بسم‌الله الرحمن الرحیم

محتویات گزارش، به شرح زیر است:

- تعریف و بیان مسئله
- مجموعه دادگان و معیار ارزیابی
- شرح فعالیت‌های صورت گرفته

۱ تعریف و بیان مسئله

پروژه TDT از دو sub-task اصلی تشکیل می‌شود:

- Topic Detection
- Topic Tracking

Topic Detection به دنبال کشف topic یا event جدید از Text corpus است و Topic Tracking به دنبال پیگیری تغییرات یک topic پیش‌فرض در طول زمان است.

از طرفی Topic Detection خود به دو task مختلف تقسیم می‌شود:

- New event detection
- Retrospective Event Detection

در New Event Detection به دنبال کشف رخداد جدید و از قبل دیده نشده در یک Stream از اسناد هستیم. یعنی اسناد ازلحاظ زمانی مرتب‌شده، به سیستم وارد می‌شود و برای هر سند، در مورد جدید یا تکراری بودن آن سند با توجه به اسناد از قبل دیده‌شده تصمیم‌گیری می‌شود[۱].

عبارات Topic و Event و Story در این پژوهش بسیار مورداستفاده قرار می‌گیرد، بنابراین، ارائه یک تعریف این واژه‌ها ضروری به نظر می‌رسد:

Topic: یک رخداد و یا فعالیت اولیه به همراه تمام رخدادها و فعالیت‌هایی که مستقیماً با آن مربوط است[۱].

Event: یک اتفاق منحصربه‌فرد که در یک‌زمان خاص رخ می‌دهد[۲]. در این تعریف، خصوصیت "زمان" باعث ایجاد تمایز بین Event و Topic می‌شود. برای مثال، سقوط هواپیما به‌عنوان یک Topic و سقوط هواپیمای ATR-72 ایران در تاریخ ۲۹ بهمن ۱۳۹۶ را می‌توان یک Event از Topic سقوط هواپیما در نظر گرفت.

Story: هر رخداد را مجموعه‌ای از Story ها تشکیل می‌دهند. درواقع، یک قطعه منسجم است که در یک یا چند گزاره، در مورد یک رخداد مشخص ، اطلاعاتی ارائه می‌دهد[۱].

مسئله‌ای که در این پژوهش به آن پرداخته می‌شود، **کشف زیر رخداد در جریان داده خبری** است. ابتدا باید زیر رخداد تعریف شود:

در این پژوهش فرض شده است که یک رخداد خبری می‌تواند به یک توالی دنباله‌دار از Event ها شکسته شود[۳]. برای مثال، رخداد سقوط هواپیمای ATR-72 یک توالی دنباله‌دار از مجموعه event هایی است که آغاز آن با خبر سقوط هواپیما شروع‌شده و در ادامه، خبرهای مربوط به شروع عملیات نجات، کشف محل سقوط، ناکامی دررسیدن گروه‌های امدادی به محل سقوط، رسیدن اولین گروه‌های امدادی به محل و درنهایت به اخباری در مورد دلیل سقوط ختم خواهد شد.

بنابراین می‌توان نتیجه گرفت که برای یک رخداد، می‌توان زیر رخداد متصور شد. پژوهش‌هایی که تاکنون بر روی کشف زیر رخداد تمرکز کرده‌اند[۴]–[۸]، در دو مقوله تعریف Sub-event و نیز روش‌های کشف آن، کاملاً به شبکه‌های اجتماعی وابسته بوده‌اند، به‌عنوان‌مثال، در [۹] زیر رخداد به‌عنوان یک موضوعی که در توییتر در یک بازه زمانی کوتاه به‌طور شدید موردبحث قرارگرفته و سپس محو می‌شود، تعریف‌شده است. بر این اساس، روش‌های کشف زیر رخداد بر مدل کردن انسجام[۱] و حالت انفجاری[۲] توییت‌ها استوار است. بنابراین، ارائه تعریف و روش‌هایی عمومی‌تر برای کشف زیر رخداد در سطح سند ضروری به نظر می‌رسد[۱۰].

