استفاده از خوشه‌بندی تکاملی برای تشخیص موضوع در بلاگ‌نویسی کوچک با لحاظ‌نمودن اطلاعات شبکه اجتماعی

الهام سادات علوی، هدی مشایخی، حمید حسن‌پور و باقر رحیم‌پور کامی

چکیده: متون کوتاه رسانه‌های اجتماعی مانند توییتر اطلاعات زیادی در مورد موضوع‌های داغ و افکار عمومی ارائه می‌دهند. برای درک بهتر اطلاعات دریافتی از شبکه‌های اجتماعی، شناسایی و ردیابی موضوع امری ضروری است. در بسیاری از روش‌های ارائه‌شده در این زمینه، تعداد موضوع‌ها باید از پیش مشخص باشد و نمی‌تواند در طول زمان تغییر کند. از این منظر، این روش‌ها برای داده‌های در حال افزایش و پویا مناسب نیستند. همچنین مدل‌های تکاملی موضوعی غیر پارامتری به دلیل مشکل کمبود داده‌ها، بر روی متون کوتاه عملکرد مناسبی ندارند. در این مقاله، یک مدل خوشه‌بندی تکاملی جدید ارائه کرده‌ایم که به طور ضمنی از فرایند رستوران چینی وابسته به فاصله (dd-CRP) الهام گرفته است. در روش ارائه‌شده برای حل مشکل کمبود داده‌ها، از اطلاعات شبکه اجتماعی در کنار شباهت متنی، برای بهبود ارزیابی شباهت بین توییت‌ها استفاده شده است. همچنین در روش پیشنهادی، برخلاف اکثر روش‌های مطرح‌شده در این زمینه، تعداد خوشه‌ها به صورت خودکار محاسبه می‌شود. در واقع در این روش، توییت‌ها با احتمالی متناسب با شباهتشان به هم متصل می‌شوند و مجموعه‌ای از این اتصال‌ها یک موضوع را تشکیل می‌دهد. برای افزایش سرعت اجرای الگوریتم، از یک روش خلاصه‌سازی مبتنی بر خوشه‌بندی استفاده نموده‌ایم. ارزیابی روش بر روی مجموعه داده واقعی که در طول دو ماه و نیم از شبکه اجتماعی توییتر جمع‌آوری شده است، انجام می‌شود. ارزیابی به صورت خوشه‌بندی متون و مقایسه بین آنها می‌باشد. نتایج ارزیابی نشان می‌دهد که روش پیشنهادی نسبت به روش‌های مقایسه‌شده دارای انسجام موضوعی بهتری بوده و می‌تواند به طور مؤثر برای تشخیص موضوع بر روی متون کوتاه رسانه‌های اجتماعی استفاده گردد.

کلیدواژه: **تشخیص موضوع، خوشه‌بندی تکاملی، شبکه اجتماعی، مدل احتمالاتی.**

۱- مقدمه

در سال‌های اخیر شاهد پیشرفت سریع فناوری اطلاعات و ارتباطات هستیم. تعداد منابع خبری در این سال‌ها به صورت چشم‌گیری افزایش یافته است. اخبار بسیار زیادی روزانه به صورت برخط[1] در سراسر اینترنت منتشر شده و باعث انفجار اطلاعات الکترونیکی می‌شود. گستردگی چنین اطلاعاتی باعث شده که عملاً برای هر کاربر، آگاهی از بسیاری از این اطلاعات غیر ممکن شود. این مشکل وقتی جدی‌تر می‌شود که بخواهیم در زمان مناسبی در مورد کاری تصمیم‌گیری نماییم. به منظور تصمیم‌گیری به موقع باید این اطلاعات در حال تغییر، به طور گسترده نظارت و پیگیری شوند. به این منظور می‌توان از یک سیستم که به طور خودکار اطلاعات را گروه‌بندی و سازمان‌دهی می‌کند، برای کشف اطلاعات مرتبط استفاده نمود. کاربرانی که قصد دنبال‌کردن رویداد خبری مورد نظر خود از میان حجم بزرگی از اسناد را دارند، می‌توانند به موضوع خبری مورد نظر خود که به کمک سازمان‌دهی اطلاعات بر پایه رویداد انجام می‌شود، دسترسی داشته باشند [۱].

تشخیص موضوع و ردیابی (TDT)[2] به عنوان راه‌حلی برای حل مشکل فوق پیشنهاد شده و کار تشخیص یک رویداد جدید و پیگیری موضوع را بر عهده دارد. TDT با هدف سازمان‌دهی جریان‌های خبری با ترتیب زمانی به کار می‌رود و این سازمان‌دهی باعث تسهیل کار و کمک به کاربران خبری می‌شود و آنها را از سردرگمی در بین جریان‌های خبری نامرتبط رهایی می‌بخشد [۲].

بلاگ‌نویسی کوچک به عنوان یک رسانه اجتماعی به کاربران اجازه می‌دهد تا اطلاعات را به سرعت و بدون محدودیت خاصی گسترش دهند. این امر محبوبیت زیادی را در میان کاربران، سازمان‌ها و محققان در رشته‌های مختلف ایجاد نموده است. توییتر به عنوان دومین شبکه اجتماعی شناخته‌شده در جهان است [۳] که در این پژوهش برای ارزیابی کار تشخیص رویداد استفاده شده است. هر توییت حداکثر می تواند شامل ۲۸۰ کاراکتر باشد که این نشان‌دهنده کوتاه‌بودن متن هر توییت می‌باشد.

یکی از چالش‌های تشخیص موضوع، کثرت منابع اطلاعاتی و مشخص‌نبودن تعداد موضوع‌ها یا رویدادها می‌باشد. بنابراین استفاده از خوشه‌بندی تکاملی که ساختار و تعداد خوشه‌ها را به صورت خودکار مشخص می‌کند، برای این داده‌ها مناسب است. همچنین معمولاً هر توییت شامل متنی کوتاه است که با روش‌های تشخیص موضوع و ردیابی معمول نمی‌توان به خوبی موضوع‌ها را استخراج و توییت‌ها را خوشه‌بندی نمود [۴]. بنابراین در این پژوهش از برخی اطلاعات شبکه اجتماعی در کنار متن توییت‌ها برای تسهیل ردیابی موضوع استفاده نموده‌ایم. به این صورت که علاوه بر شباهت متنی که در اکثر روش‌ها از آن استفاده می‌شود، از شباهت به دست آمده از اطلاعات شبکه اجتماعی بین جفت توییت‌ها نیز استفاده می‌کنیم. تشخیص موضوع با استفاده از خوشه‌بندی تکاملی که برای حل مشکل کمبود داده از اطلاعات شبکه اجتماعی استفاده نماید، در کنار نقاط قوت بیان‌شده، در کارهای مرتبط کمتر مشاهده شده است.

در این مقاله یک مدل خوشه‌بندی تکاملی جدید ارائه کرده‌ایم که به

این مقاله در تاریخ ۲ خرداد ماه ۱۳۹۸ دریافت و در تاریخ ۲۹ مرداد ماه ۱۳۹۸ بازنگری شد.

الهام سادات علوی، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، ایران، (email: elham_alavi66@yahoo.com).

هدی مشایخی (نویسنده مسئول)، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، ایران، (email: hmashayekhi@shahroodut.ac.ir).

حمید حسن‌پور، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، ایران، (email: h.hassanpour@shahroodut.ac.ir).

باقر رحیم‌پور کامی، دانشکده مهندسی برق و کامپیوتر، دانشگاه علوم و فنون مازندران، بابل، ایران، (email: rc_bagher@yahoo.com).

1. Online

2. Topic Detection and Tracking

طور ضمنی از فرایند رستوران چینی وابسته به فاصله (dd-CRP)[1] الهام گرفته است. در این روش دیگر نیازی به تنظیم دستی تعداد خوشه‌های اولیه نمی‌باشد و تعداد خوشه‌ها به صورت خودکار مشخص می‌شوند. این روش بر پایه شباهت بین توییت‌ها بنا شده و از ترکیب شباهت متنی و شباهت شبکه اجتماعی برای این کار استفاده می‌نماید. توییت‌ها به صورت احتمالاتی متناسب با معیار شباهت با لینک‌هایی به هم متصل می‌شوند که مجموعه‌ای از توییت‌های به هم متصل شده به عنوان یک خوشه و هم‌موضوع در نظر گرفته می‌شوند. الگوریتم خوشه‌بندی به صورت تکراری عمل کرده و بهترین خوشه‌بندی در تکرارهای متوالی به عنوان نتیجه انتخاب می‌شود. همچنین برای دستیابی به زمان اجرای کمتر، قبل از اعمال مدل احتمالاتی تکاملی، بر روی داده‌های اولیه یک خلاصه‌سازی انجام می‌دهیم و خلاصه‌های حاصل را به عنوان ورودی به مدل احتمالاتی تکاملی می‌دهیم.

برای ارزیابی کار از پایگاه داده واقعی جمع‌آوری شده از توییتر که شامل ۲۶۱۲۷۸ توییت می‌باشد استفاده شده است. روش پیشنهادی با دو روش LDA[2] و k-means به عنوان دو روش پرکاربرد در تشخیص موضوع در شبکه اجتماعی [۵] تا [۸] مقایسه شده و نتایج حاکی از عملکرد بهتر و داشتن انسجام موضوعی بهتر روش پیشنهادی نسبت به دو روش دیگر می‌باشد.

در ادامه این مقاله ابتدا انواع کلی روش‌های تشخیص موضوع و ردیابی را بیان می‌کنیم و به کارهای مرتبطی که در این زمینه وجود دارد اشاره می‌نماییم. بخش سوم شامل جزئیات مدل پیشنهادی برای تشخیص موضوع و ردیابی بر روی مجموعه بلاگ‌نویسی‌های کوچک می‌باشد. ارزیابی مدل پیشنهادی، بیان نتایج آزمایش‌ها و تحلیل میزان کارایی روش پیشنهادی در بخش چهارم آمده است. بخش پنجم نیز شامل جمع‌بندی و پیشنهاد کارهای آتی می‌باشد.

