الجمهورية العربية السورية
المعهد العالي للعلوم التطبيقية والتكنولوجيا
قسم النظم المعلوماتية
العام الدراسي 2024/2025

مشروع أُعِدَّ

لنيل درجة الإجازة في هندسة البرمجيات والذكاء الصنعي

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

تقديم

حسن بهجت خضور

إشراف

د. رياض سنبل

11/8/2025

الإهداء

إلى من أيقظني في كل صباح ودعا لي في كل مساء ... أمي الحبيبة.

إلى من كان دائما بجانبي وأوصلني إلى ما أنا عليه ... أبي الغالي.

إلى من قال بهم الرحمن سنشد عضدك بأخيك ... إخوتي حسين، علي.

إلى من علمني في صغري ووقفت إلى جانبي ... عمتي الغالية.

إلى روح المرحوم عمي أبو المثنى الذي تركنا فينا أثراً لا يُمحى، وسيرةً طيبةً تخلدها القلوب قبل الذاكرة. وإلى عائلته الكريمة التي كانت داعماً لي منذ بداية دراستي، وأخص بالشكر ابنة عمي فارعة لما قدمته من دعمٍ وعون.

إلى روح المرحوم عمي فؤاد الذي ترك في بصمةً لا تنسى في حياتي ومسيرتي، وإلى عائلته الكريمة.

إلى روح عمي الدكتور نزار أحمد خضور وأسرته وأخص بالشكر بديع، وديع، علي نزار خضور

إلى الذين كانوا سنداً وعوناً لي في كل مرحلة من حياتي، ... أخوالي الأعزاء

ناصر حسن خليل وعائلته.

رمضان حسن خليل وعائلته.

حافظ حسن خليل وعائلته.

إلى خالتي التي كانت داعماً صادقاً لي، ...

ردينة حسن خليل وعائلتها.

إلى من جمعتني معهم أيام جميلة وكان يحلو برفقتهم الطريق

خضر حسون، رواد ربيع.

كلمة شكر

أتقدم بخالص الشكر **للدكتور رياض سنبل** على المتابعة الدائمة لخطوات العمل، والملاحظات القيمة، والنصائح المستمرة، كما أشكر كل من كان له دور في مسيرتي التعليمية عموماً.

الخلاصة

إنّ وسائل التواصل الاجتماعي هي المصدر الأول والرئيسي للمعلومات والأحداث اليومية، وتحوي على معلومات حول ما يحصل حولنا، ومع تزايد اعتماد مجتمعنا عليها، كمصدر للمعلومات والأخبار، كان لابد من بناء نظام متكامل يقوم بتحليل محتوى هذه الوسائل ويستخلص منها الأحداث اليومية في بلدنا. أغنينا هذا المجال البحثي بمجموعة بيانات باللغة العربية تم جمعها خلال شهر أيار من العام 2025، كما قدمنا في هذا العمل، منهجية لكشف الأحداث في وسائل التواصل الاجتماعي، تعتمد على استخراج البيانات من خلال البحث بكلمات مفتاحية عامة، وتعزيزها بنتائج البحث، وتستفيد من تقنيات شبكات البيان العصبونية والتعلم المقارن لتعلم تمثيل جيد للبيانات بشكل خاضع للإشراف ذاتياً. وتم توظيف هذه المنهجية في نظام متكامل يعتمد على معمارية الخدمات المصغرة ليكون قادراً على التعامل مع كميات ضخمة من البيانات. ويسمح هذا النظام للمستخدم باستعراض الأحداث اليومية وإضافة مناطق اهتمام لتلقي إشعارات عندما يقع حدث فيها.

Abstract

Social media platforms have become the primary and most important source of information and daily events, containing valuable insights about what is happening around us. With the growing reliance of our society on these platforms for news and information, it has become essential to build an integrated system capable of analyzing their content and extracting daily events occurring within our country As part of this work, we enriched the research field by Collecting an Arabic-language dataset, collected during May 2025, specifically curated for the task of event detection from social media content.. In this work, we present a methodology for event detection on social media that relies on data extraction through general keyword searches, enhanced with search result expansion. It leverages graph neural network-based representation learning and contrastive learning techniques to obtain high-quality, self-supervised data embeddings. This methodology is employed within a fully integrated system based on a microservices architecture, enabling it to process large volumes of data efficiently. The system allows users to explore daily events and define regions of interest in order to receive notifications when relevant events occur.

المحتويات

قائمة الأشكال ..... **VI**
قائمة الجداول ..... **X**
الفصل الأول **1**
التعريف بالمشروع **1**
1.1– مقدمة **2**
2.1– الهدف من المشروع **2**
3.1– المتطلبات الوظيفية **2**
4.1– المتطلبات غير الوظيفية **3**
الفصل الثاني **4**
الدراسة النظرية **4**
1.2– العنقدة Clustering **5**
1.1.2 العنقدة القائمة على الكثافة Density Based Clustering 5
2.1.2- العنقدة الحية Online Clustering 6
3.1.2- خوارزمية العنقدة Single Pass 6
2.2– شبكات البيان العصبونية Graph Neural Network **6**
3.2– التعلم المقارن Contrastive Learning **7**
1.3.2- التعلم المقارن على البيان Graph Contrastive Learning 8
4.2– تقنيات تضمين الكلمات Word Embedding **8**
1.4.2 المحولات Transformers 8
2.4.2 Word2Vec 9
3.4.2 FastText 9
4.4.2 نموذج Bert (Bidirectional Encode Representation from Transformers) .. 9
5.2– LDA (Latent Dirichlet Allocation) **10**
6.2– شبكات المعلومات غير المتجانسة (HIN) **10**
7.2– البنية المعمارية النظيفة **11**

8.2– معمارية الخدمات المصغرة Micro Services Architecture 11
9.2– ممارسة التكامل المستمر والتسليم المستمر CI/CD 12
10.2– التعرف على الكيانات المسماة Named Entities Recognition 12
الفصل الثالث 13
الدراسة المرجعية 13
1.3– مقدمة 14
2.3– مسألة كشف الأحداث في دفق بياناتٍ 14
1.2.3 الإطار العام للمسألة 15
2.2.3 المراحل الرئيسية لسير العمل في نظام لكشف الأحداث 17
3.2.3 تعريف المسألة Problem Definition 18
4.2.3 طرق معالجة المسألة 18
5.2.3 المنهجيات والمقاربات لمعالجة المسألة 19
6.2.3 النموذج HCRC 22
7.2.3 معايير التقييم وقياس الأداء 23
8.2.3 مجموعات البيانات المتاحة 24
3.3– مسألة الاستدلال على الموقع الجغرافي من النص 27
1.3.3 الإطار العام للمسألة 27
3.3.3 المنهجيات والمقاربات المتبعة للتعرف على الموقع 28
4.3.3 تعريف المسألة Problem Definition 29
5.3.3 مجموعات البيانات المتاحة 30
6.3.3 معايير التقييم وقياس الأداء 30
4.3– أنظمة مشابهة 31
1.4.3 Liveuamap.com Live Universal Awareness Map 31
2.4.3 Globalincidentmap.com Global Incidents Map 32
3.4.3 rsoe-edis.org Emergency and Disaster Information Service (EDIS) . 33
5.3– خلاصة الدراسة المرجعية 34
الفصل الرابع 35
الدراسة التحليلية 35

1.4– مقدمة **36**
2.4– مخطط حالات الاستخدام **36**
3.4– السرد النصي لحالات الاستخدام **37**
4.4– العلاقات بين الفاعلين **40**
الفصل الخامس **41**
المنهجية المقترحة **41**
1.5– النهجية المقترحة **42**
1.1.5 وحدة البحث واستخراج دفق الرسائل 43
2.1.5 وحدة فلترة دفق الرسائل 43
3.1.5 وحدة كشف الأحداث 44
4.1.5 وحدة تلخيص الحدث 45
5.1.5 وحدة الاستدلال على الموقع 45
الفصل السادس **46**
تصميم النظام **46**
1.6– مقدمة **47**
2.6– مخطط خدمات النظام التصميمي **48**
3.6– تصميم خدمات النظام **49**
الفصل السابع **54**
الأدوات المستخدمة **54**
1.7– Docker **55**
2.7– ASP.NET WEB API **55**
3.7– Jenkins **55**
4.7– YARP **55**
5.7قاعدة المعطيات SQL Server **55**
6.7 JSON Web Token **56**
7.7إطار العمل Angular **56**
8.7الأداة Grafana K6 ..... **56**

الفصل الثامن **57**
التجارب ومقارنة النماذج **57**
1.8– مجموعة البيانات **58**
2.8– تجارب تصنيف التغريدات **62**
التجربة الثانية: تثقيل شعاع الكيانات المسماة 65
3.8– تجارب كشف الأحداث (نمذجة المواضيع) **70**
1.3.8 تقييم النماذج بنمط عدم الاتصال Offline Evaluation 70
2.3.8 تقييم النماذج بنمط الحي Online Evaluation 80
4.8الاستدلال على الموقع الجغرافي 93
الفصل التاسع **94**
تنجيز النظام **94**
1.9– مقدمة **95**
2.9– البنية التصميمية العامة للخدمات **95**
3.9– تنجيز الخدمات **97**
4.9– مخطط النظام التنجيزي **111**
5.9– التكامل المستمر **112**
الفصل العاشر **113**
تحليل ومناقشة النتائج **113**
1.10– مناقشة النموذج **114**
الفصل الحادي عشر **117**
اختبارات النظام **117**
1.11– اختبار خدمات النظام **118**
اختبار خدمة إدارة الأحداث 118
اختبار خدمة سحب البيانات 120
2.11– اختبارات الأداء **120**
اختبار الحمل Load Test 120
اختبار الضغط Stress Test 122
اختبار الارتفاع المفاجئ Spike Test 122

