الجمهورية العربية السورية

المعهد العالي للعلوم التطبيقية والتكنولوجيا

قسم النظم المعلوماتية

العام الدراسي 2024/2025

مشروع تخرج

أعد للنجاح في السنة الخامسة

كشف الأحداث ومراقبتها في وسائل التواصل الاجتماعي اعتماداً على تقنيات التعلم العميق

تقديم

حسن بهجت خضور

إشراف

د. رياض سنبل

8/10/2010

الخلاصة

يهدف هذا المشروع إلى

Abstract

Translate your abstract here.

المحتويات

الفصل الأول **1**
التعريف بالمشروع **1**
1.1– مقدمة **2**
2.1– الهدف من المشروع **2**
3.1– نطاق المشروع **2**
4.1– المتطلبات الوظيفية **3**
5.1– المتطلبات غير الوظيفية **4**
الفصل الثاني **5**
الدراسة المرجعية **5**
1.1– مقدمة **6**
1.1– مسألة كشف الأحداث في دفق بياناتٍ **6**
1.1.1 الإطار العام للمسألة 7
1.1.1 المراحل الرئيسية لسير العمل في نظام لكشف الأحداث 9
1.1.1 تعريف المسألة Problem Definition 10
1.1.1 طرق معالجة المسألة 10
1.1.1 المنهجيات والمقاربات لمعالجة المسألة 11
1.1.1 معايير التقييم وقياس الأداء 13
أ. Normalized Mutual Information (NMI) 13
ب. Adjusted Mutual Information (AMI) 13
ت. B-Cubed 14
1.1.1 مجموعات البيانات المتاحة 14
3.1– مسألة الاستدلال على الموقع الجغرافي من النص **16**
1.1.1 الإطار العام للمسألة 16
1.1.1 المنهجيات والمقاربات المتبعة للتعرف على الموقع 17
1.1.1 استخدام نماذج اللغة الكبيرة لاستدلال الموقع المذكور 18
1.1.1 تعريف المسألة Problem Definition 19

1.1.1مجموعات البيانات المتاحة 20
1.1.1 معايير التقييم وقياس الأداء 20
4.1– أنظمة مشابهة 21
1.4.1 Liveuamap.com Live Universal Awareness Map 21
2.4.1 rsoe-edis.org Emergency and Disaster Information Service (EDIS) .22
3.4.1 Globalincidentmap.com Global Incidents Map 23
الفصل الثالث 25
الدراسة النظرية 25
1.1– مقدمة 26
3.1– منصات وسائل التواصل الاجتماعي 26
3.1– العنقدة CLUSTERING 26
1.3.1 العنقدة القائمة على الكثافة Density Based Clustering 26
3.1– العنقدة ONLINE CLUSTERING 27
3.1– شبكات المعلومات غير المتجانسة (HIN) 27
2.1– التعرف على الكيانات المسماة NAMED ENTITIES RECOGNITION 28
3.1– GRAPH NEURAL NETWORK 28
الفصل الرابع 29
الدراسة التحليلية 29
1.1– مقدمة 30
3.1– مخطط حالات الاستخدام 30
4.1– السرد النصي لحالات الاستخدام 33
4.1– العلاقات بين الفاعلين 37
4.1– مخطط المكونات 38
الفصل الخامس 39
المنهجية المقترحة 39
1.1– مقدمة 40
2.1– المقاربة المتبعة لكشف الأحداث ..... خطأ! الإشارة المرجعية غير معرّفة.
3.1– المقاربة المتبعة لتصنيف الأحداث ..... خطأ! الإشارة المرجعية غير معرّفة.

4.1– منهجية الاستدلال على الموقع الجغرافي..... خطأ! الإشارة المرجعية غير معرّفة.
الفصل السادس 43
تصميم النظام 43
1.1– مقدمة 44
2.1– تصميم خدمات النظام..... 46
3.1– مخطط النظام التصميمي..... 45
4.1– مخطط النشر DEPLOYMENT DIAGRAM..... خطأ! الإشارة المرجعية غير معرّفة.
الفصل السابع..... 51
الأدوات المستخدمة 51
1.1– مقدمة 52
2.1– الهدف من المشروع 52
الفصل الثامن 90
تنجيز النظام 90
1.1– مقدمة 91
2.1– تنجيز الخدمات 92
3.1– مخطط النظام 95
الفصل التاسع..... 96
تحليل ومناقشة النتائج..... 96
1.1– مقدمة 97
2.1– معايير التقييم..... 97
الفصل العاشر..... 98
اختبارات النظام 98
1.1– مقدمة 99
2.1– اختبار خدمات النظام 99
3.1– اختبارات الأداء..... 99

مقدمة عامة

أصبحت وسائل التواصل الاجتماعي اليوم أحد المصادر الرئيسية لنشر المعلومات حول الأحداث اليومية، حيث يقوم المستخدمون بمشاركة الأخبار العاجلة، الكوارث الطبيعية، الجرائم، الحوادث والأحداث الاجتماعية بمجرد وقوعها. ومع النمو السريع في حجم المحتوى المنشور على هذه المنصات، أصبحت هناك حاجة ملحة إلى أنظمة ذكية قادرة على استخراج وتحليل هذه البيانات بشكل آلي لتوفير معلومات دقيقة وتقارير عن الأحداث الجارية.

يعتمد كشف الأحداث على تقنيات معالجة اللغة الطبيعية (NLP) والتعلم العميق لفهم وتحليل المحتوى النصي المنشور على وسائل التواصل الاجتماعي، واستخراج المعلومات ذات الصلة مثل نوع الحدث، وموقعه الجغرافي .كما يتضمن ذلك استخدام تقنيات التعرف على الكيانات المسماة (NER) لاستخلاص الأسماء، الأماكن، والمنظمات المذكورة في النصوص، بالإضافة إلى تقنيات التحليل الجغرافي لتحديد مواقع الأحداث حتى في حال عدم توفر إحداثيات جغرافية مباشرة.

تعد هذه العملية ذات أهمية كبيرة في العديد من المجالات، مثل:

- الاستجابة للكوارث والطوارئ :تمكين فرق الإغاثة والجهات الحكومية من الحصول على معلومات دقيقة وفورية حول الكوارث الطبيعية مثل الزلازل والفيضانات والحرائق.
- تحليل الجرائم والأمن العام :مساعدة الأجهزة الأمنية في تتبع الجرائم وحوادث العنف والسرقات فور الإبلاغ عنها عبر وسائل التواصل الاجتماعي.
- رصد الأحداث الاجتماعية والسياسية :تمكين الصحفيين والباحثين من متابعة التغيرات المجتمعية والتطورات السياسية في الزمن الحقيقي.
- إدارة المدن الذكية :تحسين إدارة البنية التحتية والخدمات العامة من خلال مراقبة شكاوى المواطنين وتفاعلهم مع الأحداث المحلية.

يهدف هذا المشروع إلى تقديم نظام متكامل قادر على تحليل تدفقات البيانات الاجتماعية، استخراج المعلومات منها، وعرضها بشكل مرئي يسهل فهمها، مما يسهم في تعزيز الوعي بالأحداث الجارية واتخاذ قرارات مبنية على بيانات دقيقة مرئية.

الفصل الأول

التعريف بالمشروع

نبيّن في هذا الفصل هدف المشروع ونطاقه كما نورد المتطلبات الوظيفية وغير الوظيفية.

1.1- مقدمة

في وقتنا الحال، أصبحت وسائل التواصل الاجتماعي مصدرنا رئيسياً للأخبار والمعلومات المحلية، فغدت تقدم لنا دفقاً من البيانات من المهم معالجته واستخلاص المعلومات المهمة منه وعرضها بشكل مرئي يساعد الجهات والمنظمات المعنية على اتخاذ القرارات.

2.1- الهدف من المشروع

يندرج عملنا في هذا المشروع ضمن سياقين اثنين: (1) الأول نتطرق فيه للمسألة من منظور الذكاء الصنعي، حيث نحاول توظيف تقنياته وأدواته في الوصول إلى مقاربة لحل مسألتين مطروحتين هما أ. كشف الأحداث من دفق بياناتٍ وتصنيفها، ب. الاستدلال على الموقع الجغرافي من رسالة نصية، (2) والثاني نتطرق فيه إلى بناء تطبيق برمجي يستفيد من المقاربة المقترحة في تنفيذ التطبيق عملياً مع مراعاة الأسس والمبادئ المتبعة في هندسة البرمجيات وصولاً إلى تطبيقٍ قابلٍ للتوسع وسهل الصيانة.

3.1- نطاق المشروع

نطاق هذا المشروع يقتصر على تطوير نظام لاكتشاف الأحداث المحلية باستخدام بيانات وسائل التواصل الاجتماعي وتحليلها باستخدام تقنيات الذكاء الاصطناعي والتعلم العميق. يركز المشروع على تطبيق تقنيات التعلم العميق لتحليل النصوص المستخلصة من وسائل التواصل الاجتماعي، مع الهدف الرئيسي لاكتشاف الأحداث المحلية مثل الحوادث، الأزمات، الجرائم.

مكونات المشروع:

1. جمع البيانات: سيتم جمع البيانات من وسائل التواصل الاجتماعي، مثل تويتر، تيليغرام، باستخدام واجهات برمجة التطبيقات (APIs) المتاحة لهذه الشبكات وباستخدام الزواحف.
2. تحليل النصوص: استخدام تقنيات التعلم العميق، لتحليل النصوص وتصنيفها إلى أنواع مختلفة من الأحداث (جرائم، حوادث، ...). بالإضافة إلى استخراج الكيانات الأساسية مثل الأماكن والأشخاص المتعلقين بهذه الأحداث باستخدام تقنيات التعرف عل الكيانات المسماة (NER).
3. التحليل الجغرافي: التعرف على المواقع الجغرافية من البيانات المستخلصة، سواء عبر البيانات الجغرافية المرفقة (الموسومة جغرافياً) أو من خلال الاستدلال على المواقع. وتطوير آلية لتحويل النصوص الجغرافية إلى إحداثيات جغرافية عبر مرمزٍ جغرافي (Geocoder).

4. النظام التفاعلي: تطوير تطبيق ويب تفاعلي لعرض الأحداث المكتشفة على خرائط تفاعلية في الزمن الحقيقي، مما يسمح للمستخدمين بمراقبة الأحداث وتقييمها على الخريطة. وتقديم تقارير وبيانات إحصائية حول الأحداث المكتشفة.

4.1- المتطلبات الوظيفية

نبين هنا المتطلبات الوظيفية التي يجب أن يحققها النظام.

يجب على النظام أن يسمح للمستخدمين بما يلي:

1. استعراض الأحداث حسب المنطقة الجغرافية
2. استعراض الأحداث التي قد مرت (حسب التاريخ)
3. تحديد منطقة جغرافية، لتلقي إشعارات بالأحداث التي تقع بها
4. استخراج مخططات ورسوم توضيحية حول الأحداث وتسلسلها الزمني.

يجب على النظام أن يسمح للمراقب بما يلي:

1. إدارة الأحداث، أي تعديل تصنيفها، حذفها، أهميتها، وإضافة معلومات عنها
2.

يجب على النظام ان يسمح لمدير النظام بما يلي:

1. إدارة مصادر البيانات، أي تعديل مصادر سحب البيانات، إضافة مصادر جديدة، حذف مصادر قديمة
2.

يجب على النظام أن يقوم بما يلي:

1. جدولة الأحداث بشكل يومي، بحيث يعرض على المستخدم الأحداث اليومية على الخريطة.
2. إرسال إشعارات للمستخدمين عند وقوع أحداث ضمن مناطقهم.
3. استخراج (تجريف Scrape) البيانات من وسائل التواص الاجتماعي مثل تويتر وتيليغرام والمنصات الإخبارية
4. التخاطب مع APIs وسائل التواصل الاجتماعي وجمع البيانات منها
5. تمرير البيانات في بيئة دفقية streaming إلى الخدمات الأخرى

5.1- المتطلبات غير الوظيفية

نورد هنا المتطلبات غير الوظيفية أي القيود على النظام لكي يتم قبوله. حيث قمنا بتقسيمها إلى عدة محاور وهي متطلبات الأمان ومتطلبات الأداء ومتطلبات خاصة ومتطلبات التشغيل.

أ. متطلبات الأمان

1. يجب أن يكون الدخول إلى النظام آمناً، أي يجب أن يسمح فقط للمستخدمين المسجلين بالدخول إليه.
2. أن يسمح للمستخدمين القيام بالعمليات وفقاً لما هم مخولين به من صلاحيات فقط.

ب. متطلبات الأداء

1. يجب أن يستجيب النظام لطلبات المستخدم في غضون زمن محدد لا يتجاوز 2 ثانية لمعظم العمليات.
2. أن يكون النظام قادرً على معالجة دفق البيانات وفق السعة الموضحة بالدراسة الكمية.
3. أن يكون النظام قادرً على استخراج دفق بيانات (data streams) بسعة لا تقل عن X Message /Second

ت. متطلبات خاصة

1. أن تكون الواجهات باللغة – .

ث. متطلبات التشغيل

1.
2. أن يتم.

الفصل الثاني

الدراسة المرجعية

نبيّن في هذا الفصل الأبحاث والأعمال والمنهجيات المشابهة لهذا العمل.

1.1- مقدمة

تعد مسألة اكتشاف الأحداث وتصنيفها من وسائل التواصل الاجتماعي من التحديات البحثية الهامة التي تجمع بين عدة مجالات في الذكاء الاصطناعي، مثل معالجة اللغة الطبيعية (NLP)، واستخراج المعلومات، وتحليل البيانات الضخمة. تعتمد هذه المسألة على تحليل دفق من النصوص المنشورة لاستخلاص الأحداث المهمة وتصنيفها وفقًا لنوعها، مثل الجرائم، والحوادث المرورية.

إضافةً إلى ذلك، يُعتبر استنباط الموقع الجغرافي للأحداث من النصوص غير المهيكلة تحديًا إضافيًا، حيث يتطلب تقنيات متقدمة لاستخلاص المعلومات الجغرافية وربطها بإحداثيات دقيقة. يهدف هذا القسم إلى تعريف المسألة ومراجعة الأعمال البحثية السابقة في مجال كشف الأحداث وتصنيفها، استخراج الموقع الجغرافي من النصوص، بالإضافة إلى استعراض الأنظمة المشابهة ومجموعات البيانات المتاحة لدعم هذه المهام.

1.1- مسألة كشف الأحداث في دفق بياناتٍ

إنَّ التطور السريع لمنصات وسائل التواصل الاجتماعي أدى إلى تضخم في البيانات المنشورة من قبل المستخدمين على الإنترنت. حيث مكنت هذه البيانات الضخمة من دراسة العديد من المشكلات البحثية، وواحد من مواضيعها المهمة هو كشف الأحداث (Q. Li et al., 2022).

وبالانتقال إلى تعريف الحدث، فإن التعريف يختلف قليلاً في الدراسات السابقة حيث يعرفه McMin وآخرون بشكل عام على أنه "شيء مهم يحدث في زمان ومكان محددين"، بينما يعرفه Xie وآخرون على أنه "حدث في العالم الحقيقي، وقع في زمان ومكان معينين، شارك به أو اثر بمجموعة من الأشخاص فحفزهم على التكلم عنه". وبشكل عام يمكننا أن نصفه بما يسمى الـ 4Ws (What, Where, Who & When). بينما يعرف Chao وآخرون الحدث المحليّ كما يلي "الحدث المحلي (على سبيل المثال، الاحتجاج، الجريمة، الكوارث الطبيعية) هو نشاط غير عادي ينفجر في منطقة محلية وخلال مدة محددة ويشترك به (أو يتفاعل معه) عدد كبير من المشاركين".

ولطالما تم تناول كشف الأحداث من الوسائط التقليدية في سياق كشف المواضيع وتتبعها Topic Detection and Tracking (TDT)، ومع ذلك فإنّ اكتشاف الأحداث من وسائل التواصل الاجتماعي يطرح تحديا جديدة تختلف عن تلك الموجودة في وسائل الإعلام التقليدية. فعلى عكس المقالات الإخبارية، فإن التغريدات تكون مقيدة الطول مما يجعل المعلومات النصية فيها محدودة. كما تشمل كميات كبيرة من الكلمات غير الرسمية والأخطاء الإملائية، وهياكلها غير المنظمة. وتحتوي أيضاً عبى كميات من الرسائل الوهمية والشائعات (Q. Li et al., 2022).

1.1.1 الإطار العام للمسألة

نستطيع النظر إلى مسألة كشف الأحداث من عدة زوايا، استناداً إلى نوع الحدث، يمكننا تصنيف هذه الطرق إلى اكتشاف حدث غير محدد وحدث محدد Unspecified Vs. Specified. وفي مجال ال TDT، يصنف كشف الأحداث إلى فئتين (1) كشف الأحداث الجديدة New Event Detection (NED) و (2) كشف الأحداث الرجعية Retrospective Event Detection (RED). ففي حالة NED يكون هدفنا هو اكتشاف الاحداث الجديدة في دفق بيانات في الزمن الحقيقي، بينما في حالة RED فإننا نركز على التعرف واكتشاف الأحداث في بيانات تاريخية (Q. Li et al., 2022)

يمكننا الآن مناقشة كل نوع على حدى، فنبدأ إذن بالتصنيف القائم على النوع:

- **كشف الحدث المحدد Specified Event Detection**

في هذا النوع من الأحداث يكون لدينا علم مسبق بمكان وزمان وقوع الحدث. أي عندما يكون الحدث الاجتماعي معروفاً أو مخططاً له بالفعل، فإن معالجة البيانات المتعلقة بالمعلومات المعروفة (مثل الموقع والوقت والكلمات الرئيسية والمستخدمين) لاستخراج وصف الحدث تسمى اكتشاف الأحداث المحددة (SED). ويعالج في هذا النوع، المعلومات المحددة مسبقاً والميزات التي من المتوقع أن تظهر في البيانات لتمثيل حدث. تعمل هذه المعلومات المحددة مسبقا كبذرة لسياق الحدث الفعلي. يمكن التعبير عن المعلومات المتعلقة بالحدث كليا أو جزئيا لجمع البيانات وتحليلها (Q. Li et al., 2022).

- **كشف الحدث غير المحدد Unspecified Event Detection**

في هذا النوع من الأحداث، لاعِلمَ لنا بمكان وزمان وقوع الحدث ولا بطبيعته، فعندها يكون مصدرنا الوحيد لتحديده هو مراقبة الخصائص الزمانية والكلمات المنتشرة. أي عندما تكون المعلومات الاجتماعية حول حدث ما غير معروفة، يعتمد النهج الأساسي ل UED على تحليل الأنماط الزمنية لتدفق البيانات من خلال مراقبة الدفق لتحديد الكلمات الرئيسية والمفاهيم المتكررة ذات الصلة بتسليط الضوء على الأحداث (Q. Li et al., 2022).

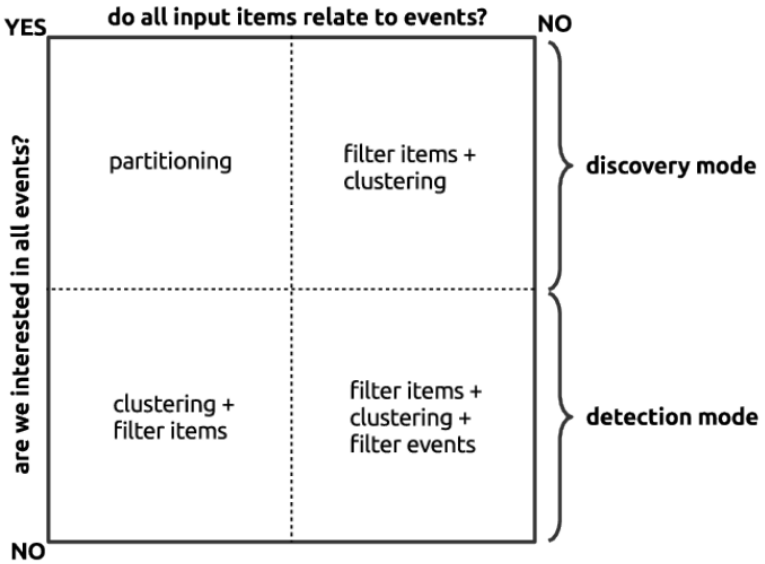
أما بالنسبة للتصنيف الثاني:

- **كشف الأحداث الجديدة والرجعية New Event Detection & Retrospective Event Detection**

اعتماد على المهمة ونوع الحدث، يمكننا أيضاً تصنيف الحدث إلى نوعين NED وRED. ونظراً لأن تقنيات NED تتضمن المراقبة المستمرة لدفق البيانات لكشف الأحداث الجديدة في الزمن الفعلي، فإنه يكون مناسباً بطبيعته لكشف أحداث العالم الحقيقي غير المعروفة والأخبار العاجلة. كما يمكن استخدام تقنيات ال NEDمن أجل كشف الأحداث المحددة، على الرغم من أن معظم الدراسات ترك على الأحداث غير المحددة. وعندما تتضمن المهمة المرادة أحداثاً محددة أو معلومات محددة

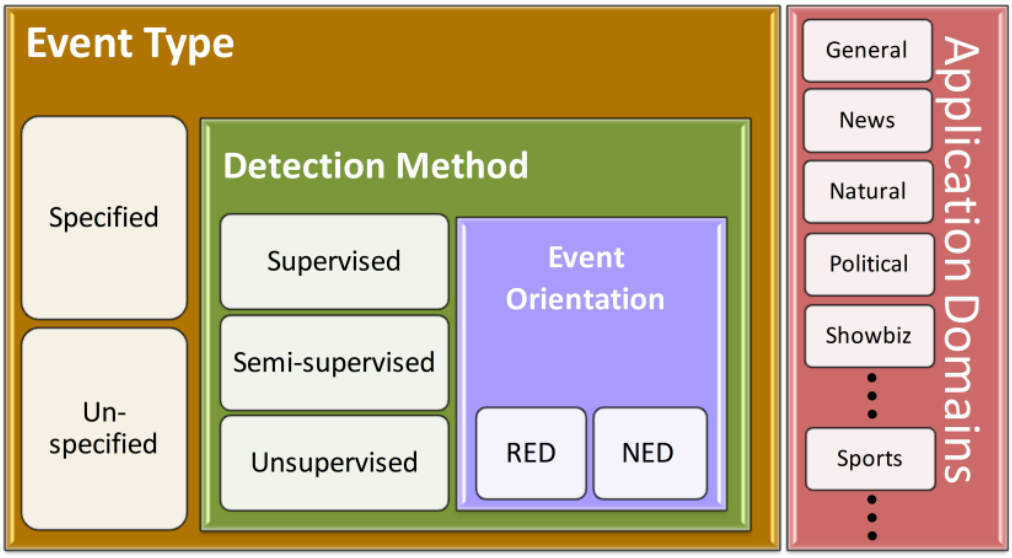
حول الحدث (عل سبيل المثال، منظمة محددة، شخص أو مكان ما) يمكن الاستفادة من هذه المعلومات لمكاملتها مع تقنيات ال NED من خلال تقنيات الفلترة أو التصنيف.

كما ويوضح الشكل أدناه بعض نطاقات المعالجة المهمة (تجزئة، فلترة، عنقدة) وذلك حسب نمط المسألة. أي إذا كنا نهتم بجميع الأحداث أو إذما كانت جميع الرسائل تنتمي إلى أحداث (Schinas et al., 2018).



الشكل 1: نطاقات معالجة كشف الأحداث حسب نمطها **(Schinas et al., 2018)**.

ويختصر الشكل أدناه مفردات المسألة ويصنفها بشكل هرمي يوضحها.



الشكل 2: تصنيف هرمي للأحداث **(Saeed, Abbasi, Maqbool, et al., 2019)**.

1.1.1 المراحل الرئيسية لسير العمل في نظام لكشف الأحداث

يقسم Li وآخرون في بحثهم مراحل عمل نظام كشف الأحداث لثلاثة مراحل كما هو موضح في الشكل أدناه، هي (1) مرحلة المعالجة المسبقة والتي تتم على دفق البيانات الذي يرد إلى النظام، (2) مرحلة كشف الأحداث، (3) مرحلة مابعد كشف الحدث.

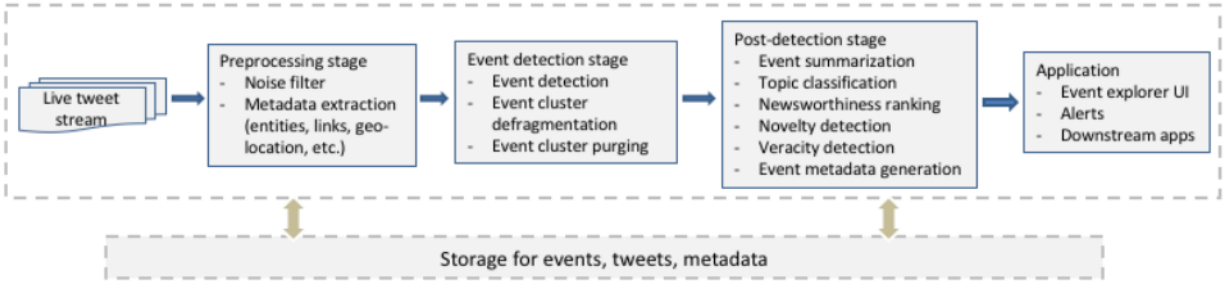


Figure**1** : The typical workflow for a NED system for unspecified events **(Q. Li et al., 2022)**.

ومهمة كل مرحلة هي كما يلي:

- ♣ مرحلة المعالجة المسبقة Preprocessing Stage، تحوي هذه المرحلة عدة مكونات مثل مكون فلترة الضجيج الذي يقوم بإزالة الرسائل غير المهمة، الإعلانات وماشابه. وتحوي على مكونات لاستخراج الكيانات، المعلومات الجغرافية.
- ♣ مرحلة كشف الحدث Event Detection Stage، اعتماداً على نوع الحدث المراد كشفه (Specified, Unspecified, RED, NED) قد تختلف تقنية الكشف الفعلية (على سبيل المثال، الطرق المعتمدة على العنقدة، المعتمدة على المصطلحات، ...). وتهدف مكونات Event Cluster Defragmentation إلى دمج مجموعات الأحداث ذات الصلة، و Purging Event إلى إزالة الأحداث القديمة التي لم يعد هناك حاجة إليها.
- ♣ مرحلة مابعد الكشف Post-Detection Stage، حسب التطبيق الذي نهتم به، قد نحتاج إلى بعض المكونات مثل تلخيص الأحداث Event Summarization، تحديد الموضوع Topic Detection، كشف الشائعات.

1.1.1 تعريف المسألة Problem Definition

أ- تعريف: دفق البيانات الاجتماعي A Social Stream

نعرف دفق البيانات $\mathfrak{S}$ على أنه تسلسل مرتب زمنياً ومستمر لكتل الرسائل الاجتماعية Social Messages Blocks، ممثلة ك $\mathfrak{S} = M_0, \dots, M_{i-1}, M_i, \dots$ حيث كل كتلة رسائل M_i تحوي كل الرسائل التي وصلت خلال الشريحة الزمنية $[t_i, t_{i+1}[$ أي M_i هي مجموعة من الرسائل الفردية، أي $M_i = \{m_j; 1 \leq j \leq |M_i|$، حيث كل رسالة m_j تحوي نص الرسالة بالإضافة إلى المرسل وزمن الإرسال (Abagissa et al., 2024).

ب- تعريف: الحدث An Event

نرمز له بِ e، وهو مجموعة من الرسائل الاجتماعية المترابطة المرتبطة بحدث في العالم الحقيقي. ونفترض أن كل رسالة مرتبط بحدث اجتماعي واحد على الأكثر. أي $e = \{m_j; 1 \leq j \leq |e|\}$ (Abagissa et al., 2024).

ت- تعريف: خوارزمية كشف الأحداث Event Detection Algorithm

من أجل كتلة رسائل M_i، يتم تدريب خوارزمية كشف الأحداث Social ED Algorithm لتعلم نموذج $\mathfrak{F}$، بحيث $\mathfrak{F}(M_i; \Theta) = E_i$ حيث E_i هي مجموعة الأحداث التي تظهر في الكتلة M_i. أي $E_i = \{e_j; 1 \leq j \leq |E_i|\}$.

و Θ هو موسط كدخل للنموذج (Abagissa et al., 2024).

1.1.1 طرق معالجة المسألة

هنالك تنوع كبير في منهجيات كشف الأحداث، ونستطيع تصنيف هذه المنهجيات إلى أربع فئات اعتمادً على مبدأ عملهم:

- ♣ مرتكزة على السمة Feature-Pivot، تعتمد هذه الطرائق على كشف أنماط غير طبيعية في بعض سمات النص (على سبيل المثال، بعض العبارات المتفجرة Bursty Phrases)، أي عندما يقع حدث ما فإن بعض الكلمات والعبارات سيتم تداولها بشكل كثير مما يجعل توزعها غير طبيعي بالنسبة لتوزعها تاريخياً، وبناءً على ذلك يتم الكشف.
- ♣ مرتكزة على المستند Document-Pivot، تعتمد هذه الطرائق على تحويل المستند (message & its meta data) إلى شعاع، ومن ثم عنقدة الأشعة وتجميعها بناءً على التشابه. وفي عدد من الدراسات الحديثة، تم تعزيز هذه المنهجية باستخدام الشبكات العصبية البيانية (GNNs) ، حيث يتم الاستفادة من شبكات المعلومات غير المتجانسة لتحويل بيانات وسائل التواصل الاجتماعي من شكلها غير المهيكل إلى شكل مهيكل ذو علاقات دلالية، ومن ثم الاستفادة من البيان الناتج ومن قدرات GNNs في تمثيل العلاقات غير الخطية والتفاعلات المعقدة بين الكيانات. لتحسين وزيادة فعالية تمثيل اشعة التضمين، على سبيل المثال، وظفت أبحاث مثل *(KPGNN - Cao et al., 2021)*، *(HCRC - Guo et al., 2024)*و *(QSGNN - Ren et al., 2022)*تقنيات GNNs.

- ♣ مرتكزة على الموضوع Topic-Pivot، تضمن هذه الطرائق استخدام نماذج إحصائية لكشف الحدث على اعتباره متحول ضمني في المستند.(Schinas et al., 2018).

وفي بعض الأبحاث لجاء الباحثون إلى طريقة قائمة على البيان Graph Based، حيث يتم الاستفادة من الهيكلية الدلالية للرسائل ضمن وسائل التواصل الاجتماعي حيث يتم تحويل مجموعة الرسائل غير المهيكلة Unstructured Messages إلى رسائل مهيكلة ذات دلالة Semantic Structured Data (وذلك باستخدام الكيانات المسماة، hashtags، عمليات الذكر) وذلك بإنشاء بيان، وبعد إنشاء هذا البيان تتحول المسألة حينها على مسألة تجزئة البيان Graph Partitioning.

1.1.1 المنهجيات والمقاربات لمعالجة المسألة

يوضح الجدول أدناه الأبحاث ذات الصلة والمقاربات المتبعة لكشف الأحداث في وسائل التواصل الاجتماعي.

#	Paper (Name)	Techniques	Reference	Supervision	Event Type	Detection Task
1	Petrovi´c et al., 2010	Online Clustering & LSH		Unsupervised	U	NED
2	Sakaki et al., 2010	SVM		Supervised	S	RED
3	Long et al., 2011	Hierarchical Clustering		Unsupervised	U	NED
4	Cordeiro, 2012	Wavelet Analysis & LDA		Unsupervised	U	NED
5	McCreadie et al., 2013	K-Mean Clustering & LSH			U	NED
6	Corney et al., 2014	Online Clustering	(Corney et al., 2014)		S	NED
7	Alsaedi et al., 2015	Online Clustering	(Alsaedi et al., 2015)	Unsupervised	U	NED
8	Patil et al., 2015	NN Classification	(Patil et al., 2015)	Supervised	S	-
9	Guille & Favre, 2015	Mention, Anomaly Detection	(Guille & Favre, 2015)		U	NED
10	Zhang et al., 2016	Spatial-temporal Clustering	(C. Zhang et al., 2016)	Unsupervised	U	NED
11	Alsaedi et al., 2016	Temporal, Spatial, Textual Features, Online Clustering	(Alsaedi et al., 2016)	Unsupervised	U	NED
	Hasan et al., 2016 (TwitterNews)	Incremental Clustering	(Hasan et al., 2016)	Unsupervised	U	NED
12	Chen et al., 2017	NN, Online Clustering			U	NED

13	Zhang et al., 2017	Spatial-Temporal Clustering	(S. Zhang et al., 2017)	Unsupervised	U	NED
14	Edouard, 2018	Domain Vocabulary and Knowledge Based			S	NED
15	Comito et al., 2019	Text & Temporal Feature Clustering		Unsupervised	U	NED
16	Hossny et al., 2019	Keyword Volume Approach, SVM	(Hossny & Mitchell, 2019)	Supervised	S	NED
17	Fedoryszak et al., 2019	Cluster Chain, Graph	(Fedoryszak et al., 2019)	Unsupervised	U	NED
18	Nguyen et al., 2019	Entity Clustering			U	NED
19	Saeed et al., 2019	Graph, Clustering	(Saeed, Abbasi, Razzak, et al., 2019)	Unsupervised	U	NED
20	Morabia et al, 2019 (SEDTWik)	Segments, Hierarchical Clustering	(Morabia et al., n.d.)	Unsupervised	U	NED
21	Liu et al., 2020 (EventX)	Community Detection		Unsupervised	U	NED
22	Cao et al., 2020 (IED KCN)	Knowledge Consolidation Network (KCN)	(P. Cao et al., 2020)	Supervised	-	NED
23	Li and Zhang, 2021	Semantic Term, GCN, Clustering	(Q. Li & Zhang, 2021)		U	NED
24	Ghaemi et al., 2021	Spatial-Temporal Clustering, VDBSCAN	(Ghaemi & Farnaghi, 2023)	Unsupervised	U	NED
25	Cao et al., 2021 (KPGNN)	GNN, Clustering	(Y. Cao et al., 2021)	Supervised	-	NED
26	George et al., 2021	Spatial-Temporal Clustering	(George et al., 2021)	Unsupervised	U	NED
27	Mojiri et al., 2021	Burst Keywords, Time Series Analysis		Unsupervised	U	NED
28	Yamani et al., 2021	Knowledge Based, MLM	(Yamani et al., n.d.)	Supervised	-	-
29	Peng et al., 2022 (FineEvent)	GNNs	(Peng et al., 2023)	-	-	-
30	Hettiarachchi et al., 2022 (Embed2Detect)	NN, hierarchical clustering	(Hettiarachchi et al., 2022)	Unsupervised	U	NED
31	Afyouni Khan et al., 2022	Spatial-Temporal Clustering, Classification	(Afyouni, Khan, et al., 2022)	-	U	NED

32	Ren et al., 2022 (QSGNN)	GNNs, Online Clustering	(Ren et al., 2022)	Supervised	-	-
33	Bok et al.,2023	Graph, Online Clustering	(Bok et al., 2023)	Unsupervised	U	NED
34	Cao et al., 2024 (HISEvent)	GNNs, Graph Clustering	(Y. Cao et al., 2023)	Unsupervised	U	
35	Yue et al., 2023	MLM	(Yue et al., 2023)	Supervised	-	-
36	Pradhan et al., 2024 (EDTBert)	Graph Clustering, PLMs	(Pradhan et al., 2024)			
37	Guo et al., 2024 (HCRC)	GNNs, Incremental Clustering	(Guo et al., 2024)	Unsupervised	U	NED
38	Ray et al., 2024	Incremental Clustering	(Ray et al., 2024)	Unsupervised	U	NED
39	P. Li et al., 2024 (RPLMSED)	PLMs	(P. Li et al., 2024)	Supervised	S	NED
40	Abagissa et al., 2024 (DistilBert – GNN)	GNNs, Clustering	(Abagissa et al., 2024)		U	NED
41	Yang et al., 2024 (ADPSEM)	Structural Entropy Minimization	(Yang et al., 2024)	Unsupervised	-	-
42	Yu et al., 2024 (HyperSED)	Hyperbolic Embedding, Graph Partitioning	(Yu et al., 2024)	Unsupervised	U	
43	Kovalchuk et al., 2025 (SemConvTree)	Semantic Convolutional Tree	(Kovalchuk et al., 2024)	Unsupervised	U	NED

1.1.1 معايير التقييم وقياس الأداء

أ. **Normalized Mutual Information (NMI)**

المعلومات المتبادل المعيرة، يقيس مقدار المعلومات المشتركة بين المجموعات الحقيقية (Ground Truth) والمجموعات المكتشفة (Predicted Clusters). يتم تقييسه باستخدام الإنتروبي ليصبح المقياس بين 0 و 1. وهو مفيد لتقييم جودة العنقدة مع أخذ احتمالية الاتفاق العشوائي بعين الاعتبار.

$$NMI(U,V) = \frac{I(U;V)}{H(U) + H(V)}$$

حيث I(U;V) هي المعلومات المتبادلة بين المجموعتين وH(U)، H(V) هي الإنتروبي لكل من المجموعات الحقيقية والمتوقعة.