۲- کارهای مرتبط

در ادامه به معرفی انواع روش‌های تشخیص موضوع به عنوان یکی از اصلی‌ترین وظایف TDT می‌پردازیم. تشخیص رویداد به دو روش کلی تقسیم می‌شود. در یک روش ابتدا مهم‌ترین کلمات کلیدی انتخاب شده و سپس خوشه‌بندی انجام می‌شوند [۹] و در روش دیگر اسناد مجموعه خوشه‌بندی می‌شوند [۱۰]. در روش اول، کلمات (ویژگی‌ها) به عنوان محور شناخته شده و یک موضوع توسط مجموعه‌ای از کلمات کلیدی نشان داده می‌شود. اما در روش دوم به عنوان محور، اسناد بررسی می‌شوند و یک موضوع توسط یک دسته از اسناد شناخته می‌شود.

همچنین روش‌های تشخیص رویداد یا موضوع را می‌توان با توجه به نوع متن، کاربرد یا نوع دسترسی به اطلاعات، دسته‌بندی نمود. به عنوان مثال برای تشخیص رویدادهای مشخص که شامل شناسایی رویدادهای از پیش شناخته شده می‌باشند، معمولاً از تکنیک‌های ویژگی‌محور استفاده می‌شود. زیرا این رویدادها می‌توانند تا حدی یا به طور کامل، با اطلاعات و ویژگی‌های خاص (مانند محل، زمان، نوع، شرح و غیره) که توسط کاربر یا از متن رویداد فراهم می‌شوند، مشخص شوند.

تشخیص رویداد همچنین بر اساس نوع دسترسی به اطلاعات، به دو گروه "تشخیص رویداد گذشته‌نگر" (RED)[3] و "تشخیص رویداد جدید" (NED)[4] تقسیم می‌شوند. در تشخیص رویداد گذشته‌نگر رویدادها از مجموعه داده‌هایی که از پیش جمع‌آوری شده‌اند استخراج می‌شوند، در حالی که تشخیص رویداد جدید شامل کشف رویدادهای جدید از جریان زنده در زمان واقعی است. تشخیص رویداد گذشته‌نگر از الگوریتم‌های خوشه‌بندی تکراری استفاده می‌کنند و نیاز به تمامی مجموعه داده برای سازماندهی اسناد به خوشه موضوع‌ها را دارند. روش خوشه‌بندی سلسله‌مراتبی به طور گسترده‌ای برای این کار استفاده شده است. در تشخیص رویداد جدید اطلاعات رویداد از قبل شناخته‌شده نیستند. روش‌های خوشه‌بندی به کار گرفته شده در این گروه معمولاً به صورت الگوریتم‌های افزایشی می‌باشد، زیرا تعداد موضوع‌ها مشخص نیستند و با گذشت زمان افزایش می‌یابند [۱۱].

۲-۱ روش ویژگی‌محور

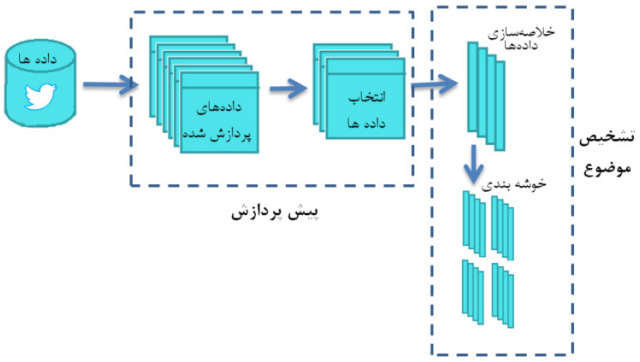
در این روش کلمات با توجه به الگوهای هم‌رخداد، خوشه‌بندی می‌شوند تا تعریف موضوع تولید شود [۱۲] و [۱۳]. بنابراین در تکنیک ویژگی‌محور، یک رویداد در جریان متن با توجه به ویژگی‌های خاصی که از نظر تعداد تکرار در حال افزایش است، پدید می‌آید. فرض اساسی این است که برخی کلمات مرتبط که زیاد استفاده می‌شوند یک رویداد را تعریف می‌کنند [۱۱]. روش‌های مبتنی بر گراف، BNgram، FPM[5] و SFPM[6] در این گروه قرار دارند [۱۴]. یکی از معایب روش‌های ویژگی‌محور این است که با توجه به این که بر اساس تجزیه و تحلیل ارتباط‌های بین کلمات انجام می‌شوند، اغلب همبستگی‌های گمراه‌کننده کلمات را نیز ضبط می‌کنند [۱۲].

مدل‌های احتمالاتی بر پایه روش ویژگی‌محوری، از روش‌های آماری و استنتاج بیزی برای استخراج مجموعه کلمات نماینده هر موضوع استفاده می‌کنند. تخصیص پنهان دیریکه (LDA) به عنوان یک مدل موضوعی برای این کار بسیار استفاده می‌شود [۱۲]. یکی از مشکلاتی که در این روش وجود دارد این است که تعداد موضوع‌ها باید از پیش به صورت دستی مشخص باشد و نمی‌تواند در طول اجرا تغییر کند [۱۳]. در روش لوو و همکارانش [۵] یک رویکرد مبتنی بر مدل موضوعی جدید برای تجزیه و تحلیل جریان آنلاین بر روی داده‌های توییتر ارائه شده که از LDA استفاده کرده است. این مدل دارای مکانیسم به روز رسانی مبتنی بر برش زمانی و پیاده‌سازی واژگان پویا است. همچنین آنها با ارائه یک روش برای اندازه‌گیری تغییرات در مدل موضوع، رویدادهای در حال ظهور را ردیابی می‌کنند.

روشی که یانگ و همکارانش [۱۵] ارائه کردند بر اساس کلمات کلیدی ساختار گرافی و تشخیص جامعه موضوعی است. آنها از معماری KeyGraph استفاده نموده‌اند و ساختار گراف را بر اساس رابطه همبستگی بین کلمات کلیدی در سند ساخته‌اند. الگوریتم تشخیص جامعه برای تقسیم‌بندی گراف کلمات کلیدی و استخراج ویژگی‌های موضوع استفاده شده و در نهایت، شباهت بین سند و موضوع‌ها محاسبه می‌گردد.

چویی و همکارانش [۱۶] روشی برای تشخیص موضوع‌های در حال ظهور در توییتر با استفاده از ابزار استخراج الگو با سود بالا (HUPM) پیشنهاد کردند. به منظور اعمال HUPM بر روی جریان‌های توییتر، یک تکنیک پنجره کشویی بر روی داده اعمال شده است. در این روش هر

1. Distance Dependent Chinese Restaurant Processes
2. Latent Dirichlet Allocation
3. Retrospective Event Detection
4. New Event Detection
5. Frequent Pattern Mining
6. Soft Frequent Pattern Mining



شکل ۱: مراحل اجرای روش پیشنهادی.

توییت به عنوان یک تراکنش و کلمات توییت به عنوان آیتم مورد استفاده قرار می‌گیرند. در نتیجه HUPM برای پیداکردن مجموعه‌ای از کلمات که دارای فرکانس بالا و سود بالا– به دست آمده بر اساس رشد فرکانس ظهورش– هستند استفاده می‌شود. نهایتاً به کمک یک درخت موضوعی، الگوهای موضوعی واقعی از الگوهای موضوع نامزد استخراج می‌شود.

۲-۲ روش سندمحور

در روش‌های سندمحوری، اسناد با استفاده از معیارهای شباهت، خوشه‌بندی می‌شوند. برای این کار شباهت یک سند ورودی را با تمام اسناد دیگر که تا کنون پردازش شده‌اند محاسبه می‌کنیم. اگر بیشترین شباهت با خوشه‌های موجود از آستانه‌ای بیشتر بود، سند به همان خوشه اختصاص داده می‌شود و در غیر این صورت یک خوشه جدید با آن سند ایجاد می‌شود. مقدار این آستانه باید از پیش تعیین شود و تنظیم مقدار مناسب برای آستانه، یکی از مشکل‌های این روش به حساب می‌آید [۱۱].

لی و همکارانش [۱۷] یک الگوریتم تشخیص رویداد جدید آنلاین مبتنی بر سند معرفی کردند. آنها از چهار ویژگی اصلی زمان، مکان، اسم مفعول و اسم فاعل برای مدل‌سازی اسناد و موضوع‌ها استفاده کرده‌اند. آنها دو مرحله "تشخیص تکرار در آغاز حوادث" و "روند دوباره تجزیه و تحلیل‌کردن حوادث پایدار" را برای حل مشکل تکه‌تکه‌شدن و ادغام موضوع‌ها به الگوریتم اضافه نمودند. نوان و همکارانش [۱۸] یک مدل برای استخراج و ردیابی رویدادها از یک جریان داده اجتماعی در زمان واقعی ارائه کرده‌اند که ویژگی‌های مبتنی بر محتوای متنی و اطلاعات به دست آمده از انتشار اخبار بین کاربران را با هم ترکیب کرده و به سیگنال‌های گسسته تبدیل نموده است. آنها ناهنجاری‌هایی را که در نوسان‌های سیگنال در طول زمان وجود دارد برای شناسایی دوره زمانی تقریبی حوادث مورد تجزیه و تحلیل قرار دادند.