قائمة الأشكال

الشكل 1: مثال على تمرير الرسائل في شبكات البيان العصبونية. 7
الشكل 2: مثال توضيحي على التعلم المقارن. 7
الشكل 3: نموذج Skip Gram و CBOW. 9
الشكل 4: مثال على شكة معلومات غير متجانسة. 10
الشكل 5: مثال على البنية المعمارية النظيفة. 11
الشكل 6: نطاقات معالجة كشف الأحداث حسب نمطها (Schinas et al., 2018). 16
الشكل 7: تصنيف هرمي للأحداث (Saeed, Abbasi, Maqbool, et al., 2019). 16
الشكل8 :مراحل المعالجة في نظام لكشف الأحدث (Q. Li et al., 2022). 17
الشكل 9: مخطط العمل العام ضمن نظام استدلال الموقع (Afyouni, Aghbari, et al., 2022). 27
الشكل 10: أنماط لاستدلال على الموقع الجغرافي (Afyouni et al., 2022). 28
الشكل 11: تصنيف طرائق التعرف على الموقع. 28
الشكل 12: مخطط حالات الاستخدام للضيف والمستخدم المسجل. 36
الشكل 13: مخطط حالات الاستخدام للمستخدم. 36
الشكل 14: مخطط حالات الاستخدام للمراقب، ولمدير النظام. 37
الشكل 15: العلاقات بين الفاعلين. 40
الشكل 16: البينة العامة للمنهجية. 42
الشكل 17: منهجية البحث واستخراج الرسائل 43
الشكل 18: آلية التضمين المقترحة. 44
الشكل 19: عملية التدريب الدورية للنموذج. 44
الشكل 20: مثال على آلية الاستدلال على الموقع. 45
الشكل 21: مخط خدمات النظام التصميمي. 48
الشكل 22: مراحل العمل ضمن سحب تجريف البيانات. 49
الشكل 23: مخطط الصفوف لخدمة تجريف البيانات. 49
الشكل 24: توزع التغريدات خلال فترة السحب. 59
الشكل 25: تواتر التغريجات خلال ساعات اليوم. 59
الشكل 26: مجموعة الكلمات المتداولة خلال 8 أيار– الفترة 9AM – 12 AM. 59

الشكل 27: المدة الزمنية للأحداث..... 60
الشكل 28: مجال عدد التغريدات ضمن الحدث. 60
الشكل 29: توزع لغات التغريدات. 61
الشكل 30: تحليل خطاب الكراهية والعنف..... 61
الشكل 31: توزع نمط التغريدات المصنفة..... 62
الشكل 32: نتائج التصنيف لنموذج mBert في التجربة الأولى..... 63
الشكل 33: مصفوفة الارتباك للنموذج mBert في التجربة الأولى. 64
الشكل 34: مخطط مقارنة النتائج حسب معيار F1..... 64
الشكل 35: مخطط مقارنة نتاج التجربة الثانية حسب معيار F1..... 66
الشكل 36: مقارنة TF-IDF +NE مع mBert..... 66
الشكل 37: مقارنة نتائج التجربة الثالثة حسب معيار F1..... 67
الشكل 38: أدناه الفرق بين نتائج النموذج الأفضل في هذه التجربة مع mBert. 68
الشكل 39: مقارنة نتائج التجربة الرابعة..... 69
الشكل 40: مقارنة نموذج التجربة الرابعة مع mBert..... 69
الشكل 41: أثر المعامل eps على النتائج باستخدام mBert لتمثيل الرسائل. 71
الشكل 42: أثر عدد الأبعاد على نتائج العنقدة..... 71
الشكل 43: أثر معامل الكثافة على جودة العنقدة. 72
الشكل 44: نتائج التجربة الأولى..... 72
الشكل 45: أثر معامل حجم العنقود الأصغري على نتائج العنقدة. 73
الشكل 46: أثر معامل حجم العينات على النتائج. 74
الشكل 47: أثر عدد الأبعاد على نتائج العنقدة..... 74
الشكل 48: نتائج العنقدة في التجربة الثانية..... 75
الشكل 49: الفرق بين نتائج الضبط والاختبار..... 75
الشكل 50: مثال عن طريقة دمج شعاع تمثيل الكيانات المسماة. 76
الشكل 51: نتائج أخذ الكيانات المسماة في التضمين مقارنةً مع النتائج السابقة. 77
الشكل 52: الفرق بين نتاج الضبط والتقييم، في التجربة الثانية في مرحلة كشف الأحداث..... 77
الشكل 53: مقارنة النموذج المرجعي مع النموذج الذي توصلنا إليه في المرحلة السابقة. 78
الشكل 54: نتائج تحليل النموذج المرجعي إلى مكوناته. 78

الشكل 55: مقارنة النموذج المقترح مع النموذج المرجعي في الوضع Offline. 79
الشكل 56: نتاج التجربة الرابعة في مرحلة كشف الأحداث. 79
الشكل 57: أثر معامل حجم النافذة الزمنية على نتائج النموذج. 83
الشكل 58: أثر معامل حجم النافذة على النتائج. 84
الشكل 59: أسقاط تمثيل الرسائل على فضاء ثنائي الأبعاد. 84
الشكل 60: مقارنة النموذج المرجعي مع أفضل نموذج ي التجارب السابقة. 89
الشكل 61: نتائج تحليل النموذج المرجعي إلى مكوناته. 89
الشكل 62: مقارنة النموذج المقترح مع النموذج المرجعي وفق معيار NMI. 90
الشكل 63: مقارنة النموذج المقتر مع النموذج المرجعي وفق مقياس AMI. 91
الشكل 64: مقارنة النموذج المقترح مع النموذج المرجعي وفق مقياس ARI. 91
الشكل 65: أثر حجم النافذة على نتاج العنقدة. 92
الشكل 66: أثر حجم النافذة الزمنية على نتائج العنقدة. 92
الشكل 67: مثال على النبية العامة للخدمات المطورة بلغة Python. 96
الشكل 68: البنية العامة للخدمات المنجزة بإطار العمل Asp.Net. 96
الشكل 69: بنية خدمة سحب البيانات. 97
الشكل 70: بنية شريحة ساحبي البيانات. 98
الشكل 71: عينة من الرماز، تبين آلية استخدام الأدوات التي نجزناها. 99
الشكل 72: عينة من خرج هذه الخدمة. 99
الشكل 73: بنية خدمة إدارة عمليات سحب البيانات. 100
الشكل 74: بنية طبقة التطبيق في خدمة إدارة عمليات سحب البيانات. 101
الشكل 75: مخطط قاعدة المعطيات الخاص بخدمة إدارة عمليات سحب البيانات. 102
الشكل 76: بنية خدمة كشف الأحداث. 103
الشكل 77: بنية شريحة الكشف. 103
الشكل 78: بنية خدمة الاستدلال على الموقع. 104
الشكل 79: بنية خدمة إدارة الأحداث. 105
الشكل 80: بنية طبقة التطبيق 106
الشكل 81: بنية مجلدات الاختبارات. 106
الشكل 82: تصميم قاعدة المعطيات لخدمة إدارة الأحداث. 107

الشكل 83: بنية خدمة استخراج الكيانات المسماة. 108
الشكل 84: بنية خدمة الهوية. 109
الشكل 85: خطط قاعدة المعطيات لخدمة الهوية. 109
الشكل 86: بنية خدمة واجهة المستخدم. 110
الشكل 87: مكتبة Shared Kernel التي قمنا برفعها على Nuget.org. 110
الشكل 88: مخطط النظام التنجيزي. 111
الشكل 89: مثال على خطوط الأنابيب. 112
الشكل 90: مقارنة النموذج المقترح مع النموذج المرجعي وفق معيار NMI. 114
الشكل 91: مقارنة النموذج المقتر مع النموذج المرجعي وفق مقياس AMI. 115
الشكل 92: مقارنة النموذج المقترح مع النموذج المرجعي وفق مقياس ARI. 115
الشكل 93: مقارنة الزمن اللازم لمعالجة كامل البيانات والتعلم عليها، وفق النموذج المقتحر والنموذج المرجعي. 116
الشكل 94: اختبارات الوحدة والتكامل. 118
الشكل 95: اختبارات البنية المعمارية. 119
الشكل 96: تشغيل خط الأنابيب الخاص بخدمة إدارة الأحداث. 119
الشكل 97: اختبار قدرة خدمة سحب البيانات على معالجة الرسائل. 120
الشكل 98: نتائج اختبار الحمل. 121
الشكل 99: نتائج اختبار الحمل الثاني. 121
الشكل 100: نتائج اختبار الضغط. 122
الشكل 101: نتائج اختبار الارتفاع المفاجئ في الطلبات. 123

