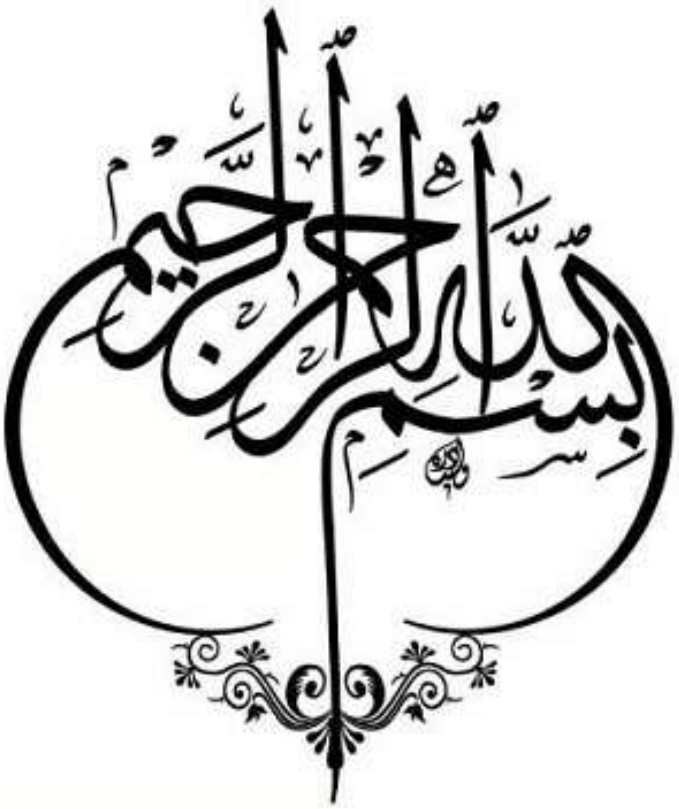

بسم الله الرحمن الرحيم

دانشگاه صنعتی شاهرود

دانشکده پردیس خوارزمی

رشته مهندسی کامپیوتر، گرایش هوش مصنوعی و رباتیک

رساله دکتری

پیش‌بینی رویدادهای اجتماعی جهان واقعی با تحلیل شبکه‌های اجتماعی مجازی

دانشجو

ابوالفضل یاوری خلیل آباد

استاد راهنما

دکتر حمید حسن پور

اساتید مشاور

دکتر محمدباقر رحیم پور

دکتر مهرگان مهدوی

دی ۱۴۰۰

سپاس و ستایش مر خدای را جل و جلاله، که آثار قدرت او بر چهره روز روشن، تابان است و انوار حکمت او در دل شب تار، درفشان. آفریدگاری که خویشتن را به ما شناساند و درهای علم را بر ما گشود و عمری و فرصتی عطا فرمود تا بدان، بنده ضعیف خویش را در طریق علم و معرفت بیازماید.

تقدیم به همسرم

به پاس قدردانی از قلبی آکنده از عشق و معرفت که محیطی سرشار از سلامت و امنیت و آرامش و آسایش برای من فراهم آورده است.

تقدیر و تشکر شایسته از استاد فرهیخته جناب دکتر **حمید حسن پور** که با نکته‌های دلاویز و گفته‌های بلند، صحیفه‌های سخن را علم‌پرور نمود و همواره راهنما و راه‌گشای نگارنده در اتمام و اکمال پایان‌نامه بوده است.

همچنین از جناب دکتر محمدباقر رحیم‌پور و دکتر مهرگان مهدوی و کلیه اساتیدم در مقطع دکتری، که در طول این مسیر پژوهشی یاریگر من بوده‌اند، تشکر و قدردانی می‌کنم.

معلما مقامت ز عرش برتر باد همیشه توسن اندیشه‌ات مظفر باد.

تعهدنامه

اینجانب ابوالفضل یاوری خلیل آباد دانشجوی دوره دکتری رشته مهندسی کامپیوتر گرایش هوش مصنوعی و باتیک دانشکده پردیس خوارزمی دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **پیش‌بینی رویدادهای اجتماعی جهان واقعی با تحلیل شبکه‌های اجتماعی مجازی** تحت راهنمایی استاد حمید حسن پور

متعهد می‌شوم:

- تحقیقات این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققهای دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا افراد دیگر برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج پایان‌نامه تأثیرگذار بوده‌اند رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

امضاء:

تاریخ:

چکیده

در دهه اخیر، ضریب نفوذ بالای شبکه‌های اجتماعی مجازی در بین افراد جوامع و فراوانی و در دسترس بودن داده‌هایشان، توجه پژوهشگران زیادی را به تحلیل شبکه‌های اجتماعی معطوف کرده است. یکی از کاربردهای مهم تحلیل شبکه‌های اجتماعی مجازی، تشخیص و پیش‌بینی وقوع رویداد می‌باشد. روشهای موجود در زمینه پیش‌بینی رویداد، تنها بر روی یک رویداد خاص تمرکز دارند، در حالیکه آنچه در این پایان‌نامه به آن پرداخته شده است، پیش‌بینی کلیه رویدادهای قابل پیش‌بینی می‌باشد. در این پژوهش رویداد به چیزی اطلاق می‌گردد که تغییر و تحولات قابل توجه، در برخی پارامترها و ویژگیهای مهم شبکه اجتماعی مجازی ایجاد کند. ویژگی که در روش پیشنهادی به آن توجه شده است، تغییرات نرخ ارسال پیامها به شبکه اجتماعی می‌باشد. در روش پیشنهادی، ابتدا بر مبنای یک پنجره زمانی ثابت، عملیات تقسیم‌بندی توییتها و سپس پیش‌پردازشهای معمول حوزه متن‌کاوی بر روی هر دسته انجام می‌شود. پس از آن با استفاده از روش تجزیه غیرمنفی ماتریس (NMF) و خوشه‌بندی افزایشی فرایند رستوران چینی مبتنی بر فاصله(ddCRP)، خوشه‌بندی توییت‌ها صورت می‌پذیرد. همزمان بر مبنای میزان تغییرات نرخ ورود توییتها به هر خوشه، وقوع رویداد پیش‌بینی می‌گردد. نهایتا توصیفی از رویدادها بصورت چند کلمه پرتکرار در هر خوشه، به عنوان رویدادهای آینده، نمایش داده می‌شوند.

روش پیشنهادی، بر روی یک مجموعه داده از توییتر، شامل تقریبا ۱۲ میلیون توییت که ۱۹۴۰ رویداد را در بر می‌گیرد، مورد ارزیابی قرار گرفته است و با دقت ۸۴ درصد موفق به پیش‌بینی وقوع رویدادها در آینده شده است.

کلمات کلیدی:

تحلیل شبکه‌های اجتماعی مجازی، تشخیص رویداد، پیش‌بینی رویداد، پیش‌بینی نتیجه انتخابات، نرخ ارسال توییتها، خوشه‌بندی افزایشی

لیست مقالات مستخرج از پایان نامه

A. Yavari, H. Hassanpour, B. Rahimpour Cami, M. Mahdavi, "Election Prediction Based on Sentiment Analysis Using Twitter Data", **International Journal of Engineering**, **Transaction B: Applications** Vol 35, No 02, (2022). Doi: 10.5829/ije.2022.35.02b.13

A. Yavari, H. Hassanpour, B. Rahimpour Cami, M. Mahdavi," Event prediction in social network through Twitter messages Analysis".AJSE, (), -. (Submitted on March 01, 2022)

فهرست مطالب

فهرست اشکال

فهرست جداول

واژه‌ها و اختصارات

واژه اصلی	علامت اختصاری
Application Programming Interface	API
British Broadcasting Corporation	BBC
the Cable News Network	CNN
Chinese Restaurant Process	CRP
Distance Dependent Chinese Restaurant Process	DDCRP
Feferal Bureau of Investigation	FBI
First Story Ditection	FSD
Internet Movie Database	IMDb
Latent Dirichlet Allocation	LDA
Locality Sensitive Hashing	LSH
Least Squares Support Vector Regression	LSSVR
Mention Anomaly Based Event Detection	MABED
Non-negative Matrix Factorization	NMF
Singular Value Decomposition	SVD
Support Vector Machine	SVM
Term Frequency – Inverse Document Frequency	TF-IDF
Valence Aware Dictionary and sEntiment Reasoner	VADER

۱ مقدمه

۱-۱ مقدمه

شبکه اجتماعی مفهوم جدید و تازه‌ای نیست. از زمانی که انسانهای نخستین گرد آتش می‌نشستند و با یکدیگر به گفتگو می‌پرداختند، شبکه‌های اجتماعی تشکیل می‌دادند. امروزه بعد از گذشت هزاران سال و گسترش فیزیکی جوامع بشری و ظهور ابزارهای دیجیتال در این زمینه، بطور گسترده‌تری با شبکه‌های اجتماعی مواجه هستیم.

شبکه اجتماعی در حقیقت ساختاری اجتماعی است که از گره‌هایی که عموما فردی یا سازمانی هستند تشکیل شده است و این گره‌ها بر اساس یک یا چند وابستگی خاص مانند ایده‌ها، تبادلات مالی، دوستان و خویشاوندان، اتصالات وب یا انتقال بیماریها به هم متصل می‌شوند[۱]. بطورکلی اگر به پژوهش‌ها و مطالعات مختلف در زمینه شبکه‌های اجتماعی رجوع شود، دو نوع شبکه اجتماعی با ویژگیهای مختلف اما با مفهوم و معنای تقریبا مشابه وجود دارد. یکی از این زمینه‌ها، قدمتی بیش از یکصد سال دارد که بیشتر در حوزه‌های علوم انسانی مطرح است و از آن برای اشاره ضمنی به مجموعه روابط پیچیده میان افراد در سیستم‌های اجتماعی در تمامی مقیاسها، از روابط بین فردی گرفته تا بین‌المللی استفاده می‌شود. علاوه بر این وجه، امروزه مفهوم دیگری از شبکه‌های اجتماعی وجود دارد که در فضای مجازی و دیجیتالی رخ می‌دهد و از آن به عنوان شبکه‌های اجتماعی آنلاین یا مجازی نام برده می‌شود. شبکه‌های اجتماعی مجازی همان ارتباط و رابطه افراد یا سازمانها با واسطه رسانه‌های اجتماعی دیجیتال در بستر اینترنت هستند. رسانه‌های اجتماعی متنوعی در حال حاضر فعال هستند که همگی علیرغم کاربردهای مختلفی که دارند در یکسری ویژگیها با یکدیگر مشترک‌اند[۲]. رسانه‌های اجتماعی مجازی برنامه‌های کاربردی مبتنی بر وب ۲ می‌باشند. این به این معنی است که کاربران قادرند خود به ایجاد و خلق محتوا اقدام نمایند، آن را ساماندهی و تنظیم کنند، دیگران را در اطلاعات و داشته‌های خود شریک و سهیم سازند، و یا به انتقاد و تغییر بپردازند. در رسانه‌های اجتماعی، هر کاربر پروفایل ویژه خود را تولید می‌کند و در نهایت امکان ارتباط افراد مختلف با یکدیگر را فراهم می‌نماید.

امروزه با توسعه فناوری اطلاعات، مردم وقت بیشتری را در اینترنت سپری می‌کنند و بسیاری از فعالیتهای روزانه مانند خرید، مطالعه اخبار، حضور در شبکه‌های اجتماعی مجازی و جستجوی اطلاعات را بصورت آنلاین انجام می‌دهند. در این بین، استفاده از شبکه‌های اجتماعی مجازی از همه بیشتر مورد توجه کاربران اینترنت قرار گرفته است و حجم زیادی از ترافیک اینترنت در قالب متن، ویدئو و تصویر به آن اختصاص دارد[۳].

در جدول ۱-۱ چندین شبکه اجتماعی مجازی بصورت دسته‌بندی شده بر اساس نوع کاربردشان آمده است. کاربردهایی که در اینجا آورده شده‌اند شامل کاربرد عمومی، موضوعی، کسب و کار، میکروبلاگ و پیام‌رسان می‌باشند. کاربرد عمومی به این معنی است که از شبکه اجتماعی در زمینه‌های مختلف اجتماعی، سیاسی و یا اقتصادی در بین افراد مختلف در قالب متن و یا تصویر می‌توان بهره برد. منظور از کاربرد موضوعی این است که تنها در زمینه‌های خاصی مانند عکاسی، کتابخوانی و آشپزی مورد استفاده قرار می‌گیرند. کاربرد کسب و کار که صاحبان کسب و کار می‌توانند با پیوستن به این دسته شبکه‌ها، کسب و کارشان را بهبود داده و از دیگر افراد متخصص مطالب جدیدی بیاموزند. کاربرد میکروبلاگ، آن دسته از شبکه‌های اجتماعی را در بر می‌گیرد که کاربران در آنها امکان نوشتن متن‌های کوتاه و انتشار آن را دارند. در نهایت، هدف از کاربرد پیام‌رسان انتقال پیام‌های متنی، صوتی یا تصویری از مبدا به مقصد خاصی می‌باشد. لازم به ذکر است برخی از این رسانه‌های اجتماعی را می‌توان در چندین دسته مختلف قرار داد.

جدول۱-۱) چند شبکه اجتماعی مجازی رایج

عمومی	موضوعی	کسب و کار	میکروبلاگ	پیام رسان
facebook.com	*fliker.com*	*linkedin.com*	*twitter.com*	*telegram.com*
hi5.com	*foursquare.com*	*xing.com*	*yammer.com*	*whatsapp.com*
orkut.com	*goodreads.com*	*Pinterest.com*	*tumblr.com*	*viber.com*
myspaces.com	*bookcrossing.com*	*Angle.co*	*jaiku.com*	*line.me*
cloob.com	*librarything.com*	*www.nexxt.com*	*Identi.ca*	*skype.com*

در جدول ۱-۲ چند ویژگی از برخی شبکه اجتماعی رایج در ایران و جهان آورده شده است. این ویژگیها شامل سال توسعه، تعداد کاربران فعال ماهیانه که می‌تواند نشانه‌ای از محبوبیت آنها باشد، نوع داده رایجی که در آنها به اشتراک گذاشته می‌شود و آدرس اینترنتی آنها می‌باشد.

جدول ۱-۲) ویژگیهای چند شبکه اجتماعی رایج در جهان و ایران

نام	سال ایجاد	تعداد کاربران فعال ماهیانه	نوع داده رایج	آدرس اینترنتی
فیس‌بوک	۲۰۰۴	۲,۸ میلیارد	متن، عکس، ویدئو	www.facebook.com
اینستاگرام	۲۰۱۲	یک میلیارد	عکس و ویدئو	www.instagram.com
اسنپ‌چت	۲۰۱۱	۲۹۳ میلیون	عکس و متن	www.snapchat.com
توییتر	۲۰۰۶	۳۸۰ میلیون	متن	www.twitter.com
پینترست	۲۰۰۹	۴۷۸میلیون	عکس و ویدئو	pinterest.com
تیک‌تاک	۲۰۱۸	یک میلیارد	ویدئو	tiktok.com
یوتیوب	۲۰۰۵	یک میلیارد	ویدئو	YouTube.com
بله	۱۳۹۵	۱,۳۸۲ میلیون	مالی و متن	Bale.ai
آپارات	۱۳۸۹	۱۴ میلیون	ویدئو	Aparat.com
سروش	۱۳۹۵	۲,۳ میلیون نفر	متن پیا	hi.sapp.ir
نزدیکا		یک میلیون	متن	nazdika.com
ایتا	۱۳۹۶	۳,۳۸ میلیون	متن	eitaa.com
گپ	۱۳۹۳	۴,۴ میلیون	متن و عکس	Gap.im

داده‌های تولیدی در شبکه‌های اجتماعی از این بابت که جنبه‌های مختلفی از جوامع دنیای واقعی را منعکس می‌کنند بسیار ارزشمند هستند. از طرفی این داده‌ها به سادگی از طریق خزشگرهای[1] وب و یا APIهای عمومی دردسترس می‌باشند. این دو ویژگی، مهمترین دلایلی هستند که پژوهشگران به مطالعه و بررسی شبکه‌های اجتماعی مجازی می‌پردازند[۴].

[1] Crawler

کاربردهای فراوانی را می‌توان در حوزه تحلیل شبکه‌های اجتماعی مجازی مورد توجه قرار داد. از جمله این موارد می‌توان به شناسایی افراد یا تیم موثر[۲][۵]، توسعه و بهبود خزشگرها [۶]، تعیین موقعیت مکانی [۷] ، تشخیص فعالیتهای خرابکارانه[۳] [۸] ، تشخیص گروه یا انجمن[۴] [۹]، سیستم‌های پیشنهادگر[۵] [۱۰]، تشخیص و پیش‌بینی رویداد [۱۱]، کاوش پیوند و اتصالات [۱۲]، و آنالیز احساسات[۶] [۱۳] اشاره نمود.

از بین تمامی این کاربردها، مسئله "تشخیص و پیش‌بینی رویداد" به دلیل پیچیدگی و اثرات اجتماعی آن نظیر تاثیر بر رسانه‌های خبری و مدیریت بحران در حوادث طبیعی و غیرطبیعی بیشتر مورد توجه قرار گرفته است[۴]. روشهای متعددی برای تشخیص و پیش‌بینی رویداد پیشنهاد شده است که در فصل دوم به آنها اشاره خواهد شد.

لازم به ذکر است که تحلیل شبکه‌های اجتماعی مجازی معمولا از دو جنبه مورد بررسی قرار می‌گیرند. جنبه اول، از لحاظ ساختاری شبکه اجتماعی مجازی ا ست که از یک سری ویژگیهای ساختاری مانند انواع مرکزیت شبکه استفاده می‌کنند و جنبه دوم از لحاظ محتوای شبکه اجتماعی مجازی است. رو شه پی شنهادی این ر ساله بر جنبه تحلیل محتوایی شبکه اجتماعی مجازی تاکید دارد و برا ساس ترخ ارسال توییتها به شبکه اجتماعی به پیش بینی رویداد می‌پردازد.

۱-۲) تعریف رویداد

بطورکلی تعاریف متعددی از رویداد در مقالات مختلف وجود دارد. در [۱۴] به چیزی که در زمان و مکان خاصی اتفاق می‌افتد و منجر به پیامدهایی نیز می‌شود رویداد می‌گویند. در [۱۵] رویداد به چیزی گفته می‌شود که در زمان و مکان خاصی رخ می‌دهد و مورد توجه رسانه‌های خبری نیز واقع می‌شود. نویسندگان در [۱۶] رویداد را بعنوان یک اتفاق جهان واقعی می‌دانند که در یک دوره زمانی خاص ادامه

[2] Influencer
[3] malicious activities
[4] Community Detection
[5] Recommender System
[6] Sentiment analysis

می‌یابد و در توئیتر نیز در همان دوره زمانی، مورد بحث قرار می‌گیرد. در [۱۷] رویداد به هر چیز مهم و چشمگیری که در زمان و مکان خاص رخ می‌دهد گفته می‌شود. در آنجا بیان شده است که چیزی را مهم و چشمگیر می‌نامند که در رسانه‌های خبری مورد بحث واقع شود. در [۱۸] رویداد به مجموعه‌ای از پست‌های به اشتراک گذاشته شده اطلاق می‌گردد که با موضوع و عبارات یکسانی در زمان کوتاه منتشر می‌شوند. هر چیزی که باعث تغییر در حجم داده‌های متنی در زمانی خاص شود، تعریفی از رویداد است که در [۱۹] آمده است. تمامی این موارد تعاریف رویداد هستند که توسط محققین مختلف ارائه شده‌اند. البته همه این تعاریف در دو چیز اتفاق نظر دارند: توجه به زمان و مکان رخ دادن رویداد. در این پژوهش، رویداد در شبکه‌های اجتماعی به چیزی اطلاق می‌شود که تغییرات و تحولاتی چشمگیر، در برخی پارامترها و ویژگیهای شبکه اجتماعی مجازی ایجاد کند. طبیعتا این تغییرات متناسب با زمان و مکان خاصی خواهند بود. منظور از تغییرات چشمگیر در پارامترها و ویژگیهای شبکه اجتماعی این است که، مقدار آنها در هنگام وقوع رویداد، اختلاف قابل توجه‌ای با مقدارشان در وضعیت عادی شبکه (بدون رویداد) دارد. در این رساله، این اختلاف قابل توجه با استفاده از حدود آستانه شناسایی می‌شود. در پژوهش‌های انجام شده در زمینه تشخیص و پیش‌بینی رویداد، به انواع مختلفی از رویدادها اشاره شده است. از جمله این انواع می‌توان به رویدادهای طبیعی، رویدادهای ساخت بشر، رویدادهای تازه خبری، رویدادهای عمومی، رویدادهای اجتماعی و غیراجتماعی، رویدادهای محلی و سراسری، رویدادهای ناگهانی و یا حتی خاص‌تر مثل ترافیک و شیوع بیماری اشاره نمود[۴][۲۰][۲۱][۲۲].

در این پژوهش رویدادها به دو گروه رویدادهای قابل پیش‌بینی و رویدادهای غیرقابل پیش‌بینی یا ناگهانی تقسیم می‌شوند. رویدادهای قابل پیش‌بینی رویدادهایی هستند که تحت تاثیر عواملی از داخل شبکه اجتماعی بوجود می‌آیند. در حقیقت با بررسی رفتار این عوامل است که می‌توان به تشخیص و پیش‌بینی رویداد پرداخت. بعنوان مثال اعضای یک شبکه اجتماعی با ارسال توییتهایی در مورد یک رویداد، بحث و گفتگو می‌کنند. اما برخی رویدادها که در گروه رویدادهای غیرقابل پیش‌بینی دسته‌بندی می‌شوند، آنهایی هستند که عواملی خارج از شبکه منجر به وقوع آنها می‌شوند. اینگونه رویدادها را تنها

می‌توان در حین یا بعد از وقوع آنها با تحلیل شبکه اجتماعی تشخیص داد. بعنوان مثال وقوع زلزله و یا مرگ ناگهانی یک شخصیت مشهور جهانی از این نوع می‌باشند.

۱-۳ تشخیص رویداد

تشخیص رویداد عبارت است از: استخراج و نمایش خودکار یک رویداد در زمینه‌های مختلف مثل محیط زیست، امنیت، سلامت و اقتصاد. اما در[۴] تعریفی فنی‌تر بصورت زیر بیان شده است:

اگر A_x دنباله‌ای از اعمال ممکن در شبکه اجتماعی x باشد، مسئله تشخیص رویداد را می‌توان شناسایی تمامی اعضاء چندتایی E تعریف نمود:

$$E = \{e_1, e_2, \dots, e_m\} \qquad 1-1$$

در رابطه ۱-۱، e_i، i امین رویداد شناسایی شده در شبکه اجتماعی x است و m تعداد کل رویدادهای شناسایی شده است و هر رویداد شامل اطلاعات زیر می‌باشد:

$$\mathrm{e_i} = \langle\ R(e_i)\ , A^{e_i}, T_A^{e_i},\ loc_{e_i}\ , I^{e_i}\ \rangle \qquad 2-1$$

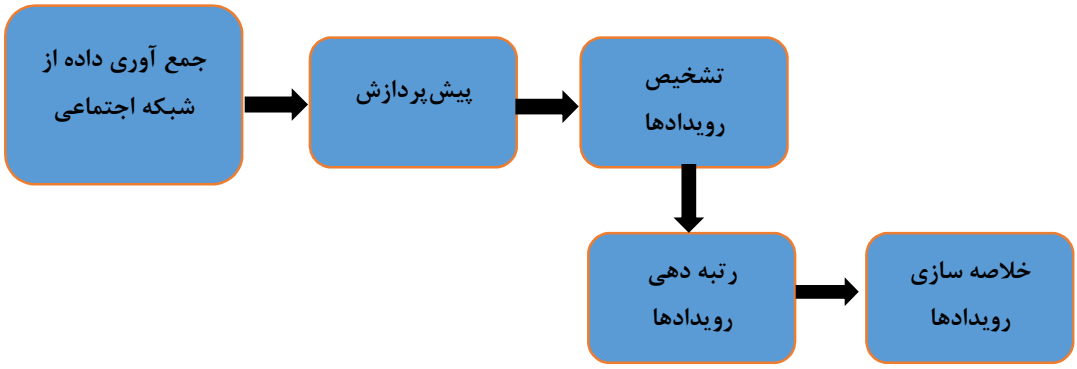
در رابطه ۱-۲، $R(e_i)$ نمایشی متنی از رویداد i ام است، A^{e_i} مجموعه اعمالی در شبکه است که به رویداد i مربوط می‌شود و $A^{e_i} \subseteq A_x$ می‌باشد، $T_A^{e_i}$ بازه زمانی است که رویداد i در آن رخ داده است، loc_{e_i} مکان وقوع رویداد i و I^{e_i} شامل کنشگران درگیر در رویداد i می‌باشد. لازم به ذکر است که اغلب سیستم‌های تشخیص رویداد لزوما تمامی اطلاعات ذکر شده را استخراج و نمایش نمی‌دهند.

ساختار کلی یک سیستم تشخیص رویداد

بطور کلی یک سیستم تشخیص رویداد همانطور که در شکل ۱-۱ قابل مشاهده است از چند جزء اساسی تشکیل می‌شود. اولین جزء یک سیستم تشخیص رویداد، واحد اخذ و جمع‌آوری داده از شبکه اجتماعی است. جریان داده‌های شبکه اجتماعی که بصورت متنی می‌باشند، پس از جمع‌آوری به واحد پیش‌پردازش منتقل می‌شوند تا برای پردازش در مراحل بعدی آماده گردند. مشخص‌سازی اجزای گرامری تشکیل دهنده متن پیام [۲۴]، مشخص کردن نهاد و اسامی خاص در متن [۲۵]، تبدیل کلمات

عامیانه[۷] به رسمی [۲۶]، فیلترکردن پیامها بر مبنای برخی معیارها [۲۷] و بدست آوردن ریشه لغات[۸] با حذف پسوند و پیشوندها [۲۸]، از جمله تکنیک‌های رایج در مرحله پیش‌پردازش می‌باشند. جزء بعدی، واحد تشخیص رویداد است. در فصل بعدی تکنیکهای مختلف تشخیص رویداد معرفی خواهند شد. جزء چهارم، رتبه‌دهی رویداد می‌باشد. در برخی سیستم‌های تشخیص رویداد، به رویدادهای شناسایی شده رتبه‌هایی تخصیص می‌یابد تا از نظر اهمیت درجه‌بندی شوند. آخرین جزء یک سیستم تشخیص رویداد واحد خلاصه‌نویسی یا ارائه گزارش است که رویدادهای شناسایی شده را بصورت یک متن قابل درک، به خروجی ارسال می‌کند.

لازم به ذکر است که لزوما یک سیستم تشخیص رویداد شامل تمامی اجزا نام برده شده نیست. مثلا می‌تواند هیچگونه رتبه‌دهی به رویدادها و یا خلاصه‌سازی رویداد در آن انجام نگیرد.



شکل۱-۱) اجزاء اصلی یک سیستم تشخیص رویداد[۲۸]

در اینجا لازم است تا تفاوت دو مفهوم تشخیص موضوع[۹] و تشخیص رویداد مشخص گردد. تشخیص موضوع که به نامهای دیگری چون تجزیه و تحلیل موضوع[۱۰]، مدلسازی موضوع[۱۱] و استخراج موضوع[۱۲] نیز

[7] Slang-Word Conversion
[8] Stemming
[9] Topic Detection
[10] Topic Analysis
[11] Topic Modeling
[12] Topic Extraction

شناخته می‌شود، یک تکنیک یادگیری ماشین است که مجموعه‌های برزگی از داده‌های متنی را سازماندهی برچسب گذاری می‌کند و موضوعات مورد بحث در آنها را استخراج می‌نماید.

تشخیص موضوع از پردازش زبانهای طبیعی برای تجزیه تحلیل زبانهای محاوره‌ای انسانی استفاده می‌کند تا بتواند الگوها و ساختارهای معنایی در متون را باز کند و منجر به ایجاد بینش بیشتر و کمک به تصمیم‌گیری شود. بعنوان رایجترین روشهای تشخیص عنوان بوسیله یادگیری ماشین می‌توان به دو روش مدلسازی عنوان و رده بندی عنوان اشاره کرد. روشهای مدلسازی عنوان از تکنیکهای بدون نظارت یادگیری ماشین بهره می‌برند در حالی که روش دوم جزء روشهای نظارت شده است و نیازمند وجود داده‌های برچسب دار جهت آموزش می‌باشد.

تشخیص عنوان می‌تواند در سه سطح مختلف مورد استفاده قرار گیرد. اولین سطح، سطح سند می‌باشد که منجر به تشخیص عناوینی در کل متن می‌گردد. برای مثال موضوع یک ایمیل یا یک مقاله خبری. دومین سطح، سطح جمله است که بمنظور درک موضوع تنها یک جمله مانند سر خط خبرها مورد استفاده قرار می‌گیرد و سطح سوم، سطح زیرجملات است. در این سطح، عبارات فرعی درون یک جمله استخراج می‌شود. بعنوان مثال کسب موضوعات مختلف از یک جمله در مورد توصیف یک محصول. در ادامه دو مثال برای کمک به درک بهتر کاربردهای تحلیل خودکار موضوع آورده شده است.

برای تعیین موضوع یک خبر می‌توان از شناسایی موضوع استفاده کرد. به عبارت دیگر هدف این است که مشخص شود یک مقاله خبری در مورد چه چیزی صحبت می‌کند. آیا سیاسی است، یا ورزشی و یا اقتصادی؟ مثلا برای این خبر " با بدست آوردن سهم هواوی از بازار، فروش آیفون در چین ۲۰ درصد کاهش یافت". یک مدل موضوعی با شناسایی کلمات و عبارات مرتبط با این موضوع (فروش، کاهش، درصد، چین، سود، سهم بازار) موضوع کلی این خبر را اقتصادی تشخیص می‌دهد.