در میان مدل‌های احتمالاتی بر پایه روش سندمحوری، می‌توان به روش‌های غیر پارامتری اشاره کرد که به طور خودکار شماره موضوع را در دوره‌های زمانی مختلف تعیین می‌کنند و می‌توان از آن برای مدل‌های تکاملی موضوعی استفاده کرد. این مدل‌ها معمولاً با گسترش فرایند دیریکله سلسله‌مراتبی (HDP)[1] ساخته می‌شوند [۱۹]. روش دیگری که برای این منظور می‌توان استفاده نمود، فرایند رستوران چینی (CRP)[2] می‌باشد که در این پژوهش از آن الهام گرفته‌ایم.

فرایند رستوران چینی یک فرایند تصادفی زمانی گسسته است که از آن برای ایجاد خوشه‌ها و برای مدل‌سازی فاکتورهای شباهت پنهان که در مجموعه داده‌ها وجود دارد، استفاده می‌شود. به عبارت دیگر، داده‌هایی را که یک فاکتور تشابه پنهان یکسان دارند در یک خوشه قرار می‌دهد. در فرایند رستوران چینی، یک رستوران با تعداد نامحدودی میز تصور می‌شود که هر میز نماینده یک خوشه است و عمل خوشه‌بندی عبارت است از تخصیص مشتریان به تعدادی از این میزها. مشتریان نشسته در میز یکسان در یک خوشه قرار دارند [۲۰].

استراتژی تخصیص در CRP می‌تواند به روش‌های مختلفی اجرا شود. دو روش اصلی وجود دارد که به عنوان فرایند رستوران چینی پایه (CRP پایه) و فرایند رستوران چینی مبتنی بر فاصله (dd-CRP) شناخته می‌شوند. در فرایند رستوران چینی مبتنی بر فاصله، تقسیم‌بندی با تخصیص مشتری، به جای تخصیص میز نشان داده می‌شود. dd-CRP یک تعمیم از CRP است که در آن توزیع احتمالی به جای جمعیت در هر میز، به شباهت مشتری بستگی دارد و مشتری‌های مشابه به یکدیگر متصل می‌شوند. تقسیم‌بندی داده‌ها از این اتصال‌های بین مشتری‌ها به وجود می‌آید.

روش ارائه‌شده در این مقاله در زمره روش‌های سندمحور است که بر خلاف بسیاری از روش‌های موجود [۵] و [۷] نیازی به تعیین تعداد موضوع‌ها از پیش ندارد و از اطلاعات شبکه اجتماعی برای تقویت معیار شباهت استفاده می‌کند.

۳- روش پیشنهادی برای تشخیص موضوع در بلاگ‌نویسی‌های کوچک

در این بخش، ما یک روش مدل‌سازی موضوع تکاملی برای متون کوتاه ارائه می‌دهیم. اگر بخواهیم روش پیشنهادی را در یکی از چهار گروه اصلی الگوریتم‌های خوشه‌بندی سلسله‌مراتبی[3]، مبتنی بر جزءبندی[4]، مبتنی بر چگالی[5] و مبتنی بر توری[6] جایی دهیم می‌توانیم روش پیشنهادی را تا حدودی جزء الگوریتم‌های خوشه‌بندی مبتنی بر چگالی به حساب آوریم. در روش پیشنهادی از فرایند رستوران چینی الهام گرفته‌ایم و همچنین آن را با اطلاعات شبکه اجتماعی ترکیب کرده‌ایم تا برای مدل‌سازی متن‌های کوتاه مؤثرتر شود. مراحل این روش به صورت نمایش داده شده در شکل ۱ می‌باشد. در ادامه این بخش با معرفی این مراحل به توضیح روش پیشنهادی و جزئیات استفاده از اطلاعات شبکه اجتماعی می‌پردازیم.

۳-۱ پیش‌پردازش

یکی از مراحل اصلی و تأثیرگذار در نتایج کار، مرحله پیش‌پردازش داده‌ها است. داده‌های اولیه به صورت مجموعه پست‌های ارسال‌شده به زبان لاتین در توییتر می‌باشند که علاوه بر خود توییت، شامل زمان ارسال، هشتگ‌ها، شناسه کاربری و مکان جغرافیایی توییت، به همراه دنبال‌کنندگان[7] هر کاربر است. پیش‌پردازش اصلی بر روی اطلاعات متنی انجام می‌شود. در توییتر همچون سایر اطلاعات شبکه اجتماعی دیگر، در کنار متن نوشتاری از بسیاری از علایم و شکلک‌ها استفاده می‌شود. بنابراین ابتدا اقدام به حذف این علایم و شکلک‌ها می‌نماییم و پس از

1. Hierarchical Dirichlet Processes
2. Chinese Restaurant Processes
3. Hierarchical
4. Partitioning Based
5. Density Based
6. Grid based
7. Followers

جدول ۱: پیش‌پردازش بر روی پست‌های توییتر.

پس از پزدازش	پیش از پردازش
belat prof inspir think critic write clearli lead authoritativel	Belated #InternationalWomensDay s/o to profs who inspire me to think critically, write clearly, lead authoritativel
nick stagnaro explain cooper spill over pictur crowd intern via good	SPSP Nick Stagnaro explains how cooperative spill-overs (pictured) # can be "crowded-in" (internalized) via good
seem mani busi card templat student degre understand	There don't seem to be many business card templates for students between degrees. I can understand why

حذف ایست‌واژه‌ها[1]، کلمات باقیمانده را ریشه‌یابی[2] و ریشه هر کلمه را جایگزین آن می‌کنیم. نمونه‌هایی از پست‌های پیش‌پردازش شده در جدول ۱ نشان داده شده است.

پس از تمیزکردن این اطلاعات متنی، کلمات کلیدی به دست آمده از تمامی توییت‌ها را استخراج می‌نماییم. برای این کار از کلماتی استفاده می‌کنیم که نه در اکثر اسناد تکرار شده باشند و نه در تعداد محدودی سند موجود باشند. کلماتی که در اکثر اسناد یا توییت‌ها موجود باشند نمی‌توانند تمایز بین اسناد را به خوبی نمایش دهند. همچنین با توجه به این که کاربران توییتر در توییت‌های خود ممکن است از کلمات یا اصطلاحات خاصی استفاده کنند که مختص به خودشان هستند و در توییت‌های دیگر به ندرت استفاده می‌شود، از به کارگیری این کلمات خاص نیز اجتناب می‌کنیم. تعداد این دست کلمات بسیار زیاد است و حتی ممکن است یک کلمه از این دست، در تنها یک توییت رخ داده باشند. بنابراین این کلمات نمی‌توانند گزینه مناسبی برای مقایسه بین توییت‌ها به حساب آید. به این ترتیب هر توییت به صورت مجموعه‌ای از کلمات کلیدی نمایش داده می‌شود. سپس به کمک روش tf-idf، هر متن را به یک بردار وزنی با طول تعداد کل کلمات کلیدی تبدیل می‌کنیم. ما از فرمول زیر برای وزن‌دهی به کلمات استفاده نموده‌ایم

$$w_{ij} = (1+\log tf_{ij}) \times \log \frac{N}{n_j} \qquad (1)$$

در (۱) w_{ij} وزن کلمه t_i در سند d_i، tf_{ij} تعداد تکرار کلمه t_i در سند d_i و N تعداد کل اسناد یا همان توییت‌ها می‌باشد. همچنین n_j برابر تعداد اسنادی است که کلمه t_j حداقل یک بار در آن موجود باشد.

لازم به ذکر است با توجه به آن که هدف ما در این پژوهش استخراج موضوعاتی از این توییت‌ها می‌باشد به گونه‌ای که هر موضوع شامل تعدادی توییت معنادار و مرتبط باشد، در انتهای مرحله پیش‌پردازش اندازه بردار هر توییت را محاسبه می‌نماییم و از توییت‌هایی که دارای اندازه بردار بیش از یک مقدار مشخص می‌باشد استفاده می‌نماییم. با این کار از توییت‌های کوتاه و بی‌معنا صرف نظر می‌شود.

۳-۲ خوشه‌بندی

یکی از مهم‌ترین مراحل کار تشخیص موضوع، خوشه‌بندی داده‌ها است. در روش پیشنهادی خوشه‌ها بر مبنای لینک بین توییت‌ها مشخص می‌شوند و بنابراین برای هر توییت یک لینک تعریف می‌شود که از طریق آن می‌تواند به هر توییت دیگری متصل شود. هر اتصال بین دو توییت به معنای نزدیکی و داشتن شباهت بین آنها می‌باشد. لازم به ذکر است که از هر توییت تنها یک لینک باید خارج و به توییت دیگری وارد شود، اما ممکن است چندین توییت از طریق این اتصال‌ها به یک توییت مشترک متصل شوند. پس از مشخص‌شدن لینک تمامی توییت‌ها، مجموعه‌ای از اتصال‌ها بین توییت‌ها ایجاد می‌شود که هر مجموعه اتصال یک خوشه را تشکیل می‌دهد. نمونه‌ای از این اتصال‌ها و خوشه‌بندی‌ها در شکل ۲ نشان داده شده است.

برای استخراج لینک‌های مذکور از یک روش تکراری استفاده می‌کنیم. این روش تکراری در عین خوشه‌بندی بر اساس شباهت توییت‌ها، به صورت احتمالاتی امکان ایجاد خوشه‌های جدید را ایجاد می‌کند. ابتدا مقدار اولیه‌ای برای لینک هر توییت تعریف می‌شود، مثلاً می‌توان هر توییت را به ترتیب ورود به توییت بعدی متصل کرد. به این معنا که توییت اول به توییت دوم و سپس توییت دوم به توییت سوم و الی آخر متصل می‌شوند. با این کار تمامی توییت‌ها در یک خوشه قرار داده می‌شوند (خط ۷ الگوریتم در شکل ۴). در ادامه، در هر تکرار اتصال‌ها به صورت احتمالاتی تغییر می‌کنند و در نهایت بهترین نتیجه خوشه‌بندی در تکرارهای متوالی به عنوان خوشه‌بندی نهایی ارائه می‌شود. همان طور که بیان شد هدف از این عملیات تکراری ایجاد لینک‌های جدید برای اتصال توییت‌ها به یکدیگر می‌باشد. برای این کار در هر تکرار به ترتیب، لینک هر توییت حذف و مجدداً به صورت احتمالی به یک توییت دیگر متصل می‌شود. تمامی این عملیات را می‌توان به سه بخش "جداسازی"، "محاسبه احتمالات" و "ادغام‌سازی" تقسیم نمود. توضیح‌های مربوط به این سه بخش برای نمونه توییت i در ادامه شرح داده می‌شود.