قائمة الجداول

جدول1 : المقاربات المتبعة لكشف الاحداث في وسائل التواصل الاجتماعي. 19
جدول2 : مجموعات البيانات المتاحة لمسألة كشف الأحداث. 24
جدول3 : مجموعات البيانات المستخدمة في الأبحاث المدروسة. 25
جدول4 : مجموعات البيانات المتاحة لمسألة التعرف على الموقع المذكور 30
جدول 5: السيناريو الناجح لحالة الاستخدام استعراض الأحداث 38
جدول6 : السيناريو الناجح لحالة الاستخدام إدارة مناطق الاهتمام. 39
جدول7: مثال على تغريدات ضمن حدث ما. 58
جدول8 : نتائج التجربة الأولى في مرحلة تصنيف التغريدات. 63
جدول9 : نتائج التجربة الثانية في مرحلة تصنيف التغريدات. 65
جدول10 : نتائج التجربة الثالثة. 67
جدول11 : نتائج التجربة الرابعة في مرحلة تصنيف التغريدات. 68
جدول12 : إحصائيات كتل الرسائل. 80
جدول13 : نتائج التجربة الأولى. 81
جدول14 : نتائج التجربة الثالثة. 87
جدول15 : نتائج النموذج المقترح والنموذج المرجعي وفق مقياس NMI. 92
جدول16 : نتائج المحول المضبوط امتخدم في التعرف على المواقع المذكورة. 93
جدول17 : مقاررنة النموذج المقترح مع النوذج المرجعي. 114
جدول18 : نتائج التجربة الأول في مرحلة تصنيف التغريدات. 136
جدول19 : نتائج التجربة الثانية في مرحلة تصنيف التغريدات. 137
جدول20 : النتائج التفصيلية للتجربة الثالثة في مرحلة تصنيف التغريدات. 138
جدول21 : النتائج التفصيلية للتجربة الرابعة، في مرحلة تصنيف التغريدات. 139
جدول22 : النتائج التفصيلية للتجربة الأولى في وضع عدم الاتصال، في مرحلة كشف الأحداث. 140
جدول23 : النتائج التفصيلية للتجربة الثانية في وضع عدم الاتصال، في مرحلة كشف الأحداث. 141
جدول24 : النتائج التفصيلية للتجربة الثالثة في وضع عدم الاتصال، في مرحلة كشف الأحداث. 142
جدول25 : النتائج التفصيلية للتجربة الرابعة في ووضع عدم الاتصال، في مرحلة كشف الأحداث. 143

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، 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، 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- 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 حذفها، تعديل موقعها، أهميتها، وإضافة معلومات عنها.
2. إدارة التصنيفات، أي إضافة- حذف- تعديل تصنيف.

يجب على النظام ان يسمح لمدير النظام بما يلي:

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

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 مناطق اهتمامهم.
3. استخراج 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 ضمن النظام.

4.1- 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 (100 رسالة في الدقيقة)، وفقاً للدراسة الكمية (اضغط هنا).
3. أن يكون النظام قادرً على استخراج دفق بيانات (data streams) بسعة لا تقل عن 1000 رسالة في الساعة.

ت. متطلبات التشغيل

1. أن تعمل الخدمات ضمن حاويات مستقلة لكل خدمة.

الفصل الثاني

الدراسة النظرية

يوضح هذا الفصل بعض المفاهيم المستخدمة في هذا العمل.

1.2- 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).

1.1.2 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). ومن بين الخوارزميات، تُعد HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) امتدادًا محسنًا لـ DBSCAN، حيث تتفوق في عدد من الجوانب. تعتمد HDBSCAN على بناء تسلسل هرمي من العناقيد ثم اختيار المستوى الأمثل منه بناءً على استقرار كل عنقود. تتميز بقدرتها على: (1) اكتشاف عناقيد بأشكال وأحجام متنوعة دون الحاجة إلى تحديد عددها مسبقًا، (2) التعامل بكفاءة مع البيانات غير المتجانسة والمناطق الفارغة، و(3) تقليل الحاجة إلى ضبط المعاملات يدويًا مقارنةً بـ DBSCAN، حيث تعتمد بشكل أساسي على معاملين فقط: min cluster size الذي يحدد الحجم الأدنى للعنقود، وmin samples الذي يؤثر على حساسية تحديد الضجيج. بفضل بنيتها الهرمية ومعالجتها الذكية للكثافة، تعتبر HDBSCAN مناسبة جدًا للبيانات المعقدة والعالية الأبعاد مثل البيانات النصية أو الجغرافية (Campello et al., 2013).

2.1.2- العنقدة الحية 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).

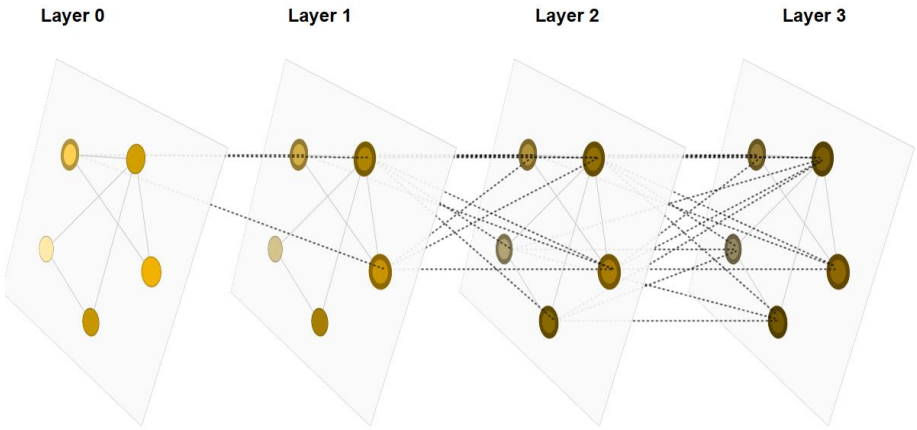
3.1.2- خوارزمية العنقدة Single Pass

تعتمد خوارزمية Single-pass Clustering على نموذج غير هرمي وغير تكراري للعنقدة، حيث يتم التعامل مع دفق البيانات النصية بشكل تسلسلي دون الرجوع للمستندات السابقة. حيث لكل مستند وارد (مثل خبر أو تغريدة)، تُحسب درجة التشابه بين تمثيله الشعاعي وكل تمثيل لعناقيد موجودة مسبقًا. وغالبًا ما يُستخدم تشابه cosine كمقياس. يتم مقارنة هذه القيم مع عتبة ثابتة (θ)، فإذا تجاوزت درجة التشابه بين المستند وأحد العناقيد هذه العتبة، يُضم المستند إلى ذلك العنقود، ويتم تحديث الخصائص الإحصائية أو التمثيل الشعاعي لذلك العنقود (مثل متوسط التمثيلات). أما إذا لم يتجاوز أي تشابه هذه العتبة، فيُعتبر أن المستند يُمثل موضوعًا جديدًا، ويُنشأ له عنقود مستقل جديد.

تُعرف هذه الخوارزمية بقدرتها على الكشف الفوري عن الأحداث الجديدة (First Story Detection) كونها لا تتطلب عمليات إعادة تدريب أو إعادة عنقدة. كما أنها مناسبة للبيئات الزمنية الحية (online settings) مثل تحليل تدفقات الأخبار أو محتوى وسائل التواصل الاجتماعي، نظرًا لأنها تعمل بزمن خطي O(n) بالنسبة لعدد المستندات وتستهلك ذاكرة محدودة نسبيًا. ومع ذلك، فإن أدائها يعتمد بشكل كبير على ضبط العتبة θ وطريقة تمثيل النصوص (Packa et al., 2003).

2.2- شبكات البيان العصبونية Graph Neural Network

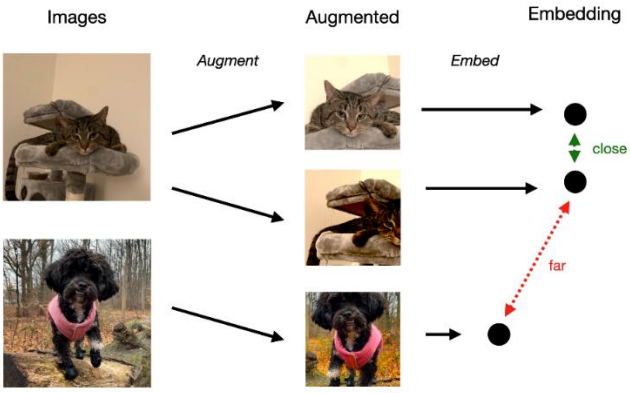
تعد شبكات البيان العصبونية GNNs، نوعاً من الشبكات العصبونية المصمم للتعامل مع بيانات مهيكلة على شكل بيان Graph، الفكرة الرئيسية هنا هي تمرير الرسائل Message Passing، حيث كل تقوم كل عقدة في البيان بتجميع المعلومات من العقدة المجاورة، وتستخدم تلك المعلومات لتحديث تمثيلها، ومن خلال تكديس عدد من طبقات GNN، تستطيع العقدة التعلم من عقد مجاورة أوسع، مما يسمح للنموذج بتعلم والتقاط معلومات وعلاقات معقدة من الهيكلية الكلية للبيان.



الشكل 1: مثال على تمرير الرسائل في شبكات البيان العصبونية.