بعنوان یک مثال دیگر یک موسسه حمل و نقل را در نظر بگیرید که برای پشتیبانی از مشتریان خود، پیامهایی از آنها را دریافت می‌کند و آنها را در حوزه‌های مختلف دسته بندی می‌نماید. پیام " سفارش

من هنوز به دستم نرسیده است" توسط یک مدلساز موضوع به عنوان یک مشکل حمل و نقل برچسب گذاری می‌شود.

آنچه در اینجا بیان شد معرفی کلی کاربرد تشخیص موضوع بود. می‌توان گفت که تشخیص موضوع دامنه وسیعتری از تشخیص رویداد را در بر می‌گیرد. در واقع تشخیص موضوع می‌تواند چیزی را شناسایی کند که ویژگی‌های یک رویداد را نداشته باشد. یعنی آن چیزی که شناسایی می‌شود، منجر به تغییرات چشمگیر برخی پارامترها در مجموعه کل اسناد یا شبکه اجتماعی نشده باشد. اما ازطرفی، یکی از روشهای تشخیص رویداد می‌تواند تشخیص عنوان باشد. بعنوان مثال موضوعات استخراج شده از میلیون‌ها توییت می‌تواند همان رویدادهای اتفاق افتاده باشد. بنابراین با توجه با همه مباحث بالا می‌توان اینگونه نتیجه‌گیری کرد که تشخیص عنوان یک موضوع جامعتر از تشخیص رویداد است.

۱-۴) پیش‌بینی رویداد

علاوه بر کاربرد تشخیص رویداد، پیش‌بینی رویداد نیز از اهمیت ویژه‌ای برخوردار است. پیش‌بینی رویداد در حقیقت به معنی بررسی رویدادهای گذشته، شناسایی عوامل موثر بر آنها و بیان وقوع رویدادهایی در آینده است. بنابراین تفاوت اصلی یک سیستم تشخیص رویداد با یک سیستم پیش‌بینی رویداد در زمان وقوع رویداد می‌باشد. سیستم تشخیص رویداد، رویدادهایی را به خروجی می‌فرستد که در زمان جاری یا قبل از آن به وقوع پیوسته‌اند در حالیکه سیستم پیش‌بینی رویداد، رویدادهایی را مشخص می‌کند که هنوز به وقوع نپیوسته‌اند و در آینده رخ می‌دهند. بعنوان مثال یک شخصیت مشهور را درنظر بگیرید که در بیمارستانی برای مدتها بستری است و نهایتا می‌میرد. سیستم تشخیص رویداد هنگام وقوع مرگ مثلا با افزایش ناگهانی تعداد توییتها، وقوع رویداد مربوطه را تشخیص می‌دهد. اما یک سیستم پیش‌بینی رویداد مانند روش پیشنهادی، به تشخیص روند رو به رشد تعداد توییتها، پیش‌بینی می‌کند یک رویداد مرتبط با آن شخص مشهور در آینده به وقوع خواهد پیوست. این پیش‌بینی می‌تواند برای خبرگزاری‌ها و حتی مدیریت بحران موثر باشد. این کار توسط افراد خبره در حوزه‌های مختلف

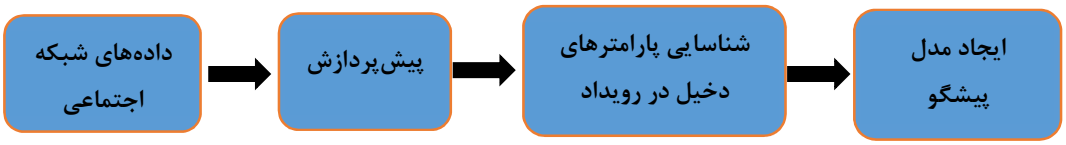
انجام می‌شود، اما آیا این پیش‌بینی‌ها توسط ماشین نیز قابل انجام است؟ پیش‌بینی توسط ماشین مزیتهای متعددی دارد که از آن‌جمله می‌توان به این موارد اشاره نمود: هزینه پیش‌بینی توسط ماشین نسبت به افراد خبره کمتر است، امکان پیش‌بینی طیف وسیعی از رویدادها توسط ماشین وجود دارد، پیش‌بینی انسان ممکن است تحت تاثیر گرایشات و منافعش باشد، و بررسی حجم زیاد اطلاعات با سرعت بالا برای انسان غیر ممکن است، در حالیکه ماشین با سرعت و دقت بهتری به انجام آن می‌پردازد.

اصولا به منظور انجام یک پیش‌بینی با کمترین خطا، در گام نخست باید یک مدل پیشگو، مبتنی بر پارامترهای موثر در آن چیزی که می‌خواهد پیش‌بینی شود، ایجاد شود (شکل ۱-۲ الف). به عبارت دیگر باید با نظارت بر داده‌های گذشته، پارامترهای موثر در وقوع رویداد شناسایی شود و نقش آنها در قالب یک مدل ترسیم شود. سپس در گام دوم، با اعمال مقادیر پارامترهای موثر در زمان جاری بر مدل پیشگو، عمل پیش‌بینی انجام شود (شکل ۱-۲ ب). بعنوان مثال در پیش‌بینی هواشناسی، پارامترهایی چون دما، رطوبت، فشار جو، فشردگی ابرها، سرعت و جهت باد همواره نظارت و بررسی می‌شوند. از این رو باید پارامترهای مهم و دخیل در وقوع رویداد را شناسایی و نظارت کرد. از جمله این پارامترها می‌توان به نرخ ثبت‌نام کاربران جدید، نرخ ارسال پیام به شبکه، تعداد هشتگها، تعداد بازتوئیت‌ها، تعداد یادکردها[۱۳]، تعداد لینکها، تعداد پاسخ کاربران به پیامها و نرخ ارسال متن توسط افراد موثر اشاره نمود. پارامتری که در این پژوهش برای پیش‌بینی رویداد مورد استفاده قرار گرفته است نرخ ارسال توییت به شبکه اجتماعی می‌باشد. در شکل ۱-۲ مراحل پیش‌بینی یک رویداد آورده شده است.

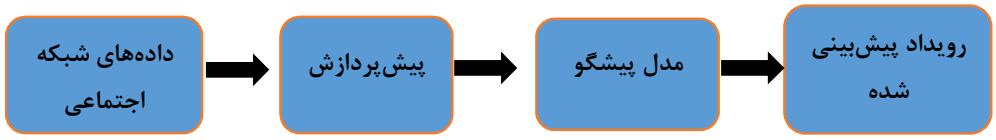
۱-۵) چالش‌های مسئله

در مسئله تحلیل شبکه‌های اجتماعی و خصوصا تشخیص رویداد و پیش‌بینی آن، چالشهای زیادی وجود دارد که از مهمترین آنها می‌توان به این موارد اشاره کرد[۴]:

[13] Mentions



الف) گام نخست پیش‌بینی: ایجاد مدل پیشگو



ب) گام دوم پیش‌بینی: پیش‌بینی رویداد بر اساس تغییرات پارامترها

شکل ۱-۲) مراحل پیش‌بینی یک رویداد

چالش اول، حجم زیاد داده‌ها است. همانطور که قبلا بیان شد، شبکه‌های اجتماعی حجم زیادی از اینترنت را به خود اختصاص داده‌اند. بعنوان مثال در توئیتر روزانه بالغ بر ۵۰۰ میلیون توئیت توسط کاربران ارسال می‌شود. این مسئله از طرفی شاید مزیت به حساب بیاید چرا که داده‌ها فراوانند اما عیب مسئله اینجاست که نگهداری و مدیریت این حجم داده‌ها به سادگی امکانپذیر نیست.

چالش دوم، سرعت ایجاد و پویایی داده‌هاست. داده‌هایی که در شبکه‌های اجتماعی تولید و جریان می‌یابند حوزه‌ها و مباحث مختلف و متنوعی را در بر می‌گیرند. یک کاربر در طول مدت شبانه روز ممکن است در زمینه‌های مختلف به تولید محتوا بپردازد. از اینرو در یک بازه زمانی محدود، پیامهای مرتبط به موضوعات و رویدادهای متنوعی به شبکه اجتماعی وارد می‌شود.

چالش سوم، نیاز به ارائه نتایج با دقت بالا در برخی کاربردها وجود دارد. بعنوان مثال تشخیص رویداد شیوع یک بیماری واگیر خطرناک، یا وقوع زلزله یا حوادث اجتماعی به منظور کنترل به موقع بحران از جمله این کاربردها می‌باشند.

چالش چهارم، محتوای کوتاه یا خلاصه شده کلمات در شبکه اجتماعی است. این مسئله خصوصا در شبکه‌های اجتماعی میکروبلاگ، که کاربران در تعداد کلمات محدودیت دارند شدیدتر است. کاربران مجبورند پیامها و نظرات خود را به صورت خلاصه، گاها با سبک دستوری غیررسمی و کلمات مخفف بیان کنند.

چالش پنجم، وجود نظرات گاها به زبانهای محاوره‌ای مختلف است. از آنجایی که در مورد موضوعات مطرح شده توسط کاربران، افراد مختلف از هر جای دنیا می‌توانند نظرات خودشان را بیان کنند، ممکن است نظرات افراد با زبانهای محاوره‌ای مختلف بیان شده باشد که تحلیل محتوا را با مشکلاتی مواجه می‌کند.

ششمین چالش این است که برای یک کاربرد خاص در زمینه تحلیل شبکه‌های اجتماعی، از کدام ویژگی آنها استفاده شود؟ به عبارت دیگر برای یک کاربرد، استفاده از پردازش متن مناسبتر است یا تصویر، هشتگ مناسبتر است و یا ویژگیهای کاربران مانند موقعیت مکانی آنها.

چالش هفتم، عدم وجود مجموعه داده استاندارد متناسب با کاربردها است. معمولا در متون مختلف، پژوهشگران از دادهایی که خودشان از شبکه اجتماعی در بازه زمانی خاص استخراج می‌نمایند برای تحلیل شبکه استفاده می‌کنند، که این خود می‌تواند مقایسه تکنیک‌های مختلف با یکدیگر را با مشکل مواجه سازد.

چالش هشتم، وجود داده‌ها و اخبار جعلی است که می‌تواند به شدت عملکرد سیستم‌های تشخیص و پیش‌بینی رویداد را با مشکل مواجه نماید.

۱-۶) اهمیت و ضرورت انجام تحقیق

تحلیل شبکه‌های اجتماعی واقعی عمری چند صد ساله دارد و بسیار مورد استفاده قرار گرفته است. اما عمر شبکه‌های اجتماعی مجازی به دو دهه نیز نمی‌رسد. به عنوان مثال، توئیتر بعنوان شبکه اجتماعی مورد استفاده در این پژوهش، از ژوئیه ۲۰۰۶ آغاز به کار کرده است. بنابراین، مبحث تحلیل شبکه‌های اجتماعی بسیار جدید و نوپاست. اغلب مطالعات و پژوهشها مربوط به سالهای ۲۰۱۰ به بعد می‌باشند. در آینده نیز با توجه به گسترش فیزیکی جوامع بشری و ظهور مسائل اجتماعی مختلف و وجود داده‌های فراوان در این شبکه‌ها، تحلیل این شبکه‌های اجتماعی بیشتر مورد توجه قرار خواهد گرفت.

حال سوال اینجاست که، چرا با وجود چالشهای مطرح شده در بخش قبل و فراوان بودن رسانه‌های رادیویی، تلویزیونی و خبری، هنوز تشخیص و پیش‌بینی رویداد از طریق شبکه‌های اجتماعی مهم است

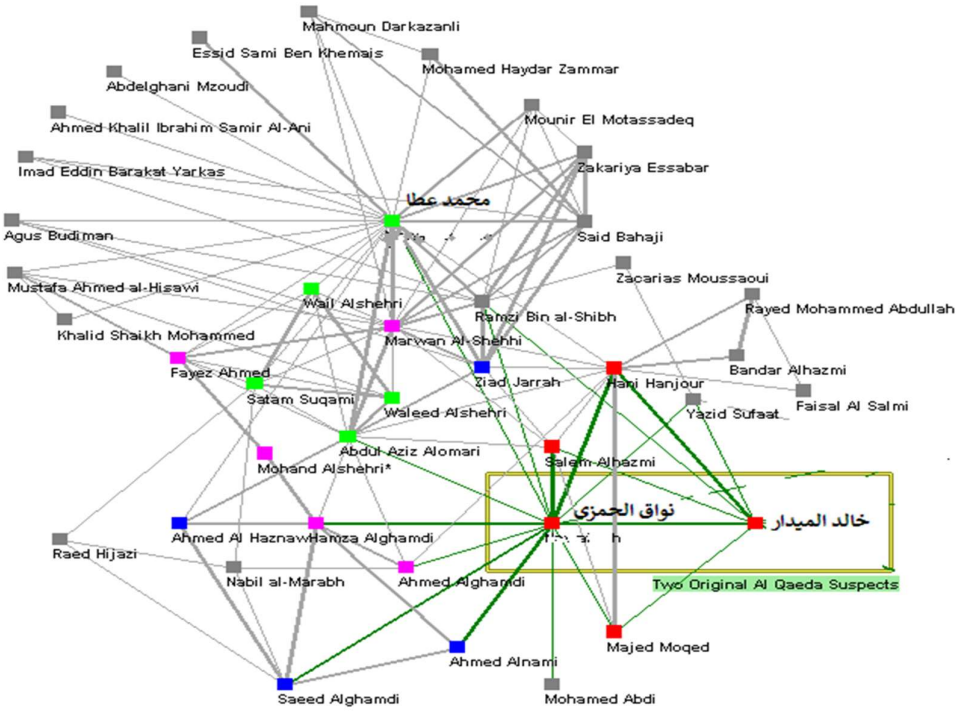
و مورد توجه پژوهشگران قرار می‌گیرد؟ می‌توان در پاسخ به این سوال دلایل زیر را متذکر شد: امروزه با گسترش شبکه‌های اجتماعی مجازی و دسترسی به دستگاه‌های دیجیتال، اغلب افراد به شبکه‌های اجتماعی متصل هستند و از آن استفاده می‌کنند به عبارت دیگر ضریب نفوذ رسانه‌های اجتماعی مجازی در بین مردم رو به افزایش است. در شبکه‌های اجتماعی هرکاربر بعنوان یک خبرنگار عمل می‌کند و کوچکترین وقایع محل حضورش را در شبکه ارسال می‌کند. در برخی مقالات، بر اساس این واقعیت، از کاربران بعنوان حسگرهایی یاد شده که اتفاقات مختلف در شبکه‌های اجتماعی را نشر می‌دهند. انتشار اطلاعات در شبکه‌های اجتماعی به سرعت انجام می‌شود. اخبار و داده‌ها بر اساس اعتماد متقابل افراد به همدیگر گسترش می‌یابند. یعنی یک فرد معمولا زمانی دست به انتشار یک پیام یا متن رسیده از دیگری می‌زند که اطمینان کافی به آن شخص داشته باشد. ممکن است خبرگزاری‌ها و رسانه‌های خبری سلیقه‌ای عمل کنند و برخی رویدادها را کلا پوشش ندهند و یا مطابق با سلیقه و دیدگاه خودشان انتشار دهند. علاوه بر موارد فوق می‌توان مثالهای زیادی ذکر کرد که تحلیل شبکه‌های اجتماعی واقعی و مجازی گاها رهگشای مشکلات مختلف جوامع بشری بوده است.

بعنوان مثال، در زمینه تحلیل شبکه‌های اجتماعی واقعی در سالهای ۱۹۳۱ و ۱۹۳۲ ، مطالعه‌ای توسط التون مایو[۱۴] به منظور بررسی اثربخشی انگیزه پرداخت حقوق در بهره‌وری تولید انجام شد. این مطالعه تحت عنوان اتاق سیم‌کشی بانک مشهور است و بر روی گروهی از ۱۴ کارگر در یک سالن تولید تجهیزات سوئیچینگ تلفن انجام گردید. بدین منظور در سالن کار، ناظرانی قرار گرفتند و تمامی ارتباطات کلامی و کاری افراد را ثبت می‌کردند و در نهایت این ارتباطات را در قالب گرافهایی به تصویر کشیدند و با تحلیل این گرافها موفق به شناسایی اجتماع‌های کوچک و موثر بین آنها شدند. در نهایت به این نتیجه رسیدند که کارگران در مقابل کنترل و انگیزه‌های مدیریتی و مالی، بیشتر به خواسته‌های گروه‌های همسالان خود پاسخ می‌دهند. مثال فوق نشان می‌دهد که با تحلیل رفتار افراد در جامعه‌هایی

[14] Elton Mayo

حتی کوچک می‌توان حقایق مفیدی در مورد آنها درک کرد و بطور خاص برای مثال فوق، جهت بهبود بهره‌وری سیستم، اقداماتی را انجام داد.

شناسایی مضنونین حادثه یازده سپتامبر، مثال دیگری از تحلیل شبکه‌های اجتماعی واقعی است که توسط FBI [15] انجام شد. آنها دو نفر به نامهای «خالد المیدار» و «نواق الحمزی» که دو سال قبل از حادثه در جلسه‌ای با سران القاعده شرکت کرده بودند را شناسایی کردند. سپس تمامی افرادی که بطور مستقیم و غیرمستقیم با آنها در ارتباط بودند را شناسایی نمودند و بر مبنای این ارتباطات به گرافی مانند شکل۱-۳ رسیدند. با تحلیل این گراف، به رهبری «محمد عطا» و شناسایی کلیه افراد دخیل در هواپیماربایی پی‌بردند. البته حرف و حدیث در مورد عامل یا عاملین حادثه یازده سپتامبر و نقش مستقیم دولت آمریکا در وقوع آن زیاد است. از جمله این موارد، تیری میسان نویسنده فرانسوی در کتاب "دروغ بزرگ"، با دلایل قوی دست داشتن دولت وقت آمریکا در وقوع حادثه را مطرح می‌کند.

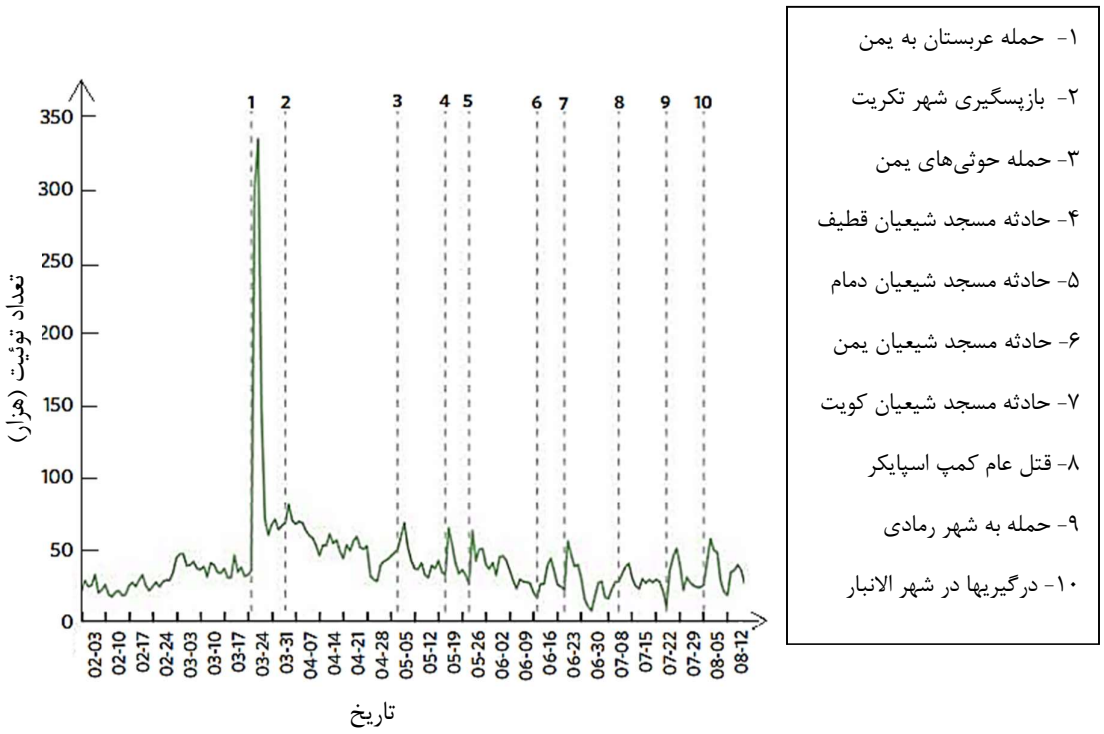


شکل۱-۳) گراف مربوط به مضنونین حادثه ۱۱ سپتامبر

[15] Federal Bureau of Investigation

شناسایی منشا شیوع یک بیماری واگیر مثل کوویید۱۹ با تحلیلی مشابه مثال قبل امکانپذیر است. چنانچه برای افراد مبتلا، مراجعات هفته اخیرشان به مراسمات، فروشگاه‌ها، رستورانها و یا مسافرتها ثبت شود، با رسم یک گراف ارتباطی می‌توان به منشا همه‌گیری در شهر و همچنین پیش‌بینی روند آن پی برد. از جمله موارد دیگر که بر اهمیت موضوع تحلیل شبکه‌های مجازی تاکید دارد، تاثیر متقابل شبکه‌های اجتماعی مجازی و دنیای واقعی بر یکدیگر است. نمونه‌های متعددی را می‌توان نشان داد که با بروز رخدادها و بحرانهای اجتماعی مثل حوادث طبیعی و غیرطبیعی، تغییرات شگرفی در شبکه‌های اجتماعی مجازی بوجود می‌آید. این تغییرات می‌تواند بطور خیلی ساده در تعداد پیامهای رد و بدل شده در شبکه اجتماعی مجازی نمود پیدا کند.

بعنوان یک مثال در [۲۹] تغییرات تعداد پیامهای ارسالی ضد شیعی در شبکه اجتماعی توئیتر و ارتباط آن با حوادث دنیای واقعی بررسی شده است. در پژوهش یادشده با استفاده از کلمات کلیدی که یکی از تکنیک‌های اولیه در شناسایی پیامها و رویدادهای خاص می‌باشد، ابتدا پیامهای ضد شیعی شناسایی می‌شوند. از جمله این کلمات کلیدی می‌توان به "شیعه"، "حزب شیطان" و "علوی" اشاره نمود. در نهایت مشخص شد که افزایش ناگهانی هر کدام از این نوع پیامها با یک رخداد واقعی مطابقت دارد. در شکل۱-۴ هر یک از قله‌ها متناظر با یک رویداد در دنیای واقعی است. بنابراین می‌توان گفت که وقایع دنیای واقعی بر عملکرد شبکه‌های اجتماعی مجازی موثر هستند.



شکل۱-۴) تعداد توئیت ضد شیعی از فوریه تا آگوست ۲۰۱۵ در توئیتر و ارتباط آنها با رویدادهای جهان[۲۹]

این اثرگذاری را می‌توان در جهت معکوس نیز بررسی نمود. شاید بتوان بارزترین رویداد دنیای واقعی که تحت تاثیر شبکه‌های اجتماعی مجازی اتفاق افتاد را در انقلاب مصر دید. گرچه می‌توان نمونه‌های دیگر مثل انتخابات‌ها و بازاریابی را نیز مثال زد. در حادثه انقلاب مصر، این اثرگذاری به قدری ملموس بود که آن را انقلاب فیس‌بوکی می‌نامند. یک نسل فیس‌بوکی از افراد تحصیلکرده تصمیم گرفتند تا با سنت‌گرایی، دیکتاتوری و ناامنی اقتصادی اجتماعی مبارزه کنند و از شبکه اجتماعی جهت اطلاع رسانی و اشتراک‌گذاری تصاویر استفاده نمودند. قتل خالد سعید توسط پلیس و ایجاد صفحه‌ای در فیس‌بوک بنام "ما همه خالد سعید هستیم"[۱۶] به تحریک مردم در انقلاب مصر کمک کرد. این صفحه فیس‌بوک توانست هزاران طرفدار را جذب کند و جامعه بین الملل، دولت مصر را تحت فشار قرار داد تا دو پلیس متهم به قتل را محاکمه کند. چند ماه بعد در حادثه انقلاب مصر به گفته تحلیلگران سیاسی، این صفحه نقش مهمی‌در تحریک مردم و حس نوع دوستی آنها داشت. از این رو با توجه به دلایل و مثالهای فوق،

[16] We are all Khaled Said

می‌توان اهمیت تحلیل شبکه‌های اجتماعی مجازی و تشخیص و پیش‌بینی خودکار رویدادها از آنها را درک کرد.

۱-۷) شبکه اجتماعی توئیتر و ویژگیهای آن

از آنجاییکه روش پیشنهادی بر روی شبکه اجتماعی توییتر مورد ارزیابی قرار گرفته است، در این بخش شبکه اجتماعی توئیتر و ویژگیهای آن معرفی می‌شود. دلایل متعددی برای انتخاب توییتر به عنوان شبکه اجتماعی مورد تحلیل این رساله می‌توان نام برد. این دلایل شامل: داده رایج در توییتر متن است، اغلب روشهای تحلیل شبکه‌های اجتماعی خصوصا در مورد تشخیص و پیش بینی رویداد مبتنی بر متن هستند، کار کردن بر روی داده‌های متنی در مقابل عکس و ویدئو راحتتر و عمومیت بیشتری دارد، پایگاه داده‌های متنی توییتر مرتبط با کاربرد تشخیص رویداد فراوانتر و در دسترس‌ترند، از توییتر برای تشخیص موضوعات ترند و مطرح معمولا استفاده می‌شود، اخبار و رویدادها معمولا در توییتر سریعتر منتشر می‌شوند و نهایتا دسترسی به داده‌های برخی شبکه‌های اجتماعی همانند شبکه‌های اجتماعی ایرانی امکانپذیر نیست. استفاده شده است این است که اولا بیشتر به داده‌های متنی استوار است و ثانیا نسبت به دیگر شبکه‌های اجتماعی، رویدادهای جهان واقع را سریعتر انتشار می‌دهد[۳۰].

توئیتر یک سرویس خبری و شبکه اجتماعی آمریکایی است و در حال حاضر محبوبترین سرویس میکروبلاگ نویسی است که به سرعت در حال رشد می‌باشد. بر اساس رتبه بندی Alexa[۱۷] در سال ۲۰۲۱، چهارمین سایت محبوب مرتبط با شبکه‌های اجتماعی بین کاربران اینترنت بوده است. این محبوبیت را می‌توان در تعداد پژوهشهای انجام شده در مورد توئیتر نیز مشاهده نمود. توئیتر به کاربران امکان ارسال و خواندن پیامهای کوتاه را می‌دهد. همانند وبلاگها، توئیتر اجازه اشتراک‌گذاری هر نوع اطلاعات، عقاید، نظرات و ایده‌ها را برای در دسترس بودن تمام آنچه در دنیا در حال اتفاق افتادن است

[17] https://www.alexa.com/topsites/

می‌دهد. تا نوامبر ۲۰۱۷ حداکثر طول یک توئیت ۱۴۰ نویسه بود که از آن موقع به بعد، غیر از زبانهای چینی، ژاپنی و کره‌ای، در مابقی زبانها می‌توان تا دو برابر نویسه، یعنی ۲۸۰ نویسه، در یک توئیت نیز استفاده کرد. بدلیل همین محدودیت است که خلاصه‌نویسی کلمات در توئیتر بسیار رایج است. مثلا از کلمه Twttr به جای Twitter و از "پاد" به جای " پایگاه اطلاع رسانی دولت" استفاده می‌شود. علاوه بر ارسال و دریافت توئیتها، هر کاربر می‌تواند به تعریف پروفایل و ارتباط با دوستان خود بپردازد و از توئیتر بعنوان یک شبکه اجتماعی نیز استفاده نماید. بطور پیش فرض کاربران می‌توانند بدون تایید کاربر دیگر به دنبال کردن آن بپردازند و کارهایی که انجام می‌دهد را رصد کنند. هیچ محدودیتی در تعداد دنبال کنندگان[۱۸] وجود ندارد هرچند هر شخص تنها می‌تواند ۲۰۰۰ نفر را دنبال نماید. پیام‌ها بطور پیش فرض قابل دسترس برای همگان است حتی افرادی که در توئیتر عضویت ندارند، مگر اینکه در تنظیمات کاربری تغییراتی ایجاد شده باشد که پیامها تنها توسط دنبال‌کنندگان مشاهده شود. اشتراک‌گذاری تنها چهار عکس، یک تصویر متحرک و یک فیلم با طول دو دقیقه و ۲۰ ثانیه نیز در هر توئیت امکانپذیر است.

اعمالی چون fave[۱۹] کردن که همان لایک یا پسندیدن می‌باشد، باز توئیت کردن[۲۰]، یادکِرد[۲۱] نمودن، و پاسخ دادن[۲۲]، مهمترین کارهایی است که کاربران در محیط توئیتر انجام می‌دهند. امکان ارسال پیام مستقیم به فردی خاص نیز میسر است.

اما یکی از مهمترین امکانات دیگر توئیتر که در پژوهشها نیز از آن زیاد استفاه شده است، هشتگ[۲۳]ها هستند که البته منحصر به توئیتر نیستند و در دیگر شبکه‌های اجتماعی نیز کاربرد دارند. هشتگ یک کلمه یا عبارت است که در کنار علامت # قرار می‌گیرد و از آن برای شناسایی کلمات کلیدی یا

[18] Follower
[19] Favorite
[20] Retweet
[21] Mention
[22] Reply
[23] Hashtag

دسته‌بندی کردن متن استفاده می‌شود. در واقع کلمه‌ای که در کنار # قرار می‌گیرد تبدیل به یک متاتگ شده و کلیک کردن بر روی آن، کاربر را به سمت دیگر محتواهای مرتبط با آن هدایت می‌کند. در بسیاری از پژوهشها از هشتگ به عنوان یک ویژگی موثر استفاده شده است. در واقع می‌توان به هشتگ بعنوان موضوع یا عنوان یک توییت اشاره کرد و در فیلتر کردن توییتهای مرتبط با یک رویداد خاص مورد استفاده قرار داد.