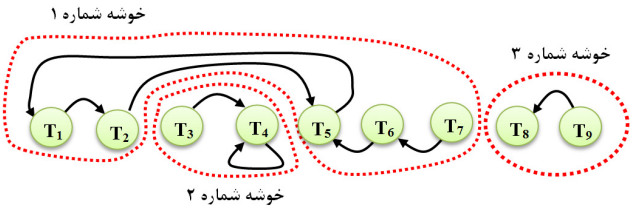
۳-۲-۱ جداسازی

در این مرحله (خط ۱۰ الگوریتم) لینک خارج‌شده از توییت i را حذف می‌کنیم. همان طور که پیش‌تر بیان شد هر توییت می‌تواند فقط یک لینک خروجی داشته باشد ولی مجاز است چندین لینک ورودی داشته باشد. بنابراین با حذف لینک خروجی توییت i که به توییت j متصل بود، ممکن است خوشه‌ای که پیش از آن، توییت i در آن قرار داشت به دو خوشه تقسیم شود. این تقسیم در صورتی ایجاد می‌شود که لینک حذف‌شده تنها ارتباط بین توییت i و j باشد. به عنوان مثال در شکل ۲ در صورتی که لینک مربوط به توییت ۱ یا ۲ یا حتی ۵ حذف شود هیچ تغییری در خوشه‌بندی ایجاد نمی‌شود ولی اگر لینک سایر توییت‌ها حذف شود در خوشه‌بندی تغییری ایجاد می‌شود و خوشه ابتدایی تقسیم می‌شود. به عنوان مثال اگر لینک توییت ۶ حذف شود خوشه ابتدایی که متشکل از توییت‌های ۱، ۲، ۵، ۶ و ۷ بود به دو خوشه با توییت‌های ۱، ۲ و ۵ در خوشه اول و توییت‌های ۶ و ۷ در خوشه دوم تقسیم می‌شود. بنابراین در صورتی که این تقسیم‌بندی صورت گرفته باشد، به خوشه‌بندی جدیدی دست می‌یابیم.

۳-۲-۲ محاسبه احتمالات

در مرحله جداسازی پس از آن که لینک خارج‌شده از توییت i حذف شد، برای ایجاد لینک جدید برای این توییت، احتمال اتصال این توییت با تمامی توییت‌ها بررسی می‌شود (خط ۱۱ الگوریتم). توییت i می‌تواند به

1. Stop Words
2. Stemming



شکل ۲: نمونه‌ای از اتصال‌ها میان توییت‌ها و خوشه‌های تشکیل‌شده توسط این لینک‌ها.

سایر توییت‌ها و یا حتی به خود متصل شود. در صورتی که توییت i تمایل به ایجاد لینک به خود را داشته باشد و هیچ لینک دیگری به توییت i وارد نشده باشد، این توییت به تنهایی یک خوشه تکی را ایجاد می‌کند. احتمال لینک به خود، بر اساس فاکتور غلظت α به دست می‌آید. هرچه اندازه این فاکتور بزرگ‌تر باشد، تمایل اتصال به خود بیشتر است و توییت‌ها به تعداد خوشه‌های بیشتری تقسیم می‌شوند.

همچنین برای محاسبه احتمال اتصال توییت i به سایر توییت‌ها، شباهت بین هر جفت از آنها را محاسبه می‌کنیم. برای این کار هم از شباهت متنی و هم از اطلاعات شبکه اجتماعی بین توییت‌ها استفاده می‌کنیم. اطلاعات متنی به دلیل محدودیت تعداد واژگان در توییت بسیار محدود است و به تنهایی قادر به ارائه اطلاعات آماری کافی برای اندازه‌گیری شباهت بین دو توییت نیست. بدین منظور از دیگر اطلاعات موجود در شبکه اجتماعی برای بهبود کار روی توییتر استفاده می‌نماییم.

– شباهت متنی

برای محاسبه شباهت متنی ابتدا فاصله کسینوسی بین دو توییت را محاسبه می‌کنیم. فاصله کسینوسی به عنوان زاویه بین دو بردار توییت تعریف می‌شود

$$distance\_cos(i,j) = 1 - \frac{w_i \cdot w_j}{|w_i||w_j|} \qquad (۲)$$

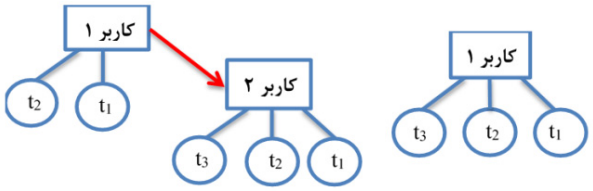
در (۲)، w_i و w_j بردارهای tf-idf دو توییت می‌باشند. سپس یک تابع تخریب بر روی این فاصله اعمال می‌نماییم. ما از عکس تابع لاجستیک برای این کار استفاده می‌کنیم تا شباهت بین داده‌هایی با فاصله کسینوسی کم و زیاد به خوبی تفکیک شود. بنابراین شباهت متنی به طور کلی به صورت زیر محاسبه می‌شود

$$content\_sim(i,j) = inverse\_logistic(distance\_cos(i,j)) \qquad (۳)$$

– شباهت شبکه اجتماعی

در این بخش برای بهبود محاسبه شباهت بین دو توییت، از اطلاعات شبکه اجتماعی شامل کلمات هشتگ‌دار، زمان ارسال توییت و ارسال‌کننده توییت برای محاسبه چهار معیار شباهت استفاده می‌کنیم. چهار متغیر متناظر به صورت v_{ij}^h، v_{ij}^t، v_{ij}^c و v_{ij}^f برای هر دو توییت t_i و t_j در مدل تعریف شده و در ادامه شرح داده می‌شوند. هر کدام از این متغیرها یک ضریب کاهنده متناظر به صورت w^h، w^t، w^c و w^f دارند که اعدادی بین صفر و یک انتخاب می‌شوند. تعیین این ضرایب به فرد متخصص و آزمایش‌های تجربی نیاز دارد که در بخش چهارم به آن می‌پردازیم.

کلمات هشتگ‌دار: هشتگ یک نماد پیشوندی با علامت "#" می‌باشد و در خدمات شبکه‌های اجتماعی و میکروبلاگینگ استفاده می‌شود. هر هشتگ با یک برچسب به صورت "#برچسب" همراه است که این



شکل ۳: نمونه روابط اجتماعی به کار برده شده در روش پیشنهادی، (الف) ارسال‌کننده مشترک و (ب) دنبال‌کردن.

برچسب می‌تواند یک اختصار یا یک کلمه یا ترکیبی از چند کلمه باشد. این عبارت هشتگ‌دار یا برچسب اجتماعی برای دسته‌بندی و به اشتراک‌گذاری پست‌ها و نظرات درباره موضوعی خاص در سطح جهانی به کار می‌رود و به این دلیل دارای اهمیت زیادی است. بررسی می‌نماییم که آیا دو توییتی که می‌خواهیم تشابه آنها را با هم محاسبه نماییم از برچسب اجتماعی یکسانی برخوردار هستند یا خیر. متغیر v_{ij}^h که نشان‌دهنده این نوع از شباهت اجتماعی می‌باشد به صورت زیر تعریف می‌شود

$$v_{ij}^h = \begin{cases} 1 & t_i \text{ and } t_j \text{ contain a similar hashtag} \\ w^h & \text{otherwise} \end{cases} \qquad (۴)$$

رابطه زمانی: در توییتر اطلاعات زمانی یکی از ویژگی‌های مهم توییت به حساب می‌آید. بسیاری از موضوع‌های طبیعی با یک چرخه عمر خاص تولید می‌شوند. به عنوان مثال بعد از یک رویداد مربوط به طوفان، توییت‌های مربوط معمولاً با یک دوره زمانی کوتاهی همراه هستند. بنابراین هرچه دو توییت از نظر زمانی به یکدیگر نزدیک‌تر باشند، با احتمال بیشتری به یک موضوع اشاره می‌کنند. همچنین در صورتی که دو توییت از نظر متنی به یکدیگر بسیار شبیه باشند ولی با فاصله زمانی زیادی از یکدیگر منتشر شده باشند، با احتمال کمتری در کنار یکدیگر در یک خوشه قرار داده می‌شوند. روش محاسبه شباهت زمانی دو توییت t_i و t_j به صورت زیر است

$$v_{ij}^t = \begin{cases} 1 & t_i^d = t_j^d \\ w^t \times (1 - e^{-\|t_i^d - t_j^d\|^2}) & \text{otherwise} \end{cases} \qquad (۵)$$

که در آن t_i^d و t_j^d نشان‌دهنده زمان ارسال توییت t_i و t_j می‌باشد.

ارسال‌کننده مشترک: اگر دو توییت از طرف یک کاربر مشترک ارسال شده باشند، این دو توییت ارسال‌کننده مشترک دارند. برای درک بهتر به شکل ۳– الف مراجعه کنید. در این شکل t_iها نشان‌دهنده توییت‌های ارسال‌شده از طرف کاربر متصل به آن می‌باشد. با توجه به این که پست‌های مربوط به یک کاربر، معمولاً از موضوع‌های مشابه تشکیل شده است، در صورتی که دو توییت توسط یک کاربر ارسال شده باشند، احتمال این که آنها از موضوع یکسانی پیروی کنند بیشتر است.