3.2- التعلم المقارن Contrastive Learning

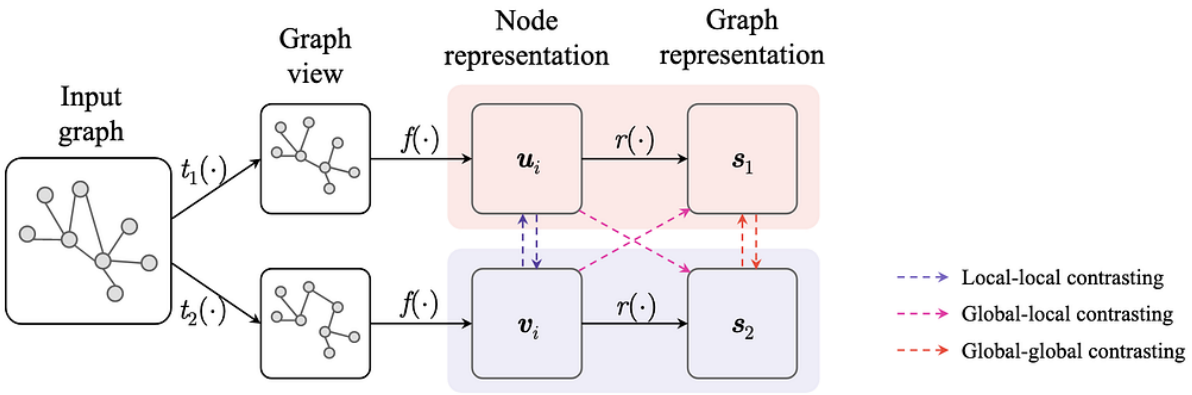
يعد التعلم المقارن نموذجاً قوياً صمن مجال التعلم المشرف عليه ذاتياً Self-Supervised Learning Paradigm، حيث يتم تدريب النموذج على التمييز بين العينات المتشابهة، عن العينات غير المتشابهة (Positive Samples, Negative Samples)، الفكرة الأساسية هنا تكمن في سحب التضمين الضمني للثنائيات الموجبة Positive Pair، بالقرب من بعضهم البعض، وبآن معاً إبقاؤهم بعيداً عن العينات السالبة. هذه الطريقة تجبر النموذج على استخراج سمات تمييزية من البيانات من دون الحاجة إلى بيانات منمطة. أي باختصار يحاول التعليم المقارن تعظيم الاتفاق بين التمثيلات الضمنية للعينات المتشابهة، تحت مجموعة بيانات معززة بطريقة مدروسة.



الشكل 2: مثال توضيحي على التعلم المقارن.

1.3.2- التعلم المقارن على البيان Graph Contrastive Learning

هو حالة خاصة من التعلم المقارن، حيث يكون على بيان، وهو غالباً ما يتكون من مكون لإنشاء وجهات نظر مختلفة للبيان، من أجل تشكيل عينات سالبة وموجبة، بالإضافة إلى هدف مقارنة لتمييز العينات الموجبة عن السالبة. ويتم تعزيز البيان (توليد وجهات نظر موجبة) من خلال إحدى الطرق التالية، إضافة وصلات بين العقد، حذف وصلات بين العقد، حذف عقد، تعمية عقد ما، إضافة خصائص للعقد.



4.2- تقنيات تضمين الكلمات Word Embedding

يعتبر تضمين الكلمات تقنية هامة في معالجة اللغات اللغات الطبيعية وتعلم الآلة، حيث تهدف هذه التقنية إلى تمثيل الكلمات بصورة رقمية (شعاعية) قابلة للمعالجة من قبل النماذج الحاسوبية. حيث يتم تحويل الكلمات إلى أشعة متعددة الأبعاد تعكس المعاني والعلاقات الدلالية والنحوية فيما بينها. حيث تعتمد تقنيات تضمين الكام على فكرة أساسية وهي أن الكلمات ذات المعاني المتشابهة يكون لها تمثيلات شعاعية متقاربة في فضاء التضمين، وبالمثل الكلمات ذات المعاني المختلفة تكون بعيدة في فضاء التضمين. وتتم عملية تضمين الكلمات باستخدام نماذج التعلم العميق، ويتم تدريب هذه النماذج على مجموعة كبير من المدونات، وتنقسم نماذج التضمين إلى نوعين، ثابتة _تعطي شعاع تضمين للكلمة بغض النظر عن سياقها، سياقية _تعطي شعاع تضمين حسب سياق الجملة حيث يتغير شعاع التضمين للكلمة حسب سياقها.

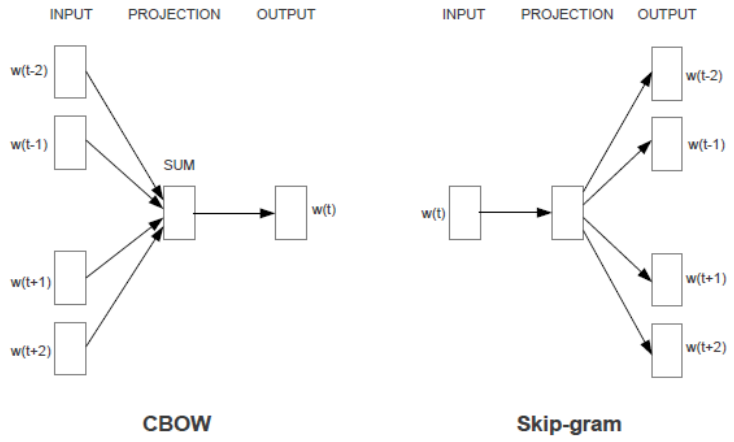
1.4.2 المحولات Transformers

تُعد تقنية المحولات (Transformers) من الابتكارات الرائدة في مجال معالجة اللغة الطبيعية (NLP)، وقد تم تقديمها لأول مرة في عام 2017. وتُعتبر هذه التقنية بمثابة ثورة في هذا المجال نظرًا لقدرتها الفائقة على فهم النصوص وتوليدها بشكل فعّال ودقيق. تعتمد المحولات على نموذج تعلم عميق يتألف من عدة طبقات من الشبكات العصبية، ويتم تدريبه على مجموعات

ضخمة من البيانات النصية بهدف تمكينه من فهم البنية اللغوية وتوليد نصوص ذات معنى. ما يميز هذه التقنية هو قدرتها على التعامل مع السياقات الطويلة في النصوص، حيث تستطيع فهم العلاقات بين الكلمات، والجُمل، والفقرات بشكل متكامل. كما تعتمد المحولات بشكل أساسي على آلية الانتباه (Attention Mechanism)، التي تُتيح للنموذج تركيز اهتمامه على الأجزاء الأهم في النص أثناء المعالجة (Vaswani et al., 2023).

Word2Vec 2.4.2

يعد هذا النموذج الأول من نوعه في هذا المجال، حيث تم إصداره من Google عام 2013، وهو نموذج تضمين ثابت، يمكنه التعلم على مدونات ضخمة غير منمطة، عن طريق إحدى مهمتين، الأولى Continuous Back of Words، حيث يتم تدريب النموذج على التنبؤ بالكلمة بناءً على الكلمات المحيطة بها، والأخرى هي Skip Gram Model، بحيث يتم التدريب على التنبؤ بالكلمات المحيطة بالكلمة الحالية، والشكل أدناه يوضح ذلك (Mikolov et al., 2013).



الشكل 3: نموذج **Skip Gram** و **CBOW**.

FastText 3.4.2

هو نموذج تضمين ثابت مشابه لـ Word2Vec، تم إصداره من قبل Facebook عام 2017، وما يتميز به هذا النموذج هو قدرته على فهم الكلمات حتى لو كان بها أخطاء إملائية أو نحوية، بالإضافة إلى فهم اللهجات، وذلك لأنه يأخذ المحارف في عين الاعتبار عند تعلم التمثيل character n-grams. وكما في Word2Vec يتم تدريب النموذج بأحد الطريقتين Skip Gram، CBOW (Bojanowski et al., 2017).

4.4.2 نموذج Bert (Bidirectional Encode Representation from Transformers)

يعتبر واحد من أحدث نماذج التضمين الدلالي، ويعتمد في بنيته على المحولات (Transformers)، ما يميز هذا النموذج عن غيره من النماذج التي سبقته هو قدرته على توليد أشعة تضمين تأخذ السياق بعين الاعتبار، أي تضمين الكلمة يختلف حسب

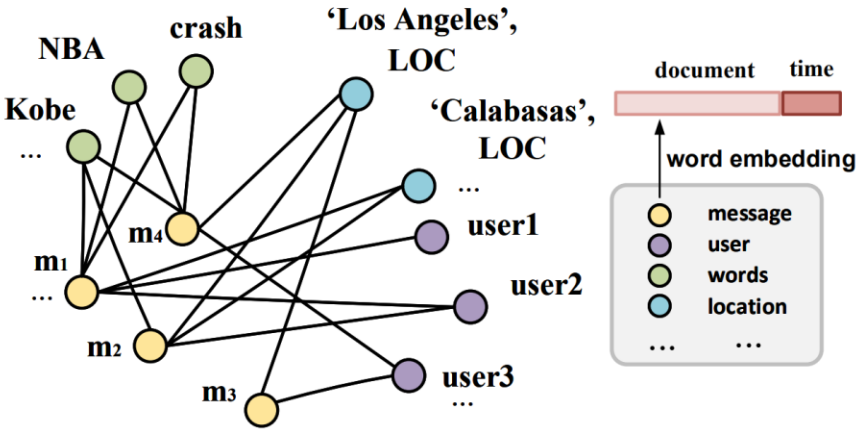
سياقها. حيث يتم تدريب هذا النموذج على مهمتين هما، مهمة تعبئة الفراغات MLM (Masked Language Model) بحيث يتم استبدال مجموعة من كلمات الجملة برمز خاص Mask ويكون هدف النموذج هو التنبؤ بهذه الكلمات، والثاني هو التنبؤ بالجملة التالية NSP (Next Sentence Prediction) والدخل هنا هو جملتين والخرج هو 0 أو 1 إذا كانت الجملة الثانية هي الجملة التالية أم لا (Devlin et al., 2019).

