DOR :20.1001.1.2538161.1401.7.4.2.3

تاریخ دریافت: ۱۴۰۱/۰۷/۲۲
تاریخ پذیرش:۱۴۰۱/۰۹/۱۹

نوع مقاله: پژوهشی

ارائه مدلی مبتنی بر ویژگی‌های زمانی برای تشخیص شایعات فارسی در شبکه اجتماعی توییتر

محمدرضا حسین‌پور نفوتی
کارشناسی ارشد مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر(پلی‌تکنیک)، تهران
پست الکترونیکی:hosseinpoor@aut.ac.ir

فاطمه کشاورز کوهجردی*
استادیار گروه علوم کامپیوتر، دانشگاه شاهد، تهران، ایران
پست الکترونیکی:f.keshavarz@shahed.ac.ir

علیرضا باقری
دانشیار گروه مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر(پلی‌تکنیک)، تهران
پست الکترونیکی:ar_bagheri@aut.ac.ir

چکیده

امروزه با پیشرفت فناوری و رشد چشم‌گیر شبکه‌های اجتماعی برخط، بسیاری از مردم جهت اطلاع از آخرین اخبار روز به این دسته از رسانه‌ها رجوع می‌کنند. این شبکه‌ها با تسهیل در ایجاد و انتشار پیام‌ها، امکان پخش شایعات را افزایش داده و در نتیجه اخبار حاصل از آن‌ها، از اعتبار کمتری نسبت به رسانه‌های سنتی برخوردار شده است. تشخیص خودکار شایعات از میان پیام‌های منتشر شده در شبکه‌های اجتماعی برخط، نیازی است که در سال‌های اخیر مطرح شده و بسیاری از دانشمندان در تلاش بوده‌اند تا راه‌حلی برای آن ارائه دهند. بنابراین ارائه روش‌هایی برای تشخیص شایعات، بیشتر از گذشته مورد توجه قرار گرفته است. در این مقاله، روشی برای تشخیص شایعات فارسی در شبکه اجتماعی توئیتر ارائه می‌شود. روش ارائه شده ابتدا کل پیام‌های مجموعه‌داده را پردازش و پیام‌های مشابه را خوشه‌بندی می‌کند و برچسب اکثریت پیام‌ها در هر خوشه را به‌عنوان برچسب آن خوشه در نظر می‌گیرد. سپس یکسری ویژگی از این خوشه‌ها استخراج می‌شود. در این مقاله، برای اولین بار ویژگی‌های مربوط به زمان در این مسئله استفاده شده است. با استفاده از این خوشه‌های برچسب‌خورده و ویژگی‌های استخراج شده یک مدل درخت تصمیم آموزش داده می‌شود. مدل حاصل نهایتا قابلیت این را دارد که شایعه بودن خوشه پیام‌های منتشر شده در شبکه اجتماعی توئیتر را تشخیص دهد. ارزیابی‌های انجام شده نشان می‌دهد که این مدل قادر است شایعه بودن خوشه پیام‌های شامل پنج پیام یا بیشتر را با دقت و فراخوانی ۰/۸۴۷ و ۰/۸۶۶ تشخیص دهد.

واژه‌های کلیدی: شبکه اجتماعی، شبکه اجتماعی توئیتر، تشخیص شایعات فارسی، استخراج ویژگی، ویژگی‌های مربوط به زمان

* نویسندهٔ مسئول

۱. مقدمه

در سال‌های اخیر و با افزایش استفاده از شبکه‌های اجتماعی برخط، آن‌ها به بستری مناسب جهت انتشار اطلاعات تبدیل شده‌اند. پیدایش نمونه‌هایی از این شبکه‌ها مانند توییتر[1]، فیسبوک[2] و اینستاگرام[3] و محبوبیت بالایی که به دست آورده‌اند، در کنار دسترسی آسان به آن‌ها و عدم کنترل محتوای به اشتراک گذاشته شده به واسطه کاربران آن‌ها به شدت بر نحوه انتشار اخبار تاثیر گذاشته است. با وجود این‌که غالبا از شبکه‌های اجتماعی برای انجام صحبت‌های روزمره استفاده می‌شود، این شبکه‌ها در انتشار اخبار و سایر اطلاعات مهم نیز نقش بسزایی دارند [۱].

بر اساس تعریف دیفونزو و بوردیا [۲]، یک پیام برای این‌که در دسته شایعه قرار گیرد باید سه ویژگی اصلی داشته باشد (۱) پیام بیان شده باید جنبه خبری داشته باشد (۲) صحت آن خبر در زمان انتشار از سوی مقامات معتبر تایید یا رد نشده باشد و (۳) آن پیام در میان مردم مورد توجه قرار گرفته باشد.

اخبار را صرف نظر از مکان و نحوه انتشارشان می‌توان به سه دسته کلی اخبار صحیح، اخبار غلط و شایعه به‌عنوان خبری که هنوز صحت یا عدم صحت آن تایید نشده است تقسیم بندی کرد [۳]. خود شایعات را نیز می‌توان به دو دسته شایعات بلندمدت و کوتاه مدت تقسیم کرد. شایعات بلندمدت آن‌هایی هستند که برای دوره‌های زمانی طولانی در بین مردم چرخیده‌اند، بدون این‌که صحت یا عدم صحت آن‌ها توسط نهادی معتبر یا شخص مربوطه تایید یا رد شود. دسته دوم از شایعات، آن‌هایی هستند که معمولا در طول شرایط بحرانی ایجاد شده، به سرعت بین مردم پخش می‌شوند و با گذشت زمانی اندک صحت آن‌ها تایید یا رد می‌شوند [۳].

تشخیص زودهنگام شایعات برای خبرگزاری‌ها نیز اهمیت بالایی دارد. در جهان امروز، خبرگزاری‌ها برای اول بودن در پوشش اخبار مختلف رقابت سنگینی را در پیش دارند زیرا اکثر این اخبار خیلی زود اهمیت خود را در میان مردم از دست داده و با گذشت زمانی اندک از انتشار آن‌ها، خوانندگان کمتری به این اخبار توجه خواهند کرد. همچنین خبرگزاری که پیش از سایرین، به بررسی صحت شایعات مورد توجه مردم بپردازد، خوانندگان بیشتری را جذب خود خواهد کرد. بنابراین رقابت بین خبرگزاری‌ها باعث شده است تا آن‌ها در جهت این‌که در انتشار یک خبر اولین باشند، به تشخیص زود هنگام شایعات علاقه‌مند باشند و در عین حال اخبار معتبر را منتشر کنند. در نتیجه، نیاز به سیستمی که بتواند شایعات منتشر شده در بین شبکه‌های اجتماعی برخط را شناسایی کرده و آن‌ها را از اخبار معتبر جدا سازد بیش از پیش احساس می‌شود. با افزایش حجم و سرعت پیام‌های منتشر شده در این شبکه‌ها، امر تشخیص شایعات توسط نیروی انسانی با تاخیر بالایی مواجه شده و علاوه بر این تاخیر، امکان پنهان ماندن بسیاری از این شایعات از دید ناظران نیز افزایش یافته است. در نتیجه سیستم‌های ارائه شده‌اند تا به روش‌های گوناگون تشخیص شایعات را به‌صورت خودکار انجام دهند.

روش پیشنهادی ما در زمره پژوهش‌های انجام شده در تشخیص شایعات منتشر شده در میان پیام‌های فارسی قرار دارد. پژوهش‌های انجام شده در این زمینه، عمدتا به استخراج ویژگی‌های جداکننده پیام‌های شایعه از سایر پیام‌ها در دسته‌های ظاهری، ساختاری، جمعیت‌شناسی، گرافی و زبانی پرداخته‌اند و با آموزش مدل‌های یادگیری ماشین مناسب، در جهت تشخیص خودکار پیام‌های شایعه از سایر پیام‌ها اقدام کرده‌اند. ما در روش پیشنهادی خود، ابتدا پیام‌های منتشر شده در شبکه را بر اساس مشابهت بین آن‌ها خوشه‌بندی می‌کنیم. بدین منظور پیام‌هایی که بیش از ٤ بار در شبکه اجتماعی بازنشر شده باشند را به‌عنوان خوشه‌های اولیه در نظر گرفتیم. در گام بعدی و در جهت ادغام خوشه‌های اولیه مربوط به یک رویداد خبری، ابتدا فاصله هر خوشه با سایر موضوعات را طوری محاسبه

1- Twitter
2- Facebook
3- Instagram

کردیم که دو خوشـــه با کلمات کلیدی مشـــترک بیشتر و زمان رخداد نزدیک‌تر، فاصله کمتری از هم بگیرند. سپس خوشـــه‌های اولیه که به یکدیگر نزدیک‌تـــر بودند را با هم ادغام کرده و خوشه‌های نهایی مختلف موجود در مجموعه پیام‌ها را از میان آن‌ها اســـتخراج کردیم. در گام بعد، با در اختیار داشتن خوشـــه‌های نهایی و پیام‌های تشکیل‌دهنده آن خوشه‌ها، به استخراج ویژگی‌های جداکننده خوشه‌های شـــایعه پرداختیم. ویژگی‌های مطالعه‌شـــده در این مقاله شـــامل ویژگی‌های مربوط به متن (مانند تعداد هشـــتگ‌ها و موجودیت‌هـــای نامی مجزای هر خوشـــه)، ویژگی‌های مربوط به کاربر (مانند میانگین طول عمر و دنبال‌شونده‌ها یا دنبال‌کننده‌های کاربران منتشرکننده پیام‌های هر خوشه)، ویژگی‌هـــای مربوط به انتشـــار (مانند میانگیـــن پیام‌های بازنشـــر یا پاسخ تشـــکیل‌دهنده هر خوشه) و ویژگی‌های مربوط به زمان (مانند طول عمر پیام‌های یک خوشه) بودند که از میان آن‌ها ویژگی‌هـــای مربوط به زمان برای اولین بار در این مقاله معرفی و بررســـی شدند. در گام آخر نیز با در دست داشتن ویژگی‌های معرف هر خوشه، یک مدل درخت تصمیم بر روی این ویژگی‌ها آموزش دادیم.

در جهـــت ارزیابی روش پیشـــنهادی در این مقاله، از مجموعه‌داده ارائه شـــده توســـط دهقانی و همکاران [۱۱] استفاده شده است. این مجموعه داده شامل ۳۵۹۸۰۴۹ عدد پیام مختلف جمع‌آوری شده از شبکه اجتماعی توییتر بوده که تعداد ٤۳٤٥ عدد از آن‌ها به‌عنوان شایعه برچسب‌گذاری شـــده‌اند. ارزیابی‌های انجام شـــده در یـــک مجموعه‌داده نامتقارن با بیش از سه و نیم میلیون پیام مختلف نشان داد که مدل ما قادر اســـت دسته پیام‌های شایعه را تنها پس از انتشار ٥ پیام مختلف از آن‌ها، با دقت و فراخوانی ۸٤۷/. و ۸٦٦/. از سایر خوشه‌ها تشخیص دهد.

ساختار مقاله به این صورت است: در بخش ۲، پیشینه تحقیق بیان می‌شـــود. در بخش ۳، روش پیشـــنهادی ارائه می‌شود. در بخش ٤، روش پیشنهادی مورد ارزیابی قرار می‌گیرد. نتیجه‌گیری نیز در بخش ٥ بیان می‌شود.