۱-۸) هدف اصلی تحقیق

هدف از انجام این تحقیق ارائه روشی جهت شناسایی و پیش‌بینی رویدادهای جهان واقع از طریق تحلیل شبکه اجتماعی مجازی می‌باشد. انجام این تحقیق، می‌تواند در حوزه‌های مختلفی چون سیاسی، اقتصادی، اجتماعی، بهداشت و سلامت و رسانه‌های خبری مفید واقع شود. علاوه بر این می‌تواند منجر به شناسایی و پیش‌بینی رویدادهای مهم، معضلات و ناهنجاریهای مختلف و نتیجتا به پیشگیری و راه حل بیانجامد.

۱-۹) نوآوریها و دستاوردها

نوآوریهایی که می‌توان در روش پیشنهادی به منظور پیش‌بینی رویداد در نظر گرفت شامل این موارد است: نوآوری اول، طراحی و پیاده‌سازی روشی برای پیش‌بینی کلیه رویدادهایی که می‌توانند در آینده اتفاق بیافتند می‌باشد. در حالیکه روشهای موجود در زمینه پیش‌بینی رویداد، به پیش‌بینی رویدادهای خاص تاکید دارند. نوآوری دوم، در نظر گرفتن ویژگی نرخ ارسال توییتها، از بین ویژگیهای مختلفی است که می‌توانند برای تحلیل شبکه‌های اجتماعی مورد استفاده قرار گیرند، می‌باشد. این انتخاب، پیاده‌سازی روش پیشنهادی را قابل فهم و ساده کرده است. سومین نوآوری، استفاده از روشهای NMF [۳۱] و ddCRP[۲۴] [۳۲] به ترتیب برای مقداردهی اولیه خوشه‌ها و خوشه‌بندی افزایشی توییتها است که به تفضیل در فصل‌های بعدی معرفی خواهند شد. نوآوری چهارم که در بخش پیاده‌سازی روش

[24] distance dependent Chinease Resturant Process

پیشنهادی استفاده شد، در نظرگرفتن یک حد آستانه برای حداقل نرخ ورود توییتها به خوشه‌ها است که باعث حذف تعداد زیادی از خوشه‌های غیرمفید می‌شوند. به این ترتیب بر دقت و سرعت روش پیشنهاد نقش موثری دارد. نوآوری پنجم، معرفی یک اندیکاتور جدید و موثر مبتنی بر نرخ ارسال توییتها برای پیش‌بینی نتیجه انتخابات بعنوان یک رویداد خاص و امکان پیش‌بینی نتیجه انتخابات در بازه‌های زمانی دلخواه قبل از برگزاری آن می‌باشند.

پیش‌بینی وقوع با دقت ۸۵ درصد برای مجموعه داده موجود و پیش‌بینی نتیجه انتخابات در بازه‌های زمانی دلخواه قبل از آن با یک دقت بسیار خوب، مهمترین دستاوردهای رساله حاضر می‌باشد.

۱-۱۰) سازماندهی طرح تحقیق

طراحی و توصیف روش پیشنهادی برای پیش‌بینی رویداد بر اساس تحلیل شبکه اجتماعی در ادامه این رساله در چهار فصل آمده است:

۱- فصل دوم بر کارهای انجام شده در زمینه تشخیص و پیش‌بینی رویداد تمرکز دارد و علاوه بر این روشهای خوشه‌بندی افزایشی ddCRP و تجزیه ماتریسی NMF با جزییات معرفی می‌گردند.

۲- در فصل سوم، روش پیشنهادی برای پیش‌بینی رویداد و همچنین روش پیشنهادی برای پیش‌بینی نتیجه انتخابات آورده شده است.

۳- در فصل چهارم، پیاده‌سازی و آزمایشات انجام شده بر اساس روش پیشنهادی در این رساله آورده شده است.

۴- نهایتا در فصل پنجم، نتیجه‌گیری و کارهای آتی در رابطه با تحلیل شبکه‌های اجتماعی و پیش‌بینی رویداد بیان شده است.

۱-۱۱) جمع بندی

در این فصل مفهوم رویداد، تشخیص و پیش‌بینی آن بر اساس تحلیل شبکه‌های اجتماعی بطور کامل تشریح شد. اهمیت تحلیل شبکه‌های اجتماعی واقعی و مجازی در قالب چندین مثال واقعی نشان داده شد. وجود ارتباط و همبستگی بین وقایع دنیای واقعی و تغییرات در برخی ویژگیهای شبکه‌های اجتماعی مطرح گردید. اجزاء یک سیستم تشخیص رویداد شامل: جمع آوری و کسب داده‌ها، پیش‌پردازش، تشخیص رویداد، رتبه‌بندی و نمایش رویداد معرفی گردید. سپس گامهای مطرح در پیش‌بینی رویداد شامل ایجاد یک مدل پیشگو و سپس پیش‌بینی رویداد بیان شد. چالشهایی که در حوزه تحلیل شبکه‌های اجتماعی مجازی مطرح هستند و دلیل اهمیت موضوع و هدف اصلی تحقیق دیگر مباحثی بودند که در این فصل پوشش داده شد. با توجه به مباحث مطرح شده در این فصل، پیش‌بینی رویداد با استفاده از تحلیل شبکه‌های اجتماعی امکانپذیر است و می‌تواند نقش موثری در اطلاع‌رسانی، پیشگیری و کنترل بحرانهای اجتماعی داشته باشد.

۲ روشهای تشخیص و پیش‌بینی رویداد

۲-۱ مقدمه

در این فصل کارهای مرتبط با روش پیشنهادی برای پیش‌بینی رویداد، در سه بخش ارائه شده‌اند. از آنجاییکه پیش‌بینی رویداد و تشخیص رویداد دو کاربرد نزدیک و مرتبط با یکدیگر در زمینه تحلیل شبکه‌های اجتماعی می‌باشند، در بخش نخست، انواع روشهای تشخیص رویداد آمده است. اما چون روش پیشنهادی مبتنی بر خوشه‌بندی افزایشی می‌باشد، چندین روش تشخیص رویداد مبتنی بر خوشه‌بندی افزایشی با جزییات بیشتری در بخش دوم معرفی شده‌اند. نهایتا، بخش سوم به پژوهشهای انجام شده در زمینه پیش‌بینی نتیجه انتخابات اختصاص دارد.

۲-۲) روشهای تشخیص رویداد در شبکه‌های اجتماعی مجازی

در دهه اخیر تکنیک‌های متعددی جهت تشخیص رویداد پیشنهاد شده است که می‌توان آنها را به چهار دسته کلی تقسیم‌بندی نمود[۴][۲۸]: روشهای مبتنی بر تشخیص وقوع بی نظمی یا جذابیت لغت، روشهای مبتنی بر خوشه‌بندی افزایشی، روشهای مبتنی بر مدل‌بندی موضوعی و روشهای تشخیص اولین وقوع. در ادامه، این روشها مختصرا توضیح داده شده‌اند. البته از آنجایی که روش پیشنهادی مبتنی بر خوشه‌بندی افزایشی می‌باشد، در بخشی مجزا به آن پرداخته شده است.

روشهای مبتنی بر تشخیص وقوع بی‌نظمی یا جذابیت لغت، معمولا بصورت خودکار به شناسایی لغاتی که به کرات توسط کاربران مورد استفاده قرار می‌گیرند، یا به عبارتی انفجاری هستند می‌پردازند. در این روشها، بروز هر گونه شرایط خارج از وضعیت عادی در شبکه به منزله وقوع یک رویداد در نظر گرفته می‌شود. از جمله این شرایط غیر عادی می‌توان به مواردی همچون، استفاده غیرطبیعی از یک لغت توسط کاربران، فعالیتهای زیاد در یک موقعیت مکانی و تغییر وضعیت احساسات پیام‌ها اشاره کرد. این روشها جریان شبکه را دنبال می‌کنند و احتمال وقوع رویداد را هنگام مشاهده بی‌نظمی اعلان می‌کنند. بعنوان مثال چنانچه تعداد پیامهای مرتبط با یک موقعیت مکانی از حد نرمال قبلی افزایش داشته باشد، می‌تواند نشانه وقوع یک رویداد در آن منطقه باشد. یا چنانچه وضعیت احساسات نهفته در پیامهای

کاربران برای یک کالا یا نهاد خاص دچار تغییر شود، مثلا از احساسات مثبت به منفی تغییر کند، می‌تواند نشانه وقوع یک رویداد در مورد آن کالا یا نهاد باشد.

یکی از روشهایی که بر اساس شناسایی بی‌نظمی به تشخیص رویداد می‌پردازد تکنیک MABED [۳۳] است. این تکنیک یک روش مبتنی بر تشخیص بینظمی یادکردها در شبکه اجتماعی توییتر می‌باشد. یادکرد قابلیتی در برخی شبکه‌های اجتماعی است که یک کاربر، توجه کاربری دیگر را به یک موضوع جلب می‌کند. بمنظور استفاده از یادکرد، نام کاربر موردنظر به همراه علامت @ در متن آورده می‌شود. روش MABED بر اساس بی‌نظمی که در تعداد یادکردهای کاربران در بازه‌های زمانی مختلف روی می‌دهد، به تشخیص رویداد می‌پردازد. چنانچه تعداد یادکردها در یک بازه زمانی خاص، از یک مقدار قابل انتظار بیشتر باشد، به معنی بی‌نظمی و وقوع یک رویداد است. سپس کاماتی که بیشترین هم وقوعی با یادکردها در یک بازه زمانی را داشته باشند، بعنوان توصیفی از رویداد موردنظر انتخاب می‌گردند. از آنجاییکه یادکردها بعنوان اتصالات پویا بین کاربرن در نظر گرفته می‌شوند، می‌توان گفت، روش MABED علاوه بر محتوای متنی، جنبه اجتماعی توییتر را نیز درنظر می‌گیرد.

از جمله روشهای دیگری که مبتنی بر تشخیص وقوع بی‌نظمی می‌باشند، می‌توان به [۳۷-۳۴ و ۱۸] اشاره نمود.

روشهای مبتنی بر مدل‌سازی موضوعی[۲۵]، به مدلهای موضوعی آماری وابسته هستند. این روشها با شناسایی موضوعات پنهان از جریان دادههای شبکه اجتماعی به تشخیص وقایع جهان می‌پردازند. روشهای مبتنی بر مدل‌سازی موضوع، هر پیام را با یک توزیع احتمالی بر روی موضوعات پنهان متنوع همراه می‌کند، تا ساختارهای محتوایی مخفی را از یک مجموعه توئیت پیدا کند و در تشخیص رویداد استفاده نماید. این روشها مبتنی بر مدلهای پیچیده هستند تا موضوعات پنهان را استخراج نمایند. از جمله این روشها می‌توان به [۴۰-۳۸] اشاره نمود.

[25] Topic Modeling

روشهای مبتنی بر شناسایی اولین وقوع معمولا در اخبار و رسانه‌های خبری کاربرد دارند و از آنها با عنوان تشخیص رویداد جدید[۲۶] نیز نام می‌برند. هدف این است که خبرها و محتویاتی که به تازگی در شبکه شروع به انتشار یافته‌اند را شناسایی کنند، چرا که می‌توانند نشانه وقوع رویداد باشند. معمولا برای کاهش احتمال تشخیص اشتباه یک رویداد با این روش، از رسانه‌های دیگری چون ویکی‌پدیا نیز استفاده می‌شود.

در [۴۱]، الگوریتمی مبتنی بر جستجو در بین نزدیکترین همسایگان به نام FSD جهت یافتن رویدادهایی که تاکنون ظاهر نشده‌اند، ارائه شده است. به منظور افزایش سرعت در جستجوی یافتن نزدیکترین همسایگان، از روش LSH[۲۷] [۴۲] استفاده شده است. LSH تکنیکی برای یافتن کارامد اسناد مشابه در مجموعه داده‌های بزرگ است. با استفاده از یک عملیات درهم‌سازی بر روی اسناد، هر توییت به یک سطل[۲۸] تخصیص می‌یابد و بدین ترتیب احتمالا توییتهای مشابه، در یک سطل یکسان قرار می‌گیرند و فرایند یافتن نزدیکترین همسایه با سرعت بالایی انجام می‌شود. با رسیدن توییتها، نزدیکترین همسایه آنها محاسبه می‌شود. اگر یک توییت جدید T به اندازه کافی با نزدیکترین همسایگان خود متفاوت باشد، در اینصورت منحصر به فرد است و ممکن است اولین موردی باشد که بعنوان یک داستان جدید ارسال شده است. توییتهای بعدی که نزدیکترین همسایه آنها همان توییت T می‌باشد در مجموعه یکسانی قرار می‌گیرند. [۴۴ و ۴۳] از جمله دیگر روشهای مبتنی بر شناسایی اولین وقوع می‌باشند.

۲-۳) روشهای تشخیص رویداد مبتنی بر خوشه‌بندی افزایشی

الگوریتم‌های خوشه‌بندی سنتی معمولا نیازمند مشخص بودن تعداد ثابت خوشه‌ها هستند. اما پیش‌بینی تعداد کل خوشه‌ها برای حجم زیاد و بلادرنگ داده‌های توئیتر، که تنوع زیادی از موضوعات را پوشش می‌دهند، دشوار است. روشهای تشخیص رویداد که در این بخش مورد بحث قرار می‌گیرند، به منظور

[26] New Event Detection (NED)
[27] Locality Sensitive Hashing
[28] Bucket

اجتناب از تعداد ثابت خوشه‌ها، بصورت ماهیتی افزایشی دنبال می‌شوند. در این بخش ابتدا چندین روش تشخیص رویداد مبتنی بر خوشه‌بندی افزایشی آورده شده است و سپس دو تکنیک NMF و ddCRP که برای خوشه‌بندی افزایشی در روش پیشنهادی مورد استفاده قرار گرفته‌اند معرفی خواهند شد.

در [۴۵] از یک الگوریتم خوشه‌بندی افزایشی جهت تشخیص رویدادها، در جریان داده‌های توییتر استفاده شده‌است. شباهت هر توییت با خوشه‌های موجود مقایسه می‌شود. اگر میزان شباهت توییت ورودی با هیچکدام از خوشه‌های موجود بر اساس یک حد آستانه برقرار نباشد، خوشه‌ای جدید حاوی آن توییت ایجاد می‌شود. در انتهای فرایند خوشه‌بندی، از یک الگوریتم SVM جهت شناسایی خوشه‌هایی که به رویدادهای جهان واقع اشاره دارند، استفاده شده‌است.

در[۴۶] روش خوشه‌بندی افزایشی مورد استفاده در [۴۵] را بر اساس اطلاعات معنایی موجود در هر توییت توسعه داده‌اند. نویسندگان از روش TwitterLDA [۴۷] برای تخصیص موضوع معنایی به هر توییت استفاده کرده‌اند. علاوه بر این بر اساس برچسبهای مبتنی بر هشتگ توانسته‌اند دقت و فراخوانی بهتری نسبت به روش [۴۵] بدست آورند. از روش خوشه‌بندی افزایشی در [۴۸] نیز بمنظور تشخیص رویداد مبتنی بر اولین وقوع استفاده شده است. هر توییت جدید که توسط کاربران به شبکه اجتماعی ارسال می‌شود، با تنها یک توییت از هر خوشه که معمولا اولین توییت ورودی به آن بوده است و همچنین k کلمه برتر هر خوشه مقایسه می‌شود و چنانچه شباهت از یک حد آستانه بیشتر باشد در آن خوشه درج می‌شود و اگر با هیچ خوشه‌ای مشابه نباشد، در خوشه‌ای جدید درج می‌گردد.

در [۴۹]، خوشه‌بندی افزایشی بر مبنای نهادهای نامدار موجود در توییتها انجام شده است. نهادهای نامدار شامل افراد، مکانها و سازمانها می‌باشند. به عبارت دیگر خوشه‌ها بر اساس نهادهای نامدار تشکیل می‌شوند. در پژوهش مذکور از تکنیک تشخیص بی‌نظمی برای شناسایی خوشه‌های مربوط به رویدادهای در حال وقوع استفاده شده است. یعنی خوشه‌هایی که تغییرات حجم آنها از حالت نرمال گذشته خارج

می‌شوند را بعنوان رویداد درنظر می‌گیرد. البته دقت این روش به شدت به سیستم تشخیص نهادهای نامدار بستگی دارد.

کاری که توسط TwitterNews+ [۵۰] برای تشخیص رویداد انجام می‌گیرد براساس یک عملیات دو گامی است. گام اول وظیفه شناسایی لغات پرتکرار یا انفجاری را برعهده دارد و گام دوم شامل خوشه‌بندی توییتهایی است که در مورد رویداد یکسانی بحث می‌کنند.گام اول بوسیله یک ماژول جستجو انجام می‌شود. ماژول جستجو از یک شاخص معکوس بهره می‌برد که هر مدخل آن شامل یک لغت و یک لیست متناهی از آخرین توئیتهایی است که لغت را در بر داشته‌اند و بطور منظم بروزرسانی می‌شوند. عملیات گام دوم توسط ماژولی به نام EventCluster انجام می‌شود. این ماژول مسئول تصمیم‌گیری سریع برای تخصیص توییت به یک خوشه با استفاده از روشی مبتنی بر خوشه‌بندی افرایشی است. یک مقایسه شباهت کسینوسی با مرکز این خوشه‌های رویداد انجام می‌شود تا خوشه با بیشترین انطباق بالاتر از یک حد آستانه مشخص پیدا شود. اگر چنین خوشه‌ای پیدا نشد، یک خوشه جدید ایجاد می‌شود و توئیت به آن تخصیص می‌یابد. از مزایای این روش، تشخیص رویدادهای خبری با دقت بالا و آنلاین می‌باشد.

تجزیه ماتریسی NMF و خوشه‌بندی افزایشی ddCRP

در مبحث جبر خطی، از تجزیه ماتریسی به عنوان ابزاری برای تحلیل داده استفاده می‌شود. هر تجزیه تعبیر مختلفی را از ساختار ضمنی داده‌ها ارائه می‌سازد که البته این تعابیر از نظر ریاضی هم ارز هستند. تجزیه مقدار تکین مانند SVD[۲۹][۵۱]، تحلیل مولفه اصلی و تجزیه نامنفی ماتریس (NMF) از جمله معروف‌ترین روشهای تجزیه ماتریس هستند. در الگوریتم NMF، ماتریس نامنفی $A \in \mathbb{R}^{m*n}$ به عنوان داده ورودی است. این الگوریتم دوماتریس نامنفی $W \in \mathbb{R}^{m*r}$ و $H \in \mathbb{R}^{r*n}$ را بدست می‌آورد، بطوری که تقریب $A \approx WH$ برقرار باشد. عدد r نسبت به کمترین مقدار اعداد m و n کوچکتر است. بطور

[29] Singular Value Decomposition

مثال در مبحث پردازش متن اگر ماتریس A به عنوان یک ماتریس سند×کلمه در نظر گرفته شود، خروجی الگوریتم NMF از نظر مفهومی دو ماتریس سند×موضوع و موضوع×کلمه می‌باشد، که در این پژوهش از ماتریس سند×موضوع به عنوان اولین گام خوشه‌بندی استفاده شده است. بر اساس ارزیابی‌های صورت گرفته در [۵۲] الگوریتم تجزیه ماتریس *NMF* در شناسایی موضوعات اسناد کوتاه، عملکردی بهتر نسبت به الگوریتم احتمالی *LDA*[۳۰] دارد.

الگوریتم ddCRP یک روش خوشه‌بندی افزایشی است که بر اساس تئوری احتمالات و چارچوب شبکه بیزین بنا نهاده شده است. الگوریتم ddCRP با بهره‌گیری از فرایند مدل سازی دریکله [۵۳] به لحاظ تئوری قادر به تولید خوشه‌هایی با تعداد نامتناهی است. این الگوریتم توسعه‌ای از الگوریتم CRP می‌باشد که خود یکی از توسعه‌های اصلی چارچوب شبکه بیزین محسوب می شود. فرایند رستوران چینی (CRP) ابزاری ساده و قدرتمند است. اما متاسفانه در متون آماری به خوبی توصیف نشده است. نام این فرایند از نحوه سرویس‌دهی به مشتریان در رستورانهای چینی گرفته شده است. در اغلب رستورانها، از جمله ایران مشتریان ابتدا پشت میز خالی می‌نشینند و سفارش غذای مورد علاقه‌اشان را می‌دهند. اما در رستورانهای چینی، غذاها از قبل بر روی میزها قرار دارند و مشتریان بنا بر علاقه‌اشان میزی را انتخاب می‌کنند. رستوران چینی که ابداع کنندگان این فرایند از آن الهام گرفته‌اند، در شهر سانفرانسیسکو قرار داشت که در ظاهر تعداد میز و صندلی نامحدود داشت. مشتریان در هنگام ورود به رستوران، میزی را برای غذا خوردن انتخاب می‌کردند. ملاک انتخاب آنها معمولا این بود که میزهایی که تعداد مشتری بیشتری دورش نشسته‌اند، احتمالا غذای بهتر و محبوبتری دارند. به عبارت دیگر، تعداد افراد پشت میزها در انتخاب شخص ورودی به رستوران موثر است، گرچه آن فرد می‌تواند یک میز بدون مشتری نیز انتخاب کند.

[30] Latent Drichlet Allocation

در معادل سازی عمل خوشه بندی داده‌ها با سرویس دهی رستورانهای چینی، هر خوشه معادل یک میز پذیرایی بی نهایت بزرگ می‌باشد. هر داده مانند یک مشتری است که وارد رستوران می‌شود و در یک میز می‌نشیند. در این تشابه سازی، فرض می‌شود که مشتریان در یک میز محبوب بنشینند، اما با این وجود، همواره یک احتمال غیر صفر نیز وجود دارد که مشتری جدید در میزی بنشیند که هنوز توسط کسی اشغال نشده است.

برای اینکه ببینیم چطور این کار امکانپذیر است، فرض کنید در حال حاضر N مشتری در رستوران هستند و z_i نشان دهنده میزی است که مشتری iام نشسته است. بنابراین یک بردار شامل تخصیص مشتریان به میزها داریم : $Z = (z_1, z_2, \dots, z_N)$. برای مثال اگر N=6 باشد، بردار می‌تواند $Z = (1,1,2,1,3,4)$ باشد، یعنی سه مشتری در میز اول، و در هر کدام از میزهای شماره دو، سه و چهار یک مشتری نشتسته‌اند. لازم به ذکر است، شماره میزها در روند انتخاب آنها هیچ تاثیری ندارند و تنها برای شاخص گذاری استفاده می‌شوند. از اینرو دو بردار $Z = (1,1,2,1,3,4)$ و $Z = (2,2,1,2,4,3)$ معادل یکدیگرند. اگر n_k نشان دهنده تعداد افراد نشسته در میز kام باشد و K تعداد کل میزهای غیر تهی را نشان دهد، آنگاه بردار شمارش $n = (n_1, n_2, \dots, n_K)$ نشان دهنده این است که پشت هر میز چه تعداد مشتری نشسته است. برای مثال قبل اگر $Z = (1,1,2,1,3,4)$ باشد آنگاه $n = (3,1,1,1)$ خواهد بود. بنابراین $\sum_{k=1}^{K} n_k = N$ خواهد بود.

اکنون با معرفی این نشانه گذاریها، توصیف CRP ساده است. اگر تاکنون N مشتری در رستوران نشسته باشند، آنگاه احتمال اینکه مشتری N+1 در میز k اام بنشیند، متناسب با محبوبیت آن میز است (رابطه ۲-۱)

$$P(z_{N+1} = k | n, \alpha) = \frac{n_k}{N + \alpha} \qquad 1 - 2$$

که α را پارامتر تراکم[31] می‌نامند و تنظیم کننده میزان فشـــردگی افراد پشـــت میزها اســـت. با تحلیل رابطه ۲-۱ متوجه می‌شـــویم که مقداری از احتمال کل گم شـــده اســـت! زیرا اگر مجموع رابطه یک را برای K میز محاســـبه کنیم، مقدار $\frac{N}{N+\propto}$ بدســـت خواهد آمد که از مقدار یک کمتر اســـت. دلیل آن این است که مقداری احتمال هم وجود دارد که مشتری N+1 تصمیم بگیرد تا در میز جدیدی بنشیند. اگر این میز جدید بعنوان میز K+1 شناخته شود، آنگاه احتمال نشستن پشت آن برابر رابطه ۲-۲ می‌شود:

$$P(z_{N+1} = K + 1|n, \alpha) = \frac{\alpha}{N + \alpha} \qquad 2-2$$

تجمیع دو رابطه ۲-۱ و ۲-۲ را می‌توان بصورت رابطه ۲-۳ نشان داد:

$$P(z_{N+1} = k|\alpha, n) = \begin{cases} \frac{n_k}{N + \alpha} & k \in K \\ \frac{\alpha}{N + \alpha} & k = K + 1 \end{cases} \qquad 3-2$$

یکی از جذابترین ویژگیهای CRP این است که امکان استفاده از یک توزیع ناپایدار[32] بر روی خوشه‌ها را فراهم می‌آورد بدون اینکه نیاز به مشخص بودن تعداد خوشه‌ها باشد. این ویژگی بسیار مفید است. برای نشان دادن اهمیت این ویژگی، مثال ساده‌ای آورده شده است. فرض کنید از یک جامعه یکصد نفری پرسیده می شود که کدام عدد بدشانسی می‌آورد. از این افراد ۵۰ نفر عدد ۱۳، ۴۰ نفر عدد ۴ و ۱۰ نفر هم عدد ۸۷ را انتخاب کرده اند. این پاسخ‌ها احتمالا شانسی و تصادفی بوجود نیامده اند. آنچه داریم سه گروه مجزا از افراد هستند. چگونه باید این داده‌ها را مدل کنیم؟ برای مدلسازی این مساله پیشنهادات زیر را مطرح می‌کنیم:

۱- تمام خوشه‌ها دارای احتمال یکسان باشند. این پیشنهاد به این معناست که ویژگی ناپایداری را حذف کنیم و یک توزیع یکنواخت بر روی خوشـــه‌ها ایجاد کنیم. در اینصـــورت تعداد روشـــهایی که

[31] concentration
[32] clumpy

می‌شود N شی را در بین خوشه‌ها تقسیم کرد برابر Nامین عدد بل[۳۳] می‌باشد. ایجاد این توزیع خیلی سخت نیست اما مشکل اصلی این است که نفر ۱۰۱ در پاسخ به سوال فوق عدد ۹۱ را با همان احتمال عدد ۱۳ انتخاب می‌کند. به عبارتی دیگر، اینکه افراد زیادی قبلا عدد ۱۳ را انتخاب کرده اند، اهمیتی ندارد. بنابراین ناپایداری توزیع، ضروری است. ناپایداری بیان می‌دارد که مشاهدات جدید به احتمال زیادتر به مشاهدات اخیر مشابه اند.

۲- تعداد خوشه‌ها ثابت باشند. اگر تعداد ثابتی خوشه داشته باشیم و از یک الگوریتم مثل Kmeans استفاده کنیم، چه اتفاقی می‌افتد؟ در این حالت، علاوه بر اینکه انتخاب عدد K خود یک چالش است، مشکل این پیشنهاد دقیقا عکس مشکل پیشنهاد اول است! به طور خاص، برای وضعیت غیر محتمل اما ممکنی که شرکت کننده شماره ۱۰۱ واقعا بخواهد ۹۱ را انتخاب کند، راه حلی وجود ندارد، چون تعداد خوشه‌ها ثابت است و اماکن تعریف یک خوشه جدید وجود ندارد.

۳- داشتن یک دانش قبلی در مورد تعداد خوشه‌ها: در این حالت برای مقدار K یکسری توزیع‌های احتمالاتی تعیین می‌شود و سپس فرض می‌شود که تمام خوشه‌هایی که N عنصر را در خود جای داده اند، دارای احتمال یکسان هستند. به این تزتیب برای شرکت کننده شماره ۱۰۱ امکان انتخاب عدد ۹۱ بعنوان یک عدد غیر محتمل وجود دارد. این در حالی است که با تعریف یک پیشین بر روی K، آن شخص با احتمال بیشتر عدد ۱۳، ۴۰ یا ۸۷ را خواهد گفت. این روش نیز یک مشکل دارد: شرکت کننده ۱۰۱ احتمالا دقیقا ۱۳ را انتخاب کند، حتر اگر ۸۷ متداول تر باشد.