ب. **Adjusted Mutual Information (AMI)**

مقياس معدل من NMI يأخذ بالحسبان مقدار المعلومات المتبادلة المتوقعة (Expected Mutual Information) عند التوزيع العشوائي للتسميات. بالتالي، يوفر مقياسًا أكثر دقة في الحالات التي يكون فيها الاتفاق العشوائي مرتفعًا.

$$AMI(U,V) = \frac{I(U;V) - \mathbb{E}[T(U;V)]}{\max(H(U),H(V)) - \mathbb{E}[T(U;V)]}$$

ت.B-Cubed

مقاييس B-Cubed Precision وB-Cubed Recall تستخدم لتقييم نتائج العنقدة (Clustering) بحيث تأخذ بعين الاعتبار كل عنصر ومقارنته مع عناصر أخرى في نفس المجموعة. أي بعكس المقاييس الأخرى التي تقيم التوزيع الكامل للمجموعات دفعة واحدة، يقوم B-Cubed بتقييم كل عنصر على حدة، ويحسب دقة واسترجاعه بناءً على الانتماءات الفعلية والمتوقعة.

$$BCubed\ Precision = \frac{1}{N}\sum_{e_i \in E} \frac{|Correct\ Labels(e_i) \cap Cluster(e_i)|}{|Cluster(e_i)|}$$

$$BCubed\ Recall = \frac{1}{N}\sum_{e_i \in E} \frac{|Correct\ Labels(e_i) \cap Cluster(e_i)|}{|Correct\ Labels(e_i)|}$$

1.1.1 مجموعات البيانات المتاحة

يوضح الجدول أدناه مجموعات البيانات المتاحة والمستخدمة في معظم المراجع لتدريب النماذج لهذه المسألة وقياس أداءها.

	Dataset	Size	Events	Language	Paper
1	Kawarith	~9K tweet	7	Arabic	(Alharbi & Lee, n.d.)
2	Events 2012	~68K tweet	503	English	(McMinn et al., 2013)
3	Maven	~10K tweet	164	English	(Wang et al., 2020)
4	Events 2018	~64K twee	257	French	-
5	FA Cup	~670K tweet		English	-
6	CrisisLex	~60K tweet	6	English	(Olteanu et al., 2014)
7	Few Events	~8K tweet	64	English	
8	KBP	~85K tweet	100	English	(Deng et al., 2020)
9	FloDust	~20K tweet	3	Arabic	(Hamoui et al., 2021)
10	CrisisMMD	~18K tweet	7	English	-

جدول 1: مجموعات البيانات المتاحة لمسألة كشف الأحداث.

كما ويوضح الجدول أدناه مجموعات البيانات التي تم استخدامها في الأبحاث، حيث يلاحظ أن معظم الأبحاث يعتمد على مجموعة بيانات خاصة بالبحث (غير متوفرة).

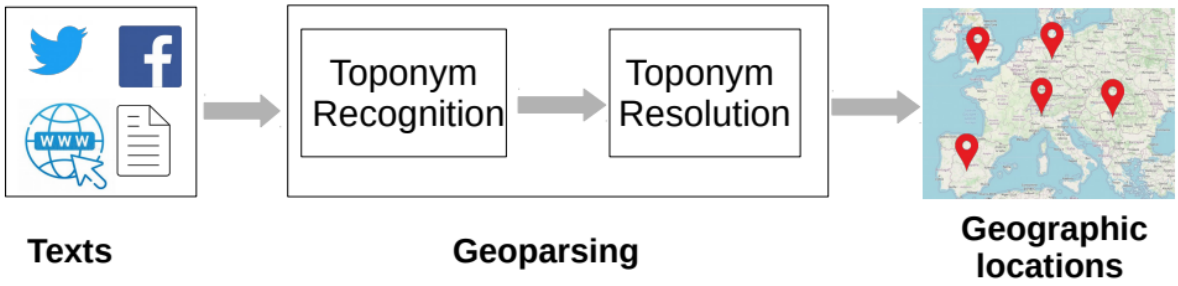
	Paper | Dataset	Events 2012	Events 2018	Maven	Few Events	CrisisLex	Kawarith	FA Cup	Other	Private
1	Alsaedi et al. , 2015									Y
2	Zhang et al. , 2017									Y
3	Hossny et al. , 2019									Y
4	Saeed et al. , 2019							Y	Y	
5	Fedoryszak et al. ,2019	Y								
6	Morabia et al. ,2019	Y								
7	Zhang et al.,2021									Y
8	Hettiarachchi et al. ,2021							Y		
9	Ghaemi et al. ,2021									Y
10	Mojiri et al. ,2021							Y	Y	
11	Yamani et al. ,2021								Y	
12	Cao et al. ,2021	Y		Y						
13	Peng et al. ,2021								Y	
14	Afyouni et al. ,2022									Y
15	Ren et al. ,2022	Y	Y							
16	Kolajo et al. ,2022									Y
17	Peng et al. ,2022									Y
18	Bok et al. ,2023									Y
19	Cao et al. ,2023	Y	Y							
20	Ren et al. ,2023 a	Y	Y				Y			
21	Ren et al. ,2023 b	Y	Y			Y				
22	Guoa et al. ,2023			Y						
23	Yue et al. ,2023			Y	Y					
24	Yang et al. ,2024	Y	Y							
25	Qiu et al. ,2024	Y				Y	Y			
26	Ray et al. ,2024	Y	Y							
27	Yu et al. ,2024	Y	Y				Y		Y	
28	Abagissa et al. ,2024	Y								
29	Pradhana et al. ,2024							Y	Y	
30	Ma et al. ,2024	Y	Y		Y	Y	Y			
31	Li et al. ,2024	Y	Y				Y			
32	Kovalchuk et al. ,2024 a									Y
33	Kovalchuk et al. ,2024 b	Y							Y	
34	Kargupta et al. ,2024									Y
35	Murthy et al. ,2025									Y

جدول 2: مجموعات البيانات المستخدمة في الأبحاث المدروسة.

3.1- مسألة الاستدلال على الموقع الجغرافي من النص

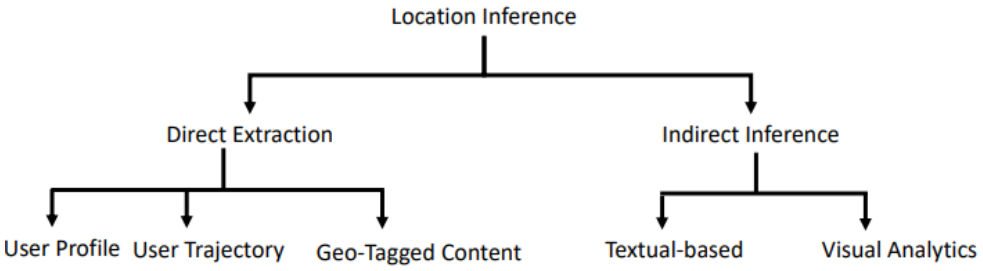
1.1.1الإطار العام للمسألة

يعرف الاستدلال على الموقع من النصوص الفردية –ويسمى أيضاً (GeoParsing)– بعملية تقدير الإحداثيات الجغرافية للنص بناءً على محتواها أو بياناتها الوصفية أو كليهما. ونظرًا لأن نسبة ضئيلة فقط من النصوص في مواقع التواصل الاجتماعي تحمل علامات جغرافية صريحة geotag in metadata فعلى سبيل المثال، في منصة توتير فقط 3% من التغريدات تحمل علامات جغرافية صريحة، ولذلك يعتمد الباحثون على المعلومات المستخرجة من اللغة الطبيعية – مثل أسماء الأماكن المذكورة في نص التغريدة – بالإضافة إلى المعلومات السياقية مثل الوسوم المعلومات الزمنية وبيانات المستخدم. ويتكون نظام الاستدلال على الموقع الجغرافي من عمليتين هما (1) التعرف على الموقع المذكور Location Mention Recognition (LMR)، و (2) تحديد الموقع وفك غموضه Location Mention Disambiguation and Detection (LMD). وبالنسبة لعملية التعرف على الموقع المذكور LMR والتي تسمى في بعض المراجع أيضاً بـ Toponym Recognition وهو حالة خاصة من مسألة التعرف على الكيانات المسماة Named Entities Recognition (NER) حيث يتم فيها تحديد الكلمات التي تدل على الموقع واستخراجها. أما في عملية تحديد الموقع وفك غموضه والتي أيضاً تسمى في بعض المراجع ب Toponym Resolution، ففيها يتم تحويل المواقع المستخرجة من المرحلة السابقة إلى إحداثيات جغرافية أو معرفات في قاموس جغرافي Gazetteer، وذلك بعد فك الغموض الذي يحدث نتيجة وجود عدد من العناوين المتشابهة (مثل مدينة بانياس في محافظة طرطوس، ومدينة بانياس في محافظة القنيطرة) التي تحل من خلال أخذ المعلومات السياقية بالاعتبار أو من خلال بعض التجريبيات. ويوضح الشكل أدناه آلية العمل تلك بشكل عام.



رسم توضيحي 1: مخطط العمل العام ضمن نظام استدلال الموقع **(Afyouni, Aghbari, et al., 2022)**.

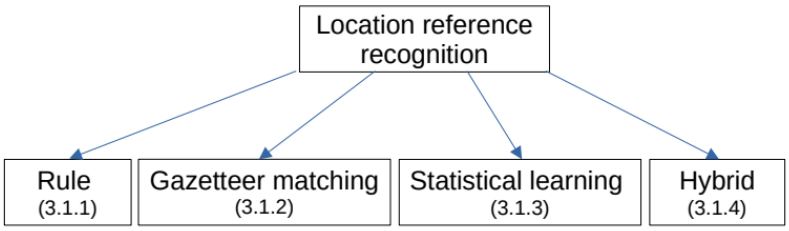
كما ويصنف الباحثون مصادر المعلومات المستخدمة للاستدلال على الموقع، إلى فرعين، الأول استخراج مباشر يتم عن طريق استخراج المعلومات المكانية الموسومة من ملف المستخدم الشخصي أو عن طريق المعلومات المكانية الموسومة مع النص المنشور حيث تقدم بعض وسائل التواصل مثل تويتر وسوم مكانية في حال فعل المستخدم ال GPS، والفرع الثاني هو غير مباشر ويكون إما عن طريق الاستدلال من خلال النص المنشور (وهو المناسب في حالتنا) أو عن طريق تحليل المحتوى المرئي (مثل الصور) وهذا لن نتطرق له في حالتنا. والشكل أدناه يوضح هذه الأنماط.



الشكل 3: أنماط لاستدلال على الموقع الجغرافي (**Afyouni et al.**, **2022**).

1.1.1 المنهجيات والمقاربات المتبعة للتعرف على الموقع

تصنف المقاربات المتبعة للتعرف على الموقع ضمن أربع أنماط هي (1) معتمدة على القواعد، (2) معتمدة على المطابقة من Gazeteer، (3) معتمدة على التعلم، (4) طرق هجينة بين الثلاث السابقة.



الشكل 4: تصنيف طرائق التعرف على الموقع.

- مقاربات قائمة على القواعد Rule-Based Approaches: في هذا النمط من الأنماط يتم معالجة المسألة من خلال الأنماط المنتظمة Regular Expressions، وذلك لأن الموقع الذي يذكر في النص غالباً ما يحتوي على سمات تميزه لذلك يتم الاعتماد على قواعد (سواء كانت موضوعة من قبل خبراء، أم تم تعلمها) لاستخراج النص الذي يشير إلى موقع، وهذه الطرق قليلة الاستخدام نظراً لأن النص المكتوب في وسائل التواصل الاجتماعي يحوي على ميزات تجعل تطبيق هذا النوع غير مجدٍ، ولكن تبقى هذه الطرق مستخدمة لتحسين أداء المقاربات الأخرى عند استخدامها معها.

- مقاربات مطابقة القاموس الجغرافي Gazetteer Matching Approaches: في المقاربات القائمة على المطابقة من قاموس جغرافي، يعتمد الباحثون في هذه المقاربة على أخذ الـ n-gram للنص، ومن ثم لكل سلسلة يقومون بمحاولة لمطابقتها مع القاموس الجغرافي، وبعد ذلك يتم فلترة النتائج وفك غموضها بناء على مجموعة تجريبيات heuristics.
- مقاربات التعلم Statistical Learning Approaches: في هذه الطرائق يتم الاعتماد على تعلم الآلة Machine Learning وعلى مدونة منمطة، لتعلم نموذج يقوم بتصنيف المفردات في النص. وفي معظم الأبحاث يقوم الباحثون بالاعتماد على نماذج التعرف على الكيانات المسماة NER ومن ثم ضبطها وتدريبها على هذه المسألة Fine-Tuning، وفي أبحاث أخرى كما (Bobadilla et al., Cadorel et al.، Chen et al.) يقوم الباحثون بالتعامل مع المسألة بشكل منفصل عن NER وتدريب نماذجهم لهذه المسألة خصيصاً.
- المقاربات الهجينة Hybrid Approaches: في هذا النوع يلجأ الباحثون إلى استخدام المقاربات السابقة مع بعضها البعض بحيث يتم الاستفادة من ميزات كل مقاربة لتحسين الحل المقترح.

1.1.1 استخدام نماذج اللغة الكبيرة لاستدلال الموقع المذكور

لقد عزز ظهور نماذج اللغة الضخمة (LLMs) بشكل كبير قدرة الآلة على فهم استفسارات المستخدم المعقدة، وعزز فعالية معالجة اللغة وفهمها لها، مما أفاد العديد من المجالات، بما في ذلك مجال العلوم الجغرافية المكانية. وقد بحثت العديد من الدراسات في إمكانات وقيود نماذج اللغة الضخمة في هذا المجال (Ji and Gao 2023, Mooney et al. 2023, Tao and Xu 2023, Xie et al. 2023, Yin et al. 2023, Hochmair et al. 2024). فعلى سبيل المثال، درس Xie وآخرون حدود تطبيق نماذج الذكاء الاصطناعي في السياقات الجغرافية المكانية، ودعوا إلى نماذجٍ ذات أساس جغرافي مُصممة خصيصًا لها. حيث يسلط هذا البحث الضوء على الاختلاف بين أنواع البيانات التقليدية، مثل نصوص اللغة الطبيعية والفيديو، التي صُممت من أجلها نماذج الذكاء الاصطناعي، والبيانات الجغرافية المكانية، مُشيرًا إلى ضرورة الابتكار في هذا المجال. وقيّم Ji وGao مدى كفاءة نماذج اللغة الضخمة في تفسير وتمثيل المفاهيم المكانية المذكورة في النصوص. وقارن Hochmair وآخرون أداء نماذج ChatGPT-4 وBard وClaude-2 وCopilot في مهام جغرافية مكانية مُختلفة، مثل فك الغموض المكاني، ومفاهيم نظم المعلومات الجغرافية، ورسم الخرائط. كما واستكشف Juhasz وآخرون إمكانية استخدام GPT 3.5-turbo لإثراء Open Street Map (OSM) من خلال اقتراح أفضل الوسوم لكل طريق في OSM بناءً على أوصاف مشتقة من صور Mapillary. وبعض الأبحاث أيضًا حول تطبيق نماذج اللغة الضخمة (LLM) في التحليل الجغرافي، على سبيل المثال، قدّم Li وآخرون نموذج لغة مكاني GeoLM، وهو مدرّب مسبقًا باستخدام بيانات من مصادر مثل ويكيبيديا و Wikidata وOpenStreetMap. يمكن تكييف GeoLM لدعم مهام متعددة في مراكز المدن، مثل التعرف على أسماء المواقع الجغرافية، والتحليل، واستخراج العلاقات، وكتابة الكيانات الجغرافية. كما واستكشف Mai وآخرون إمكانية الاستفادة من نماذج اللغة الضخمة، في مجال الذكاء الاصطناعي الجغرافي. وأثبتوا أن نماذج اللغة الضخمة تتفوق على النماذج المخصصة لمهام محددة في مهام الدلالات

الجغرافية المكانية، بما في ذلك التعرف على أسماء المواقع الجغرافية والتعرف على وصف الموقع. وقام أيضاً Huوآخرون بدراسة جدوى نماذج GPT المختلفة، لاستخراج أوصاف المواقع من وسائل التواصل الاجتماعي أثناء الكوارث. وقارنوا هذه النماذج بأدوات التعرف على الكيانات المسماة (NER) التقليدية ونموذج BERT المعدّل. استخدمت الدراسة مجموعة بيانات من التغريدات المنمطة من إعصار هارفي. وتُظهر النتائج أن نماذج GPT الموجّهة بالمعرفة الجغرافية تفوقت بشكل ملحوظ على أدوات NER ونموذج BERT المعدّل في التعرف على أوصاف المواقع الكاملة والفئات المرتبطة بها. كما وقيم Hu وآخرون أداء نماذج اللغة الضخمة المخففة Lightweight في مهمة التعرف على الموقع المذكور وفك الغموض.

1.1.1 تعريف المسألة Problem Definition

كما ذكرنا آنفاً أنّ المسألة تتكون من جزءين التعرف على الموقع المذكور (LMR)، وتحديد الموقع وفك الغموض (LMD)، ونستطيع صياغة المسألة لهما بشكل رياضي كما يلي:

- تصاغ مسألة التعرف على الموقع المذكور كما يلي

ليكن لدينا نص t، يهدف نظام LMR إلى تحديد جميع المواقع المذكورة في النص $L_t = \{l_i; i \in [1, n_t]$ حيث n_t هي العدد الكلي للمواقع المذكورة في النص t. وكل l_i قد تغطي أكثر من مفردة في النص لذلك يتم استخدام نموذج الوسم BILOU، حيث B توسم بداية الموقع و I توسم أننا ضمن ذكر موقع و L توسم نهاية ذكر الموقع و O أن المفردة خارج الرمز ينما ترمز U إلى أن ذكر الموقع يغطي مفردة واحدة فقط (Suwaileh et al., 2023).

- تصاغ مسألة تحديد الموقع وفك غموضه كما يلي

بعد أن يتم التعرف على جميع المواقع المذكورة في النص $L_t = \{l_i; i \in [1, n_t]$، يتم هنا ربط كل ذكر موقع l_i بتمثيل جغرافي دقيق وفريد، وهو يكون إما عبارة عن إحداثيات $(l_{atitude}, l_{ongitude})$ أو معرف في قاعدة بيانات جغرافية. نُعرّف هذه المسألة رياضياً كتابع $LMD : L \rightarrow \mathcal{G}$، حيث L هو مجموعة المواقع المذكورة المستخرجة من النص t، $\mathcal{G}$ هي مجموعة الكيانات الجغرافية الممكنة (على سبيل المثال، Gazeeter)، حيث تعطي لكل مدخل جغرافي إحداثياته $(l_{atitude}, l_{ongitude})$. حيث يهدف التابع LMD إلى إيجاد المطابقة الأمثل لكل l_i بحيث يُسند إلى l_i الكيان الجغرافي $g \in \mathcal{G}$ الذي يمثل بدقة المقصود في السياق النصي (Suwaileh et al., 2023).

1.1.1 مجموعات البيانات المتاحة

يوضح الجدول أدناه مجموعات البيانات المتاحة والمستخدمة في معظم المراجع لتدريب النماذج لهذه المسألة وقياس أداءها.

	Dataset	Size	Region	Language	Paper
1	IDRISI-R (Gold)	~20K En, ~4.6K Ar	Global (41 disaster events)	Ar, En	(Suwaileh et al., 2023)
2	IDRISI-R (Silver)	~57K En, ~1.2M Ar	Global (41 disaster events)	Ar, En	(Suwaileh et al., 2023)
3	GeoCorpora	~6,648 tweets	USA	En	
4	GeoCoV19	~524M tweets	Global (COVID-19 related)	Multilingual	(Qazi et al., 2020)
5	UTGEO	~670K tweets	USA	Multilingual	
6	TweeLoc	~325K tweets	Rome, Netherlands	Un Specified	

جدول 3: مجموعات البيانات المتاحة لمسألة التعرف على الموقع المذكور.

1.1.1 معايير التقييم وقياس الأداء

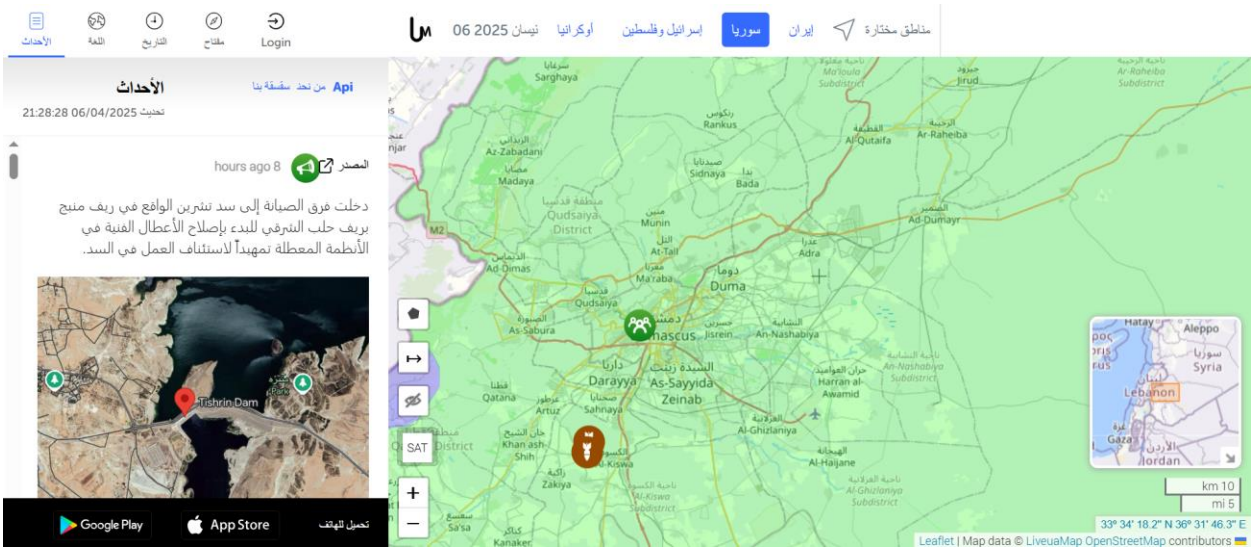
من أجل تقييم النتائج لهذه المسألة وقياس أداء النماذج يوجد العديد من معايير التقييم، التي تختلف حسب الهدف من النموذج فمن أجل النماذج التي تهدف إلى التنبؤ بـ (المنطقة، المدينة، ...) فيتم في هذه الحالة استخدام المعايير التقليدية كالصحة، الدقة والاستذكار (Accuracy, Precision, Recall)، أي يتم التعامل معا كأنها مسألة تصنيف طبيعية. (أي هي مناسبة في حالة التعرف على الموقع المذكور (LMR))، أما بالنسبة للنماذج التي تتنبأ بالإحداثيات الجغرافية للرسالة، ففي هذه الحالة يتم استخدام معايير التقيم المستخدمة لمسائل الانحدار متوسط خطاء المسافة Mean Error Distance، الصحة في الكيلومتر Accuracy@KM (أي بتكميم المسألة وتحويلها إلى تصنيف ضمن خلايا)، مسافة هافيرسين (Haverrsin Distance) وهي صيغة مسافة بين الإحداثيات الجغرافية ويستخدم لقياس المسافة على سطح الأرض، وصيغته هي كما هو موضح أدناه.

4.1- أنظمة مشابهة

يوجد عدد من المنصات التي تعالج المسألة المطروحة سواء بشكل شبه كامل أو جزء منها، على مستويات مختلفة، أي يوجد بعض المنصات التي تعالج المسألة على مستوى عالمي وأخرى على مستوى محلي.

1.4.1 Liveuamap.com Live Universal Awareness Map

خريطة الوعي العالمي الحي، هو موقع إلكتروني تفاعلي يعرض خريطة عالمية تحتوي على أخبار وأحداث جارية يتم تحديد مواقعها جغرافيًا. المنصة تُستخدم لتتبع الأحداث السياسية، العسكرية، الكوارث الطبيعية، وأخبار الأمن والنزاعات في الوقت الحقيقي، حيث يسمح للمستخدم بتحديد نوع الأحداث التي يريد الاطلاع عليها (كوارث طبيعية، نزاعات، ...) ومن ثم يعرض عليها بشكل حي هذه الأحداث ويتم تحديث المعلومات بالزمن شبه الحقيقي. حيث يعد مصدرًا قيمًا للمعلومات، ويوفر وصولاً مفتوحًا عبر الإنترنت إلى أرشيف زمني كامل للمعلومات، مما يتيح للمشاهدين البحث في الأحداث الماضية والاتجاهات التاريخية



♣ كيف يعمل الموقع؟

- يعتمد على مصادر متعددة لجمع البيانات مثل وسائل التواصل الاجتماعي، وكالات الأنباء، والمصادر المحلية.
- تستخدم المنصة خوارزميات آلية ومحررين بشريين للتحقق من صحة الأخبار وربطها بالموقع الجغرافي المناسب.
- يتم تصنيف الأحداث إلى أنواع متعددة (مثل: نزاع، سياسة، طقس، صحة، كوارث، إلخ) باستخدام رموز مختلفة.
- يمكن للمستخدمين استعراض الأخبار حسب الموقع الجغرافي، أو حسب التصنيف الزمني أو الموضوعي.

♣ كيف يجمع البيانات؟

- يعتمد على زواحف زكية AI Crawlers لتجميع البيانات وانتقاء الأخبار الجديرة بالذكر.

- يتم جمع البيانات من عدة مصادر من بينها وسائل التواصل الاجتماعي.
- بالإضافة إلى خوارزميات والنماذج، يعتمد الموقع على خبراء بشريين.

2.4.1 rsoe-edis.org Emergency and Disaster Information Service (EDIS)

موقع RSOE EDIS (نظام معلومات الطوارئ والكوارث)، هو خدمة تدار من قبل الجمعية الوطنية المجرية للإغاثة اللاسلكية والاتصالات، ويهدف إلى تتبع وتوثيق وتحليل الكوارث والحوادث الطارئة حول العالم. ويهدف إلى إبقاء العامة، والمنظمات الإنسانية، والباحثين، والحكومات على اطلاع دائم بالأحداث الطارئة، حيث يعرض بينات حيّة تم جمعها من وكالات الأنباء، وسائل الإعلام، المراصد، ... إلخ، ويقوم بتحليلها وعرض الجدير بالذكر منها.

صورة 1: الصفحة الرئيسية للموقع **rsoe-edis.org**

♣ كيف يعمل الموقع؟

- يقوم بجمع البيانات من جهات رسمية ومنظمات إنسانية ووسائل إعلام موثوقة.
- يجمع الموقع البيانات من مجموعة كبيرة من المصادر مثل:

  – وكالات حكومية – وسائل الإعلام الرسمية – منظمات دولية – شبكات المتطوعين

- يعرض الأحداث في الوقت الحقيقي عبر خريطة تفاعلية عالمية.
- يوفر اشعارت للمستخدمين وتنبيهات.

- يوفر تقارير تفصيلية حول أنواع مختلفة من الكوارث مثل:
  - الكوارث الطبيعية (زلازل، فيضانات...) – التهديدات المتعلقة بالبنية التحتية
  - الحوادث البيولوجية – الأزمات الاجتماعية

3.4.1 Globalincidentmap.com Global Incidents Map

هو موقع إلكتروني يعرض خريطة تفاعلية لحوادث وتهديدات أمنية من مختلف أنحاء العالم. يركّز على التقارير المتعلقة التهديدات الأمنية، الحوادث البيولوجية، اختطاف الطائرات، وأحداث أخرى تمس الأمن العام.

صورة 2: الصفحة الرئيسية للموقع **globalincidentmap.com**

♣ كيف يعمل الموقع؟

- يعرض الأحداث على خريطة زمنية–مكانية تفاعلية تُحدث بشكل مستمر.
- يتم تصنيف الحوادث في فئات متعددة (مثل: تهديدات أمنية، إطلاق نار، حوادث طيران، حرائق كبيرة، كوارث طبيعية، الخ).
- يمكن للمستخدمين تصفح الحوادث حسب الفئة، التاريخ، أو الموقع الجغرافي.
- يوجد عدد أنماط للموقع أحدها هو الحوادث، بينما يوجد أيضاً العديد من الأنماط
- منها الكوارث الطبيعية، أزمات الغذاء، الأزمات السيبرانية إلخ.

♣ كيف يجمع البيانات؟

- يعتمد على مزيج من المصادر الإخبارية الرسمية والمصادر المفتوحة مثل تقارير الشرطة، وكالات الأنباء العالمية، والمصادر عبر الإنترنت.
- يتم فلترة وتصنيف البيانات من خلال أنظمة ذكية بمساعدة محررين بشريين وخوارزميات تحليل نصوص.
- بعض الخرائط يتم تحديثها يدويًا، وبعضها يتم تغذيته عبر تدفقات بيانات شبه فورية.

♣ ما الفائدة من هذا النظام؟

- أداة قوية لمتابعة التهديدات العالمية والأحداث الأمنية لحظة بلحظة.
- تُستخدم من قبل الصحفيين، مراكز الدراسات الأمنية، وفرق إدارة الأزمات.
- تُوفر تنبيهات في الوقت الحقيقي للمستخدمين المسجلين.

الفصل الثالث

الدراسة النظرية

يوضح هذا الفصل بعض المفاهيم المستخدمة في هذا العمل.

1.1- مقدمة

3.1- منصات وسائل التواصل الاجتماعي

3.1- العنقدة Clustering

العنقدة، وهي عملية من تقنيات تعلم الآلة غير الخاضعة للإشراف unsupervised، حيث تقوم على تجزئة مجموعة من البيانات (العناصر) إلى مجموعات جزئية --أي عناقيد-- بحيث تكون العناصر الموجودة في العنقود الواحد متشابهة مع بعضها البعض، وتختلف عن العناصر الموجودة في المجموعات الأخرى. وتصنف طرق العنقدة بشكل عام في أربع مجموعات:(1) طرائق التجزئة Partitioning Based، حيث يتم تحديد عدد العناقيد مسبقا، أو (2) القائمة على الشبكة Grid Based، حيث يتم تقسيم فضاء العناصر إلى عدد محدد مسبقاً من الخلايا، أو (3) هرمية Hierarchical، حيث يتم تنظيم البيانات في مستويات متعددة، أو (4) قائمة على الكثافة Density Based، حيث يتم أخذ الكثافة بالاعتبار (Igartua et al., 2020).

1.3.1 العنقدة القائمة على الكثافة Density Based Clustering

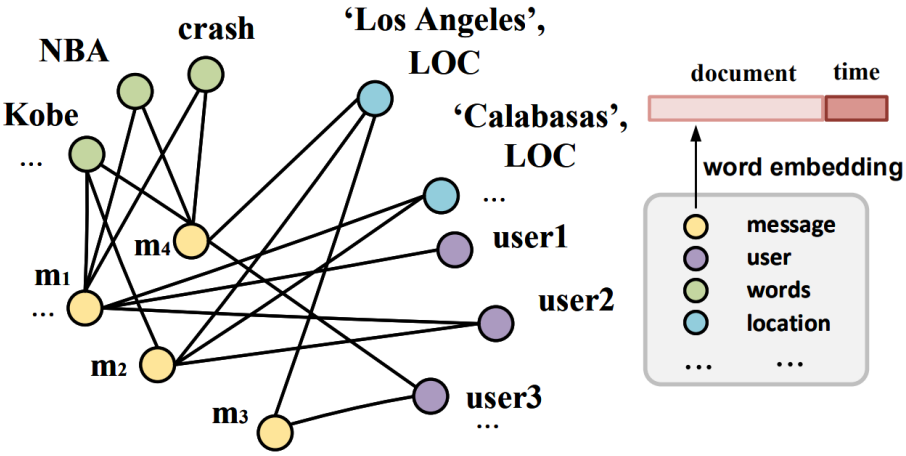
تنظم خوارزميات العنقدة القائمة على الكثافة البيانات في المناطق (العناقيد) حيث تكون العناصر كثيفة والتي تفصل عن يعضها بمناطق ذات كثافة عناصر منخفضة، والتي تعتبر ضجيج. وهذه الخوارزميات قادرة على (1) اكتشاف مجموعات من الأشكال الاعتباطية arbitrary، (2) التعامل مع المناطق المفرغة sparse (التي تعتبر مناطق ضجيج) و (3) العمل دون معرفة عدد العناقيد مسبقاً (Igartua et al., 2020). ومن بين المقترحات المختلفة في الأدبيات يوجد DBSCAN (التجميع المكاني القائم على الكثافة للتطبيقات ذات الضجيج Density-Based Spatial Clustering of Applications with Noise). DBSCAN هي خوارزمية شائعة للعنقدة القائم على الكثافة، تجمع النقاط المتراصة مع بعضها البعض، مع تمييز النقاط في المناطق منخفضة الكثافة كقيم شاذة. حيث وبخلاف K-means، لا تتطلب DBSCAN تحديد عدد العناقيد مسبقًا، ويمكنها تحديد العناقيد ذات الأشكال المختلفة. يعتمد أداؤها على معاملين: ε (نصف قطر الجوار) وMinPts (الحد الأدنى لعدد النقاط اللازمة لتكوين منطقة كثيفة). تتميز DBSCAN بفعالية خاصة في قواعد البيانات المكانية والتحليل الجغرافي المكاني (Ester et al., 1996).

3.1- العنقدة Online Clustering

تشير العنقدة الحية، والمعروفة أيضًا بالعنقدة التزايدية (Incremental Clustering)، أو العنقدة التدفقية (Streaming Clustering)، إلى تقنيات العنقدة التي تحدّث العناقيد ديناميكيًا مع وصول بيانات جديدة، دون الحاجة إلى إعادة معالجة مجموعة البيانات بأكملها. ويكتسب هذا أهمية خاصة في الحالات التي تتضمن بيانات واسعة النطاق أو آنية، مثل شبكات الاستشعار، والتوصيات عبر الإنترنت، وتدفقات وسائل التواصل الاجتماعي. تهدف خوارزميات مثل خوارزمية K-means التزايدية النسخ الحية من DBSCAN إلى التكيف بكفاءة مع انحراف المفاهيم وضمان قابلية التوسع مع ذاكرة محدودة (Aggarwal et al., 2003).

3.1- شبكات المعلومات غير المتجانسة (HIN)

شبكات المعلومات غير المتجانسة Heterogeneous Information Networks (HIN)، وهي حالة خاصة من البيان Graph، حيث يحتوي هذا على أكثر من نوع من العقد والوصلات. أي كما يلي، $G = (V, E, \mathcal{N}, T)$ حيث V و E هما مجموعات العقد و الوصلات و $\mathcal{N}$ و T هما مجموعات أنواع العقد والوصلات، حيث لكل عقدة نوع من $\mathcal{N}$ و لكل وصلة نوع من T، فعلى سبيل المثال في حالتنا، تكون الرسالة m وهي عقدة (من نمط رسالة) مرتبطة مع عقد الأماكن المذكورة في نصها (وهذه العقد تكون من نوع أماكن)، وتكون أيضاً هذه الرسالة مرتبطة مع عقد ال Hashtags المذكورة فيها (وهي عقد من نوع Hashtag)، وبذلك وبالمرور على كل الرسائل الواردة في هذه الشريحة المنية، نكون قد شكلنا هذه الشبكة. ويوضح الشكل أدناه مثالاً عن ذلك.



رسم توضيحي 2: مثال على شبكة معلزمات غير متجانسة.

2.1- التعرف على الكيانات المسماة Named Entities Recognition

يعرف الكيان المسمى Named Entity بأنّه كائن واقعي يملك تعريفاً مناسباً، ويمكن الإشارة له باسمه المناسب. حيث يمكن أن تكون الكيانات المسماة مكاناً أو شخصاً أو منظمةً أو عنصراً أو كياناً جغرافياً أو وقتاً. وكمثال عن الكيانات المسماة ف "المعهد العالي" هو كيان مسمى من نوع مؤسسة. وتتم عملية التعرف على الكيانات المسماة باستخدام أدوات وتقنيات معالجة اللغات الطبيعية، حيث تقوم بتصنيف مفردات النص المدخل إلى فئات محددة مسبقاً تمثل الكيانات المسماة.

3.1- Graph Neural Network

.

الفصل الرابع

الدراسة التحليلية

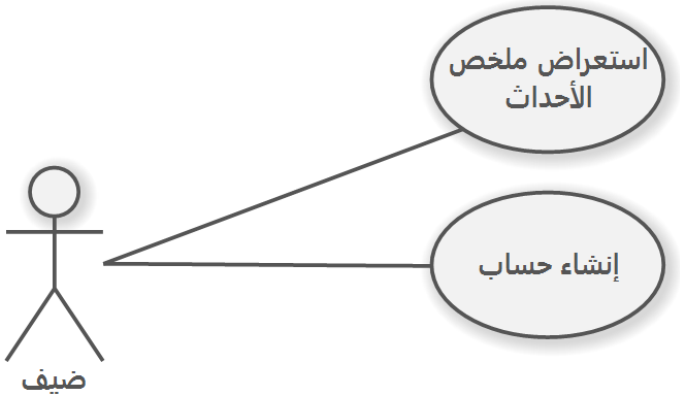
يقدم هذا الفصل تحليلاً للمتطلبات التي أوردناها في الفصل الأول.

1.1- مقدمة

نقدم في هذا الفصل دراسة تحليلية للمسألة المطروحة، ونبين حالات الاستخدام ووصفها السردي وبعض مخططات UML الداعمة.