۲. پیشینه تحقیق

در جهت تشـــخیص خودکار شایعات از میان پیام‌های منتشر شده در شبکه‌های اجتماعی برخط، در دنیا مدل‌های مختلفـــی معرفی شـــده‌اند که اکثر آن‌ها بـــر پایه یادگیری ماشین و یادگیری عمیق هستند. ما و همکاران [٤] در سال ۲۰۱٦ روشـــی ارائه کردند که با استفاده از تنوع اطلاعات جمع‌آوری شـــده در فواصـــل زمانی مختلـــف مربوط به هر رویداد به یادگیری مدلی مبتنی بر شـــبکه‌های عصبی بازگشـــتی می‌پرداخت. پژوهش انجام شده توسط یانگ و همکاران [٥]، اولین پژوهشـــی بود که در زمینه تشخیص شایعات بر روی شبکه اجتماعی ویبو صورت می‌گرفت. این شبکه اجتماعی، سرویسی رسمی برای تشخیص شایعات دارد که تمرکز اصلی آن بر شناسایی شایعاتی است که به میزان وسیعی در شـــبکه پخش شده‌اند. آن‌ها در پژوهش خود، از این ســـرویس استفاده کردند و مجموعه‌داده‌ای از پیام‌های مرتبط با شایعات جمع‌آوری کردند. سپس هر یک از این پیام‌ها با توجه به این‌که در تایید آن شـــایعه بوده یا جهت رد آن نوشـــته شده بود، به ترتیب برچسب اطلاعات غلـــط و صحیح دریافت کردند. مونـــدال و همکاران [٦] در پژوهش خود به مطالعه نحوه انتشار شایعات و ویژگی‌های مربـــوط به آن‌ها در وضعیت پـــس از فاجعه پرداختند. در ادامـــه بارنوال و همـــکاران [۷] در پژوهش خود به مطالعه تشـــخیص شـــایعات در زمان‌های بحرانـــی پرداختند و با تحلیل مجموعه‌دادگانی که در طول زمین لرزه سال ۲۰۱٥ نپال جمع‌آوری شده بود، عملکرد روش‌های مختلف را در تشخیص شـــایعات موجود مقایسه کردند. چن و همکاران [۸] با بیان این اصل که ویژگی‌های متنی مرتبط با شایعات ممکن است در طول زمان تغییر کنند مدل پیشنهادی خود را بـــا نام CallAtRumours ارائه کردند. این مدل که بر پایه یک شـــبکه عصبی بازگشتی مبتنی بر توجه عمیق طراحی شده اســـت قادر است تا به جای اســـتفاده از ویژگی‌های استخراج شده توسط نیروی انسانی به‌صورت خودکار به یادگیری جلوه شـــایعات بپردازد، دشواری حفظ وابستگی

بین پیام‌هایی از یک شـــایعه که با فاصله بالا از هم منتشر می‌شـــوند را کاهش دهد و مشکلات مربوط به تکرار بالای پیام‌ها را برطرف ســـازد. لیو و همکاران [۹] با اشـــاره بر گروه‌بندی بدون نظارت پیام‌های منتشر شده پیرامون یک رخداد در طول زمان، روشـــی تحت نظارت برای یادگیری انجام این گروه‌بندی معرفی کردند که به‌طور چشـــمگیری کارایی مـــدل را افزایش داد. علاوه بر افزایش کارایی مدل، توسط این روش به راحتی قادر خواهیم بود پست‌های غیر مرتبط را از گروه‌های جدا شده حذف کنیم.

به دلیل تفاوت‌هـــای فرهنگی و زبانـــی موجود در هر جامعه، مدل‌ها ارائه شـــده به‌صورت مستقیم در شناسایی پیام‌های فارسی منتشر شده در شبکه‌های اجتماعی عملکرد خوبی از خود نشان نخواهند داد. پژوهش‌های انجام شده در شناسایی شایعات منتشرشده در میان پیام‌های فارسی بسیار محدود است و اغلب بر روی حجم کوچکی از پیام‌ها که به‌صورت متوازن در دو گروه شایعه و غیرشایعه قرار گرفته‌اند، صورت گرفته اســـت. اولین پژوهش انجام شده در این زمینه توسط زمانی و همکاران [۱۰] در سال ۱۳۹۵ انجام شد. به دلیل عدم وجود ابزارهای دقیق پردازش زبان طبیعی برای زبان فارسی، تمرکز اصلی این مقاله بر روی استخراج ویژگی‌های ساختاری شایعات از طریق استخراج گراف ارتباطی کاربران تولید کننده این پیام‌ها است. آن‌ها شـــبکه‌اجتماعی توییتر را برای انجـــام آزمایش‌های خود انتخاب کردند. مدل آن‌ها، با روش پیشـــنهادی ارائه شـــده در این مقاله چندین تفاوت عمـــده دارد. اولین تفاوت آنجا ظاهر می‌شـــود که مدل آن‌ها در یک مجموعه داده متقارن و کم‌حجم اجرا و ارزیابی شـــده است. اما در این مقاله، در جهت نزدیک کردن شـــرایط اجرای مـــدل به دنیای واقعی، مدل در میان دادگانی نامتقارن و با حجم بسیار بالا ارزیابی شده است. همچنین آن‌ها در جهت دوری از پردازش زبان فارسی، ویژگی‌های زبانی را نادیده گرفتند و برای جبران آن ویژگی‌هـــای گرافی را به مدل خـــود اضافه کردند. این دســـته ویژگی‌ها مانند رتبه صفحه کاربر برای شکل‌گیری نیازمند زمان بیشـــتری هســـتند. در مقابل در این مقاله با به کارگیری ویژگی‌های زمانی، نشـــان داده شده است که تشخیص خوشه پیام‌های شـــایعه تنها پس از انتشار پنج پیام مختلف در آن‌ها امکان‌پذیر است.

پژوهـــش بعدی در زمینه تشـــخیص شـــایعه در میان پیام‌های فارسی در سال ۱۳۹۶ توسط دهقانی محمودآباد و همکاران [۱۱] انجام شد. آن‌ها با تاکید بر این نکته که در دنیای واقعی نسبت پیام‌های حاوی شایعه به پیام‌های غیرشایعه بسیار کوچک‌تر است مدل پیشنهادی خود را در جهت حذف بایاس[4] دسته‌بند به سمت داده‌های ردهٔ غیرشایعه، در میان مجموعه دادگانی نامتوازن آموزش دادند.

در ایـــن مقالـــه، از مجموعه دادگان جمع‌آوری شـــده توســـط دهقانی محمودآباد و همکاران استفاده شده است. با این حـــال، در مقاله آن‌ها به معرفی و اســـتخراج برخی ویژگی‌های محتوایی و ساختاری از متن و کاربران منتشر کننده تک تک پیام‌ها پرداخته شده است. در این مقاله، ابتدا به خوشـــه‌بندی این پیام‌هـــا در رویدادهای خبری مختلف اقدام شده است، سپس ویژگی‌های متمایز کننده خوشه‌های شایعه از سایر خوشه‌ها از میان آن‌ها استخراج شده است. به این ترتیب خوشـــه‌های شایعه به‌صورتی ساختارمندتر شکل خواهند گرفت و جداسازی آن‌ها از سایر خوشه‌ها با دقتی بالاتر انجام خواهد شد.

پژوهش بعدی توســـط زرهران و همـــکاران [۱۲] و در جهت تشـــخیص پیام‌های خبری نسبت به پیام‌های شایعه انجام شده است. جهانبخش نقده و همکارانش [۱۳] در سال ۱۳۹۸ با اســـتخراج عمل‌بیان‌های[5] موجـــود در میان متون فارسی، نشان دادند که شایعات فارسی معمولا در سه ردهٔ عمل‌بیان داستانی، سوالی و تهدیدی و در برخی موارد در ردهٔ درخواستی منتشر می‌شوند. عمل‌بیان عملی است که توســـط گوینده با بیان آن انجام می‌شود. برای مثال وقتی کسی بیان می‌کند «راس ساعت شش آنجا خواهم بود» تنها یک جمله معمولی گفته نشـــده اســـت بلکه یک عمل‌بیان از

4- Bayes
5 - Speech act

نوع «وعده دادن» صورت گرفته اســت. در کار آن‌ها برای تشــخیص عمل‌بیان‌های یــک متن از مجموعــه داده تهیه شده توسط همایونپور و ســلطانی‌پناه [۱٤] استفاده شده است. این مجموعه داده ۹۱٤٥ جمله فارسی را به هفت ردهٔ عمل‌بیان (۱) ســوالات، (۲) درخواست‌ها، (۳) بخشنامه‌ها، (٤) تهدیدها، (٥) نقل قول‌ها، (٦) اخبار و (۷) روایت‌ها تقسیم کرده اســت. آن‌ها مبنای پژوهش خود را بر نتایج تحقیق انجام شده توسط مرکز مدیریت امداد و هماهنگی عملیات رخدادهای رایانه (CERTCC[6]) که شــایعات منتشر شده در فضــای مجازی ایران را به پنج دســته (۱) اخبار دروغ از چهره‌های مشهور، (۲) پیام‌های جعلی با محتوی احساسی، (۳) شایعات انتخاباتی، (٤) اخبار کذب در مورد شبکه‌های اجتماعی و (٥) شــایعات مرتبط با خطرات تقســیم کرده است، قرار دادند.

تفــاوت اصلی کار آن‌هــا با روش پیشــنهادی در این مقاله این اســت که در کار آن‌ها شــایعات و اخبار صحیح از وبگاه‌های معتبر خبری جمع‌آوری شــده است، درحالی که در این مقاله از بین پیام‌های منتشــر شــده در شــبکه اجتماعی توییتر به تشــخیص خوشه‌های شایعه پرداخته شده اســت. پیام‌های منتشر شــده در توییتر ممکن است شــکل نوشــتاری کاملا متفاوتی از اخبار منتشر شده در وبگاه‌های رســمی داشته باشــند و همچنین طول محدود کلمات آن‌ها نیز پیچیدگی‌های تشخیص شایعات را بیشتر خواهد کرد. علاوه بر این، تعداد زیادی از پیام‌های منتشر شده در شبکه اجتماعی توییتر، به کل ماهیت خبری ندارد و از مکالمات روزمره مردم شکل گرفته است که بر نامتقارن بودن پیام‌های شایعه نسبت به سایر پیام‌ها می‌افزاید.

جهانبخش نقــده و همــکاران [۱٦-۱٥] از ویژگی‌های محتوایــی با در نظر گرفتن ســه دیــدگاه (۱) تعبیه کلمات محتوایــی[7]، (۲) عمل بیان‌ها و (۳) ســبک نــگارش[8] برای تشخیص و تایید شایعات فارسی استفاده کردند. در مقاله آن‌ها فقط از ویژگی‌های محتوایی اســتفاده شده است در حالی کــه در این مقاله علاوه بــر ویژگی‌های محتوایی از ویژگی‌هــای دیگــری مانند ویژگی زمان اســتفاده شــده اســت. همچنین در روش ارائه شــده در این مقاله، پیام‌ها را به‌صورت گروهی در دو دســته شــایعه و غیرشــایعه طبقه‌بندی می‌کند.

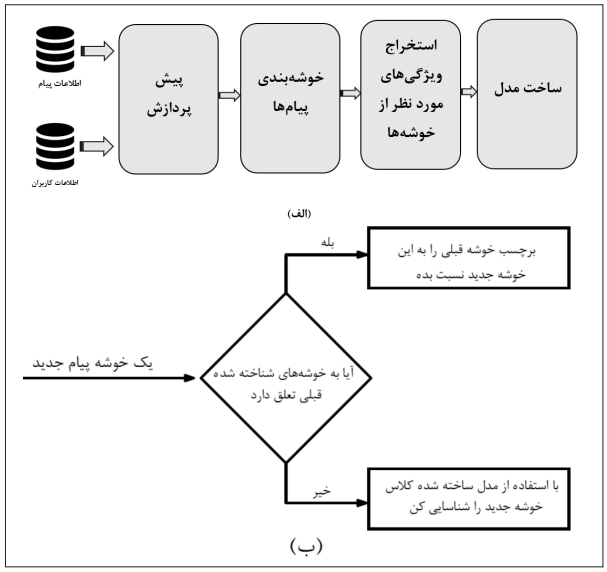
صمدی و همــکاران [۱۷] دو مدل عمیق مبتنی بر مدل BERT برای تشــخیص اخبار جعلی فارسی ارائه کرده‌اند. طبــق ادعای آن‌ها این دو روش نســبت به روش زمانی و همکاران [۱۰] بهبود داشته است. خسروی و همکاران [۱۸] با اســتفاده از مدل ترکیبی LSTM – CNN یک روش برای تشخیص شایعات مبتنی بر تحلیل ویژگی‌های محتوایی در شبکه‌های اجتماعی ارائه کردند.

