملکرد کلی بدین ترتیب است که در ابتدا یک بردار از اعداد تصادفی با اندازه مشخص تولید می‌شود. اگر اعداد انتخابی بردار، یکسان باشند توزیع دیریکله یکنواخت و در غیر این صورت غیر یکنواخت نامیده می‌شود. اعداد تصادفی بردار در بازه [۰،۱] نرمال‌سازی می‌گردند. سپس به ازای هر نمونه‌گیری تصادفی، یک عدد تصادفی در بازه [۰،۱] انتخاب می‌شود که این عدد نشانگر یکی از حالات فضای بین [۰،۱] است که توسط بردار تقسیم شده است. کاربرد این توزیع در مقداردهی اولیه روش تخصیص نهفته دیریکله می‌باشد.

۳-۲ تخصیص نهفته دیریکله

روش تخصیص نهفته دیریکله یک روش بدون نظارت جهت دسته‌بندی متون و تخصیص کلمات و سندها می‌باشد [۱۷]. در این روش هر سند به صورت مجموعه‌ای از کلمات در نظر گرفته می‌شود که در فرایند دیریکله، کلمات مرتبط و در یک حوزه را به صورت بدون ناظر دسته‌بندی می‌کند. ورودی این روش سندها (d) و موضوعات (k) موجود در کل اسناد است. مدل مولد این فرایند به صورت زیر است:

۱) به ازای هر موضوع یک توزیع دیریکله ایجاد می‌شود (ϕ_k) که بیانگر میزان تعلق هر کلمه به موضوع k است.

۲) به ازای هر سند (d) یک توزیع دیریکله ایجاد می‌شود (θ_d) که این توزیع بیانگر تعلق سند (d) به هر موضوع است.

۲-۱) به ازای هر کلمه (w_i) از سند (d):

۲-۱-۱) یکی از موضوعات از توزیع سند (θ_d) انتخاب می‌شود (z_i).

۲-۱-۲) بر اساس موضوع منتخب (z_i) سند، کلمه جدیدی از توزیع (ϕ_{z_i}) انتخاب شده و در سند (d) جای گیرد.

در نهایت این فرایند، کلمات مرتبط در یک گروه ظاهر می‌شوند. در اینجا متغیرهای نهفته ϕ_k, θ_d, z_i بوده و متغیر آشکار کلمات دیده‌شده می‌باشند. هدف محاسبه (۱) خواهد بود. w, z, θ, ϕ حالت کلی متغیرهای اشاره‌شده در قسمت ۳-۱ هستند

$$p(z,\theta,\phi|w) = \frac{p(w,z,\theta,\phi)}{p(w)} \qquad (۱)$$

متأسفانه (۱) به راحتی قابل محاسبه نمی‌باشد چرا که $p(w)$ در دست نیست. جهت حل چنین احتمالاتی روش‌هایی وجود دارد که یکی از این روش‌ها، روش نمونه‌برداری Gibbs است که در ادامه توضیح داده شده است. شکل ۲ گراف احتمالی متغیرهای فرایند دیریکله را نمایش می‌دهد. در این شکل دو عامل در تعیین احتمال اختصاص کلمه (w) به هر یک از موضوعات، تأثیرگذار است: ۱) توزیع احتمالی موضوعی کلمه در دیگر اسناد که با ϕ نشان داده شده است. ۲) توزیع احتمالی موضوعی کلمه در هر سند که با θ نمایش داده شده است. البته از روی توزیع هر سند، به ازای w_i انتخاب‌شده از آن سند، موضوعی نمونه‌برداری می‌شود که با z_i نمایش داده می‌شود و در کل، بردار z بر تعیین موضوع بردار w تأثیر اصلی را دارد.

۳-۳ روش نمونه‌برداری Gibbs

در این الگوریتم که یک روش از MCMC[1] است، جهت محاسبه توابع احتمالی که فرم مشخص ندارند استفاده می‌گردد. جهت محاسبه (۱) می‌توان از روش نمونه برداری Gibbs استفاده کرد، ولی ۲ متغیر ϕ و θ بر اساس متغیر z قابل محاسبه هستند چرا که در صورت مشخص‌بودن موضوع هر کلمه (z_i)، دو متغیر ۱) موضوع سند (θ_d) و ۲) توزیع موضوعی کلمات (ϕ_k)، قابل محاسبه می‌باشد. بنابراین این دو متغیر می‌توانند از فرایند محاسبه حذف شوند و در نتیجه می‌توان از نمونه‌برداری Gibbs فروپاش[2] استفاده کرد. بدین ترتیب محاسبه (۱) به صورت محاسبه $p(w,z)/p(w)$ خواهد بود. البته طی این شرایط می‌توان به جای محاسبه $p(w,z)/p(w)$ فقط قسمت صورت کسر را محاسبه کرد چرا که داریم

$$\frac{p(w,z)}{p(w)} = \frac{p(w,z)}{\sum_{all\,z} p(w|z)} \sim p(w,z)$$

فرایند اثبات در (۲) تا (۴) ذکر شده‌اند

$$p(w,z) = p(w|z) \times p(z) \qquad (۲)$$

$$p(w_i|z=j) = \int p(w_i|z,\phi)p(\phi)\mathrm{d}\phi = \frac{n_j^{w_i} + \beta}{n_j^{w_\cdot} + W\beta} \qquad (۳)$$