5.2- LDA (Latent Dirichlet Allocation)

LDA، هو خوارزمية شائعة في معالجة اللغات الطبيعية من أجل نمذجة المواضيع، وهو عبارة عن نموذج إحصائي يساعد على اكتشاف المواضيع ضمن مجموعة من المستندات، بحيث كل مستند هو عبارة عن مجموعة من المواضيع، وكل موضوع هو عبارة عن مجموعة من الكلمات. وبشكل أساسي يهدف LDA إلى الكشف عن الهياكل الموضوعية الضمنية ضمن مدونة نصية.

6.2- شبكات المعلومات غير المتجانسة (HIN)

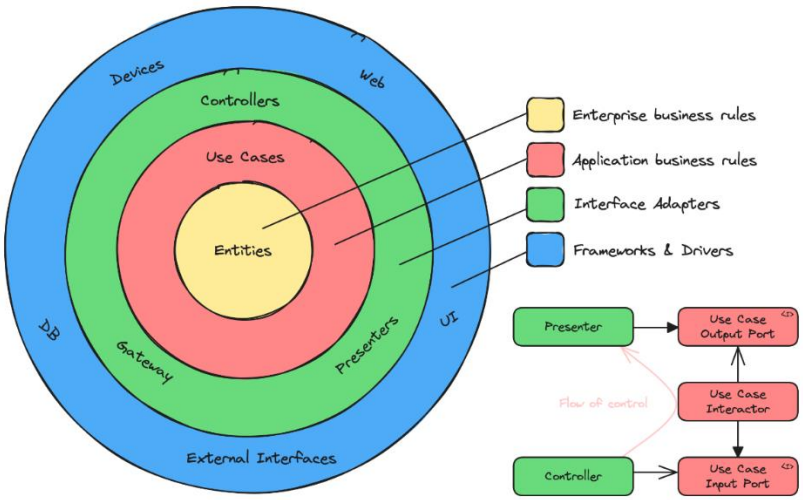
شبكات المعلومات غير المتجانسة Heterogeneous Information Networks (HIN)، وهي حالة خاصة من البيان Graph، حيث يحتوي هذا على أكثر من نوع من العقد والوصلات. أي كما يلي، $G = (V, E, \mathcal{N}, T)$ حيث V و E هما مجموعات العقد و الوصلات و $\mathcal{N}$ و T هما مجموعات أنواع العقد والوصلات، حيث لكل عقدة نوع 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.



الشكل 4: مثال على شكة معلومات غير متجانسة.

7.2- البنية المعمارية النظيفة

تُعد البنية النظيفة (Clean Architecture) فلسفة تصميم برمجي تهدف إلى بناء أنظمة قابلة للتوسع، سهلة الصيانة، وقابلة للاختبار. تستند هذه البنية إلى مبدأ فصل مكونات النظام إلى طبقات مستقلة، بحيث يكون لكل مكون مسؤولية واضحة ومعزولة عن تفاصيل التنفيذ، مثل قاعدة البيانات أو واجهة المستخدم. تُركّز البنية النظيفة على عزل منطق الأعمال (Business Logic) عن الأطر الخارجية (مثل قواعد البيانات، الشبكة، أو واجهات العرض)، مما يتيح تطوير كل مكون أو تغييره دون التأثير على باقي النظام. يؤدي هذا النهج إلى تقليل التعقيد العام للنظام، ويعزز من مرونته وقدرته على التكيّف مع التغيّرات المستقبلية. ومن أبرز خصائص البنية النظيفة: (1) أن منطق التطبيق مستقل عن إطار العمل المستخدم، (2) أن آلية تخزين البيانات يمكن تغييرها دون المساس بمنطق الأعمال، و(3) أن النظام يكون أكثر قابلية للاختبار والتطوير المستمر



الشكل 5: مثال على البنية المعمارية النظيفة.

8.2- معمارية الخدمات المصغرة Micro services Architecture

تمثل الخدمات المصغرة أسلوباً معمارياً حديثاً يحلل التطبيقات التقليدية ذات الكتلة الواحدة Monolithic Application، إلى مجموعة من الخدمات المصغرة الموزعة القابلة للنشر بشكل مستقل عن بعضها البعض. يعزز هذا النهج المرونة في بناء التطبيقات وقابلية التوسع ويزيد من إنتاجية المطورين، من خلال السماح للفرق بالعمل بشكل مستقل. وتتميز الخدمات المصغرة بتماسكها الداخلي العالي (High Cohesion) وارتباطها الخفيف (Low Coupling) مما يسهل تطوير التطبيقات المعقدة.

9.2- ممارسة التكامل المستمر والتسليم المستمر CI/CD

يُعد التكامل المستمر (CI) والنشر المستمر (CD) من الممارسات الأساسية في تطوير البرمجيات الحديثة، لاسيّما ضمن بيئات وأطر عمل DevOps. تهدف تقنيات CI/CD إلى أتمتة عمليات اختبار البرمجيات ونشرها، مما يتيح للفرق التقنية إمكانية تقديم التحديثات بسرعة وكفاءة بعد حدوث أي تعديل في الشيفرة المصدرية (Source Code). وتُساهم هذه الممارسات في تقليل الأخطاء البشرية، وتحسين جودة المنتج، وتسريع دورة التطوير والنشر.

10.2- 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.

الفصل الثالث

الدراسة المرجعية

نبيّن في هذا الفصل الأبحاث والأعمال والمنهجيات المشابهة لهذا العمل.

1.3- 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. وإضافةً 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، استخراج الموقع الجغرافي من النصوص، بالإضافة إلى استعراض الأنظمة المشابهة ومجموعات البيانات المتاحة لدعم هذه المهام.

2.3- 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 أثر 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).

1.2.3 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 الـ 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، لا عِلمَ 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).

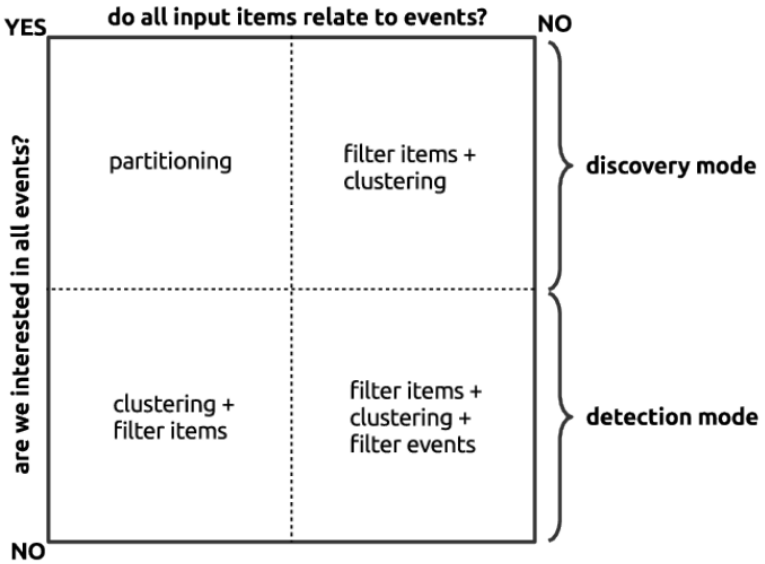
أما بالنسبة للتصنيف الثاني:

- **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 تركز على الأحداث غير المحددة. وعندما تتضمن المهمة المرادة أحداثاً محددة أو معلومات محددة

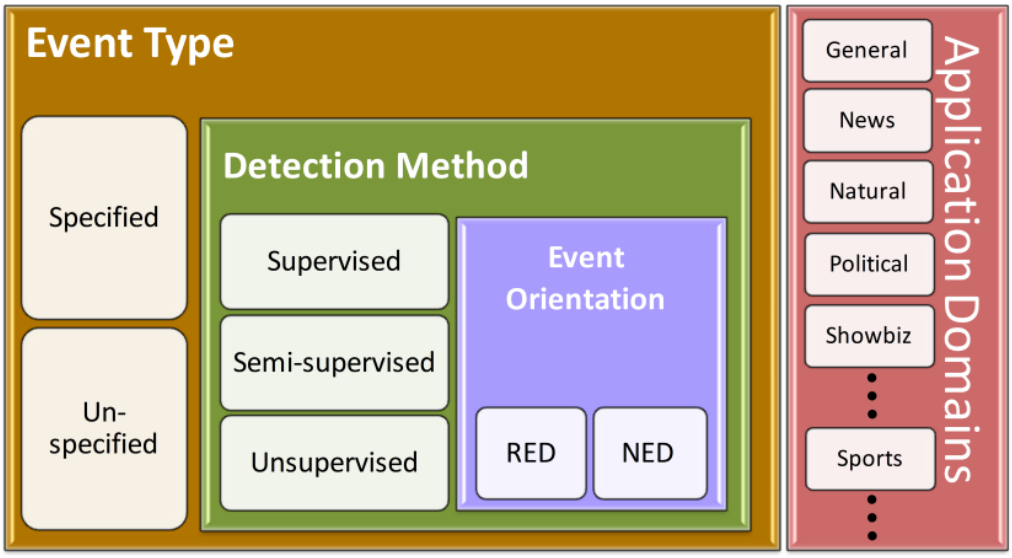
حول الحدث (عل سبيل المثال، منظمة محددة، شخص أو مكان ما) يمكن الاستفادة من هذه المعلومات لمكاملتها مع تقنيات الـ NED من خلال تقنيات الفلترة أو التصنيف.