از جملــه نقاط ضعف کارهای پیشــین کــه در زمینه تشخیص شــایعه فارســی انجام شده اســت می‌توان به مــوارد زیر اشــاره کــرد. کارهــای انجام شــده یا روی مجموعه داده‌های متوازن و کم‌حجم اجرا و ارزیابی شــده اســت یا فقط ویژگی‌های محتوایی را در نظر گرفتند و به ویژگی‌های ســاختاری بخصوص ویژگی‌های مربوط به زمان کمتر توجه شده است یا تشخیص شایعه تک پیامی و غیرخوشه‌ای بوده است. اما در روش پیشنهادی علاوه بر ویژگی‌های محتوایی از ویژگی‌های ساختاری نیز استفاده می‌شود و استفاده از ویژگی‌های ساختاری مربوط به زمان برای اولین بار پیشــنهاد می‌شود. همچنین ارزیابی روش روی مجموعه داده‌های با حجم نســبتا زیــاد و نامتوازن انجام می‌شود. در روش پیشنهادی، به جای این‌که تک‌تک پیام‌ها برچســب شایعه یا غیرشــایعه بگیرند، ابتدا پیام‌ها خوشه‌بندی می‌شوند و سپس یک خوشه برچسب خواهد گرفت، که این کار باعث افزایش دقت می‌شود.

۳. روش پیشنهادی

در ایــن بخش، بــه معرفی روش پیشــنهادی در جهت تشــخیص شــایعات از میان پیام‌های عــادی می‌پردازیم. در این روش پس از پیش‌پردازش، پیام‌ها ابتدا بر اســاس

6- https://www.certcc.ir/
7- Contextual word embedding
8- Writing style



شکل ۱: (الف) ساخت مدل و (ب) روند تشخیص شایعه در روش پیشنهادی

شباهت به خوشه‌های مختلف خوشه‌بندی می‌شوند. هر خوشه با برچسب اکثریت پیام‌های آن خوشه برچسب زده می‌شود. سپس ویژگی‌های مختلفی از این خوشه‌ها استخراج شده و یک مدل درخت تصمیم بر روی این خوشه‌ها آموزش داده می‌شود. پیش‌بینی ردهٔ شایعه پیام‌های جدید به دو صورت خواهد بود. ممکن است این پیام‌ها از قبل به یکی از خوشه‌های شناخته شده تعلق داشته‌اند که در این حالت بدون نیاز به پردازش بیشتر، برچسب همان خوشه را دریافت خواهند کرد. در حالت دوم، این پیام‌ها خوشه جدیدی را شکل خواهند داد که در مجموعه‌داده آموزش نبوده‌اند. در این حالت مدل درخت تصمیم برچسب خوشه جدید را پیش‌بینی خواهد کرد. در شکل ۱، روند کلی روش پیشنهادی را مشاهده می‌کنید. در ادامه، به بررسی بخش‌های مختلف روش پیشنهادی خواهیم پرداخت.

۱،۳. پیش‌پردازش

همان‌طور که بیان شد، تمرکز روش پیشنهادی بر روی طبقه‌بندی جداگانه هر پیام نیست. در عوض این پیام‌ها ابتدا به خوشه‌های مختلفی تقسیم می‌شوند و در ادامه این خوشه‌ها هستند که برچسب شایعه یا غیرشایعه دریافت خواهند کرد. پیش‌پردازش این پیام‌ها از جنبه یافتن کلیدواژه‌های هر پیام حائز اهمیت است؛ چراکه شباهت پیام‌های یک خوشه به یکدیگر از این کلیدواژه‌ها به دست خواهد آمد. در ادامه، به بررسی قسمت‌های مختلف پیش‌پردازش پیام‌ها در روش پیشنهادی و دلیل لزوم آن‌ها خواهیم پرداخت. برای پردازش متون از ابزارهای کتابخانه hazm[9] با برخی اصلاحات جزئی استفاده شده است.

حذف علامات و حروف غیرفارسی: در ابتدای پیش‌پردازش پیام‌ها، تمام حروف غیر فارسی از آن‌ها حذف می‌شوند. این حروف غالبا در سه دسته مختلف قرار می‌گیرند. در دسته اول شکلک‌ها قرار دارند که از لحاظ معنایی ارزشی برای ما ندارند و در فرایند تشخیص پیام‌های نزدیک به یک خوشه انحراف ایجاد می‌کردند. در دسته دوم علامت‌هایی مانند # یا ؟ و ! قرار دارند. این علامت‌ها اگر قبل یا بعد از کلمه‌ای چسبیده باشند، آن کلمه را از اصل خود جدا خواهند کرد. برای مثال دو پیام زیر را در نظر بگیرید:

- درگیری کشاورزان روستای جالیزی در استان ایلام با ماموران سرکوبگر نیروی انتظامی
- بحث #کشاورزان #روستای_جالیزی در #استان_ایلام با ماموران #نیروی_انتظامی به درگیری! کشید.

این پیام‌ها مشخصا به یک خوشه تعلق دارند ولی کلمات کلیدی پیام اول در پیام دوم به‌صورت هشتگ ظاهر شده‌اند که بدون پیش‌پردازش صحیح، اشتراکی با یکدیگر نخواهند داشت. موضوع دیگری که از مثال بالا قابل مشاهده است تفاوت میان علامت «#» و «_» است. در یک پیش‌پردازش صحیح علامت اول باید حذف شود ولی با حذف علامت دوم عبارتی مانند «استان_ایلام» به کلمه بی‌مفهوم «استانایلام» تبدیل خواهد شد. برای رفع این مشکل در پیش‌پردازش پیشنهادی همه علامات ابتدا با یک فاصله جایگزین شدند؛ سپس تمام فواصل قبل یا بعد از هر پیام حذف و از هر چند فاصله پشت هم، تنها یکی از آن‌ها برای مراحل بعدی حفظ شده است.

همچنین در صورتی که این علامت‌ها به‌صورت جداگانه

9- https://sobhe.ir/hazm/

نیز در متن ظاهر شـــوند یک کلمه مستقل را شکل خواهند داد که ممکن است متن‌های غیر مربوط را به یکدیگر نزدیک کند. به‌عنـــوان نمونه دو پیام زیر هیچ شـــباهت معنایی با یکدیگر ندارند ولی تعداد چهار علامت مجزا را با یکدیگر به اشتراک گذاشته‌اند که می‌تواند باعث انحراف در تشخیص پیام‌های شـــبیه به یکدیگر شود. علامت‌های هر پیام که در صورت عدم حذف به‌عنوان کلمات مستقلی شناخته خواهند شد در مثال‌های زیر با پس‌زمینه رنگی مشخص شده‌اند.

استاندار کرمانشاه : «سلبریتی‌ها باید درباره کمک‌های مردمـــی به زلزله‌زدگان گزارش دهنـــد .» من که بهش حق میدم ! ...

جهانگیـــری در ادامـــه تاکید کرد : «گرد و غبار شـــدید خوزستان هم ناشـــی از فرو ریختن مسکن مهر در زلزله کرمانشاه است.» حرفی نمیمونه برام ...

دســـته ســـوم نیز حروف غیر فارســـی مانند حروف الفبای انگلیســـی هســـتند. این حروف نیز از لحاظ معنایی در تشخیص پیام‌های شـــبیه به یکدیگر مفید نبوده و تنها احتمال خطا در این فرایند را افزایش می‌دهند. دسته اول از این حروف انگلیسی توسط خود شبکه اجتماعی به پیام‌ها اضافه می‌شـــود. بـــرای مثال در شـــبکه اجتماعی توییتر، در ابتدای بازنشـــر هر پیام، عبارت "RT" اضافه می‌شود. بخش دیگری از متون انگلیســـی نیز از نام افراد در پیام‌ها آمده اســـت. به این ترتیب در صورت حذف نکردن حروف انگلیسی از پیام‌ها، دو پیام مختلف که هر دو بازنشر بوده و کاربر یکســـانی را ذکر کرده‌اند ممکن است به اشتباه در یک خوشه قرار گیرند.

نرمال‌ســـازی متن: نرمال‌سازی متون فارسی از جهت یکسان‌ســـازی نگارش فاصله‌هـــا و نیم‌فاصله‌ها از اهمیت ویژه‌ای برخوردار است. به‌عنوان نمونه دو پیام زیر را در نظر بگیرید:

- خداحافظی ارتشـــی‌ها و مردم زلزله‌زده سرپل‌ذهاب و قدردانی مردم و اشک طرفین.
- خدا حافظی ارتشی ها و مردم زلزله زده سر پل ذهاب و قدردانی مردم و اشک طرفین.

پیام دوم به واقع همان پیام اول اســـت با این تفاوت که در آن فاصله‌گذاری بین کلمات به درســـتی رعایت نشـــده است. به این ترتیب تعداد کلمات کلیدی مشترک بین این دو پیام به شدت کم شده و بدون پیش‌پردازش صحیح ممکن است این دو پیام در یک خوشه قرار نگیرند.

نکته مهـــم دیگری که بایـــد ذکر شـــود قرارگیری این مرحله پس از بخش «حذف علامات و حروف غیرفارسی» اســـت. در آن بخش بیان کردیم که علاماتـــی مانند «–» و «_» به‌عنوان علامت‌های واصل کلمات شـــناخته شده و با یک فاصله جایگزین خواهند شد. این در صورتی است که ممکن اســـت در حقیقت بیـــن آن دو کلمه به یک نیم فاصله نیاز بوده اســـت. برای مثال عبارت «زلزله_زده» را در نظر بگیرید. این عبـــارت در بخش قبلی تبدیل به عبارت «زلزله زده» شـــده اســـت. در صورتی که پس از «حذف علامات و حروف غیرفارســـی «مرحله» نرمال‌سازی متون» انجام شده باشـــد، خطاهای تولید شده در آن بخش نیز در کنار خطاهای موجود در متون اولیه اصلاح خواهند شد.

تبدیل متن محاوره به رسمی: شـــبکه‌های اجتماعی، محلی هستند که بسیاری از مردم به زبان محاوره اقدام به انتشـــار مطالبی در آن‌ها می‌کنند. علاوه بر این بخش قابل توجهـــی از این پیام‌ها نیز به زبان رســـمی بوده یا حداقل بخشـــی از آن پیام به زبان رسمی نوشته شده است. برای مثال، نقل قول از افراد غالبا به زبان رسمی نوشته می‌شود. شکل املایی برخی کلمات ممکن است در صورت محاوره و رسمی آن متفاوت باشـــد. این تفاوت می‌تواند پیام‌هایی از یک خوشـــه را نســـبت به هم دور کرده و در محاســـبه شـــباهت پیام‌ها به یکدیگر خطاهایی را وارد کند. به‌عنوان نمونه عباراتی که در زبان محاوره به جای اســـت با حرف «ه» خاتمه می‌یابند را در نظر بگیرید. در جمله‌ای مانند «این ماشـــین دکتر مصدق تو احمدآبـــاده» کلمه احمدآباد جزو کلمات کلیدی متن محســـوب می‌شود که در شکل محاوره نوشته شده با شـــکل اصلی خود تفاوت داشته و منجر به

ایجاد اختلاف با آن خواهد شد.

علاوه بر این بسیاری از افعال وجود دارند که باید تمام حالات نوشتار رسمی و محاوره‌ای آن‌ها به یک صورت مشخص درآیند و در محاسبه شباهت پیام‌ها، تعداد تکرار، و یافتن ایست‌واژه‌ها به یک صورت مشخص پردازش شوند. در ادامه، این بخش‌ها به‌صورت دقیق‌تر شرح داده خواهند شد.

ریشه‌یابی افعال: افعال در متون فارسی می‌توانند شکل‌های مختلفی به خود بگیرند. در بخش قبلی، شکل‌های محاوره‌ای این افعال به‌صورت رسمی درآمدند. با این حال همچنان یک فعل با معنای مشخص ممکن است تنوع بالایی در جملات داشته باشد. به‌عنوان نمونه فعل گفت را در نظر بگیرید. در عباراتی مانند «جهانگیری گفت»، «جهانگیری می‌گوید»، «جهانگیری گفته است» و ... این فعل معنای مشخصی دارد اما صورت‌های املایی مختلف آن، کلماتی متفاوت از هم را شکل داده است.