هدف اصلی در این پژوهش کشف اسنادی است که ازلحاظ محتوایی، نسبت به اسناد قبلی دیده‌شده، دارای محتوای جدید باشند، ازآنجایی‌که اخبار مرتبط با یک رویداد، توسط رسانه‌های مختلفی پوشش داده می‌شود، وجود اخبار با محتوای تکراری در جریان خبرهایی از منابع و خبرگزاری مختلف، امری طبیعی محسوب می‌شود. بنابراین، افزونگی در خبرها و به عبارتی، تعداد زیاد خبرهایی با محتوای یکسان، می‌تواند موجب اتلاف زمان کاربر نهایی شود. بنابراین ارائه روشی برای کشف اطلاعات جدید و مرتبط با یک رویداد خاص و حذف خبرهای تکراری، می‌تواند منجر به تولید رشته‌ای از اخبار شود که بدون افزونگی، کل اتفاقات و اطلاعات درباره یک رخداد را شامل شود.

در تعریف واژه "جدید"، ابهاماتی وجود دارد و تابه‌حال، جدید بودن یک محتوا نسبت به محتواهای قبلی، به‌طور دقیق تعریف‌نشده است[۳]، برای مثال، خبری را در نظر بگیرید که نسبت به اخبار قبلی، تنها اطلاعاتی درباره سن یک فرد مهم افزوده‌شده است. این موضوع سبب می‌شود که دو متن با شباهات متنی بسیار بالا، نسبت به هم متفاوت باشند و این خود یک چالش مهم در این پژوهش است، بنابراین تکیه‌بر اطلاعات محتوایی خبر، به‌تنهایی موجب خطاهایی در کشف زیر رخداد جدید شود. به‌منظور حل این مشکل، از اطلاعات موجود در شبکه‌های اجتماعی درباره رخداد موردعلاقه و یا نظرات منتشرشده کاربران در هر مقاله خبری می‌توان استفاده کرد. کاربران شبکه‌های اجتماعی در واکنش به یک خبر جدید، محتواهایی را تولید می‌کنند که این محتواها می‌تواند نسبت به محتواهای تولیدشده درباره اخبار قبلی، متفاوت باشد[۸]. همین تفاوت محتوا می‌تواند به کشف زیر رخداد جدید کمک کند.

۲ مجموعه دادگان و معیار ارزیابی

برای آزمایش مجموعه روش‌هایی که برای کشف زیر رخداد، ارائه‌شده است، نیاز به مجموعه دادگانی از اخبار مرتبط به یک رخداد یا event واحد نیاز داریم. مجموعه داده‌ای شامل اخبار مرتبط به "قتل روح‌الله داداشی"، شامل ۲۴۰ خبر است. در این پژوهش، هر خبر می‌تواند ۵ برچسب داشته باشد:

- جدید

---

[۱] cohesiveness

[۲] burstiness

- تقریباً جدید
- تکراری
- بدون ارتباط
- تحلیلی

این نوع برچسب‌گذاری، در [۱۱] نیز استفاده‌شده است. توصیف هرکدام از برچسب‌ها بدین شرح است:

جدید: خبری که حاوی اطلاعات جدید نسبت به اخبار قبلی است و این اطلاعات جدید فراوان است.

تقریباً جدید: خبری که حاوی اطلاعات جدید نسبت به خبر قبلی است اما این اطلاعات جدید، نسبت به برچسب "جدید" کمتر است.

تکراری: خبری که حاوی اطلاعات جدید نیست و خبری تکراری محسوب می‌شود.

تحلیلی: خبری که به رخداد ارتباط دارد اما حاوی اطلاعات جدید نیست و یا تحلیلی بر رخداد است.

بدون ارتباط: خبری که به رخداد ارتباطی ندارد.

در این پژوهش، اخباری که دارای برچسب "تقریباً جدید" هستند درنهایت، به‌عنوان خبر جدید در نظر گرفته‌شده‌اند.

برای ارزیابی روش‌های کشف زیر رخداد، علاوه بر معیار ارزیابی F-measure یک معیار دیگر نیز استفاده‌شده است[۸] که در ادامه به توصیف آن پرداخته می‌شود:

هزینه کشف یا Cost detection که به‌اختصار با C_{Det} نمایش داده می‌شود. این معیار ارزیابی، درواقع جمع وزن‌دار بین دو نوع احتمال است:

		Reference Annotation	
		Target	Non-Target
System Response	YES (a Target)	Correct	*False Alarm*
	NO (Not a Target)	*Missed Detection*	Correct

احتمال miss: خبری که جدید است، به خطا، تکراری اعلام شود

احتمال false alarm: خبری که تکراری است، به خطا، جدید اعلام شود.