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

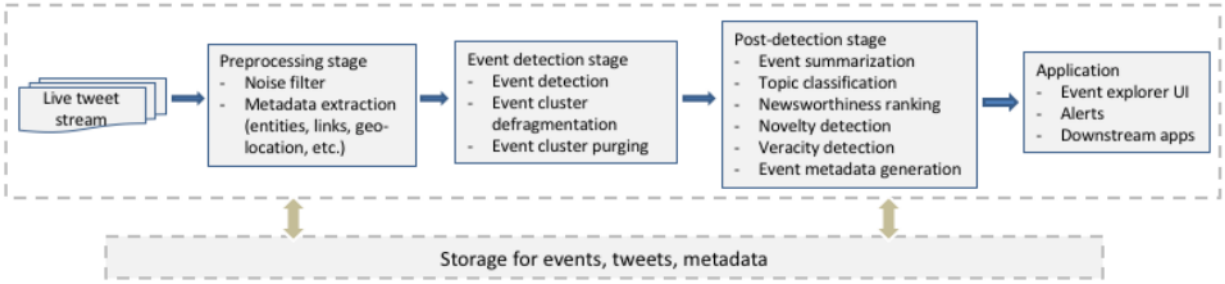
ويختصر الشكل أدناه مفردات المسألة ويصنفها بشكل هرمي يوضحها.



الشكل 7: تصنيف هرمي للأحداث **(Saeed, Abbasi, Maqbool, et al., 2019)**.

2.2.3 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 ما بعد كشف الحدث.



الشكل**8** :مراحل المعالجة في نظام لكشف الأحدث **(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 وما شابه. 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 ما بعد 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.

3.2.3 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 كـ $\mathfrak{S} = M_0, \dots, M_{i-1}, M_i, \dots$ 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).

4.2.3 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 شبكات البيان العصبونية (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 أشعة التضمين، على سبيل المثال، وظفت أبحاث مثل *(KPGNN - Cao et al., 2021)*، (QSGNN - Ren et al., 2022) *(HCRC - Guo et al., 2024)* 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 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.

5.2.3 المنهجيات والمقاربات لمعالجة المسألة

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

جدول**1** : المقاربات المتبعة لكشف الاحداث في وسائل التواصل الاجتماعي.

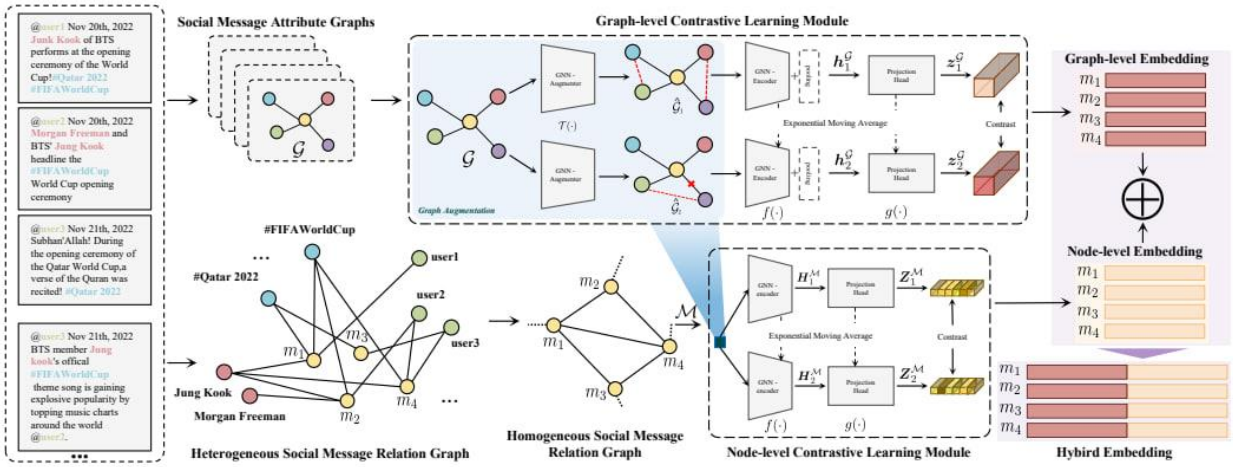
#	Paper (Name)	Techniques	Reference	Supervision	Event Type	Detection Task
1	Petrovi´c et al., 2010	Online Clustering & LSH	(Petrovi´c et al., 2010)	Unsupervised	U	NED
2	Sakaki et al., 2010	SVM	(Sakaki et al., 2010)	Supervised	S	RED
3	Long et al., 2011	Hierarchical Clustering	(Long et al., 2011)	Unsupervised	U	NED
4	Cordeiro, 2012	Wavelet Analysis & LDA	(Cordeiro, 2012)	Unsupervised	U	NED
5	McCreadie et al., 2013	K-Mean Clustering & LSH	(McCreadie et al., 2013)	Unsupervised	U	NED
6	Corney et al., 2014	Online Clustering	(Corney et al., 2014)	Unsupervised	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	RED
9	Guille & Favre, 2015	Mention, Anomaly Detection	(Guille & Favre, 2015)	Unsupervised	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

12	Hasan et al., 2016 (TwitterNews)	Incremental Clustering	(Hasan et al., 2016)	Unsupervised	U	NED
13	Chen et al., 2017	NN, Online Clustering	(Chen et al., 2017)	Unsupervised	U	NED
14	Zhang et al., 2017	Spatial-Temporal Clustering	(S. Zhang et al., 2017)	Unsupervised	U	NED
15	Edouard, 2018	Domain Vocabulary and Knowledge Based	(Edouard, 2018)	Supervised	S	NED
16	Comito et al., 2019	Text & Temporal Feature Clustering	(Comito et al., 2019)	Unsupervised	U	NED
17	Hossny et al., 2019	Keyword Volume Approach, SVM	(Hossny & Mitchell, 2019)	Supervised	S	NED
18	Fedoryszak et al., 2019	Cluster Chain, Graph	(Fedoryszak et al., 2019)	Unsupervised	U	NED
19	Nguyen et al., 2019	Entity Clustering	(Nguyen et al., 2019)	Unsupervised	U	NED
20	Saeed et al., 2019	Graph, Clustering	(Saeed, Abbasi, Razzak, et al., 2019)	Unsupervised	U	NED
21	Morabia et al, 2019 (SEDTWik)	Segments, Hierarchical Clustering	(Morabia et al., n.d.)	Unsupervised	U	NED
22	Liu et al., 2020 (EventX)	Community Detection	(Liu et al., 2020)	Unsupervised	U	NED
23	Cao et al., 2020 (IED KCN)	Knowledge Consolidation Network (KCN)	(P. Cao et al., 2020)	Supervised	-	NED
24	Li and Zhang, 2021	Semantic Term, GCN, Clustering	(Q. Li & Zhang, 2021)	Supervised	U	NED
25	Ghaemi et al., 2021	Spatial-Temporal Clustering, VDBSCAN	(Ghaemi & Farnaghi, 2023)	Unsupervised	U	NED
26	Cao et al., 2021 (KPGNN)	GNN, Clustering	(Y. Cao et al., 2021)	Supervised	-	NED
27	George et al., 2021	Spatial-Temporal Clustering	(George et al., 2021)	Unsupervised	U	NED
28	Mojiri et al., 2021	Burst Keywords, Time Series Analysis	(Mojiri et al., 2021)	Unsupervised	U	NED