با توجه به این که متغیر ϕ توزیع دیریکله دارد، انتگرال محاسبه‌شده امید ریاضی توزیع دیریکله می‌باشد که در آن $n_j^{w_i}$ تعداد دفعاتی است که کلمه

1. Markov Chain Monte Carlo
2. Collapsed Gibbs Sampling

شکل ۳: فرایند کلی پیشنهاد هشتگ در روش پیشنهادی.

w_i به موضوع j ام اختصاص داده شده و n_j^w تعداد کل کلماتی است که در موضوع j ام هستند. ثابت β برای جلوگیری از صفرشدن صورت و مخرج استفاده می‌شود. ثابت W نیز تعداد کل کلمات موجود در سندها است

$$p(z=j)=\int p(z|\theta)p(\theta)\mathrm{d}\theta=\frac{n_j^d+\alpha}{n_\cdot^d+K\alpha} \qquad (۴)$$

در (۴) نیز امید ریاضی توزیع θ که توزیعی دیریکله است، محاسبه شده و ثابت α جهت جلوگیری از صفرشدن و ثابت K تعداد موضوعات است. متغیر n_j^d نمایانگر تعداد کلماتی است که در توئیت d آمده و متعلق به موضوع j هستند. همچنین متغیر $n_\cdot^d$ نمایانگر تمامی کلمات موجود در توئیت d است.

الگوریتم Gibbs تضمین می‌کند که پارامترهای w,θ,z,ϕ به درستی صورت تخمین زده خواهند شد. در نتیجه به ازای هر کلمه، در هر یک از سندها، احتمالات تخصیص آن کلمه به هر یک از موضوعات محاسبه می‌شود. در پایان محاسبات، پارامترهای w,θ,z,ϕ تقریب زده خواهند شد [۱۸].

٤- روش پیشنهادی

در این بخش در ارتباط با بردار موضوعی بحث می‌شود که بر اساس ذائقه کاربر، توئیت‌های مشابه را شناسایی کرده و بر اساس آنها هشتگ پیشنهاد خواهد شد.

٤-۱ بردار موضوعی (TV)

جهت تشکیل بردار موضوعی هر توئیت، ابتدا مدل bag of words بر روی توئیت‌ها اعمال شده است. در این مدل، هر توئیت به صورت مجموعه‌ای از کلمات در نظر گرفته می‌شود. جزئیات بیشتر درباره تصفیه توئیت‌ها در بخش ۴-۳ (داده) آورده شده است. پس از آماده‌سازی داده‌ها با اعمال روش LDA بر روی داده‌ها با تعداد موضوع مشخص، میزان تعلق هر توئیت به هر یک از موضوعات قابل محاسبه می‌باشد. در اجرای روش مذکور، تعداد موضوعات به صورت تصاعدی از ۲۰۰۰ تا ۱۶۰۰۰ در نظر گرفته شده است

$$K=\{۲۰۰۰,۴۰۰۰,۶۰۰۰,۸۰۰۰,۱۰۰۰۰,۱۲۰۰۰,۱۴۰۰۰,۱۶۰۰۰\}$$

بدین ترتیب در هر اجرای روش LDA، برای هر توئیت یک بردار با تعداد موضوعات تعیین‌شده در نظر گرفته می‌شود و هر خانه بردار با میزان تعلق آن سند به موضوع در نظر گرفته می‌شود که در (۵) نحوه محاسبه آمده است. به عبارت دیگر خواهیم داشت

$$E_j^d=p(K_j|d)=\frac{n_j^d}{n_\cdot^d} \qquad (۵)$$

که در عبارت بالا E_j^d نمایانگر احتمال عضویت سند d به موضوع j می‌باشد. تشکیل بردار معنایی برای هر توئیت در (۶) آورده شده است

$$TV(d)=\langle E_۱^d,E_۲^d,\ldots,E_{|K|}^d\rangle \qquad (۶)$$

در این شرایط هر توئیت به یک بردار عددی قابل نگاشت است. جهت یافتن شباهت دو بردار از فاصله کسینوسی استفاده شده که جواب مناسبی در مسئله پیشنهاد هشتگ در میکروبلاگ‌ها دارد [۱۹]. فاصله کسینوسی برای دو بردار مفروض $TV_۱$ و $TV_۲$ طبق (۷) قابل محاسبه است. طبق (۷) فاصله دو بردار بین ۱- و ۱+ خواهد بود که ۱- تفاوت کامل و ۱+ شباهت کامل را نشان خواهد داد. به سبب این که عناصر بردار موضوعی را احتمالات تشکیل می‌دهند و این که مقدار احتمال عددی بین ۰ و ۱ است، بنابراین امکان وجود عدد منفی در بردار موضوعی وجود ندارد. از این رو در فاصله کسینوسی استفاده‌شده، عدد ۰ تفاوت کامل و عدد ۱+ شباهت کامل را نشان خواهد داد