ریشه‌یابی افعال از دو جنبه مورد اهمیت است. اول آن که یک فعل عام ممکن است در تشخیص پیام‌های مربوط به یک خوشه کمکی نکند و در خوشه‌های مختلف تکرار شود. در این صورت این فعل باید از میان کلمات کلیدی پیام حذف شود. ریشه‌یابی و یکی کردن صورت املایی حالات مختلف آن فعل کمک می‌کند تا اگر آن فعل به‌عنوان یک ایست‌واژه از میان واژگان حذف نشد، به دلیل تکرار بالا در میان کل مجموعه دادگان شناسایی و حذف شود. در بخش‌های بعدی به توضیح این عملکرد خواهیم پرداخت. حالت دوم زمانی است که آن فعل واقعا در پیدا کردن پیام‌های یک خوشه موثر باشد. در این حالت نیز یکی کردن صورت‌های املایی مختلف آن فعل، شباهت میان پیام‌های مختلف را افزایش خواهد داد.

حذف ایست‌واژه‌ها: ایست‌واژه‌ها حروف و افعال پر تکرار و بی‌معنی‌ای هستند که در متون فارسی پیدا می‌شوند که باید حذف شوند. این دسته از کلمات، کمکی در پیدا کردن پیام‌های نزدیک به یکدیگر نمی‌کنند اما در صورتی که دو پیام متفاوت از تعداد بالایی از آن‌ها تشکیل شده باشند، ممکن است به اشتباه نزدیک به هم و متعلق به یک خوشه تشخیص داده شوند. در جهت حذف این خطا از میان پیام‌ها، ایست‌واژه‌های آن‌ها حذف خواهند شد. در زیر مثالی از دو پیام متفاوت که به دلیل تعداد بالای ایست‌واژه‌ها میان آن‌ها، احتمال خطا در تشخیص خوشه‌ها را افزایش خواهند داد مشاهده می‌کنید. در این مثال‌ها، کلمات ایست‌واژه با پس‌زمینه رنگی مشخص شده‌اند.

مثال ۱) خبر خوب این که کمک‌های بلاعوض به زلزله‌زدگان کرمانشاه از امروز شروع شده است.

مثال ۲) این تصاویر هوایی است که از شهر البوکمال پس از آزادسازی به دست سپاه گرفته شده است.

استخراج کلمات کلیدی متن: کلمات کلیدی هر پیام استخراج می‌شوند. این کلمات کلیدی در خوشه‌بندی پیام‌ها استفاده خواهد شد. کلمات کلیدی هر پیام، آن دسته از کلماتی هستند که با حفظ آن‌ها موضوع آن پیام قابل تشخیص باشد. در جهت استخراج این کلمات از پیام‌ها دو دسته از کلمات شناسایی و حذف خواهند شد. دسته اول کلماتی هستند که در میان کل مجموعه داده تکرار بالایی داشته باشند. این کلمات که ممکن است در بخش قبلی شناسایی نشده باشند می‌توانند همچون ایست‌واژه‌ها عمل کرده و پیام‌های غیرمربوط به هم را به یکدیگر نزدیک‌تر سازند. در جهت حذف این دسته کلمات ابتدا تعداد تکرار هر کلمه در پیام‌های مختلف شناسایی شد. پس از آن کلماتی که در بیش از یک هزارم تعداد کل پیام‌ها تکرار شده بودند از میان واژگان حذف شدند.

به همین ترتیب کلماتی که در کمتر از پنج پیام مختلف حضور داشتند نیز به‌عنوان کلمات خاص شناسایی و حذف شدند. حضور این کلمات در بین پیام‌ها می‌تواند دو پیامی که از لحاظ معنی به یکدیگر شبیه‌اند را از هم دور سازد. همچنین تعداد پنج به این دلیل انتخاب شده است که در این نوشتار حداقل اندازه یک موضوع برای بررسی ویژگی‌ها

و دسته‌بندی آن در ردهٔ شایعه یا غیر شایعه پنج در نظر گرفته شده است. بدین ترتیب می‌توان مطمئن بود که کلمات کلیدی یک خوشه استخراج شده، حذف نشده باشند.

۲،۳. خوشه‌بندی پیام‌ها

همان‌طور که بیان شد، روش پیشنهادی برای تشخیص شایعات به خوشه‌بندی پیام‌های مختلف در دو ردهٔ شایعه و غیرشایعه می‌پردازد. بنابراین، در ابتدا نیاز است تا پیام‌های مختلف به خوشه‌های مربوط به خود منتسب شوند. پس از این‌که در بخش قبلی کلمات کلیدی هر پیام شناسایی شد، اکنون می‌توانیم به کمک آن‌ها شباهت پیام‌های مختلف به یکدیگر را محاسبه و خوشه‌های مرتبط به هر پیام را تشخیص دهیم. به این جهت ابتدا یک سری خوشه‌های اولیه از کل مجموعه‌داده استخراج می‌کنیم. پس از آن نیاز است تا خوشه‌هایی که به یکدیگر شبیه بوده‌اند را تجمیع کنیم تا خطای استخراج ویژگی از میان آن‌ها به حداقل برسد. پس از این‌که خوشه‌های مختلف مورد بحث در مجموعه‌داده را شناسایی کردیم نیاز است تا کل پیام‌های موجود را مرور کرده و هر پیام را به نزدیک‌ترین خوشه مربوطه متصل کنیم. در ادامه، جزئیات این مراحل به تفکیک شرح داده شده است.

مرحله ۱. انتخاب خوشه‌های اولیه: در ابتدا نیاز است تا مجموعه‌ای از خوشه‌های اولیه از میان پیام‌ها استخراج شود. به این منظور هر پیامی که بیشتر از چهار بار بازنشر شده باشد شناسایی و استخراج می‌شود. هر کدام از این پیام‌ها در کنار بازنشرهایش یکی از خوشه‌های اولیه ما را شکل خواهند داد. بدین ترتیب هر خوشه‌ای حداقل حاوی پنج پیام خواهد بود. تعداد پنج پیام برای هر خوشه از دو جنبه حائز اهمیت است. اول آن که ویژگی‌هایی که قصد استخراج آن‌ها را از خوشه‌های مختلف داریم نیاز به زمان جهت شکل‌گیری صحیح دارند. خوشه‌هایی که زمان بیشتری از آن‌ها گذشته و شامل تعداد پیام‌های بیشتری هستند ساختار دقیق‌تری را شکل داده‌اند و امکان شناسایی دقیق‌تر آن‌ها افزایش خواهد یافت. نکته دوم آن است که این مقدار نباید خیلی زیاد باشد. زیاد بودن بیش از حد این مقدار باعث می‌شود تا بسیاری از پیام‌ها در هیچ خوشه‌ای قرار نگیرند و تنها امکان شناسایی و دسته‌بندی خوشه‌هایی وجود داشته باشد که زمان بیشتری از آن‌ها گذشته باشد. تشخیص شایعات امری است که نیازمند سرعت حداکثری خواهد بود چراکه شایعات جدید ممکن است پس از گذشت زمان، اهمیت خود را در اذهان مردم از دست داده یا حتی صحتشان تایید یا تکذیب شده و شکل خبری به خود گیرند. بدین ترتیب مقدار پنج به‌عنوان حداقل پیام‌های مورد نیاز برای شکل‌گیری یک خوشه جدید در نظر گرفته شده تا هم دید اولیه‌ای از ساختار آن خوشه در دست باشد و هم زمان زیادی از شکل‌گیری آن نگذشته باشد.

مرحله ۲. تشکیل گراف تشابه خوشه‌ها و تجمیع خوشه‌های مشابه: در مرحله انتخاب خوشه‌های اولیه، هر پیام که بیش از چهار بازنشر داشت به‌عنوان یک خوشه مستقل شناسایی شد. بسیار محتمل است که در یک موضوع مورد بحث چندین پیام مختلف وجود داشته باشند که هرکدام شامل تعداد بالایی بازنشر باشند. به‌عنوان نمونه پیام‌های زیر را در نظر بگیرید. این پیام‌ها همگی در مورد یک موضوع مشخص صحبت می‌کنند، با این حال هرکدام به‌صورت جداگانه در شبکه اجتماعی بازنشر شده و در نتیجه خوشه‌های مختلفی را شکل داده‌اند.

بالایی ماشین دکتر مصدق تو احمدآباده و پایینی آدامسیه که آلکس فرگوسن از دهنش تف کرده

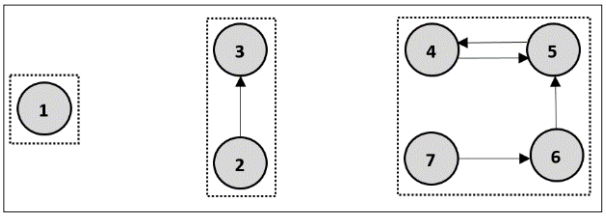
تصویر بالا ماشین دکتر مصدق در احمدآباد و تصویر پایین آدامس تف کرده فرگوسن مربی سابق منچستر که در موزه نگهداری می‌شود

ماشین محمد مصدق نخست وزیر سابق این مملکت در احمدآباد مستوفی رها شده است اما در انگلیس آدامس جویده آلکس فرگوسن را در موزه نگهداری می‌کنند

در جهت تجمیع خوشه‌های یکسان، فاصله هر خوشه از خوشه‌های دیگر بر اساس فرمول (۱) (فرمول مشابهت) محاسبه خواهد شد:

$$topicsDisctance(x.y) = \frac{|tlast_x - tlast_y| + |tfirst_x - tfirst_y|}{N(x.y)} \quad (۱)$$

در فرمـــول (۱)، $N(x.y)$ برابـــر با تعـــداد کلمات کلیدی مشـــترک میان دو خوشه x و y است. مقدار $tlast_x$ برابر با زمان نشر آخرین پیام از خوشه x و به همین ترتیب $tlast_y$ برابر با زمان انتشـــار آخرین پیام از خوشه y است. مقدار $tfirst_x$ و $tfirst_y$ نیز به ترتیب بیانگر زمان انتشار اولین پیام از خوشه‌های x و y هستند. در جهت کاهش خطای موجود در تجمیع خوشه‌ها مختلف، به ازای هر خوشه تنها امتیاز خوشه‌هایی محاســـبه شده اســـت که تعداد کلمات کلیدی مشترک بیشتر از دو با خوشه اولیه داشته و زمان ابتدایی یکی بعد از زمان انتهایی دیگری نباشد. شرط دوم اطمینان حاصل می‌کند که زمان وقوع این دو خوشه حتما با یکدیگر همپوشانی داشـــته باشند چراکه خوشه‌های یکسان عمدتا در بازه زمانی مشخصی رخ خواهند داد. بنابر فرمول (۱)، فاصله دو خوشـــه با یکدیگر رابطه عکـــس با تعداد کلمات مشـــترک آن‌ها و رابطه مستقیم با اختلاف زمان شروع و پایان این دو خوشـــه با هم دارد. به عبـــارت دیگر فاصله دو خوشه از یکدیگر هرچه تعداد کلمات مشترک بیشتری داشته و تقریبا همزمان شـــروع شده و پایان یافته باشند، کمتر خواهد بود. پس از محاسبه فاصله هر خوشه به ازای یک خوشـــه مشخص، آن خوشـــه‌ای که کمترین فاصله را با آن داشـــته به‌عنوان نزدیک‌ترین همســـایه به آن انتخاب می‌شـــود و اینکار به ازای تمام خوشـــه‌های موجود تکرار خواهد شد. توجه داشته باشید ممکن است برای یک خوشه هیچ همســـایه‌ای پیدا نشـــود. پس از محاسبه نزدیک‌ترین همســـایه هر خوشـــه، یک گراف جهت‌دار از خوشـــه‌های مختلف ساخته خواهد شد. در این گراف هر گره نمایشگر یکی از خوشـــه‌های ابتدایی و هر یـــال بین دو گره x و y بیانگر این خواهد بود که خوشه y نزدیک‌ترین همسایه به خوشه x بوده است. به عبارت دیگر، اگر شبیه‌ترین خوشه به خوشه x خوشه y باشد، آنگاه یک یال جهت‌دار از x به y وجود دارد. شکل ۲ گرافی نمونه از شباهت خوشه‌ها را به تصویر کشیده است.