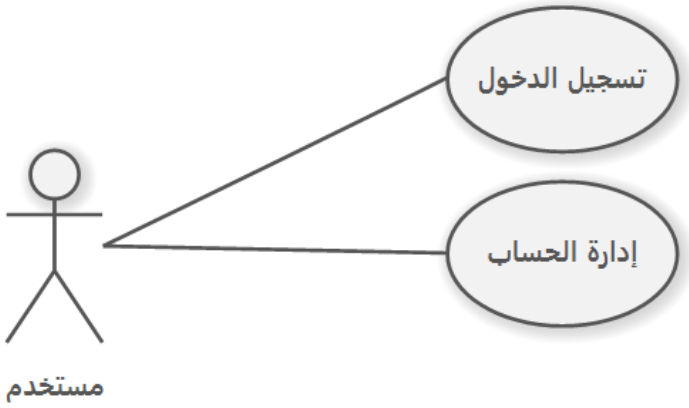
3.1- مخطط حالات الاستخدام

أ- حالات الاستخدام الخاصة بالضيف



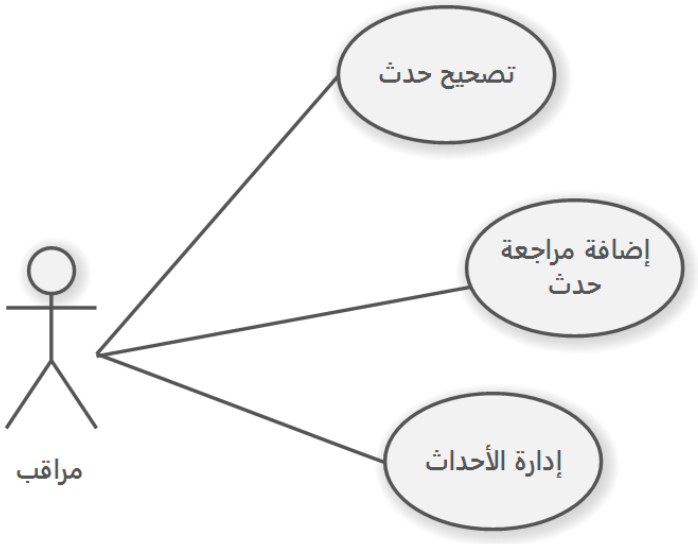
رسم توضيحي 3: مخطط حالات الاستخدام للضيف.

ب- حالات الاستخدام الخاصة بالمستخدم



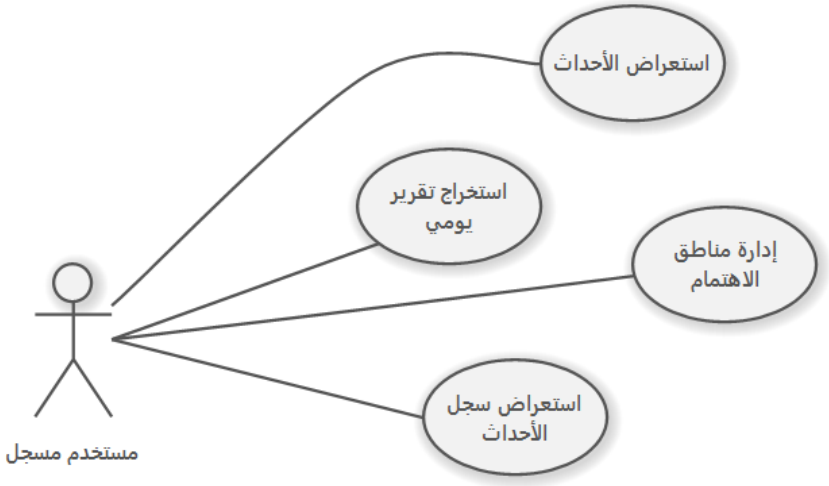
رسم توضيحي 4: مخطط حالات الاستخدام للمستخدم.

ت- حالات الاستخدام الخاصة بالمراقب



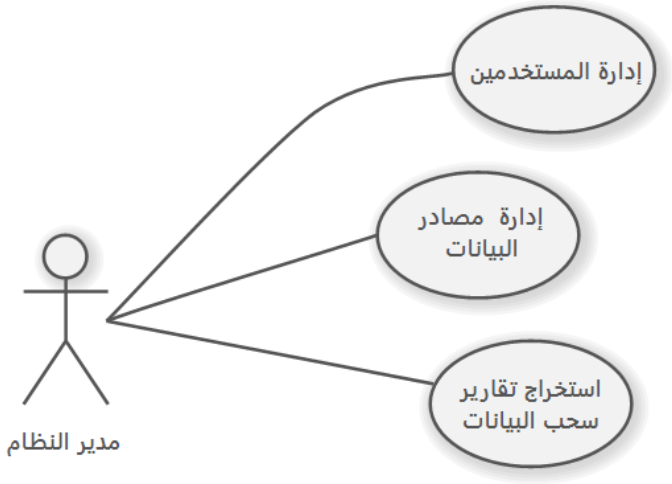
رسم توضيحي 5: مخطط حالات الاستخدام للمراقب.

ث- حالات الاستخدام الخاصة بالمستخدم المسجل



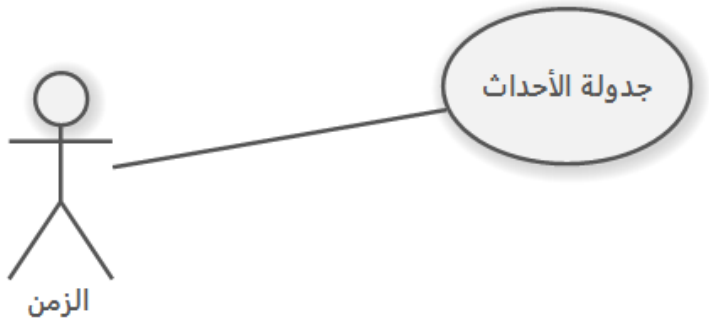
رسم توضيحي 6: مخطط حالات الاستخدام للمستخدم المسجل.

ج. حالات الاستخدام الخاصة بمدير النظام



رسم توضيحي 7: مخطط حالات الاستخدام لمدير النظام.

ح. حالات الاستخدام التي يفعلها الزمن



رسم توضيحي 8: مخطط حالات الاستخدام التي يفعلها الزمن.

4.1- السرد النصي لحالات الاستخدام

نوضح هنا السرد النصي لحالات الاستخدام الرئيسية وهما جدولة الأحداث، إدارة مناطق الاهتمام، استعراض الأحداث.

أ. حالة الاستخدام جدولة الأحداث

اسم حالة الاستخدام: **جدولة الأحداث**	
الفاعلون الأوليون	الزمن
الشروط المسبقة	• لايوجد.
الشروط اللاحقة	• تمت جدولة الأحداث التي حصلت اليوم، أي إزالتها من قائمة الأحداث المعروضة وإضافتها إلى سجل الأحداث التاريخية.
الوصف	يتم جدولة الأحداث التي حصلت اليوم وإضافتها إلى سجل الأحداث، وذلك في ساعة محددة مسبقاً .

سير الأحاث

السيناريو الرئيسي الناجح – Main Success scenario

جدول 4: السيناريو الناجح لحالة الاستخدام جدولة الأحداث

النظام	الزمن
	1. يبدأ عملية جدولة الحدث.
2. يستخرج قائمة الأحداث اليومية	
3. يضيف قائمة الأحداث اليومية إلى سجل الأحداث	
4. يستخرج التقرير اليومي للأحداث	
5. يرسل إشعاراً إلى مدير النظام بنجاح الجدولة، ويرفق معه التقرير المستخرج.	
6. ينهي النظام علية الجدولة.	

المسارات البديلة

A1: في المرحلة رقم 2-أو 3 أو 4، في حال فشل إحدى عمليات الاستخراج، فإنّ النظام يرسل إشعاراً إلى مدير النظام بحدوث خطاء ويرفق العملية التي تسببت بالخطاء، ومن ثم يعيد المحاولة بعد الانتظار لزمن محدد، ويتابع من المرحلة رقم 1. وفي حال تكرار المحاولة أكثر من ثلاث مرات، فعندئذٍ ينتقل إلى حالة الخطاء E1.

المسارات الخاطئة

E1: في حال تكرار السيناريو البديل أكثر من ثلاث مرات، فعندئذٍ ينهي النظام عملية الجدولة.

ب. حالة الاستخدام استعراض الأحداث

اسم حالة الاستخدام: استعراض الأحداث	
الفاعلون الأوليون	المستخدم المسجل
الشروط المسبقة	• المستخدم مسجل بالنظام، ومعروف لدى النظام، أي تم تنفيذ حالة الاستخدام "تسجيل الدخول" بنجاح.
الشروط اللاحقة	• لا يوجد.
الوصف	يقوم المستخدم المسجل باستعراض الأحداث على الخريطة أو بشكل قائمة، والاطلاع على ملخصها، واختيار أحدها والاطلاع عليه بشكل مفصل.

سير الأحاث

السيناريو الرئيسي الناجح – Main Success scenario

جدول 5: السيناريو الناجح لحالة الاستخدام استعراض الأحداث

النظام	**المستخدم المسجل**
	1. يقوم المستخدم ببدء عملية الاستعراض
2. يعيد قائمة بأنماط العرض الممكنة (خريطة، قائمة، تقرير report).	
	3. يختار نمط العرض المراد.

4. يعيد مجموعة الأحداث اليومية وفق نمط العرض المختار.	
	5. يستعرض الأحداث، وبعد ذلك ينهي عملية الاستعراض

المسارات البديلة

A1: في المرحلة رقم 5، إذا أراد المستخدم استعراض حدث ما بشكل تفصيلي فيتابع كما يلي ويكمل من المرحلة رقم 5:

	1. يختار الحدث الذي يريده، ويرسل معرفه
2. يرسل معلومات الحدث بشكل تفصيلي (ملخص، تقييم، مراجعات المراقب، ...)	

المسارات الخاطئة

لا يوجد.

ت. حالة الاستخدام إدارة مناطق الاهتمام

اسم حالة الاستخدام: إدارة مناطق الاهتمام	
الفاعلون الأوليون	المستخدم المسجل
الشروط المسبقة	• المستخدم مسجل بالنظام، ومعروف لدى النظام، أي تم تنفيذ حالة الاستخدام "تسجيل الدخول" بنجاح.
الشروط اللاحقة	• لا يوجد.
الوصف	يقوم المستخدم المسجل باستعراض المناطق التي يهتم بتلقي اشعارات بالأحداث التي تقع بها، إضافة منطقة جديدة ، أو إزالة منطقة اهتمام.

سير الأحاث

السيناريو الرئيسي الناجح – Main Success scenario

جدول 6: السيناريو الناجح لحالة الاستخدام إدارة مناطق الاهتمام

النظام	المستخدم المسجل
	1. يقوم المستخدم ببدء عملية الاستعراض
2. يعيد النظام قائمة المناطق التي يهتم بها المستخدم.	
	3. ينهي المستخدم عملية الاستعراض.

المسارات البديلة

A1: في المرحلة رقم 3، إذا أراد المستخدم إضافة منطقة جديدة، فيتابع كما يلي ومن ثم يكمل من المرحلة رقم 3:

	1. يختار بدء عملية إضافة منطقة اهتمام
2. يعيد خريطة المناطق، والنطاقات الممكنة.	
	3. يختار المنطقة والنطاق (الإحداثيات، ونصف القطر)
4. يضيف النظام المنطق الجديدة إلى قائمة المناطق.	
	5. ينهي عملية إضافة المنطقة.

A1: في المرحلة رقم 3، إذا أراد المستخدم إزالة منطقة جديدة، فيتابع كما يلي ومن ثم يكمل من المرحلة رقم 3:

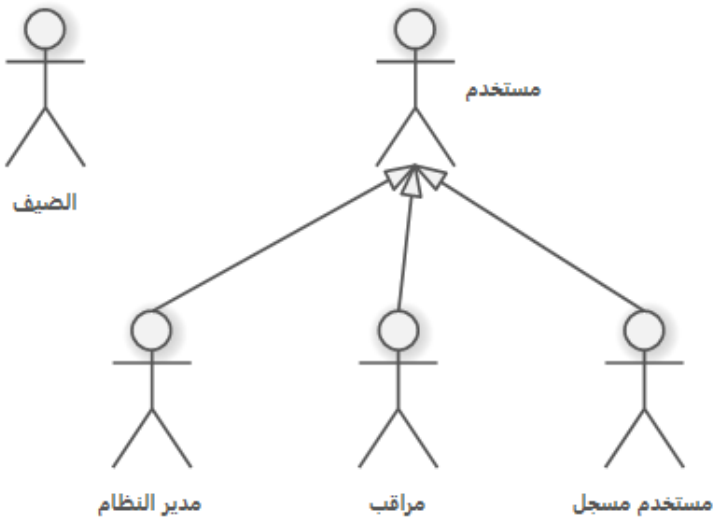
	1. يختار بدء عملية إزالة منطقة اهتمام
2. يعيد قائمة المناطق المهتم بها	
	3. يختار المنطقة التي يريد إزالتها
4. يزيل النظام المنطقة من قائمة المناطق.	
	5. ينهي عملية الإزالة المنطقة.

المسارات الخاطئة

لا يوجد.

4.1- العلاقات بين الفاعلين

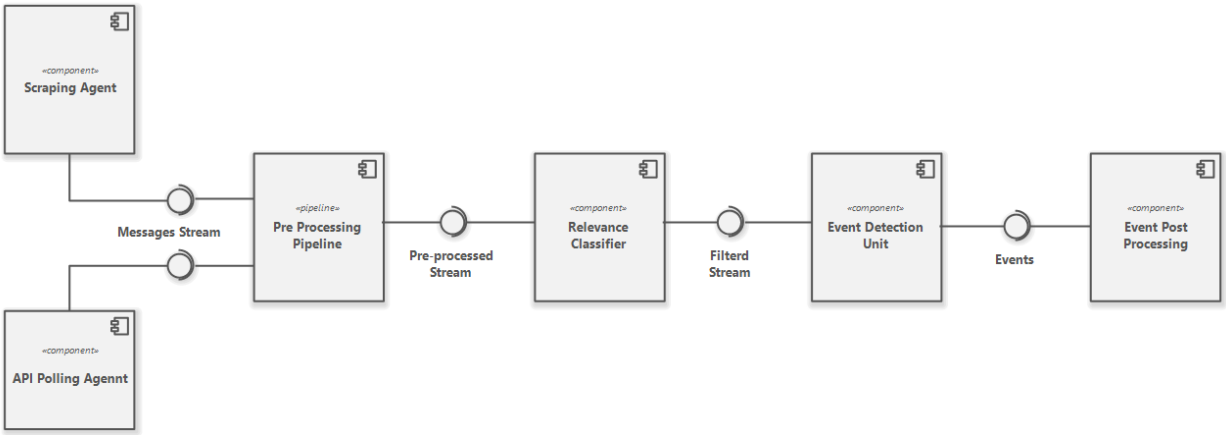
يوضح المخطط أدناه العلاقات بين الفاعلين في هذا النظام.



رسم توضيحي 9: العلاقات بين الفاعلين.

4.1- مخطط المكونات

يساعد المخطط أدناه في فهم مكونات النظام، حيث يتكون من عدة مكونات، وهي مكونان من أجل سحب البيانات الأول يقوم بسحب البيانات من خلال الاستقصاء النشط لـ API المقدمة من وسائل التواصل (تويتر وتيلغرام) والثاني يقوم بسحبها من خلال تقليد سلوك مستخدم حقيقي، ومن ثم يقدمان دفق البيانات للمكون التالي الذي يقوم بتطبيق المعالجات المسبقة على دفق الرسائل ومن ثم يأتي دور المكون الثالث الذي يقوم بفلترة الدفق بالاستعانة بالمصنف المدرب على التعرف على الرسائل ذات الصلة، ويمررها بدوره لمكون كشف الأحداث، الذي يصنفها إلى عناقيد ويمررها إلى المكون الأخير الذي يقوم بمعالجة لاحقة عليها. وهذا يوضح مكونات النظام بشكل عام ومجرد بعيداً عن التفاصيل التقنية.



الشكل 5: مخطط مكونات النظام.

الفصل الخامس

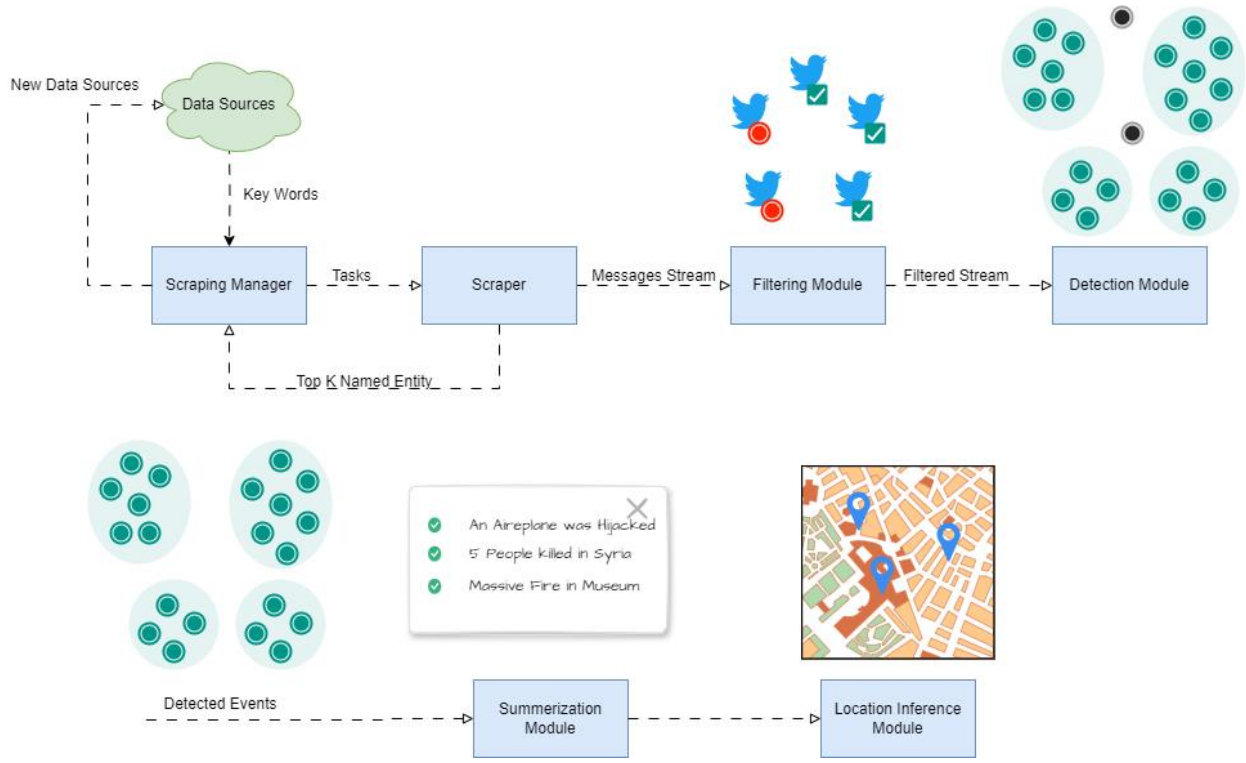
المنهجية المقترحة

يعرض هذا الفصل منهجية العمل المقترحة لتحقيق النظام بناءً على المنهجيات التي عرضناها في الفصل الثاني.

1.1- مقدمة

بعد الاطلاع على الأبحاث في مجال كشف الأحداث من وسائل التواصل الاجتماعي، تم اتخاذ المنهجية التي سنورد شرحها أدناه، حيث اعتمدنا على مقاربة معتمدة على المستند (Document Pivot Approach)، ولم نلجأ إل العائلة الأخرى (Feature Pivot Approaches)، لأن ذلك النمط من المقاربات لن يكشف الأحداث الجديرة بالذكر غير ذات الأثر الكبير. وبالإضافة على ذلك وبعد اطلاعنا على تلك الأبحاث وجدنا أنّ معظم الأبحاث قد لجأت إلى استخدام مجموعة بيانات خاصة بها، وأنّه لا يوجد مجموعة باللغة العربية تدعم مسألتنا، حيث أن المجموعة الوحيدة المتوفرة باللغة العربية كانت KAWARITH 7 وهي تضم فقط سبعة أحداث ذات أثر كبير (كزلزال سورية 2023)، لذلك هي غير مناسبة، وبالإضافة إلى ذلك كان لابد من استخدام بيانات عربية لضبط معاملات النماذج والخوارزميات المستخدمة، لذلك قمنا بجمع مجموعة البيانات الخاصة بنا واستخدامها. كما اعتمدنا على فصل منهجية سحب البيانات عن منهجية كشف الأحداث مما ساعد على جعل قناة دفق الأحداث أقل ضجيجاً (30% من التغريدات فقط غير ذي صلة).

ويوضح الشكل أدناه البنية العامة للمنهجية التي سنعمل بها.

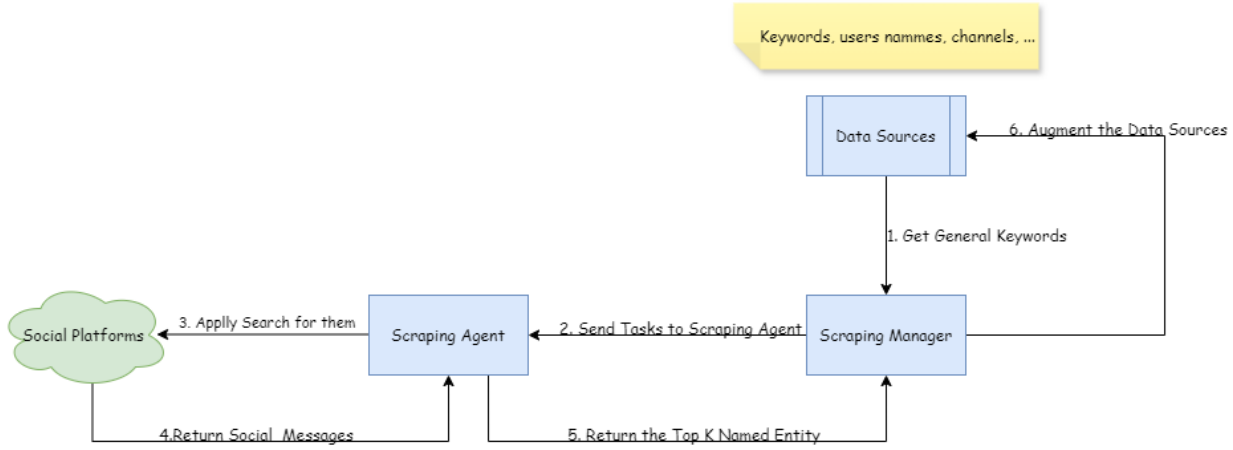


الشكل 6: البينة العامة للمنهجية.

حيث تتكون منهجيتنا من خمسة وحدات، (1) وحدة البحث واستخراج دفق المعطيات، (2) وحدة فلترة دفق الرسائل، (3) وحدة الكشف، (4) وحدة التلخيص، (5) وحدة الاستدلال على الموقع.

وحدة البحث واستخراج دفق الرسائل

مسؤولية هذا الجزء، هو استخراج دفق الرسائل، من خلال الاستفادة من مصادر البيانات (كلمات مفتاحية، أسماء مستخدمين، قنوات، ...)، وإرسالهم إلى خدمة سحب البيانات، التي نستفيد من نتيجتها في تغذية مصادر البيانات (Feedbacks)، من أجل تعزيز دفقة الرسائل، كما هو موضح بالشكل أدناه.



الشكل 7: منهجية البحث واستخراج الرسائل

تبدأ العملية بأخذ قائمة من الكلمات المفتاحية العامة (مثل، سوريا، دمشق، حادثة مروعة، ...)، وإرسالها إلى خدمة سحب البيانات التي بدورها تسحب البيانات وترسلها إلى المرحلة التالية وهي مرحلة الكشف وفي نفس الوقت تعيد إلى خدمة الـ Scraping Manager قائمة بأكثر K كيان مسمى مكرر في نتائج البحث (مثل، مشفى ابن النفيس، قرية بلابلابلا، الشابة فلانة، ...) التي بدورها تستفيد منهم لتغذية مصادر البحث وتوليد كلمات مفتاحية جديدة وإرسالها من جديد للبحث عنها. حيث بهذا الشكل نكون قد غذينا مصادر البيانات بمصادر رائجة حالياً (بناءً على فرضية أنّه عند وقوع حادث في قرية او مكان ما فإنّ اسمه سيرد بشكل أكثر من بقية الكيانات المسماة.

وحدة فلترة دفق الرسائل

مسؤولية هذا الجزء هو فلترة الرسائل وإزالة الرسائل غير المرغوب بها (مثل الإعلانات والمنشورات الشخصية غير ذات الصلة بالمسألة)، وفي هذا الجزء اعتمدنا على بناء مصنف ثنائي، باستخدام المجموعة المنمطة التي جمعناها، حيث لجأنا لبناء هذا المصنف إلى نماذج تعلم الآلة التقليدية مثل (SVM, KNN, ...) لكون مجموعة البيانات غير كافية لاستخدام نماذج كبيرة ومن

أجل بناء مصنف صغير بحيث يتم دمجه مع خدمة سحب البيانات لنحصل على وكيل سحب بيانات ذكي (Scraping Agent).

وحدة كشف الأحداث

مهمة هذا المكون هو عنقدة الرسائل (التغريدات)، بحيث تجمع الرسائل التي تناقش (تقدم معلومات) عن أمر ما، ضمن عناقيد بحيث يعبر العنقود عن حدث ما، في هذه الجزئية وبعد اطلاعنا على الأبحاث الأخرى، تبين لنا أنّ المشكلة التي تكمن في هذه الجزئية هي بتمثيل الرسائل بأشعة تحمل معلومات تمييزية (Discriminative Representation)، وبعد ذلك يتم عنقدة الرسائل باستخدام خوارزمية عنقدة تزايدية.

ومن أجل العنقدة التزايدية يوجد خوارزميتين مستخدمتان في الأبحاث التي تم الاطلاع عليها وهما SinglePass وتم استخدامها في الأبحاث (Abagissa et al., 2024; Alsaedi & Burnap, 2015; Guo et al., 2024; Ray et al., 2024)، وDBSCAN التي تم استخدامها في الأبحاث (Y. Cao et al., 2021; Ren et al., 2022; Yu et al., 2024).

ومن أجل طرائق التمثيل المستخدمة فإنّ الطرائق المتبعة هي Word2Vec, LDA(Latent Dirichlit Allocation) BERT, SBERT,WMD,.

وحدة تلخيص الحدث

مهمة هذا المكون هو تلخيص مجموعة الرسائل وإعطاء ملخص للحدث، وفي هذا الجزء اعتمدنا على نموذج لغة ضخم لتلخيص الرسائل، كما بالإضافة إلى ذلك ولأنه قد يكون هنالك عدد كبير من التغريدات ضمن الحدث الواحد لذلك ليس من المنطقي أن نمررها كلها سوية، لأن ذلك غير ممكن بسبب محدودية طول الأمر Prompt المرسل إليه، ولأنها قد تحوي على معلومات مكررة، لذلك لجأنا إلى أخذ طول التغريدة، مدى الاهتمام بها (التفاعلات)، المفردات الجديدة المستخدمة، التشابه مع تغريدات في نفس الحدث، لتعريف تابع ترتيب للتغريدات ضمن الحدث، من أجل أخذ أفضل K تغريدة حسب سعة نموذج اللغة الضخم.

وحدة الاستدلال على الموقع

مهمة هذا المكون هو أخذ مجموعة التغريدات ضمن الحدث والاستدلال على الموقع المذكور فيها وإعادة إحداثياته الجغرافية، وفي هذا الجزء اعتمدنا على مجموعة البيانات IDRIS (Suwaileh et al., 2023)، من أجل تدريب ووضع تقييم لجودة النموذج، ومقارنته مع إمكانية استخدام نماذج اللغة الضخمة في هذه المهمة.

الفصل السادس

تصميم النظام

يوضح هذا الفصل تصميم خدمات النظام والأنماط التصميمية المرتبطة بها.

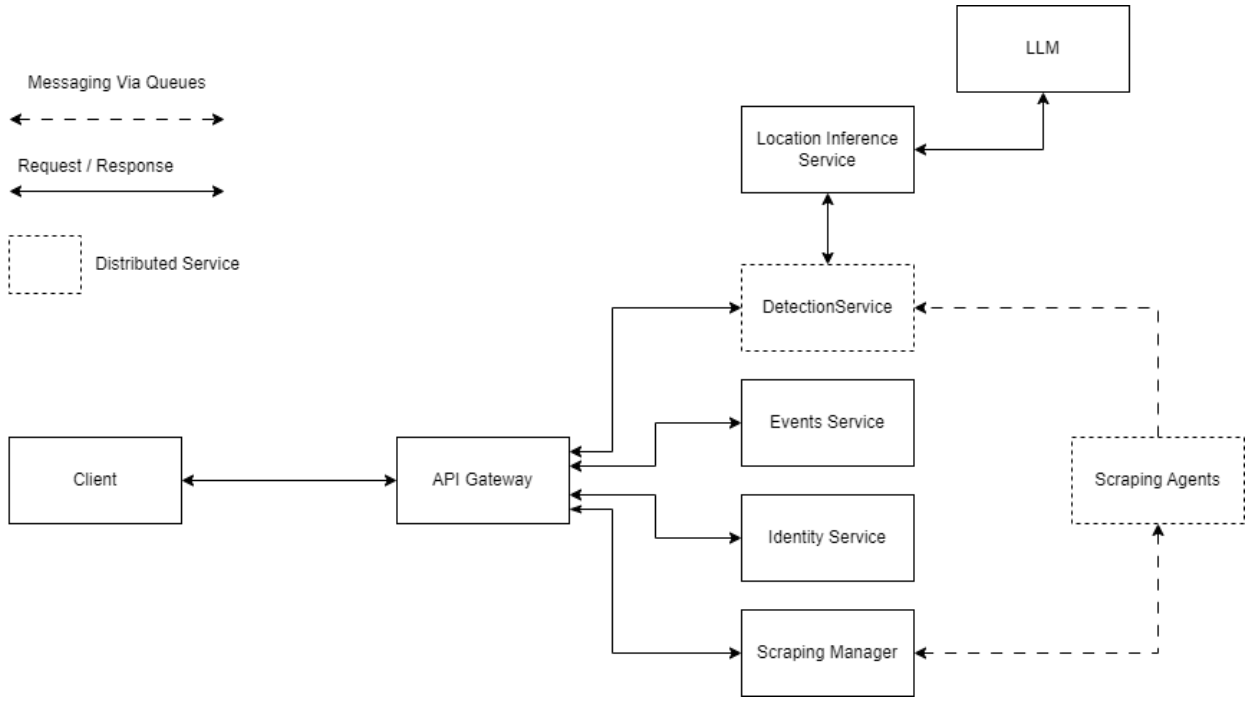
1.1- مقدمة

تم اعتماد بنية الخدمات المصغرة (micro-services) لبناء النظام، حيث يتكون انظام من مجموعة من الخدمات التي تعمل مع بعضها بشك متكامل، وتم اعتماد هذه البنية بسبب الحاجة لتطبيق قابل للتوسع بسهولة وقادر على العمل مع دفق كبير من المعطيات ومعالجتها، ولأننا وبعد تحليلنا للنظام في الفصل الثالث وبعد دراستنا المرجعية في الفصل الثاني تبين لنا أن النظام قادر على التعامل مع المعطيات بشكل متوازٍ وقابل للفصل بين مكوناته بسهولة بحيث لا يوجد تتداخل فيما بينها، وأيضاً أنه يتعامل مع معطيات كبيرة فهو إذن بحاجة للتوسع أفقياً. لذلك اخترنا تلك البنية فهي تقدم الفوائد التالية:

- التوسع المستقل: يمكن توسيع كل خدمة مصغرة بشكل مستقل بناءً على الطلب، حيث تسمح هذه المرونة باستخدام الموارد بكفاءة، إذ يتم توسيع أجزاء التطبيق التي تحتاج إلى المزيد من الموارد فقط، بدلاً من النظام كله.
- المرونة في استخدام التقنيات: يمكن استخدام لغات البرمجة وأطر العمل وقواعد البيانات والتقنيات المختلفة للخدمات المختلفة، وتحسين كل خدمة بناءً على متطلباتها المحددة.
- النشر المستقل: يمكن نشر الخدمات المصغرة بشكل مستقل، مما يسمح بالتكامل المستمر والتسليم المستمر (CI/CD) مما يعطي إمكانية إصدار التحديثات وإصلاح الأخطاء على حدة.

ويوضح الشكل أدناه المخطط التصميم للنظام حيث نعرض فيه الخدمات وهي، خدمة تجريف (كشط) البيانات والتي تقوم بسحب البيانات بناءً على المهمات التي تردها من خدمة إدارة عمليات التجريف، ومن ثم تقوم بنشر البيانات المسحوبة (عبر رتل تراسل) إلى خدمة كشف الأحداث المرتبطة بمجالها، والتي بدورها تتواصل مع خدمة الاستدلال على الموقع مع نموذج لغة كبير (من أجل تعيين الموقع، وإعطاء ملخص للحدث) ومن ثم ترسل الأحداث المكتشفة إلى خدمة الأحداث التي بدورها تتولى عملية إدارة الأحداث وتفضيلات المستخدمين والإشعارات. كما ويوجد خدمة الهوية التي تتولى عملية إدارة المستخدمين والصلاحيات، وأيضاً خدمة طرفية API (API Gateway) التي تجعل المستخدم النهائي يتعامل مع النظام كما لو أن قطعة واحدة. وفي النهاية يتفاعل العميل مع خدمة الواجهات ويرسل الطلبات إلى طرفية النظام.

3.1- مخطط خدمات النظام التصميمي



الشكل 8: مخط خدمات النظام التصميمي.

حيث يشير السهم المنقط إلى أن التواصل يكون عبر رتل تراسل (يمكن أن يكون Kafka Queue أو Rabbit MQ).

ويشير الصندوق المخطط الحواف، إلى أن هذه الخدمة تعمل بشكل موزع (أي على سبيل المثال تكون كل نسخة من خدمة كشف الأحداث مرتبطة مع مجال معين أي مثلاً خدمة كشف أحداث مرتبطة مع مجال "أخبار محلية ضمن دمشق").

وبالنسبة لخدمة المجرف وخدمة إدارة عمليات السحب فهما على النمط السيد–العبد (master-slave) حيث تتولى خدمة إدارة عمليات النظام مهمة إرسال الأعمال إلى خدمات التجريف وهم بدورهم العمال.

ويشير السهم المتصل إلى أن التراسل يكون عبر طلبات–ردود مباشرة متزامنة.

2.1- تصميم خدمات النظام

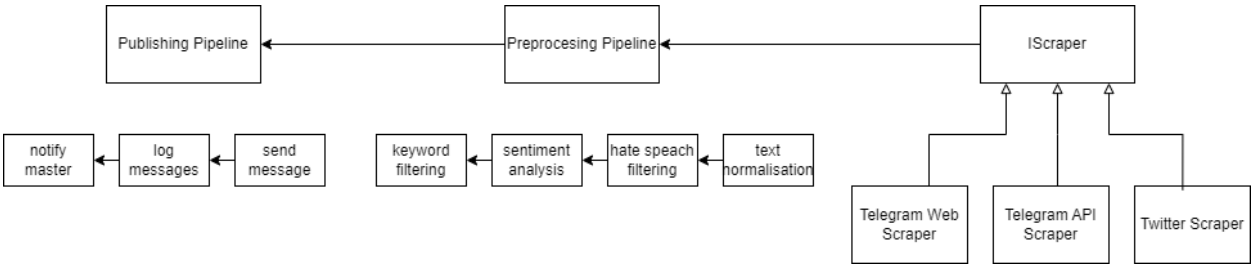
نعرض فيما يلي تصميمنا المقترح لخدمات النظام كل على حدة.

1) خدمة التجريف Scraping Service (Scraping Agent)

تقوم هذه الخدمة بسحب المهمات (وهي قائمة من مصادر البيانات المرتبطة بمجال معين ومنصة تواصل محددة) من رتل التراسل، وتقوم بتنفيذها كما يلي:

تقوم بتحميل عميل التجريف Scraping Agent الموافق لمنصة التواصل التي تتبع لها المهمة، ومن ثم يبدأ بسحب البيانات.

