"Listening In": Social Signal Detection for Crisis Prediction

۱ مسئله اصلی چیست و چرا مهم است؟

مقاله روی این مسئله تمرکز دارد که بحران‌ها معمولا سیگنال‌های هشدار اولیه تولید می‌کنند؛ اما این سیگنال‌ها اغلب ضعیف هستند و در نویز زندگی روزمره گم می‌شوند. در مدیریت بحران، اگر بتوان این سیگنال‌های اولیه را زودتر تشخیص داد، سازمان‌ها می‌توانند پیش‌دستانه آماده شوند و قبل از وقوع بحران واکنش‌های مناسب دهند.
در عین حال، تشخیص سیگنال از توییت‌ها در عمل رایج نشده چون حجم نویز و توییت‌های نامرتبط بالاست و همچنین مشکلاتی مثل سوگیری زبانی/فرهنگی و کمبود کانتکست به‌خاطر کوتاه بودن توییت‌ها وجود دارد.

ایده این مقاله این است که با یک روش Open-domain و چندزبانه، بتوان توییت‌ها را به‌صورت موثر پالایش کرد تا تنها بخش مرتبط با بحران باقی بماند و سپس با تحلیل روند (Burst/Trend) رشدهای غیرعادی را به‌عنوان سیگنال شناسایی کرد.

۲ ورودی‌ها و خروجی‌های مدل/سیستم چیست؟

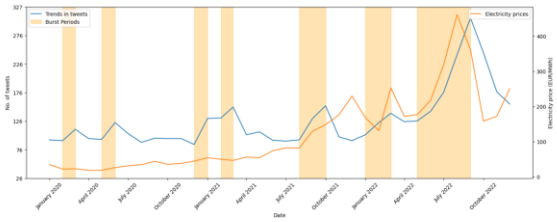
در پیاده‌سازی مقاله، ورودی/خروجی‌ها به‌صورت زیر تعریف می‌شوند:

ورودی‌ها

1. **کلیدواژه‌های دامنه:** (متن ساده) که کاربر برای موضوع بحران می‌دهد؛ مثال حوزه انرژی: rising و Blackout و ...
2. **تنظیمات اجرایی**: کشور/زبان/بازه زمانی سیستم قابل تعیین است.
3. **داده توییتر**: توییت‌های زنده/تاریخی از طریق Twitter/X API .

خروجی‌ها

1. **توییت‌های فیلترشده و مرتبط** با بحران: توییت‌هایی که informative تشخیص داده شده‌اند و احساس negative/neutral دارند.
2. **هشدار سیگنال**: وقتی روند فراوانی توییت‌های مرتبط وارد فاز رشد پایدار شود.
3. **نمایش روند:** در سرویس نمونه، خروجی شامل ویژوال‌سازی شبیه شکل زیر است.



(a) Crisis event e_1:'Peak in electricity price', Germany, Jan 2020 - Dec 2022.

۳ داده مورد استفاده (نوع، منبع، اندازه) چیست؟

مقاله دو منبع داده اصلی برای کار دارد:

الف) داده برای فاین‌تیون مدل پالایش

- دیتاست: CrisisBench
- اندازه: 156,899 توییت
- تقسیم‌بندی: train = 109,796 ، dev = 16,008، test = 31,095
- وظیفه یادگیری: طبقه‌بندی دودویی informative در برابر not informative .

ب) داده برای ارزیابی سرویس سیگنال‌دهی

- منبع: Twitter/X از طریق API
- اندازه خام: 46,963 توییت از آلمان، بریتانیا، اسپانیا

۴ روش پیشنهادی مقاله به زبان ساده + شماتیک یا شبه‌کد

مقاله روش **OSOS** را معرفی می‌کند: یک pipeline سه‌مرحله‌ای که از داده توییتر + کلیدواژه‌های دامنه شروع می‌کند و با هشدار سیگنال پایان می‌دهد. در توضیح مقاله، خروجی پیش‌پردازش وارد پالایش می‌شود و خروجی پالایش وارد آشکارساز سیگنال می‌گردد.