شکل ۲: گراف تشابه خوشه‌های نمونه

در شـــکل ۲، برای خوشـــه ۱ هیچ خوشه مشابهی پیدا نشـــده اســـت؛ خوشه ۳ شـــبیه‌ترین خوشـــه به خوشه ۲ اســـت بنابراین یک یال جهت‌دار از خوشـــه ۲ به خوشه ۳ وجود دارد؛ خوشه ۵ شـــبیه‌ترین خوشه به خوشه‌های ٤ و ٦ اســـت، خوشـــه ٤ شبیه‌ترین خوشـــه به ۵ و خوشه ٦ شبیه‌ترین خوشـــه به ۷ است. در جهت تجمیع خوشه‌های مشـــابه، هر بخش متصل این گراف (که در شکل ۲ با نقطه چین مشخص شده است) به‌عنوان یک خوشه جدید ایجاد شـــده که زمان اولین پیام منتشر شده در آن، کوچک‌ترین زمان کل پیام‌های منتشر شده در تمام خوشه‌ها آن بخش، زمـــان آخرین پیام آن بزرگترین زمـــان تمام این پیام‌ها و کلمات کلیدی آن اجتماع کلمات کلیدی تمام خوشـــه‌ها آن بخش خواهد بود. پس از ایجاد خوشـــه‌های تجمیع شده از هـــر بخش نیز، تمام خوشـــه‌های آن بخش از مجموعه کل خوشه‌ها حذف خواهند شد. بدین ترتیب پس از انتهای این مرحله خوشـــه‌های استخراج شده از بین پیام‌ها از یکدیگر مجزا بوده و خوشه‌های یکسان، به‌صورت جداگانه در نظر گرفته نخواهند شد.

مرحله ۳. خوشه‌بندی پیام‌های باقی‌مانده در خوشه‌های به دســـت آمده: پس از آن که از مرحله ۲ خوشه‌های نهایی مجموعه‌داده استخراج شـــدند، اکنون نیاز است تا هر پیام را به یکی از این خوشـــه‌ها منتســـب کنیم. در صورتی که یک پیام بازنشر یکی از خوشـــه‌ها هر بخش گراف تشابه خوشه‌ها باشد، بدون نیاز به پردازش بیشتر به آن خوشه منتســـب خواهد شـــد. در غیـــر این صورت نیاز اســـت تا شباهت آن پیام به تک‌تک خوشه‌های تجمیع شده از مرحله ۲ محاسبه شـــود. به این منظور، تعداد کلمات مشترک آن پیام با هر خوشه محاسبه خواهد شد با این شرط که زمان

انتشار آن پیام از ۲۴ ساعت قبل از اولین پیام آن خوشه تا ۲۴ ســاعت بعد از آخرین پیام آن قرار گیرد. این شرط به این دلیل لحاظ شده است تا خوشه‌هایی که از زمان انتشار آن پیام فاصله زیادی دارند از محاسبات حذف شده و دقت در تعیین خوشه هر پیام افزایش یابد. به‌عنوان نمونه فرض کنید خوشه‌ای امروز به‌عنوان شایعه در شبکه‌های اجتماعی مطرح شــده و هفته بعد اخبار اصلاحیه این شایعه منتشر شوند. در صورتی که شرطی روی محدودسازی خوشه‌ها مورد بررســی برای یک پیام نداشته باشیم، پیام شایعه‌ای که امروز منتشــر شده ممکن اســت در خوشه اصلاحیه آن در هفته بعد قرار گیرد. پس از محاســبه کلمات کلیدی مشترک هر پیام با خوشه‌های نزدیک به زمان انتشار آن، خوشه‌ای که بیشترین کلمه کلیدی مشترک با آن را داشته باشــد، با این شــرط که تعداد کلمات کلیدی محاسبه شده بیشتر از دو باشند، به‌عنوان خوشه نهایی آن پیام انتخاب خواهد شد. شــرط تعداد کلمات کلیدی حداقلی نیز به دلیل کم کردن خطای موجود در تشخیص خوشه هر پیام لحاظ شده اســت تا پیام‌های بی‌ربط در یک خوشه قرار نگیرند. توجه داشــته باشید ممکن است هیچ خوشه‌ای به برخی از پیام‌ها منتسب نشود.

۳،۳. استخراج ویژگی هر خوشه

در مرحله دسته‌بندی پیام‌ها، هرکدام از پیام‌های موجود به یک خوشه مشخص منتسب شدند (فعلا از پیام‌هایی که در هیچ خوشــه‌ای قرار نگرفتند، صرف‌نظر می‌کنیم). حال زمان آن اســت تا ویژگی‌های ســاختاری این خوشه‌ها را از آن‌ها اســتخراج کرده و مجموعــه‌دادگان خود را برای آموزش مدل تشخیص شایعات آماده سازیم. به این منظور ابتــدا برخــی از ویژگی‌هایی که در کارهای پیشــین برای دســته‌بندی شایعات از میان خوشــه‌های موجود، معرفی شده بودند را از خوشه‌ها به دست آمده استخراج می‌کنیم. این ویژگی‌ها در سه دسته ویژگی‌های مربوط به متن(مانند تعداد هشتگ‌ها و موجودیت‌های نامی مجزای هر خوشه)، ویژگی‌هــای مربوط به کاربر(ماننــد میانگین طول عمر و دنبال‌شــونده‌ها یا دنبال‌کننده‌های کاربران منتشــرکننده پیام‌های هر خوشه) و ویژگی‌های مربوط به انتشار(مانند میانگین پیام‌های بازنشر یا پاسخ تشکیل‌دهنده هر خوشه) قرار گرفته‌اند که در ادامه به تشریح آن‌ها خواهیم پرداخت. پس از آن که برخی از مهمترین ویژگی‌های معرفی شده در کارهای پیشین را از مجموعه‌دادگان خود استخراج کردیم دسته‌ای جدید از این ویژگی‌ها را معرفی خواهیم کرد. این ویژگی‌ها که با عنوان ویژگی‌هــای مربوط به زمان(مانند طول عمر پیام‌های یک خوشــه) مشخص شده‌اند در ادامه شرح داده خواهند شــد. ویژگی‌های مربوط به زمان برای اولین بار در این مقاله معرفی و بررسی شدند. ویژگی‌های معرفی شده در این مقاله از شبکه اجتماعی توییتر استخراج شــده و در آن ادبیات تعریف شــده‌اند. با این حال امکان ترجمه بســیاری از این ویژگی‌ها در شــبکه‌های اجتماعی مختلف وجود خواهد داشت.

ویژگی‌های مربوط به متن: دســته اول از ویژگی‌های اســتخراج شده از خوشه‌های به دســت آمده، ویژگی‌های مربوط به متن پیام‌های منتشــر شــده در آن خوشــه‌ها هستند. این ویژگی‌ها را می‌توان بدون نیاز به هیچ اطلاعات اضافی دیگری، به‌طور مستقیم از متن پیام‌ها استخراج کرد. در نامگذاری این ویژگی‌هــا غالبا واژه توییت به کار رفته است. فهرســت کامل ویژگی‌های مربوط به متن و تعریف آن‌ها در جدول ۱ آمده است.

نحوه اســتخراج ویژگی‌هــای بیان شــده در جدول ۱ مستقیما از تعریفشان قابل استنتاج است. برای مثال ویژگی avg_tweets_with_hashtag از تقسیم تعداد پیام‌هایی که در آن‌هــا حداقل یکبار علامت «#» دیده شــده بر تعداد کل پیام‌های منتشر شــده در هر خوشه به دست خواهد آمد. ویژگی num_unique_hashtags برابر با تعداد هشتگ‌های مجزای اســتخراج شــده از میان کل پیام‌های هر خوشــه خواهد بود. نکته قابل توجــه در مورد ویژگی‌های مربوط به موجودیت‌های نامی اســت. در روش پیشنهادی در این مقاله، انواع مختلف موجودیت‌های نامی از جمله نام افراد،

جدول ۱: فهرست ویژگی‌های مربوط به متن

ردیف	نام ویژگی	تعریف
۱	num _unique_nes	تعداد کل موجودیت‌های نامی مجزا در بین پیام‌های هر خوشه
۲	num _unique_hashtags	تعداد کل هشتگ‌های مجزا در بین پیام‌های هر خوشه
۳	avg_tweets_len	میانگین طول پیام‌های هر خوشه
۴	avg_tweets_with_url	نسبت پیام‌های حاوی url هر خوشه به کل پیام‌های آن
۵	avg_tweets_with_mention	نسبت پیام‌های حاوی ذکر کاربر هر خوشه به کل پیام‌های آن
۶	avg_tweets_with_hashtag	نسبت پیام‌های حاوی هشتگ هر خوشه به کل پیام‌های آن
۷	avg_tweets_with_ne	نسبت پیام‌های حاوی موجودیت‌های نامی هر خوشه به کل پیام‌های آن
۸	avg_tweets_with_media	نسبت پیام‌های حاوی تصویر یا ویدئو هر خوشه به کل پیام‌های آن
۹	avg_nes_per_tweet	متوسط تعداد موجودیت‌های نامی هر پیام
۱۰	avg_hashtags_per_tweet	متوسط تعداد هشتگ‌های هر پیام

سازمان‌ها، واحدهای پول، مکان‌ها، تاریخ‌ها و زمان‌های مختلف استخراج شده است. همچنین، در روش پیشنهادی میان موجودیت‌های نامی از انواع مختلف تفاوتی قائل نشده و تمام آن‌ها را به یک صورت پردازش خواهیم کرد.

ویژگی‌های مربوط به کاربر: پس از استخراج ویژگی‌های مختلف از متن پیام‌ها، نیاز به بررسی کاربران منتشر کننده آن پیام‌ها خواهیم داشت. در جدول ۲ فهرست ویژگی‌های استخراج شده از کاربران منتشر کننده پیام‌های هر خوشه به همراه تعریفی از این ویژگی‌ها آمده است.

محاسبه ویژگی‌های تعریف شده در جدول ۲ با در دست داشتن اطلاعات خام استخراج شده از شبکه اجتماعی توییتر، فرایندی مستقیم از تعریف آن ویژگی خواهد بود. به‌عنوان نمونه تعداد پیام‌های منتشر شده توسط هر کاربر در ۱۵ روز اخیر از شبکه اجتماعی توییتر قابل دریافت است. حال محاسبه ویژگی avg_users_tweets_in15 با میانگین‌گیری از این مقدار به ازای کاربر منتشر کننده هر پیام در خوشه مورد بررسی به دست خواهد آمد. نکته قابل

جدول ۲: فهرست ویژگی‌های مربوط به کاربر

ردیف	نام ویژگی	تعریف
۱	avg_users_with_description	متوسط تعداد کاربران شرکت کننده در هر خوشه که متن توضیحات پروفایلشان خالی نبوده
۲	avg_users_verified	متوسط تعداد کاربران تایید شده شرکت کننده در هر خوشه
۳	avg_users_followers	متوسط تعداد دنبال کنندگان کاربران شرکت کننده در هر خوشه
۴	avg_users_friends	متوسط تعداد دوستان کاربران شرکت کننده در هر خوشه
۵	avg_users_popularity	متوسط نسبت دوستان به دنبال کننده‌های کاربران شرکت کننده در هر خوشه
۶	avg_users_favorits	متوسط تعداد توییت‌های مورد علاقه کاربران شرکت کننده در هر خوشه
۷	avg_users_age	متوسط طول عمر فعالیت کاربران شرکت کننده در هر خوشه
۸	avg_users_tweets	متوسط توییت‌های منتشر شده توسط کاربران شرکت کننده در هر خوشه
۹	avg_users_tweets_in15	متوسط توییت‌های منتشر شده توسط کاربران شرکت کننده در هر خوشه در ۱۵ روز اخیر
۱۰	avg_users_retweets_in15	متوسط توییت‌های بازنشر شده توسط کاربران شرکت کننده در هر خوشه در ۱۵ روز اخیر

توجه آن است که ممکن است در یک خوشه تعداد زیادی از پیام‌ها، توسط یک یا چند کاربر محدود ایجاد شده باشد. در چنین شرایطی هر کاربر به تعداد پیام‌هایی که در آن خوشه منتشر کرده در میانگین‌گیری شرکت خواهد کرد تا وزن ویژگی‌های آن کاربر نسبت به کاربران دیگری که تنها یک پیام در آن خوشه منتشر کرده‌اند، بیشتر در نظر گرفته شود.