$$Sim(TV_۱,TV_۲)=\frac{TV_۱.TV_۲}{|TV_۱|\times|TV_۲|} \qquad (۷)$$

٤-۲ پیشنهاد هشتگ

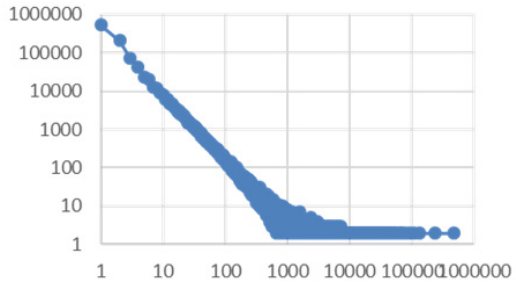
جهت پیشنهاد هشتگ از روش شبیه‌ترین N توئیت[1] استفاده شده است. در این روش به ازای هر توئیت، N توئیتی که بیشترین شباهت را دارند، پیدا شده و هشتگ‌های آنها پیشنهاد داده می‌شود. شباهت هر دو توئیت بر اساس (۷) پس از اعمال روش LDA قابل محاسبه خواهد بود. در سیستم پیشنهادی تحلیل حساسیت بر $N\in\{۱,\ldots,۵\}$ انجام شده و نتایج هر یک به دست آمده است. فرایند کلی پیشنهاد هشتگ در روش پیشنهادی در شکل ۳ قابل مشاهده می‌باشد. خانه‌های با رنگ قرمز نوآوری‌های این مقاله هستند.

ایده نوی مقاله شامل تشکیل بردار معنایی و یافتن شباهت بر اساس بردار معنایی مربوط می‌باشد بدین صورت که بردار معنایی تشکیل‌شده می‌تواند شباهت توئیت‌ها را به صورت میزان تعلق به دسته‌های مختلف بسنجد.

٤-۳ داده

برای اعمال روش پیشنهادی، از داده استفاده‌شده در [۲۰] استفاده گردیده و در ابتدا متن خالص توئیت‌ها جداسازی شده است. طی این فرایند ۶۱۷۳۲۹۶۹ توئیت از ۱۴۷۹۰۹ کاربر به دست آمده که به طور متوسط هر کاربر ۴۱۷ توئیت داشته است. سپس توئیت‌هایی که حداقل یک هشتگ در متنشان بوده است از توئیت‌هایی که هیچ هشتگی نداشته‌اند جدا شده‌اند چرا که توئیت‌های بدون هشتگ در پیشنهاد هشتگ و ارزیابی سیستم قابل استفاده نخواهند بود. به علاوه توئیت‌های بازنشرشده که متن دقیقاً یکسانی دارند و منجر به عدم کارایی سیستم پیشنهاد هشتگ می‌شوند نیز حذف شده‌اند. طی این فرایند ۱۲۳۰۹۹۱۱ توئیت به دست آمد. به طور تقریبی ۲۰٪ توئیت‌ها حداقل یک هشتگ داشته‌اند. سپس جهت پیشنهاد هشتگ، توئیت‌هایی انتخاب شده‌اند که به زبان انگلیسی باشند. این قضیه از آن جنبه اهمیت دارد که مدل LDA بر روی داده‌هایی با یک زبان می‌تواند نتیجه بهتری داشته باشد. طی این فرایند نیز ۸۳۹۶۷۴۴ توئیت باقی ماند. برای ارزیابی نیز ۸۰٪ داده‌ها برای یادگیری و ۲۰٪ نیز برای ارزیابی به صورت تصادفی انتخاب شدند. این

1. Top N Similar Tweet



شکل ۴: تعداد تکرار هشتگ‌های استفاده‌شده [۲۰].

جدول ۱: معیارهای اندازه‌گیری سیستم پیشنهادی.

	هشتگ‌های موجود در توئیت	هشتگ‌های غیر موجود در توئیت
هشتگ‌های پیشنهادی	TP	FP
هشتگ‌های غیر پیشنهادی	FN	TN

کار ۵ مرتبه صورت گرفته و نتایج به صورت میانگین نتایج به دست آمده از هر مرحله آورده شده‌اند. نهایتاً به ازای هر یک از توئیت‌های باقیمانده ۱ تا ۵ توئیت با بیشترین شباهت به دست آمده و هشتگ‌های آنها پیشنهاد شده است.