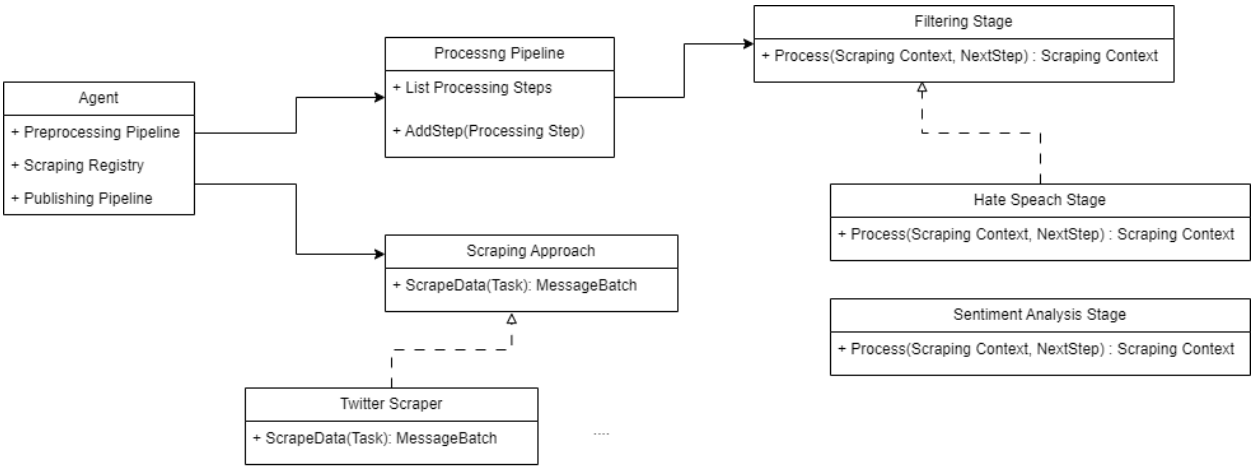
عندما يسحب دفعة من البيانات، ترسل إلى خط أنابيب المعالجة المسبقة ومن ثم إلى خط أنابيب النشر (Pre Processing Pipeline & Publishing Pipelin). ويوضح الشكل أدناه تلك العملية.



الشكل 9: مراحل العمل ضمن خدمة تجريف البيانات.

وفي هذه الخدمة يتم استخدام النمطين التصميميين Pipes and Filters وStrategy Pattern.

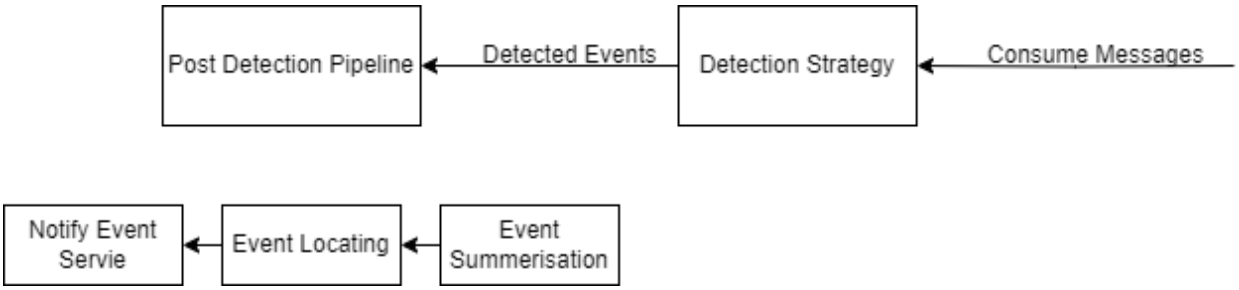
ويوضح الشك أدناه مخطط الصفوف للخدمة.



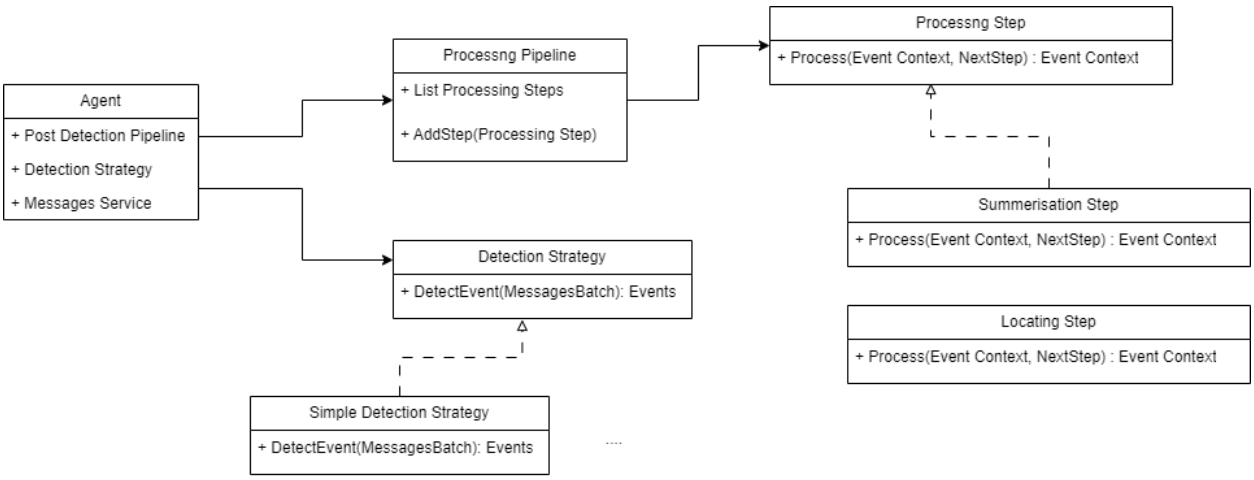
الشكل 10: مخطط الصفوف لخدمة تجريف البيانات.

2) خدمة كشف الأحداث Event Detection Service

تقوم هذه الخدمة بقراءة دفعات الرسائل messages batches من رتل التراسل، وتقوم باستخدام خوارزمية كشف الأحداث، بكشف الأحداث الجديدة وتعديل الأحداث القديمة (في حال أتت رسائل جديدة لها)، ومن ثم كلما اكتشفت دفعة أحداث، تقوم بتنفيذ خط أنابيب المعالجة اللاحقة، حيث تقوم بتعيين موقع الحدث من خلال التواصل مع خدمة الاستدلال على الموقع، ومن ثم كتابة ملخص الحدث بالاستعانة بنموذج لغة ضخم، ... إلخ. وبعد ذلك ترسل الحدث ورسائله إلى خدمة الأحداث. كما أن هذه الخدمة تعمل بشكل موزع حيث تكون مرتبطة بمجال معين تقرأ رسائله من رتل تراسل خاص به. ويوضح الشكل أدناه مخطط مكونات هذه الخدمة.

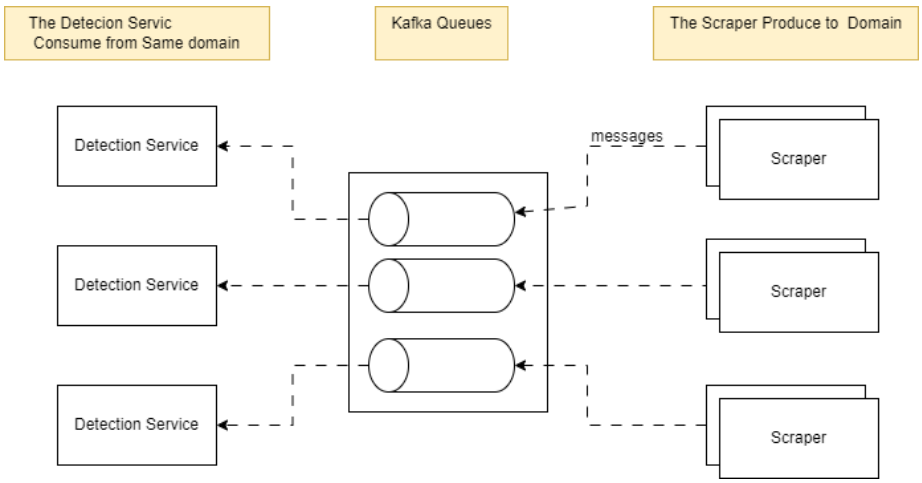


ويوضح الشكل أدناه مخطط الصفوف لها.



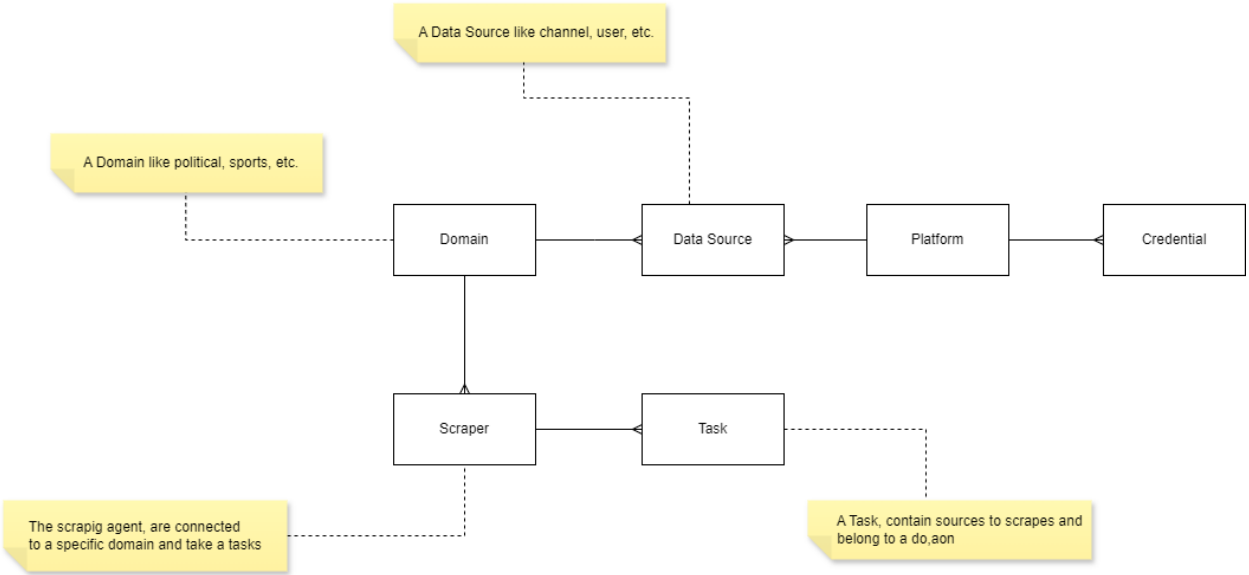
وعند معاينة هيكليتي تلك الخدمتين، نكون قد حققنا النمط Vertical Slice Architecture، أي تقسيم النظام عمودياً إلى خدمات مختلفة مستقلة عن بعضها البعض (فصل المكونات والتداخل بين ملفاتهما) ومتشابهة بمعماريتها الداخلية أي لكلا الخدمتين نفس البينة المعمارية.

كما ويوضح الشكل أدناه طريقة التواصل وتوزيع خدمات الكشف وتجريف البيانات، حيث وكما نرى لكل مجال، ما عدد من نسخ خدمات تجريف البيانات التي ترتبط بهذا المجال وتقوم بسحب البيانات ووضعها في رتل التراسل الخاص بهذا المجال، ومن ثم تقوم نسخة خدمة الكشف المرتبطة بهذا المجال بأخذ البيانات ومعالجتها. كما هو موضح أدناه.



3) خدمة إدارة عمليات السحب Scraping Management Service

هذه الخدمة مسؤولة عن توليد المهمات وجدولتها، وإرسالها إلى رتل التراسل مع العمال، الذين بدورهم وعند انتهائهم من مهمة ما، يقومون بسحب مهمات أخرى من رتل التراسل وكذا دواليك. ويوضح الشكل أدناه مخطط مبسط للعلاقات مابين الكيانات ERD في هذه الخدمة.

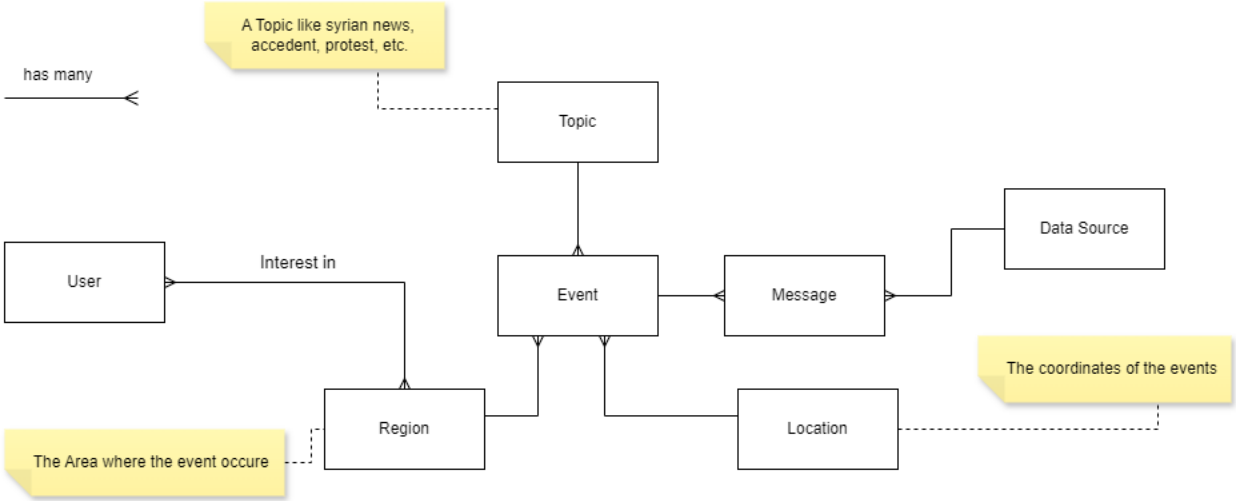


حيث وكما نرى فإنّ لكل مجال بيانات (أخبار سياسية سورية، أخبار رياضية، أحداث محلية ضمن مدينة) مرتبط مع عدد من مجرفي البيانات، الذين يستلمون المهمات. وكل مجال أيضاً مرتبط بمجموعة من مصادر البيانات (كلمات مفتاحية، معرفات قنوات، ... إلخ)، ومصدر البيانات مرتبط مع منصة تواصل اجتماعي ما. ولكل منصة تواصل معلومات لتسجيل الدخول Credential. ولكل مجرف عدد من المهمات.

(لم أنتهي بعد من تصميم الخدمة)

4) خدمة إدارة الأحداث Events Service

تتولى هذه الخدمة عملية إدارة الأحداث (يتخاطب معها العميل للحصول على معلومات الحدث، وتتخاطب معها خدمة كشف الأحداث لتسجيل الأحداث) حيث تدير عمليات إرسال الإشعارات وإدارة مناطق اهتمام المستخدمين، والمواضيع. ويوضح الشكل أدناه مخط العلاقات بين الكيانات ERD لهذه الخدمة.



حيث وكما نرى، فإن للحدث موضوع (سياسي، رياضي، أزمة، حوادث، ... إلخ)، وموقع (إحداثيات)، ومنطقة حدث بها. كما وله عدة رسائل تتحدث عنه وهذه الرسائل تتبع لمنصة تواصل ما. وأيضاً فإن المستخدم يبدي اهتمامه بعدة مناطق.

(لم أنتهي بعد من تصميم الخدمة).

5) خدمة الاستدلال على الموقع Location Inference Service

مهمة هذه الخدمة هي استقبال طلبات الاستدلال على الموقع وهي إما رسالة نرغب في معرفة الموقع المذكور بها أو حدث (مجموعة رسائل) نرغب بمعرفة موقعه، أو عنوان (اسم الموقع) نرغب بتحديد إحداثياته. في هذه لخدمة نحتاج إلى النمط التصميمي Strategy، لتعريف أنماط التعرف على الموقع المختلفة (باستخدام نموذج لغة ضخم أو أدوات NER). وقد تتواصل هذه الخدمة مع نموذج لغة ضخم (حسب طريقة الاستدلال).

6) خدمة الهوية Identity Service

تتولى هذه الخدمة مهمة إدارة المستخدمين وصلاحياتهم وأدوارهم.

7) خدمة البوابة Gateway Service

بعد النظر إلى الخدمات التي تكون النظام الكلي، نجد أنّ تواصل النظام مع كل خدمة على حدة له مجموعة من المشاكل مثل عدم وجود واجهة موحدة تمثل النظام يستطيع المستخدم التخاطب معها والتعامل مع النظام ككتلة واحدة، فلذلك ظهرت الحاجة لوجود مخدم يعمل كواجهة تخاطبية للنظام ويوجه الطلبات للخدمات الداخلية حسب الطلب.

8) خدمة الواجهة الأمامية Frontend Service

هذه الخدمة مسؤولة عن التفاعل والتعامل مع المستخدم.

الفصل السابع

الأدوات المستخدمة

نعرض في هذا الفصل الأدوات وأطر العمل المستخدمة في تنجيز النظام.

1.1- مقدمة

.

2.1-

الفصل الثامن

التجارب ومقارنة النماذج

نعرض في هذا الفصل التجارب التي قمنا بها، والنماذج المقترحة وتقييمها وضبطها.

1.1- مجموعة البيانات

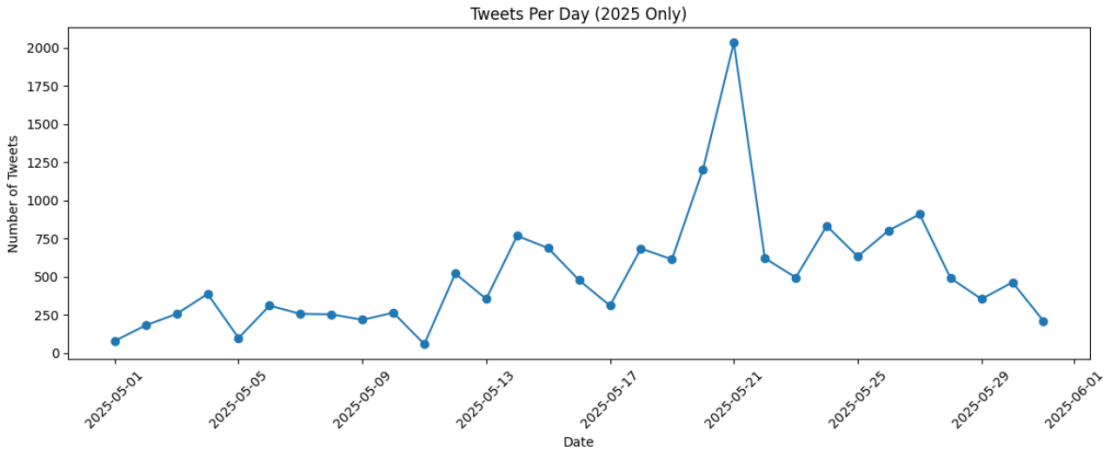
تم جمع مجموعة البيانات من منصة تويتر عبر البحث بكلمات مفتاحية عامة وخاصة (مثل، سوريا، دمشق، اعتصام الميدان، ...)، من تاريخ 1 أيار ولغاية 31 أيار 2025، ومن خلالها جمعنا مجموعة بيانات حجمها 20 ألف تغريدة (23 ألف قبل إزالة التكرارات) بالإضافة على معلوماتها metadata. سميناها SAS (Golden)، وبعد ذلك قمنا بتنميط مجموعة جزئية منها يدوياً بحجم 4500 تغريدة إلى ذات صلة وغير ذات صلة (Relevant, Un Relevant)، وتم ذلك وفق دليل التنميط المبين في الملحق ب (اضغط هنا) سميناها SAS (Labeled).

بعد ذلك تم النظر إلى الجزء من مجموعة البيانات الذي تم جمعه من خلال البحث بكلمات مفتاحية خاصة (أحداث وقعت في شهر أيار، مثل، اعتصام في مكان فلاني، رفع العقوبات الأميركية، حادث في مكان فلاني، ...)، وحجمه هو عبارة عن 6998 تغريدة بعد إزالة التكرارات، تنتمي إلى 109 أحداث وقعت في شهر أيار، تم وضع كل تغريدات ناتجة عن بحث معين في ملف منفصل (مبوبة)، وحيث بعد النظر إلى تلك النتائج (المبوبة ضمن ملف لكل حدث) كانت نسبة أكثر من 98% من التغريدات منتمية لذلك الحدث، وبالتالي كان باستطاعتنا استخدام المقاييس (NMI، AMI، ARI) المبينة في فصل الدراسة المرجعية لتقييم جودة نموذج كشف الحدث. وتمت تسمية هذا الجزء من المجموعة بـ SAS (Silver). ويوضح الجدول أدناه مثالاً على تغريدات من نفس الحدث.

جدول 7: مثال على تغريدات ضمن حدث ما.

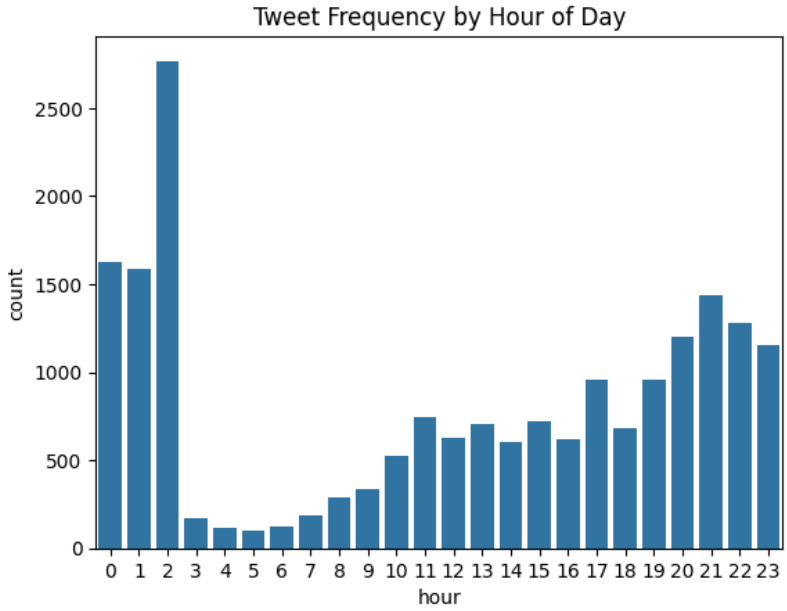
Author	Type	Tweet Text	Tweet ID
Grok	Reply	رامز الخياط هو رجل أعمال قطري، الرئيس التنفيذي لشركة أورباكون القابضة، وهي جزء من مجموعة باور إنترناشيونال القابضة التي يملكها ويديرها عائلة الخياط في قطر. تأسست أورباكون عام 2011، وتتخصص في مشاريع الطاقة والبنية التحتية. الخ	19281218
الإخبارية السورية	Tweet	مؤتمر صحفي مشترك لوزير الطاقة محمد البشير والرئيس التنفيذي لشركة أورباكون رامز الخياط على هامش توقيع الاتفاقيات الاقتصادية مع شركات الطاقة #سوريا_نحو_الريادة #الإخبارية_السورية	92811483
شبكة أخبار سوريا	Tweet	مؤتمر صحفي لوزير الطاقة المهندس محمد البشير والرئيس التنفيذي لشركة أورباكون القابضة رامز الخياط في قصر الشعب بعد توقيع مذكرات تفاهم مع شركات دولية في مجال الطاقة.	19281069
ssy	Tweet	الرئيس التنفيذي لشركة أورباكون القابضة رامز الخياط: - ما يتم إبرامه اليوم هو بداية مشروع وطني باستثمار أجنبي قيمته 7 مليار دولار - المشروع سيؤمن الاكتفاء الذاتي من الطاقة الكهربائية في سوريا - المشروع يتضمن تطوير 4 محطات إنتاجية باستخدام أحدث ما توصلت إليه التكنولوجية الأميركية	19281044
ssy	Tweet	أعجبني برتوكول توقيع الاتفاقيات	19281013
ABO_OBIDA	Tweet	الرئيس التنفيذي لشركة أورباكون القابضة رامز الخياط: 📌الاتفاقيات التي تم توقيعها اليوم ستحول سوريا من دولة لديها عجز في مجال الطاقة إلى دولة مصدرة لها	19280948

ويوضح الشكل أدناه توزع التغريدات طوال شهر أيار، حيث يلاحظ استقرار في عدد التغريدات، باستثناء عدد قليل من الأيام التي يحدث فيها أمور مهمة (كيوم رفع العقوبات، أو أثناء الكوارث أو النزاعات المسلحة).



الشكل 11: توزع التغريدات خلال فترة السحب.

ويوضح الشكل أدناه تواتر التغريدات خلال ساعات اليوم، حيث نلاحظ أن معظم التغريدات تكون في فترة المساء $(0 - 3)$، بينما ينخفض تواترها صباحاً، ويكون مسقراً ظهراً، ويفسر سبب ارتفاعها ليلاً لوجود عدة أحداث مهمة وقعت ليلاً (مثل رفع العقوبات الأميركية)، كما أن أغلب المواطنين يبدون رأيهم عما حدث خلال اليوم ليلاً. ويساعد هذا المخطط على تحديد آلية وتوقيت سحب البيانات عند وضع النظام بالخدمة.

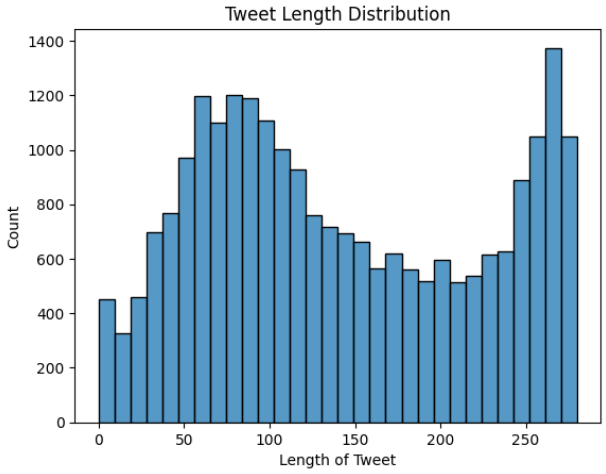


الشكل 12: تواتر التغريجات خلال ساعات اليوم.

ويوضح الشكل أدناه مجموعة الكلمات الأكثر تداولاً خلال فترة من أحد الأيام (8 أيار- الفترة 9AM – 12 AM).

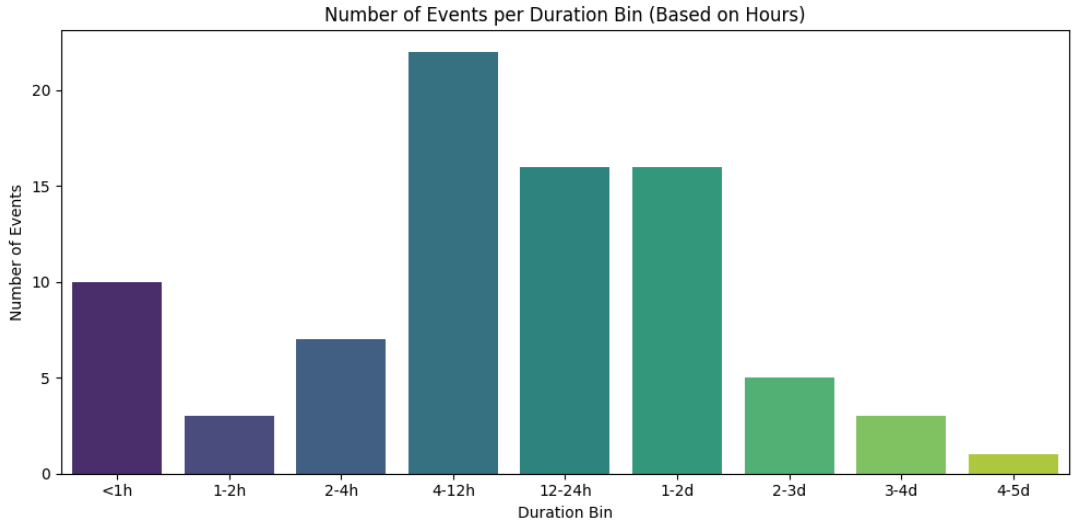
الشكل 13: مجموعة الكلمات المتداولة خلال 8 أيار- الفترة 9AM – 12 AM.

ويوضح الشكل أدناه، توزع طول التغريدات (بالمحرف).



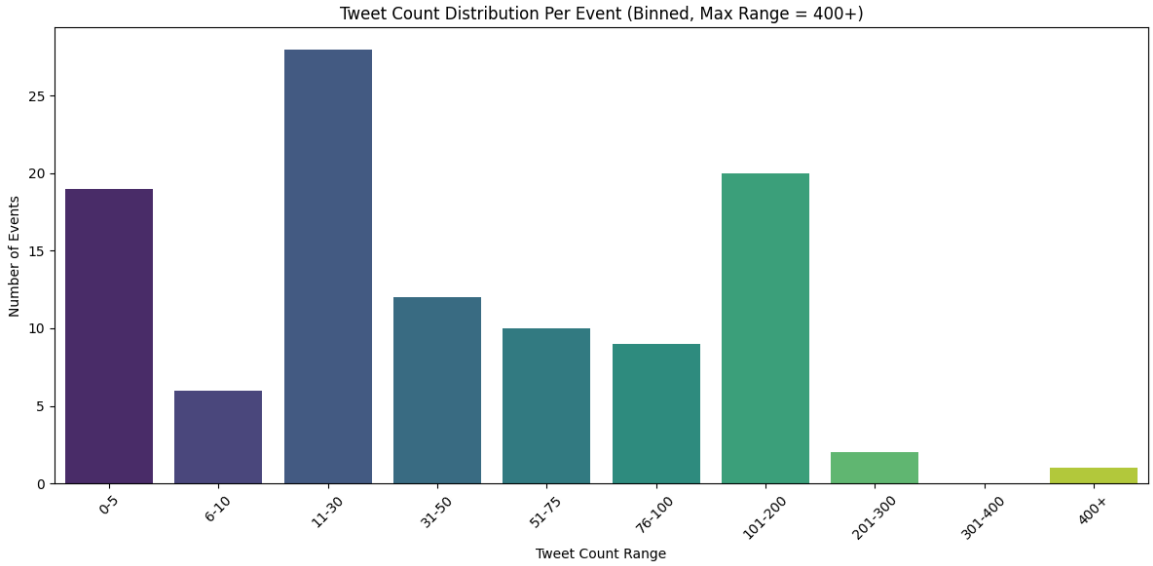
الشكل 14: توزع طول التغريدات.

ويبين الشكل أدناه المدة الزمنية للأحداث، حيث كان هنالك ثلاثة أحداث استمرت لمدة أكثر من أربعة أيام، بينما كانت معظم الأحداث لمد يوم واحد فقط، كما كان هنالك عشرة أحداث تم تداولها لمدة أقل من ساعة.



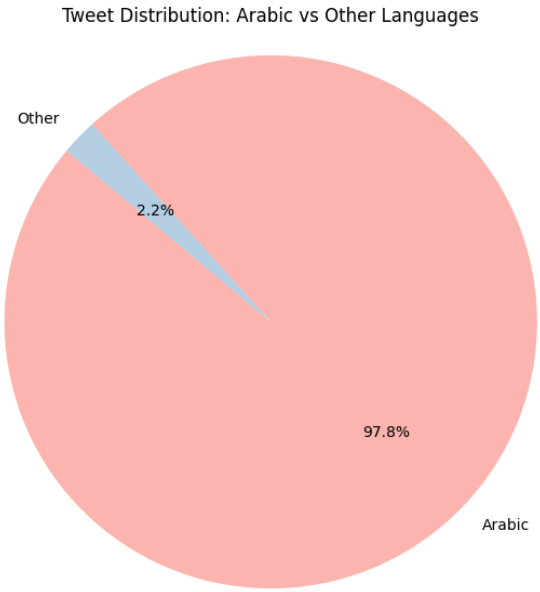
الشكل 15: المدة الزمنية للأحداث.

ويوضح الشكل أدناه عدد التغريدات ضمن الحدث، حيث كانت معظم الأحداث بين ثلاثين ومئة تغريدة، بينما كان هنالك 18 حدثاً بأقل من خمس تغريدات، وكان هنالك ثلاث أحداث بأكثر من 400 تغريدة.



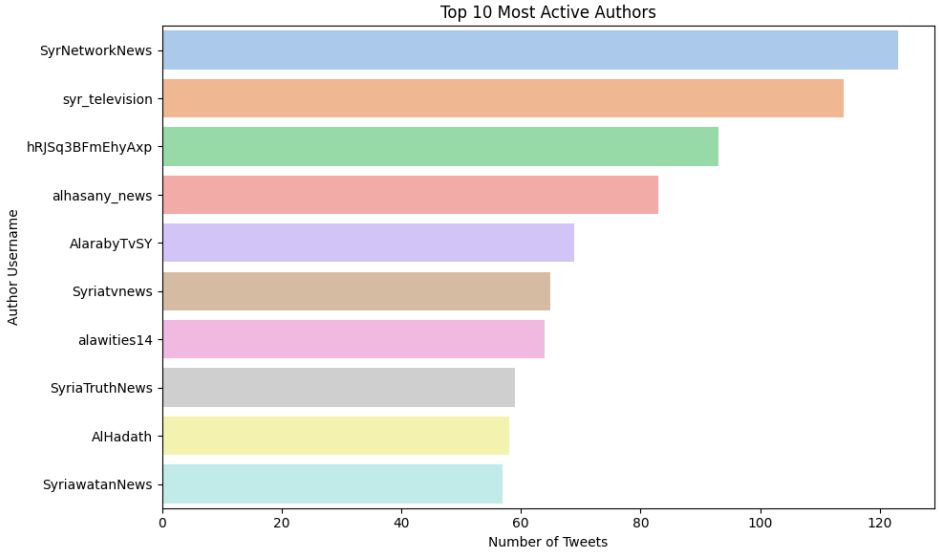
الشكل 16: مجال عدد التغريدات ضمن الحدث.

ويبين الشكل أدناه توزع لغات التغريدات حيث نلاحظ أن 97.8% بالمئة منها باللغة العربية، بينما 2.2% من بقية اللغات.



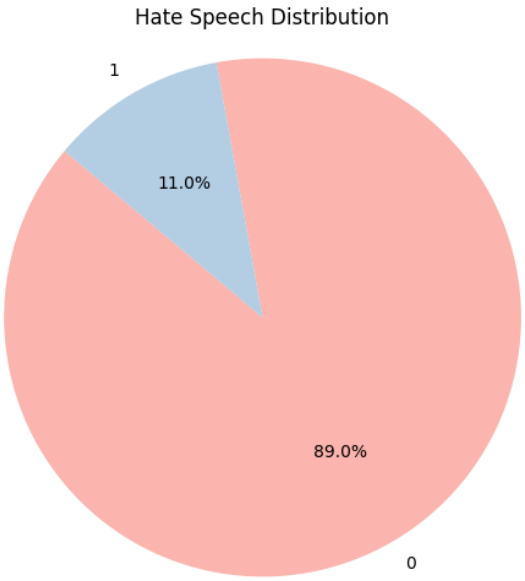
الشكل 17: توزع لغات التغريدات.

ويبين الشكل أدناه المصادر الأكثر نشراً للتغريدات ضمن المجموعة، وتيفيد معرفتها بتحديد الجهات والمؤسسات التي تنقل الأحداث السورية.



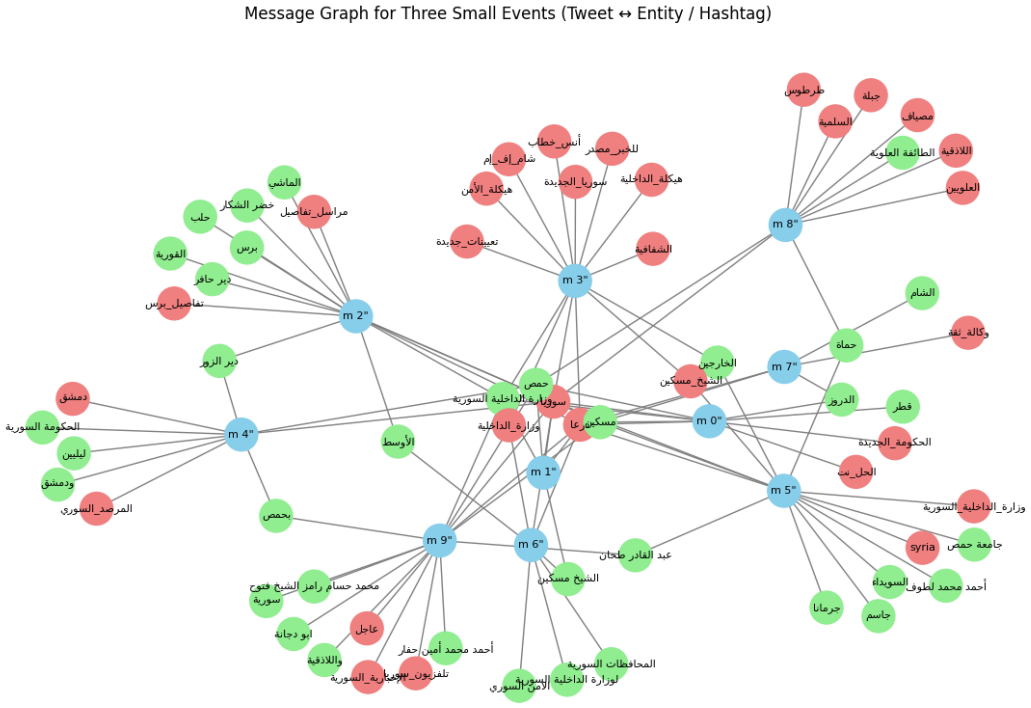
الشكل 18: مصادر البيانات الأكثر نشراً للأحداث المتداولة.

ويبين الشكل أدنا تحليل خطاب الكراهية في التغريدات، ويلاحظ أن 11% بالمئة من التغريدات تحض على العنف والكراهية.



الشكل 19: تحليل خطاب الكراهية والعنف.