29	Yamani et al., 2021	Knowledge Based, MLM	(Yamani et al., n.d.)	Supervised	-	-
30	Peng et al., 2022 (FineEvent)	GNNs	(Peng et al., 2023)	Unsupervised	U	NED
31	Hettiarachchi et al., 2022 (Embed2Detect)	NN, hierarchical clustering	(Hettiarachchi et al., 2022)	Unsupervised	U	NED
32	Afyouni Khan et al., 2022	Spatial-Temporal Clustering, Classification	(Afyouni, Khan, et al., 2022)	-	U	NED
33	Ren et al., 2022 (QSGNN)	GNNs, Online Clustering	(Ren et al., 2022)	Supervised	-	-
34	Bok et al.,2023	Graph, Online Clustering	(Bok et al., 2023)	Unsupervised	U	NED
35	Cao et al., 2024 (HISEvent)	GNNs, Graph Clustering	(Y. Cao et al., 2023)	Unsupervised	U	NED
36	Yue et al., 2023	MLM	(Yue et al., 2023)	Supervised	-	-
37	Pradhan et al., 2024 (EDTBert)	Graph Clustering, PLMs	(Pradhan et al., 2024)	Unsupervised	U	NED
38	Guo et al., 2024 (HCRC)	GNNs, Incremental Clustering	(Guo et al., 2024)	Unsupervised	U	NED
39	Ray et al., 2024	Incremental Clustering	(Ray et al., 2024)	Unsupervised	U	NED
40	P. Li et al., 2024 (RPLMSED)	PLMs	(P. Li et al., 2024)	Supervised	S	NED
41	Abagissa et al., 2024 (DistilBert – GNN)	GNNs, Clustering	(Abagissa et al., 2024)	Unsupervised	U	NED
42	Yang et al., 2024 (ADPSEM)	Structural Entropy Minimization	(Yang et al., 2024)	Unsupervised	-	-
43	Yu et al., 2024 (HyperSED)	Hyperbolic Embedding, Graph Partitioning	(Yu et al., 2024)	Unsupervised	U	NED

6.2.3 النموذج HCRC

تم في هذه الورقة اقترح طريقة جديدة للكشف عن الأحداث بطريقة غير خاضعة للإشراف من خلال التعلم المقارن للبيان الهجين (Hybrid Graph Contrastive Learning)، حيث يتم استخدام التعلم المقارن على البيان لتعلم معلومات تمييزية للرسائل. ويوضح الشكل أدناه البينة العامة للنموذج، حيث يتم تعلم التضمين على مستويين، مستوى البيان، ومستوى العقدة.



ويتم ذللك وفق التالي، عندما تصل كتلة رسائل M_i، يتم تنفيذ مسار العمل الموضح بالشكل أدناه.

- عند وصول كتلة الرسائل، يتم إنشاء بيانين، (1) بيان السمات Social Messages Attribute Graph، بحيث يتم إنشاء بيان لكل رسالة في كتلة الرسائل، ويبنى هذا البيان كعقدة مركزية تمثل الرسالة وتربط بالكيانات المسماة الخاصة بها (مثل أسماء المواقع، أسماء الأشخاص، .... إلخ) على شكل بيان نجمي، ويكرز هذا البيان على المعلومات الخاصة بكل رسالة. (2) بيان العلاقات Social Messages Relation Graph، حيث تحول الرسائل وعلاقاتها إلى بيان معلومات غير متجانس (HIN)، ومن ثم يتم تحويله إلى بيان متجانس يحوي على الرسائل فقط كعقد، وترسم الوصلات بين الرسائل من خلال الكيانات المسماة المشتركة بينها.

- التعلم المقارن: لتعلم تضمينات فعالة من هذه التمثيلات، يطبق التعلم المقارن على مستويين –مستوى البيان ومستوى العقدة – حيث في مستوى البيان، يتم تعزيز على البيان G لتوليد نسختين محسنات عبر تقنيات تعزيز مثل (حذف وصلات، إضافة وصلات) حيث يتم تمرير هاتين النسختين عبر شبكة GNN لاستخراج التمثيلات الكلية، ثم تسقط هذه التمثيلات عبر رأس إسقاط غير خطي Nonlinear Projection Head (.)g، لتنتج التمثيلات الضمنية لهما. وفي مستوى العقدة، يطبق على البيان M (بيان العلاقات)، التعزيز بإضافة وحذف وصلات، لينتج نسختين منه M1, M2، ويتم استخدام GNN لاستخراج تمثيلات العقد، ويتم إسقاطها إلى فضاء ضمني.

- ثم يطبق هدف التعلم المقارن Contrastive Loss، بين الأزواج الموجبة (تمثيل نفس الرسالة بين نسختين)، والأزواج السالبة (تمثيل الرسائل المختلفة)، وتوضح العبارة أدناه تابع الخسارة المستخدم.

$$L(u_k, v_k) = \log\left(\frac{e^{\frac{\theta(u_k,v_k)}{\tau}}}{e^{\frac{\theta(u_k,v_k)}{\tau}} + \sum_{m\neq k} e^{\frac{\theta(u_k,v_m)}{\tau}} + \sum_{m\neq k} e^{\frac{\theta(u_k,u_m)}{\tau}}}\right)$$

حيث:

u_k, v_k هو تمثيل نفس الرسالة في النسختين.

$\theta(S(g(u), g(v))$ بحيث S هي تابع التشابه Cosine Similarity.

g(.) وهو إسقاط غير خطي من طبقتي Fully Connected.

ويتم تعظيم المتوسط عبر جميع الأزواج

$$L_{contrastive} = \frac{1}{2N}\sum_{m\neq k} l(u_k, v_k) + l(v_k, u_k)$$

- بعد ذلك يتم التضمين الهجين Hyprid Embedding، ممن خلال دمج Concatenate التمثيل الناتج عن مستوى العقدة، ومع تثميل مستوى البيان.
- وبعد ذلك يتم عنقدة الرسائل باستخدام خوارزمية العنقدة Single Pass.

7.2.3 معايير التقييم وقياس الأداء

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

المعلومات المتبادل الممعيرة، يقيس مقدار المعلومات المشتركة بين المجموعات الحقيقية (Ground Truth) والمجموعات المكتشفة (Predicted Clusters). يتم تقييسه باستخدام الإنتروبي ليصبح المقياس بين 0 و1. 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. Rand Index (RI)

بهو مقياس يستخدم لتقييم جودة العنقدة من خلال مقارنة العنقدة الناتجة مع التجميع الحقيقي (أو التجميع المرجعي ground truth). الفكرة الأساسية هي حساب مدى الاتفاق بين التجميعين في تصنيف الأزواج من العناصر.

ث. Adjusted Rand Index (ARI)

مقياس معدل من Rand Index يأخذ في الحسبان احتمال الاتفاق العشوائي بين العناقيد. بخلاف RI، يقوم ARI بتصحيح القيمة لتجنب التضليل الناتج عن الاتفاق الذي قد يحدث بالصدفة، مما يجعله أكثر موثوقية عند مقارنة خوارزميات التجميع، خاصةً عندما يكون عدد المجموعات كبيرًا أو التوزيع غير متوازن. ويحسب من خلال العلاقة المبينة أدناه.

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

8.2.3 مجموعات البيانات المتاحة

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

جدول**2** : مجموعات البيانات المتاحة لمسألة كشف الأحداث.

	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	-

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

جدول**3** : مجموعات البيانات المستخدمة في الأبحاث المدروسة.

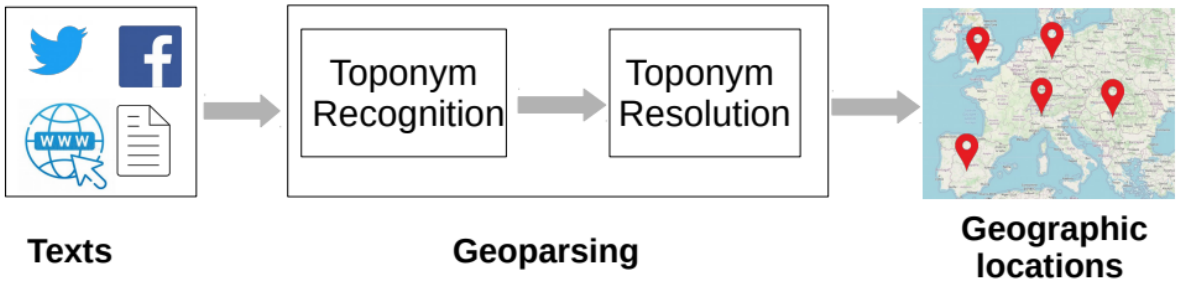
	Paper | Dataset	Event 2012	Event 2018	Maven	Few Events	Crisis Lex	Kawarith	FA Cup	Other	Private
1	Petrovi´c et al., 2010								Y	
2	Sakaki et al., 2010									Y
3	Long et al., 2011									Y
4	Cordeiro, 2012									Y
5	McCreadie et al., 2013									Y
6	Corney et al., 2014									Y
7	Alsaedi et al. , 2015									Y
8	Patil et al., 2015									Y
9	Zhang et al. , 2017									Y
10	Hossny et al. , 2019									Y
11	Saeed et al. , 2019							Y	Y	
12	Fedoryszak et al. ,2019	Y								
13	Morabia et al. ,2019	Y								
14	Zhang et al.,2021									Y
15	Hettiarachchi et al. ,2021							Y		
16	Ghaemi et al. ,2021									Y
17	Mojiri et al. ,2021							Y	Y	
18	Yamani et al. ,2021								Y	
19	Cao et al. ,2021	Y		Y						