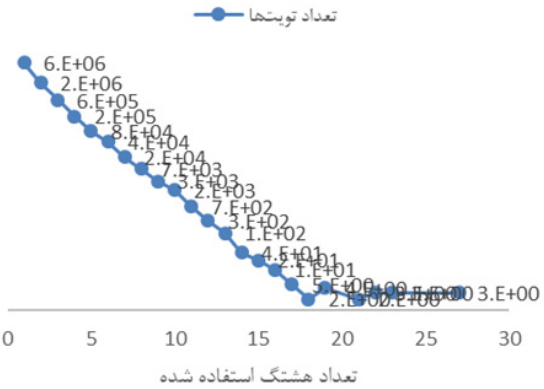
۵- تحلیل نتایج

در این بخش نتایج به دست آمده تحلیل خواهند شد. کار ارائه‌شده با روش پیشنهاد محتمل‌ترین هشتگ‌ها که در [۸] آورده شده و در بین کارهای بررسی‌شده که از روش LDA بدون تغییر استفاده کرده‌اند، بهترین نتیجه را داده است و روش SV[1] [۱۵] مقایسه شده و نتایج به دست آمده بهبود در پیشنهاد هشتگ را نشان می‌دهد. نتایج نشان‌دهنده این است که کارایی روش SV بهتر یا یکسان با محتمل‌ترین هشتگ مشابه بوده و دقت روش پیشنهادی TV[2] بالاتر از دو مورد دیگر است. همچنین لازم به ذکر است که نتایج روش [۸] و [۱۵] در شکل‌های ۷ تا ۹ تأیید می‌شود چرا که در [۸] از معیار مستقیم دقت و یا بازخورد استفاده نشده است ولی در حالت پیشنهاد ۵ هشتگ، درصد پیشنهاد ۱ هشتگ صحیح از همه بالاتر، سپس پیشنهاد ۲ هشتگ صحیح، سپس ۳ هشتگ صحیح، سپس ۴ هشتگ صحیح و پیشنهاد ۵ هشتگ تقریباً برابر با صفر می‌باشد.

شکل ۴ نمایش‌دهنده تعداد تکرارهای هشتگ‌ها است. به عبارت دیگر محور Y نمایش‌دهنده تکرار هشتگ و نمودار X تعداد هشتگ‌هایی است که این مقدار تکرار داشته‌اند. به طور مثال ۵۲۹۲۴۹ هشتگ وجود دارند که فقط یک بار تکرار شده‌اند. چنانچه در شکل نیز مشهود است این توزیع power law است که تگ‌های یک بار استفاده شده بیشترین تعداد و تگ‌هایی که بیشترین استفاده را دارند، کمترین تکرار را دارند.

در روش پیشنهادی و دیگر روش‌ها، مرحله خوشه‌بندی از مراحل ثابت می‌باشد. در دیگر روش‌ها [۸] تا [۱۱] مرحله پیشنهاد هشتگ فقط شامل خوشه‌بندی توئیت مورد بحث است. زمان اجرای روش LDA از مرتبه $O(n^2km)$ است که در آن m تعداد توئیت‌ها (سندها)، k تعداد موضوعات و n نیز تعداد کل کلمات می‌باشد. ولی در روش [۱۵] و روش کنونی، بردار TV (و SV در روش [۱۵])، هر توئیت با بقیه توئیت‌ها جهت یافتن

1. Semantic Vector
2. Topic Vector



شکل ۵: تعداد توئیت‌ها با تعداد هشتگ مشخص [۲۰].

نزدیک‌ترین توئیت مقایسه می‌گردد و لذا یک الگوریتم چینش[3] نیز اضافه می‌شود. هر توئیت با دیگر توئیت‌ها مقایسه شده و در صورتی که شباهتش بیشتر از شباهت‌های یافته‌شده باشد در لیست توئیت‌های کاندید قرار می‌گیرد. این روش چینش (چینش سریع) مرتبه اجرایی $O(m)$ دارد که به فرایند پیشنهادی اضافه می‌شود. در نتیجه زمان کلی روش پیشنهادی $O(n^2km+m)$ خواهد بود که زمان الگوریتم چینش در مقابل زمان LDA قابل چشم‌پوشی است.

۵-۱ پارامترهای ارزیابی

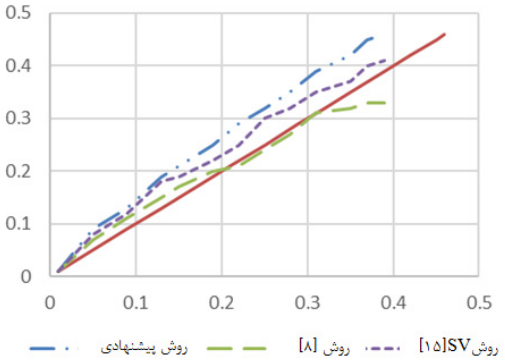
پارامترهای ارزیابی استفاده‌شده در این مقاله، شامل دقت، فراخوانی و f-score است و هر یک از این معیارها نشان‌دهنده میزان صحت پیشنهاد برای توئیت می‌باشد. در ارزیابی نتایج، هشتگ به توئیت‌هایی پیشنهاد شده که خود دارای هشتگ هستند یعنی ۲۰٪ توئیت‌های آزمایشی در هر مرحله دارای هشتگ بوده‌اند. البته هشتگ خود توئیت در فرایند پیشنهاد دخیل نبوده و در خارج از سیستم نگهداری شده است. بعد از پیشنهاد، هشتگ‌های پیشنهادی با هشتگ‌های موجود خود توئیت مقایسه شده‌اند و با بررسی صحت هشتگ‌های پیشنهادی، سیستم پیشنهاد هشتگ مورد ارزیابی قرار گرفته است.