ویژگی‌های مربوط به انتشار: برخی از ویژگی‌های استخراج شده از هر خوشه، نه مستقیما از متن پیام‌های آن و نه از کاربران منتشر کننده آن پیام‌ها قابل استخراج است. این ویژگی‌ها در دسته ویژگی‌های مربوط به انتشار آن خوشه قرار گرفته‌اند. فهرست کامل ویژگی‌های استخراج شده در این دسته به همراه تعریف هر کدام در جدول ۳

جدول۳: فهرست ویژگی‌های مربوط به انتشار

ردیف	نام ویژگی	تعریف
۱	num_of_tweets	تعداد کل پیام‌های تشکیل دهنده هر خوشه
۲	avg_retweets	میانگین پیام‌های بازنشر تشکیل دهنده هر خوشه
۳	avg_replies	میانگین پیام‌های پاسخ تشکیل دهنده هر خوشه
۴	avg_like_per_followers	میانگین نسبت تعداد لایک پیام‌های تشکیل دهنده هر خوشه به تعداد دنبال‌کننده‌های کاربر منتشر کننده آن
۵	avg_rt_per_followers	میانگین نسبت تعداد بازنشر پیام‌های تشکیل دهنده هر خوشه به تعداد دنبال‌کننده‌های کاربر منتشر کننده آن

نمایش داده شده است.

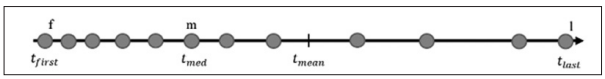
ایـن ویژگی‌هـا نیز پـس از خوشـه‌بندی پیام‌ها و به کمـک اطلاعات اولیـه در دسـترس از شـبکه اجتماعی توییتر به سـادگی قابل محاسـبه خواهند بود. برای مثال از شـبکه اجتماعی توییتر نسبت بازنشر هر پیام به تعداد دنبال‌کننده‌های کاربر منتشر کننده آن قابل استخراج است. حال ویژگی avg_rt_per_follower تنها یک میانگین‌گیری بر روی این مقدار به ازای تمام پیام‌های منتشر شده در هر خوشه خواهد بود. به‌عنوان مثالی دیگر از داده‌های استخراج شده از شـبکه اجتماعی توییتر مشخص است که یک پیام آیا بازنشـر پیام دیگری بوده یا در پاسخ به پیامی منتشر شـده اسـت یا خیر. به این ترتیب ویژگی avg_retweets و avg_replies به ترتیب بیانگر تعداد پیام‌های بازنشـر یا پاسخ هر خوشـه به تعداد کل پیام‌های آن خوشه خواهند بود.

ویژگی‌های مربوط به زمان: آخرین دسته از ویژگی‌های اسـتخراج شـده از میان خوشـه‌های پیام‌ها، ویژگی‌های مربوط به زمان انتشـار آن‌ها اسـت. این ویژگی‌ها برای اولیـن بـار در این مقاله اسـتفاده می‌شـود. در جدول ٤ فهرسـت ویژگی‌های استخراج شـده از میان خوشه‌ها به همراه تعریف مختصری از آن‌ها آمده است.

در بین این ویژگی‌ها، ویژگی liveness عمر آن خوشه و ویژگی avg_liveness_per_tweet عمر متوسط هر پیام

جدول ۴: فهرست ویژگی‌های مربوط به زمان

ردیف	نام ویژگی	تعریف
۱	liveness	فاصله زمانی اولین پیام منتشر شده در هر خوشه تا آخرین پیام آن
۲	avg_liveness_per_tweet	متوسط فاصله زمانی انتشار هر پیام تا پیام بعدی در هر خوشه
۳	time_ration_over_tweets	نسبت فاصله انتشار پیام میانی تا آخرین پیام هر خوشه به فاصله انتشار پیام میانی تا اولین پیام آن
۴	tweets_ratio_over_time	نسبت تعداد پیام‌های منتشر شده در بازه دوم عمر هر خوشه به تعداد پیام‌های منتشر شده در بازه اول عمر آن



شکل ۳: نمونه انتشار پیام‌های یک خوشه در طول زمان

از آن خوشـه را نشان می‌دهد که با تقسیم عمر آن خوشه بر تعداد کل پیام‌های آن خوشـه نیز به دست خواهد آمد. اما محاسـبه دو ویژگی بعدی نیاز به کار بیشـتری دارد. فرض کنید شـکل ۳ نمایش دهنده انتشار پیام‌های مختلف از یک خوشه در طول زمان باشد. در این شکل محور افقی نمایشگر گذشت زمان و هر دایره بیانگر انتشار یک پیام در آن نقطه از زمان است.

در شکل ۳، اولین پیام با حرف f و آخرین پیام با حرف l در بالای آن مشخص شده‌اند. پیام میانی نیز با حرف m نشان داده شـده که صرف نظر از زمان انتشار آن، میانه رشته پیام‌های منتشر شـده بوده است. اگر مانند شکل ۳ تعداد پیام‌های منتشر شده در یک خوشه زوج باشد، آخرین پیام در محاسبه میانه لحاظ نخواهد شد. در این شکل t_{first} زمان انتشـار اولیـن و t_{last} زمان انتشـار آخرین پیام از خوشـه مورد نظر است. زمان انتشار پیام میانه نیز با t_{med} نشان داده شده است. ویژگی time_ratio_over_tweets از نسبت فاصله زمان t_{med} تا t_{last} به فاصله زمان t_{first} تا t_{med} به دست خواهد آمد (فرمول (۲)).

$$timeRatioOverTweets = \frac{(t_{last} - t_{med})}{(t_{med} - t_{first})} \qquad (۲)$$

در این شـــکل t_{mean} نشان دهنده میانگین دو زمان t_{first} و t_{last} است. ویژگی tweets_ratio_over_time از تقسیم تعداد پیام‌های منتشـــر شـــده در بازه t_{mean} تا t_{last} به تعداد پیام‌های منتشر شده در بازه t_{first} تا t_{mean} به دست خواهد آمد (فرمول (۳)). به‌عنوان نمونه در شکل ۳ این مقدار برابر با $\frac{4}{8}$ یا همان ۰٫۵ خواهد بود.

$$tweetsRatioOverTime = \frac{\#tweets(t_{last}, t_{mean})}{\#tweets(t_{mean}, t_{first})} \quad (۳)$$

٤,۳. ساخت مدل

پس از آن که پیام‌ها را خوشه‌بندی کردیم و ویژگی‌های مورد نظر از این خوشه‌ها را استخراج کردیم، اکنون زمان آن است تا یک مدل دسته‌بندی برای تشخیص ردهٔ هر پیام انتخاب کرده و آن را با مجموعه‌داده موجود آموزش دهیم. مدل‌هـــای زیادی برای دســـته‌بندی داده‌هـــا در زمینه‌های مختلف ارائه شـــده و بســـیاری از آن‌هـــا در پژوهش‌های مربوط به پیش‌بینی ردهٔ شایعات مطالعه و بررسی شده‌اند. بایـــد چالش‌هایی که در مجموعه‌دادگان وجود دارد نیز در نظر گرفته شود و با توجه به آن‌ها بهترین مدل موجود برای آموزش آن بر روی مجموعه داده انتخاب شود. حجیم بودن داده‌ها و نامتوازن بودن تعداد پیام‌های ردهٔ شـــایعه نسبت به ردهٔ غیرشـــایعه از مهم‌ترین این چالش‌هاســـت. پس از بررسی‌های انجام شده، مدل درخت تصمیم برای آموزش بر روی دادگان در روش پیشـــنهادی انتخاب شـــده است. درخت تصمیم یک الگوریتم حریصانه اســـت که به تقسیم بازگشـــتی دودویی فضای ویژگی‌ها می‌پردازد. این روش، برخلاف بسیاری از روش‌های متداول در دسته‌بندی داده‌ها نیازی به نرمال‌سازی ویژگی‌ها و انتخاب ویژگی‌های موثر نداشته و در برابر نامتوازن بودن وزن داده‌های موجود از رده‌های مختلف به شـــدت مقاوم است. درخت ایجاد شده توسط این الگوریتم به ازای تمام داده‌های تقسیم شده در هر برگ، یک ردهٔ یکســـان را پیش‌بینی خواهد کرد. در این مقالـــه، از الگوریتم درخت تصمیـــم CART که در کتابخانه mllib اسپارک پیاده‌سازی شده است، استفاده می‌شود.

ردهٔ پیش‌بینی‌شده

ردهٔ حقیقی		مثبت	منفی
	مثبت	TP	FN
	منفی	FP	TN

شکل ۴: ماتریس درهم‌ریختگی

۴. ارزیابی روش پیشنهادی

در تشـــخیص شـــایعات، روشی بهینه اســـت که تمام شایعات موجود را به درستی شناسایی کرده و هیچ خوشه غیرشـــایعه‌ای را به اشتباه به‌عنوان شایعه برچسب‌گذاری نکند. به این منظور در ارزیابی روش پیشنهادی خود از دو معیار دقت[10] و فراخوانی[11] استفاده خواهیم کرد تا به ترتیب میزان خطای الگوریتم در شناسایی خوشه‌های غیرشایعه به‌عنوان شایعه و خوشه‌های شایعه به‌عنوان غیرشایعه را محاســـبه کنیم. همچنین معیار اف[12]، به‌عنوان معیاری که کیفیت کلی روش پیشنهادی را نشان خواهد داد نیز مطالعه شده است.

ماتریس درهم‌ریختگی یکی از ابتدایی‌ترین و ساده‌ترین معیارها برای ارزیابی کیفیت مدل‌های ارائه شـــده اســـت. ماتریس درهم‌ریختگی، جدولی دوبعدی اســـت که یک بعد آن بیانگر ردهٔ حقیقی و بعد دیگر ردهٔ پیش‌بینی شده توسط مـــدل را بیان می‌کنـــد و در هر دو بعـــد آن مجموعه‌ای از رده‌های موجود در مســـئله قرار گرفته اســـت. در شکل ٤ نمونه‌ای از ماتریس درهم‌ریختگی برای یک مســـئله با دو ردهٔ مثبت و منفی را مشاهده می‌کنید [۱۹].

در شکل ٤، TP[13] یانگر تعداد مواردی است که ردهٔ آن‌ها به درســـتی، ردهٔ مثبت پیش‌بینی شده اســـت؛ TN[14] تعداد مواردی را نشـــان می‌دهد که ردهٔ حقیقی آن‌ها منفی بوده و مدل نیز به درســـتی کلاس آن‌ها را منفی پیش‌بینی کرده اســـت. به همین ترتیب FP[15] بیانگر تعداد مواردی است که

10- Precision
11- Recall
12- F-measure
13- True positive
14- True negative
15- False positive

ردهٔ آن‌ها به اشتباه، ردهٔ مثبت پیش‌بینی شده و FN[16] تعداد مواردی را نشان می‌دهد که ردهٔ حقیقی آن‌ها مثبت بوده و به اشتباه ردهٔ منفی برای آن‌ها پیش‌بینی شده است.

دقت معیاری است برای بیان نسبت تعداد مواردی که به درستی به‌عنوان مواردی از ردهٔ مثبت پیش‌بینی شده‌اند به کل مواردی که مثبت پیش‌بینی شده‌اند و به‌صورت زیر محاسبه می‌شود.

$$precision = \frac{TP}{TP + FP} \qquad (٤)$$

فراخوانی معیاری است که بیان می‌کند چه نسبتی از موارد واقعا به ردهٔ مثبت تعلق داشته‌اند و به درستی در این ردهٔ دسته‌بندی شده‌اند به کل موارد واقعا مثبت و به‌صورت زیر محاسبه می‌شود.

$$recall = \frac{TP}{TP + FN} \qquad (٥)$$

معیار اف به ما کمک می‌کند بدون این‌که نیاز به تحلیل هردو معیار دقت و فراخوانی به‌صورت جداگانه داشته باشیم، هر دوی آن‌ها را در قالب یک مقدار ببینیم. معیار اف به‌صورت زیر محاسبه می‌شود.