بنابراین، در این پژوهش، برای هر سند، تصمیم‌گیری مشود که سند جاری، جدید است و یا تکراری.

$$C_{Det} = C_{Miss}P_{miss}P_{target} + C_{FA}P_{FA}(1 - P_{target})$$

Detection Cost is normalized to lie between 0 and 1:

$$(C_{Det})_{norm} = \frac{C_{Det}}{min\{C_{miss}P_{target}\ ,\ C_{FA}(1-P_{target})\}}$$

$C_{Miss} = 1$ And $C_{FA} = 0.1$ and $P_{target} = 0.02$ are pre-specified.

$$P_{miss} = \frac{\#missed\ detectios}{\#targets},\ P_{FA} = \frac{\#false\ alarms}{\#non-targets}$$

۳ شرح فعالیت‌های صورت گرفته

- روش پایه‌ای:

برای ارزیابی و مقایسه با روش‌های موجود، روش پیشنهادی در [۱۲] مورداستفاده قرارگرفته است. روش ارائه‌شده بر شباهت کسینوسی و روش وزن دهی TF.IDF افزایشی استوار است. بدین‌صورت که شباهت کسینوسی سند جاری با تمام اسناد قبلی که پردازش‌شده‌اند محاسبه‌شده و چنانچه این شباهت، از آستانه‌ی مشخصی فراتر رود، سند جاری به‌عنوان سند تکراری و در غیر این صورت به‌عنوان سند جدید معرفی خواهد شد.

- شباهت خبرگزاری و فاصله زمانی

این ویژگی به شکل سخت‌گیرانه‌ای در [۱۳] استفاده‌شده است:

$$sim(d_i, d_j) = \cos(d_i, d_j).\tau(d_i, d_j).\delta(d_i, d_j)$$

$$\tau(d_i, d_j) = 1 - \frac{|d_i.t - d_j.t|}{T}$$

$$\delta(d_i, d_j) = \begin{cases} 0\ , if\ \ d_i.d = d_j.d \\ 1\ , otherwise \end{cases}$$

رابطه اول، شباهت کسینوسی تو قطعه از متن را با ویژگی فاصله زمانی انتشار دو قطعه خبری و ویژگی شباهت خبرگزاری درهم‌آمیخته است، به این صورت که هرچه فاصله زمانی دو خبر منتشرشده، بیشتر باشد، شباهت این دو خبر کمتر خواهد شد، دلیل این امر نیز این است که اخباری که در فاصله زمانی نزدیک به هم منتشر می‌شوند، احتمالاً محتواهایی درباره یک رخداد واحد دارند و هرچه فاصله زمانی دو خبر بیشتر باشد، این شباهت کمتر خواهد بود. علاوه بر این، در این روش فرض شده است که یکسان بودن خبرگزاری منتشرکننده دو خبر ($d_i.d = d_j.d$)، باعث صفر شدن شباهت بین خبرها خواهد شد، معنی این عبارت در این فرض نهفته است که یک خبرگزاری، اخبار تکراری را منتشر نخواهد کرد. اما بر مبنای آزمایش‌هایی که در این پژوهش انجام‌شده است، استفاده از این ویژگی‌ها، بدین شکل سخت‌گیرانه می‌تواند منجر به افزایش خطا شود. به همین دلیل، روابط بالا به شکل دیگری بازنویسی شده است:

$$sim(d_i, d_j) = \cos(d_i, d_j).\tau(d_i, d_j).\alpha.\delta(d_i, d_j)$$

$$\tau(d_i, d_j) = \exp(-\frac{\max(0, |d_i.t - d_j.t| - offset)^2}{2\sigma^2})$$

$$\sigma^2 = \frac{-T^2}{(2.\ln(decay))}$$

$$\delta(d_i, d_j) = \begin{cases} 0\,, if\ \ d_i.d = d_j.d \\ 1\,, otherwise \end{cases}$$

پارامترها:

Offset: تا چه حدی از فاصله زمانی در محاسبه شباهت دخالت داده نشود.

T: در چه میزانی از فاصله زمانی، امتیاز فاصله زمانی برابر صفر قرار داده شود.