هشتگ‌های پیشنهادی و هشتگ‌های موجود در پست، ۴ حالت نسبت به هم می‌توانند داشته باشند که در جدول ۱ نشان داده شده که در آن TP نشان‌دهنده True Positive یا همان هشتگ‌هایی است که هم توسط سیستم پیشنهاد شده‌اند و هم در خود توئیت موجود هستند. FP نشان‌دهنده False Positive است که توسط سیستم پیشنهاد داده شده ولی در توئیت موجود نیستند. FN معادل False Negative است و به معنی هشتگ‌هایی است که در توئیت هستند ولی پیشنهاد نشده‌اند. TN معادل True Negative بوده و نشانگر هشتگ‌هایی است که در توئیت موجود نیستند و توسط سیستم نیز پیشنهاد داده نشده است. با این تفاسیر هر یک از معیارهای ارزیابی قابل تعریف به صورت ریاضی می‌باشند.

تعریف ریاضی معیار دقت (P) در (۸) و تعریف معیار فراخوانی (R) در (۹) آورده شده است. در حقیقت دقت میزان درستی هشتگ‌های پیشنهادی را نشان می‌دهد و فراخوانی میزان هشتگ‌هایی که بایستی پیشنهاد شوند. معیار F-Score که در (۱۰) آورده شده نیز ترکیب میانگین معکوس این دو پارامتر است و وقتی مقدار بالاتری می‌گیرد که هر دو معیارها عدد بالایی داشته باشند [۲۱]

$$P=\frac{TP}{TP+FP} \qquad (۸)$$

3. Sort Algorithm



شکل ۶: نمودار ROC.

$$R = \frac{TP}{TP + FN} \quad (۹)$$

$$F - Score = \frac{۲ \times P \times R}{P + R} \quad (۱۰)$$

۵-۲ ارزیابی روش پیشنهادی

نمودار ROC در شکل ۶ نیز نشانگر تشخیص بهتر هشتگ‌های اشتباه است و آنها را به کاربر پیشنهاد نداده و طی این حالت نتایج در شکل‌های ۷ تا ۹ آورده شده است. چنانچه در شکل ۷ مشاهده می‌شود، دقت سیستم پیشنهادی بالاتر از دقت محتمل‌ترین هشتگ است. هنگامی که فقط یک هشتگ پیشنهاد داده می‌شود، روش محتمل‌ترین هشتگ جواب بهتری نسبت به سیستم پیشنهادی دارد. علت این است که کاربرانی که فقط از یک هشتگ استفاده می‌کنند، یا از هشتگ‌های منحصر به فرد استفاده می‌کنند که در دیگر توئیت‌ها استفاده نشده و یا از هشتگ‌های بسیار مشهور استفاده می‌کنند که اکثر کاربران از آنها استفاده کرده‌اند. مورد اول در شکل ۴ قابل مشاهده است چرا که تعداد هشتگ‌هایی که یک بار استفاده شده‌اند بسیار بالاست. به عبارت دیگر هشتگ‌هایی در داده موجود هستند که یک بار استفاده شده‌اند و این باعث کاهش دقت خواهد شد چرا که مشابه این هشتگ حتی توسط یک کاربر دیگر نیز استفاده نشده و این مسئله موجب می‌شود که فرایند یادگیری سیستم با مشکل مواجه شده و نتواند هشتگ‌های یک بار مصرف شده را پیشنهاد دهد. شکل ۸ نشان‌دهنده مقدار عددی دقت می‌باشد که مقدار دقت در هر سه روش به نسبت پایین است و عموما در بازه ۰/۰۵ تا ۰/۳ می‌باشد. علت این امر نیز از مشکل روش LDA در متون کوتاه است. همچنین شکل ۹ نیز میزان عددی بازخوانی را نمایش می‌دهد که این اعداد در حدود ۰/۵ برای روش TV است و نمایشگر کارایی روش TV می‌باشد. تحت شرایطی که LDA در متون کوتاه با مشکل مواجه است، ولی توانسته بازخوانی قابل قبولی داشته باشد و تقریباً نیمی از هشتگ‌هایی که می‌بایست پیشنهاد گردند، پیشنهاد داده شده‌اند.

البته سیستم توئیتر خود پیشنهاد می‌کند که برای هر توئیت بیش از ۲ هشتگ استفاده نشود. از این رو حداکثر ۵ هشتگ به کاربران پیشنهاد شده که از بین آنها انتخاب کنند. شکل ۵ نیز مؤید همین مورد است که حدود ۶ میلیون توئیت فقط از یک هشتگ استفاده کرده‌اند. به علاوه از شکل ۷ نیز معلوم است که بهترین نتایج در پیشنهاد ۳ تا ۴ هشتگ بوده که مانند گفته سیستم توئیتر است. یعنی حدود ۳ الی ۴ هشتگ پیشنهادی یک یا دو هشتگ اصلی استفاده‌شده در خود توئیت بوده‌اند. البته این نتایج در نگاه اول پایین به نظر می‌رسند چرا که حداکثر دقت به ۰/۳۵ و حداکثر فراخوانی ۰/۴۶ بوده است، روش TV برای مدل‌سازی در یافتن میزان شباهت توئیت‌ها کاربرد دارد، در حالی که از LDA استفاده کرده است. به عبارتی با اینکه روش LDA در مدل سازی متون کوتاه با مشکل مواجه است، ولی نحوه استفاده از نتایج LDA در روش TV باعث ایجاد تغییرات قابل ملاحظه در نتایج بوده است. با این که روش LDA بدون تغییر نتایج مناسبی برای متون کوتاه ندارد و در شکل‌ها نشان داده شده است، روش TV توانسته نتایج بسیار بالاتری در مقایسه با دیگر روش‌ها بدهد. افزایش حدود ۲۵٪ در F-Score مؤید این قضیه می‌باشد.