۴- مدل سازی احتمالات همراه هر گروه: یک ایده خلاقانه می‌تواند تعیین یک توزیع پیشین بر روی K و تخصیص احتمال θ_k به هر مشاهده که می‌تواند در خوشه K قرار بگیرد باشد. پس می‌توان یک پیشین نیز بر روی مقادیر θ_k تعیین کرد و آنها را یکپارچه کرد. این بدین معنی است که مشاهدات جدید با احتمال بیشتر به خوشه‌هایی که قبلا تعداد بیشتری شی به آنها تخصیص یافته

[33] Bell Number

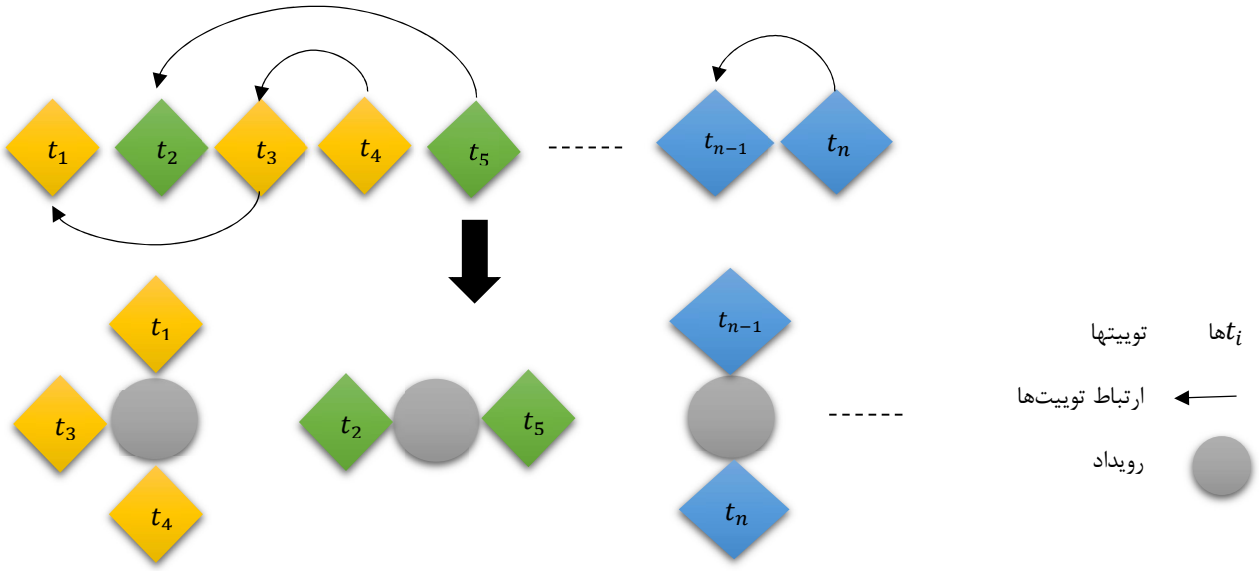
نسبت داده می‌شوند. علاوه بر این، امکان افزایش مقدار K هنگام مشاهدات جدید نیز وجود دارد. بدین ترتیب تمام مشکلات پیشنهادات قبلی حل می‌شود.

نکته پشت این بحث این است که CRP روشی برای تعیین یک توزیع پیشین بر روی خوشه‌هاست که دو پیش نیاز دارد: (۱) تعداد خوشه K می‌تواند همانطور که تعداد مشاهدات N افزایش می‌یابد، زیاد شوند و (۲) مشاهدات جدی با احتمال زیاد به خوشه‌های محبوب مشاهدات قبلی تخصیص خواهند یافت.

الگوریتم ddCRP به جای وابسته نمودن احتمال به تعداد افراد دور میزها، آن را به میزان شباهت بین افراد وابسته می‌کند (رابطه ۲-۴) . شکل ۲-۱ نشان‌دهنده نحوه عملکرد ddCRP است.

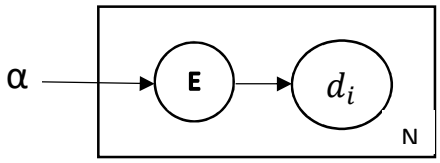
$$P(c_{N+1} = k|\alpha, S) = \begin{cases} s_{N+1,c_k} & k \in K \\ \alpha & k = K+1 \\ \hline \alpha + N & \end{cases} \qquad 4-2$$

در رابطه ۲-۴، s_{N+1,c_k} میزان شباهت بین مشتری جدید c_{N+1} و مشتریان اطراف میز k است و S مجموعه تمام مقادیر شباهت بین مشتریان می‌باشد. در واقع در مسئله خوشه‌بندی اسناد متنی، هر سند بعنوان مشتری و خوشه‌ها به عنوان میزها در نظر گرفته می‌شوند.



شکل ۲-۱) خوشه‌بندی افزایشی ddCRP

با توجه به شـکل ۲-۲ هر توییت همچون d_i، معلول یک فاکتور پنهان (رویداد) همچون E اسـت و با توجه به اینکه سـهم توییتهای هر رویداد متفاوت اسـت و تعداد رویدادها مشـخص نیسـت، از توزیع پیشین دریکله برای بیان توزیع این تابع احتمال استفاده می‌شود.



شکل ۲-۲) مدل شبکه بیزی استفاده شده

همانند فرایند رسـتوران چینی، الگوریتم ddCRP دارای پارامتر α اسـت که برای تنظیم میزان تراکم در توزیع پیشین دریکله استفاده می‌شود. بطوری که با افزایش α، فرض می‌شود که اسناد از موضوعات بیشتری تشکیل شده اند. در واقع با افزایش این پارامتر، موضوعات به ازای هر سند خاصتر و جزئی تر می‌گردند.

در این پژوهش به منظور محاسبه شباهت بین دو توییت از شباهت کسینوسی[۳۴] [۵۴] استفاده شده است. برای این منظور شباهت بین دو توییت t_i و t_j مطابق رابطه ۲-۵ محاسبه می‌گردد.

$$s(t_i, t_j) = \frac{t_i . t_j}{\|t_i\| \|t_j\|} \qquad 5-2$$

با ورود توییتهای جدید، تابع احتمال ddCRP که مبتنی بر مقادیر شباهت بین توییتها و پارامتر α است، برای توییت ورودی اجرا می‌شود. با توجه به رابطه ۲-۴، در صورت انتخاب مقدار احتمال از بین بردار شباهت، توییت d به توییت انتخاب شده متصل می‌گردد. در ابنصورت توییت جدید به یکی از خوشه‌های موجود ملحق می‌شود. اما در صورت انتخاب پارامتر تراکم α، خوشه جدید که حاوی توییت d است، تشکیل می‌گردد.

[34] Cosine Similarity

۲-۴) کارهای انجام شده در زمینه پیش‌بینی رویداد

تاکنون پژوهش‌های متعددی در زمینه پیش‌بینی رویدادهای خاص در شبکه‌های اجتماعی انجام شده است. پیش‌بینی نتایج انتخابات، پیش‌بینی نحوه انتشار بیماری، پیش‌بینی بازار سهام، پیش‌بینی میزان فروش محصولی خاص، پیش‌بینی تعداد ارجاعات به یک مقاله و پیش‌بینی نتایج مسابقات ورزشی از جمله این موارد ه ستند. در ادامه، ابتدا مهمترین رو شهای پیش‌بینی در حوزه یادگیری ما شین آورده شده است و سپس چندین کار انجام شده در زمینه پیش‌بینی رویدادهای مختلف، و پس از آن در یک بخش مجرا به پژوهشهای مرتبط با پیش‌بینی انتخابات اشاره شده است.

۲-۴-۱) پیش‌بینی در حوزه یادگیری ماشین

در یادگیری ماشین، دو هدف پیش‌بینی و خوشه‌بندی دنبال می‌شود. در این بخش به پیش‌بینی پرداخته خواهد شد. پیش‌بینی فرآیندی است که طی آن، با بهره‌گیری از یک مجموعه متغیرهای ورودی، ارزش یک متغیر خروجی تخمین زده می‌شود. برای مثال، با استفاده از مجموعه‌ای از مشخصه‌های یک خانه، می‌توان قیمت آن را تخمین زد. مسائل پیش‌بینی به دو دسته کلی رگرسیون و دسته‌بندی تقسیم می‌شوند. در مسائل رگرسیون، متغیری که باید پیش‌بینی شود، عددی است (برای مثال قیمت یک خانه). اما در مسائل دسته‌بندی، متغیرها برای پیش‌بینی یکی یا برخی از دسته‌های از پیش تعریف شده مورد استفاده قرار می‌گیرند. این دسته‌ها می‌توانند به سادگی «بلی» یا «خیر» باشند (برای مثال، پیش‌بینی اینکه آیا یک قطعه مکانیکی دچار نقص فنی خواهد شد یا خیر). بطور کلی الگوریتم‌های پیش‌بینی در سه دسته مدل‌های خطی، مدل‌های مبتنی بر درخت و شبکه‌های عصبی قرار می‌گیرند.

مدل‌های خطی

یک مدل خطی از فرمولی ساده برای پیدا کردن خط دارای بهترین برازش در میان یک مجموعه از نقاط داده استفاده می‌کند. پیشینه روش‌های خطی به بالغ بر ۲۰۰ سال پیش باز می‌گردد و البته این روش‌ها

در حال حاضر به طور گسترده‌ای در آمار و یادگیری ماشین مورد استفاده قرار می‌گیرند. مدل‌های خطی به دلیل سادگی برای کارهای آماری بسیار مفید هستند و در آن‌ها متغیری که قصد پیش‌بینی آن وجود دارد تحت عنوان متغیر وابسته، به عنوان معادله‌ای از متغیرهای شناخته شده یا متغیرهای مستقل ارائه می‌شود و بنابراین پیش‌بینی نهایی تنها به متغیرهای ورودی مستقل و داشتن معادله‌ای که پاسخ را به دست دهد بستگی دارد. برای مثال، فرد ممکن است بخواهد بداند که پختن یک کیک چقدر زمان می‌برد و با بهره‌گیری از تحلیل‌های رگرسیون معادله ای مانند رابطه ۲-۱ را تولید کند:

$$t = 0.5x + 0.25y \qquad 6 - 2$$

که در رابطه ۲-۶، t زمان پخت کیک به ساعت، x وزن خمیر کیک به کیلوگرم، و y متغیری است که اگر کیک شکلاتی باشد برابر با یک و اگر غیر شکلاتی باشد برابر با ۰ است. اگر فرد یک کیلوگرم خمیر کیک شکلاتی داشته باشد، با قرار دادن مقادیر متغیرها در معادله بالا می‌تواند مقدار زمان پخت کیک یعنی t را به دست آورد که برای این مثال به صورت $t = (0.5 \times 1) + (0.25 \times 1) = 0.75$ ساعت یا ۴۵ دقیقه است.

رگرسیون خطی یا به طور مشخص‌تر رگرسیون کم‌ترین مربعات، استاندارترین شکل یک مدل خطی است. برای مساله رگرسیون، رگرسیون خطی ساده‌ترین و قابل فهم‌ترین نوع مدل‌های خطی محسوب می‌شود. اشکال این روش آن است که گرایش به بیش‌برازش کردن مدل دارد. بیش‌برازش به بیان ساده یعنی مدل بیش از اندازه با داده‌های برچسب‌دار آموزش داده شود و خروجی‌هایی عینا مثل داده‌های آموزش به دست دهد که مانعی در راه عمومی‌سازی مدل محسوب می‌شوند. به همین دلیل، رگرسیون خطی (همراه با رگرسیون لوجستیک که در ادامه به آن پرداخته خواهد شد) در یادگیری ماشین اغلب تنظیم[۳۵] می‌شوند، بدین معنا که مدل در صورت انجام بیش‌برازش دارای جریمه مشخصی خواهد شد.

[35] regularized

دیگر اشکال مدل رگرسیون خطی سادگی بسیار زیاد آن است که موجب می‌شود هنگام پیش‌بینی رفتارهای پیچیده‌تر که متغیرهای ورودی در آن‌ها مستقل نیستند دچار مشکل می‌شود.

رگرسیون لوجستیک تطبیق دادن رگرسیون خطی برای مسائل دسته‌بندی است. مشکلات رگرسیون لوجستیک درست مانند رگرسیون خطی است با این تفاوت که این روش گرایش زیادی نیز به بیش برازش دارد. از آنجا که این رگرسیون مقادیر بین ۰ و ۱ را نگاشت می‌کند، راهکار مناسبی برای مسائل دسته‌بندی است، زیرا احتمال وجود داشتن نقاط داده در هر کلاس را به دست می‌دهد.

مدل‌های مبتنی بر درخت

درخت تصمیم گرافی است که از روش انشعابی برای نمایش کلیه خروجی‌های ممکن برای یک تصمیم استفاده می‌کند. برای مثال، فرد برای سفارش یک سالاد، ابتدا نوع کاهو، سپس چاشنی و در نهایت سس روی آن را انتخاب می‌کند. می‌توان همه خروجی‌های ممکن را در درخت تصمیم ارائه کرد. در یادگیری ماشین، انشعاب‌های استفاده شده پاسخ‌های دودویی بله/خیر هستند.

برای آموزش دادن درخت تصمیم، مجموعه داده آموزش از ورودی دریافت می‌شود (که مجموعه داده‌ای است که برای آموزش دادن مدل مورد استفاده قرار می‌گیرد) و متغیرهایی که مجموعه داده آموزش را به بهترین شکل منشعب می‌کنند با توجه به هدف انتخاب می‌شوند. برای مثال، در تشخیص کلاهبرداری، می‌توان ویژگی(هایی) را پیدا کرد که به بهترین شکل خطر کلاهبرداری در یک کشور را پیش‌بینی می‌کند. پس از اولین انشعاب، دو زیرمجموعه حاصل می‌شوند، سپس منشعب کردن بر اساس دومین ویژگی برای هر زیر مجموعه تکرار می‌شود و این کار تا هنگامی که از تعداد کافی از متغیرها برای ارضای نیازهای کاربر استفاده شده باشد ادامه پیدا می‌کند. درخت‌های تصمیم در صورت مشخص بودن اولین ویژگی در انجام پیش‌بینی عالی هستند.

یک جنگل تصادفی میانگینی از تعداد بسیار زیادی درخت تصمیم است که هر یک با نمونه‌ای تصادفی از داده‌ها آموزش داده شده‌اند. هر درخت تنها در کل جنگل از یک درخت تصمیم کامل ضعیف‌تر است، اما با کنار هم قرار گرفتن همه آن‌ها، به لطف تنوع خروجی‌ها کارایی کلی بهتری حاصل می‌شود. جنگل تصادفی امروزه یکی از محبوب‌ترین الگوریتم‌های یادگیری ماشین محسوب می‌شود. آموزش دادن این مدل بسیار آسان و کارایی آن خوب محسوب می‌شود. مشکل این الگوریتم آن است که برای ارائه پیش‌بینی‌های خروجی نسبت به دیگر الگوریتم‌ها کندتر عمل می‌کند، بنابراین ممکن است هنگام نیاز به پیش‌بینی‌های خیلی سریع از این روش استفاده نشود.

روش گرادیان بوستینگ نیز مانند جنگل تصادفی از درخت‌های تصمیم ضعیف استفاده می‌کند. تفاوت بزرگ این دو روش آن است که در روش گرادیان بوستینگ درخت‌ها یکی پس از دیگری آموزش داده می‌شوند. هر درخت زیرمجموعه در درجه اول با داده‌هایی که به اشتباه توسط درخت پیشین پیش‌بینی شده‌اند آموزش داده می‌شوند. این امر موجب می‌شود مدل کمتر بر مسائلی که پیش‌بینی در آن‌ها آسان است و بیشتر روی موارد پیچیده متمرکز شود.

شبکه‌های عصبی

شبکه عصبی به پدیده بیولوژیکی مربوط می‌شود که شامل نورون‌های متصلی است که به مبادله پیام با یکدیگر می‌پردازند. این ایده در حال حاضر در جامعه یادگیری ماشین پذیرفته شده و به آن شبکه‌های عصبی مصنوعی گفته می‌شود. یادگیری عمیق که یکی از مباحث داغ روز است را می‌توان با چندین لایه از شبکه‌های عصبی روی هم قرار گرفته پیاده‌سازی کرد.

شبکه‌های عصبی مصنوعی خانواده‌ای از مدل‌ها هستند که آموخته‌اند مهارت‌های شناختی را کسب کنند. هیچ الگوریتم دیگری نمی‌تواند وظیفه فوق‌العاده پیچیده‌ای همچون بازشناسی تصویر را به خوبی

شبکه‌های عصبی انجام دهد. اگرچه، درست مانند مغز انسان، زمان زیادی لازم است تا این مدل آموزش ببیند و نیازمند توان محاسباتی زیادی نیز هست

۲-۴-۲) پیش‌بینی رویدادهای مختلف

در این بخش تعدادی از پژوهشهای انجام شده در حوزه پیش‌بینی رویداد آورده شده است.

در [۵۵] سیستمی جهت پیش‌بینی انتشار کووید ۱۹ بر مبنای تشخیص و ردیابی رویدادها در توییتر پیشنهاد شده است. این سیستم ابتدا گرافهای دانش مبتنی بر جریانات توییتری مربوط به کووید ۱۹ ایجاد می‌کند. به عبارت دیگر، بر اساس اطلاعات و اخبار جدیدی که در مورد شناسایی موارد جدید ابتلا در موقعیتهای مکانی مختلف در شبکه انتشار می‌یابد گراف مورد نظر بروز می‌شود و سپس بر اساس آن پیش‌بینی نحوه انتشار بیماری انجام می‌گیرد. در [۵۶] مدل جدیدی بر مبنای تحلیل شبکه اجتماعی توییتر به نام tweetluenza بمنظور پیش‌بینی تعداد مراجعات افراد مبتلا به آنفولانزا به بیمارستانهای کشور امارات متحده عربی معرفی شده است. مدل مذکور ابتدا بر اساس یکسری کلمات کلیدی، توییتهای مرتبط با آنفولانزا را شناسایی می‌کند و سپس توییتهایی که محتوی گزارش وقوع انتشار جدید بیماری می‌باشند را تشخیص می‌دهد. سپس با استفاده از این نوع توییتها و یک مدل رگرسیون خطی به پیش‌بینی تعداد مراجعات بیمارستانی می‌پردازد. همبستگی بالا بین خروجی مدل و داده‌های واقعی بیمارستان، عملی بودن مدل مذکور را نشان می‌دهد.

نویسندگان در [۵۷] با استفاده از مدل رگرسیون بردار پشتیبان حداقل مربعات (LSSVR[۳۶]) به پیش‌بینی میزان فروش فیلم‌ها پرداخته‌اند. آنها مدل مذکور را بر روی سه نوع داده مختلف شامل داده‌های پایگاه اطلاعاتی فیلم‌ها همانند Box Office و IMDB[۳۷]، داده‌های توییتری در مورد فیلمها و داده‌های ترکیبی اعمال کردند. نتایج حاصله نشان می‌دهند که اعمال مدل بر روی داده‌های ترکیبی

[36] Least Square Support Vector Regression
[37] Internet Movie DataBase

نتایج دقیقتری نسبت به سایر داده‌ها بدست می‌آورد. [۵۸]و [۵۹] نیز روشهایی می‌باشند که با استفاده از تحلیل احساسات بعد از انتشار یک فیلم، به پیش‌بینی میزان درآمد و موفقیت آن پرداخته‌اند.

در [۶۰] با تحلیل وبلاگها، یک همبستگی مثبت بین حجم پست‌های کاربران در مورد یک خواننده و آلبوم خاص و فروش آن آلبوم موسیقی یافتند. نویسندگان در [۶۱] با تحلیل توئیتر نیز به چنین همبستگی بین پیامهای توئیتر و فروش آلبوم موسیقی دست یافتند. آنها از پیامهایی که محتوی نام خواننده و یا نام آلبوم بود در بازه زمانی دو هفته قبل از آغاز فروش تا یک هفته بعد از آن استفاده نمودند.

۲-۴-۳) پیش‌بینی انتخابات

در سالهای اخیر چندین پژوهش در مورد پیش‌بینی نتایج انتخابات بر مبنای تحلیل شبکه‌های اجتماعی انجام شده است که از صفتهای مختلفی چون فراوانی نام نامزدهای انتخاباتی و یا نام احزاب سیاسی استفاده کرده‌اند.

برای اولین بار در [۶۲]، پیش‌بینی انتخابات ۲۰۰۹ آلمان با استفاده از تحلیل داده‌های توییتر انجام شده است. نویسندگان بر اساس تعداد توییت‌های مرتبط با هر کدام از احزاب سیاسی به پیش‌بینی نتایج انتخابات پرداختند. همانطور که دراین رساله نیز نشان داده خواهد شد، مقایسه تعداد توییتهای طرفداران یک حزب نسبت به حزب دیگر نمی‌تواند معیار مناسبی برای پیش‌بینی نتیجه انتخابات باشد. چون ممکن است طرفداران یک نامزد خاص در شبکه‌های اجتماعی فعالیت بیشتری داشته باشند.

در نظرگرفتن احساسات توییتهای افراد، می‌تواند دقت پیش‌بینی نتایج انتخابات را افزایش دهد [۶۳]. هرچند در این رساله نشان داده خواهد شد که تنها درنظر گرفتن تعداد توییتهای احساسی، نمی‌تواند نتیجه انتخابات را بدرستی تخمین بزند. Burnap و همکاران [۶۴] مدلی بر مبنای تحلیل احساسات

برای پیش‌بینی نتیجه انتخابات ۲۰۱۵ انگلستان پیشنهاد دادند. آنها بر اساس روش تحلیل احساسات که در [۶۵] ارائه شد، به هر توییت یک نمره بین ۵- تا ۵+ به ترتیب به معنای احساس منفی شدید تا احساس مثبت شدید اختصاص دادند و بر اساس مجموع نمرات احساس برای هر حزب؛ به پیش‌بینی تعداد کرسی احزاب در پارلمان انگلستان پرداختند. نویسندگان در [۶۶] بر اساس یک روش تحلیل احساسات مبتنی بر واژه[۳۸] به پیش‌بینی نتیجه انتخابات ریاست جمهوری آمریکا پرداختند. آنها ابتدا بر اساس اطلاعات جغرافیایی موجود در هر توییت، ایالتی که توییت از آنجا ارسال شده است را استخراج نمودند. سپس برای هر ایالت، تعداد توییتهای احساسی مثبت و منفی هر دو حزب را شمردند و هر حزب که تعداد توییت بیشتری داشت را بعنوان برنده آن ایالت مشخص نمودند. نهایتا بر اساس شمارش تعداد ایالتها، برنده نهایی را پیش‌بینی کردند. البته داده‌هایی که آنها از توییتر جمع‌آوری کردند تنها به یک هفته منتهی به انتخابات محدود می‌شد.

اتان و همکاران [۶۷]، مجموعه داده‌ای شامل ۲۶۰۰۰۰ توییت در مورد انتخابات ریاست جمهوری آمریکا در سال ۲۰۲۰ را گردآوری کردند. آنها با بررسی تعداد توییتهای احساسی مثبت و منفی، نزدیک بودن نتایج انتخابات را به درستی پیش‌بینی کردند. علاوه بر این نشان دادند که تحلیل احساسات با استفاده از داده‌های رسانه‌های اجتماعی، یک روش کم هزینه و دقیق برای کسب عقاید کلی نسبت به نامزدهای انتخاباتی و پیش‌بینی نتایج آن می‌باشد. در [۶۸] نیز یک روش مبتنی بر چندین ویژگی مختلف شبکه اجتماعی توییتر برای پیش‌بینی نتیجه انتخابات ریاست جمهوری امریکا در سال ۲۰۲۰ پیشنهاد شد. نویسندگان از ویژگیهای متعددی همچون تعداد توییتهای مثبت و منفی، تعداد افراد، تعداد لایک‌ها و بازتوییتها بر روی یک پایگاه‌داده مشتمل بر ۱۰ میلیون توییت استفاده کردند. روشهای دیگری نیز مانند [۷۱-۶۹]، نتایج انتخابات را با استفاده از تحلیل احساسات پیش‌بینی کرده‌اند.

[38] Lexicon

در نظر گرفتن عواملی خارج از شـــبکه اجتماعی مجازی در کنار تحلیل‌های شـــبکه اجتماعی نیز برای پیش‌بینی نتیجه انتخابات اســـتفاده شـــده اســـت. Ruowei و همکاران [۷۲] یک مدل ترکیبی برای پیش‌بینی نتیجه انتخابات پیشـنهاد دادند. آنها روش مبتنی بر تحلیل احسـاسـات را با مدلهای سـنتی علوم ســـیاســـی ترکیب کردند و نتایج انتخابات را بصـــورت محلی در ایالت جورجیا پیش‌بینی نمودند. منظور از مدلهای سـنتی علوم سـیاسـی در نظر گرفتن مواردی همچون نرخ رشـد اقتصـادی یا میزان کاهش نرخ بیکاری است که در دولت جاری اتفاق افتاده است، و یا بصورت وعده‌های انتخاباتی توسط نامزدها مطرح می‌گردد. به عبارت دیگر آنها تحلیل احســـاســـات در شـــبکه اجتماعی را در کنار شاخص‌های رشد و توسعه اقتصادی سیاسی کشور مورد استفاده قرار دادند.

سه روشی که در ادامه معرفی شده‌اند، از جهاتی بیشترین شباهت را به روش پیشنهادی این رساله برای پیش‌بینی نتیجه انتخابات دارند. این روشها برای پیش‌بینی انتخابات، شاخصی معرفی کرده‌اند که بر مبنای تحلیل احســـاســـات به پیش‌بینی انتخابات می‌پردازد. ســـینگ و همکاران [۷۳]، از داده‌های توییتر برای پیش‌بینی انتخابات مجلس پنجاب (ایالتی از هند) در ســـال ۲۰۱۷ اســـتفاده کردند. آنها رابطه ۲-۴ را بعنوان یک اندیکاتور برای پیش‌بینی تعداد کرسـی‌های کسـب شـده برای هر حزب یا به عباراتی دیگر پیش‌بینی برنده انتخابات مورد استفاده قرار دادند.

$$SS(A) = \frac{pos(A) - neg(A)}{T(A) + T(B)} \qquad 4-2$$

در رابطه ۲-۴، *SS(A)* امتیاز احساسی حزب *A* است، *pos(A)* و *neg(A)* به ترتیب تعداد کل توییت‌های مثبت و منفی برای حزب *A* و *T(A)* و *T(B)* تعداد کل توییت‌های مربوط به احزاب *A* و *B* می‌باشـند. هر حزبی که مقدار اندیکاتور بیشتری را داشته باشد برنده انتخابات خواهد بود.

در [۷۴]، میزان محبوبیت نامزدهای انتخاباتی، بر اساس رابطه ۲-۵ که مبتنی بر تحلیل احساسات می‌باشد، محاسبه شده است. روش پیشنهادی مقاله مذکور برای پیش‌بینی نتیجه انتخابات ریاست جمهوری فرانسه در سال ۲۰۱۷ استفاده شده است.

$$P(A) = \left[\frac{pos(A)}{pos(A) + neg(A)}\right]\left[\frac{T(A)}{T(A) + T(B)}\right] \qquad 5-2$$

P(A) در رابطه ۴-۴ میزان محبوبیت حزب *A* است.

در [۷۵]، روشـــی مبتنی بر تحلیل احســـاســـات و رابطه ۲-۶ برای پیش‌بینی نتیجه انتخابات ۲۰۱۶ ریاست جمهوری آمریکا پیشنهاد شده است. بر اساس این رابطه میزان موفقیت هر حزب در انتخابات محاسبه می‌شود و حزبی که نمره بیشتری کسب کرده باشد، پیروز انتخابات خواهد بود.

$$SR(A) = \frac{pos(A) + neg(B)}{T(A) + T(B)} \qquad 6-2$$

SR (A) میزان موفقیت حزب *A* است.

۲-۵ جمع بندی

در این فصل ابتدا انواع روشهای تشخیص رویداد در شبکه‌های اجتماعی مجازی همچون خوشه‌بندی افزایشی، تشخیص وقوع بی‌نظمی، تشخیص اولین وقوع و روشهای مبتنی بر مدلبندی موضوعی مطرح گردید. سپس کارهای انجام شده در زمینه پیش‌بینی رویداد و پیش‌بینی نتیجه انتخابات معرفی شدند. با توجه به مرور روشهای مشابه قبلی و ماهیت داده‌های شبکه اجتماعی، استفاده از تکنیک خوشه‌بندی افزایشی انتخاب موثری خواهد بود. در زمینه پیش‌بینی نتیجه انتخابات نیز، تحلیل‌های مبتنی بر احساسات می‌تواند نتایج با دقت بهتری ارائه دهند.

۳ روش پیشنهادی

۳-۱ مقدمه

در این فصل روش پیشنهادی برای پیش‌بینی رویداد با تحلیل شبکه اجتماعی توییتر تشریح شده است. همانطور که قبلا بیان شد، به منظور انجام یک پیش‌بینی با دقت مناسب، باید پارامترهای موثر در پیش‌بینی شناسایی و مورد نظارت و ارزیابی قرار گیرد. ویژگی که در این پژوهش برای پیش‌بینی رویداد مورد استفاده قرار گرفته است، تغییرات نرخ ارسال پیام‌ها به شبکه اجتماعی می‌باشد. در ادامه، ابتدا روش پیشنهادی برای پیش‌بینی رویدادها آورده شده است. لازم به ذکر است که این روش، خاص منظوره نیست. این بدین معنی است که، تنها برای پیش‌بینی یک رویداد خاص، طراحی نشده است، بلکه قادر است تا رویدادهای مختلفی را پیش‌بینی کند. این توانایی یکی از نوآوریها و نقاط قوت روش پیشنهادی می‌باشد. نهایتا در بخش سوم این فصل، از همین ویژگی تغییرات نرخ ارسال پیامها، برای پیش‌بینی نتیجه انتخابات استفاده شده است، تا مسئله پیش‌بینی به کمک تحلیل شبکه اجتماعی توییتر، برای یک رویداد خاص نیز بررسی شود. در این روش نیز که مبتنی بر تحلیل احساسات می‌باشد، بر اساس تغییرات نرخ ارسال پیامها، یک اندیکاتور معرفی می‌شود که بر اساس مقدار آن برنده انتخابات پیش‌بینی می‌گردد.

۳-۲ روش پیشنهادی برای پیش‌بینی رویداد

در این بخش ابتدا نشان داده خواهد شد که قبل از وقوع رویدادهای قابل پیش‌بینی، توییتهایی مرتبط با آن رویداد به شبکه ارسال می‌شوند. سپس روش پیشنهادی برای پیش‌بینی رویدادها شرح داده می‌شود.