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \qquad (٦)$$

در جهت ارزیابی روش پیشنهادی در این مقاله، از مجموعه‌داده ارائه شده توسط دهقانی و همکاران [۱۱] استفاده شده است. این مجموعه داده شامل ۳۵۹۸۰۴۹ عدد پیام مختلف جمع‌آوری شده از شبکه اجتماعی توییتر بوده که تعداد ٤٣٤٥ عدد از آن‌ها به‌عنوان شایعه برچسب‌گذاری شده‌اند. نحوه برچسب‌گذاری توییت‌های جمع‌آوری شده در این مجموعه داده نیز به‌صورت خودکار و با مطابقت کلمات موجود در متن این پیام‌ها با کلیدواژه‌های شناسایی شده به‌عنوان شایعه از وبگاه شایعات بوده است. این وبگاه به‌صورت روزانه در فرایندی غیرخودکار، اقدام به شناسایی شایعات منتشر شده در شبکه‌های اجتماعی برخط و معرفی آن‌ها می‌کند. مجموعه‌داده تهیه شده توسط دهقانی و همکاران، علاوه بر متن و برچسب هر پیام، اطلاعات دیگری را نیز از شبکه اجتماعی توییتر

16- False negative

جدول ۵: ویژگی‌های موجود از هر پیام در مجموعه داده تهیه شده توسط دهقانی محمودآباد و همکاران [۱۱]

ردیف	ویژگی‌های موجود از هر پیام در مجموعه داده
۱	شناسه هر پیام
۲	شناسه منبع انتشار پیام‌های بازنشر شده
۳	زمان انتشار هر پیام
۴	آیا پیام شامل تصویر یا ویدئو نیز بوده یا خیر
۵	آیا پیام شامل یک url بوده یا خیر
۶	آیا پیام شامل منشن افراد در آن بوده یا خیر
۷	آیا پیام شامل حداقل یک هشتگ بوده یا خیر
۸	آیا پیام در پاسخ به پیام دیگری منتشر شده یا خیر
۹	آیا پیام بازنشر پیام دیگری بوده یا خیر
۱۰	متن توضیحات موجود در پروفایل کاربر منتشرکننده پیام
۱۱	تعداد دنبال‌کننده‌های کاربر منتشرکننده پیام
۱۲	تعداد دوستان کاربر منتشرکننده پیام
۱۳	تعداد پیام‌های مورد علاقه کاربر منتشرکننده پیام
۱۴	عمر فعالیت کاربر منتشرکننده پیام در شبکه
۱۵	آیا حساب کاربر منتشرکننده پیام تایید شده است یا خیر
۱۶	تعداد کل پیام‌های منتشر شده توسط کاربر منتشرکننده پیام
۱۷	تعداد پیام‌های منتشر شده در ۱۵ روز اخیر توسط کاربر منتشرکننده پیام
۱۸	تعداد پیام‌های بازنشر شده در ۱۵ روز اخیر توسط کاربر منتشرکننده پیام
۱۹	تعداد بازنشرهای پیام
۲۰	نسبت بازنشرهای پیام به تعداد دنبال‌کننده‌های کاربر منتشرکننده آن
۲۱	نسبت پسندهای پیام به تعداد دنبال‌کننده‌های کاربر منتشرکننده آن

برای آن پیام‌ها استخراج می‌کند. برخی از این اطلاعات که به‌صورت ویژگی‌هایی مختلف در پیاده‌سازی روش پیشنهادی استفاده شده است در جدول ۵ فهرست شده‌اند. از آنجا که حجم داده‌های مجموعه‌داده استفاده شده زیاد است به منظور کاهش زمان، روش پیشنهادی در بستر اسپارک[17] برای فراهم کردن مقیاس‌پذیری پیاده‌سازی شده است.

پس از خوشه‌بندی این پیام‌ها، تعداد کل ۸۹۱٤۷ خوشه مختلف استخراج شد که از این تعداد، ٦٦۷۵۱ عدد از آن‌ها، شامل بیش از چهار پیام بودند. دلیل آن که با وجود انتخاب اولیه خوشه‌ها از میان پیام‌هایی با بیش از چهار بازنشر، همچنان تعداد قابل توجهی از خوشه‌ها در

17- Spark

انتهای دسته‌بندی، تعداد کمتری پیام داشتند این است که خوشه‌های اولیه با توجه به ویژگی ردیف ۱۹ از جدول بالا استخراج شده بودند. این ویژگی بیانگر تعداد واقعی بازنشرهای یک پیام در شبکه اجتماعی توییتر است، با این حال ممکن است این بازنشرها در زمانی پس از زمان انتخاب شده برای جمع‌آوری توییت‌ها از شبکه رخ داده یا در جمع‌آوری این داده‌ها از سطح شبکه جا مانده باشند.

لازم به ذکر است، API فراهم شده توسط شبکه اجتماعی توییتر، تنها درصد کمی از تمام پیام‌های منتشر شده در بازه زمانی انتخابی برای استخراج پیام‌ها را برمی‌گرداند، لذا حذف شدن برخی از بازنشرهای یک پیام در بین توییت‌های استخراج شده کاملا محتمل خواهد بود. با این حال ما در روش پیشنهادی نیاز داشتیم تا حداقل ۵ پیام از هر خوشه داشته باشیم تا خطای موجود در استخراج ویژگی‌های یک خوشه را به حداقل برسانیم. لذا تنها این خوشه‌ها برای دسته‌بندی انتخاب شده و تمام خوشه‌های دیگر به‌عنوان غیرشایعه برچسب‌گذاری شدند.

پس از خوشه‌بندی پیام‌ها و بررسی آن‌ها متوجه تعداد قابل توجهی از پیام‌ها شدیم که به اشتباه برچسب غیرشایعه خورده بودند. برای مثال در جدول ٦ برخی از پیام‌هایی که در یک خوشه قرار گرفته‌اند اما برچسب متفاوتی دارند فهرست شده است. در این جدول پیام‌هایی که در ستون شایعه برچسب خورده قرار گرفته‌اند به درستی برچسب شایعه داشتند و پیام‌هایی که در همان سطر و در ستون شایعه برچسب نخورده فهرست شده‌اند، توسط روش پیشنهادی در همان موضوعات شایعه دسته‌بندی شدند اما در مجموعه‌داده موجود برچسب غیر شایعه داشتند.

همان‌طور که از جدول ٦ مشخص است، خوشه‌بندی ما به خوبی عمل کرده و دو پیام مختلف که هردو راجع به یک شایعه مشخص صحبت می‌کردند را در یک خوشه قرار داده است. با این حال برخی از این پیام‌ها به اشتباه در مجموعه‌داده برچسب غیرشایعه دریافت کرده بودند. حدس ما این است که با توجه به روش مطرح شده مبتنی

جدول ۶: مقایسه شایعات برچسب خورده در مجموعه داده با شایعات مشابه که برچسب غیرشایعه داشته‌اند

	شایعه برچسب نخورده	شایعه برچسب خورده
خوشه ۱	RT @PMOIRAN: https://t.co/6ccNHSWtPZ درگیری #کشاورزان روستای جالیزی در استان #ایلام با مأموران سرکوبگر #نیروی_انتظامی #اهواز #اصفهان #اخبار	درگیری کشاورزان روستای جالیزی در استان ایلام با مأموران سرکوبگر نیروی انتظامی #اهواز #خراسان #تهران #کرمانشاه... https://t.co/KUFgABTQoz
خوشه ۲	RT @1hoseim: روستای کوئیک حسن، اطراف سرپل ذهاب، آسیب خیلی زیادی دیده بود. دو روز قبل کامیونت سیار تئاتر کانون پرورش فکری رفته برای اجرای نمایش	حضور کامیونت سیار نمایش #کانون_پرورش_فکری در مناطق زلزله‌زده #کرمانشاه و استقبال #کودکان از آن https://t.co/lIJwL5DVuo
خوشه ۳	استاندار کرمانشاه در گفت‌وگو با ایلنا: سلبریتی‌ها و سمن‌ها باید درباره کمک‌های مردمی به زلزله‌زدگان گزارش دهند https://t.co/7drbZyFwbJ	استاندار کرمانشاه: سلبریتی‌ها و سمن‌ها باید درباره کمک‌های مردمی به زلزله‌زدگان گزارش دهند. https://t.co/xxwN-QkANXa

بر جستجو کلمات کلیدی پیدا شده از وبگاه شایعات در مجموعه پیام‌ها جهت برچسب‌زنی آن‌ها، احتمالا مرحلهٔ پیش‌پردازش بر روی این پیام‌ها به درستی انجام نشده و منجر به در نظر گرفته نشدن تعداد زیادی از شایعات واقعی موجود در مجموعه‌داده شده باشند. به منظور اصلاح برچسب‌های موجود در مجموعه‌داده، تمام خوشه‌هایی که به‌عنوان شایعه در نظر گرفته شده بودند با نظارت فردی بررسی شده و پیام‌های غیرشایعه‌ای که در این خوشه‌های دسته‌بندی شده بودند مورد مطالعه قرار گرفتند. پس از این بررسی برچسب تمام پیام‌هایی که به اشتباه درج شده بود به شایعه تغییر پیدا کرد و در نهایت تعداد ٤٣٤٥ عدد شایعه اولیه به میزان ٤٦٢٠ شایعه نهایی رسید. همان‌طور که بیان شد در مجموع ٦٦٧٥١ خوشه مختلف با بیش از چهار پیام در هر خوشه استخراج شده بود که تعداد ۱۷۰ عدد از این خوشه‌ها برچسب شایعه دریافت کردند. در این خوشه‌بندی، برخی پیام‌ها نیز در خوشه‌های اشتباهی دسته‌بندی شدند. تعداد ۷۸ پیام شایعه از مجموعه‌داده موجود در خوشه‌بندی در هیچ خوشه استخراج شده‌ای قرار نگرفتند. تعداد ٦٨ پیام شایعه شناخته شده نیز به اشتباه در خوشه‌های غیرشایعه قرار گرفته بودند. با این

جدول ۷: ارزیابی روش پیشنهادی در تشخیص خوشه‌های شایعه

ارزیابی تشخیص خوشه‌ها	
./۸۴۷	دقت
./۸۶۶	فراخوانی
۰/۸۵۶	معیار اف

حال پس از برچسب‌زنی مجدد هیچ پیام غیرشایعه‌ای نبود که در خوشه‌های شایعه قرار گرفته باشد.

روش پیشنهادی ارائه شده بر پایه خوشه‌بندی پیام‌ها و جداسازی خوشه‌های شایعه از خوشه‌های غیر شایعه بنا شده است. بنابراین، در گام اول به ارزیابی تشخیص این خوشه‌ها خواهیم پرداخت. پس از آن، با فرض در نظر گرفتن برچسب یک خوشه برای تمام پیام‌های منتشر شده در آن، به ارزیابی جداسازی تک تک پیام‌های مربوط به خوشه‌های شایعه از سایر پیام‌ها اقدام خواهیم کرد. پژوهش‌های انجام شده در تشخیص شایعات میان پیام‌های فارسی، هر یک بر روی مجموعه‌دادگانی مجزا صورت گرفته است و بنابراین مقایسه کیفیت روش پیشنهادی تنها با روش پایه صورت خواهد گرفت.

۱,٤. ارزیابی تشخیص خوشه‌ها

به منظور ارزیابی روش پیشنهادی در برچسب‌گذاری صحیح خوشه‌ها به دو گروه شایعه و غیرشایعه، خوشه‌های ساخته شده را به‌صورت تصادفی به مجموعه‌دادگان آموزش و ارزیابی تقسیم کردیم چراکه به دلیل طول عمر مختلف خوشه‌ها با یکدیگر، امکان انتخاب ترتیبی آن‌ها وجود نداشته است. حتی با در نظر گرفتن ترتیب خوشه‌ها بر اساس اولین پیام منتشر شده در آن‌ها نیز، انتخاب ترتیبیشان منجر خواهد شد تا خوشه‌هایی با تعداد پیام‌های بیشتر، در مجموعه‌داده آموزش مدل قرار گرفته (چراکه خوشه‌های انتهایی ممکن است هنوز به اتمام نرسیده باشند) و به این ترتیب در ارزیابی روش دچار اشتباه شویم.

تمام مقادیر گزارش‌شده در جدول ۷، با اجرای ۱۰ باره مدل‌ها و میانگین‌گیری از نتایج ارزیابی آن‌ها به دست آمده است. در هر اجرا ۷۰ درصد از خوشه‌ها به‌صورت تصادفی به‌عنوان مجموعه‌داده آموزش و ۳۰ درصد باقی‌مانده آن‌ها به‌عنوان مجموعه داده ارزیابی انتخاب شده‌اند. به این ترتیب، ارزیابی‌های ما توانایی ارزیابی کلی روش پیشنهادی را خواهند داشت.