متغیر v_{ij}^c برای نشان‌دادن این نوع از شباهت اجتماعی استفاده می‌شود و به صورت زیر تعریف می‌گردد

$$v_{ij}^c = \begin{cases} 1 & t_i \text{ and } t_j \text{ are posted by one user} \\ w^c & \text{otherwise} \end{cases} \qquad (۶)$$

دنبال‌کردن: اگر کاربر ارسال‌کننده توییت اول دنبال‌کننده کاربر ارسال‌کننده توییت دوم باشد، دو توییت با هم رابطه دنبال‌کنندگی دارند. به شکل ۳– ب مراجعه نمایید که در آن پیکانی که از کاربر یک به کاربر دوم متصل است نشان‌دهنده دنبال‌کردن کاربر اول از کاربر دوم می‌باشد. بنابراین در صورتی که دو توییت با هم رابطه دنبال‌کنندگی داشته باشند،

1. **Input:** Tweets(T), $\phi = \{D(\cdot), \alpha, w^h, w^t, w^c, w^f\}$
2. **Output:** topic (cluster) number of each tweet
3. **Initialization:**
4. $N = |T|$.
5. α = mean of similarity matrix
6. iterations = Number of program iterations
7. Initialize clusters
8. **for** $k = 1$ **to** iterations **do**
9. **for** each $t_i \in T$ **do**
10. remove link of t_i
11. **for** each $t_j \in T$, $p_{i,j}$ = calculate the probability linking t_i to t_j (Eq. (4))
12. t_j^* = randomly choose a tweet with probability P
13. link t_i to t_j^* ▷ reassign the tweet
14. **end for**
15. Recalculate the topics (clusters) from links
16. L = calculate LCP for this clustering in this iteration
    Update α
17. **end for**
18. Select the topics (clustering) with maximum LCP (L_{max})

شکل ۴: الگوریتم پیشنهادی تشخیص موضوع.

به دلیل این که کاربر اولیه منافع کاربر ثانویه را به اشتراک می‌گذارد، پست‌هایشان بیشتر از نظر موضوعی مشابه هستند. متغیر متناظر به صورت زیر تعریف می‌شود

$$v_{ij}^f = \begin{cases} 1 & \text{user of } t_i \text{ follows user of } t_j \\ w^f & \text{otherwise} \end{cases} \quad (۷)$$

پس از محاسبه چهار مورد از شباهت‌های اطلاعات شبکه اجتماعی ذکرشده، مطابق (۸) مقادیر این متغیرها در هم ضرب می‌شوند

$$factor\_social(i,j) = v_{ij}^h \times v_{ij}^t \times v_{ij}^c \times v_{ij}^f \quad (۸)$$

پس از محاسبه ضریب اجتماعی دو توییت که از طریق (۸) حاصل می‌شود، این ضریب را در شباهت متنی به دست آمده از آن دو توییت ضرب می‌نماییم. هرچه ضریب اجتماعی کوچک‌تر باشد، شباهت کلی دو توییت کمتر می‌شود و احتمال اتصال این دو توییت به یکدیگر کاهش پیدا می‌کند. در نهایت پس از محاسبه شباهت متنی و ضریب اجتماعی و تعیین فاکتور غلظت α، احتمال اتصال هر توییت از (۹) به دست می‌آید

$$p(c_i^{(new)} | D, \alpha) \propto \begin{cases} \alpha \ , \text{ if } c_i^{(new)} \text{ is equal to } i \\ Factor\_social(i,j) conent\_sim(d_{ij}) \ , \text{ otherwise} \end{cases} \quad (۹)$$

که در آن $c_i^{(new)}$ تخصیص جدید یا همان لینک جدید توییت i است. با توجه به این معادله، احتمال اتصال هر توییت به خود برابر α و احتمال اتصال به سایر توییت‌ها برابر حاصل‌ضرب شباهت متنی در ضریب اجتماعی می‌باشد.

۳-۲-۳ ادغام‌سازی

پس از محاسبه احتمال‌های اتصال لینک توییت i به هر یک از توییت‌ها که در مرحله گذشته بیان شد، یکی از این لینک‌ها به صورت تصادفی متناسب با احتمال‌ها انتخاب می‌شود. بنابراین پس از برقراری این اتصال ممکن است دو خوشه با یکدیگر ادغام شوند. این ادغام در صورتی رخ می‌دهد که این اتصال به توییتی خارج از خوشه‌ای که توییت i در آن قرار دارد، برقرار شود (خطوط ۱۲ و ۱۳ الگوریتم).

در انتها پس از آن که سه مرحله "جداسازی"، "محاسبه احتمالات" و "ادغام‌سازی" برای هر یک از توییت‌ها اجرا شد، یک تکرار برنامه به اتمام می‌رسد. سپس بر روی خوشه‌بندی ایجادشده در انتهای هر تکرار معیار "انسجام موضوعی[1]" اعمال می‌شود (خط ۱۶ الگوریتم).

انسجام موضوعی یک ارزیابی کیفی از ماهیت معنایی موضوع‌ها به ما می‌دهد. این معیار بر اساس L بالاترین کلمات کلیدی یک موضوع محاسبه می‌شود و بررسی می‌نماید که آیا این کلمات به صورت جمعی موضوع را انتقال می‌دهند یا خیر. ما از LCP[2] به عنوان معیار انسجام موضوعی استفاده نموده‌ایم که به صورت زیر محاسبه می‌شود

$$LCP(t) = \sum_{j=2}^{L} \sum_{i=1}^{j-1} \log \frac{P(w_i, w_j)}{P(w_i)} \quad (۱۰)$$

در این رابطه L تعداد بالاترین کلمات کلیدی است که از هر موضوع در نظر گرفته می‌شود. $P(w_i)$ احتمال کلمه w_i و متناسب با تعداد اسنادی است که این کلمه در آن رخ داده است. $P(w_i, w_j)$ احتمال هم‌رخدادی دو کلمه کلیدی می‌باشد و متناسب با تعداد اسنادی است که هر دو کلمه کلیدی w_i و w_j با هم در آن رخ داده است. نمره بالاتر LCP برای یک موضوع، کیفیت بهتر آن موضوع را نشان می‌دهد [۲۱].

پیچیدگی زمانی این الگوریتم برابر $O(kn^2)$ می‌باشد که k تعداد تکرارها و n تعداد توییت‌ها می‌باشد. هرچه n بیشتر باشد برای رسیدن به جواب بهتر باید تعداد تکرارها را افزایش دهیم.

۳-۳ بهبود زمان اجرا

با توجه به این که مرحله خوشه‌بندی روش پیشنهادی، یک روش تکراری می‌باشد و در هر تکرار باید شباهت هر توییت با همه توییت‌های دیگر محاسبه شود، زمان اجرای این مرحله از مرتبه توان دوم نسبت به تعداد توییت‌ها می‌باشد. به منظور کاهش زمان اجرای این بخش، از یک روش خلاصه‌سازی استفاده می‌کنیم به این صورت که ابتدا کل داده‌ها را به تعدادی خوشه تقسیم می‌کنیم. ابتدا داده‌ها را به ترتیب زمان ورودشان مرتب می‌کنیم. سپس داده‌ها را به بازه‌های زمانی یکسان تقسیم کرده و بر روی هر بازه، الگوریتم خوشه‌بندی (k-means) را با توجه به شباهت متنی اعمال می‌کنیم. تعداد خوشه‌های استخراج‌شده از هر بازه، متناسب با تعداد داده‌هایی است که در آن قرار گرفته است. بدین معنا که اگر تعداد توییت‌های بازه اول بیشتر از بازه دوم باشد، بازه اول به تعداد خوشه‌های بیشتری تقسیم می‌شود. حال داده‌هایی که در خوشه یکسان قرار گرفته‌اند را با هم تجمیع می‌کنیم تا به عنوان یک داده خلاصه منفرد در نظر گرفته شوند. بدین منظور بردارهای tf-idf تمام توییت‌های قرارگرفته در یک خوشه را با هم تجمیع می‌کنیم تا بردار وزنی داده خلاصه‌شده محاسبه شود.

با توجه به خلاصه‌سازی و تجمیع توییت‌ها، باید محاسبه برخی موارد مرتبط با شباهت شبکه اجتماعی بازتعریف شوند. هر توییت شامل یک یا چند کلمه هشتگ‌دار می‌باشد که پس از تجمیع خوشه، کلمات هشتگ‌دار خوشه تجمیع‌شده برابر کلمات هشتگ‌دار تمام توییت‌های متعلق به آن خوشه می‌باشد. برای خوشه تجمیع‌شده، زمان ارسال داده خلاصه را کمترین زمان ارسال توییت متعلق به آن خوشه در نظر می‌گیریم. همچنین با توجه به این که هر توییت توسط یک فرد با شناسه کاربری واحد ارسال می‌شود، برای مشخص‌کردن شناسه کاربری خوشه تجمیع‌شده، شناسه‌های کاربری تمام توییت‌های قرارگرفته در آن خوشه را

1. Coherence
2. Log Conditional Probability

در کنار هم قرار می‌دهیم و فرض مسئله به این صورت می‌شود که هر داده جدید که از تجمیع داده‌های متعلق به یک خوشه به دست آمده، دارای چند فرستنده یا همان شناسه کاربری متفاوت می‌باشد.

در آزمایش‌های انجام‌گرفته در این مقاله توییت‌های اولیه به ۱۰۰۰ داده خلاصه‌سازی شده تبدیل شدند. هرچه تعداد توییت‌های اولیه بیشتر باشد کار خلاصه‌سازی مهم‌تر و در نتیجه نهایی تأثیر بیشتری خواهد گذاشت.