Decay: در فاصله زمانی برابر با T چه امتیازی به شباهت سند داده شود.

α: تنظیم‌کننده میزان تأثیر شباهت خبرگزاری بر شباهت سند

- PLM

برای برخی اسناد، شباهت کسینوسی به‌تنهایی نمی‌تواند معیار خوبی باشد، دلیل این امر این است که برخی اسناد خبری که از چندین بند تشکیل‌شده‌اند، روایتی تکراری از خبرهای گذشته و مرور رخداد را شامل می‌شوند و در یک پاراگراف و یا در چند جمله، اطلاعات جدیدی را ارائه می‌دهند، در این حالت، شباهت کسینوسی بسیار بالاست و خطای miss رخ می‌دهد. برای حل این چالش از مدل زبانی اسناد استفاده‌شده است. بدین‌صورت که هر سند جدید، بر اساس پاراگراف، قطعه‌بندی شده و هر قطعه به‌عنوان یک query در نظر گرفته می‌شود. برای هر سند جدید، شبیه‌ترین n سند ازلحاظ کسینوسی نیز به‌عنوان مجموعه اسناد در نظر گرفته می‌شود. پس برای هر سند جدید که به p قطعه تقسیم‌شده است، n سند مشابه وجود دارد. به ازای هرکدام از قطعات سند جاری، با استفاده از PLM امتیاز شباهت محاسبه می‌شود. فرض شده است که اگر قطعه‌ای از سند جاری، حاوی اطلاعات جدید باشد، امتیاز آن با دیگر قطعات سند، متفاوت باشد. برای مثال، نتایج زیر از سندی که به ۴ قطعه تقسیم‌شده و ۳ سند شبیه به سند جاری، استخراج‌شده است، q1 تا q4 قطعات متن جاری است و Doc1 تا Doc3 نیز، شبیه‌ترین اسناد، به سند جاری است. همان‌طور که مشخص است، واریانس امتیاز q4 بسیار کمتر است و می‌توان نتیجه گرفت که این قطعه از متن، حاوی اطلاعات جدیدی است که قبلاً در ۳ سند شبیه به سند جاری، مشاهده نشده است.

PML	
q1 Doc1 -1.18391 q1 Doc2 -2.37288 q1 Doc3 -2.40587	0.4846
q2 Doc2 -1.04139 q2 Doc3 -2.14024 q2 Doc1 -2.16265	0.4108
q3 Doc3 -1.95715 q3 Doc1 -3.20701 q3 Doc2 -3.21143	0.5225

q4 Doc2 -2.53407 **q4 Doc3 -2.58379** **q4 Doc1 -2.70037**	**0.0072**

- نظرات منتشرشده برای هر خبر

همان‌طور که در [۸] عنوان‌شده است، هنگامی‌که رخدادی جدید و متفاوت از قبل اتفاق می‌افتد، انتظار می‌رود که کاربران شبکه‌های اجتماعی و کاربران مخاطب خبر، نظراتی با محتواهای جدید در ارتباط به رخداد جدید منتشر کنند. از این ایده برای کشف زیر رخداد جدید نیز می‌توان استفاده کرد. این ایده در مجموعه دادگان اخبار و نظرات منتظر شده در هر خبر نیز در این پژوهش مورد آزمایش قرار گرفت. هر نظر، به‌عنوان یک tweet فرض شده است و گرافی از مجموعه نظرات ساخته می‌شود. هر گره در این گراف، متناظر با یک واژه و یال‌های وزن‌دار بین گره‌های گراف، تابعی از هم رخدادی واژگان است. (بدون نتیجه)

- تغییرات تعداد نظرات

در مجموعه دادگان، مشاهده شد که برای اخبار جدید، تعداد نظرات به‌تنهایی می‌تواند ویژگی خوبی باشد، می‌توان از این ویژگی به شکل نرمال شده استفاده کرد. خبرگزاری‌ها با سیاست‌های متفاوتی نسبت به درج نظرات خوانندگان اقدام می‌کنند و بنابراین، نمی‌توان از تعداد نظرات به شکل خام استفاده کرد. برای مثال، متوسط تعداد نظرات در سایت "خبرآنلاین" بسیار بالاتر از خبرگزاری‌های دیگر است.