۶- نتیجه‌گیری و کارهای آینده

در این مقاله، روشی مبتنی بر Latent Dirichlet Allocation ارائه شده که جهت پیشنهاد هشتگ در محیط‌های میکروبلاگ کاربرد دارد. با استفاده از تکنیک یافتن شبیه‌ترین N توئیت و پیشنهاد هشتگ‌های آنها بر روی ۸ میلیون توئیت واقعی از سیستم توئیتر، نتیجه این بوده که سیستم پیشنهادی بهتر از روش مرسوم پیشنهاد محتمل‌ترین هشتگ جوابگو است. جهت پیشنهاد از تعداد موضوع ۲۰۰۰ تا ۱۶۰۰۰ استفاده شده که بهترین جواب در ۸۰۰۰ موضوع بوده است. نتیجه، افزایش حدود ۲۵٪ F-Score پیشنهاد هشتگ در مقایسه با دیگر روش‌های LDA برای پیشنهاد هشتگ است که عدد مناسبی می‌باشد. علت پایین‌بودن عددی نتایج شامل یکتابودن اکثر هشتگ‌ها و کوتاه‌بودن متن توئیت و استفاده کاربر از هشتگ‌های مختلف برای نشان‌دادن یک مفهوم می‌باشد که همین دقت پایین نیز از دیگر سیستم‌ها ۲۵٪ بهتر جواب داده است. در روش پیشنهادی فقط توئیت‌های یک زبان (انگلیسی) مورد بررسی قرار گرفته است. در ادامه این کار می‌توان از روش‌های دیگر مدل‌سازی موضوعی چون مدل‌سازی موضوعی بر اساس زبان و مدل‌سازی موضوعی بر اساس نویسنده آزمایش شود که در این حالت می‌توان توئیت‌های چند زبان را مقایسه کرد. به علاوه تشخیص تعداد موضوع نیز می‌تواند ادامه کار باشد که از روش‌های HDA می‌توان برای تشخیص خودکار تعداد موضوعات استفاده کرد.

مراجع

[1] N. Eltantawy and J. B. Wiest, "Social media in the Egyptian revolution: reconsidering resource mobilization theory," *International J. of Communication*, vol. 5, no. 1, 18 pp, 2011.

[2] D. Laniado and P. Mika, "Making sense of twitter," in *Proc. 9th Inte. Semantic Web Conf., ISWC'10,* pp. 470-485, Shanghai, China, 7-10 Nov. 2010.

[3] E. Otsuka, S. A. Wallace, and D. Chiu, "Design and evaluation of a twitter hashtag recommendation system," in *Proc. of the 18th International Database Engineering & Applications Symposium*, pp. 330-333, Yokohama, Japan, 7-9 Jul. 2014.

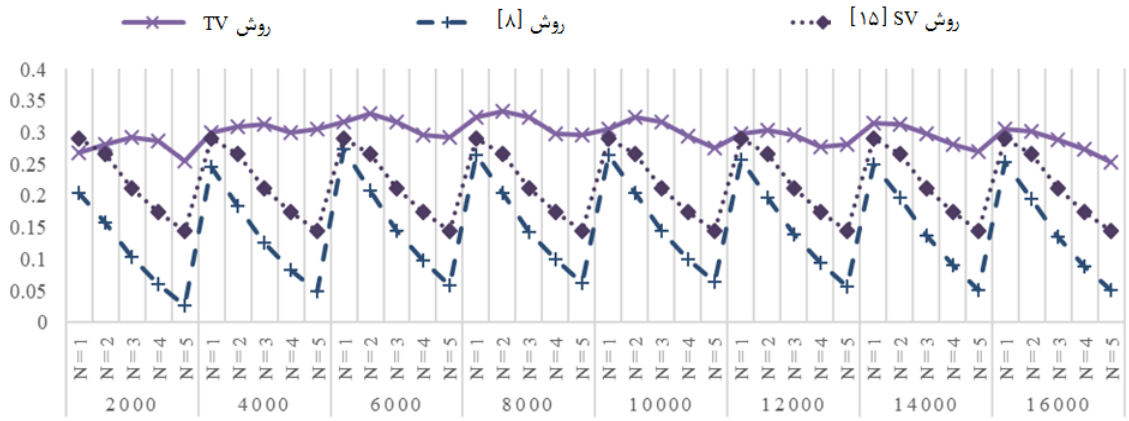
[4] J. Hillebrand, *Twitter Hashtag Analysis: Do People Really Use Them?*, 08 Aug 2016, https://www.quintly.com/blog/2014/08/twitter-hashtag-analysis/[Accessed 8 Aug. 2016].