٤- آزمایش‌ها و تحلیل نتایج

در این بخش، آزمایش‌های مختلفی جهت بررسی عملکرد مدل پیشنهادی تشریح می‌شوند. در انجام آزمایش‌ها، ابعاد مختلف مدل پیشنهادی و پارامترهای مؤثر در عملکرد مدل مورد بررسی قرار گرفته است. همچنین با مقایسه این روش با روش‌های k-means و LDA که در بسیاری از مقاله‌ها برای تشخیص موضوع در شبکه اجتماعی استفاده می‌شوند [۵] تا [۸] به ارزیابی کارمان می‌پردازیم.

٤-۱ پایگاه داده و محیط پیاده‌سازی

برای ارزیابی روش پیشنهادی از پایگاه داده واقعی که از طریق Twitter API[1] جمع‌آوری شده است استفاده می‌کنیم. در این واسط محدودیت‌هایی در باب تعداد توییت‌های درخواستی به ازای هر کاربر در طول پنجره زمانی مشخص وجود دارد. به طور مثال در طول سه ساعت می‌توان ۳۰۰ توییت را جمع‌آوری نمود. پایگاه داده جمع‌آوری شده شامل ۲۶۱۲۷۸ توییت می‌باشد که از تاریخ ۳۱ دسامبر ۲۰۱۷ تا تاریخ ۱۸ مارس ۲۰۱۸ جمع‌آوری شده است. تعداد ۵۵۴۵ کاربر متفاوت در پایگاه وجود دارد. همچنین اطلاعات دیگری که شامل لیستی از دنبال‌کنندگان هر کاربر می‌باشد، همراه دیگر اطلاعات در دسترس است. تعداد کلمات کلیدی منتخب از کل توییت‌ها برابر حدود ۵۰۰۰ کلمه کلیدی است. در مرحله خلاصه‌سازی، توییت‌ها به ۱۰۰۰ توییت خلاصه شده‌اند.

پیاده‌سازی روش پیشنهادی به زبان پایتون[2] انجام شده و سیستم سخت‌افزاری به کار گرفته شده دارای پردازنده Intel Core i۷ و حافظه ۳۲ گیگابایتی می‌باشد. برای انجام پیش‌پردازش‌ها از پلتفرم موجود در پایتون به نام جعبه ابزار زبان طبیعی (NLTK)[3] استفاده نموده‌ایم.

٤-۲ معیار ارزیابی

مدل‌های موضوعی معمولاً با اندازه‌هایی نظیر سرگشتگی[4] یا احتمال حاشیه‌ای[5] مورد ارزیابی قرار می‌گیرند اما این اندازه‌ها نمی‌تواند تفسیرپذیری[6] موضوع‌ها را به خوبی نشان دهد. مطالعه‌های اخیر [۲۱] و [۲۲] نشان می‌دهد که انسجام موضوعی می‌تواند تفسیرپذیری موضوع را به خوبی نشان دهد. بنابراین برای ارزیابی کیفیت موضوع، از انسجام موضوعی استفاده می‌شود که نسبت به معیارهای دیگر نزدیک‌تر به قضاوت‌های انسانی می‌باشد. ما از LCP به عنوان معیار انسجام موضوعی استفاده می‌نماییم و متوسط نمره LCP در تمام موضوع‌ها به عنوان معیار نهایی خوشه‌بندی استفاده می‌شود.

1. https://dev.twitter.com
2. Python
3. Natural Language Toolkit
4. Perplexity
5. Marginal Likelihood
6. Interpretability

جدول ۲: تأثیر انفرادی هر یک از شباهت‌های اجتماعی.

معیار شباهت	هیچ کدام	هشتگ	دنبال‌کنندگی	ارسال‌کننده مشترک	زمانی
LCP ($L=۵$)	-۲٫۸۷	-۲٫۸۲	-۲٫۷۳	-۲٫۷۰	-۲٫۵۵

٤-۳ ارزیابی پارامترها

اگرچه در روش پیشنهادی نیازی به تعیین تعداد خوشه‌ها از پیش نیست اما سایر پارامترهای این روش از قبیل ضرایب شبکه اجتماعی $(w^h - w^f)$ و فاکتور غلظت α بر روی نتایج خوشه‌بندی تأثیر بسزایی دارد و باید از پیش مشخص شود. بنابراین پارامترهای $w^h - w^f$ توسط فرد متخصص و بنا به کاربرد مورد نظر مقداردهی می‌شوند. همان طور که پیش‌تر بیان شد هر ضریب عددی ما بین صفر و یک است و در اندازه‌گیری یک شباهت شبکه اجتماعی به کار می‌رود. پیش از مقداردهی این ضرایب به صورت دقیق، ابتدا به بررسی تأثیرگذاری هر یک از این شباهت‌ها به صورت انفرادی می‌پردازیم. بدین منظور تأثیرگذاری همه شباهت‌ها به جز یک شباهت را حذف می‌کنیم. به عنوان مثال در ستون اول جدول ۲ تأثیر همه شباهت‌ها به جز شباهت به دست آمده از رابطه زمانی را حذف می‌کنیم. همچنین در آزمایش‌های انجام‌شده در این جدول، وزن شباهت مورد بحث برابر ۰٫۵ قرار داده می‌شود به این معنا که اگر دو توییت، آن نوع از شباهت اجتماعی مورد نظر را دارا نباشند، شباهت متنی بین آنها ۵۰٪ کاهش می‌یابد. تأثیر انفرادی هر یک از شباهت‌های اجتماعی در نتیجه خوشه بندی بر روی این داده‌های آزمایشی به کمک معیار LCP در جدول ۲ نمایش داده شده است.

لازم به ذکر است که با توجه به احتمالاتی‌بودن روش و داشتن جواب‌های متفاوت در اجراهایی با تنظیم‌های یکسان، نتایج دست‌یافته در تمام جداول و نمودارها، حاصل میانگین ۲ الی ۳ اجرای پشت هم با پارامترهای یکسان می‌باشد. همان طور که در جدول ۲ مشاهده می‌کنید، شباهت حاصل از رابطه زمانی، تأثیر بسزایی در این روش ایفا می‌کند. همچنین "ارسال‌کننده مشترک" و "رابطه دنبال‌کنندگی" تقریباً تأثیر یکسانی می‌گذارند و در انتها تأثیرداشتن هشتگ مشترک از همه کمتر می‌باشد. اما تأثیرگذاری تمامی این شباهت‌ها، حتی به صورت انفرادی، نسبت به حالتی که از هیچ کدام از این شباهت‌ها استفاده نکرده‌ایم بهتر است. در جدول ۲ که تأثیر هیچ کدام از اطلاعات شبکه اجتماعی در آن دخیل نمی‌باشد LCP از همه کوچک‌تر می‌باشد و در نتیجه خوشه‌بندی بدتری نسبت به چهار وضعیت دیگر به دست آمده است.

حال با در نظر گرفتن تأثیر انفرادی هر مورد، چند آزمایش با تأثیرهای ترکیبی و ضرایب متفاوت انجام می‌دهیم. در جدول ۳ نمونه‌ای از آزمایش‌های انجام‌شده با تأثیرهای ترکیبی و ضرایب متفاوتی که به هر اطلاع اجتماعی داده‌ایم نشان داده شده است. همان طور که مشاهده می‌شود، بهترین نتیجه از خط دوم به دست می‌آید و بنابراین برای آزمایش‌های بعدی از این مقادیر استفاده می‌نماییم.

پارامتر مورد بحث بعدی، فاکتور غلظت α است که یکی از تأثیرگذارترین پارامترها در تعیین تعداد خوشه‌ها می‌باشد. می‌توان این فاکتور را در طول تکرارهای متفاوت اجرا، ثابت نگه داشت یا پس از هر تکرار، آن را به روز رسانی نمود. این به روز رسانی به گونه‌ای است که پس از اجرای هر تکرار، درصد خاصی از آلفا کاهش پیدا می‌کند. برای بررسی ارزیابی این دو روش آزمایشاتی با مقادیر اولیه متفاوت آلفا به

جدول ۳: تأثیرهای ترکیبی شباهت‌های اجتماعی.

شماره سطر	ضریب رابطه زمانی (W^t)	ضریب رابطه ارسال‌کننده مشترک (W^c)	ضریب رابطه دنبال‌کنندگی (W^f)	ضریب هشتگ (W^h)	LCP ($L=5$)
۱	۰/۵	۰/۷	۰/۷	۰/۸	-۲/۵۰
۲	۰/۶	۰/۷	۰/۷	۰/۸	-۲/۴۲
۳	۰/۵	۰/۷	۰/۷	۰/۹	-۲/۵۵
۴	۰/۶	۰/۶	۰/۶	۰/۹	-۲/۶۶
۵	۰/۶	۰/۶	۰/۶	۰/۸	-۲/۴۹
۶	۰/۶	۰/۸	۰/۸	۰/۸	-۲/۴۵
۷	۰/۶	۰/۶	۰/۶	۰/۷	-۲/۸۰
۸	۰/۵	۰/۸	۰/۸	۰/۷	-۲/۴۷
۹	۰/۶	۰/۸	۰/۸	۰/۷	-۲/۵۰

جدول ۴: انتساب پارامترهای روش پیشنهادی.

مقدار پارامتر	نوع پارامتر	توضیحات
میانگین ماتریس شباهت متنی	α	فاکتور غلظت
۰/۶	W^t	وزن شباهت زمانی در شباهت شبکه اجتماعی
۰/۷	W^c	وزن شباهت ارسال‌کننده مشترک
۰/۷	W^f	وزن شباهت دنبال‌کنندگی
۰/۸	W^h	وزن شباهت هشتگ مشترک

صورت ثابت و به صورت متغیر، انجام می‌دهیم. در آزمایش‌های انجام‌شده به این منظور، سایر مقادیر همچون ضرایب موجود در رابطه‌های شباهت‌های اجتماعی، ثابت در نظر گرفته شده است. در شکل ۵ نتیجه چند اجرا با توجه به پنج مقدار اولیه متفاوت آلفا انجام شده است.