۳-۲-۱ رفتار داده‌های شبکه اجتماعی قبل از وقوع رویداد

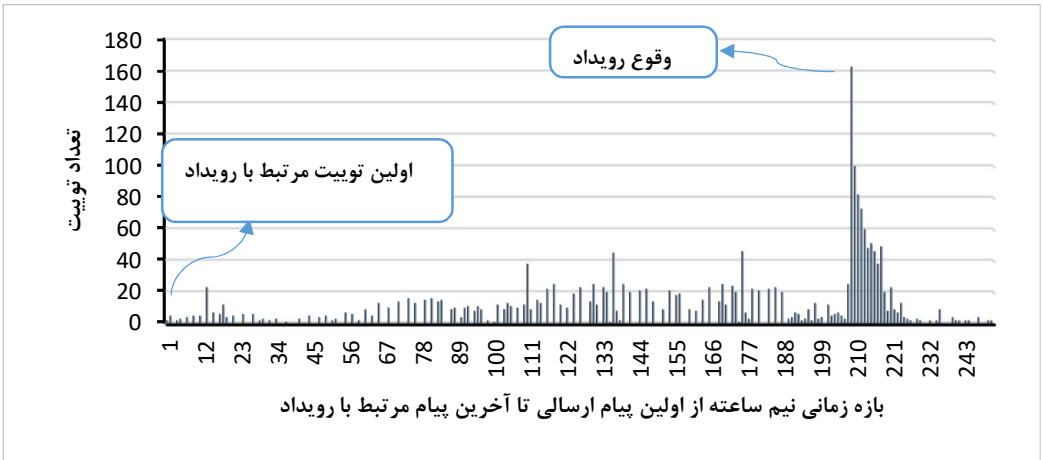
در دنیای واقعی، معمولا قبل از وقوع یک رویداد، جریانات، گفتگوها و تحرکات متعددی در مورد آن رویداد به وقوع می‌پیوندد. بعنوان مثال، ماه‌ها قبل از شروع انتخابات، صحبتها در مورد آن شروع می شود. قبل از وقوع سیل، هشدارها و پیامهای متعددی در مورد وضعیت آب و هوا صادر می‌گردد.

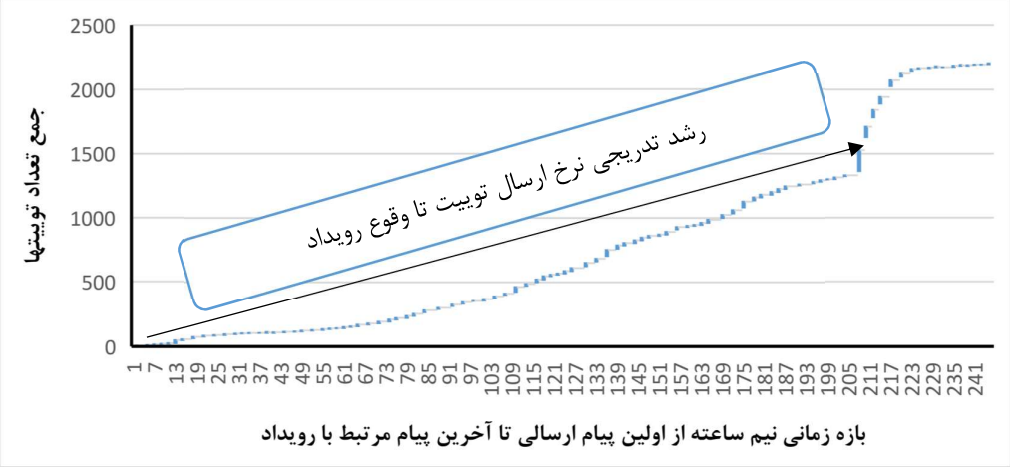
قبل از آغاز اغتشاشات خیابانی، پیامها و جلسات هماهنگی در مورد آن به راه می‌افتد. وقایعی همچون جام‌جهانی و کنسرت نیز نمونه‌هایی از همین موارد هستند. بنابراین قبل از وقوع رویدادهای قابل پیش‌بینی، نشانه‌هایی از وقوع آنها در آینده، در جامعه نمایان می‌شود. طبیعتا برخی رویدادها نیز که قابل پیش‌بینی نیستند، همانند سقوط هواپیما یا زلزله، قبل از وقوع نمی‌توان نشانه‌هایی از آنها در جامعه یافت و در واقع بلافاصله بعد از وقوع مطرح می‌شوند.

وجه مشترک دیگری که بین رویدادهای قابل پیش‌بینی می‌توان دید، این است که از لحظه بروز اولین نشانه‌های وقوع رویداد، هر چه زمان وقوع آن نزدیکتر می شود، بروز نشانه‌ها سرعت بیشتری می‌گیرد تا اینکه در حین وقوع و تا حدودی بعد از آن به اوج خودش می‌رسد و سپس این نرخ، یک روند نزولی می‌گیرد تا تقریبا به طور کامل از بین برود. این اوج‌گیری و تنزل نشانه‌های وقوع رویداد، هنگام وقوع رویدادهای جهان واقع قابل م شاهده ه ستند. در این بخش هدف این ا ست که وجود چنین رفتارهای مشابهی در شبکه اجتماعی توییتر نیز مورد بررسی قرار گیرد.

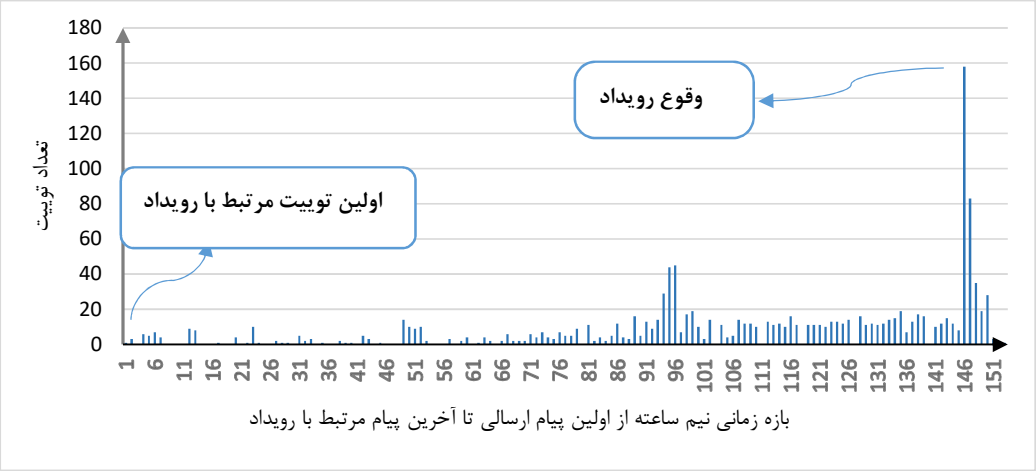
در شکل ۳-۱ به عنوان نمونه، نمودار مربوط به تعداد توییتهای ارسالی مرتبط با دو رویداد مختلف نشان داده شده‌است. این رویدادها از مجموعه رویدادهای قابل پیش‌بینی انتخاب شده‌اند. اولین رویداد، به رسمیت شناختن ازدواج همجنسگرایان در یوتا یک از ایالات غربی آمریکا است که در تاریخ ۲۰ دسامبر ۲۰۱۳ روی داده است، در حالی که اولین توییتهای مربوط به این رویداد در تاریخ ۱۶ دسامبر ۲۰۱۳ به شبکه ارسال شده است (شکل ۳-۱ الف). رویداد دوم مربوط به گفتگوهای صلح بین فلسطین و ا سرائیل ا ست که در تاریخ ۲۳ آوریل ۲۰۱۴ به نتیجه ر سید و اولین نشانه‌های مربوط به این رویداد چهار روز قبل از آن، توییت شده است (شکل ۳-۱ ب). محور افقی در نمودارها، بازه‌های زمانی نیم ساعته از اولین توییت تا آخرین توییت مربوط به رویداد مورد نظر را نشان می‌دهد. بازه با حداکثر تعداد توییت ارسالی نیز زمان وقوع رویداد را نشان می‌دهد. همانطور که در شکل ۳-۱ قابل مشاهده‌است، رفتار مورد انتظار در این نمودارها با شدت‌های مختلفی دیده می‌شود. یعنی قبل از اینکه

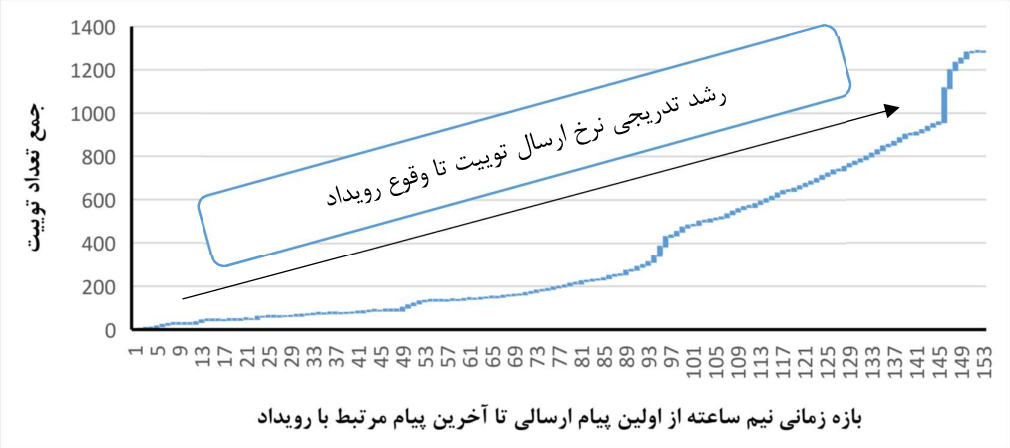
رویداد اتفاق بیافتد، پیامهایی در مورد آن از روزها قبل در شبکه منتشر شده است. مدت زمان بین اولین نشانه هر رویداد و آخرین نشانه آن و تعداد توییتهای مرتبط با آن، به نوع رویداد بستگی دارد. برخی رویدادها اثرات اجتماعی کمتر و یا بیشتری می‌گذارند، بنابراین تغییرات تعداد پیامهای ارسالی یا به عبارتی دیگر، تغییرات نرخ ارسال پیام‌ها، روند متفاوتی خواهد داشت.





(الف) رویداد به رسمیت شناختن ازدواج همجنسگرایان در ایالت یوتا

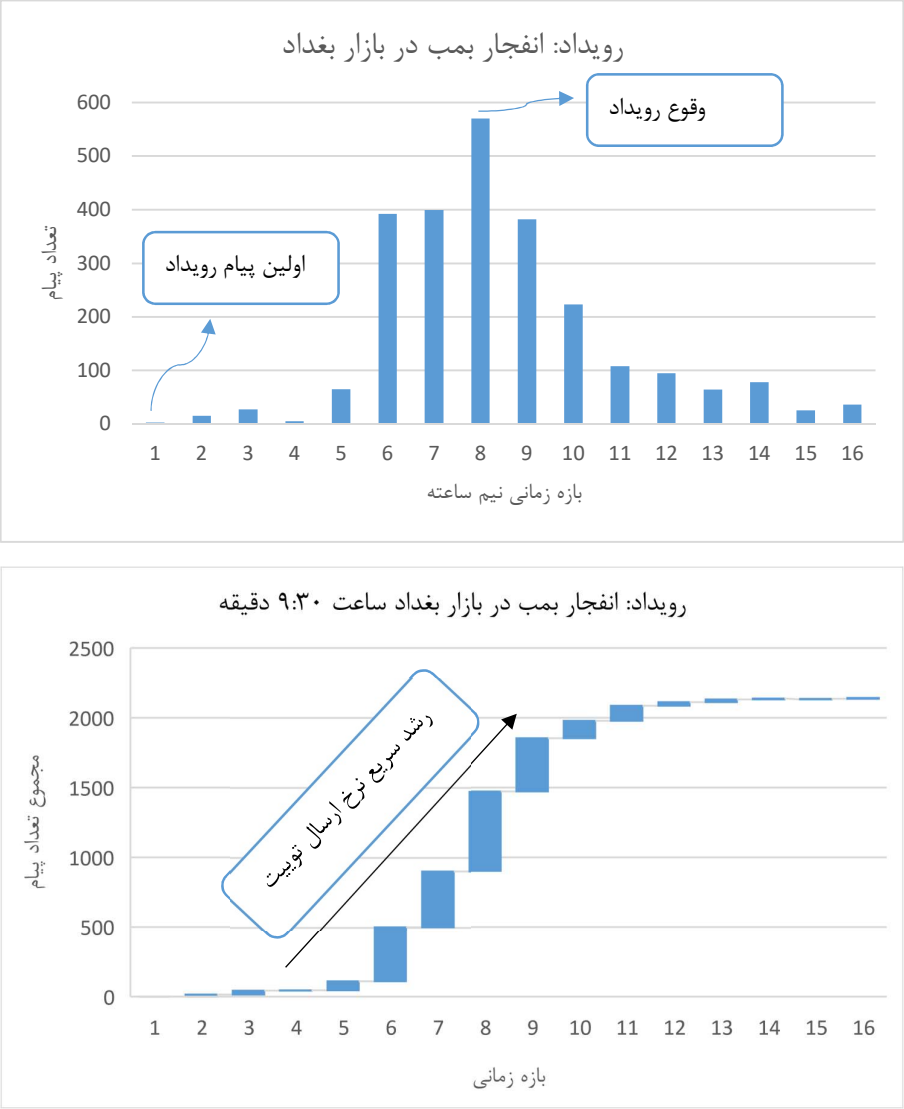




(ب) رویداد گفتگوهای صلح فلسطین و اسرائیل

شکل ۳-۱) تغییرات نرخ ارسال توییتهای مربوط به دو رویداد قابل پیش‌بینی الف) به رسمیت شناختن ازدواج همجنسگرایان در یوتا ب) گفتگوهای صلح فلسطین و اسرائیل

بعنوان مثالی از حوادث غیر قابل پیش‌بینی، در شـکل ۳-۲، نمودار تعداد پیامهای ارسـالی در بازه‌های زمانی نیم ساعته مربوط به یک رویداد غیرقابل پیش‌بینی آورده شده است. این رویداد مربوط به انفجار بمب در بازار بغداد در تاریخ ۲۵ دسامبر ۲۰۱۳ ساعت ۹:۳۰ قبل از ظهر است. در اینجا نیز آنچنان که مشاهده می‌شود، قبل از وقوع رویداد تقریبا هیچگونه نشانه در مورد رویداد وجود ندارد، اما با وقوع آن، توییتها با نرخ بالایی به شبکه ارسال شده است.



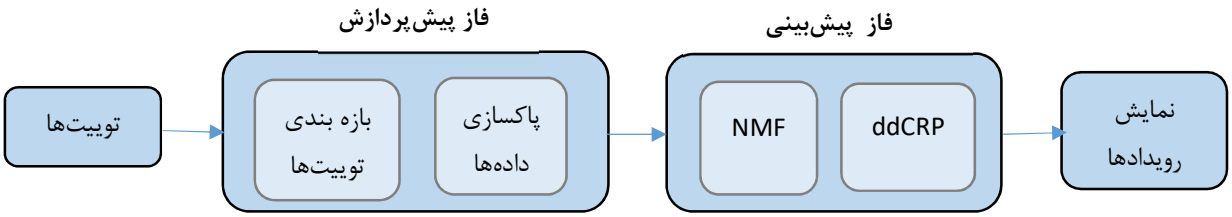
شکل ۳-۲) تغییرات نرخ ارسال توییتهای مربوط به یک رویداد غیرقابل پیش‌بینی

در فصـــل ۴ یک مطالعه آماری بر روی مجموعه داده موجود برای همه رویدادهای قابل پیش‌بینی و غیرقابل پیش‌بینی به منظور بررسی این رفتار شبکه اجتماعی انجام شده است.

۳-۲-۲ پیش‌بینی وقوع رویداد

در بخش قبل، وجود تغییرات نرخ ارسال توییتهای مرتبط با یک رویداد در شبکه اجتماعی توییتر قبل از وقوع رویداد مورد برر سی قرار گرفت. در اینجا هدف این ا ست تا روش پی شنهادی جهت پیش‌بینی رویدادها بر اساس ویژگی تغییر نرخ ارسال توییتهای مرتبط با آنها معرفی گردد.

معماری روش پیشنهادی، در شکل ۳-۳ آمده است. این سیستم شامل سه بخش اصلی است. بخش اول با عنوان پیش‌پردازش، خود شامل دو گام است. ابتدا توییتهای ورودی، بر اساس یک پنجره زمانی با طول ثابت گروه‌بندی می شوند. سپس عملیات معمول پیش‌پردازش داده‌های متنی در سیستم‌های داده کاوی انجام می‌گیرد و در نهایت بردار وزن TF-IDF ایجاد می‌گردد. در بخش دوم سیستم پیشنهادی، ضمن عمل خوشه‌بندی با استفاده از دو الگوریتم NMF و ddCRP، ویژگی تغییرات نرخ ورود توییتها به هر خوشه مورد بررسی قرار می‌گیرد تا بتوان به کمک آن پیش‌بینی وقوع رویداد را انجام داد.



شکل ۳-۳) معماری کلی سیستم پیشنهادی برای پیش‌بینی رویداد

نهایتا در آخرین بخش سیستم پیشنهادی، رویدادهای پیش‌بینی شده، نمایش داده می‌شوند. در ادامه این قسمت، هر یک از بخش‌های اصلی سیستم با جزییات بیشتر توضیح داده شده است.

الف) پیش‌پردازش

بطور کلی در اغلب اسناد متنی، یکسری داده‌های غیرمفیدی ممکن است وجود داشته باشد که علاوه بر اینکه بر دقت نهایی تاثیر منفی می گذارند، حجم داده ها و زمان پردازش را نیز تحت تاثیر قرار می‌دهند. بنابراین لازم است تا قبل از پردازش اسناد، عمل پاکسازی بر روی آنها اعمال شود. در این پژوهش قبل از عمل پاکسازی داده‌ها، بر مبنای یک پنجره زمانی، که طول آن در شروع به کار سیستم قابل تنظیم است، توییت‌ها دسته‌بندی می‌شوند. مثلا اگر طول پنجره زمانی برابر نیم ساعت تنظیم شود، توییتهایی که در یک فاصله زمانی نیم ساعت از سال می شوند، در یک گروه قرار می‌گیرند. این

گروه‌بندی توئیتها از دو جنبه اهمیت دارد. اولا، از آنجاییکه پایه تحلیل‌های بعدی سیستم بر اساس تغییرات نرخ ورود توییت‌ها به سیستم می‌باشد، محاسبات را ساده‌تر می‌کند. چون مدت زمان در طول بازه‌ها ثابت است و تغییرات تعداد توییتهای ارسالی می‌تواند به تنهایی نمایانگر تغییرات نرخ ارسال توییت‌ها با شد. ثانیا، خواندن برخط توییت‌ها برای سیستم نیز بدین‌گونه شبیه سازی می شود. بعد از گروه‌بندی توییتها بر اساس زمان انتشار آنها، به ترتیب هر گروه خوانده می‌شود و اعمال معمول پاکسازی و پیش‌پردازش بر روی داده‌ها همانند شکستن رشته متنی، حذف کلمات متوقف کننده، حذف اعداد و لینک‌ها، و استخراج بن واژه‌ها بر روی توییت‌ها انجام می‌گیرد. در انتهای این مرحله، با استفاده از روش TF-IDF (رابطه ۳-۱) ماتریس وزن سند × کلمه که محتوی میزان اهمیت هر کلمه موجود در توییتها است، ایجاد می‌شود و بعنوان ورودی مرحله پیش‌بینی مورد استفاده قرار می‌گیرد.

$$TF - IDF(w, t, N, n_w) = f_{w,t}.\log\left(\frac{N}{n_w}\right) \qquad 1-3$$

در رابطه ۳-۱، $f_{w,t}$ فراوانی کلمه w در توییت t است، N تعداد کل توییت‌ها و n_w تعداد توییتهایی است که کلمه مورد نظر در آنها وجود داشته است.

ب) پیش‌بینی رویداد

روشهای خوشه‌بندی ساده همانند خوشه‌بندی k-means در موضوع خوشه‌بندی داده‌های متنی جریانی چندان کاربرد ندارند. چون در ابتدای اجرای الگوریتم نیاز به مشخص بودن تعداد خوشه‌ها می‌باشد و علاوه بر این برای داده‌های جریانی، پیچیدگی محاسباتی بالایی دارند. دلیل دیگری که می‌توان از دیدگاه منطقی به آن اشاره کرد این است که در محیطهای عملیاتی پویایی همچون شبکه‌های اجتماعی، اطلاعات همواره در حال افزایش و تکامل است، بنابراین در نظر گرفتن یک حد بالا برای تعداد خوشه‌ها صحیح به نظر نمی‌رسد و باید راهکاری را در پیش گرفت که با تکامل اطلاعات قبلی و وقوع اطلاعات جدید سازگار با شد. بنابراین از الگوریتم‌های خو شه‌بندی افزایشی در این زمینه استفاده می‌شود. در روش خوشه‌بندی افزایشی بدین گونه عمل می‌شود که با ورود داده‌ی جدید،

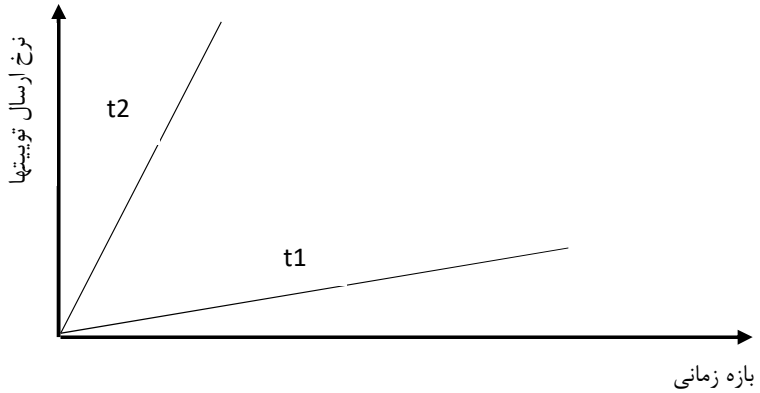
شباهت این داده با خوشه‌های موجود بر اساس یک معیار شباهت سنجیده می‌شود. اگر مقدار شباهت داده جدید به یکی از خوشه‌ها بالاتر از یک مقدار حد آستانه باشد، داده به آن خوشه اختصاص می‌یابد. اگر مقدار شباهت برای هیچ کدام از خوشه‌ها برقرار نباشد، خوشه‌ای جدید ایجاد می‌شود و داده در آن درج می‌گردد. و این روال برای تمام داده‌های جدید بعدی تکرار می‌شود.

در این پژوهش نیز از روش خوشه‌بندی افزایشی استفاده شده‌است. در آغاز فرایند خوشه‌بندی افزایشی، گروه نخست توییت‌ها بر اساس الگوریتم NMF خوشه‌بندی می‌شوند. تا بدین ترتیب اولا مقداردهی اولیه خوبی برای خوشه‌های اولیه و محتوای آنها ایجاد شود و ثانیا از نظر زمانی نیز سریعتر ساختار اولیه مناسب ایجاد گردد. سپس با خواندن هر مجموعه توییت جدید، روال خوشه‌بندی افزایشی اجرا می‌گردد. در این روش از خوشه‌بندی افزایشی ddCRP استفاده شده است. در پایان هر مرحله خوشه‌بندی افزایشی ، مجموعه‌ای از خوشه‌ها موجود است. در واقع می‌توان هر خوشه را بعنوان یک رویداد در نظر گرفت. پس از خوشه‌بندی هر گروه از توییت‌ها، مقدار نرخ ورود توییت‌ها به هر خوشه از زمان تشکیل آن خوشه محاسبه و ذخیره می‌گردد. البته از آنجاییکه گروه‌بندی توییت‌ها در مرحله پیش‌پردازش بر مبنای یک پنجره زمانی با طول ثابت انجام شده بود، نرخ ورود توییت‌ها در بازه زمانی i ، $(V_{t,i})$، با تعداد توییت‌های ورودی به هر خوشه در همین بازه، $n_{t,i}$، رابطه مستقیمی خواهد داشت (رابطه ۳-۲).

$$V_{t,i} = {n_{t,i}}/{\|i\|} \qquad 2-3$$

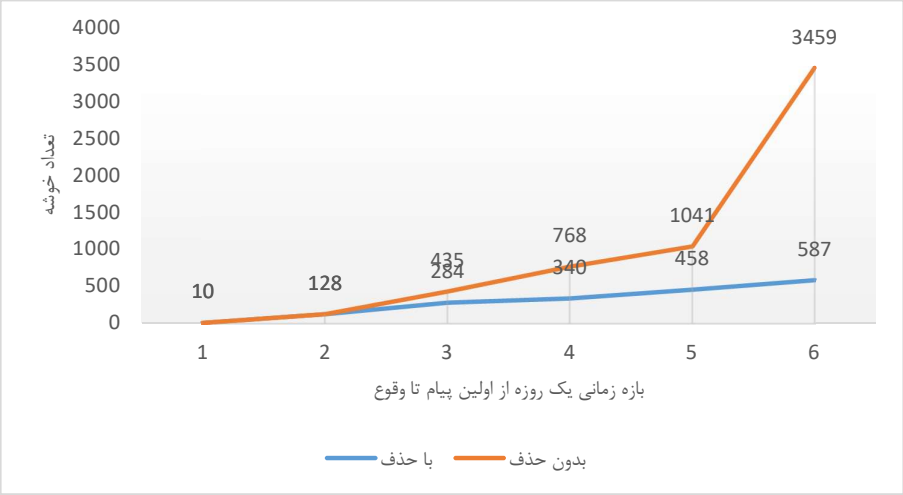
در رابطه ۳-۲، $\|i\|$ طول بازه زمانی می‌باشد.

حال می‌توان بر مبنای میزان تغییرات نرخ ورود توییت‌ها به هر خوشه و مقایسه این میزان تغییرات با یکسری حدود آستانه که بصورت تجربی بدست می‌آیند، وقوع رویداد در آینده را پیش‌بینی نمود. همانطور که در شکل ۳-۴ مشخص است، دو حد آستانه t1 و t2 بعنوان حدود آستانه برای میزان شیب رشد خوشه‌ها به منظور تعیین نوع خوشه‌ها در روش پیشنهادی مورد استفاده قرار گرفته اند.



شکل ۳-۴) دو حد آستانه میزان تغییرات رشد خوشه‌ها در هر بازه زمانی از ابتدا

برای هر خوشه در انتهای هر بازه زمانی، تغییرات نرخ ورود توییتها با آن خوشه از زمان اولین توییت ورودی به آن محاسبه می‌شود. چنانچه تغییرات نرخ از حد آستانه t2 بیشتر باشد، آن خوشه بعنوان خوشه مرتبط با رویدادهای ناگهانی شناسایی می‌شود، در غیر اینصورت اگر تغییرات نرخ از حد آستانه t2 کمتر و از t1 بیشتر باشد، خوشه بعنوان رویداد قابل پیش بینی تعیین می‌گردد. اما اگر شیب رشد خوشه در تعداد تعداد بازه زمانی معین کمتر از t1 باشد، خوشه موردنظر بعنوان یک رویداد بی اهمیت بطور کامل حذف می‌گردد. لازم به ذکر است که خوشه‌های نوع اول و دوم بعد از برچسب گذاری به رویدادهای ناگهانی و قابل پیش بینی، حذف نمی‌شوند و توییتهای مرتبط به خود را جذب می‌کنند. اگر این خوشه‌ها حذف شوند، در ادامه مجددا تشکیل می‌شوند و بنابراین منجر به شناسایی رویدادهای تکراری می‌گردند. حذف خوشه‌های نوع سوم، تاثیر مهمی در زمان اجرای برنامه و حافظه استفاده شده سیستم و میزان رشد تعداد خوشه‌ها دارد. علاوه براین اگر واقعا مرتبط به یک رویداد باشند، در ادامه کار سیستم، مجدد ایجاد خواهند شد و به نوع اول و یا دوم برچسب گذاری خواهند شد. شکل ۳-۵ میزان رشد تعداد خوشه‌ها در دو حالت مختلف برای رویدادهایی که از اولین پیام آنها تا وقوعشان شش روز فاصله بوده است را نشان می‌دهد. حالت اول میزان رشد خوشه‌ها را در وضعیتی نشان می‌دهد که عمل حذف خوشه‌ها وجود ندارد و حالت دوم وضعیتی را نشان می‌دهد که عمل حذف خوشه‌ها بر اساس حد آستانه انجام می‌گیرد.



شکل ۳-۵) میزان رشد تعداد خوشه‌ها در دو وضعیت حذف یا عدم حذف خوشه‌های نوع سوم

ج) نمایش رویدادهای آینده

آخرین بخش سیستم پیشنهادی، مصورسازی رویدادهایی می‌باشد که در روش پیشنهادی بعنوان رویدادهای آینده پیش‌بینی شده‌اند. از هر خوشه‌ای که نشان دهنده رویدادهای آینده است، k کلمه پرتکرار بعنوان شرح رویداد، به خروجی ارسال می‌شود. جدول ۳-۱ چند ردیف از خروجی واقعی روش پیشنهادی است که با اعمال بر روی مجموعه داده‌ای از توییتر که در فصل بعدی معرفی خواهد شد، بدست آمده است.

جدول۳-۱) خروجی سیستم پیشنهادی به عنوان رویدادهای پیش‌بینی شده

ردیف	شش کلمه پرتکرار چند خوشه بعنوان شرح رویدادهای پیش‌بینی شده
۱	Chemical, Kerry, use, house, weapons, stand
۲	Divorce, old, request, wife, George, make
۳	Fire, hospital, immediate, dead, sandy, topic
۴	Merkel, exit, angel, third, car, stop
۵	Mayor, healthcare, new, india, first, york

۳-۳ روش پیشنهادی برای پیش‌بینی نتیجه انتخابات

روش پیشنهادی رساله در مواجهه با انتخابات می‌تواند به دو صورت عمل کند. اول اینکه می‌تواند رویداد انتخابات را پیش بینی کند. یعنی اینکه پیش بینی کند که در آینده رویدادی در مورد انتخابات که همان برگزاری آن است به وقوع خواهد پیوست و حتی در صورت لزوم، با تحلیل محتوای توییتها، امکان تعیین تاریخ آن میسر است. دوم اینکه نتیجه انتخابات نیز خود یک رویداد می‌باشد، چون بر اساس تعریفی که از رویداد در فصل اول ارائه شد، تغییرات چشمگیری در تعداد توییتهای ارسالی به شبکه اجتماعی خواهد داشت. به عبارت دیگر هنگامی که هر یک از احزاب پیروز انتخابات شوند، تعداد توییتها افزایش خواهند یافت و این یعنی یک رویداد که در واقع همان نتیجه انتخابات است به وقوع پیوسته است. بنابراین روش پیشنهادی در مورد پیش بینی نتیجه انتخابات دقیقا پیش بینی یک رویداد می‌باشد. پیش بینی این رویداد خاص، بر اساس تغییرات نرخ ارسال توییتهای احساسی مختلف بوقوع پیوسته است و با اصل موضوع رساله در ارتباط است.