۴.۱ گام اول: پیش‌پردازش

- حذف کاراکترهای خاص، lowercase، حذف تکراری‌ها و رشته‌های خالی، حذف punctuation .
- توکنایز و جمله‌بندی و تبدیل توییت به لیست توکن‌ها.

۴.۲ گام دوم: پالایش داده

هدف: حذف نویز و نگه داشتن توییت‌هایی که واقعا به بحران مربوط‌اند.

1. طبقه‌بندی با استفاده از GPT-3 (Curie) که روی CrisisBench فاین‌تیون شده تا توییت‌های بحران‌مرتبط را بهتر تشخیص دهد.
2. تحلیل احساسات با DistilBERT و تمرکز روی negative یا neutral چون در متن اشاره می‌شود این دو احساس معمولاً با رویدادهای بحران‌مرتبط همراه‌اند و لذا سیستم روی آن‌ها تمرکز می‌کند.

خروجی این ماژول: مجموعه‌ای از توییت‌هایی که هم informative هستند و هم negative/neutral ؛ بقیه کنار گذاشته می‌شوند.

۴.۳ گام سوم: آشکارسازی سیگنال

روش مقاله:

- فراوانی توییت‌های مرتبط فیلترشده به سری زمانی روزانه نگاشت می‌شود.
- trend component استخراج می‌شود و سپس دوره‌های رشد نمایی در روند بررسی می‌گردد.
- تعریف سیگنال :دوره‌ای از رشد نمایی که حداقل ۷ روز دوام داشته باشد و آستانه ٪۱۰ تغییر در فراوانی توییت‌ها را رد کند؛ در این صورت هشدار صادر می‌شود.

۵ نتایج اصلی، محدودیت‌ها و ایده‌های ادامه

۵.۱ نتایج اصلی

الف) کیفیت ماژول پالایش

در ارزیابی چند مدل روی CrisisBench ، مدل GPT-3FT (Curie) بهترین عملکرد را گزارش می‌کند:
Accuracy = 0.882 و F1 = 0.905 بهتر از گزینه‌های دیگر.

ب) عملکرد سرویس در اجرا

- از 46,963 توییت خام، پس از پیش‌پردازش و پالایش، 7042 توییت برای تحلیل نهایی باقی مانده است.
- سرویس سیگنال‌دهی برای سه رویداد انرژی‌مرتبط تاریخی بررسی شده و معیار تصمیم تغییر فراوانی توییت‌ها (FC%) در بازه‌های 12، 8 و 4 هفته قبل از رخداد گزارش شده است.

۵.۲ محدودیت‌ها

مقاله چند محدودیت/ریسک کلیدی را صریح بیان می‌کند:

1. روش نشان نمی‌دهد **همه انواع بحران** با یک سطح عملکرد قابل پیش‌بینی‌اند؛ تمرکز روی بحران‌های قابل پیش‌بینی اما به‌سختی قابل کنترل بوده (مثل افزایش قیمت انرژی).
2. تکیه اصلی روی **داده‌های توییتر** است و در نتیجه با مسئله **کیفیت پایین، اطلاعات غلط و توییت‌های جعلی** مواجه می‌شود؛ حتی امکان **دست‌کاری با ارسال انبوه توییت‌های مخرب** وجود دارد.
3. درباره یکی از رویدادهای تاریخیمذکور اشاره می‌کند که تغییرات کوچک‌تر بوده و می‌تواند ناشی از مسائل آستانه ۷ روزه یا عوامل ارتباطی در ادراک بحران باشد و این موضوع خودش یک پرسش پژوهشی می‌سازد.

۵.۳ ایده‌های ادامه

- افزودن **تکنیک‌های راستی‌آزمایی** برای کاهش اثر اطلاعات غلط.
- ترکیب OSOS با **منابع اطلاعاتی دیگر** (مثلا اخبار، داده‌های بازار، گزارش‌های کسب‌وکار) برای مقاوم شدن و کاهش ریسک‌های توییترمحور بودن.