مقدار آلفا باید متناسب با شباهت کلی داده‌ها باشد زیرا مقادیر بسیار بالاتر از مقدار شباهت کلی باعث می‌شود هر توییت به جای وصل‌شدن به توییت دیگر به خودش وصل شود و این امر باعث ایجاد تعداد خوشه‌های بسیار زیاد می‌شود که تعداد زیادی از آنها، خوشه‌های تک‌داده‌ای می‌باشند. مخصوصاً اگر مقدار آلفا در تکرارهای متفاوت برنامه یکسان باشد، دیگر به خوشه‌بندی بهینه دست نخواهیم یافت. بنابراین با توجه به این که بدون نگاه‌کردن به داده‌ها، اطلاعی از داده‌های داخلی آن نداریم، سعی می‌کنیم آن را متناسب با شباهت‌های بین توییت‌ها انتخاب نماییم. در داده‌های مورد آزمایش قرار داده شده میانگین ماتریس شباهت تقریباً برابر ۰/۰۰۴ می‌باشد و همان طور که در شکل ۵ مشاهده می‌نمایید، بهترین نتیجه در هر دو روش ثابت و متغیر در محدوده همین عدد است.

در نتیجه با توجه به این که با مقداردهی آلفا به صورت متغیر و هم به صورت ثابت به نتایج دلگرم‌کننده‌ای دست یافتیم، از ترکیب این دو روش استفاده نموده‌ایم و ابتدا آلفا را میانگین ماتریس شباهت متنی در نظر می‌گیریم و بعد از اجرای نیمی از تکرارها آن را برابر میانگین ماتریس شباهت کلی که از ضرب شباهت متنی و شباهت اجتماعی حاصل می‌شود قرار می‌دهیم.

پارامتر بعدی مورد بحث در روش پیشنهادی، تعداد تکرارهای اجرای برنامه می‌باشد. ما تمامی آزمایش‌های بالا را با ۴۰۰ تکرار انجام داده‌ایم. اما اگر تعداد داده‌ها افزایش یابد، باید تعداد تکرارها را افزایش داد زیرا همان طور که پیش‌تر بیان شد تعداد تکرارها با توجه به تعداد داده‌های ورودی باید تغییر پیدا کند. هرچه داده‌های ورودی بیشتر باشد برای رسیدن به نتیجه بهتر باید تکرارهای بیشتری انجام داد. در انتها برای جمع‌بندی می‌توان پارامترهای ورودی روش پیشنهادی را به صورت خلاصه‌شده در جدول ۴ نمایش داد.

٤-٤ تحلیل کارایی روش پیشنهادی

ما روش پیشنهادی را با دو روش k-means و LDA مقایسه نموده‌ایم. LDA برای مدل‌های موضوعی بسیار استفاده می‌شود و به عنوان یکی از روش‌هایی می‌باشد که در اکثر تحقیق‌ها در زمینه تشخیص موضوع و ردیابی مورد ارزیابی قرار داده می‌شود [۵] و [۶]. در این روش هر موضوع به صورت مجموعه‌ای از کلمات با درجه اهمیتش در آن موضوع نشان داده می‌شود و هر سند به موضوعی تخصیص داده می‌شود که کلمات تشکیل‌دهنده آن دارای بیشترین امتیاز با توجه به درجه اهمیت کلمات تشکیل‌دهنده آن موضوع می‌باشند. مهم‌ترین بخش در LDA تعیین درجه اهمیت کلمات نسبت به هر موضوع است که بر اساس هم‌رخدادی کلمات تعیین می‌شود.

همچنین با توجه به این که کار تشخیص موضوع، نوعی خوشه‌بندی می‌باشد و روش k-means به عنوان یکی از روش‌های پرکاربرد خوشه‌بندی برای تشخیص موضوع به حساب می‌آید [۷] و [۸]، برای ارزیابی کارمان از آن استفاده کرده‌ایم. در این روش ابتدا نقاط اولیه‌ای به عنوان مرکز خوشه تعیین می‌شود و نقاط دیگر بر اساس نزدیکیشان به هر یک از این مراکز، تخصیص داده می‌شود. در این روش تعیین نقاط اولیه، حایز اهمیت و در نتایج خوشه‌بندی بسیار تأثیرگذار است.

در روش پیشنهادی، شباهت بین توییت‌ها به عنوان یکی از فاکتورهای مهم این روش به حساب می‌آید. با توجه به این که در این روش از شباهت‌های به دست آمده از اطلاعات شبکه اجتماعی در کنار دیگر شباهت‌ها استفاده شده است، بهبود بیشتری در محاسبه شباهت حاصل گردیده است.

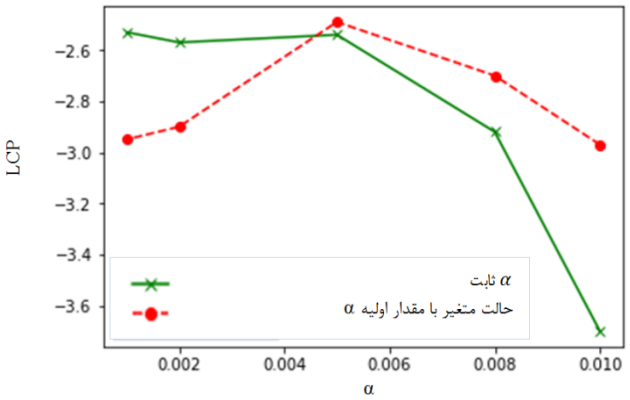
در جدول ۵ نتایج انسجام موضوعی روش‌ها بر اساس معیار LCP با سه مقدار متفاوت L نشان داده شده است. همچنین با توجه به این که تعداد موضوع‌ها در دو روش k-means و LDA باید به صورت دستی تنظیم شوند، ما این دو روش را با تعداد موضوع‌های از پیش تعریف شده متفاوت اجرا می‌کنیم. با توجه به این که تعیین دقیق تعداد موضوع‌ها برای این دو روش مشکل است، دستیابی به بهترین عملکرد این مدل‌ها سخت است. همان طور که در جدول ۵ مشاهده می‌کنید روش پیشنهادی از دو روش مورد مقایسه با توجه به هر سه مقدار متفاوت L، دارای عملکرد بهتری می‌باشد.

۵- جمع‌بندی

همان طور که اخبار موجود در بلاگ‌نویسی‌های کوچک گسترده‌تر

جدول ۵: مقایسه انسجام موضوعی روش پیشنهادی با روش‌های دیگر.

روش	k-means				LDA				روش پیشنهادی
تعداد موضوعات / LCP	۵	۱۰	۱۵	میانگین	۵	۱۰	۱۵	میانگین	–
$L=5$	−۴٫۲۳	−۴٫۷۷	−۵٫۰۷	−۴٫۶۹	−۴٫۹۹	−۳٫۵۶	−۳٫۸۳	−۴٫۱۲	−۲٫۸۷
$L=10$	−۲۲٫۵۵	−۲۵٫۵۱	−۲۶٫۹۱	−۲۴٫۹۹	−۲۱٫۵۳	−۱۴٫۵۹	−۱۶٫۴۴	−۱۷٫۵۲	−۱۴٫۱۱
$L=15$	−۵۶٫۳۳	−۶۵٫۴۵	−۶۰٫۶۶	−۶۰٫۸۱	−۴۹٫۶۰	−۳۳٫۵۲	−۴۰٫۷۵	−۴۱٫۲۹	−۳۱٫۶۰



شکل ۵: ارزیابی پارامتر آلفا.

می‌شوند، تشخیص و ردیابی موضوع از این محتواها به یک وظیفه مهم‌تر تبدیل می‌شود. این کار نه تنها به افراد کمک می‌کند که اطلاعات ضروری را درک کنند، بلکه باعث تسهیل بسیاری از برنامه‌های کاربردی می‌شود. اکثر مدل‌های تشخیص موضوعی نیاز به مشخص‌کردن تعداد موضوع‌ها به صورت دستی دارند. در این پژوهش، ما یک مدل تکامل موضوعی ارائه کرده‌ایم که تعداد موضوع‌ها را به صورت خودکار تعیین می‌نماید. مطالعه‌های تجربی بر روی مجموعه داده‌های توییتر انجام شده است. هر توییت با احتمالی به نزدیک‌ترین توییت خود متصل می‌شود و هر مجموعه از این اتصال‌ها یک خوشه را تشکیل می‌دهد. شباهت بین توییت‌ها هم از طریق شباهت متنی و هم بر اساس شباهت شبکه اجتماعی محاسبه می‌شود. این امر مشکل کمبود داده به دلیل کوتاه‌بودن طول توییت‌ها را کاهش می‌دهد. ما از چهار نوع از اطلاعات شبکه اجتماعی زمانی، هشتگ، ارسال‌کننده مشترک و دنبال‌کردن برای این کار استفاده نموده‌ایم و ارزیابی‌ها نشان داده که استفاده از این اطلاعات شبکه اجتماعی بر روی نتایج تأثیر مثبت دارد. روش پیشنهادی با روش‌های k-means و LDA مقایسه شده است. این ارزیابی بر اساس معیار انسجام موضوعی انجام شده که منعکس‌کننده تفسیرپذیری موضوع می‌باشد. نتایج تجربی حاکی از عملکرد بهتر روش پیشنهادی نسبت به دو روش دیگر می‌باشد. در این پژوهش برای محاسبه شباهت متنی دو توییت از شباهت کسینوسی استفاده شده که می‌توان برای بهبود عملکرد، از سایر روش‌های مطرح‌شده در این زمینه مانند تحلیل معنایی پنهان احتمالی (PLSA)[1] که یک شباهت مبتنی بر مجموعه سند می‌باشد استفاده نمود. همچنین با توجه به این که داده‌های اولیه به کار گرفته شده در این مقاله فاقد برخی دیگر از اطلاعات شبکه اجتماعی از قبیل نشر دوباره توییت[2]، مکان جغرافیایی و غیره می‌باشد از آنها استفاده نشده و در آینده می‌توان کارایی سایر اطلاعات شبکه اجتماعی نظیر مکان جغرافیایی [۲۳]، نشر دوباره توییت و رابطه دنبال‌کنندگی بین بیش از دو کاربر [۲۴] را در نتیجه خوشه‌بندی ارزیابی نمود. همان طور که پیش‌تر بیان شد برای بهبود زمان اجرا از یک روش خلاصه‌سازی استفاده می‌کنیم که هرچه تعداد توییت‌های اولیه بیشتر باشد کار این بخش مهم‌تر و تأثیرگذارتر می‌باشد. بنابراین برای داده‌های خیلی بزرگ می‌توان از دیگر روش‌های خوشه‌بندی به جای k-means بهره برد.