در این بخش استفاده از ویژگی تغییرات نرخ ارسال توییتها به شبکه اجتماعی به منظور پیش‌بینی رویداد انتخابات مورد بررسی قرار می‌گیرد. اگر برای پیش‌بینی نتیجه انتخابات، تنها تعداد توییتهای ارسالی طرفداران احزاب بدون توجه به فاکتورهای دیگر خصوصا احساسات نهفته در توییتها در نظر گرفته شود، نتایج دقیقی حاصل نخواهد شد [۶۳]. بنابراین از ویژگی‌های دیگر، خصوصا احساسات نهفته در هر توییت در پژوهش‌های مختلف استفاده شده است. احساسات مثبت توییت‌های منتسب به یک حزب، بعنوان یک امتیاز برای آن حزب و توییتهای حاوی احساسات منفی بعنوان یک ضعف برای آن حزب در نظر گرفته می‌شود. غالبا روشهایی که احساسات یک توییت را استخراج می‌کنند مبتنی بر یک فرهنگ لغت هستند. در این فرهنگ لغت، به هر یک از کلمات بطور تجربی نمراتی متناسب با بار احساسی آنها داده می‌شود. سپس نمره احساسی یک جمله یا یک توییت، بر اساس اینکه از کدام کلمات موجود در این فرهنگ لغت تشکیل شده‌اند، محاسبه می‌شود. از اینرو یک جمله می‌تواند دارای

احساسات مثبت، منفی یا خنثی باشد. در روش پیشنهادی از توییتهای احساسی خنثی صرفنظر شده است.

شاخص یا اندیکاتوری که در روش پیشنهادی به عنوان عامل تعیین‌کننده در پیش‌بینی انتخابات مورد استفاده قرار گرفته است، نسبت نرخ رسال پیامهای مثبت به منفی افراد در یک پنجره زمانی ثابت i می‌باشد:

$$A_i = \frac{(Positive\ Tweets\ Rate)_i}{(Negetive\ Tweets\ Rate)_i + 1} = \frac{(Positive\ Tweets\ Count)_i}{(Negetive\ Tweets\ Count)_i + 1} \qquad 3-3$$

اندیکاتور پیش‌بینی (A_i) در رابطه ۳-۳، از نسبت نرخ پیامهای احساسی مثبت به منفی در پنجره زمانی i حاصل می شود. چون نرخ رسال در یک پنجره زمانی ثابت محاسبه می شود، رابطه مذکور با نسبت تعداد توییتهای احساسی مثبت به منفی برابر است. از آنجا که ممکن است در یک پنجره زمانی، تعداد توییتهای منفی برابر صفر باشد، عدد یک در مخرج کسر قرار داده شده است.

مقایسه مقدار این اندیکاتور در هر پنجره زمانی برای نامزدهای شرکت‌کننده در انتخابات، شانس آنها برای موفقیت در انتخابات را نشان می‌دهد. بنابراین طول پنجره زمانی بر اساس اینکه بخواهیم پیش‌بینی چه مدت قبل از انتخابات انجام شود، تعیین می شود. به عبارت دیگر برای پیش‌بینی نتیجه انتخابات در d روز قبل از آن، طول پنجره زمانی، برابر d تنظیم می‌شود.

در روش پیشنهادی برای پیش‌بینی مقدار اندیکاتور در بازه زمانی بعدی، از روش تخمین سالمندی استفاده شده است. روش سالمندی بر مبنای یک میانگین‌گیری نمایی از مشاهدات قبلی عمل می‌کند:

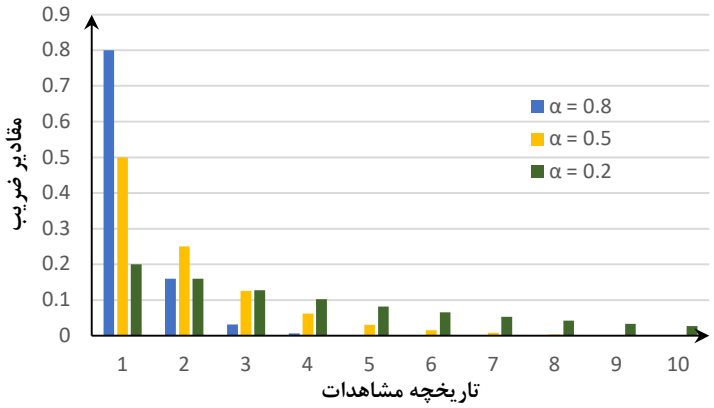
$$A_{i+1} = \alpha O_i + (1-\alpha)A_i \qquad 4-3$$

در رابطه ۳-۴، A_{i+1} مقدار پیش‌بینی اندیکاتور برای پنجره زمانی بعدی (i+1) می‌باشد. O_i مقدار واقعی مشاهده شده اندیکاتور در پنجره زمانی جاری است و A_i مقدار پیش‌بینی شده اندیکاتور در مرحله i است. α یک پارامتر بین صفر و یک است که وزن نسبی سابقه پیش‌بینی‌ها و وزن مقدار

مشاهده اخیر را در پیش‌بینی تعیین می‌کند. هر چه مقدار α به یک نزدیکتر باشد، مشاهدات اخیر اثر بیشتری در محاسبه متوسط نمایی دارند و هرچه مقدار آن به صفر نزدیکتر شود، مشاهدات قدیمی اثر بیشتری در محاسبه میانگین دارند (رابطه ۳-۵)

$$A_{i+1} = \alpha O_i + (1-\alpha)\alpha O_{i-1} + \cdots + (1-\alpha)^{i+1} A_0 \qquad 5-3$$

از آنجا که α و $(1-\alpha)$ هر دو کوچکتر از یک هستند، جمله‌های بعدی در رابطه ۳-۵ به تدریج کوچکتر می‌شوند. اندازه α به عنوان تابعی از موقعیت آن در رابطه ۳-۵ در شکل ۳-۶ نشان داده شده است. برای $\alpha = 0.8$ تقریبا تمام وزن به چهار مشاهده اخیر داده شده است. به عبارت دیگر، چهار مقدار مشاهده شده اخیر اندیکاتور بیشترین تاثیر را در پیش‌بینی مقدار بعدی اندیکاتور،خواهند داشت در حالیکه برای $\alpha = 0.5$ متوسط گیری بر روی حدود هشت مشاهده آخر پخش شده است.



شکل ۳-۶) تاثیر ضریب آلفا بر تعداد مشاهدات اخیری که برای پیش‌بینی مقدار بعدی استفاده می‌شوند.

مزیت بکارگیری مقدار نزدیک به یک برای α این است که تغییرات سریع مشاهدات، سریعتر در مقدار متوسط منعکس می‌شوند و دقت پیش‌بینی تا حد زیادی افزایش می‌یابد. بنابراین در ماه‌های نزدیک به انتخابات خصوصا در ماه منتهی به انتخابات که معمولا مناظرات انتخاباتی و تبلیغات نهایی نامزدها بیشتر می‌شوند، مقادیر α نزدیک به یک، در پیش‌بینی موثرتر خواهد بود.

۳-۴ جمع بندی

در این فصل روش پیشنهادی برای پیش‌بینی رویداد با استفاده از تحلیل شبکه اجتماعی معرفی شد. ابتدا نشان داده شد که قبل از وقوع رویدادهای قابل پیش‌بینی توییتهایی درباره آن رویدادها به شبکه ارسال می‌شود و هرچه به زمان وقوع رویدادها نزدیکتر می‌شود، ارسال توییتها با سرعت بیشتری اتفاق می‌افتد. سپس مراحل پیش‌پردازش و خوشه‌بندی توییتها بر مبنای الگوریتمهای NMF و ddCRP تشریح گردید. بعد از آن بر مبنای تغییرات نرخ ورود توییتها به هر خوشه، وقوع رویداد مرتبط با خوشه‌ها در آینده پیش‌بینی می‌گردد. نهایتا توصیفی از رویداد بر اساس چند کلمه پرتکرار خوشه‌ها ارائه می‌گردد. نهایتا روش پیشنهادی برای پیش‌بینی نتیجه انتخابات بر مبنای تحلیل احساسات و نرخ ارسال توییتها بیان شد.

۴

پیاده‌سازی و تحلیل آزمایش‌ها

۴-۱ مقدمه

در این فصل آزمایشات انجام شده بر روی مجموعه داده‌هایی از توییتر آورده شده است. ابتدا روشهای کسب داده از توییتر توضیح داده شده است. سپس مجموعه آزمایشات انجام شده به منظور ارزیابی روش پیشنهادی برای پیش‌بینی رویداد آمده است و در نهایت آزمایشات مرتبط با روش پیشنهادی برای پیش‌بینی انتخابات آورده شده است.

۴-۲ روشهای کسب داده از توییتر

با استفاده از رابط برنامه‌نویسی کاربردی و ثبت نام در سایت توییتر می‌توان داده‌ها را بصورت آنلاین دریافت کرد. رابط برنامه‌نویسی کاربردی یک مجموعه از توابع، پروتکل‌ها، و ابزارهایی است که برای ساخت یک برنامه کاربردی و یا ارتباط با خدماتی خاص، مورد استفاده قرار می‌گیرد.

توئیتر سه نوع رابط برنامه نویسی در اختیار برنامه نویسان قرار می‌دهد که دو مورد آنها عمومی و رایگان هستند و امکان دسترسی به داده‌های شبکه را فراهم می‌کنند. این اطلاعات به منظور انجام فعالیتهای پژوهشی، جستجوی اطلاعات بصورت بلادرنگ و یا تاریخی مورد استفاده قرار می‌گیرند. در ادامه این سه نوع API آمده است.

API REST

این API امکان خواندن و نوشتن داده در توئیتر را از طریق دستورات ساده HTML مثل GET و POST[۳۹] فراهم می‌کند. نمونه‌ای از خروجی این API در شکل ۴-۱ آمده است.

[39] https://dev.twitter.com/rest/public

```
“favourites_count”:757,
“Follow_request_sent”:false,,
“Followers_count”:143916
“following”:false,
“Friends_count”: 1484
“geo_enabled”:true
“id”:2244994945,
“is_str”:”2244994945”
“lang”:”en”
…
```

شکل ۴-۱ خروجی حاصل از API REST

محدودیت اصلی در کسب داده‌ها از طریق API REST، تعداد دفعاتی است که می‌توان تقاضای خواندن ارسال کرد که این نیز بستگی به نوع احراز هویتی دارد که مورد استفاده قرار می‌گیرد.

توئیتر برای هر متقاضی، یک پنجره کاری ۱۵دقیقه‌ای در نظر می‌گیرد. این بدین معنی است که هر توکن معتبر کاربر، تنها برای یک پنجره کاری قابل استفاده است و اگر قبل از اتمام ۱۵ دقیقه، توکن به خاطر محدودیت دریافت داده، منقضی شده باشد، باید صبر کرد تا در ۱۵دقیقه بعدی که توکن دوباره بازنشانی می‌شود، ادامه کسب داده انجام شود.

Streaming API

این API امکان دسترسی به جریان عمومی توئیت‌ها و دیگر موارد مثل پاسخ‌ها، بازتوئیت‌ها و موارد دیگر را می‌دهد. برخلاف API REST با این امکان، ارتباط پایداری بین متقاضی و توئیتر ایجاد می‌شود. این API همچنین فیلترهای متعددی در اختیار توسعه‌دهندگان قرار می‌دهند تا تنها داده‌های با ویژگیهای خاص مدنظرشان را دریافت کنند. بعنوان مثال دریافت توئیت‌هایی از ناحیه جغرافیایی خاص، توئیتهایی با محتوی کلماتی خاص و یا توئیت‌های محتوی هشتگی مشخص. محدودیت این شیوه این است که تنها می‌توان به یک درصد از کل داده‌ها دسترسی داشت، که البته با توجه به حجم داده‌ها (بالغ بر ۵۰۰ میلیو توئیت در روز)، این مقدار هم برای کارهای پژوهشی قابل توجه می‌باشد. در شکل ۴-۲ نمونه خروجی این نوع API نشان داده شده است.

```
"followers_count": 70,
"protected": false,
"notifications": null,
"profile_background_image_url_https": "https://si0.twimg.com/images/themes/theme1/bg.png",
"profile_background_color": "C0DEED",
"verified": false,
"geo_enabled": true,
"time_zone": "Pacific Time (US & Canada)",
"description": "Born 330 Live 310",
"default_profile_image": false,
"profile_background_image_url": "http://a0.twimg.com/images/themes/theme1/bg.png",
"statuses_count": 579,
"friends_count": 110,
"following": null,
"show_all_inline_media": false,
"screen_name": "sean_cummings"
```


شکل ۴-۲) نمونه خروجی Streaming API

Gnip

این API دسترسی به داده‌های توئیتر را در قبال پرداخت هزینه‌ای متناسب با نیاز مشتری فراهم می‌کند. اما در قبال دریافت هزینه مذکور، د ستر سی به کل جریان داده‌ها بدون هیچ گونه محدودیتی امکان پذیر است.

۴-۳ آزمایشات مربوط به پیش‌بینی رویداد

مجموعه داده مورد استفاده در این پژوهش برای پیش‌بینی رویداد از مقاله [۷۶] گرفته شده است. این مجموعه داده شامل توییت‌های برگرفته از رسانه‌های خبری معروف همچون CNN, BreakNews, BBC در شبکه اجتماعی توییتر در یک بازه زمانی شش ماهه می‌با شد. مجموعه داده شامل تقریبا ۴۳ میلیون شناسه توییت می‌باشد که به ۵۲۳۴ رویداد ا شاره دارد. هر رویداد در مجموعه داده، با تعدادی کلمه کلیدی تو صیف می شود که برای ارزیابی سیستم پیشنهادی مورد ا ستفاده قرار می‌گیرد. نحوه برچسب گذاری توییتها در مجموعه داده بدین گونه بوده است که ابتدا سرخط خبرها از رسانه‌های مختلف خبری ا ستخراج شده ا ست و سپس بطور موازی بر ا ساس یکسری کلمات کلیدی موجود در سرخط خبرها، تمامی توییتهای دارای آن کلمات کلیدی از جریان توییتر ا ستخراج و برچسب گذاری شده اند. خلاصه آماری مجموعه داده مذکور در جدول ۴-۱ آورده شده‌است.

جدول ۴-۱) خلاصه آماری مجموعه داده اولیه

آمار مجموعه داده	کمترین	متوسط	بیشترین
تعداد توییت برای هر رویداد	۱۰۰۰	۸۲۵۴	۵۱۰۹۲۰
تعداد کلمات کلیدی برای هر رویداد	۲	۳,۷۷	۳۹

برای هر توییت، اطلاعات مورد نیاز سیستم شامل شناسه توییت، متن توییت و تاریخ انتشار آن با استفاده از ابزار twitter Data Base Collector[۴۰] جمع آوری شده است. متاسفانه به دلایلی همچون مسدود بودن برخی حسابهای کاربری و حذف توییت‌ها، تقریبا ۲۷ میلیون توییت جمع آوری گردیده است. اما خوشبختانه این تعداد توییت نیز مجموعه کامل ۵۲۳۴ رویداد را پوشش می‌دهد. اما به دلیل اینکه سیستم پیشهادی به تعداد توییتهای مربوط به هر رویداد بسیار وابستهاست، تنها رویدادهایی در آزمایشها انتخاب شده‌اند که بیش از ۸۰ درصد توییتهای آنها دردسترس باشد. بنابراین تعداد رویدادهای نهایی ۱۹۴۰ رویداد و کل توییت‌ها تقریبا به ۱۲ میلیون تقلیل یافتند. خلاصه آماری مجموعه داده نهایی در جدول ۴-۲ آورده شده‌است.

جدول ۴-۲) خلاصه آماری مجموعه داده نهایی

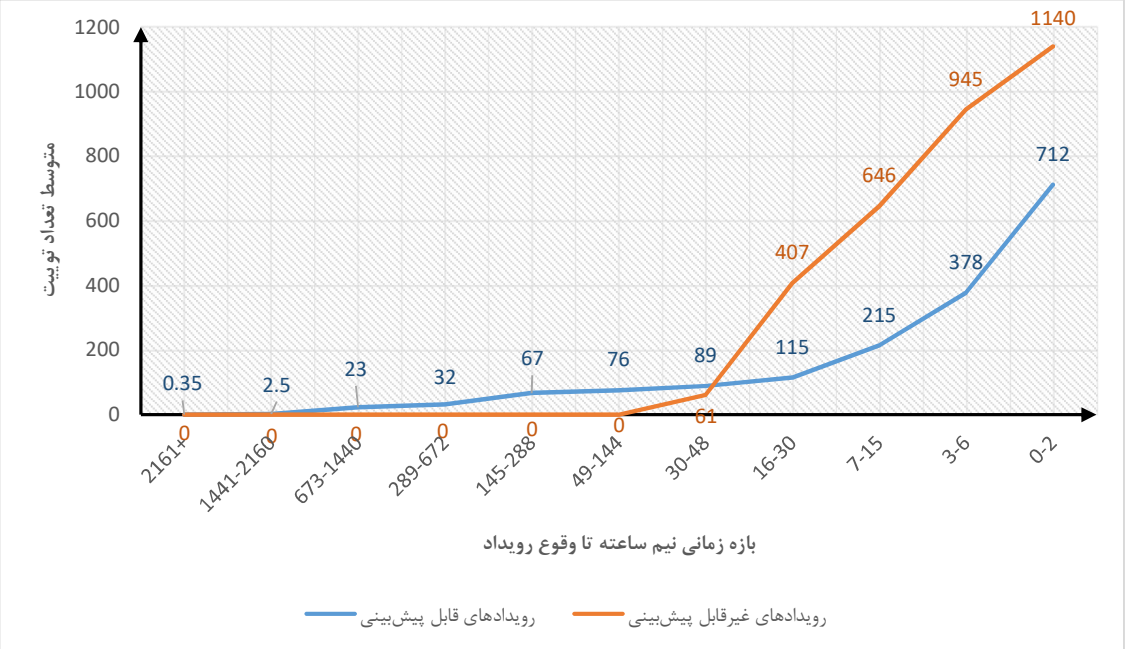
آمار مجموعه داده	کمترین	متوسط	بیشترین
تعداد توییت برای هر رویداد	۱۱۲۰	۶۶۲۵	۳۸۴۳۶
تعداد کلمات کلیدی برای هر رویداد	۲	۳,۶۵	۲۹

در آزمایش نخست، رفتار شبکه اجتماعی توییتر قبل از وقوع رویدادهای قابل پیش‌بینی و رویدادهای غیرقابل پیش‌بینی یا ناگهانی بررسی شده است. همانطور که در مقدمه بیان شد، رویدادهای قابل پیش‌بینی آن دسته از رویدادهایی هستند که تحت تاثیر عواملی از داخل شبکه اجتماعی بوجود می‌آیند. بعنوان مثال اعضا شبکه اجتماعی، توییت‌هایی در مورد یک رویداد به شبکه ارسال می‌کنند.

[40] https://github.com/socialsensor/twitter-dataset-collector

روش پیشنهادی نیز بر اساس تغییرات نرخ ارسال همین توییتهای ارسالی کاربران در مورد هر رویداد به پیش‌بینی وقوع آن می‌پردازد. رویدادهای ناگهانی رویدادهایی هستند که عواملی خارج از شبکه اجتماعی منجر به وقوع آنها می‌شوند، مانند وقوع زلزله که بلافاصله بعد از وقوع آن، شبکه اجتماعی را تحت تاثیر قرار می‌دهد.

نتایج آزمایش در شکل ۴-۳ آورده شده است. محور افقی تعداد بازه زمانی نیم ساعته تا وقوع رویداد را نشان می‌دهد. محور عمودی متوسط تعداد توییتهای ارسالی مرتبط با رویدادها می‌باشد. همانطور که در شکل قابل مشاهده است، برای رویدادهای قابل پیش‌بینی، از حداقل ۲۱۶۱ بازه زمانی نیم ساعته یا ۴۵ روز قبل از وقوع برخی از آنها، توییتهایی در مورد آنها ارسال شده است و هرچه به لحظه وقوع رویداد نزدیکتر می‌شود، نرخ ارسال توییتها افزایش می‌یابد. اما در رویدادهای ناگهانی، تنها در روز وقوع رویداد، نرخ ارسال توییتهای مرتبط با آنها، به شدت افزایش می‌یابد.



شکل ۴-۳) متوسط تعداد توییت‌های ارسالی در بازه‌های نیم ساعته بین اولین توییت ارسالی مربوط به رویدادها تا وقوع آنها

در آزمایش دوم، کارایی روش پیشنهادی بر اساس معیارها ارزیابی استاندارد دقت[41]، فراخوانی[42] و نمره F1[43] مورد ارزیابی قرار گرفته است. معیار دقت به این اشاره دارد که چند درصد از خروجی روش پیشنهادی به درستی نشاندهنده رویداد هستند. معیار فراخوانی مشخص می‌نماید که چند درصد از رویدادهای موجود در مجموعه داده توسط روش پیشنهادی به درستی پیش بینی شده است. بنابراین برای محاسبه دقت و فراخونی بر اساس کلمات کلیدی رویدادها، به ترتیب از رابطه ۴-۱ و ۴-۲ استفاده می‌شود.

$$P = \frac{|\{DS\ Events\ Keywords\} \cap \{Predicted\ Events\ Keywords\}|}{|\{Predicted\ Events\ Keywords\}|} \qquad 1-4$$

$$R = \frac{|\{DS\ Events\ Keywords\} \cap \{Predicted\ Events\ Keywords\}|}{|\{DS\ Events\ Keywords\}|} \qquad 2-4$$

معیار F1 نیز که نوعی میانگین بین معیارهای دقت و فراخوانی می‌باشد از طریق رابطه ۴-۳ بدست می‌آید.

$$F1 - Score = 2 \times \frac{P \times R}{P + R} \qquad 3-4$$

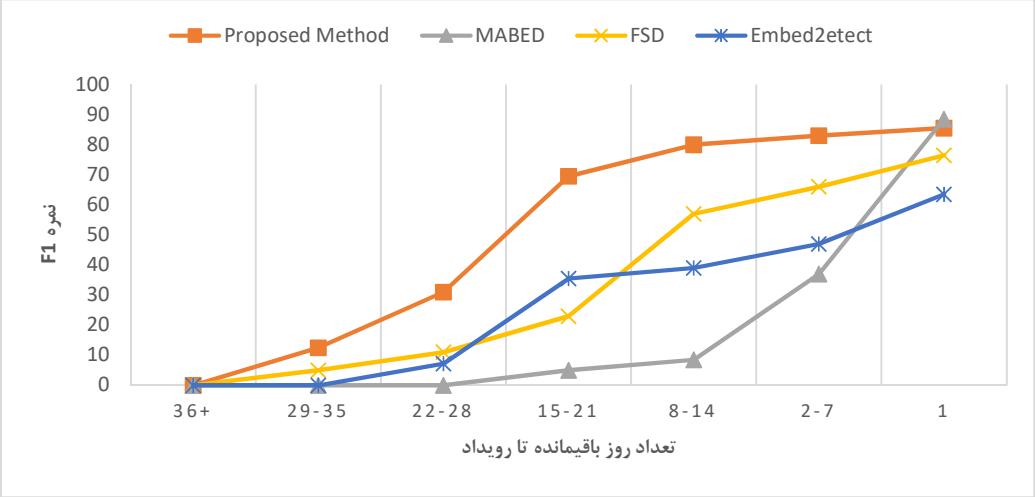
هرقدر نتایج بدست آمده برای این سه شاخص به عدد یک نزدیکتر باشند، کارایی و عملکرد روش پیشنهادی بهتر خواهد بود.

به منظور شناسایی رویدادهای خروجی سیستم، کلمات پرتکرار هر خوشه، بعنوان توصیفی از رویدادهای پیش‌بینی شده، با کلمات کلیدی رویدادها که در مجموعه داده در اختیار می‌باشند، مقایسه می‌شوند. درصورتی که کلمات کلیدی خروجی سیستم، حداقل شباهت (s) را با کلمات کلیدی مربوط به هر رویداد در مجموعه داده داشته باشد، بعنوان یک رویداد شناسایی شده درنظر گرفته می‌شود. معیارهای کارایی روش پیشنهادی بر اساس تنظیم پارامتر تراکم خوشه‌بندی افزایشی به α=۰,۲ و شباهت s =٪۷۰ محاسبه شده است. نتیجه مقایسه روش پیشنهادی با سه روش تشخیص رویداد

[41] precition
[42] Recal
[43] F1-Score

MABED[۴۴][۳۳] ، FSD[۴۵] [۴۱] و Embed2Detect [۷۷] در ادامه آورده شده است. همانطور که در فصل دوم گفته شد، MABED با استفاده از یک رویکرد آماری و تشخیص بی‌نظمی در تعداد کلمات ارسالی به توییتر، به تشخیص رویداد می‌پردازد. FSD نیز با یک رویکرد افزایشی و تشخیص اولین ظهور، وقوع رویداد را تشخیص می‌دهد. روش Embed2Detect نیز بر اساس ترکیب ویژگیهای معنایی نهفته در کلمات و خوشه بندی سلسله مراتبی به تشخیص رویداد می‌پردازد.

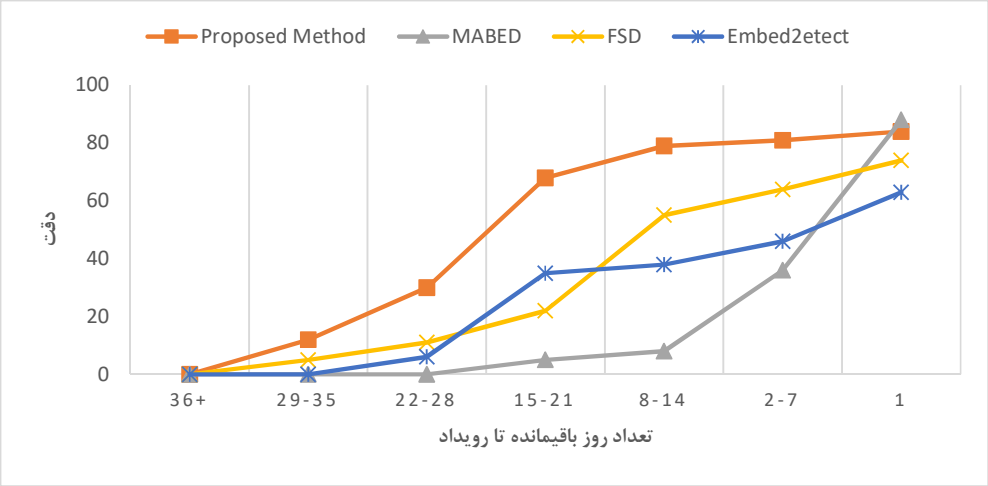
در شکل ۴-۴ معیار F1 برای روش پیشنهادی و سه روش تشخیص رویداد دیگر آورده شده است.



شکل ۴-۴) مقایسه معیار F1 روش پیشنهادی با سه روش تشخیص رویداد دیگر

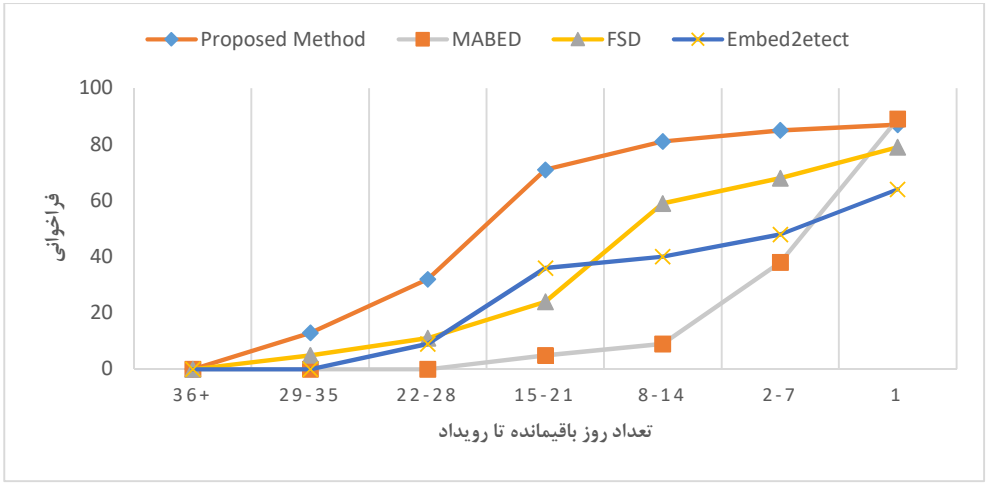
نتیجه محاسبه معیار دقت برای روش پیشنهادی و سه روش تشخیص رویداد دیگر در شکل ۴-۵ آورده شده است.

[44] Mention Anomaly Based Event Detection
[45] First Story Detection



شکل ۴-۵) مقایسه معیار دقت روش پیشنهادی با سه روش تشخیص رویداد دیگر

نهایتا معیار فراخوانی برای روش پیشنهادی و سه روش تشخیص رویداد دیگر در شکل ۴-۶ به تصویر کشیده شده است.