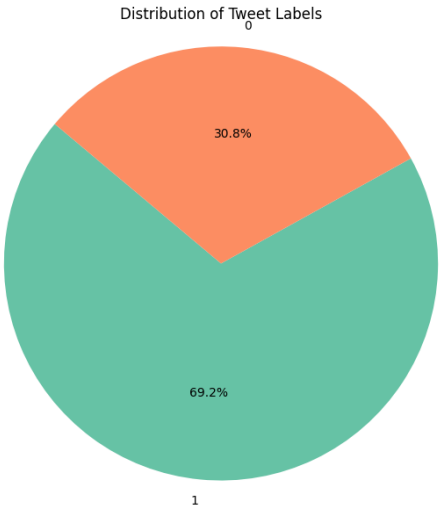
ويبين الشكل أدناه بيان المعلومات، لرسائل تتبع لأحداث مختلفة، ويلاحظ أن المستخدمين يلجؤون لإضافة hashtags ليست ذات صلة بالحدث ذات، مما يقل أهمية الـ hashtag بل وحتى الكيانات المسماة.



الشكل 20: بيان المعلومات للرسائل تتبع لأحداث مختلفة.

2.1-تجارب تصنيف التغريدات

في هذا الجزء من المنهجية، الهدف هو تصنيف التغريدات القادمة إلى ذات صلة وغير ذات صلة. ومن أجل اختيار النموذج الأفضل، تم استخدام مجموعة البيانات المنمطة التي جمعناها (SAS(Labeled. حيث تحوي مجموعة البيانات هذه على 4500 تغريدة بعد إزالة التكرارات والتغريدات باللغات غير العربية، وتكون نسبة التغريدات ذات الصلة هي 69.2% ذات صلة، و30.8% غير ذات صلة كما هو مبين بالشكل أدناه.



في المرحلة الأولى من التجارب التي قمنا بها، تم اقتراح أربع طرائق لإضافة معلومات للرسائل قبل تضمينها ولكل منها تم استخدام أربع نماذج للتمثيل Representation.

الطرائق المقترحة، هي

- (الطريقة المرجعية Base Line) استخدام نص التغريدة وحده من أجل التمثيل.
- استخدام الكيانات المسماة لوحدها (كون الرسائل ذات الصلة تتحدث عن أحداث فمن المتوقع أن تحمل الرسائل ذات الصلة كيانات مسماة أكثر من غير ذات الصلة) لتمثيل الرسائل.
- تثقيل الكيانات المسماة بأوزان لإضافة أهمية لوجود الكيان المسمى ضمن نص التغريدة.
- استبدال الكيانات المسماة بكلمات بدل Place Holders (مثل استبدال المكان المذكور بكلمة مكان أي حادث سير في دمشق => حادث سير في مكان، صرح محمد محمد الرئيس التنفيذي لشركة الشام القابضة بأنّ => صرح شخص الرئيس التنفيذي لمؤسسة بأنّ) تفيد هذه الطريقة بتجنب أن يتعلم النموذج معلومات تتعلق بمكان معين أو شخص معين وأن تكون قدرته على التعميم أفضل.

ومن أجل نماذج التمثيل تم استخدام FastText, mBert, AraBert, TF-IDF، ولكل نموذج تمثيل تم تدريب النماذج التالية SVM, KNN, Neural Network, Logistic Regression, Naïve Bayes، وتم ضبط الموسطات من خلال تحليل أثرها على حدى ومن ثم اختيار المجا المناسب لها، وبعد ذلك إجراء grid search مع k fold cross validation مع k=5.

تم التركيز على معاملين هما F1 macro وAccuracy، لأننا نريد النموذج دقيقاً ولكن ليس على حساب استبعاد تغريدات ذات صلة، لذلك ركزنا على f1 وaccuracy.

وكانت النتائج وفق الجدول المبين أدناه.

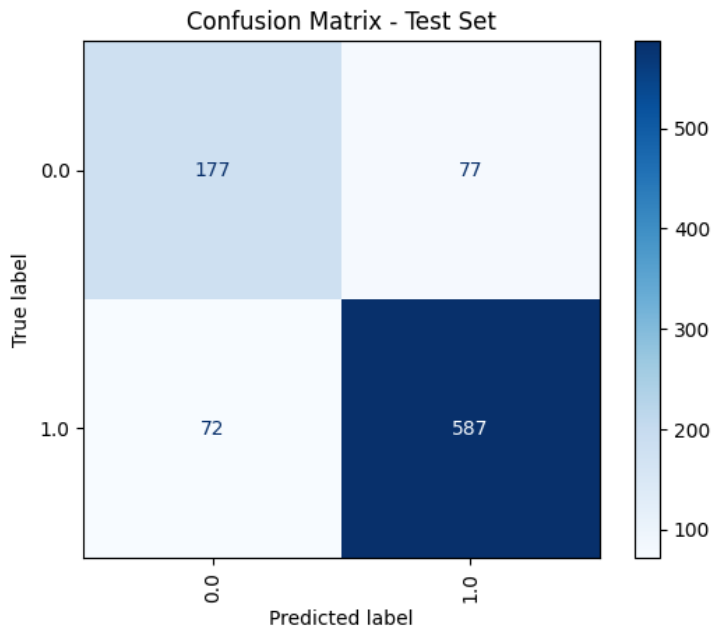
#	Name	Description	Model	Results							
				Accuracy		Recall		Precision		F1 Score	
				Train	Test	Train	Test	Train	Test	Train	Test
Exp. 1	FastText	Embedding Tweets with Fastext Model	Neural Network	0.84	0.82	0.90	0.89	0.86	0.86	0.81	0.77
			KNN	0.86	0.81	0.93	0.89	0.88	0.87	0.81	0.75
			SVM	0.89	0.84	0.95	0.90	0.91	0.89	0.86	0.79
			Naïve Bayes	0.73	0.76	0.85	0.88	0.78	0.81	0.67	0.68
			Logistic Regression	0.84	0.81	0.91	0.88	0.87	0.86	0.81	0.76
	AraBert	Embedding Tweets with AraBert Model	Neural Network	0.86	0.82	0.88	0.85	0.92	0.90	0.85	0.79
			KNN	0.85	0.81	086	0.82	0.91	0.91	0.83	0.78
			SVM	0.99	0.81	1.00	0.88	0.99	0.87	0.99	0.77
			Naïve Bayes	0.76	0.77	0.79	0.80	0.85	0.88	0.73	0.74
			Logistic Regression	0.87	0.83	0.92	0.89	0.90	0.89	0.85	0.79
	mBert	Embedding Tweets with mBert Model	Neural Network	0.93	0.80	0.93	0.82	0.97	0.89	0.92	0.76
			KNN	0.86	**0.84**	0.91	0.89	0.88	0.88	0.83	**0.80**
			SVM	0.85	0.81	0.90	0.88	0.89	0.87	0.83	0.77
			Naïve Bayes	0.74	0.71	0.78	0.70	0.83	0.87	0.71	0.68
			Logistic Regression	0.84	0.81	0.90	0.88	0.87	0.86	0.81	0.76
	TF-IDF	Embedding Tweets with TF-IDF	Neural Network	0.99	0.78	0.99	0.82	0.99	0.87	0.98	0.78
			KNN	0.84	0.81	0.88	0.83	0.89	0.90	0.82	0.78
			SVM	0.87	0.80	0.93	0.89	0.88	0.84	0.84	0.74
			Naïve Bayes	0.69	0.74	0.67	0.80	0.85	0.83	0.67	0.69
			Logistic Regression	0.89	0.79	0.93	0.87	0.90	0.84	0.87	0.73

حيث كان النموذج الأفضل للتمثيل هو mBert مع نموذج knn للتعلم، حيث كانت نتائجه لـ F1 macro هي 0.80، وصحة accuracy هي 84%، يليه نموذج FastText بـ F1 macro 0.79، وصحة 84%، ويوح الشكل أدماه نتائج التصنيف له.

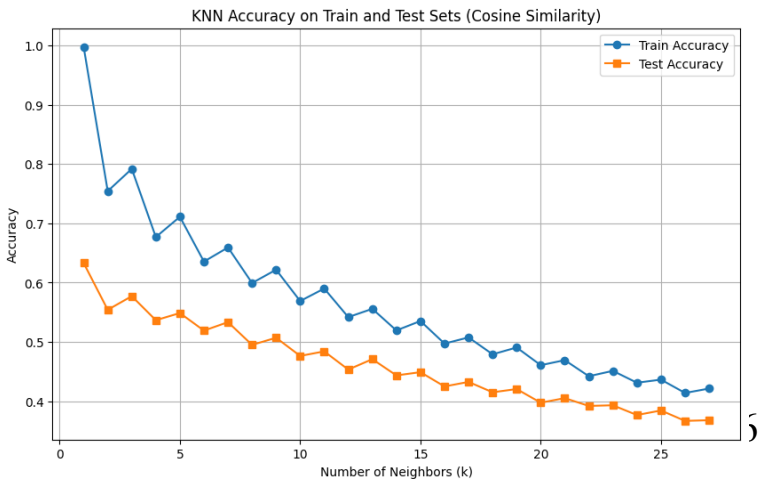
وكان الفرق بين نتائج التدريب والاختبار 3%.

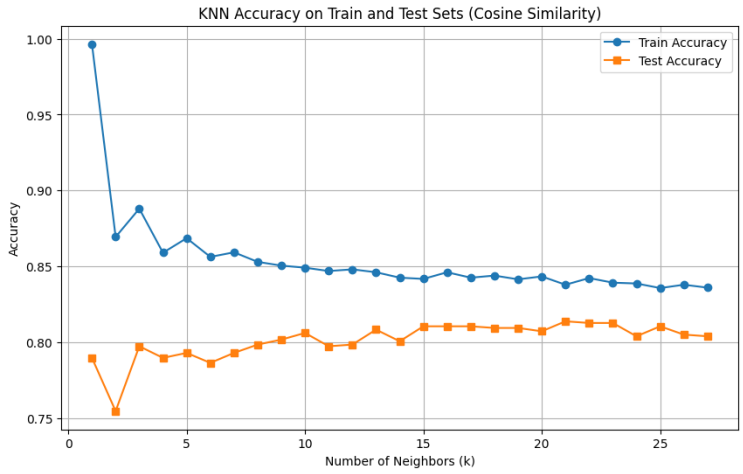
```
Classification Report - Test Set
              precision    recall  f1-score   support

         0.0       0.71      0.70      0.70       254
         1.0       0.88      0.89      0.89       659

    accuracy                           0.84       913
   macro avg       0.80      0.79      0.80       913
weighted avg       0.84      0.84      0.84       913
```

ويوضح الشكل أدناه مصفوفة الارتباك له.



ويوضح الشكل أدناه أثر اختيار تابع المسافة هو cosine، عوضاً عن تابع المسافة الإقليدية (الشكل على اليسار)، حيث أنه حسن نتائج.





التجربة الثانية: تثقيل شعاع الكيانات المسماة

في هذه التجربة أخذنا تثقيلاً لتمثيل الكيانات المسماة (شعاع التمثيل = شعاع التمثيل القديم + w * شعاع تمثيل الكيانات المسماة)، لم تضف هذه التجربة قيمة مضافة للنتائج، ولكنها خففت من الفرق بين نتائج التدريب والاختبار.

وكان النتائج وفق الجداول التالية.

#	Name	Description	Model	results							
				Accuracy		Recall		Precision		F1 Score	
				Train	Test	Train	Test	Train	Test	Train	Test
Exp. 2	FastText	Embedding Tweets with Fastext + NE Weight	Neural Network	0.83	0.80	0.92	0.90	0.84	0.83	0.79	0.74
			KNN	0.85	0.80	0.92	0.87	0.87	0.85	0.82	0.75
			SVM	0.85	0.81	0.91	0.87	0.88	0.87	0.83	0.76
			Naïve Bayes	0.73	0.76	0.86	0.88	0.77	0.81	0.66	0.68
			Logistic Regression	0.84	0.81	0.91	0.89	0.87	0.86	0.81	0.76
	AraBert	Embedding Tweets with AraBert + NE Weight	Neural Network	0.84	0.81	0.96	0.93	0.83	0.83	0.79	0.74
			KNN	0.85	**0.83**	0.90	0.88	0.88	0.88	0.82	**0.79**
			SVM	0.88	0.83	0.93	0.88	0.90	0.89	0.86	0.79
			Naïve Bayes	0.78	0.79	0.78	0.78	0.89	0.91	0.77	0.76
			Logistic Regression	0.86	0.83	0.91	0.88	0.89	0.89	0.84	0.79
	mBert	Embedding Tweets with mBert + NE Weight	Neural Network								
			KNN	0.83	0.79	0.86	0.82	0.89	0.88	0.81	0.75
			SVM	0.84	0.80	0.90	0.86	0.87	0.86	0.81	0.75
			Naïve Bayes	0.67	0.65	0.66	0.64	0.83	0.84	0.65	0.63
			Logistic Regression	0.85	0.80	0.90	0.87	0.88	0.86	0.82	0.75
	TF-IDF	Embedding Tweets with TF-IDF + NE Weight	Neural Network	0.85	0.75	0.91	0.83	0.88	0.82	0.82	0.68
			KNN	0.83	0.75	0.88	0.83	0.87	0.82	0.80	0.75
			SVM	0.75	0.74	1.00	0.98	0.73	0.74	0.60	0.53
			Naïve Bayes	0.65	0.55	0.49	0.44	1.00	0.88	0.65	0.55
			Logistic Regression	0.85	0.77	0.92	0.85	0.86	0.83	0.81	0.70

كان أفضل نموذج في هذه التجربة هو Arabert + NE، مع التصنيف باستخدام KNN وكانت النتائج لـ 0.79 f1 macro و صحة 83%.

التجربة الثالثة: تمثيل الكيانات المسماة فقط

في هذه التجربة قمنا بأخذ التمثيل فقط للكيانات المسماة (شعاع التمثيل = شعاع تمثيل الكيانات المسماة)، لم تحسن هذه التجربة النتائج، ولكنها أعط نتائج قريبة مما حصلنا عليه (6%-)، مما يشير إلى أن الكيانات المسماة لوحدها تحمل معلومات مفيدة تساهم في التصنيف.

#	Name	Description	Model	Results							
				Accuracy		Recall		Precision		F1 Score	
				Train	Test	Train	Test	Train	Test	Train	Test
Exp. 3	FastText	Embedding Tweets with Fastext Model	Neural Network								
			KNN								
			SVM								
			Naïve Bayes								
			Logistic Regression								
	AraBert	Embedding Tweets with AraBert Model	Neural Network	0.73	0.75	0.95	0.94	0.73	0.77	0.60	0.60
			KNN	0.76	0.74	0.90	0.89	0.78	0.78	0.69	0.63
			SVM	0.82	0.74	0.95	0.88	0.82	0.78	0.77	0.64
			Naïve Bayes	0.64	0.65	0.67	0.69	0.78	0.80	0.61	0.61
			Logistic Regression	0.78	0.73	0.94	0.88	0.78	0.78	0.71	0.63
	mBert	Embedding Tweets with mBert Model	Neural Network	0.73	0.75	0.95	0.94	0.73	0.77	0.60	0.60
			KNN	0.75	0.73	0.90	0.88	0.77	0.78	0.68	0.62
			SVM	0.82	0.74	0.95	0.88	0.82	0.78	0.77	0.64
			Naïve Bayes	0.64	0.65	0.67	0.69	0.78	0.80	0.61	0.61
			Logistic Regression	0.78	0.73	0.94	0.88	0.78	0.78	0.78	0.73
	TF-IDF	Embedding Tweets with TF-IDF	Neural Network	0.78	0.73	0.96	0.90	0.77	0.77	0.78	0.73
			KNN	0.83	0.74	0.92	0.87	0.85	0.79	0.80	**0.74**
			SVM	0.78	**0.76**	0.99	0.97	0.76	0.76	0.68	0.58
			Naïve Bayes	0.67	0.62	0.56	0.53	0.93	0.91	0.67	0.61
			Logistic Regression	0.76	0.73	0.92	0.88	0.77	0.77	0.68	0.61

وكان أفضل نموذج حصلنا عليه هنا هو التمثيل باستخدام TF-IDF مع تعلم التصيف باستخدام KNN، وكانت نتائجه هي 0.74 لـ F1 macro والصحة 74%. بفارق 6%- عن أفضل نموذج حصلنا عليه.

التجربة الرابعة: استبدال الكيانات المسماة بكلمات بدل Placeholders

في هذه التجربة قمنا باستبدال الكيانات المسماة بكلمات بدل، لأننا كنا نعتقد أن النموذج يتعلم معلومات مكانية غير دقيقة لا يجب تعلمها (أي أن النموذج رأى مثلاً العبارة "وقع حادث في دمشق" بينما لم يرى العبارة "وقع حادث في حمص" لذلك كنا نعتقد أن النموذج يأخذ في الاعتبار كلمة دمشق أثناء التصنيف مما لايجعله قادراً على التعميم لذلك قمنا باستبدالها بالعبارة "وقع حادث في مكان")، وبعد ذلك أجرينا التجارب التي يوضحها الجدول أدناه، وحصلنا على نتائج مطابقة لما حصلنا عليه في التجربة الأولى، مما يعني أن النموذج لم يكن يتعلم معلومات مكانية غير دقيقة.

#	Name	Description	Model	results							
				Accuracy		Recall		Precision		F1 Score	
				Train	Test	Train	Test	Train	Test	Train	Test
Exp. 4	FastText	Embedding Tweets with Fastext Model	Neural Network	0.83	0.82	0.89	0.87	0.87	0.87	0.80	0.78
			KNN	0.83	0.79	0.90	0.88	0.85	0.84	0.79	0.72
			SVM	0.84	0.82	0.91	0.88	0.87	0.87	0.81	0.77
			Naïve Bayes	0.74	0.75	0.85	0.85	0.78	0.81	0.68	0.68
			Logistic Regression	0.83	0.83	0.90	0.89	0.85	0.87	0.80	0.78
	AraBert	Embedding Tweets with AraBert Model	Neural Network	0.87	0.82	0.91	0.88	0.90	0.88	0.85	0.78
			KNN	0.85	0.81	0.89	0.86	0.89	0.88	0.83	0.77
			SVM	0.89	0.82	0.92	0.88	0.91	0.87	0.87	0.78
			Naïve Bayes	0.76	0.77	0.79	0.80	0.85	0.88	0.73	0.74
			Logistic Regression	0.87	0.83	0.92	0.89	0.90	0.88	0.85	0.79
	mBert	Embedding Tweets with mBert Model	Neural Network	0.87	0.82	0.91	0.88	0.90	0.88	0.85	0.78
			KNN	0.85	0.81	0.89	0.86	0.89	0.88	0.83	0.77
			SVM	0.89	0.82	0.92	0.88	0.91	0.87	0.87	0.78
			Naïve Bayes	0.76	0.77	0.79	0.80	0.85	0.88	0.73	0.74
			Logistic Regression	0.87	0.83	0.92	0.89	0.90	0.88	0.85	0.79
	TF-IDF	Embedding Tweets with TF-IDF	Neural Network	0.96	0.81	0.98	0.88	0.97	0.86	0.96	0.76
			KNN	0.85	**0.83**	0.89	0.86	0.90	0.90	0.83	**0.80**
			SVM	0.84	0.82	0.89	0.88	0.88	0.87	0.81	0.77
			Naïve Bayes	0.71	0.70	0.77	0.69	0.81	0.86	0.68	0.67
			Logistic Regression	0.84	0.83	0.90	0.89	0.87	0.87	0.81	0.78

حيث كان أفضل نموذج حصلنا عليه في هذه التجربة هو التمثيل باستخدام TF-IDF والتصنيف باستخدام KNN، حيث كانت نتائجه هي F1 macro 0.80 وصحة 83%.

2.1-تجارب كشف الأحداث (نمذجة المواضيع)

كما سبق وذكرنا أننا في هذا الجزء من العمل، تتلخص مهمتنا بإيجاد طريقة تمثيل وعنقدة مناسبتين للبيانات. في هذه الجزئية يلجأ الباحثون لتقييم نموذجهم بطريقتين الأولى بالوضع offline والثانية بالوضع online. ويعرف التقييم بوضع عدم الاتصال Offline Evaluation، بأنّه التقييم الذي يجرى على بيانات تاريخية، أو مجموعة مسبقاً دون الحاجة إلى تفاعل حي للمستخدم أو أخذ الزمن بعين الاعتبار (نتعامل مع كامل مجموعة البيانات دفعة واحدة) (Abagissa et al., 2024)، بينما يعرف التقييم الحي Online Evaluation، بأنّه التقييم الذي يجرى بشكل حيّ، أي بأخذ زمن وصول الرسائل بالاعتبار، حيث تصل الرسائل كتلة تلو كتلة Block By Block على التتالي M1, M2, ...

1.1.1 تقييم النماذج بنمط عدم الاتصال Offline Evaluation

هنا بدأنا بتقسيم مجموعة البيانات إلى جزئين هما مجموعة الضبط Tuning Set التي استخدمناها لضبط المعاملات الفوقية Hyper Parameters للنموذج المقترح، ومجموعة التقييم التي استخدمناها لاختبار النموذج ومقارنة النتائج. لم نأخذ مجموعة تدريب لأن العملية التي قمنا بها هي Unsupervised أي النماذج المقترحة والمستخدمة لا تحتاج إلى تدريب.

وقمنا بأخذ الرسائل من أول أسبوع (من1 أيار ولغاية 7 أيار) لمجموعة الضبط وعددها 939 رسالة تشكل 13% من مجموعة البيانات، وأخذ بقية الرسائل للتقييم وعددها 6064.

أخذنا معايير التقييم هي NMI المعلومات المتبادلة المعيرة (قيمه بين 0 و1 والأفضل هو الأعلى)، AMI المعلومات المتبادلة المضبوطة، ARI مؤشر العشوائية المضبوط، وهذه المعايير تم استخدامها في معظم الأبحاث المرجعية (Abagissa et al., 2024; Guo et al., 2024; Ray et al., 2024; Ren et al., 2022; Yang et al., 2024; Yu et al., 2024). وتم التركيز والمقارنة بين النماذج وفق المعيار NMI وهذا ما قام به الباحثون المذكورون أعلاه.

أجرينا عدداً من التجارب، حيث بدأنا باختبار النماذج المستخدمة في الأبحاث السابقة وهي LDA, Word2Vec, WMD, Bert, sBert, FastText.

من أجل Bert، استخدمه الباحثون من خلال أخذ متوسط التضمينات (أي ليس كـ Scentence Transformer) وقمنا باستخدام النسخة العربية منه AraBert.

ومن أجل sBert، أي استخدام Bert كـ Sentence Transformer قمنا باستخدام النسخة العربية منه وهي mBert.

من أجل FastText لم يستخدمه الباحثون في الأوراق السابقة، ونحن استخدمناه لأنه يعالج مشكلة Out of Vocabulary وهو يعطي تضمين أفضل للكلمة (مع وجود أخطاء إملائية) كونه يستخدم n-gramعلى مستوى المحارف.

من أجل WMD، أخذ معهFastText كنموذج تضمين.

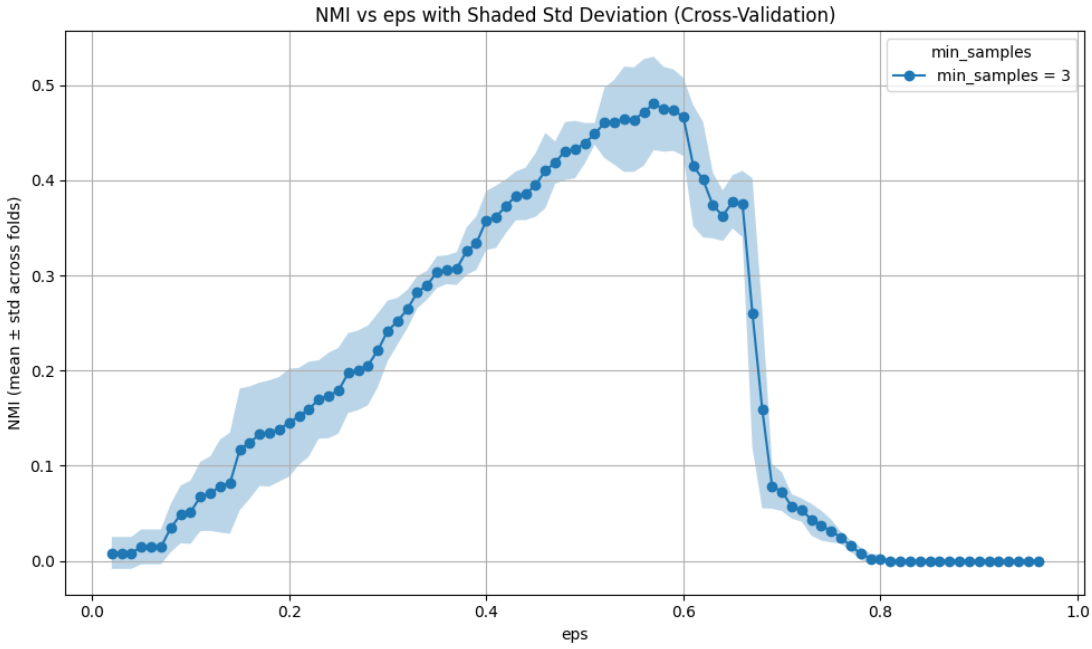
في التجربة الأولى أخذنا النماذج السابقة مع خوارزمية العنقدة DBSCAN.

بدأنا من أجل كل نموذج من النماذج السابقة بتحليل أثر كل معامل على حدى، والمعاملات لهذه النماذج هي بعد شعاع التضمين، العدد الأصغري للعناصر min Samples، معامل المسافة epsilon. وكانت أنماط النتائج متشابهة بين النماذج.

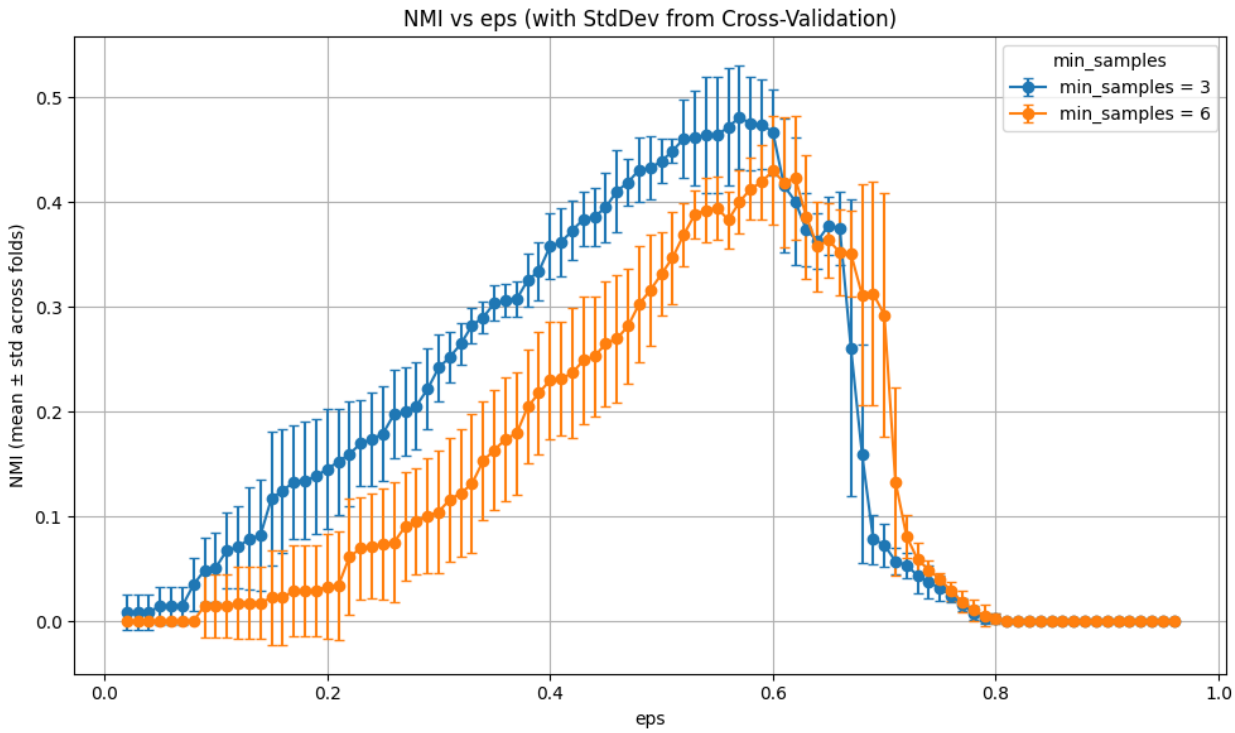
قمنا من أجل كل معامل بعمل K Cross Validation أثناء البحث عن المعامل الأفضل وأخذنا k=5.

- تحليل معامل المسافة min epsilon

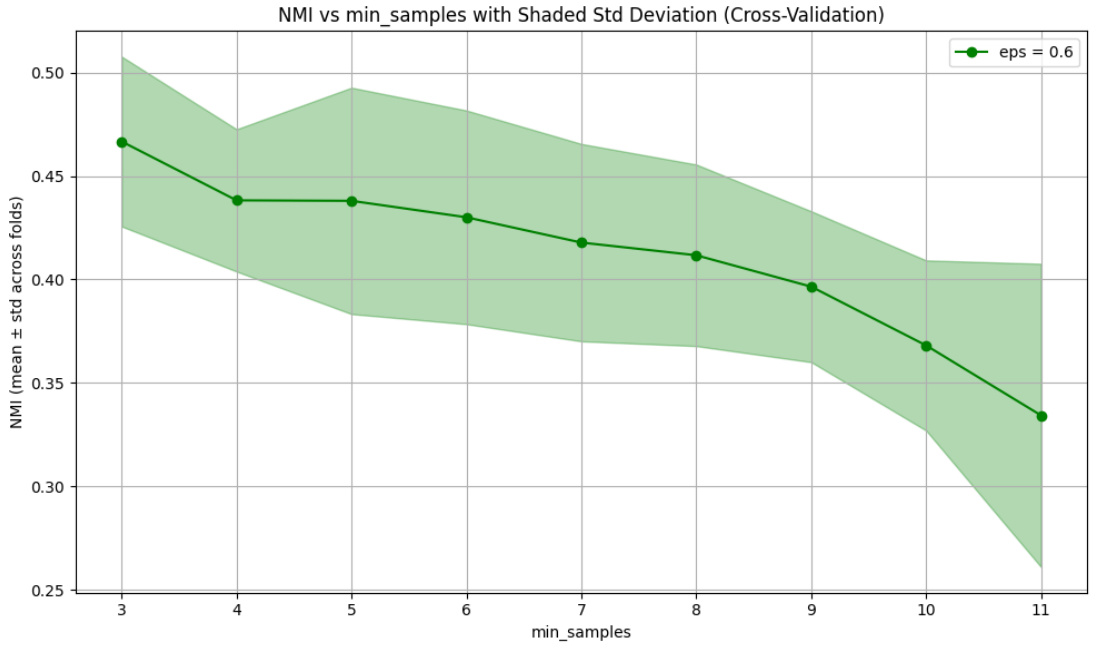
تختلف قيم هذا المعامل حسب طريقة التضمين، ولكن بمعظم التجارب كانت بين 0.1 و0.6، يوضح الشكل أدناه أثر هذا المعامل على نتائج العنقدة، حيث يبين من أجل كل قم له، متوسط نتائج الـ cross validation مع الانحراف المعياري، من أجل طريقة التمثيل mBert، (بقية الطرائق كانت النتائج مشابهة كشكل التابع مع تغير قيمة eps الأفضل)، حيث نلاحظ أنه من أجل القيم الصغير لـ eps، 0.23>، كان الانحراف المعياري كبيراً، عى عكس بقية القيم.



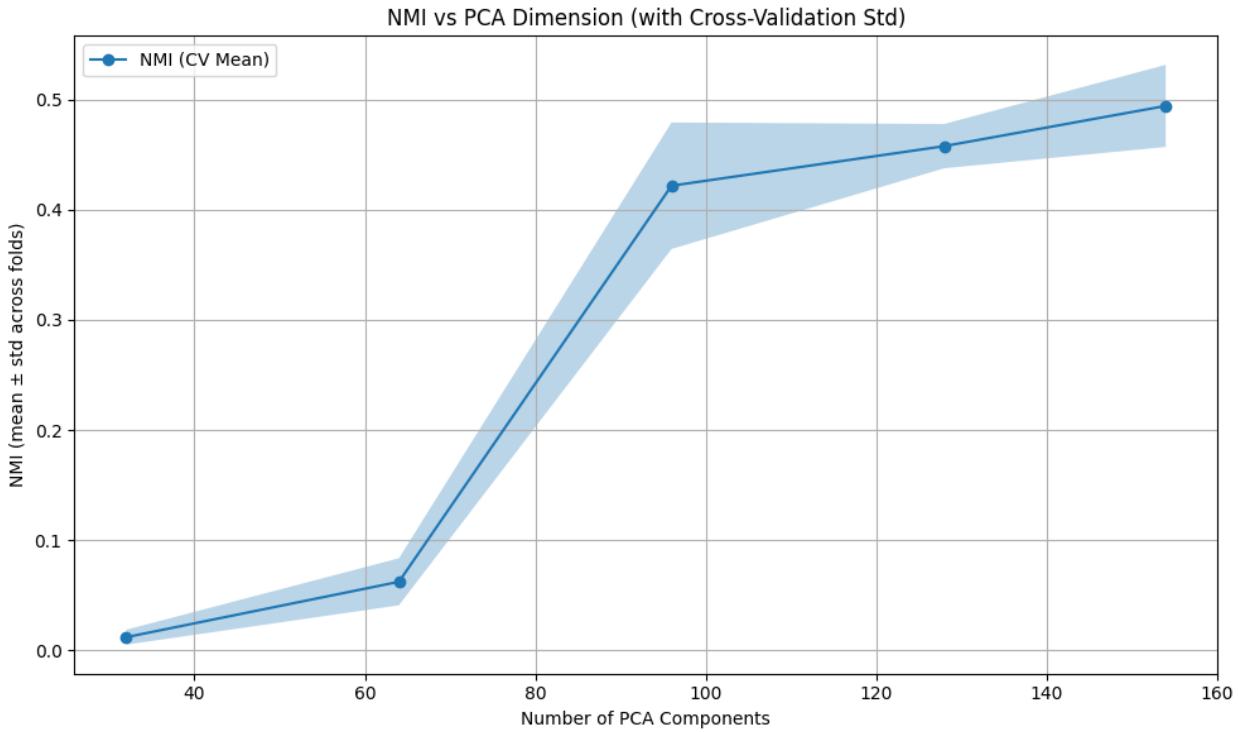
الشكل 21: أثر المعامل **eps** على النتائج باستخدام **mBert** لتمثيل الرسائل.



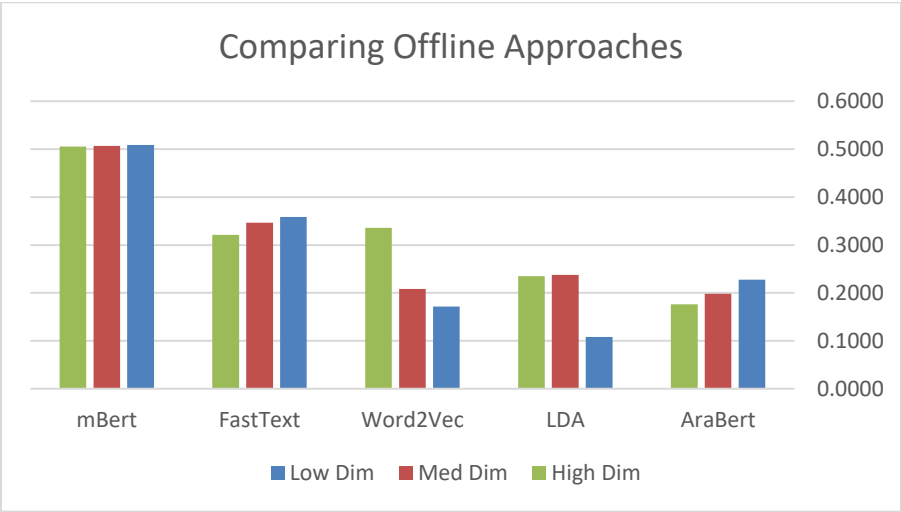
- تحليل معامل الكثافة min Samples، قيم هذا المعامل كانت في معظم الحالات بين 3 و7، القيم الكبيرة كانت تؤدي إلى نتائج غير جيدة، بسبب وجود عدد من الأحداث بعدد قليل من الرسائل، كان الإنحراف المعياري ما يقارب ∓ 0.4، ويشكل تقريباً نسبة $\mp 8\%$، مما يوحي بوجود overfitting، لذلك في التجارب القادمة سنلجأ إلى تعزيز مجموعة البيانات.



- تحليل أثر الأبعاد، يوضح الشكل أدناه أثر بعد فضاء التضمين على النتائج.



يوضح الشكل أدناه مقارنة بين النتائج التي حصلنا عليها، حيث كان أفضلها هو mBert حيث كان تحسينه عن النموذج الذي يليه هو 45%.

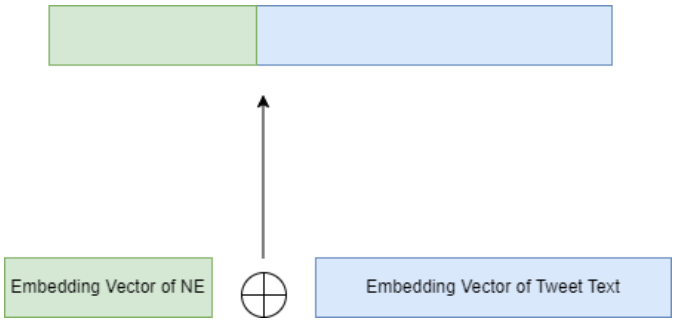


ويوضح الجدول أدناه بقية المقاييس.