1. Probabilistic Latent Semantic Analysis
2. Retweet

مراجع

[۱] ر. بهرامی و ح. مریم، "ارائه یک الگوریتم تشخیص رویداد جدید و ردیابی موضوع در اخبار فارسی،" *مجموعه مقالات دومین همایش ملی پژوهش‌های کاربردی در علوم کامپیوتر و فناوری اطلاعات*، ۸ صص.، دانشگاه جامع علمی کاربردی، تهران، ۱۳۹۳.

[2] J. Allan, "Introduction to topic detection and tracking," In: Allan J. (eds.) *Topic Detection and Tracking. The Information Retrieval Series,* vol 12, pp.1-16, Springer, Boston, MA, USA, 2002.

[3] T. R. Zaman, R. Herbrich, J. Van Gael, and D. Stern, "Predicting information spreading in twitter," in *Proc. Workshop on Computational Social Science and the Wisdom of Crowds, Nips,* vol. 104, pp. 17599-17601, 2010.

[4] V. Krishnan and J. Eisenstein, "Nonparametric Bayesian storyline detection from microtexts," arXiv preprint arXiv:1601.04580, 2016.

[5] J. H. Lau, N. Collier, and T. Baldwin, "On-line trend analysis with topic models: # twitter trends detection topic model online," in *Proc. of COLING*, pp. 1519-1534, Mumbai, India, Dec. 2012.

[6] L. AlSumait, D. Barbara, and C. Domeniconi, "On-line lda: adaptive topic models for mining text streams with applications to topic detection and tracking," in *Proc. IEEE Int. Conf. on Data Mining*, pp. 3-12, Pisa, Italy, 15-19 Dec. 2008.

[7] K. Nur'aini, I. Najahaty, L. Hidayati, H. Murfi, and S. Nurrohmah, "Combination of singular value decomposition and K-means clustering methods for topic detection on Twitter," in *Proc. In. Conf. on Advanced Computer Science and Information Systems, ICACSIS'15*, pp. 123-128, Depok, Indonesia, 10-11 Oct. 2015.

[8] S. Li, X. Lv, T. Wang, and S. Shi, "The key technology of topic detection based on K-means," in *Proc. Int. Conf. on Future Information Technology and Management Engineering*, vol. 2, pp. 387-390, Changzhou,, China, 9-10 Oct. 2010.

[9] L. M. Aiello, *et al.*, "Sensing trending topics in Twitter," *IEEE Trans. on Multimedia*, vol. 15, no. 6, pp. 1268-1282, Jun. 2015.

[10] Y. Xiaolin, Z. Xiao, K. Nan, and Z. Fengchao, "An improved single-pass clustering algorithm internet-oriented network topic detection," in *Proc. 4th IEEE Int. Conf. on Intelligent Control and Information Processing, ICICIP'13,* pp. 560-564, Beijing, China, 9-11 Jun. 2013.

[11] F. Atefeh and W. Khreich, "A survey of techniques for event detection in twitter," *Computational Intelligence*, vol. 31, no. 1, pp. 132-164, Feb. 2013.

[12] L. M. Aiello, *et al.*, "Sensing trending topics in Twitter," *IEEE Trans. on Multimedia*, vol. 15, no. 6, pp. 1268-1282, Oct. 2013.

[13] G. Petkos, S. Papadopoulos, and Y. Kompatsiaris, "Two-level Message Clustering for Topic Detection in Twitter." SNOW-DC@ WWW, pp. 49-56, 2014.

[14] R. Ibrahim, A. Elbagoury, M. S. Kamel, and F. Karray, "Tools and approaches for topic detection from Twitter streams: survey," *Knowledge and Information Systems*, vol. 54, no. 3, pp. 511-539, Mar. 2018.

[15] S. Yang, Q. Sun, H. Zhou, Z. Gong, Y. Zhou, and J. Huang, "A topic detection method based on keygraph and community partition," in *Proc. of the Int. Conf. on Computing and Artificial Intelligence, ICCAI'18,* pp. 30-34, Chengdu, China, Mar. 2018.

[16] H. J. Choi and C. H. Park, "Emerging topic detection in twitter stream based on high utility pattern mining," *Expert Systems with Applications*, vol. 115, pp. 27-36, Jun. 2019.
[17] Y. N. Li, Y. Tao, J. N. Wang, and Y. H. Fu, "A new online new event detection algorithm based on event merging and event splitting," *Applied Mechanics and Materials*, vol. 513, pp. 2024-2030, Feb. 2014.
[18] D. T. Nguyen and J. E. Jung, "Real-time event detection for online behavioral analysis of big social data," *Future Generation Computer Systems*, vol. 66, pp. 137-145, Jun. 2017.
[19] Y. Zhang, W. Mao, and J. Lin, "Modeling topic evolution in social media short texts," in *Proc. IEEE Int. Conf. on Big Knowledge, ICBK'17,* pp. 315-319, Hefei, China, 9-10 Aug. 2017.
[20] D. M. Blei and P. I. Frazier, "Distance dependent Chinese restaurant processes," *Journal of Machine Learning Research*, vol. 12, pp. 2383-2410, 2011.
[21] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: automatically evaluating topic coherence and topic model quality," in *Proc. of the 14th Conf. of the European Chapter of the Association for Computational Linguistics*, *EACL'14,* pp. 530-539, Gothenburg, Sweden, 26-30 Apr. 2014.
[22] J. H. Lau, T. Baldwin, and D. Newman, "On collocations and topic models," *ACM Trans. on Speech and Language Processing*, Article No.: 10, Jul. 2013.
[23] Y. Fang, H. Zhang, Y. Ye, and X. Li, "Detecting hot topics from Twitter: a multiview approach," *J. of Information Scienc*, vol. 40, no. 5, pp. 578-593, Jul. 2014.
[24] J. Tang and H. Liu, "Feature selection with linked data in social media," in *Proc. of the SIAM Int. Conf. on Data Mining*, pp. 118-128, Anaheim, CA, USA, 26-28 Apr. 2012.

الهام سادات علوی در سال ۱۳۸۹ مدرک کارشناسی علوم کامپیوتر خود را از دانشگاه مازندران و در سال ۱۳۹۸ مدرک کارشناسی ارشد هوش مصنوعی خود را از دانشگاه صنعتی شاهرود دریافت کرد. زمینه پژوهشی ایشان یادگیری ماشین می‌باشد.

هدی مشایخی مدرک دکترای تخصصی رشته مهندسی کامپیوتر (نرم‌افزار) را از دانشگاه صنعتی شریف در سال ۱۳۹۲ اخذ نموده است. پیش از آن، مقاطع کارشناسی و کارشناسی ارشد را نیز در همان دانشگاه به پایان رسانیده است. یادگیری ماشین، داده‌کاوی و کاوش داده‌های بزرگ، پردازش توزیع‌شده، سیستم‌های توصیه‌گر و پشتیبان تصمیم، و مدیریت دانش از جمله علایق پژوهشی وی است.

حمید حسن‌پور مدرک دکتری خود را از دانشگاه صنعتی کوئینزلند استرالیا در گرایش پردازش سیگنال در سال ۱۳۸۳ دریافت نموده‌اند. ایشان مدرک کارشناسی ارشد خود را در گرایش هوش ماشین در سال ۱۳۷۵ از دانشگاه صنعتی امیرکبیر، و مدرک کارشناسی خود را در سال ۱۳۷۲ در گرایش سخت‌افزار از دانشگاه علم و صنعت ایران اخذ نموده‌اند. دکتر حسن‌پور در طی سال‌های ۱۳۸۴ تا ۱۳۸۶ به عنوان عضو هیأت علمی در دانشکده مهندسی برق و کامپیوتر دانشگاه صنعتی بابل فعالیت داشتند؛ سپس به دانشکده کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود انتقال یافتند. پردازش سیگنال، پردازش تصویر، داده‌کاوی، و پردازش متن از جمله علایق پژوهشی وی است.

باقر رحیم‌پور کامی در سال ۱۳۸۴ مدرک کارشناسی مهندسی کامپیوتر- نرم‌افزار خود را از دانشگاه علوم و فنون مازندارن و در سال ۱۳۸۶ مدرک کارشناسی ارشد مهندسی کامپیوتر- نرم‌افزار خود را از دانشگاه صنعتی امیرکبیر دریافت نمود. ایشان در سال ۱۳۹۷ موفق به اخذ درجه دکتری مهندسی کامپیوتر- هوش مصنوعی از دانشگاه صنعتی شاهرود گردید. از سال ۱۳۸۸ تاکنون نیز به عنوان عضو هیأت علمی در دانشگاه علوم و فنون مازندران مشغول به فعالیت می‌باشد. زمینه‌های علمی ایشان شامل داده‌کاوی، سیستم‌های توصیه‌گر، شناسایی رویداد در شبکه‌های اجتماعی می‌باشد.