[5] D. A. Melis and M. Saveski, "Topic modeling in twitter: aggregating tweets by conversations," in *Proc. 10th Int. AAAI Conf. on Web and Social Media, ICWSM'16*, pp. 519-522, Cologne, Germany, 17-20 May 2016.

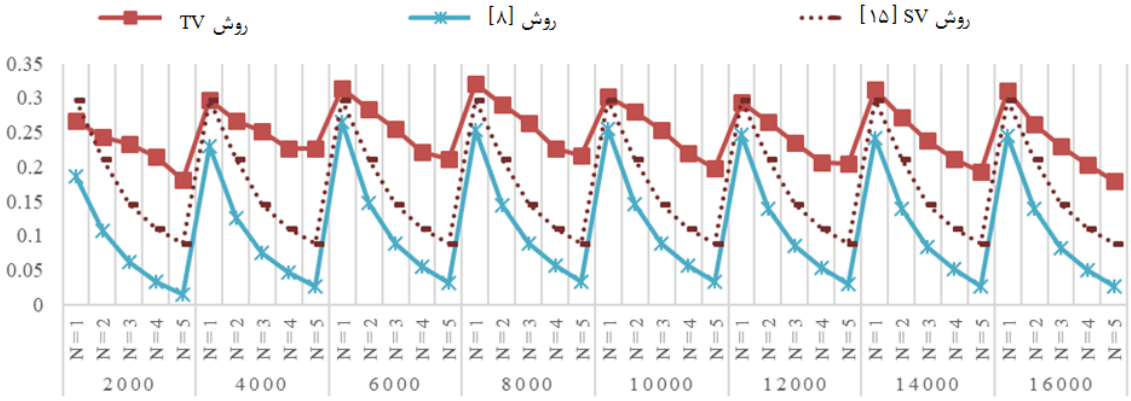
[6] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving LDA topic models for microblogs via automatic tweet labeling and pooling," in *Proc. of 36th Annual ACM Special Interest Group on Information Retrieval Conf., SIGIR'13*, pp. 889-892, Dublin, Ireland, 28 Jul.-1 Aug. 2013.

[7] N. F. N. Rajani, K. McArdle, and J. Baldridge, "Extracting topics based on authors, recipients and content in microblogs," in *Proc. of the 37th Int. ACM SIGIR Conf. on Research & Development in Information Retrieval, SIGIR'14*, pp. 1171-1174, New York, NY, USA, 6-11 Jul. 2014.

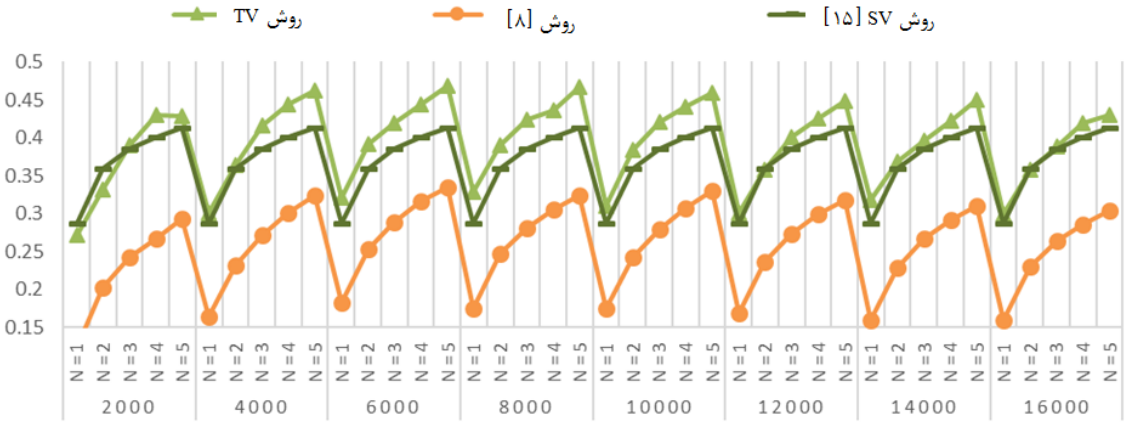
[8] F. Godin, V. Slavkovikj, W. De Neve, B. Schrauwen, and R. V. Walle, "Using topic models for twitter hashtag recommendation," in *Proc. of the 22nd Int. Conf. on World Wide Web, WWW'13 Companion*, pp. 593-596, New York, NY, USA, 13-17 May 2013.



شکل ۷: معیار F-Score روش پیشنهادی و دیگر روش‌ها.



شکل ۸: معیار دقت روش پیشنهادی و دیگر روش‌ها.



شکل ۹: معیار فراخوانی روش پیشنهادی و دیگر روش‌ها.