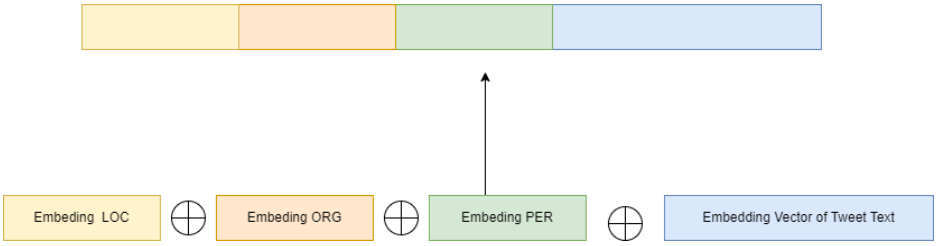
#	Approach	Description	Embedding Model	Results		
				NMI	AMI	ARI
Exp. 1	AraBert	DBSCAN + Arabert	AraBert 768	0.1765	0.0882	0.0039
			AraBert 320	0.1983	0.1030	0.0046
			AraBert 128	0.2278	0.1259	0.0059
			AraBert 32	0.2734	0.1881	0.0104
Exp. 2	mBert	DBSCAN +mBert	mBert 768	0.5056	0.4158	0.0295
			mBert 512	0.5066	0.4165	0.0297
			mBert 256	**0.5085**	**0.4182**	**0.0302**
			mBert 128	0.5060	0.4247	0.0346
			mBert 64	0.4790	0.4217	0.0518
			mBert 32	0.0578	0.0347	0.0025
Exp. 3	FastText	DBSCAN +FastText	FastText 300	0.3213	0.1972	0.0091
			FastText 160	0.3464	0.2223	0.0095
			FastText 128	0.3583	0.2354	0.0102
			FastText 64	0.3088	0.2148	0.0130
Exp. 4	Word2Vec	DBSCAN+Word2Vec	Word2Vec 300	0.3361	0.2470	0.0157
			Word2Vec 160	0.2079	0.1331	0.0101
			Word2Vec 128	0.1714	0.1059	0.0081
			Word2Vec 64	0.1018	0.0569	0.0054
Exp. 5	LDA	DBSCAN + LDA	Topics 140	0.2347	0.1442	0.0106
			Topics 130	0.2374	0.1353	0.0105
			Topics 90	0.1082	0.0506	0.0037
			Topics 50	0.0265	0.0007	0.0127
Exp. 6	WMD	DBSCAN + WMD	FastText 300	-	-	-
			FastText 160			
			FastText 128			
			FastText 64			

التجربة الثانية: إضافة شعاع الكيانات المسماة

في التجربة الثانية قمنا بإضافة شعاع تمثيل الكيانات المسماة،



التجربة الثالثة: تجزئة شعاع الكيانات المسماة



التجربة الرابعة: تعزيز مجموعة البيانات

t-SNE Projection (Colored by True Labels) — DBSCAN eps=0.4, min_samples=3

t-SNE Projection (Colored by True Labels) — DBSCAN eps=0.34, min_samples=3

1.1.1 تقييم النماذج بنمط الحي Online Evaluation

هنا قمنا كما قام (Abagissa et al., 2024; Guo et al., 2024; Ray et al., 2024; Ren et al., 2022; Yang et al., 2024; Yu et al., 2024) بتقسيم مجموعة البيانات إلى كتل $M_0, M_1, \dots M_{24}$، بحيث تحوي الكتلة الأولى M_0 الرسائل الواردة خلا الأسبوع الأوّل، بينما تحوي الكتل $M_1, \dots M_{24}$، على الرسائل الواردة يوماً بيوم، ويوضح الشكل أدناه إحصائيات هذه الكتل.

Block	Date	Size	Events
M0	Before 8 May	934	44
M1	8-May	85	10
M2	9-May	16	7
M3	10-May	81	8
M4	11-May	33	8
M5	12-May	294	10
M6	13-May	211	8
M7	14-May	287	8
M8	15-May	127	10
M9	16-May	47	9
M10	17-May	171	12
M11	18-May	164	12
M12	19-May	180	16
M13	20-May	265	19
M14	21-May	1005	20
M15	22-May	231	14
M16	23-May	417	14
M17	24-May	394	15
M18	25-May	384	19
M19	26-May	387	8
M20	27-May	209	9
M21	28-May	225	8
M22	29-May	277	14
M23	30-May	352	11
M24	31-May	165	4

جدول 8: إحصائيات كتل الرسائل.

بحيث يتم تمرير الكتل، كتلة تلو أخرى إلى النموذج المقترح، ليقوم بعنقدتها، ومن الجدير بالذكر هنا، أنه يجب تعديل خوارزمية العنقدة بحيث تقوم بإهمال الرسائل القديمة (التي يتجاوز قدمها X يوم، وأن تأخذ معامل آخر وهو حجم النافذة W (نافذة منزلقة) لمراعاة عدم أخذ كمية ضخمة من الرسائل التي تؤثر على أداء النموذج.

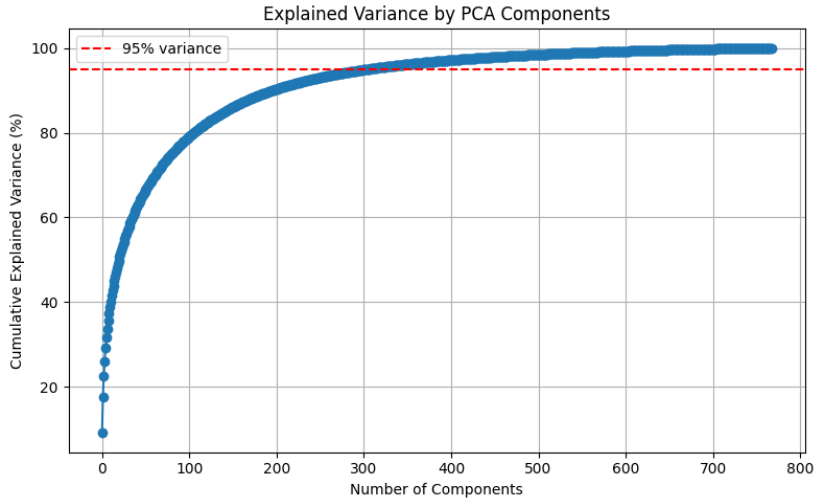
وأخذنا معايير التقييم هي NMI المعلومات المتبادلة المعيرة لمقارنة النماذج وتقييمها، كما قمنا بحساب المقاييس الأخرى AMI المعلومات المتبادلة المضبوطة، ARI مؤشر العشوائية المضبوط، وأرفقناه بالملحق، وهذه المعايير تم استخدامها في معظم الأبحاث المرجعية (المذكورون سابقاً). وأجرينا عدداً من التجارب، حيث بدأنا باختبار النماذج المستخدمة في الأبحاث السابقة وهي LDA, Word2Vec, WMD, Bert, sBert, FastText.

التجربة الأولى

في التجربة الأولى أخذنا النماذج السابقة مع خوارزمية العنقدة HDBSCAN، بعدما عدلناها لتعمل بشكل تزايدي، وأضفنا لها معاملين هما (1) معامل التذكر T الذي يحدد عمر الرسائل قبل إهمالها، (2) معامل حجم النافذة W، الذي يحدد حجم الرسائل التي يتعامل معها النموذج (بحيث يتم وضع الرسائل في Buffer وإزالة الرسائل من نهايته كلما امتلاء).

من أجل هذه التجربة اختبرنا طرائق التضمين التالية (mBert, AraBert, Word2Vec, FastText, LDA).

من أجل الـ LDA، أخذنا عدد المواضيع topics هو 60. ولاختيار عدد أبعاد بقية الطرائق، تم استخدام طريقة Explainable Variance، مع خارزمية اختزال الأبعاد PCA. ويوضح المخطط أدناه مثالاً على ذلك.



ويوضح الجدول النتائج التي حصلنا عليها.

	M_1	M_2	M_3	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
mbert	**0.440**	**0.436**	**0.460**	**0.532**	**0.480**	**0.644**	**0.463**	**0.364**	**0.373**	0.330	**0.576**	**0.544**
AraBert	0.278	0.372	0.366	0.316	0.306	0.392	0.339	0.228	0.205	0.246	0.346	0.419
Word2Vec	0.356	0.374	0.379	0.349	0.271	0.405	0.363	0.087	0.163	0.189	0.356	0.330
FastText	0.278	0.372	0.366	0.316	0.306	0.392	0.339	0.228	0.205	0.246	0.346	0.419
LDA	0.381	0.395	0.392	0.401	0.421	0.548	0.405	0.317	0.369	**0.345**	0.435	0.448
	16%	10%	17%	33%	14%	18%	14%	15%	1%	-4%	32%	21%

	M_{12}	M_{13}	M_{14}	M_{15}	M_{16}	M_{17}	M_{18}	M_{19}	M_{20}	M_{21}	M_{22}	M_{23}	M_{24}
mbert	**0.557**	**0.503**	**0.562**	**0.721**	**0.625**	**0.448**	**0.556**	**0.462**	**0.346**	**0.406**	**0.572**	**0.575**	**0.547**
AraBert	0.431	0.332	0.352	0.420	0.415	0.354	0.407	0.351	0.266	0.401	0.551	0.478	0.404
Word2Vec	0.415	0.454	0.209	0.408	0.437	0.352	0.452	0.361	0.315	0.430	0.508	0.455	0.438
FastText	0.431	0.332	0.352	0.420	0.415	0.354	0.407	0.351	0.266	0.401	0.551	0.478	0.404
LDA	0.436	0.466	0.410	0.540	0.459	0.395	0.443	0.351	0.273	0.337	0.492	0.446	0.473
	28%	8%	37%	33%	36%	13%	25%	32%	26%	21%	16%	29%	16%

جدول 9: نتذاج التجربو الأولى.

(الخط الغامق أعلى نتيجة، الرقم الذي تحته خط ثاني أعلى نتيجة)

في هذه التجربة، كانت طريقة التضمين mBert، هي الأفضل، حيث كانت أفضل من بقية الطرائق في 23 كتلة من أصل 24. وكانت أفضل من ثاني أفضل طريقة (LDA) بتحسين قدره 20%. ويعود ذلك كون mBert يعطي تضميناً للجملة بناءً على سياقها، بينما LDA هو طريقة إحصائية لنمذجة المواضيع، وبقية الطرائق تعتمد على أخذ متوسط تضمينات الكلمات.

وكانت من الأخطاء التي رأيناها من النمط

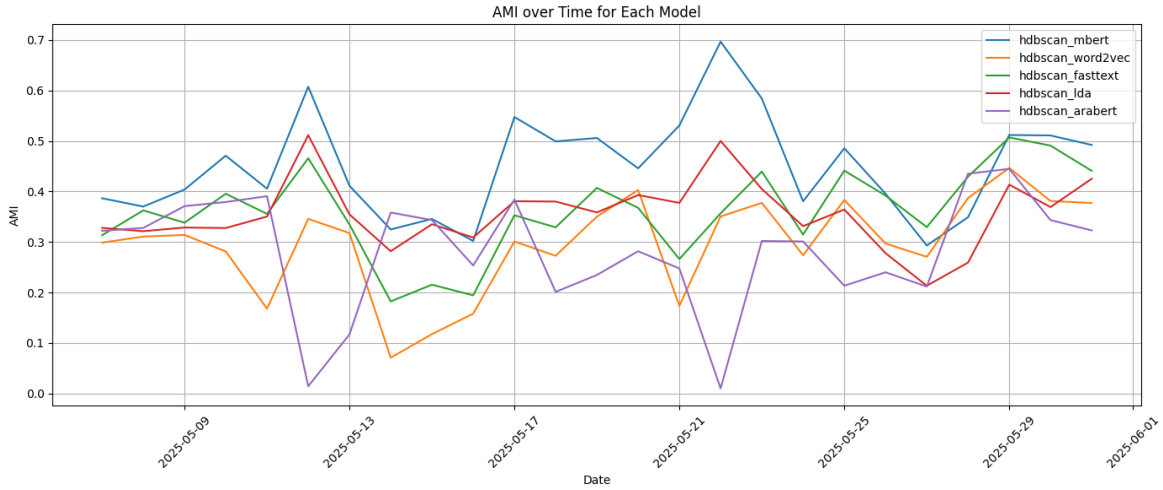
- تجميع الرسائل حسب معناها (أي مثلاً رسائل الحرائق تقع بالقرب من بعضها في فضاء التمثيل على الرغم من أن الحرائق تحدث في أماكن مختلفة، مما يؤدي إلى تداخل في الأحداث)
- كون العنقدة تحدث على كامل المجموعة كان هناك أحداث وقعت في نفس المكان ولكن بزمنين مختلفين (العثور على مستودع *** في ريف دمشق)
- كان هنالك بعض الأحداث التي تم تجميعها وفق هدف غير فعل الحدث (كتجميع تصريحات وزير الخارجية على ثلاثة أحداث مختلفة عوضاً عن أن تكون التصريحات تنتمي إلى الحدث، مثل تصريحات الوزير على رفع العقوبات الأمريكية، تصريحاته على زيارة دولة فرنسا، تصريحاته على رفع العقوبات الأوروبية)
- وجود ضجيج (رسائل غير ذات صلة كانت قليلة في الحدث الواحد، ولكن عند تجميع الأحداث كانت تدخل في بعض العناقيد)

وعليه:

الأخطاء من النمط الأول والثالث من المفترض أنه بإمكاننا تفاديها وتحسين النتائج، أذا أخذنا الكلمات المكانية، والكلمات الرئيسية بطريقة التضمين.

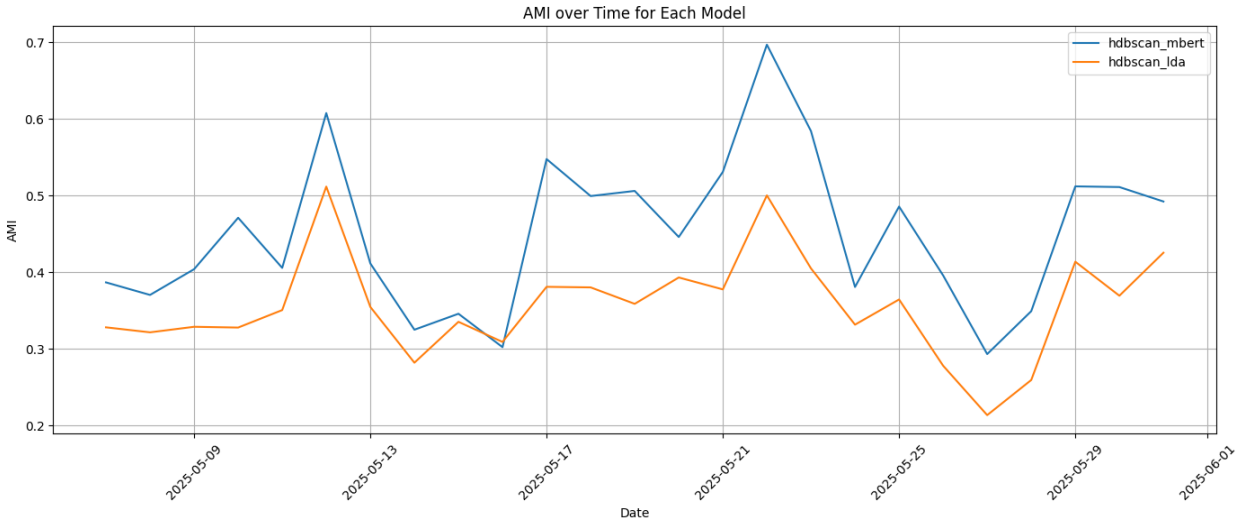
الأخطاء من النمط الثاني، يمكننا تفاديها عند أخذ الزمن بالاعتبار.

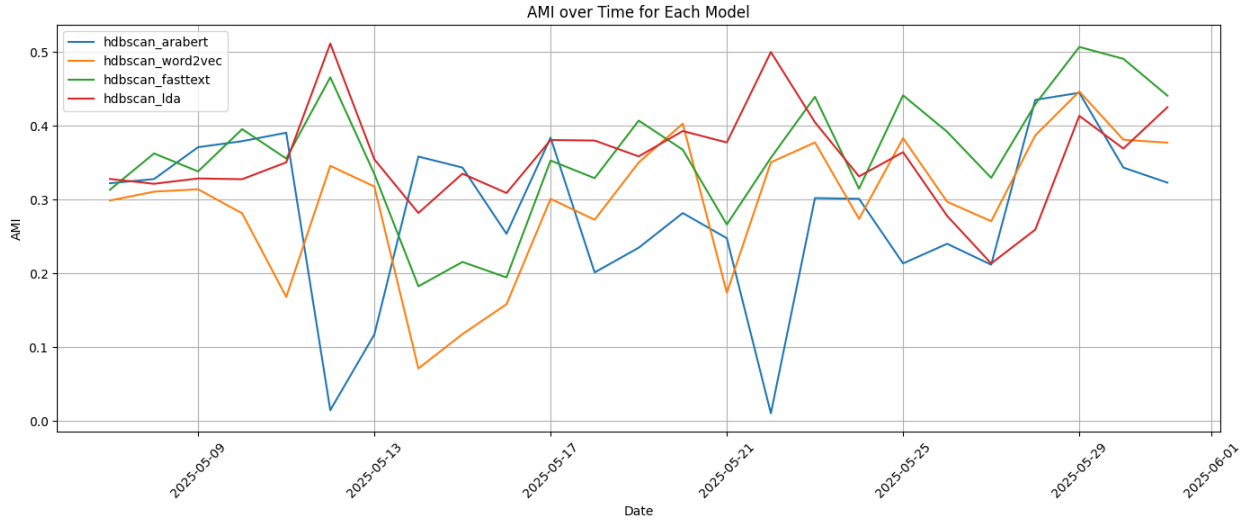
والأخطاء من النمط الرابع هي غير قابلة للتفادي في العنقدة (مما يوجب علينا إضافة مرحة معالجة لاحقة لإزالتها).

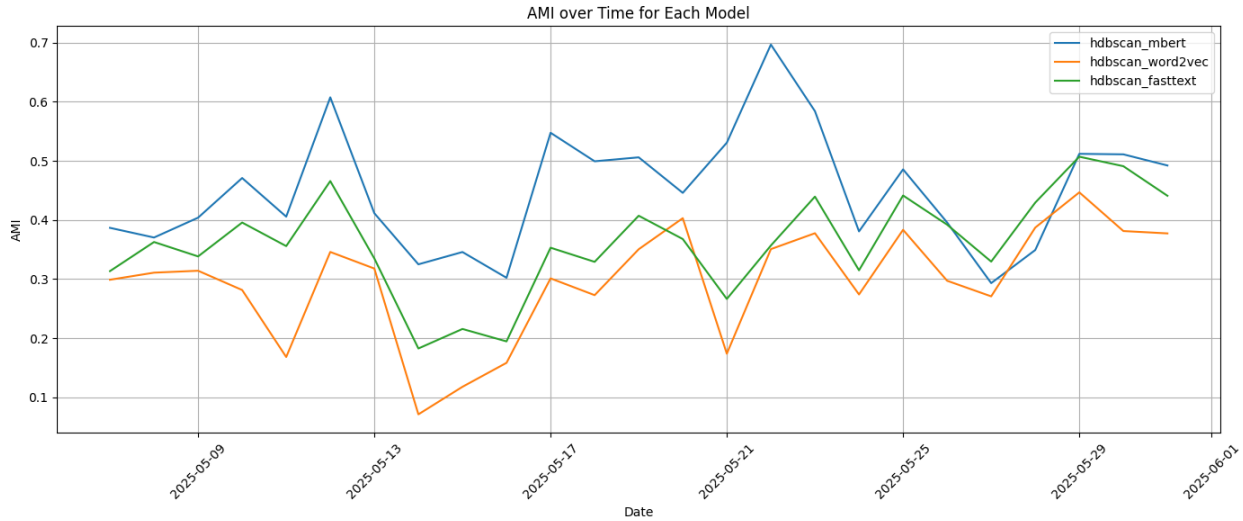


ويوضح الشكل أعلاه مخطط قيم NMI للنماذج المستخدمة، حيث وكما ذكرنا كان mBert هو الأفضل يليه LDA. وكانت نتائج بقية النماذج متقاربة.

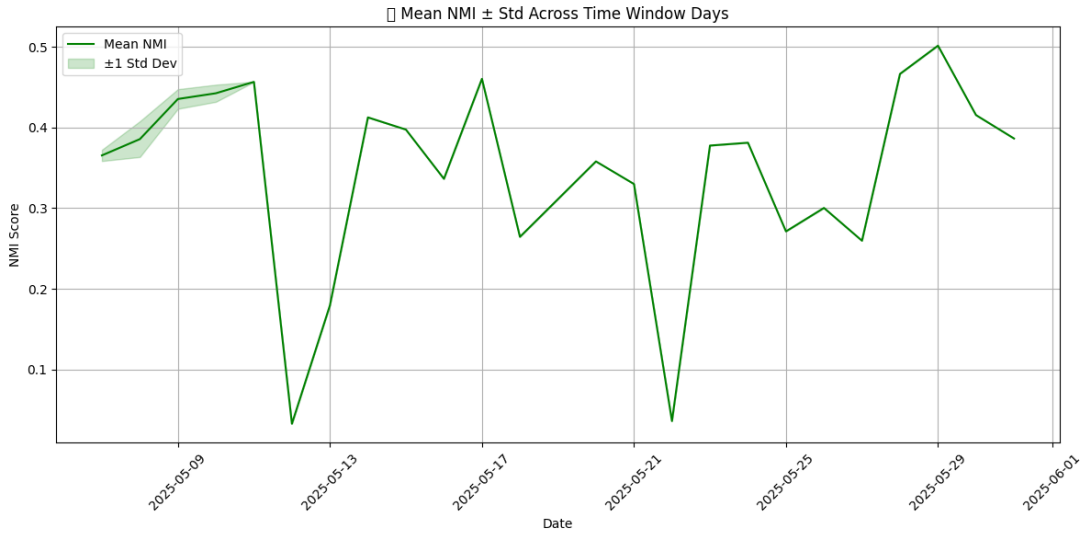
ويوضح الشكل أدناه معيار التقييم AMI، حيث كما نلاحظ أيضاً أنّ mBert كان الأفضل أيضاً بناءً عليه.





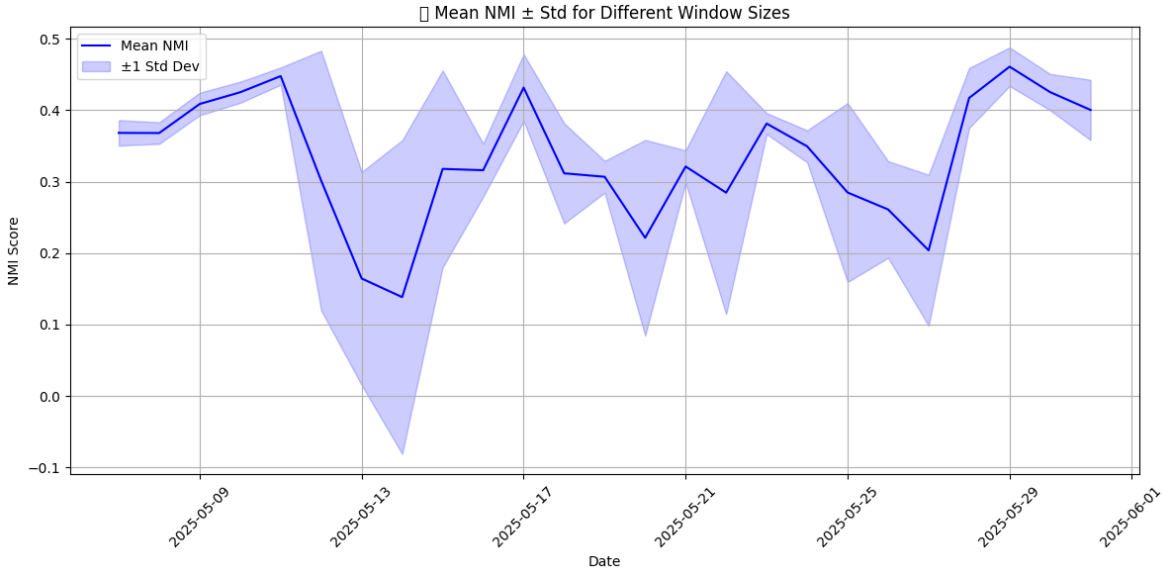


ويوضح الشكل أدناه أثر حجم النافذة الزمنية (معامل النسيان T)، على النتائج، حيث فمنا بالاختبار من أجل القيم (2 إلى خمسة أيام) وأخذنا متوسط القيم وانحرافها المعياري وكانت النتائج متقاربة جداً، حيث يوضح الشكل أدناه المتوسط والانحراف المعياري للنتائج لقيم مختلفة لـ T، حيث كانت قيم الانحراف المعياري صغيرة جداً.



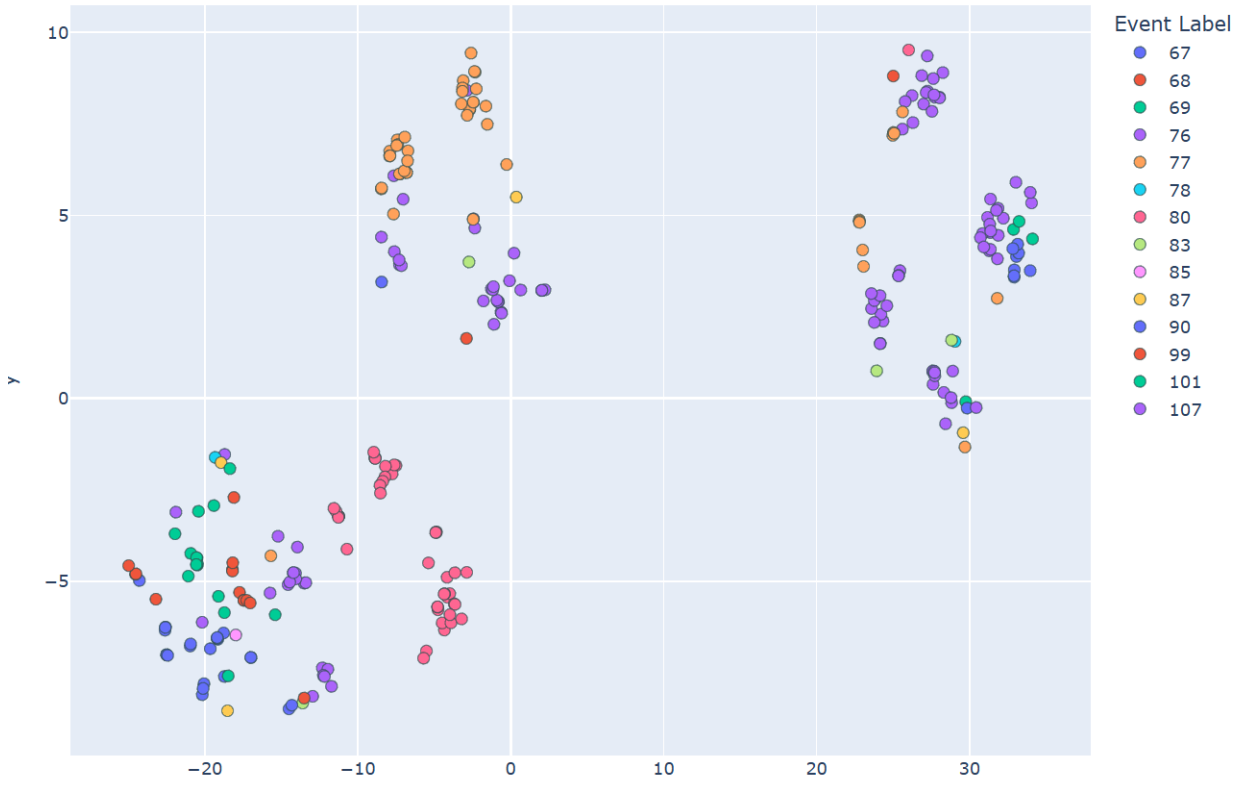
الشكل 22: أثر معامل حجم النافذة الزمنية على نتائج النموذج.

ويوضح الشكل التالي، أثر معامل حجم النافذة المنزلقة Sliding Window على النتائج، وكانت القيم الصغيرة لـ W (-400 200) تؤدي إلى نتائج متفاوتة، أي في الكتل الكبير تكون قيم nmi صغيرة جداً بسبب نسيان الأحداث الحالية بسبب كبر جدم كتلة الرسائل، ينما القيم الكبيرة لـ W (<500) كانت تؤدي إلى نتائج متقاربة فيما بينها. ويوضح الشكل أدناه المتوسط والانحراف المعياري للنتائج بعد إجراء تجارب مختلفة لقيم W.



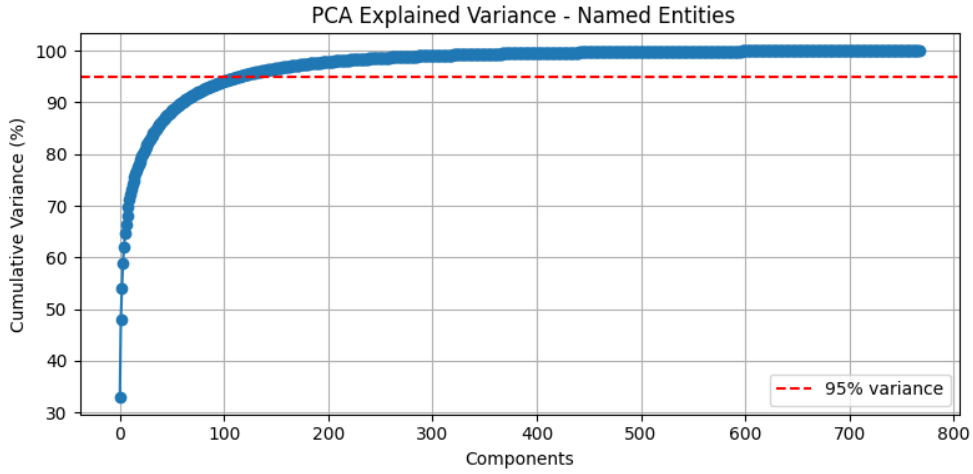
الشكل 23: أثر معامل حجم النافذة على النتائج.

ويوضح الشكل أدناه أسقاط تمثيل الرسائل على فضاء ثنائي الأبعاد.



التجربة الثانية: أضافة شعاع الكيانات المسماة

في هذه التجربة، من أجل كل نموذج من النماذج السابقة قمنا باستخراج الكيانات المسماة المذكورة في النص وتمثيلها باستخدام النموذج، وبعد ذلك اختصار البعد باستخدام PCA ومن ثم إضافتها إلى شعاع تمثيل الجملة Concatenation.

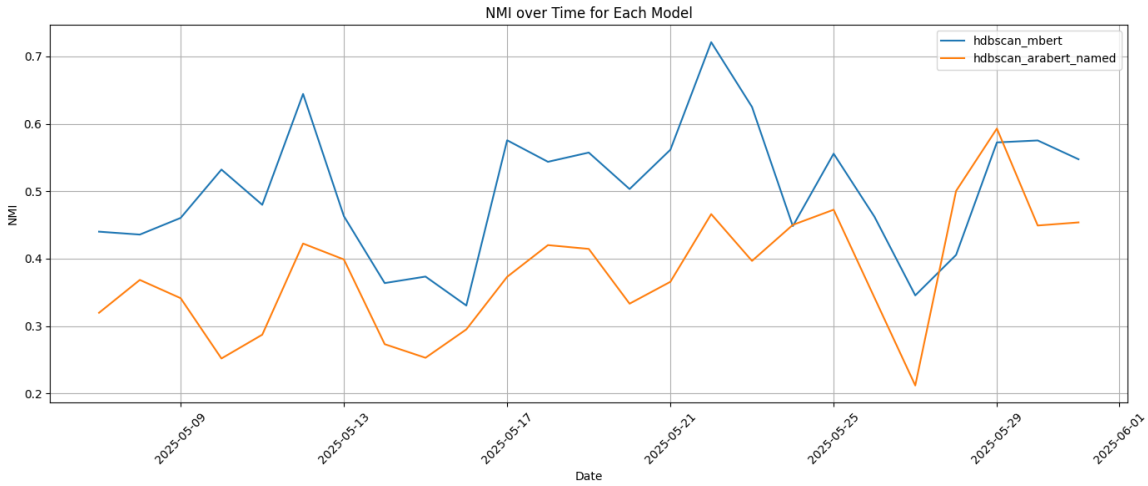


في هذه التجارب لم نحصل على قيمة مضافة على النتائج، حيث بقي mBert أفضل من AraBert + NE بـ 25%.

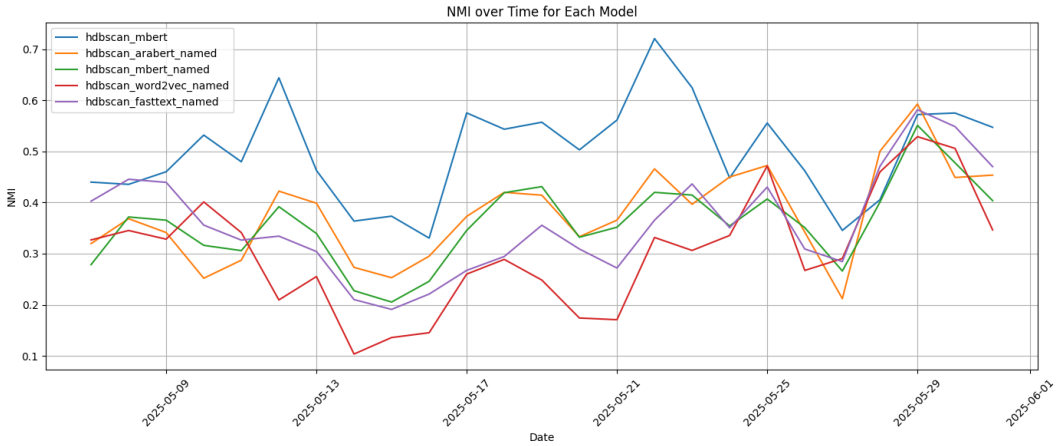
	M_1	M_2	M_3	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
mbert +NE	0.278	0.372	0.366	0.316	0.306	0.392	0.339	0.228	0.205	0.246	0.346	0.419
AraBert +NE	0.320	0.368	0.341	0.252	0.287	0.422	0.399	0.273	0.253	0.295	0.373	0.420
Word2Vec +NE	0.327	0.345	0.328	0.401	0.341	0.209	0.255	0.103	0.136	0.145	0.260	0.289
FastText +NE	0.403	0.446	0.440	0.356	0.327	0.334	0.304	0.210	0.191	0.221	0.267	0.294
Previous	0.440	0.436	0.460	0.532	0.480	0.644	0.463	0.364	0.373	0.330	0.576	0.544
	-27%	-15%	-26%	-53%	-40%	-34%	-14%	-25%	-32%	-11%	-35%	-23%

	M_{12}	M_{13}	M_{14}	M_{15}	M_{16}	M_{17}	M_{18}	M_{19}	M_{20}	M_{21}	M_{22}	M_{23}	M_{24}
mbert +NE	0.431	0.332	0.352	0.420	0.415	0.354	0.407	0.351	0.266	0.401	0.551	0.478	0.404
AraBert +NE	0.414	0.333	0.366	0.466	0.397	0.450	0.473	0.342	0.212	0.500	0.593	0.449	0.454
Word2Vec +NE	0.248	0.174	0.171	0.332	0.306	0.336	0.472	0.267	0.290	0.460	0.529	0.506	0.346
FastText +NE	0.356	0.309	0.272	0.366	0.437	0.351	0.430	0.309	0.284	0.471	0.582	0.549	0.470
Previous	0.557	0.503	0.562	0.721	0.625	0.448	0.556	0.462	0.346	0.406	0.572	0.575	0.547
	-26%	-34%	-35%	-35%	-37%	0%	-15%	-26%	-39%	23%	4%	-22%	-17%

ويضح الرسم أدناه الفارق بينهما

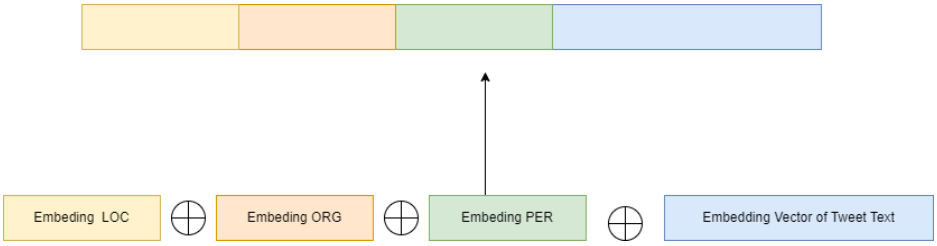


كما ويبين الشكل أدناه مقارنة الطرائق المتبعة في هذه التجربة مع mBert.