شکل ۴-۶) مقایسه معیار فراخوانی روش پیشنهادی با سه روش تشخیص رویداد دیگر

بالاتر بودن معیارهای ارزیابی روش پیشنهادی در مقایسه با سه روش دیگر، از ۳۵ روز به وقوع رویداد در نمودار مشخص است. روش پیشنهادی در سه هفته مانده به وقوع رویداد با دقت ۷۱ درصد موفق به پیش‌بینی رویداد شده است که دقت در دو هفته و یک هفته به وقوع رویداد به ترتیب به ۷۹ و ۸۱ درصد افزایش می‌یابد. در روز منتهی به وقوع رویداد، که بیشتر به رویدادهای ناگهانی اختصاص دارد،

دقت روش پیشـــنهادی در پیش‌بینی و تشـــخیص رویداد به ۸۴ درصـــد افزایش می‌یابد. البته جهش افزایش دقت در سه روش تشخیص رویداد دیگر خصوصا روش MABED که بر مبنای افزایش بی نظمی عمل می‌کند نیز در روز منتهی به رویداد، مشهود است.

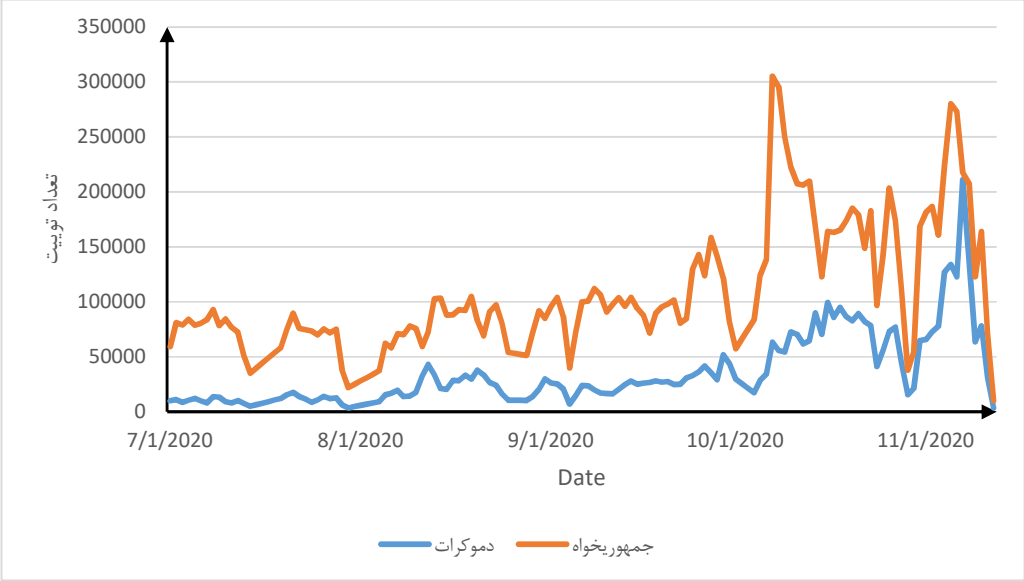
۴-۴ آزمایشات مرتبط با پیش‌بینی نتیجه انتخابات

در این بخش آزمایشات متعددی بر روی مجموعه داده‌ای دیگر از توییتر آورده شده است. مجموعه داده از [۷۸] گرفته شده است. این مجموعه داده شامل تقریبا ۲۴ میلیون توییت مرتبط با انتخابات ریاست جمهوری سال ۲۰۲۰ ایالت متحده آمریکا می‌باشد که در بازه زمانی اول جولای تا ۱۲ نوامبر گردآوری شده است. هر توییت بر اساس کلمات کلیدی مورد استفاده در آن مانند *TheDemocrats*، *JoeBiden* *realDonaldTrump* یا *Keep America Great* به دو حزب دموکرات یا جمهوریخواه نسبت داده شده‌اند. البته به برخی توییتها از آنجاییکه کلمات کلیدی مرتبط با هر دو حزب وجود داشته است، برچسب *both* زده شده است که در روش پیشنهادی مورد استفاده قرار نگرفته‌اند. در مجموعه داده مذکور با استفاده از روش تحلیل احساسات *VADER*[۴۶][۷۹]، به هر توییت یک نمره در بازه *[-1..+1]* تخصیص داده شده است. مقدار ۱+ احساسات مثبت شدید و ۱- احساسات منفی شدید را نشان می‌دهد. روشهای تحلیل احساسات به طور کلی به سه دسته روشهای یادگیری ماشین، روشهای مبتنی بر واژگان و روشهای ترکیبی تقسیم می‌شوند. *VADER* در واقع یک ابزار تحلیل احساسات مبتنی بر قاعده و واژه‌نامه است. به عبارت دیگر روش تحلیل احساسات *VADER* از یک لغتنامه که کلمات آن بر اساس شدت احساسی که دارند بصورت تجربی وزندهی شده‌اند، به همراه تعدادی قواعد نحوی و دستوری زبان که توسط انسانها هنگام بیان یا تاکید بر شدت احساسات از آنها استفاده می‌کنند، ایجاد شده است. بنابراین *VADER* بر مبنای نوع کلمات و قواعد دستوری که در یک توییت گنجانده شده است، نمره احساسی

[46] Valence Aware Dictionary and sEntiment Reasoner

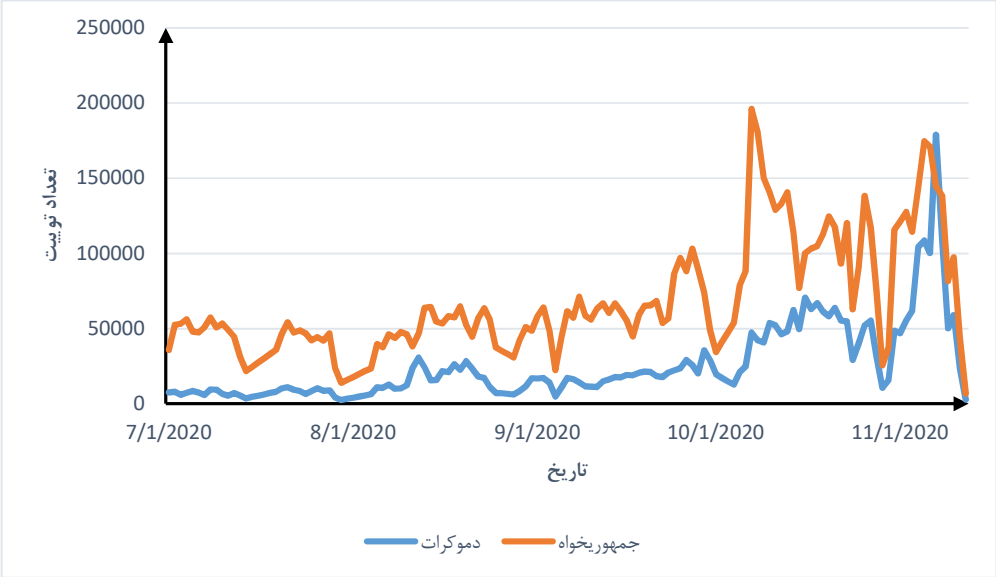
نهفته در آن را بصورت یک عدد در بازه *[-1..+1]* مشخص می‌کند. روش *VADER* به دلیل عملکرد خوبی که دارد، در حوزه تحلیل احساسات متون رسانه‌های اجتماعی زیاد مورد استفاده قرار گرفته است.

در شکل ۴-۷، تغییرات تعداد توییتهای ارسالی دو حزب در پنجره زمانی یک روز آمده است. با توجه به نتیجه نهایی انتخابات ریاست جمهوری ۲۰۲۰ آمریکا، این شکل نشان می‌دهد که تعداد توییتهای طرفداران یک حزب خاص، به تنهایی نمی‌تواند شاخص معتبری در پیش‌بینی نتیجه انتخابات باشد و باید از ویژگیهای دیگر نهفته در هر توییت نیز بهره برد.

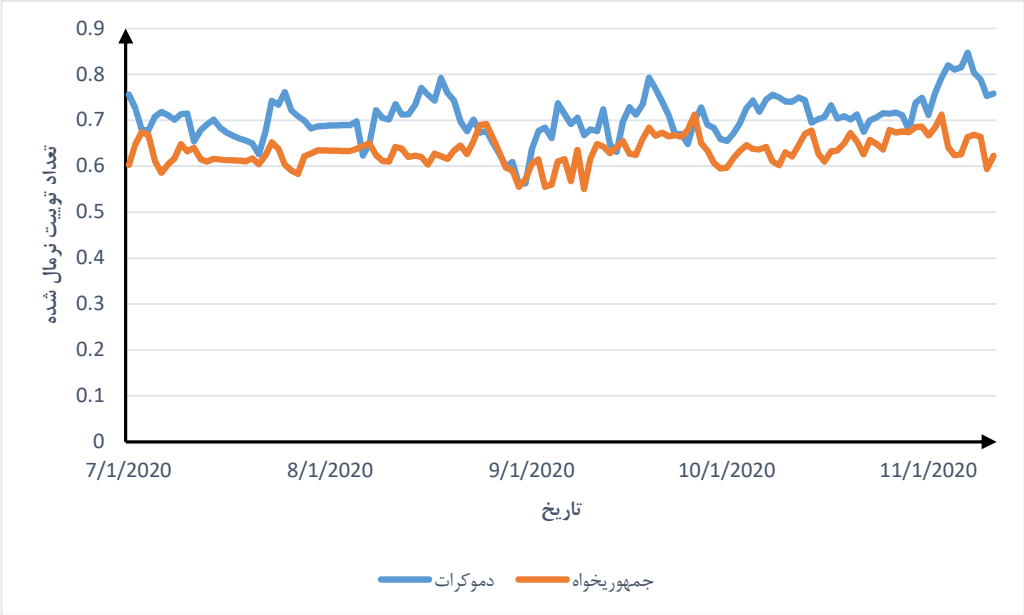


شکل ۴-۷) تعداد توییتهای ارسالی مرتبط به دو حزب دموکرات و جمهوریخواه در بازه زمانی یک روز

شکل ۴-۸ تعداد توییتهای ارسالی با بار احساسی مثبت مربوط به دو حزب را در پنجره زمانی یک روزه نشان می‌دهد. با توج به شکل ۴-۸، تنها در نظر گرفتن تعداد توییتهای احساسی نیز نمی‌تواند در پیش‌بینی نتیجه موثر باشد. چنانچه بر روی تعداد توییتهای احساسی مثبت که به شبکه ارسال شده‌اند، عمل نرمال‌سازی انجام شود، نتیجه حاصل از لحاظ همخوانی با نتیجه نهایی انتخابات که حزب دموکرات پیروز نهایی بوده است، جالب توجه می‌باشد (شکل ۴-۹)



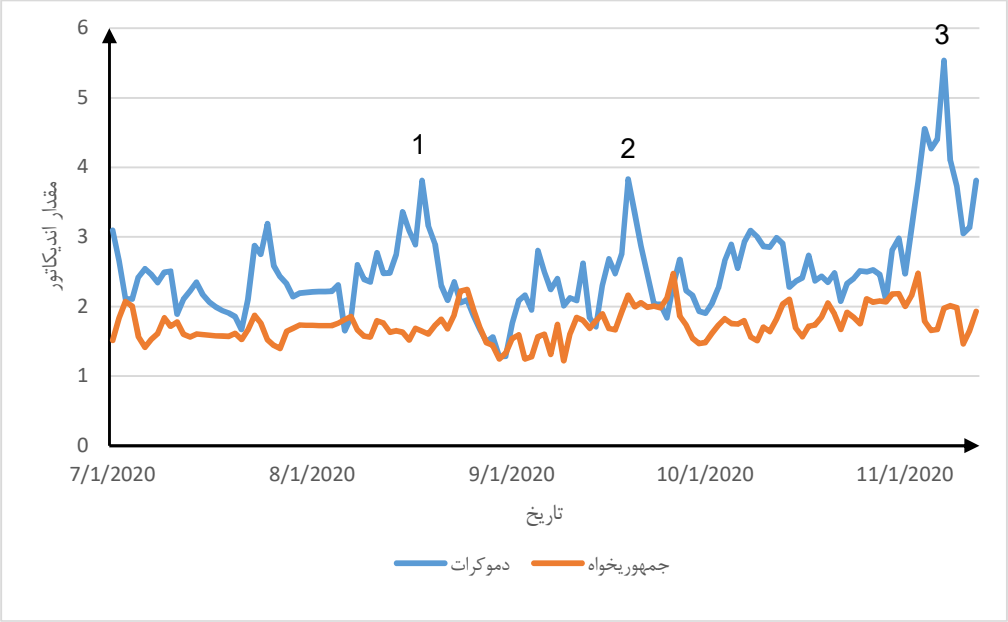
شکل ۴-۸) تعداد توییتهای روزانه با بار احساسی مثبت برای دو حزب دموکرات و جمهوریخواه



شکل ۴-۹) تعداد توییتهای مثبت روزانه دو حزب دموکرات و جمهوریخواه بصورت نرمال شده

بر اساس آنچه در فصل چهارم مربوط به روش پیشنهادی برای پیش‌بینی نتیجه انتخابات تشریح شد، از آنجاییکه توییتهای با بار احساسی مثبت نشانه‌ای از افزایش محبوبیت یک حزب و توییتهای با بار احساسی منفی نشانه‌ای از کاهش آن است، در اینجا نسبت پیامهای مثبت به منفی برای هر حزب

محاسبه شده است. در شکل ۴-۱۰ برای پنجره زمانی یک روزه، نسبت تعداد پیامهای احساسی مثبت به منفی برای دو حزب رسم شده است. در مقایسه با شکل ۴-۹، اختلافها بین دو نمودار احزاب بیشتر و قلهها و درهها واضحتر شده است. در این شکل می‌توان هر یک از قلهها و درههای موجود را به یک رویداد واقعی، مرتبط دانست. بعنوان مثال قله شماره ۱ بعد از معرفی خانم کامالا هریس بعنوان نامزد پست معاون اولی توسط جو بایدن رخ داده است. حمایت ترامپ از پیروان نظریه راست افراطی بعنوان یک حرکت جنجالی و نادرست از سوی او منجر به ایجاد قله شماره ۲ گردیده است. قله شماره ۳ مربوط به اعلام پیروزی جو بایدن توسط اسوشیتدپرس پس از انتخابات است. این آزمایش توانایی اندیکاتور معرفی شده برای پیش‌بینی انتخابات را نشان می‌دهد. حال با استفاده از روش تخمین سالمندی، امکان پیش‌بینی نتیجه انتخابات برای بازه‌های زمانی دلخواه قبل از آن فراهم می‌شود.



شکل ۴-۱۰) نسبت تعداد توییت مثبت به منفی روزانه دو حزب دموکرات و جمهوریخواه. قله شماره ۱ بعد از معرفی خانم کامالا هریس بعنوان نامزد پست معاون اولی توسط جو بایدن رخ داده است. حمایت ترامپ از پیروان نظریه راست افراطی بعنوان یک حرکت جنجالی و نادرست از سوی او منجر به ایجاد قله شماره ۲ گردیده است. قله شماره ۳ مربوط به اعلام پیروزی جو بایدن توسط اسوشیتدپرس پس از انتخابات است.

در آزمایش دیگری، دقت پیش‌بینی مقدار اندیکاتور در بازه‌های زمانی مختلف و تاثیر مقادیر مختلف α بررسی شده‌اند. جدول ۴-۳ مقدار پیش‌بینی اندیکاتور بر اساس روش سالمندی در بازه‌های زمانی مختلف به همراه معیار ارزیابی دقت را نشان می‌دهد. دقت پیش‌بینی مقدار اندیکاتور مربوط به هر کدام از احزاب بر اساس رابطه ۴-۴ بدست آمده است.

$$Accuracy = 1 - mean\left[\frac{abs(A - O)}{O}\right] \qquad 4-4$$

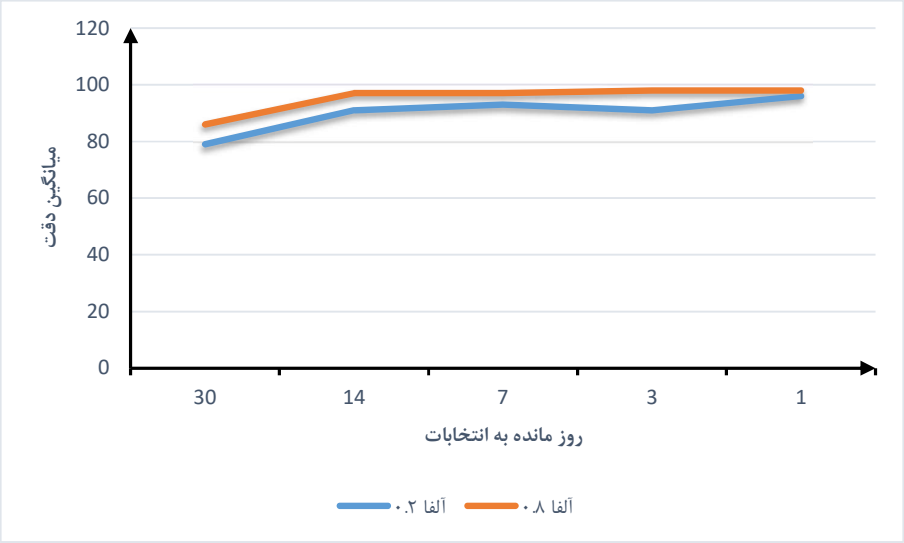
در رابطه ۴-۴، A مقادیر پیش‌بینی شده و O مقادیر مشاهده شده می‌باشد. میانگین دقت به دست آمده برای احزاب، در جدول ۴-۳ آمده است.

جدول ۴-۳) پیش‌بینی اندیکاتور در پنجره‌های زمانی مختلف در ۴ ماه

طول پنجره زمانی	مقدار آلفا	اندیکاتور دموکرات	اندیکاتور جمهوری‌خواه	میانگین دقت
۱۴ روز	۰,۸	۲,۴۷	۱,۹۴	۰,۹۷
	۰,۲	۲,۶۰	۱,۸۵	۰,۹۱
۷ روز	۰,۸	۲,۶۷	۲,۰۶	۰,۹۷
	۰,۲	۲,۵۶	۱,۹۸	۰,۹۳
۷ روز	افزایش تدریجی	۲,۷	۱,۹۸	۰,۹۶
یک روز	۰,۸	۳,۶۴	۲,۴	۰,۹۸
	۰,۲	۳,۲۹	۲,۱۵	۰,۹۶

با توجه به جدول ۴-۳، در هر پنجره زمانی برای مقادیر بزرگتر α دقت پیش‌بینی مقدار اندیکاتور بیشتر می‌باشد و هر چه طول پنجره زمانی کوتاه‌تر درنظر گرفته شود میزان دقت اغلب افزایش می‌یابد. در جدول ۴-۳ بعنوان مثال ۱۴ روز قبل از انتخابات، بیشتر بودن مقدار اندیکاتور دموکراتها از جمهوریخواهان با میانگین دقت ۹۷ درصد پیش‌بینی شده است. به عبارت دیگر، دو هفته قبل از

انتخابات، پیش‌بینی شـده اسـت که در ۱۴ روز بعدی نسـبت تعداد پیامهای مثبت به منفی دموکراتها بیشـــتر از جمهوریخواهان خواهد بود و این نشـــانه‌ای قوی از پیروزی حزب دموکرات در انتخابات می‌باشد. در ردیف سوم جدول ۴-۳، تاثیر انتخاب مقدار α افزایشی بر اندیکاتور نشان داده شده است. برای این منظور از چهار ماه به انتخابات تا یک هفته به آن، مقدار α از ۰,۲ تا ۰,۹ افزایش یافته ا ست. دقت پیش‌بینی نسبت به حالتی که α کوچک و ثابت انتخاب می شود بهتر می‌با شد. در شکل ۴-۱۱، روند افزایشـی دقت پیش‌بینی اندیکاتور نسـبت به روزهای باقیمانده به انتخابات آورده شـده اسـت. از اینرو هر چه تاریخ برگزاری انتخابات نزدیکتر می‌شـــود، مقدار اندیکاتور با دقت بهتری پیش‌بینی می‌شود.



شکل ۴-۱۱) دقت پیش‌بینی مقدار اندیکاتور در روزهای مانده به انتخابات

نهایتا در یک آزمایش دیگر، مقدار اندیکاتور، تنها با بررســـی داده‌های مربوط به یک ماه منتهی به انتخابات بررسی شده است. نتایج در جدول ۴-۴ آمده است.

جدول ۴-۴) پیش‌بینی اندیکاتور در پنجره‌های زمانی مختلف در یک ماه منتهی به انتخابات

طول پنجره زمانی	مقدار آلفا	اندیکاتور دموکرات	اندیکاتور جمهوری‌خواه	دقت
۱۴ روز	۰,۸	۲,۵۲	۱,۸۹	۰,۹۴
	۰,۲	۲,۵۹	۱,۷۶	۰,۸۹
۷ روز	۰,۸	۲,۶۷	۲,۰۵	۰,۹۷
	۰,۲	۲,۴۳	۱,۸	۰,۹۳
یک روز	۰,۸	۳,۶۴	۲,۴	۰,۹۸
	۰,۲	۲,۹	۲,۱۴	۰,۹۱

آنچه از جدول ۴-۴ قابل ا ستنباط ا ست این ا ست که برای داده‌های مربوط به ماه منتهی به انتخابات که رویدادهای انتخاباتی افزایش می‌یابد، برای داشـــتن یک پیش‌بینی با دقت خوب، انتخاب مقدار α کوچک نقش موثری خواهد داشت.

با توجه به آزمای شات انجام شده و جداول ۴-۳ و ۴-۴، که مقدار پیش‌بینی شده اندیکاتور برای حزب دموکرات بیشـــتر از جمهوری‌خواه می‌باشـــد، پیروزی جو بایدن در انتخابات ۲۰۲۰ ریاســـت جمهوری آمریکا قابل پیش‌بینی بوده است.

در ادامه، روش پیشـــنهادی برای پیش‌بینی انتخابات، با ســـه روش [۷۵-۷۳] که در فصـــل دوم معرفی شدند، مقای سه شده ا ست. برای این آزمایش، علاوه بر مجموعه داده [۷۸] که در بخش قبل تو ضیح داده شد، از دو مجموعه داده [۶۸] و [۸۰] نیز استفاده شده است. مجموعه داده دوم شامل تقریباً ۱۱ میلیون توییت مربوط به انتخابات ریاســـت جمهوری ایالات متحده در ســـال ۲۰۲۰ اســـت که از ۱ سپتامبر ۲۰۲۰ تا ۲ نوامبر ۲۰۲۰ جمع آوری شده ا ست. خلا صه آماری این مجموعه داده در جدول ۴-۵ آورده شده است.

جدول ۴-۵ خلاصه آماری مجموعه داده دوم

	تعداد کل توییت	مثبت	منفی	خنثی
دموکراتها	۳۸۵۱۲۹۳	۱۶۶۳۳۷۳	۱۶۳۹۴۹۵	۵۴۸۴۲۴
جمهوریخواهان	۷۱۰۹۹۴۱	۲۸۳۱۱۷۹	۳۷۹۱۷۳۲	۴۸۷۰۳۱
جمع کل	۵۴۳۱۲۲۷	۴۴۹۴۵۵۲	۵۴۳۱۲۲۷	۱۰۳۵۴۵۵

مجموعه داده سوم شامل تقریبا ۲۶ میلیون توییت در مورد انتخابات سال ۲۰۱۶ آمریکا ا ست که در بازه سی‌ام آگوست تا یازدهم نوامبر (روز انتخابات) جمع آوری شده است. در انتخابات سال ۲۰۱۶ اتفاقی نادر افتاد و علیرقم اینکه تعداد آراء حزب دموکرات (هیلاری کلینتون) بیشتر از تعداد آراء حزب جمهوریخواه (دونالد ترامپ) بود، اما این جمهوریخواهان بودند که پیروز انتخابات شدند. از این رو از این مجموعه داده به منظور بررسی مقاومت پذیری روش پیشنهادی استفده شده است.

جدول ۴-۶ خروجی روش پیشنهادی را با سه روش مذکور بر اساس مجموعه داده اول، برای دوره‌های زمانی مختلف قبل از انتخابات نشان می‌دهد. همانطور که در جدول ۴-۶ مشاهده می‌شود، روش پیشنهادی بر خلاف سه روش دیگر، قادر به پیش‌بینی دقیق نتیجه در تمامی دوره‌های زمانی قبل از انتخابات بوده است. جدول ۴-۷ خروجی روش پیشنهادی را با سه روش دیگر برای مجموعه داده دوم در روز قبل از انتخابات را نشان می‌دهد.

جدول ۴-۶ مقایسه خروجی روش پیشنهادی با سه روش دیگر بر اساس مجموعه داده اول. اعداد جدول در واقع مقادیر شاخص در روشهای مختلف است که بر مبنای آن نتیجه انتخابات پیش‌بینی می‌شود.

	نام حزب	تعداد روزهای مانده به انتخابات			
		۳۰	۱۴	۷	۱
روش اول [۷۰]	دموکرات	۰,۰۷۶	۰,۰۸۴	۰,۰۹۱	۰,۰۹۴
	جمهوریخواه	**۰,۲۰۷**	**۰,۲۰۶**	**۰,۲۰۳**	**۰,۲۰۶**
روش دوم [۷۱]	دموکرات	۰,۱۳۵	۰,۱۴۴	۰,۱۵۶	۰,۱۶۱
	جمهوریخواه	**۰,۵۰۶**	**۰,۵۰۳**	**۰,۴۹۰**	**۰,۴۸۹**
روش سوم [۷۲]	دموکرات	۰,۴۳۴	۰,۴۳۸	۰,۴۴۳	۰,۴۴۳
	جمهوریخواه	**۰,۵۶۵**	**۰,۵۶۱**	**۰,۵۵۶**	**۰,۵۵۶**
روش پیشنهادی	**دموکرات**	**۲,۰۷**	**۲,۴۷**	**۲,۶۷**	**۳,۶۴**
	جمهوریخواه	۱,۷۴	۱,۹۴	۲,۰۶	۲,۴

جدول ۴-۷ مقایسه خروجی روش پیشنهادی با سه روش دیگر بر اساس مجموعه داده دوم

حزب	روش اول [۷۰]	روش دوم [۷۱]	روش سوم [۷۲]	روش پیشنهادی
دموکرات	**۰,۰۰۲**	۰,۱۷۷	**۰,۴۹۸**	**۱,۰۱۴**
جمهوریخواه	-۰,۰۸	**۰,۲۷۷**	۰,۴۰۸	۰,۷۴۷

در جدول ۴-۸، خروجی روش پیشنهادی و سه روش دیگر بر اساس مجموعه داده سوم (انتخابات ۲۰۱۶ آمریکا) آورده شده است.

جدول ۴-۸) مقایسه خروجی روش پیشنهادی با سه روش دیگر بر اساس مجموعه داده سوم. اعداد جدول در واقع مقادیر شاخص در روشهای مختلف است که بر مبنای آن نتیجه انتخابات پیش‌بینی می‌شود.

	نام حزب	تعداد روزهای مانده به انتخابات		
		۱۴	۷	۱
روش اول [۷۰]	**دموکرات**	**-۰,۱۶**	**-۰,۱۷**	**-۰,۱۸**
	جمهوریخواه	-۰,۲۶	-۰,۲۶	-۰,۲۵
روش دوم [۷۱]	دموکرات	۰,۱	۰,۱	۰,۱۱
	جمهوریخواه	**۰,۱۸**	**۰,۱۷**	**۰,۱۷**
روش سوم [۷۲]	**دموکرات**	**۰,۵۵**	**۰,۵۴**	**۰,۵۳**
	جمهوریخواه	۰,۴۵	۰,۴۶	۰,۴۶
روش پیشنهادی	دموکرات	۰,۳۷	۰,۳۷	۰,۳۸
	جمهوریخواه	**۰,۴۰**	**۰,۴۰**	**۰,۴۱**

در جداول ۴-۶، ۴-۷ و ۴-۸ پیش بینی نتیجه انتخابات برای هر روش بصورت پر رنگ نشان داده شده است. روش پیشنهادی برای هر سه مجموعه داده موفق به پیش بینی صحیح در بازه‌های مختلف شده است. این در حالی است که روش دوم تنها در مجموعه داده سوم و روش‌های اول و سوم تنها در مجموعه داده دوم موفق به پیش بینی صحیح انتخابات شده اند.

۴-۵ جمع‌بندی

در این فصل، روش پیشنهادی برای پیش‌بینی رویداد و پیش‌بینی نتیجه انتخابات مورد ارزیابی قرار گرفت. ابتدا رفتار شبکه اجتماعی قبل از وقوع رویدادهای قابل پیش‌بینی و رویدادهای ناگهانی بررسی شد. مشخص گردید که برای رویدادهای قابل پیش‌بینی از روزها قبل، پیامهایی در مورد آنها به شبکه ارسال می‌شود و هرچه به روز وقوع نزدیکتر می‌شود، تعداد پیامهای ارسالی در مورد آنها به شبکه

افزایش می‌یابد، تا اینکه در روز واقعه به اوج خود می‌رسد. در آزمایش دیگری، عملکرد روش پیشنهادی با دو روش تشـــخیص رویداد مقایســـه گردید و برتری دقت روش پیشـــنهادی در پیش‌بینی وقوع رویدادهای قابل پیش‌بینی، مشخص گردید.

برای روش پیشنهادی در مورد پیش‌بینی انتخابات نیز آزمایشات متنوعی انجام شد و مشخص گردید که تعداد توییتهای طرفداران احزاب نمی‌تواند به تنهایی نشانه مناسبی برای پیش‌بینی باشد و نیاز به ویژگیهای دیگری از توییتها است. برای این منظور از احساسات نهفته در توییتها استفاده شد. سپس نشان داده شد که اندیکاتور پیشنهادی می‌تواند شاخص خوبی برای پیش‌بینی نتیجه انتخابات باشد و با روش تخمین سالمندی می‌توان مقدار این اندیکاتور را برای بازه‌های بعدی پیش‌بینی نمود.