۲,٤. ارزیابی تشخیص پیام‌ها

با توجه به این‌که روش پیشنهادی ارائه شده بر پایه خوشه‌بندی پیام‌ها قرار گرفته است، ارزیابی تشخیص شایعه بودن یا نبودن تک‌تک پیام‌ها ممکن است چندان صحیح به نظر نرسد. با این حال در جهت مقایسه این کار با روش پایه این ارزیابی را انجام می‌دهیم. اینجا همچنان خوشه‌ها تشخیص داده می‌شوند ولی وقتی یک خوشه برچسب خورد کل پیام‌های آن همان برچسب را می‌خورد ولی محاسبات روی تک‌تک پیام‌ها انجام می‌شود. ارزیابی مدل، با میانگین‌گیری از نتایج ارزیابی آن در ۱۰ اجرای مختلف صورت گرفت که در هر اجرا ۷۰ درصد خوشه‌ها به‌عنوان مجموعه‌داده آموزش و ۳۰ درصد باقی‌مانده به‌عنوان مجموعه‌داده ارزیابی انتخاب شده بودند. دلیل این‌که مجموعه‌دادگان آموزش و ارزیابی که اکنون شامل خود پیام‌ها است به کمک خوشه‌های آن‌ها جداسازی شد، این است که در صورت جداسازی مجموعه‌دادگان توسط خود پیام‌ها، تعداد قابل توجهی از پیام‌های مجموعه‌داده ارزیابی در خوشه‌هایی قرار می‌گرفتند که از قبل در مجموعه‌داده آموزش دیده شده و به این ترتیب برچسب همان خوشه را دریافت می‌کردند. با این وجود، تعداد قابل توجهی از پیام‌ها وجود داشتند که در هیچ خوشه‌ای دسته‌بندی نشده بودند. فرض ما بر این بود که تمام این پیام‌ها از دسته غیرشایعه بوده‌اند. بدین ترتیب تعداد ۷۸ پیام شایعه‌ای که در این گروه قرار گرفته بودند به اشتباه به‌عنوان غیرشایعه شناسایی شدند. علاوه بر این ۶۸ پیام شایعه نیز در گروه‌های غیرشایعه قرار گرفته بودند. این گروه‌ها، در صورتی که به‌عنوان شایعه شناسایی شوند، تعدادی از پیام‌های غیرشایعه را به اشتباه به‌عنوان شایعه،

جدول ۸: ارزیابی روش پیشنهادی در تشخیص پیام‌های شایعه

ارزیابی تشخیص پیام‌ها	
۰/۹۵۷	دقت
۰/۸۳۵	فراخوانی
۰/۸۹۱	معیار اف

و در غیـر ایـن صورت دسـته‌ای از پیام‌های شـایعه را به‌عنوان غیرشایعه برچسـب‌گذاری خواهند کرد. ما برای ارزیابـی صحیح مدل خود، تمام این خطاها را در ماتریس درهم‌ریختگی حاصل، لحاظ کردیم.

جدول ۸ معیارهـای دقت، فراخوانی و معیار اف را در ارزیابی طبقه‌بندی پیام‌ها به گروه‌های شایعه و غیرشایعه نشان می‌دهد. مشاهده می‌شود که دقت مدل در تشخیص پیام‌های شایعه نسـبت به دقت آن در تشخیص خوشه‌ها شایعه افزایش داشته است. دلیل این افزایش دقت آن است که خوشه‌هایی با تعداد پیام‌های بیشتر ساختار مشخصی را شکل داده‌اند، بهتر قابل شناسایی هستند و خطای موجود در خوشه‌هایی پیدا شده که تعداد پیام‌های کمی داشته‌اند. این مقدار معیار فراخوانی نیز به دلیل در نظر گرفتن خطای موجود در خوشه‌بندی و افزایش ۷۸ واحدی مقدار FN رخ داده اسـت. با این حال معیار اف نشان می‌دهد که عملکرد مدل در کل نسبت به ارزیابی آن در تشخیص خوشه‌ها، نه تنها بدتر نشده بلکه بهبود نیز پیدا کرده است.

اکنـون که مدل خود را در برابر تشـخیص پیام‌ها نیز ارزیابی کردیم، لازم اسـت یک مقایسـه با مدل پایه انجام دهیم. دهقانی محمودآباد و همکاران [۱۱] که مجموعه‌داده اسـتفاده شـده در این مقاله را جمع‌آوری کرده بودند در مقاله خود، با اسـتخراج ویژگی‌های مختلف از متن پیام‌ها، اقدام به آموزش مدل‌های یادگیری ماشین مختلف بر روی آن‌هـا کرده و در نهایت جنـگل تصادفی را به‌عنوان مدلی که بهترین عملکرد را در برابر این مجموعه‌دادگان داشـت انتخاب کردنـد. آن‌ها در ارزیابی‌های خود دقت، فراخوانی و معیار اف را برای مدل جنگل تصادفی به ترتیب برابر با ۰/۹۷۷، ۰/۹۹۷ و ۰/۹۹۷ گزارش کرده بودند. با این حال ما در این مقاله، با استفاده از تمام ویژگی‌هایی که توسط دهقانی محمودآبـادی و همـکاران [۱۱] در مقاله آن‌ها به ازای هر پیام استخراج شـده بود، بار دیگر همان مدل‌های استفاده شده توسط آن‌ها را آموزش دادیم و نتایج ارزیابی‌هایمان با نتایج گزارش شـده توسـط آن‌ها اختلاف داشـت. در ارزیابـی که ما از روش آن‌ها انجام دادیم برای مدل جنگل تصادفی و بر روی ۱۰۰ هزار داده اول از مجموعه پیام‌ها، دقـت و فراخوانی روش آن‌ها را به ترتیب برابر با ۸۷۵/. و ۱۹۲/. به دست آوردیم. در حالی که این معیارها در گزارش ارزیابی‌هـای آن‌ها به ترتیب مقادیـر ۷۱/. و ۲۲/. گزارش شده است.

همچنین در ارزیابی که ما از روش آن‌ها انجام دادیم، با به کارگیری الگوریتم SMOTE [۲۰] که در روش آن‌ها برای ایجاد توازن بین دو کلاس شـایعه و غیرشـایعه استفاده شـده است دقت و فراخوانی به ترتیب به ۱۴۵/. و ۲۸۸/. به دست آمده است. در حالی که این معیارها، در ارزیابی‌های انجام شده توسط دهقانی و همکارانش، به ترتیب ۹۵۹/. و ۹۶/. گزارش شده اسـت. این ارزیابی‌ها نیز بر روی ۱۰۰ هزار داده اول انجام شـده اسـت. ارزیابی‌های انجام شده نشـان می‌دهد که روش پیشنهادی ارائه شده در این مقاله قادر اسـت با تنها در اختیار داشتن پنج پیام از هر خوشه، با دقت ۸۴۷/. و فراخوانی ۸۶۶/. به جداسـازی خوشه‌های شایعه از خوشه‌های غیر شایعه بپردازد.

۵. نتیجه‌گیری

در این مقاله، روشـی برای طبقه‌بندی خوشه پیام‌های منتشـر شده در شـبکه اجتماعی توییتر ارائه شد. روش پیشنهادی پس از انجام خوشـه‌بندی پیام‌ها، به استخراج ویژگی‌های سـاختاری مختلف از این خوشـه‌ها در چهار دسـته ویژگی‌های مربوط به متن، کاربر، انتشار و زمان آن‌ها می‌پردازد. ویژگی‌هـای مربوط به زمان برای اولین بار در این مقاله معرفی و مورد مطالعه قرار گرفته است. در روش پیشنهادی برای طبقه‌بندی از درخت تصمیم استفاده

شده است. ارزیابی‌های انجام شده نشان می‌دهد که روش پیشـــنهادی قادر است تنها پس از انتشـــار پنج پیام در هر خوشـــه با دقت ۰/۸۴۷ و فراخوانی ۰/۸۶۶ به جداســـازی خوشه‌های شایعه از غیرشایعه بپردازد.

مراجع

[1] S. Vosoughi, D. Roy, and S. Aral, The spread of true and false news online, Science, vol. 1151, pp. 1146-1151, 2018.

[2] N. DiFonzo and P. Bordia, Rumor, gossip and urban legends, Diogenes, vol. 54, no. 1, pp. 19–35, 2007.

[3] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, Detection and Resolution of Rumours in Social Media: A Survey, ACM Computing Surveys (CSUR), vol. 51, no. 2, 2017.

[4] J. Ma, W. Gao, P. Mitra, and E. Al, Detecting Rumors from Microblogs with Recurrent Neural Networks, in Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp. 3818–3824, 2016.

[5] F. Yang, Y. Liu, X. Yu, and M. Yang, Automatic Detection of Rumor on Sina Weibo, Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, 2012.

[6] T. Mondal, T. Roy, I. Bhattacharya, S. Bhattacharya, and I. Das, A study on rumor propagation trends and features in a post disaster situation, in ICDCN '19 Proceedings of the 20th International Conference on Distributed Computing and Networking, pp. 381–384, 2019.

[7] D. Barnwal, S. Ghelani, R. Krishna, M. Basu, and S. Ghosh, Identifying fact-checkable microblogs during disasters, in ICDCN '19 Proceedings of the 20th International Conference on Distributed Computing and Networking, pp. 389–392, 2019.

[8] T. Chen, X. Li, H. Yin, and J. Zhang, Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection Tong, in Trends and Applications in Knowledge Discovery and Data Mining, vol. 11154, Springer International Publishing, pp. 168–173, 2018.

[9] Y. Liu and X. Chen, Supervised Group Embedding for Rumor Detection in Social Media, in International Conference on Web Engineering, 2019, pp. 139–153.

[10] S. Zamani, M. Asadpour, and D. Moazzami, Rumor detection for Persian Tweets, in Iranian Conference on Electrical Engineering (ICEE2017), pp. 1532–1536, 2017.

[11] S. Dehghani Mahmoodabad, S. Farzi, D. B. Bakhtiarvand, Persian Rumor Detection on Twitter, in 9th International Symposium on Telecommunications (IST), pp. 597–602, 2018.

[12] M. Zarharan, S. Ahangar, F.S. Rezvaninejad, M. Bidhendi, M.T. Pilehvar, B. Minaei, and S. Eetemadi, Persian Stance Classification Data Set, in Proceedings of the Conference for Truth and Trust Online, 2019.

[13] Z. Jahanbakhsh-Nagadeh, M. R. Feizi-Derakhshi, and A. Sharifi, A speech act classifier for persian texts and its application in identify rumors, Journal of Soft Computing and Information Technology (JSCIT), vol. 9, no. 1, pp. 18–27, 2019.

[14] M. M. Homayounpour and A. Soltani Panah, Speech Acts Classification of Persian Language Texts Using Three Machine Leaming Methods, International Journal of Information & Communication Technology Research, vol. 2, no. 1, pp. 65–71, 2010.

[15] Z. Jahanbakhsh-Nagadeh1, M.R. Feizi-Derakhshi and A. Sharifi, A Model for Detecting of Persian Rumors based on the Analysis of Contextual Features in the Content of Social Networks, Signal and Data Processing, vol. 18, no. 1, pp. 29-50, 2021.

[16] Z. Jahanbakhsh-Nagadeh, M. R. Feizi-Derakhshi, and A. Sharifi, a semi-supervised model for Persian rumor verification based on content information, Multimedia Tools and Applications, 2020, https://doi.org/10.1007/s11042-020-10077-3.

[17] M. Samadi, M. Mousavian, and S. Momtazi, Persian Fake News Detection: Neural Representation and Classification at Word and Text Levels, ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 21, no. 10, pp. 1-11, 2021.

[18] M. Khosravi, H. Shirazi, K. Dadahstabar, S. Hashemi Golpaygani, A New Preprocessing Method for Rumor Detection in Social Networks based on LSTM-CNN, C4I Journal, Vol. 4, no. 1, PP. 38-51, 2020.

[19] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Third Edition, 2012.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, SMOTE: synthetic minority over-sampling technique, journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.