[9] Z. Ding, X. Huang, and Q. Zhang, "Automatic hashtag recommendation for microblogs using topic-specific translation model," in *Proc. 24th Int. Conf. on Computational Linguistics COLING'12*, pp. 265-274, Bombay, India, 11-16 Dec. 2012.

[10] J. She and L. Chen, "TOMOHA: topic model-based hashtag recommendation on twitter," in *Proc. of the 23rd Int. Conf. on World Wide Web, WWW'14 Companion*, pp. 371-372, New York, NY, USA, 07-11 Apr. 2014.

[11] Y. Gong, Q. Zhang, and X. Huang, "Hashtag recommendation using dirichlet process mixture models incorporating types of hashtags," inProc. of the Conf. on Empirical Methods in Natural Language Processing, pp. 401-410, Lisbon, Portugal, 17-21 Sept. 2015.

[12] F. Ricci, L. Rokach, B. Sharpira, and P. B. Kantor, *Recommender Systems Handbook*, Springer, 2011.

[13] D. Jannach, M. Zanker, A. Felfernig, and G. Fredrich, *Recommender Systems: An Introduction*, Cambridge University Press, Sep. 2010.

[۱۴] م. رجب‌زاده و ر. رافع، "ارائه یک سیستم توصیه‌گر ترکیبی برای تجارت الکترونیک،" *مجله مهندسی برق دانشگاه تبریز*، جلد ۴۵، شماره ۴، صص. ۹۱-۸۵، زمستان ۱۳۹۴.

[15] M. S. Tajbakhsh and J. Bagherzadeh, "Microblogging hashtag recommendation system based on semantic TF-IDF: twitter use case," in *Proc. 3rd International Symposium on Social Networks Analysis, Management and Security, SNAMS'16*, pp. 252-257, Vienne, Austria, 22-24 Sept. 2016.

[۱۶] م. محسنی، م. ازوجی و ر. قادری، "قطعه‌بندی تصویر مبتنی بر برش نرمالیزه گراف از دیدگاه میزان اطلاعات جداکننده،" *مجله مهندسی برق دانشگاه تبریز*، جلد ۴۶، شماره ۱، صص. ۳۱۰-۳۰۳، بهار ۱۳۹۵.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. of Machine Learning Research*, vol. 3, no. 1, pp. 993-1022, Mar. 2003.

[18] G. Casella and E. I. George, "Explaining the gibbs sampler," *The American Statistician*, vol. 46, no. 3, pp. 167-174, Dec. 1990.

[19] E. Zangerle, W. Gassler, and G. Specht, "On the impact of text similarity functions on hashtag recommendations in microblogging environments," *Social Network Analysis and Mining*, vol. 3, no. 4, pp. 889-898, Dec. 2013.

[20] R. Li, S. Wang, H. Deng, R. Wang, and K. C. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home locations," in *Proc. of the 18th ACM SIGKDD Inte.*

Conf. on Knowledge Discovery and Data Mining, pp. 1023-1031, Beijing, China, 12-16 Aug. 2012.

[21] J. W. Perry, A. Kent, and M. M. Berry, "Machine literature searching X Machine language, factors underlying its design and development," *American Documentation*, vol. 6, no. 4, pp. 242-254, Oct. 1995.

میر سامان تاجبخش تحصیلات خود را در مقاطع کارشناسی و کارشناسی ارشد فناوری اطلاعات به‌ترتیب در سال‌های ۱۳۹۰ و ۱۳۹۲ از دانشگاه صنعتی ارومیه و در مقطع دکتری فناوری اطلاعات در سال ۱۳۹۷ از دانشگاه ارومیه به پایان رسانده است و هم اکنون هیأت علمی همکار دانشکده مهندسی کامپیوتر دانشگاه صنعتی ارومیه می‌باشد. نام‌برده سابقه تدریس در حوزه داده کاوی، شبکه‌های اجتماعی و شبکه‌های کامپیوتری را در دانشگاه‌های ارومیه و صنعتی ارومیه داشته و دارد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: شبکه‌های اجتماعی، سیستم‌های توصیه‌گر، داده کاوی، داده‌های عظیم.

جمشید باقرزاده در سال ۱۳۷۶ مدرک کارشناسی مهندسی کامپیوتر خود را از دانشگاه صنعتی شریف و در سال ۱۳۷۹ مدرک کارشناسی ارشد مهندسی کامپیوتر خود را از دانشگاه تربیت مدرس دریافت نمود. وی مقطع دکتری خود را در سال ۱۳۸۶ از دانشگاه IIT هندوستان گرفته است. پس از اتمام تحصیلات، ایشان تدریس در دانشگاه ارومیه را آغاز کرده و در سال ۱۳۹۶ ریاست دانشکده مهندسی برق و کامپیوتر دانشگاه ارومیه را بر عهده داشته و عضو هیأت علمی دانشگاه ارومیه می‌باشد. زمینه‌های علمی مورد علاقه شامل موضوعاتی مانند داده کاوی، داده‌های عظیم، شبکه‌های اجتماعی، امنیت و سیستم‌های سلامت پزشکی می‌باشد.