التجربة الثالثة: أضافة شعاع الكيانات المسماة بشكل منفصل

في هذه التجربة، من أجل كل نموذج من النماذج السابقة قمنا باستخراج الكيانات المسماة المذكورة بشكل منفصل (الأماكن المذكورة، الأشخاص، المنظمات) في النص وتمثيل كل منها بشعاع باستخدام النموذج، وبعد ذلك اختصار البعد لكل منها باستخدام PCA ومن ثم إضافتهم إلى شعاع تمثيل الجملة Concatenation، كما هو مبين بالشكل أدناه.



ولكن، وكما في التجربة السابقة، أيضاً لم نحصل على قيمة مضافة. حيث بقي نموذج mBert، ولكن كانت نتائج هذه التجربة أفضل من سابقتها حيث اقتربت النتائج من نتائج mBert التي حصلنا عليها في التجربة الأولى.

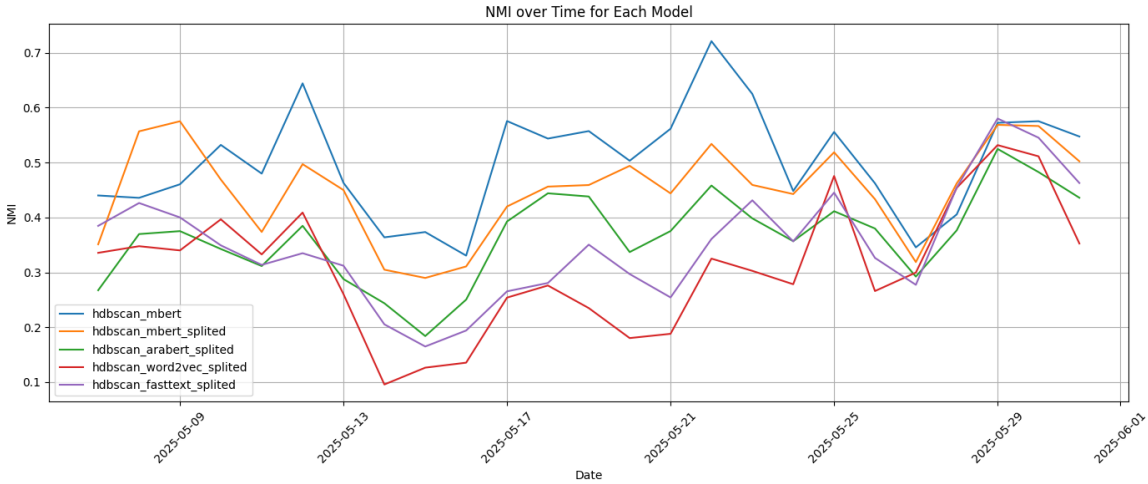
ويوضح الجدول أدناه النتائج التي حصنا عليها.

	M_1	M_2	M_3	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
mbert	0.351	0.557	0.575	0.469	0.374	0.497	0.450	0.305	0.290	0.311	0.420	0.456
AraBert	0.267	0.370	0.375	0.343	0.312	0.385	0.288	0.244	0.184	0.250	0.393	0.444
Word2Vec	0.336	0.348	0.340	0.397	0.333	0.409	0.260	0.096	0.126	0.135	0.254	0.276
FastText	0.385	0.426	0.400	0.349	0.314	0.335	0.312	0.205	0.165	0.194	0.265	0.281
Previous	0.440	0.436	0.460	0.532	0.480	0.644	0.463	0.364	0.373	0.330	0.576	0.544
	-20%	28%	25%	-12%	-22%	-23%	-3%	-16%	-22%	-6%	-27%	-16%

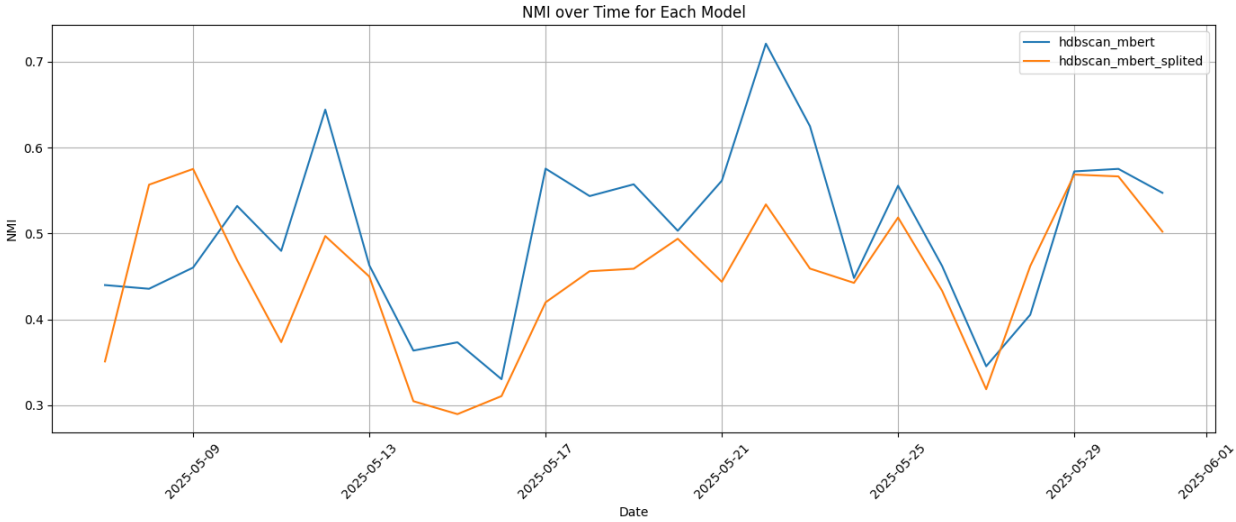
	M_{12}	M_{13}	M_{14}	M_{15}	M_{16}	M_{17}	M_{18}	M_{19}	M_{20}	M_{21}	M_{22}	M_{23}	M_{24}
mbert	0.459	0.494	0.444	0.534	0.459	0.443	0.519	0.433	0.319	0.462	0.569	0.567	0.502
AraBert	0.438	0.337	0.375	0.458	0.398	0.357	0.411	0.380	0.292	0.377	0.525	0.483	0.436
Word2Vec	0.235	0.180	0.188	0.325	0.303	0.278	0.475	0.266	0.300	0.454	0.532	0.511	0.353
FastText	0.350	0.297	0.254	0.361	0.431	0.356	0.445	0.326	0.277	0.455	0.580	0.545	0.463
Previous	0.557	0.503	0.562	0.721	0.625	0.448	0.556	0.462	0.346	0.406	0.572	0.575	0.547
	-18%	-2%	-21%	-26%	-27%	-1%	-7%	-6%	-8%	14%	-1%	-2%	-8%

جدول 10: نتائج التجربة الثالثة.

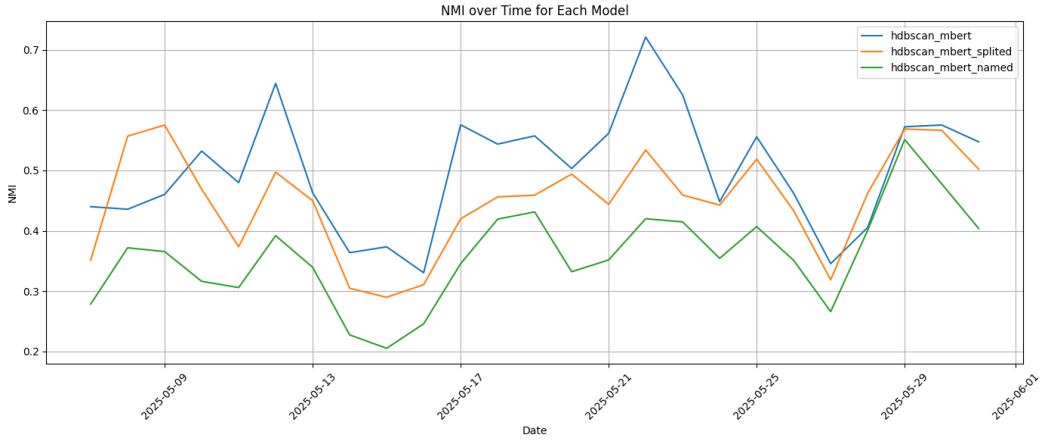
فيما يوضح الرسم البياني نتائج التجربة ومقارنتها مع mBert.

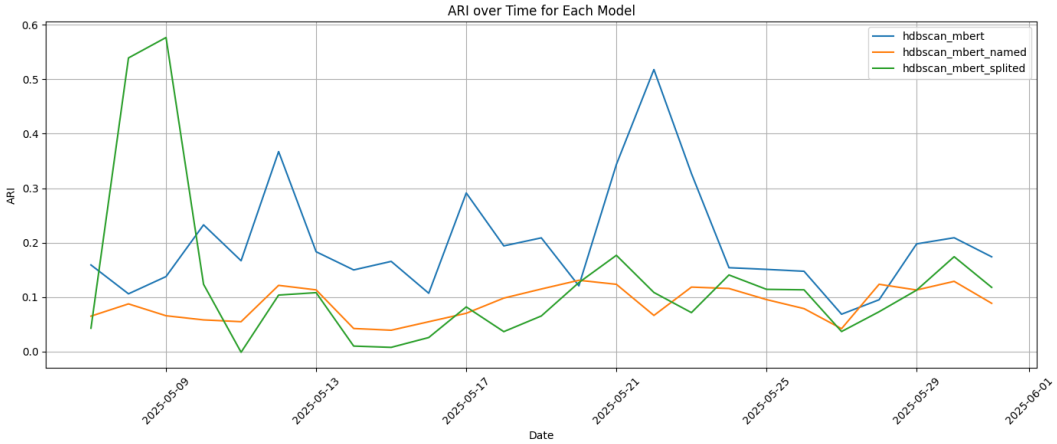


ويبين الرسم أدناه مقارنة mBert مع أفضلهم.

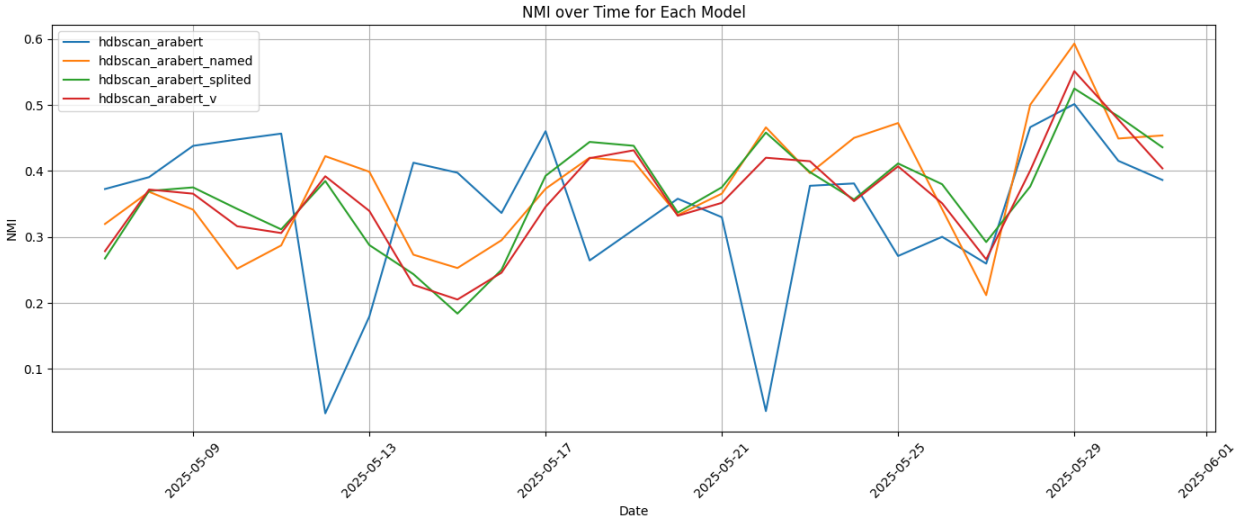


فيما يبين الرسم أدناه مقارنة نتائج النموذج مع إضافة الاشعة إليه.

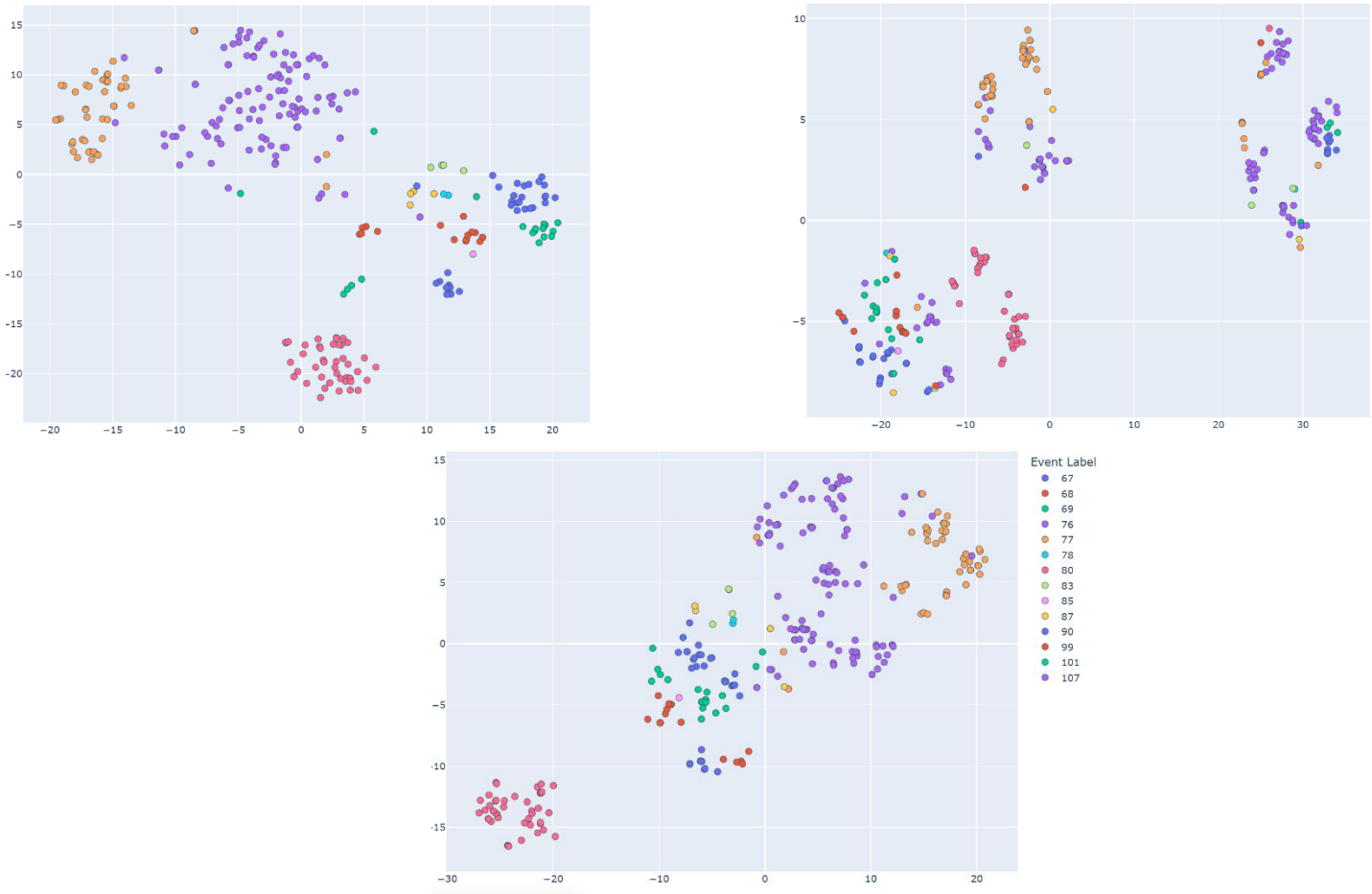




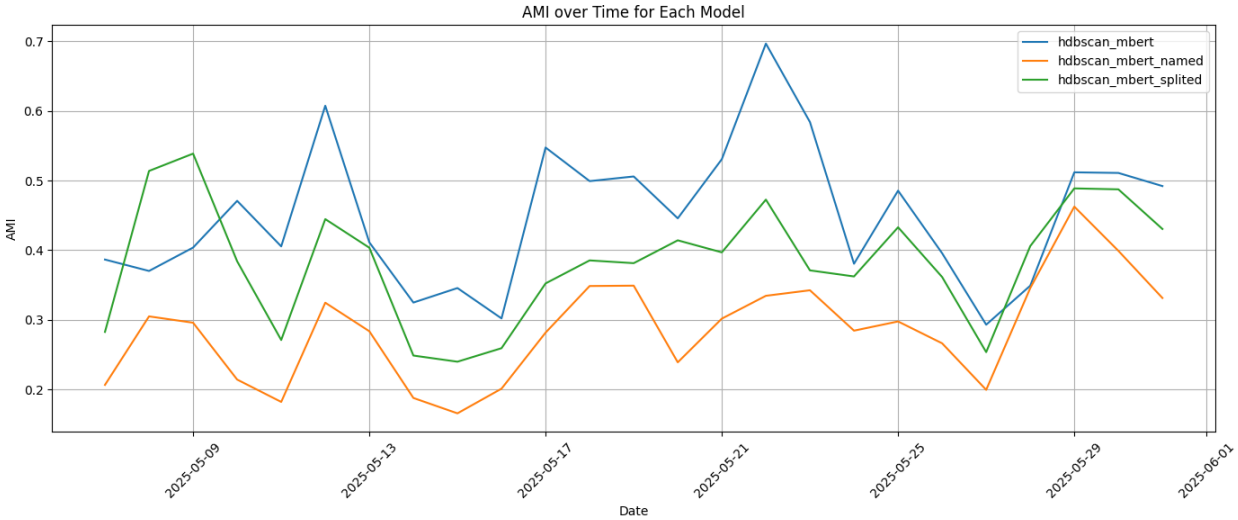
ويبين الشكل أدناه كيف أن الإضافات التي قمنا بها عل التمثيل جعلت التميل باستخدام arabert مستقراً أكثر.

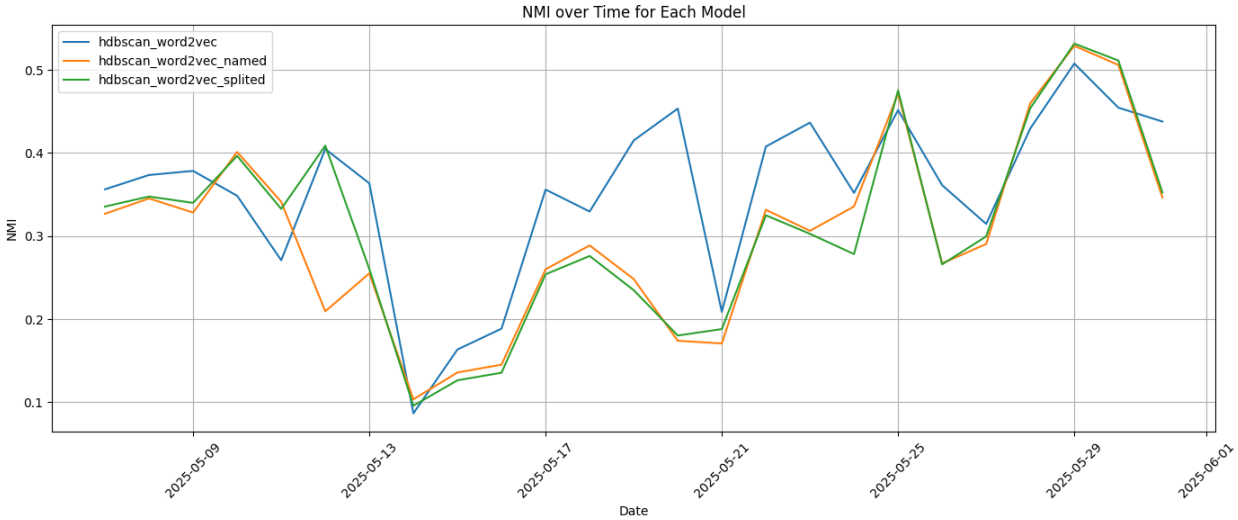


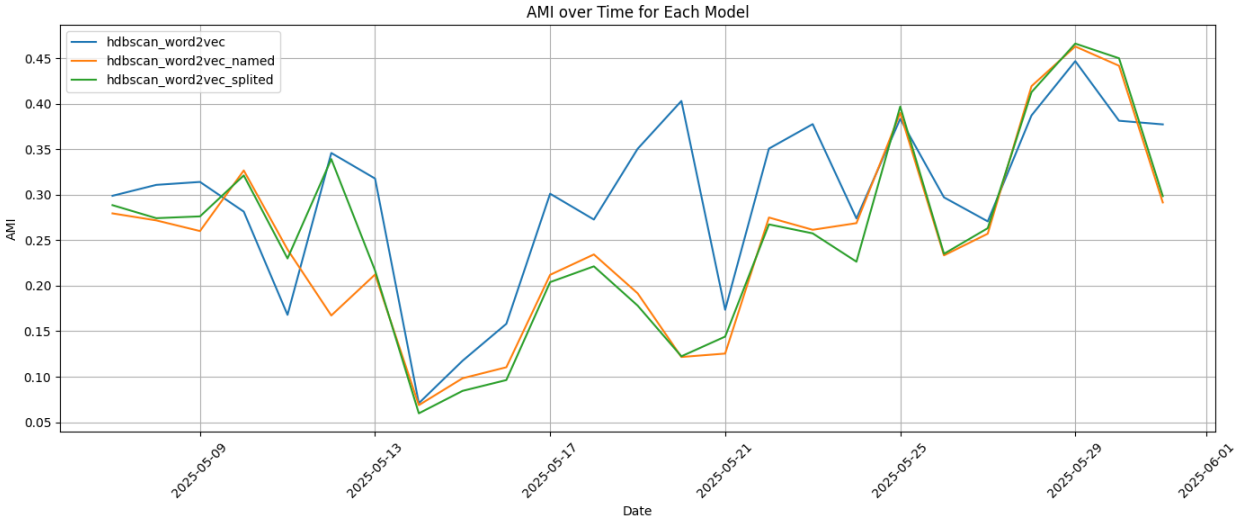
ويبين الشكل أدناه نتائج العنقدة على كتلة رسائل، في التجارب السابقة.

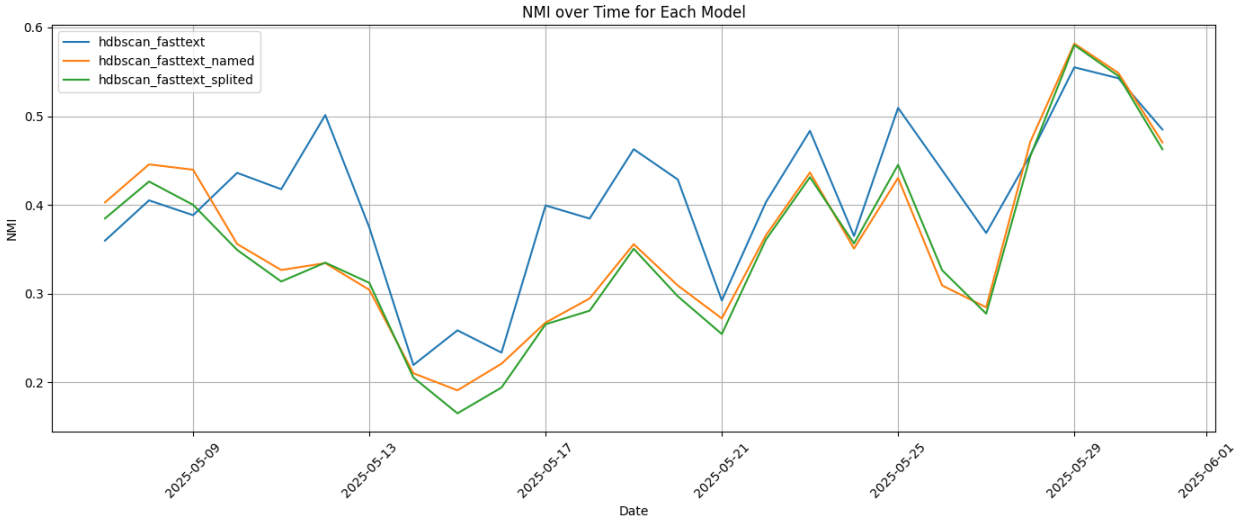


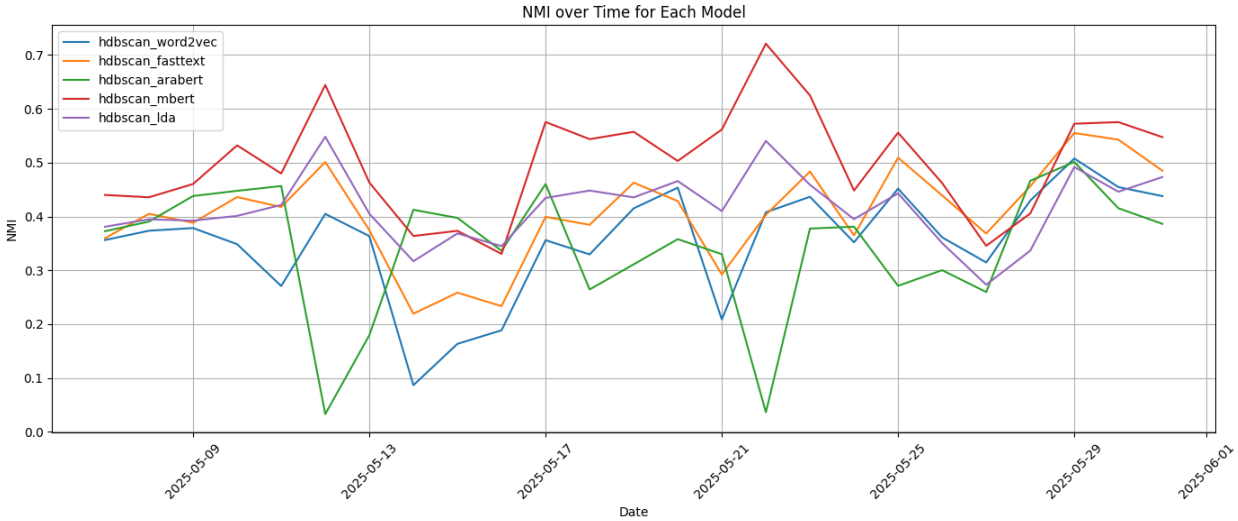
التجربة الرابعة: تعزيز مجموع بيانات الضبط

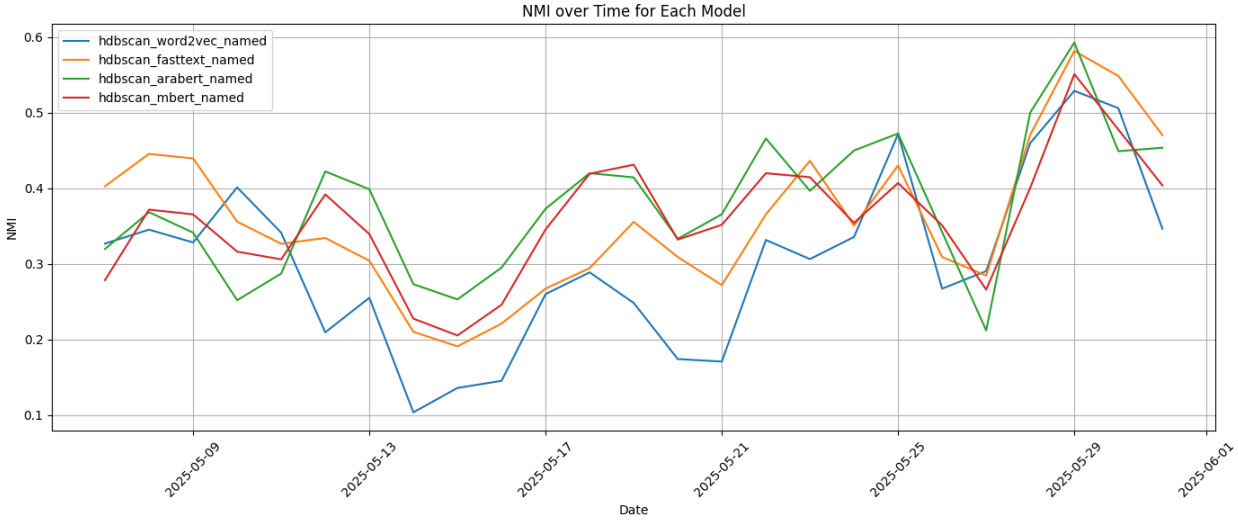

AMI over Time for Each Model
hdbscan_mbert
hdbscan_mbert_named
hdbscan_mbert_splited
AMI
Date
0.7
0.6
0.5
0.4
0.3
0.2
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01

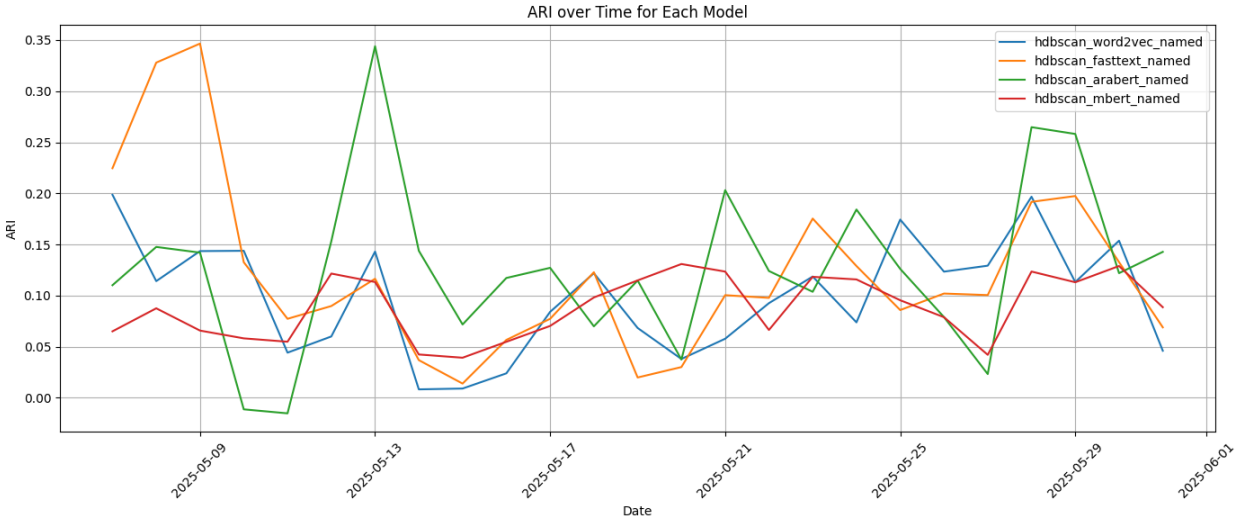

NMI over Time for Each Model
hdbscan_word2vec
hdbscan_word2vec_named
hdbscan_word2vec_splited
NMI
Date
0.5
0.4
0.3
0.2
0.1
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01

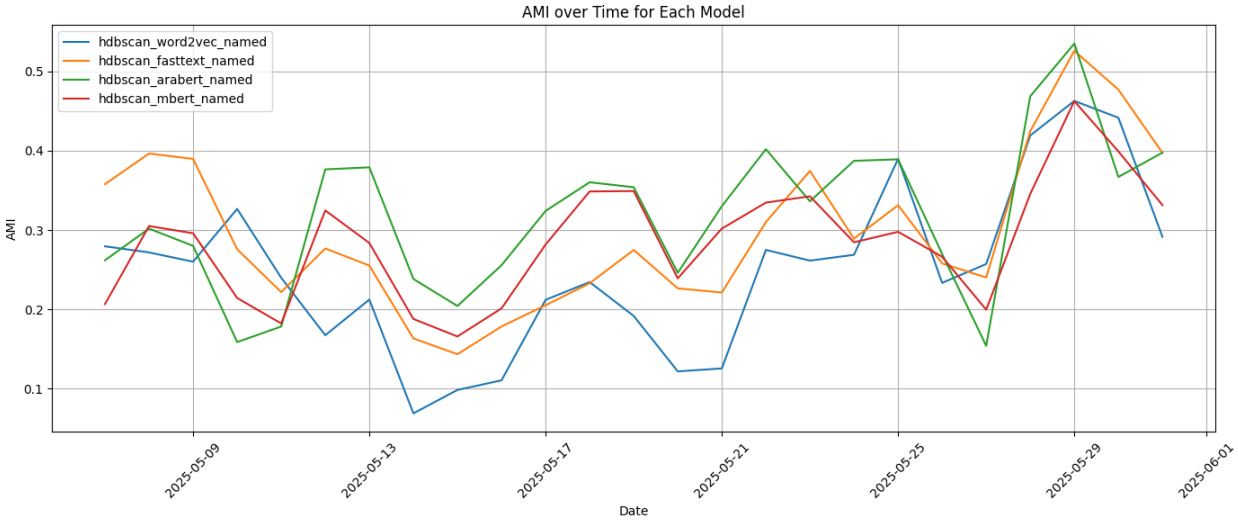

AMI over Time for Each Model
hdbscan_word2vec
hdbscan_word2vec_named
hdbscan_word2vec_splited
AMI
Date
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01

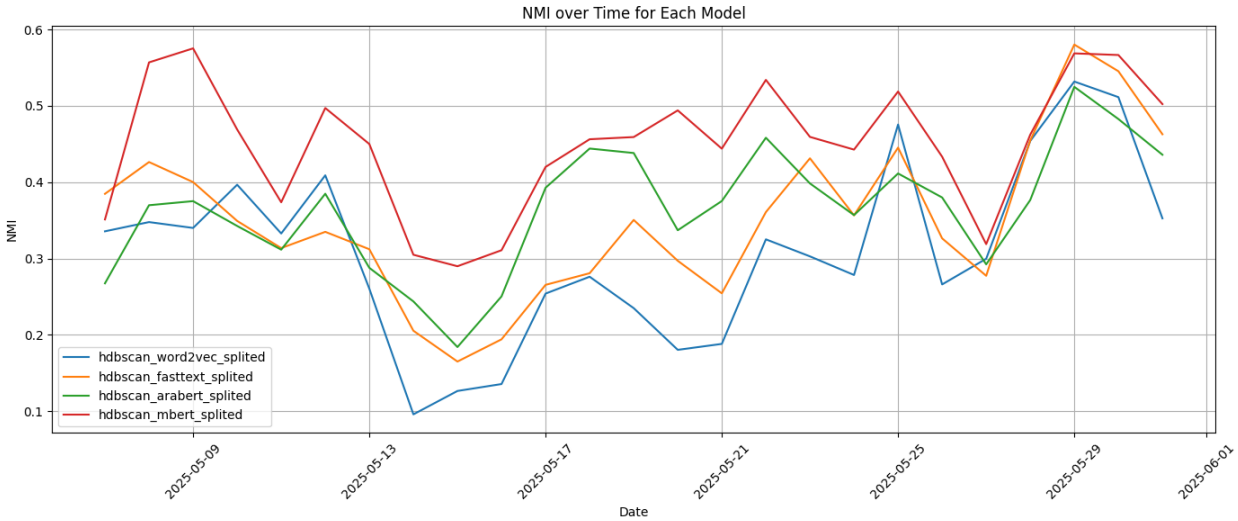

NMI over Time for Each Model
hdbscan_fasttext
hdbscan_fasttext_named
hdbscan_fasttext_splited
NMI
Date
0.2
0.3
0.4
0.5
0.6
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01

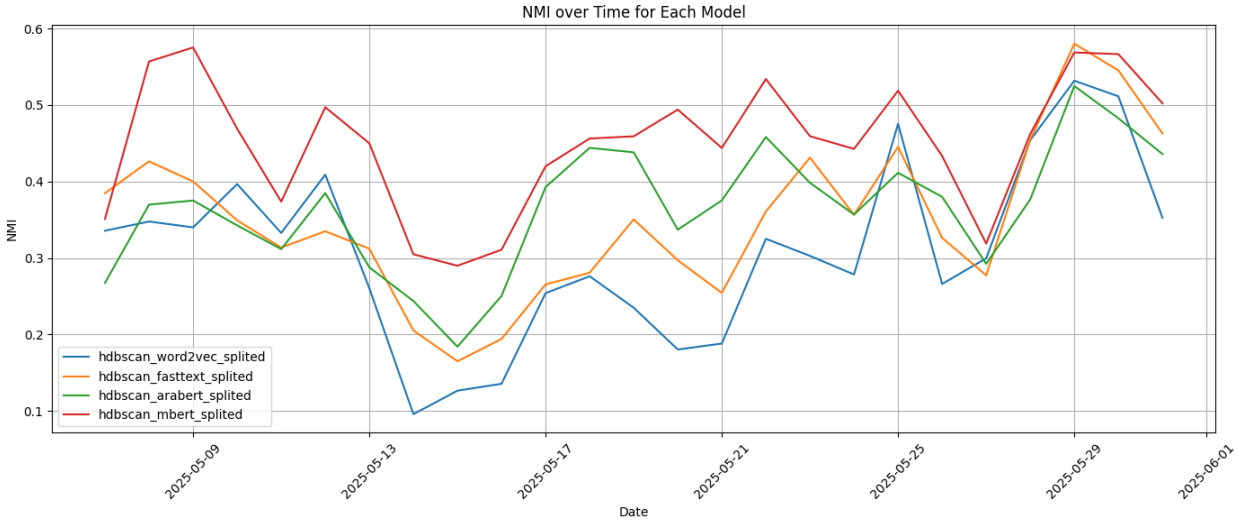

NMI over Time for Each Model
hdbscan_word2vec
hdbscan_fasttext
hdbscan_arabert
hdbscan_mbert
hdbscan_lda
NMI
Date
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01