در نهایت روش پیشنهادی با سه روش دیگر در حوزه پیش‌بینی انتخابات مقایسه شد و توانایی آن در پیش‌بینی نتیجه انتخابات برای دو مجموعه داده متفاوت نشان داده شد.

۵ جمع بندی و کارهای آینده

۵-۱) خلاصه و نتیجه‌گیری

امروزه شبکه‌های اجتماعی مجازی، به شدت بین جوامع بشری نفوذ یافته است. بطور خاص در دو سال گذشته و با شیوع بیماری کووید ۱۹، شبکه‌های اجتماعی مجازی نقش مهمی در جهت اطلاع‌رسانی، پیشگیری و آموزش داشته‌اند. کاربردهای متعددی در حوزه تحلیل شبکه‌های اجتماعی تعریف می‌شوند که تشخیص و پیش‌بینی رویداد یکی از مهمترین آنها می‌باشند. در این رساله تلاش شد تا با تحلیل شبکه اجتماعی توییتر، روشی برای پیش‌بینی رویداد معرفی شود.

در تعریفی که از رویداد در این رساله آمد، رویداد به چیزی گفته می‌شود که تغییرات چشمگیری در برخی پارامترهای شبکه اجتماعی ایجاد کند. رویدادها نیز بر اساس مبنای پیدایش آنها به دو دسته قابل پیش‌بینی و غیر قابل پیش‌بینی یا ناگهانی تقسیم شدند. پارامتری که در روش پیشنهادی برای پیش‌بینی وقوع رویداد مورد بررسی قرار گرفت، نرخ ارسال توییتهای مرتبط با یک رویداد به شبکه اجتماعی می‌باشد. ایده اصلی کار بدین گونه است که، چنانچه تغییرات نرخ ارسال توییتهای مرتبط با یک رویداد به شبکه اجتماعی در محدود یکسری حدود آستانه قرار گیرد، می‌توان آن را نشانه‌ای برای وقوع آن در آینده دانست. علاوه بر این اگر این تغییرات بسیار شدید و بالاتر از یک حد آستانه دیگر باشد، می‌توان آن را نشانه وقوع یک رویداد ناگهانی یا غیرقابل پیش‌بینی در نظر گرفت.

در پیاده سازی روش پیشنهادی، در مرحله نخست، ابتدا توییتها بر اساس یک پنجره زمانی با طول ثابت دسته بندی می شوند و سپس عمل پیش‌پردازش به منظور پاک سازی داده‌ها بر روی هر دسته اعمال می‌شود. در مرحله دوم، خوشه‌بندی توییتها انجام می‌گیرد. برای این منظور در ابتدا به جهت ایجاد خوشه‌های اولیه و مقداردهی آنها، از الگوریتم NMF استفاده شده است. بعد از اینکه خوشه‌های اولیه تنها با استفاده از گروه نخست داده‌ها ایجاد شدند، گروه‌های بعدی داده‌ها به ترتیب بر اساس الگوریتم ddCRP خوشه‌بندی می‌شوند. در حین فرایند خوشه‌بندی، میزان تغییرات رشد اندازه هر خوشه مورد نظارت قرار می‌گیرد و چنانچه در محدوده حدود آستانه مورد نظر قرار گرفتند، k کلمه

پرتکرار خوشـه موردنظر بعنوان توصـیفی از رویداد پیش‌بینی شـده به خروجی ارسـال می‌شـود. روش پیشنهادی موفق به پیش‌بینی ۸۵ درصد کل رویدادهای قابل پیش‌بینی بوده است.

علاوه بر این، ویژگی نرخ ارسال توییت‌ها برای پیش‌بینی یک رویداد خاص نیز مورد بررسی قرار گرفت. برای این منظور رویداد انتخابات و پیش‌بینی نتیجه آن با اسـتفاده از تحلیل شـبکه اجتماعی توییتر انتخاب گردید. در روش پیشنهادی برای پیش‌بینی نتیجه انتخابات، نسبت نرخ ارسال توییت‌های با بار احساسی مثبت به تعداد توییت‌های با بار احساسی منفی در یک پنجره زمانی با طول ثابت به عنوان یک اندیکاتور معرفی شـد و با اسـتفاده از روش تخمین سـالمندی، مقدار این اندیکاتور در پنجره‌های زمانی بعدی برای هر کدام از احزاب شرکت کننده در انتخابات پیش‌بینی شد. روش پیشنهادی بر روی داده‌های توییتر مربوط به انتخابات ریاسـت‌جمهوری ۲۰۲۰ ایالات متحده آمریکا به عنوان مطالعه موردی مورد ارزیابی قرار گرفت. از آنجایی که مقدار اندیکاتور پیشـنهادی، برای دموکرات‌ها بیشـتر از جمهوری خواهان بود، پیروزی دموکراتها قابل پیش‌بینی بوده است.

۵-۲) پیشنهادات ادامه کار

با توجه به مطالعات و کارهای انجام شـده در زمینه پیش‌بینی رویداد با اسـتفاده از تحلیل شـبکه‌های اجتماعی، موارد زیر جهت پژوهش در این زمینه پیشنهاد می‌شود:

۱- **تعیین زمان تقریبی وقوع رویدادهای پیش‌بینی شده**: روش پیشـنهادی در این رسـاله، تنها وقوع رویدادهای آتی را پیش‌بینی می‌کند و زمان وقوع آنها را مشخص نمی‌کند. در فرایند خوشـه‌بندی افزایشـی روش پیشـنهادی، آن دسـته از خوشـه‌ها که بیانگر رویدادهای قابل پیش‌بینی بودند و کلمات پرتکرار آنها بعنوان خروجی انتخاب می‌شـدند، سـرعت رشـدهای متفاوتی دارند. می‌توان بر اسـاس میزان رشـد این خوشـه‌ها نیز زمان تقریبی وقوعشـان را به صورت رویدادهای کوتاه مدت، میان مدت و بلند مدت پیش‌بینی نمود. لازم به ذکر است، با

توجه به تحلیل‌های انجام شـــده، پیش‌بینی زمان دقیق وقوع رویدادهای آینده غیر ممکن به نظر می‌رسد.

۲- **اســـتفاده از ویژگیهای دیگر شـــبکه اجتماعی همچون هشـــتگ‌ها، یادکردها، تعداد بازتوییت، تعداد لایک و افراد موثر به منظور بهبود سرعت و دقت روش پیـشنهادی**: هشـــتگ‌ها حاوی اطلاعات خیلی مفید در توییتها هســـتند. در برخی پژوهش‌ها بر مبنای آنها خو شه‌بندی انجام داده‌اند و بعنوان فاز نمایش رویدادها نیز مورد ا ستفاده قرار گرفته‌اند [۸۲-۸۱]. در واقع هشـــتگ‌ها را می‌توان موضـوع هر توییت در نظر گرفت. بنابراین می‌توانند در فاز خوشـه‌بندی و نمایش رویدادها بسـیار موثر باشـند. علاوه بر این چنانچه تعداد توییتهای افراد موثر مورد بررســـی قرار گیرد، علاوه بر فیلترکردن حجم زیادی از توییتها، وجود توییتهای جعلی به حد زیادی کاهش و دقت پیش‌بینی افزایش می‌یابد. رو شهای زیادی برای شنا سایی افراد موثر پیـشنهاد شده ا ست که نسبت تعداد دنبال‌کنندگان به دنبال شوندگان از آن جمله هستند. هر چه تعداد دنبال‌کنندگان یک فرد بیشتر از دنبال‌شوندگانش باشند، موثر بودن فرد بیشـــتر خواهد بود و توییتهای قابل اعتمادتری ایجاد می‌کند. تعداد بازتوییتها و لایک‌ها نیز می‌توانند بعنوان یک فاکتور در تعیین میزان اهمیت توییتها استفاده شوند.

۳- **پیش‌بینی پیا مد های یک رو یداد**: معمولا پس از وقوع رو یداد ها، افراد موثر با ارســـال پیامهایی به شـبکه اجتماعی عکس العملهایی انجام می‌دهند و برخی افراد شـبکه را گرد خود جمع می‌کنند. به عبارت دیگر تشکیل انجمن می‌دهند. تحلیل توییتهایی که در این انجمن‌ها رد و بدل می‌شـوند، می‌تواند وقوع رویدادهایی بعد از رویداد اصـلی را شـناسـایی یا پیش‌بینی کند.

۴- **دسته‌بندی توییتها بر اساس یک پنجره زمانی با طول متغیر**: روش پیشنهادی در مرحله پیش‌پردازش، توییتها را بر اســاس یک پنجره زمانی با طول ثابت دســته‌بندی می‌کند. پنجره زمانی با طول متغیر برای مواجهه با برخی انواع رویدادها می‌تواند مفید باشـد. برخی رویدادها

زودگذر و برخی دیگر ماندگاری بیشتری در شبکه اجتماعی دارند. بنابراین ایجاد یک پنجره با طول تطبیق‌پذیر می‌تواند موثرتر باشد.

۵- **نمایش خواناتر رویدادها**: در روش پیشنهادی، خروجی تنها بصورت یکسری کلمات پرتکرار در خو شه‌ها نمایش داده می شود. می‌توان خروجی را از دو منظر بهبود بخ شید. اولا خروجی بصـــورت یک جمله خوانا ایجاد شـــود و ثانیا اطلاعاتی در مورد اثرات وقوع رویداد پیش‌بینی شده، بر جوامع از نظر میزان اثر یا هشدار برای وقوع رویدادهای دیگر نیز تعیین گردد.

۶- **شنا سایی خودکار عوامل موثر در وقوع یک رویداد**: در واقع به جای اینکه همانند روش پیشـــنهادی، بر اســـاس پارامترهای شـــبکه اجتماعی به پیش‌بینی رویداد پرداخته شـــود، با شنا سایی رویدادهای جاری، رویداد آینده پیش‌بینی شوند. به عبارت دیگر جریانات دخیل در وقوع رویدادهای آتی شناسایی گردد. این پیشنهاد می‌تواند برای مواردی همچون پیشگیری از برخی رویدادها و بحرانهای اجتماعی مفید باشد.

۷- **ا ستفاده از منطق فازی در نمره اح سا سی توییتها**: در زمینه پیش‌بینی نتیجه انتخابات می‌توان از منطق فازی برای نمره احســاســی توییت‌ها اســتفاده کرد. به عبارت دیگر، در روش پیشـــنهادی توییت‌هایی با امتیاز احســـاســـی ۰,۱ و یک، هر دو به عنوان توییت‌های مثبت برچسب‌گذاری می‌شوند، در حالی که استفاده از منطق فازی می‌تواند مفید باشد.

۸- **در نظر گرفتن موقعیت مکانی ثبت شده برای توییتها در پیش‌بینی انتخابات الکترال محور**: از آنجایی که نتیجه نهایی انتخابات ریاســـت جمهوری ایالات متحده بر اســـاس آرای الکترال تعیین می‌شود و تعداد آرای الکترال برای هر دو حزب بر اساس آرای داده شده در هر ایالت تعیین می شود، پیش‌بینی تعداد آرای الکترال بر ا ساس تجزیه و تحلیل داده‌های توییتر به صورت محلی در هر ایالت می‌تواند نتایج دقیق‌تر و قابل اعتمادتری به همراه داشته باشد.

۹- **ا ستفاده از اندیکاتور پی شنهادی برای پیش‌بینی انتخابات در حوزه‌های دیگر**: بعنوان مثال برای پیش‌بینی قیمت سهام می‌توان از اندیکاتور پی‌شنهادی ا ستفاده کرد. البته به دلیل پیچیدگی داده‌ها در برخی حوزه‌ها همچون بازرهای مالی، احتمالاً باید از چندین شـــاخص مختلف بصورت ترکیب استفاده نمود.

۱۰- **اســـتفاده از انواع دیگر داده‌ها همچون تصـــویر برای پیش‌بینی رویداد**: علیرقم اینکه مجموعه داده‌های متنی فراوانتر و روشـــهای تحلیل شـــبکه‌های اجتماعی مبتنی بر متون بیشـــتر هســـتند و پردازش تصـاویر معمولا زمانبرتر هســـتند، اما بســـته به نوع رویداد می‌توان از انواع دیگر داده همچون تصویر استفاده نمود.

۱۱- **ایجاد مجموعه داده جدید**: یکی از مهمترین چالشـــهایی که در طول انجام رســـاله با آن مواجه بوده‌ایم، وجود یک مجموعه داده مناســـب بوده اســـت. بطوری که هم از نظر تعداد و هم از نظر پوشش یک بازه زمانی حداقل چهار ماهه قابل قبول باشد. از اینرو یکی از کارهای بسیار خوبی که می‌توان در ادامه این مسیر پژوهشی درنظر گرفت، فراهم ساختن یک مجموعه داده مناسب و حتی از توییت‌های فارسی است.

۶ مراجع

1. Wasserman, S., and Galaskiewicz, J., (1994), "*Advances in social network analysis: Research in the social and behavioral sciences",* SAGE Publications, Doi: 10.4135/9781452243528
2. Obar, J.A. and Wildman, S., (2015), "Social media definition and the governance challenge: An introduction to the special issue", *Telecommunications Policy*, 39(9), 745-750.
3. Deitrick, W. and Hu, W., (2013), "Mutually Enhancing Community Detection and Sentiment Analysis on Twitter Net-works", *Journal of Data Analysis and Information Processing*, 1, 19-29.
4. Panagiotou, N., Katakis, I., & Gunopulos, D., (2016), "Detecting events in online social networks: Definitions, trends and challenges", In *Solving Large Scale Learning Tasks. Challenges and Algorithms,* (pp. 42-84). Springer.
5. Ramya, G. R., and P. Bagavathi Sivakumar. (2021), "An incremental learning temporal influence model for identifying topical influencers on Twitter dataset", *Social Network Analysis and Mining* 11(1), 1-16.
6. Sohail, S. S., (2021), "Crawling Twitter data through API: A technical/legal perspective." *arXiv preprint arXiv:2105.10724*.
7. Li, R., et al., (2020), "Few-shot learning for new user recommendation in location-based social networks." *Proceedings of The Web Conference 2020*.
8. Rout, R., Greeshma L., and Durvasula S., (2020), "Detection of malicious social bots using learning automata with url features in twitter network." *IEEE Transactions on Computational Social Systems* 7.4, 1004-1018.
9. Su, C., Guan, X., Du, Y., Wang, Q., & Wang, F. (2018). A fast multi-level algorithm for community detection in directed online social networks. *Journal of Information Science*, *44*(3), 392-407.
10. Tahmasebi, H., Ravanmehr, R. and Mohamadrezaei, R., (2021), "Social movie recommender system based on deep autoencoder network using Twitter data." *Neural Computing and Applications* 33.5, 1607-1623.
11. Yang, J., Wu, Y., (2021), "An approach of Bursty event detection in social networks based on topological features", *Appl Intell* , Doi: 10.1007/s10489-021-02729-0.
12. Daud, N., Siti, H., Muntadher, S., Firdaus, A., (2020), "Applications of link prediction in social networks: A review". Journal of Network and Computer Applications, 166, Doi: 10.1016/j.jnca.2020.102716.
13. Birjali, M., Mohammed K., and Abderrahim B., (2021), "A comprehensive survey on sentiment analysis: Approaches, challenges and trends." *Knowledge-Based Systems*, 107134.
14. Allan, J., (2002), "Introduction to topic detection and tracking", *Topic Detection and Tracking* (pp. 1-16). Springer, Boston, MA.
15. Aggarwal, C. C., & Subbian, K. (2012). "Event detection in social streams". In *Proceedings of the 2012 SIAM international conference on data mining* (pp. 624-635). Society for Industrial and Applied Mathematics.
16. Becker, H., Naaman, M., & Gravano, L. (2011). "Beyond Trending Topics: Real-World Event Identification on Twitter". *11*(2011), 438-441.
17. McMinn, A. J., Moshfeghi, Y., & Jose, J. M. (2013). "Building a large-scale corpus for evaluating event detection on twitter". In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 409-418). ACM.
18. Weng, J., & Lee, B. S. (2011). "Event detection in twitter". *ICWSM*, *11*, 401-408.
19. Dou, W., Wang, X., Ribarsky, W., & Zhou, M. (2012). "Event detection in social media data". In *IEEE VisWeek Workshop on Interactive Visual Text Analytics-Task Driven Analytics of Social Media Content* (pp. 971-980).
20. Daly, E. M., & Geyer, W. (2011). "Effective event discovery: using location and social information for scoping event recommendations". In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 277-280). ACM.
21. Goswami, A., & Kumar, A. (2016). "A survey of event detection techniques in online social networks". *Social Network Analysis and Mining*, *6*(1), 107.
22. Nurwidyantoro, A., & Winarko, E. (2013). "Event detection in social media: A survey". In *ICT for Smart Society (ICISS), 2013 International Conference on* (pp. 1-5). IEEE.
23. AlKhwiter, W., and Nora A., (2021), "Part-of-speech tagging for Arabic tweets using CRF and Bi-LSTM." *Computer Speech & Language* 65, 101138.

24. Rajoria, L., (2021), "Named Entity Recognition in Tweets." *International Journal of Research in Engineering, Science and Management* 4.1, 43-50.
25. Sharma, S., et al. (2021), "Machine Learning Application: Sarcasm Detection Model." *Artificial Intelligence for a Sustainable Industry 4.0*. Springer, Cham, 125-138.
26. McMinn, A. J., & Jose, J. M., (2015). "Real-time entity-based event detection for twitter". In *International conference of the cross-language evaluation forum for european languages* (pp. 65-77). Springer, Cham.
27. Naseem, U., Imran R., and Peter W., (2020), "A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter." *Multimedia Tools and Applications*, 1-28.
28. Hasan, M., Mehmet A., and Rolf S., (2018), "A survey on real-time event detection from the Twitter data stream." *Journal of Information Science,* 44.4, 443-463.
29. Siegel, A., (2017), "Twitter Wars: Sunni-Shi a Conflict and Cooperation in the Digital Age." *Beyond Sunni and Shia*. Oxford University Press, 157-180.
30. Al-Rawi, A., Jacob G., and Li Z., (2018), "What the fake? Assessing the extent of networked political spamming and bots in the propagation of fakenews on Twitter." Online Information Review .
31. Blei, D. M., & Frazier, P. I. (2011). Distance dependent Chinese restaurant processes. Journal of Machine Learning Research, 12(8).
32. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.
33. Guille, A. and C. Favre, (2015) *Event detection, tracking, and visualization in twitter: a mention-anomaly-based approach.* Social Network Analysis and Mining, **5**(1): p. 18.
34. Li, C., Sun, A., & Datta, A., (2012), Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 155-164). ACM.
35. Gaglio, S., Re, G. L., & Morana, M. (2016). A framework for real-time Twitter data analysis. *Computer Communications*, *73*, 236-242.
36. Stilo, G., & Velardi, P. (2016). Efficient temporal mining of micro-blog texts and its application to event discovery. *Data Mining and Knowledge Discovery*, *30*(2), 372-402.
37. Zhang, X., Chen, X., Chen, Y., Wang, S., Li, Z., & Xia, J. (2015). Event detection and popularity prediction in microblogging. *Neurocomputing*, *149*, 1469-1480.
38. Srijith, P. K., et al., (2017), "Sub-story detection in Twitter with hierarchical Dirichlet processes." *Information Processing & Management* 53.4 (2017): 989-1003.
39. Li, J., Wen, J., Tai, Z., Zhang, R., & Yu, W. (2016). "Bursty event detection from microblog: a distributed and incremental approach". *Concurrency and Computation: Practice and Experience*, *28*(11), 3115-3130.
40. Cai, H., Yang, Y., Li, X., & Huang, Z. (2015). "What are popular: exploring twitter features for event detection, tracking and visualization". In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 89-98). ACM.
41. Petrović, S., M. Osborne, and V. Lavrenko. (2010). "*Streaming first story detection with application to twitter"*. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Aassociation for Computational Linguistics*.
42. Jafari, O., Maurya, P., Nagarkar, P., Islam, K. M., & Crushev, C. (2021). A survey on locality sensitive hashing algorithms and their applications. *arXiv preprint arXiv:2102.08942*.
43. Comito, C., Agostino F., and Clara P., (2019), "Bursty event detection in Twitter streams." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13.4, 1-28.
44. Choi, H., and Cheong H., (2019), "Emerging topic detection in twitter stream based on high utility pattern mining." *Expert systems with applications* 115, 27-36.
45. [28] Becker, H., Naaman, M., & Gravano, L. (2011). "Beyond trending topics: Real-world event identification on twitter". In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 5, No. 1).
46. De Boom, C., Van Canneyt, S., & Dhoedt, B. (2015). "Semantics-driven event clustering in twitter feeds". In Making Sense Of Microposts (Vol. 1395, pp. 2-9). CEUR.

47. Zhao, W. X., Jiang, J., He, J., Song, Y., Achanauparp, P., Lim, E. P., & Li, X. (2011). "Topical keyphrase extraction from twitter". In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 379-388).
48. Phuvipadawat, S., & Murata, T. (2010). "Breaking news detection and tracking in Twitter". In 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (pp. 120-123). IEEE.
49. McMinn, A. J., & Jose, J. M. (2015). "Real-time entity-based event detection for twitter". In International conference of the cross-language evaluation forum for european languages (pp. 65-77). Springer, Cham.
50. Hasan, M., Orgun, M. A., & Schwitter, R. (2019). "Real-time event detection from the Twitter data stream using the TwitterNews+ Framework". *Information Processing & Management*, *56*(3), 1146-1165.
51. De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000). "A multilinear singular value decomposition". *SIAM journal on Matrix Analysis and Applications*, *21*(4), 1253-1278.
52. Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). "Experimental explorations on short text topic mining between LDA and NMF based Schemes". *Knowledge-Based Systems*, *163*, 1-13.
53. Gershman, S. J., & Blei, D. M. (2012). "A tutorial on Bayesian nonparametric models". *Journal of Mathematical Psychology*, *56*(1), 1-12.
54. McMinn, A. J., & Jose, J. M. (2015). "Real-time entity-based event detection for twitter". In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 65-77). Springer, Cham.
55. Fu, X., Jiang, X., Qi, Y., Xu, M., Song, Y., Zhang, J., & Wu, X. (2020, August). "An event-centric prediction system for COVID-19". In 2020 IEEE International Conference on Knowledge Graph (ICKG) (pp. 195-202). IEEE.
56. Alkouz, B., Al Aghbari, Z., & Abawajy, J. H. (2019). "Tweetluenza: Predicting flu trends from twitter data". Big Data Mining and Analytics, 2(4), 273-287.
57. Huang, Y. T., & Pai, P. F. (2020). "Using the Least Squares Support Vector Regression to Forecast Movie Sales with Data from Twitter and Movie Databases". Symmetry, 12(4), 625.
58. Asur, S., & Huberman, B. A. (2010). "Predicting the future with social media". In 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (Vol. 1, pp. 492-499).
59. Asghar, M. Z., Khan, A., Khan, F., & Kundi, F. M. (2018). "Rift: a rule induction framework for twitter sentiment analysis". Arabian Journal for Science and Engineering, 43(2), 857-877.
60. Dhar, V., & Chang, E. A. (2009). "Does chatter matter? The impact of user-generated content on music sales". *Journal of Interactive Marketing*, *23*(4), 300-307.
61. Vossen, R. (2013). Does chatter matter? Predicting music sales with social media. [online] Availableat: https://www.basicthinking.de/blog/wpcontent/uploads/2013/06/Does-Chatter-Matter.pdf
62. Tumasjan, A., Sprenger, T.,Sandner, P., Welpe, I., (2010), "Predicting elections with twitter: What 140 characters reveal about political sentiment", In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 4, No. 1.
63. Ceron, A., Curini, L., Iacus, S. M., (2015), "Using sentiment analysis to monitor electoral campaigns: Method matters—evidence from the United States and Italy", Social Science Computer Review, Vol. 33, No. 1, 3-20, DOI: 10.1177/0894439314521983.
64. Burnap, P., Gibson, R., Sloan, L., Southern, R., Williams, M., (2016), "140 characters to victory?: Using Twitter to predict the UK 2015 General Election", Electoral Studies, Vol. 41, 230-233. DOI: 10.1016/j.electstud.2015.11.017.
65. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A., (2010), "Sentiment strength detection in short informal text", *Journal of the American Society for Information Science and Technology*, Vol. 61, No. 12, 2544-2558. DOI: 10.1002/asi.21416.
66. Nugroho, D. K., (2021), "US presidential election 2020 prediction based on Twitter data using lexicon-based sentiment analysis," 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 136-141. DOI: 10.1109/Confluence51648.2021.9377201.
67. Xia, E., Yue, H., & Liu, H., (2021), "Tweet Sentiment Analysis of the 2020 US Presidential Election", In *Companion Proceedings of the Web Conference 2021,* pp. 367-371. DOI: 10.1145/3442442.3452322.

68. Sabuncu, I., Balci, M. A., & Akguller, O., (2020), "Prediction of USA November 2020 Election Results Using Multifactor Twitter Data Analysis Method", *arXiv: 2010.15938*.
69. Baccianella, S., Esuli, A., Sebastiani, F., (2010), "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining", ***In Language Resources and Evaluation***, Vol. 10, pp. 2200-2204.
70. Ibrahim, M., Abdillah, O., Wicaksono, A. F., Adriani, M., (2015), "Buzzer detection and sentiment analysis for predicting presidential election results in a twitter nation", IEEE International Conference on Data Mining Workshop, 1348-1353. DOI: 10.1109/ICDMW.2015.113.
71. Sharma, P., Moh, T. S., (2016), "Prediction of Indian election using sentiment analysis on Hindi Twitter." IEEE International Conference on Big Data, 1966-1971. DOI: 10.1109/BigData.2016.7840818.
72. Liu, R., Yao, X., Guo, C., Wei, X., (2021), "Can We Forecast Presidential Election Using Twitter Data? An Integrative Modelling Approach", Annals of GIS, Vol. 27, No. 1, 43-56, DOI: 10.1080/19475683.2020.1829704.
73. Singh, P., Dwivedi, Y. K., Kahlon, K. S., Pathania, A., & Sawhney, R. S., (2020), "Can twitter analytics predict election outcome? An insight from 2017 Punjab assembly elections." ***Government Information Quarterly***, Vol. *37, No.* 2, DOI: 10.1016/j.giq.2019.101444.
74. Wang, L., Gan, J. Q., (2017), "Prediction of the 2017 French election based on Twitter data analysis", In 9th Computer Science and Electronic Engineering, 89-93. DOI: 10.1109/CEEC.2018.8674188
75. Wicaksono, A. J., (2016), "A proposed method for predicting US presidential election by analyzing sentiment in social media." *2nd Int. Conference on Science in Information Technology*, 276-280. DOI: 10.1109/ICSITech.2016.7852647.
76. Kalyanam j., Quezada M., Poblete B., and Lanckriet G., (2016), "Prediction and characterization of high-activity events in social media triggered by real-world news," PloS one, 11(12), e0166694.
77. Hettiarachchi, H., Adedoyin-Olowe, M., Bhogal, J., & Gaber, M. M. (2021). Embed2Detect: Temporally clustered embedded words for event detection in social media. *Machine Learning*, 1-39.
78. Sabuncu, I., (2020), "USA Nov.2020 Election 20 Mil. Tweets (with Sentiment and Party Name Labels) Dataset", IEEE Dataport, DOI: 10.21227/25te-j338.
79. Hutto, C.J. Gilbert, E.E., (2014), "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216-225.
80. Alex Filatov, https://data.world/alexfilatov/2016-usa-presidential-election-tweets
81. Kumar, P. M. A., Charan, K. G., Kumar, G. B. V. S., Amith, K. and Krishna, K. S., (2021), "Real-Time Hashtag based Event Detection Model with Sentiment Analysis for Recommending user Tweets," *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pp. 1437-1444, doi: 10.1109/ICICV50876.2021.9388426.
82. Guoming L., Yaqiao M., Jianbin G., Franck A.P., Chenxi L., Ruozhou W., Aiguo C., (2021), "A hashtag-based sub-event detection framework for social media", Computers & Electrical Engineering, Volume 94, 107317, https://doi.org/10.1016/j.compeleceng.2021.107317.

Abstract

In the last decade, the high penetration rate of virtual social networks among individuals in communities and the frequency and availability of their data has attracted the attention of many researchers to the analysis of social networks. One of the important applications of virtual social network analysis is event detection and prediction. Existing event prediction methods focus only on one specific event, while what is addressed in this study is the prediction of all predictable events. In this research, the event refers to something that causes significant changes in some parameters and important features of the virtual social network. A feature that has been considered in the proposed method is changes in the rate of sending messages to the social network. In the proposed method, based on a fixed time window, the operation of splitting and cleaning tweets is performed. The tweets are then categorized using the Non-negative Matrix Factorization (NMF), a matrix analysis method, and incremental clustering distance dependent Chinese Restaurant Process (ddCRP). After that, the occurrence of the event is predicted based on the changes in the rate of tweets entering each cluster. Finally, a description of the event is displayed in a few repetitive words in each cluster.

The proposed method, based on a dataset of Twitter, contains approximately 12 million tweets covering 1940 events, is evaluated, and succeeds in predicting future events with 85% accuracy.

Key words:

Virtual Social Network Analysis, Event Detection, Event Prediction, Election Prediction, Tweet Sending Rate

Shahrood University of Technology

Kharazmi Campus College

Computer engineering, artificial intelligence and robotics

Real World Social Events Prediction via Analyzing Virtual Social Networks

A thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Doctor of Philosophy in Computer Engineering

By:

Abulfazl Yavari Khalilabad

Supervisor:

Prof. Hamid Hassanpour

Advisor:

Dr. Mohammad Bagher Rahimpour Cami

Prof. Mehrgan Mahdavi

January 2022