- تعداد کلمات کلیدی مشترک با شبیه‌ترین خبر

کاهش ابعاد هر سند خبری به تعداد n کلمه کلیدی، می‌تواند پیچیدگی زمانی را کاهش دهد، در این پژوهش به‌منظور بررسی کارایی کلمات کلیدی، در هر سند، n کلمه کلیدی استخراج‌شده و اندازه مجموعه حاصل از اشتراک کلمات کلیدی سند جاری با مجموعه کلمات شبیه‌ترین سند به‌عنوان یک ویژگی در نظر گرفته‌شده است. همان‌طور که قابل پیش‌بینی است، این روش تا حد خوبی(این بیان غیررسمی است و حد خوب باید تعریف شود) در مدل کردن شباهت محتوایی در ابعاد کمتر موفق عمل کرده است.

- مقایسه شباهت کسینوسی حاصل از دو نوع بردار

استفاده از شباهت کسینوسی بر مبنای بردار حاصل از موجودیت‌های اسمی اسناد و مقایسه آن با بردار حاصل از تمام کلمات سند در [۱۴] به‌تفصیل بررسی‌شده است. در این کار، نویسنده به این نتیجه رسیده است که در برخی موارد، استفاده از بردار حاصل از موجودیت‌های اسمی می‌تواند در کشف رخداد جدید موفق عمل کند و در برخی موارد نیز، این بردار موفق عمل‌نکرده است. در این پژوهش، مقایسه شباهت کسینوسی حاصل از بردار موجودیت‌های اسمی با بردار حاصل از تمام کلمات سند، می‌تواند ویژگی موفقی در کشف رخداد جدید و یا تشخیص تکراری بودن خبر باشد. (در مجموعه دادگان روح‌الله داداشی نتیجه داده است)

۴ منابع

[1] J. Allan, “Introduction to Topic Detection and Tracking,” in *Topic Detection and Tracking: Event-based Information Organization*, J. Allan, Ed. Boston, MA: Springer US, 2002, pp. 1–16.

[2] R. Papka and J. Allan, “On-Line New Event Detection Using Single Pass Clustering TITLE2:,” 1998.

[3] J. Allan, R. Gupta, and V. Khandelwal, “Temporal Summaries of News Stories,” *Proc. ACM SIGIR 2001*, pp. 10–18, 2001.

[4] D. Abhik and D. Toshniwal, “Sub-Event Detection During Natural Hazards Using Features of Social Media Data,” in *Proceeding of the 22th International Conference on World Wide Web*, 2013, pp. 783–788.

[5] S. Katragadda, R. Benton, and V. Raghavan, “Sub-event detection from tweets,” in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 2128–2135.

[6] C. Xing, Y. Wang, J. Liu, Y. Huang, and W. Ma, “Hashtag-Based Sub-Event Discovery Using Mutually Generative LDA in Twitter,” in *Aaai*, 2016, pp. 2666–2672.

[7] P. K. Srijith, M. Hepple, K. Bontcheva, and D. Preotiuc-Pietro, “Sub-story detection in Twitter with hierarchical Dirichlet processes,” *Inf. Process. Manag.*, vol. 53, no. 4, pp. 989–1003, 2017.

[8] P. Meladianos, G. Nikolentzos, F. Rousseau, Y. Stavrakas, and M. Vazirgiannis, “Degeneracy-Based Real-Time Sub-Event Detection in Twitter Stream,” in *AAAI*, 2015.

[9] C. Shen, F. Liu, F. Weng, T. Li, and P. Alto, “A Participant-based Approach for Event Summarization Using Twitter Streams,” in *HLT-NAACL*, 2013.

[10] A. Badgett and R. Huang, “Extracting Subevents via an Effective Two-phase Approach,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 906–911.

[11] E. Gabrilovich, S. Dumais, and E. Horvitz, “Newsjunkie: providing personalized newsfeeds via analysis of information novelty,” in *WWW*, 2004, pp. 482–490.

[12] T. Brants and F. Chen, “A System for new event detection,” in *SIGIR*, 2003.

[13] J. B. P. Vuurens, R. Blanco, and P. Mika, “Online news tracking for ad-hoc Information Needs,” in *ICTIR ’15*, 2015, pp. 2–4.

[14] G. Kumaran and J. Allan, “Text classification and named entities for new event detection,” in *SIGIR*, 2004.