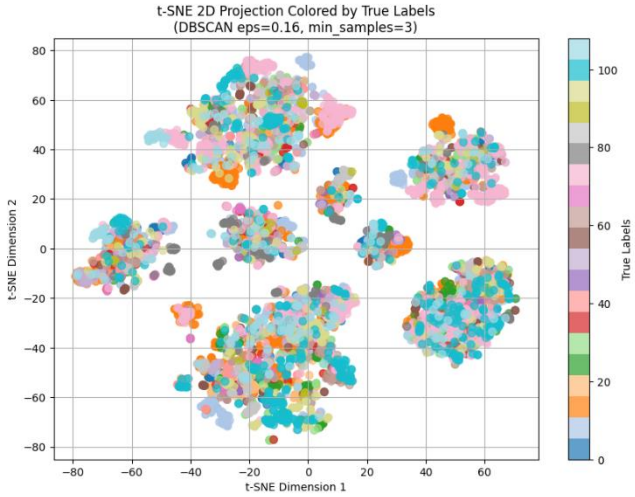

NMI over Time for Each Model
hdbscan_word2vec_named
hdbscan_fasttext_named
hdbscan_arabert_named
hdbscan_mbert_named
NMI
Date
0.6
0.5
0.4
0.3
0.2
0.1
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01

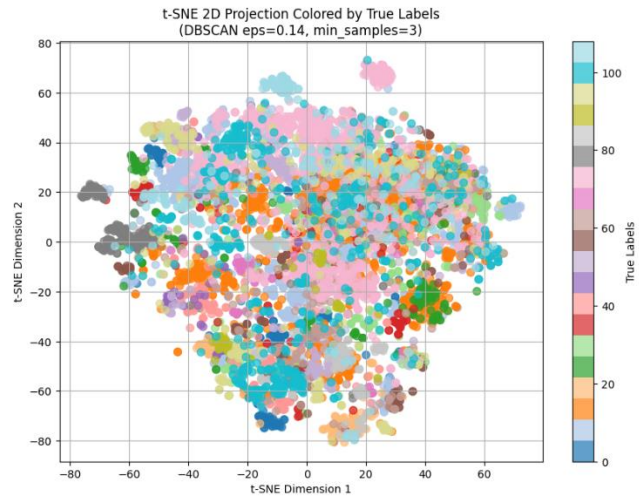

ARI over Time for Each Model
hdbscan_word2vec_named
hdbscan_fasttext_named
hdbscan_arabert_named
hdbscan_mbert_named
ARI
Date
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01

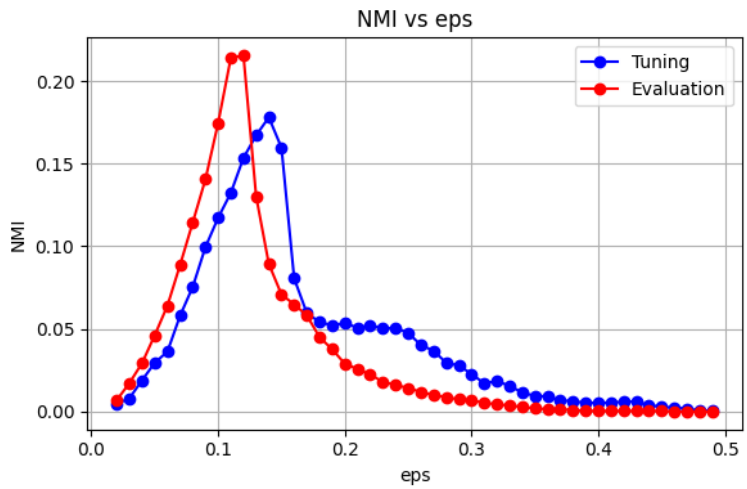

AMI over Time for Each Model
hdbscan_word2vec_named
hdbscan_fasttext_named
hdbscan_arabert_named
hdbscan_mbert_named
AMI
Date
0.1
0.2
0.3
0.4
0.5
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01

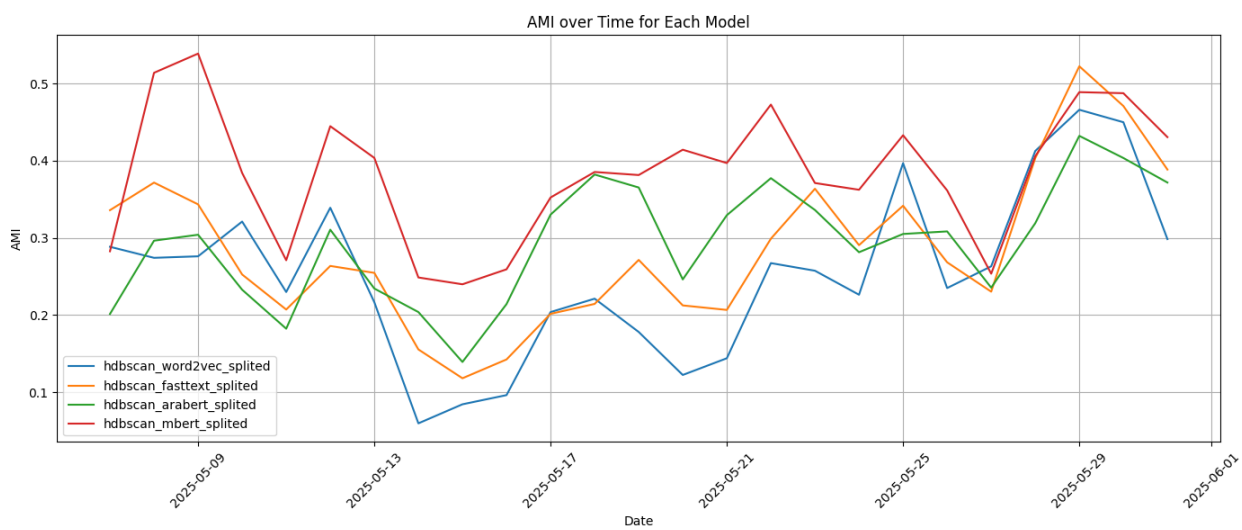

NMI over Time for Each Model
hdbscan_word2vec_splited
hdbscan_fasttext_splited
hdbscan_arabert_splited
hdbscan_mbert_splited
NMI
Date
0.1
0.2
0.3
0.4
0.5
0.6
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01


NMI over Time for Each Model
hdbscan_word2vec_splited
hdbscan_fasttext_splited
hdbscan_arabert_splited
hdbscan_mbert_splited
NMI
Date
0.1
0.2
0.3
0.4
0.5
0.6
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01


t-SNE 2D Projection Colored by True Labels
(DBSCAN eps=0.16, min_samples=3)
t-SNE Dimension 1
t-SNE Dimension 2
True Labels


t-SNE 2D Projection Colored by True Labels
(DBSCAN eps=0.14, min_samples=3)
t-SNE Dimension 1
t-SNE Dimension 2
True Labels


NMI vs eps
Tuning
Evaluation
NMI
eps


AMI over Time for Each Model
hdbscan_word2vec_splited
hdbscan_fasttext_splited
hdbscan_arabert_splited
hdbscan_mbert_splited
AMI
Date

الفصل التاسع

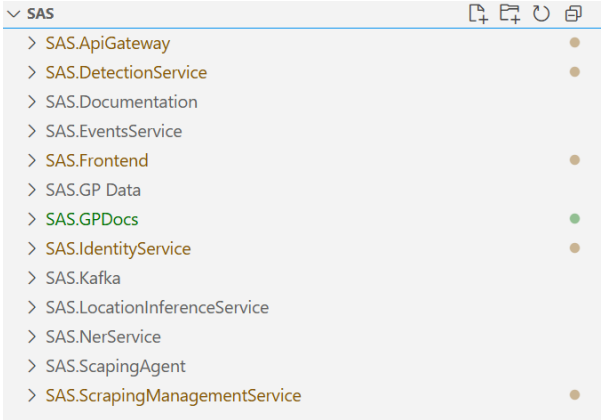
تنجيز النظام

يعرض هذا الفصل كيفية تنجيز النظام.

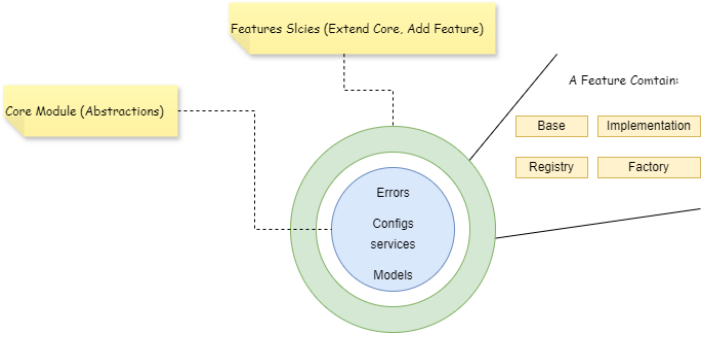
1.1- مقدمة

خلال عملية تطوير الخدمات يتم رفع الرماز المصدري الخاص بكل خدمة باستخدام Git إلى مستودع (Repository) ضمن الموقع Github، ضمن فرع Branch يسمى dev يرمز إلى مرحلة التطوير، حيث عند إضافة ميزة جديدة يتم إنشاء فرع انطلاقاً من الفرع dev، وعند الانتهاء منها يتم طلب دمج فرعها مع فرع التطوير بعد تشغيل خط أنابيب الاختبار Pipeline الخاصة باختبارات الوحدة والتكامل والجودة لها، وبذلك نضمن أن فرع التطوير dev يحوي على رماز مصدري مختبر ويعمل، وضمن الجودة المرجوة. وبهذه الطريقة نكون قد حققنا التكامل المستمر (Continues Integration).

يوضح الشكل أدناه الخدمات المنجزة، حيث كل خدمة منفصلة ولها مستودع خاص بها واختباراتها وخط أنابيب لها.

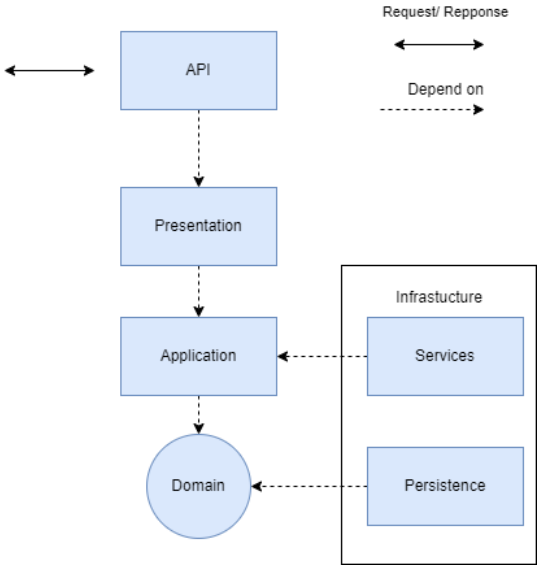


وعند تنجيزنا للخدمات المطورة باستخدام لغة Python، أملينا على أنفسنا مجموعة من القواعد، لبناء خدمة قابلة للتوسع وإعادة الاستخدام بسهولة، اعتمدنا في وضعها على البنية المعمارية النظيفة، ومعمارية الشرائح العمودية Vertical Slice Architecture. بحيث تساعدنا في وضع المكونات التي تتغير سوية بجانب بعضها، وإبقاء المكونات المستقرة بجانب بعضها.



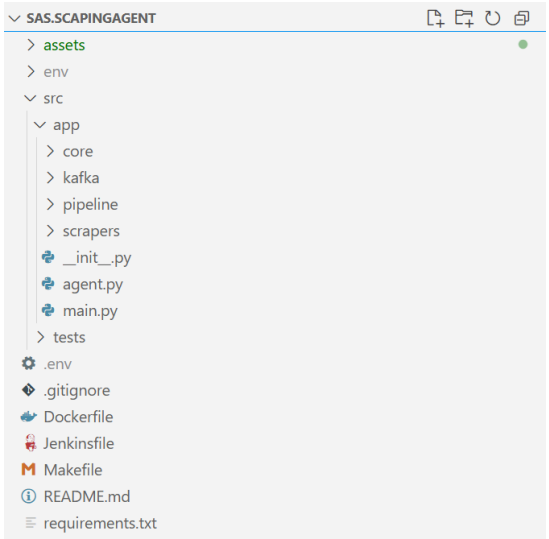
بحيث، في كل مشروع يكون هنالك مكون يحتوي على تجريدات Abstraction، وعندما نريد إضافة ميزة ما، فإن كل ما علينا هو إضافة مجلد باسمها، إضافة تجريد لها ضمن مجلدها، وكتابة التنجيزات الممكنة وتعريف المسجل، تعريف آلية بناءها.

وبالنسبة للخدمات المعتمدة على إطار العمل ASP.NET، اعتمدت البينة المعمارية النظيفة وفق المشاريع موضحة بالشكل أدناه



2.1- تنجيز الخدمات

خدمة سحب البيانات

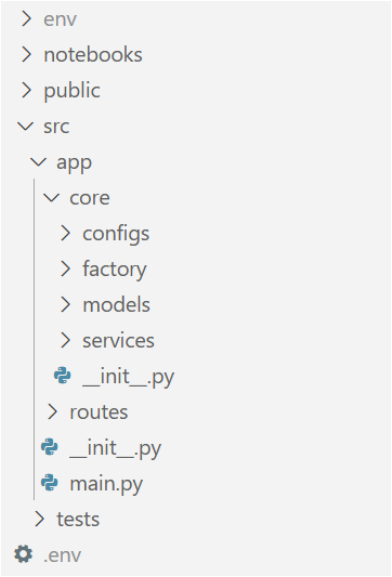



```python
async for task in tasks_service.stream_tasks():
    scraper = factory.create_scraper(task.scraping_approach)

    agent.set_scraper(scraper)
    agent.assign_task(task)
    await agent.run()
```

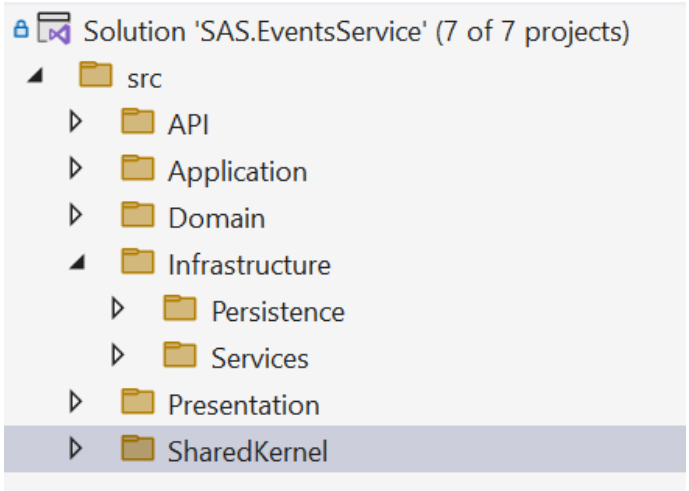
خدمة إدارة عمليات السحب

خدمة كشف الأحداث

خدمة الاستدلال على الموقع



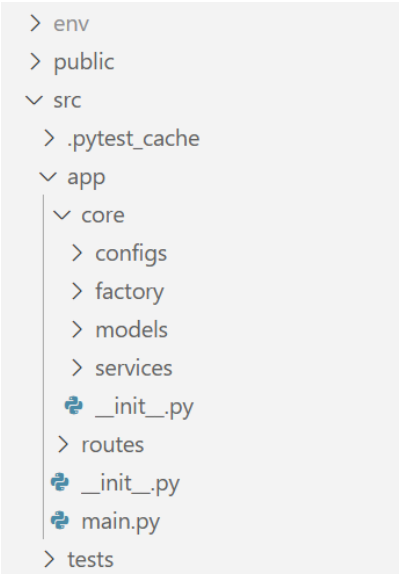
خدمة إدارة الأحداث



خدمة البوابة

خدمة واجهة المستخدم

خدمة استخراج الكيانات المسماة



3.1- مخطط النظام

يهدف

3.1- نشر الخدمات

الفصل التاسع

تحليل ومناقشة النتائج

يوضح هذا الفصل بعض المفاهيم المستخدمة في هذا العمل.

1.1- مقدمة

تعتبر الصفحات

2.1- معايير التقييم

ن

الفصل العاشر

اختبارات النظام

يوضح هذا الفصل بعض المفاهيم المستخدمة في هذا العمل.

1.1- مقدمة

تعتبر ال.

2.1- اختبار خدمات النظام

نطلب

3.1- اختبارات الأداء

ي

الخاتمة والآفاق المستقبلية

يهدف هذا الجزء كما تشير تسميته إلى تحديد الغاية من المشروع وتعريف دفتر شروطه الأولي، بالإضافة إلى لمحة عما تم إنجازه فعلياً خلال العمل. يفضل أن يخلو الملخص من المقدمات بحيث لا يتجاوز نصف صفحة. ننصح أيضاً بكتابة نفس الملخص باللغة الإنكليزية (على نفس الصفحة إن أمكن) وذلك بغية التعريف بالعم بشكل أوسع عبر محركات بحث الإنترنت إذا تم وضع التقرير إلكترونياً على الشبكة.(Zhang et al., 2021)

عادةً ما تُطبع صفحة الخلاصة أيضاً على الواجهة الخلفية للتقرير المسماة بالغلاف الرابع (Forth cover)، مما يمكّن الآخرين

المراجع

[1.] Abagissa, A. T., Saxena, S., & Chandra, J. (2024). *Distilbert-gnn: A Powerful Approach to Social Media Event Detection*. In Review. https://doi.org/10.21203/rs.3.rs-4193412/v1

[2.] Afyouni, I., Aghbari, Z. A., & Razack, R. A. (2022). Multi-feature, multi-modal, and multi-source social event detection: A comprehensive survey. *Information Fusion*, *79*, 279–308. https://doi.org/10.1016/j.inffus.2021.10.013

[3.] Afyouni, I., Khan, A., & Aghbari, Z. A. (2022). Deep-Eware: Spatio-temporal social event detection using a hybrid learning model. *Journal of Big Data*, *9*(1), 86. https://doi.org/10.1186/s40537-022-00636-w

[4.] Aggarwal, C. C., Yu, P. S., Han, J., & Wang, J. (2003). A Framework for Clustering Evolving Data Streams. In *Proceedings 2003 VLDB Conference* (pp. 81–92). Elsevier. https://doi.org/10.1016/B978-012722442-8/50016-1

[5.] Alharbi, A., & Lee, M. (n.d.). *Kawa¯rith: An Arabic Twitter Corpus for Crisis Events*.

[6.] Alsaedi, N., & Burnap, P. (2015). Arabic Event Detection in Social Media. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (Vol. 9041, pp. 384–401). Springer International Publishing. https://doi.org/10.1007/978-3-319-18111-0_29

[7.] Alsaedi, N., Burnap, P., & Rana, O. (2021). Sensing Real-World Events Using Arabic Twitter Posts. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 515–518. https://doi.org/10.1609/icwsm.v10i1.14765

[8.] Bok, K., Kim, I., Lim, J., & Yoo, J. (2023). Efficient graph-based event detection scheme on social media. *Information Sciences*.

[9.] Cao, P., Chen, Y., Zhao, J., & Wang, T. (2020). Incremental Event Detection via Knowledge Consolidation Networks. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 707–717. https://doi.org/10.18653/v1/2020.emnlp-main.52

[10.] Cao, Y., Peng, H., Wu, J., Dou, Y., Li, J., & Yu, P. S. (2021). Knowledge-Preserving Incremental Social Event Detection via Heterogeneous GNNs. *Proceedings of the Web Conference 2021*, 3383–3395. https://doi.org/10.1145/3442381.3449834

[11.] Cao, Y., Peng, H., Yu, Z., & Yu, P. S. (2023). *Hierarchical and Incremental Structural Entropy Minimization for Unsupervised Social Event Detection* (No. arXiv:2312.11891). arXiv. https://doi.org/10.48550/arXiv.2312.11891

[12.] Corney, D., Martin, C., & Göker, A. (2014). Spot the Ball: Detecting Sports Events on Twitter. In M. De Rijke, T. Kenter, A. P. De Vries, C. Zhai, F. De Jong, K. Radinsky, & K. Hofmann (Eds.), *Advances in Information Retrieval* (Vol. 8416, pp. 449–454). Springer International Publishing. https://doi.org/10.1007/978-3-319-06028-6_40

[13.] Deng, S., Zhang, N., Kang, J., Zhang, Y., Zhang, W., & Chen, H. (2020). Meta-Learning with Dynamic-Memory-Based Prototypical Network for Few-Shot Event Detection. *Proceedings of the 13th International Conference on Web Search and Data Mining*, 151–159. https://doi.org/10.1145/3336191.3371796

[14.] Fedoryszak, M., Frederick, B., Rajaram, V., & Zhong, C. (2019). Real-time Event Detection on Social Data Streams. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2774–2782. https://doi.org/10.1145/3292500.3330689

[15.] George, Y., Karunasekera, S., Harwood, A., & Lim, K. H. (2021). *Real-time spatio-temporal event detection on geotagged social media*.

[16.] Ghaemi, Z., & Farnaghi, M. (2023). Event detection from geotagged tweets considering spatial autocorrelation and heterogeneity. *Journal of Spatial Science*, *68*(3), 353–371. https://doi.org/10.1080/14498596.2021.2002201

[17.] Guille, A., & Favre, C. (2015). Event detection, tracking, and visualization in Twitter: A mention-anomaly-based approach. *Social Network Analysis and Mining*, *5*(1), 18. https://doi.org/10.1007/s13278-015-0258-0

[18.] Guo, Y., Zang, Z., Gao, H., Xu, X., Wang, R., Liu, L., & Li, J. (2024). Unsupervised Social Event Detection via Hybrid Graph Contrastive Learning and Reinforced Incremental Clustering. *Knowledge-Based Systems*, *284*, 111225. https://doi.org/10.1016/j.knosys.2023.111225

[19.] Hamoui, B., Mars, M., & Almotairi, K. (n.d.). *FloDusTA: Saudi Tweets Dataset for Flood, Dust Storm, and Traffic Accident Events*.

[20.] Hasan, M., Orgun, M. A., & Schwitter, R. (2016). *TwitterNews: Real time event detection from the Twitter data stream*. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.2297v1

[21.] Hettiarachchi, H., Adedoyin-Olowe, M., Bhogal, J., & Gaber, M. M. (2022). Embed2Detect: Temporally Clustered Embedded Words for Event Detection in Social Media. *Machine Learning*, *111*(1), 49–87. https://doi.org/10.1007/s10994-021-05988-7

[22.] Hossny, A. H., & Mitchell, L. (2019). *Event detection in Twitter: A keyword volume approach* (No. arXiv:1901.00570). arXiv. https://doi.org/10.48550/arXiv.1901.00570

[23.] Igartua, M. A., Almenares, F., Redondo, R. P. D., Martín, M. I., Forné, J., Campo, C., Fernández, A., Cruz, L. J. de la, García-Rubio, C., Marínn, A., Mezher, A. M., Díaz, D., Cerezo, H., Rebollo-Monedero, D., Arias, P., & Rico, F. (2020). INRISCO: INcident monitoRing In Smart COmmunities. *IEEE Access*, *8*, 72435–72460. https://doi.org/10.1109/ACCESS.2020.2987483

[24.] Kovalchuk, M. A., Filatova, A., Korneev, A., Koreneva, M., Nasonov, D., Voskresenskii, A., & Boukhanovsky, A. (2024). SemConvTree: Semantic Convolutional Quadtrees for Multi-Scale Event Detection in Smart City. *Smart Cities*, *7*(5), 2763–2780. https://doi.org/10.3390/smartcities7050107

[25.] Li, P., Yu, X., Peng, H., Xian, Y., Wang, L., Sun, L., Zhang, J., & Yu, P. S. (2024). *Relational Prompt-based Pre-trained Language Models for Social Event Detection* (No. arXiv:2404.08263). arXiv. https://doi.org/10.48550/arXiv.2404.08263

[26.] Li, Q., Chao, Y., Li, D., Lu, Y., & Zhang, C. (2022). Event Detection from Social Media Stream: Methods, Datasets and Opportunities. *2022 IEEE International Conference on Big Data (Big Data)*, 3509–3516. https://doi.org/10.1109/BigData55660.2022.10020411

[27.] Li, Q., & Zhang, Q. (2021). Twitter Event Summarization by Exploiting Semantic Terms and Graph Network. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), 15347–15354. https://doi.org/10.1609/aaai.v35i17.17802

[28.] McMinn, A. J., Moshfeghi, Y., & Jose, J. M. (2013). Building a large-scale corpus for evaluating event detection on twitter. *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 409–418. https://doi.org/10.1145/2505515.2505695

[29.] Morabia, K., Murthy, N. L. B., Malapati, A., & Samant, S. (n.d.). *SEDTWik: Segmentation-based Event Detection from Tweets Using Wikipedia*.

[30.] Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014). CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 376–385. https://doi.org/10.1609/icwsm.v8i1.14538

[31.] Patil, P., Patil, P., Salvi, A., & Hatiskar, M. (2015). *Social Sensor for Real Time Event Detection*. *5*(3).

[32.] Peng, H., Zhang, R., Li, S., Cao, Y., Pan, S., & Yu, P. S. (2023). Reinforced, Incremental and Cross-Lingual Event Detection From Social Messages. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 980–998. https://doi.org/10.1109/TPAMI.2022.3144993

[33.] Pradhan, A. K., Mohanty, H., & Lal, R. P. (2024). EDTBERT: Event Detection and Tracking in Twitter using Graph Clustering and Pre-trained Language Model. *Procedia Computer Science*, *233*, 481–491. https://doi.org/10.1016/j.procs.2024.03.238

[34.] Qazi, U., Imran, M., & Ofli, F. (2020). GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information. *SIGSPATIAL Special*, *12*(1), 6–15. https://doi.org/10.1145/3404820.3404823

[35.] Ray, M., Wang, Q., Mélanie-Becquet, F., Poibeau, T., & Mazoyer, B. (2024). An Incremental Clustering Baseline for Event Detection on Twitter. *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, 18–24. https://doi.org/10.18653/v1/2024.futured-1.2

[36.] Ren, J., Jiang, L., Peng, H., Cao, Y., Wu, J., Yu, P. S., & He, L. (2022). *From Known to Unknown: Quality-aware Self-improving Graph Neural Network for Open Set Social Event Detection* (No. arXiv:2208.06973). arXiv. https://doi.org/10.48550/arXiv.2208.06973

[37.] Saeed, Z., Abbasi, R. A., Maqbool, O., Sadaf, A., Razzak, I., Daud, A., Aljohani, N. R., & Xu, G. (2019). What's Happening Around the World? A Survey and Framework on Event Detection Techniques on Twitter. *Journal of Grid Computing*, *17*(2), 279–312. https://doi.org/10.1007/s10723-019-09482-2

[38.] Saeed, Z., Abbasi, R. A., Razzak, M. I., & Xu, G. (2019). *Event Detection in Twitter Stream using Weighted Dynamic Heartbeat Graph Approach* (No. arXiv:1902.08522). arXiv. https://doi.org/10.48550/arXiv.1902.08522

[39.] Schinas, M., Papadopoulos, S., Kompatsiaris, Y., & Mitkas, P. (2018). *Event Detection and Retrieval on Social Media* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1807.03675

[40.] Suwaileh, R., Elsayed, T., & Imran, M. (2023). IDRISI-D: Arabic and English Datasets and Benchmarks for Location Mention Disambiguation over Disaster Microblogs. *Proceedings of ArabicNLP 2023*, 158–169. https://doi.org/10.18653/v1/2023.arabicnlp-1.14

[41.] Wang, X., Wang, Z., Han, X., Jiang, W., Han, R., Liu, Z., Li, J., Li, P., Lin, Y., & Zhou, J. (2020). *MAVEN: A Massive General Domain Event Detection Dataset* (No. arXiv:2004.13590). arXiv. https://doi.org/10.48550/arXiv.2004.13590

[42.] Yamani, A., Alsulami, A., & Al-Zaidy, R. (n.d.). *Event-Based Knowledge MLM for Arabic Event Detection*.

[43.] Yang, Z., Wei, Y., Li, H., Li, Q., Jiang, L., Sun, L., Yu, X., Hu, C., & Peng, H. (2024). Adaptive Differentially Private Structural Entropy Minimization for Unsupervised Social Event Detection. *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2950–2960. https://doi.org/10.1145/3627673.3679537

[44.] Yu, X., Wei, Y., Zhou, S., Yang, Z., Sun, L., Peng, H., Zhu, L., & Yu, P. S. (2024). *Towards Effective, Efficient and Unsupervised Social Event Detection in the Hyperbolic Space* (No. arXiv:2412.10712). arXiv. https://doi.org/10.48550/arXiv.2412.10712

[45.] Yue, Z., Zeng, H., Lan, M., Ji, H., & Wang, D. (2023). Zero- and Few-Shot Event Detection via Prompt-Based Meta Learning. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7928–7943. https://doi.org/10.18653/v1/2023.acl-long.440

[46.] Zhang, C., Zhou, G., Yuan, Q., Zhuang, H., Zheng, Y., Kaplan, L., Wang, S., & Han, J. (2016). GeoBurst: Real-Time Local Event Detection in Geo-Tagged Tweet Streams. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 513–522. https://doi.org/10.1145/2911451.2911519

[47.] Zhang, S., Cheng, Y., & Ke, D. (2017). *Event-Radar: Real-time Local Event Detection System for Geo-Tagged Tweet Streams* (No. arXiv:1708.05878). arXiv. https://doi.org/10.48550/arXiv.1708.05878

الملاحق

ملحق أ

الدراسة الكميّة

نبين هنا الأرقام والقياسات التي يجب أن يعمل ضمنها النظام

مقدمة

نهدف من هذه الدراسة إلى تقدير الكميات المطلوبة لتصميم وتشغيل نظام كشف ومراقبة الأحداث في وسائل التواصل الاجتماعي.

عدد الأحداث اليومي

في بلد جميل مثل سوريا، وي مرحلة مفصلية في تاريخها كما اليوم، بدأت البلاد تخطو خطوات حثيثة نحو إعادة البناء، سواء على مستوى البنية التحتية أو على مستوى الإنسان والمجتمع. وفي هذا السياق، أصبحت الأحداث اليومية أكثر كثافة وتنوعًا، ليس فقط بسبب التحديات، بل لأن الحركة في المجتمع تعود تدريجيًا، ويتفاعل الناس مع الواقع من جديد عبر وسائل الإعلام التقليدي والاجتماعي. ففي مدينة مثل دمشق، أو حلب، يمكن أن نسجّل يوميًا عشرات الأحداث التي تشمل:

- إطلاق مشاريع خدمية جديدة
- فعاليات تعليمية وصحية
- حالات استجابة للطوارئ
- حوادث بسيطة أو أعطال بنية تحتية

وبناءً على البيانات التي جمعناها، يمكن تقدير عدد الأحداث اليومية في المدن الكبرى بـ 10 إلى 100 حدث يوميًا، أي في عموم البلاد يمكن أن يصل وسطياً إلى 500 حدث. وفي حالات الكوارث الطبيعية والأزمات، قد يرتفع هذا العدد إلى أكثر من ذلك بعشرات الأضعاف، ففي الأوقات ذات الطابع الطارئ أو عند وقوع كوارث طبيعية كبرى، فمثلاً، في زلزال 6 شباط 2023، حيث توالت البلاغات والمعلومات من جميع أنحاء البلاد خلال ساعات قليلة، يمكن أن يرتفع العدد إلى 5000 حدث في يوم واحد.

كمية البيانات الواجب سحبها

بناءً على تحليل الكلمات مفتاحية مرتبطة بالأحداث العامة، نُقدّر أن عدد التغريدات اليومية المرتبطة بالشأن السوري التي يتم تداولها، يتراوح بين 3,000 إلى 20,000 تغريدة في الأيام العادية، خاصة تلك المتعلقة بالخدمات، الأمن، الصحة، والنشاطات المجتمعية.

في الأوقات الخاصة مثل الكوارث أو الأزمات، يتضاعف هذا الرقم. فعلى سبيل المثال، خلال زلزال 6 شباط 2023، تم تسجيل أكثر من 500,000 تغريدة خلال أول 24 ساعة تتعلق مباشرة بالحدث ومجرياته، وفق تقارير تحليلية. وكذلك، في أحداث ذات طابع سياسي أو عسكري، مثل التصعيدات الإسرائيلية– الإيرانية، قد يصل حجم التفاعل إلى 100,000تغريدة أو أكثر في اليوم الواحد.

معدل دفق التغريدات

في الأيام الاعتيادية، ومع مراقبة كلمات مفتاحية محددة مرتبطة بالخدمات أو الأمن أو النشاطات المجتمعية، قد يصل معدل التدفق إلى 100 – 300 تغريدة في الدقيقة الواحدة، أي ما يعادل تقريبًا 2 إلى 5 تغريدات في الثانية.

عند حدوث أمر طارئ أو لافت، مثل انقطاع كبير للكهرباء، حادث مروري واسع النطاق، أو حتى حدث ثقافي ضخم، يقفز هذا الرقم إلى ما يزيد عن 1000 تغريدة في الدقيقة – أي ما يقارب 16 تغريدة في الثانية، وأحيانًا أكثر.

في حالات الكوارث الكبرى – مثل الزلازل أو التصعيدات العسكرية المفاجئة – يمكن أن يتضاعف هذا الرقم بسهولة، ليصل إلى 5000 تغريدة في الدقيقة أو ما يقارب 80 تغريدة في الثانية، موزعة بين تحديثات إخبارية، صور من شهود عيان، بلاغات غير رسمية، ومناشدات استغاثة.

عدد المستخدمين المتوقع

في مرحلة إعادة البناء، لا تُبنى البلاد بالإسمنت فقط، بل بالوعي، بالمعلومة، وبالقدرة على الوصول السريع إلى ما يحدث. ولهذا، فإن نظام كشف الأحداث ليس موجّهًا فقط لصنّاع القرار أو الجهات الأمنية، بل أيضًا للمواطن، وللصحفي، وللخدمة العامة التي تسعى لأن تكون حاضرة في المكان الصحيح في الوقت المناسب.

المستخدمون الداخليّون(المسجلون، المراقبون):

هؤلاء المستخدمون يأتون من لمستخدمين الراغبين بتلقي التنبيهات، المراسلين، الإعلاميين، ومحللي البيانات، المهتمين بالشأن السوري.

التقدير المبدئي يشير إلى ما بين 300 إلى 1000 مستخدم داخلي نشط يوميًا، بحسب حجم المنطقة الجغرافية التي يخدمها النظام وعدد الجهات المتعاونة معه.

المستخدمون الخارجيّون (الجمهور غير المسجلين):

يمكن أن يصل إلى 1000 – 50,000 تصفح نشط يوميًا في المرحلة الأولى، وقد يتجاوز 100,000 مستخدم يوميًا في حالات الأزمات الكبيرة أو عند تغطية أحداث تهمّ الرأي العام.

ملحق ب

دليل تنميط مجموعة البيانات

نرفق هنا الدليل الذي تم وضعه لتنميط مجوعة البيانات.

دليل تعليمات تصنيف التغريدات

يهدف هذا الملف إلى توضيح كيفية تصنيف التغريدات المرفقة إلى فئتين :ذات صلة بالمسألة (Relevant) وغير ذات صلة (Not Relevant).

الغرض

قم بوضع علامة على كل تغريدة على أنها ذات صلة أو غير ذات صلة استنادًا إلى ما إذا كانت تحتوي على معلومات مفيدة حول الأحداث المتعلقة بالسياق المستهدف (على سبيل المثال، الحوادث، والكوارث، والاحتجاجات، والصراعات) التي تريد اكتشافها

معايير التصنيف

ذات صلة: تتضمن معلومات عن حدث وقع بالفعل أو يجري حالياً أو سيحدث مستقبلاً، مثل تقارير عن هجمات، مظاهرات، اجتماعات سياسية، كوارث طبيعية، إلخ.

غير ذات صلة: لا تتضمن أي معلومات تتعلق بحدث فعلي، مثل الآراء العامة، النكات، النقاشات الدينية أو الطائفية دون ذكر حدث محدد.

أمثلة على تصنيف التغريدات

السبب	التصنيف	التغريدة
تقرير عن موقف سياسي	Relevant	أ ف ب عن قيادي كردي بارز :لا تنازل عن مطلب اللامركزية في محادثاتنا مع #دمشق
رأي بدون معلومات عن حدث	Not Relevant	مذيعين من الشارع، حقد طائفي وقومي -مش هيك الإعلام، مش هيك التعايش بيصير.
محتوى اجتماعي بدون علاقة بحدث	Not Relevant	وحده يمنية نزلت منشور في التيك توك إنها تخطبت لسوري .الردود من السوريين شيء يخجل

تقرير عن حدث سياسي/عسكري	Relevant	تركيا تسعى لإنشاء قاعدتين »بحرية وجوية «في سوريا
مزحة أو رأي بدون ذكر حدث	Not Relevant	@mehmod_kolo و هو راح سأل جميع السوريين هههههه 😂
تقرير عن حادثة ورد فعل	Relevant	الموساد ينفّذ عملية إنزال جوي في حمص، سوريا