	Paper | Dataset	Event 2012	Event 2018	Maven	Few Events	Crisis Lex	Kawarith	FA Cup	Other	Private
20	Peng et al. ,2021								Y	
21	Afyouni et al. ,2022									Y
22	Ren et al. ,2022	Y	Y							
23	Kolajo et al. ,2022									Y
24	Peng et al. ,2022									Y
25	Bok et al. ,2023									Y
26	Cao et al. ,2023	Y	Y							
27	Ren et al. ,2023 a	Y	Y				Y			
28	Ren et al. ,2023 b	Y	Y			Y				
29	Guoa et al. ,2023			Y						
30	Yue et al. ,2023			Y	Y					
31	Yang et al. ,2024	Y	Y							
32	Qiu et al. ,2024	Y				Y	Y			
33	Ray et al. ,2024	Y	Y							
34	Yu et al. ,2024	Y	Y				Y		Y	
36	Abagissa et al. ,2024	Y								
37	Pradhana et al. ,2024							Y	Y	
38	Ma et al. ,2024	Y	Y		Y	Y	Y			
39	Li et al. ,2024	Y	Y				Y			
40	Kovalchuk et al. ,2024 a									Y
41	Kovalchuk et al. ,2024 b	Y							Y	
42	Kargupta et al. ,2024									Y

3.3- مسألة الاستدلال على الموقع الجغرافي من النص

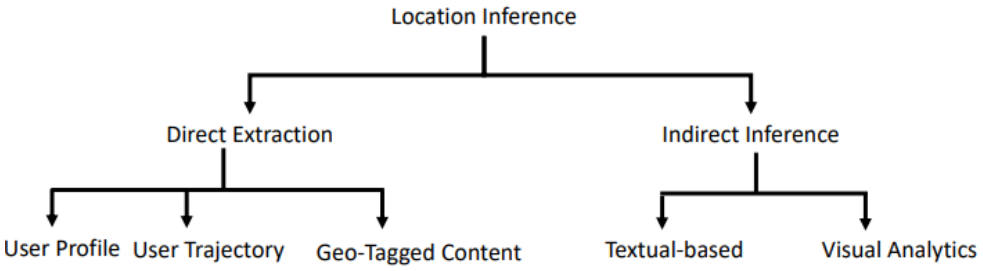
1.3.3 الإطار العام للمسألة

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.



الشكل 9: مخطط العمل العام ضمن نظام استدلال الموقع (**Afyouni, Aghbari, et al., 2022**).

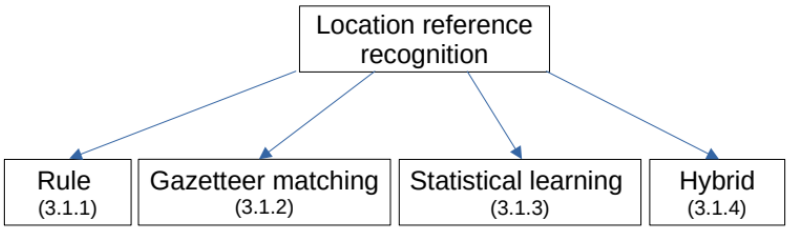
كما ويصنف الباحثون مصادر المعلومات المستخدمة للاستدلال على الموقع، إلى فرعين، الأول استخراج مباشر يتم عن طريق استخراج المعلومات المكانية الموسومة من ملف المستخدم الشخصي أو عن طريق المعلومات المكانية الموسومة مع النص المنشور حيث تقدم بعض وسائل التواصل مثل تويتر وسوم مكانية في حال فعل المستخدم ال GPS، 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 10: أنماط لاستدلال على الموقع الجغرافي (**Afyouni et al.**, **2022**).

3.3.3 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 11: 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 (Chen et al. ،Bobadilla et al., Cadorel 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.

4.3.3 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).

5.3.3 مجموعات البيانات المتاحة

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

جدول**4** : مجموعات البيانات المتاحة لمسألة التعرف على الموقع المذكور

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

6.3.3 معايير التقييم وقياس الأداء

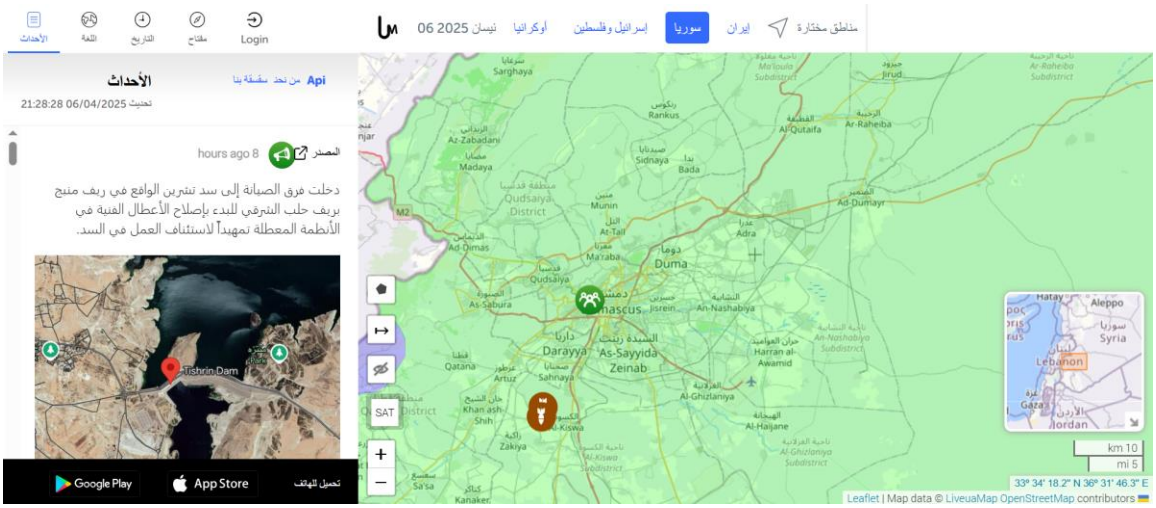
من أجل تقييم النتائج لهذه المسألة وقياس أداء النماذج يوجد العديد من معايير التقييم، التي تختلف حسب الهدف من النموذج فمن أجل النماذج التي تهدف إلى التنبؤ بـ (المنطقة، المدينة، ...) فيتم في هذه الحالة استخدام المعايير التقليدية كالصحة، الدقة والاستذكار (Accuracy, Precision, Recall)، أي يتم التعامل معا 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.

4.3- 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.

1.4.3 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، حيث يسمح للمستخدم بتحديد نوع الأحداث التي يريد الاطلاع عليها (كوارث طبيعية، نزاعات، ...) ومن ثم يعرض عليها بشكل حي هذه الأحداث. 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؟

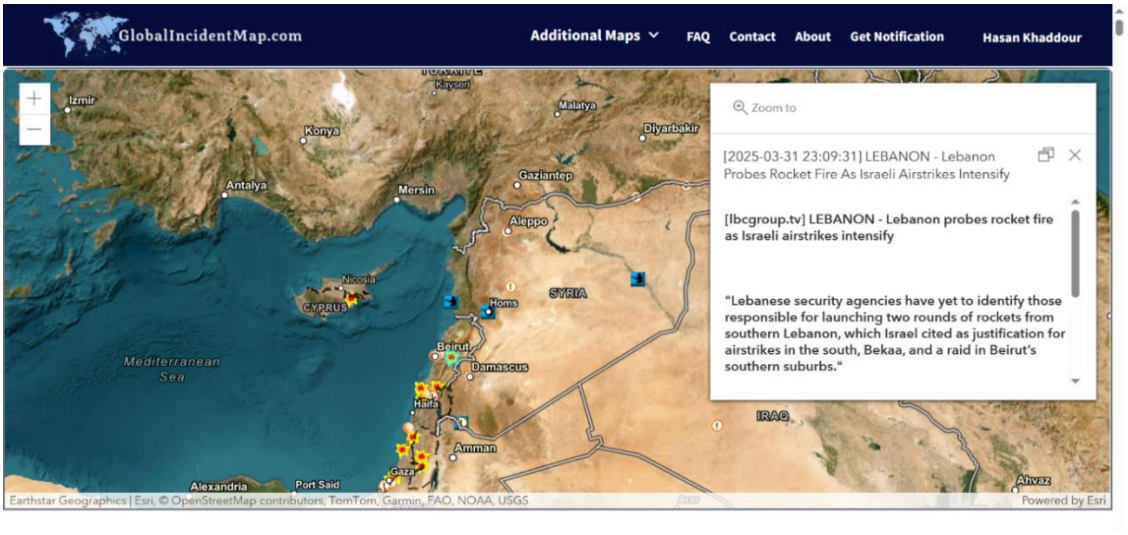
- يعتمد على مصادر متعددة لجمع البيانات مثل وسائل التواصل الاجتماعي، وكالات الأنباء، والمصادر المحلية.
- تستخدم المنصة خوارزميات آلية ومحررين بشريين للتحقق من صحة الأخبار وربطها بالموقع الجغرافي المناسب.
- يتم تصنيف الأحداث إلى أنواع متعددة (مثل: نزاع، سياسة، طقس، صحة، كوارث، إلخ) باستخدام رموز مختلفة.

♣ 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.

2.4.3 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 1: 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، كوارث طبيعية، الخ).
- يمكن للمستخدمين تصفح الحوادث حسب الفئة، التاريخ، أو الموقع الجغرافي.
- يوجد عدة أنماط للموقع أحدها هو الحوادث، كما يوجد أنماط منها الكوارث الطبيعية، أزمات الغذاء، إلخ.

♣ كيف يجمع البيانات؟

- يعتمد على مزيج من المصادر الإخبارية الرسمية والمصادر المفتوحة مثل تقارير الشرطة، وكالات الأنباء العالمية.
- 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.

3.4.3 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 2: الصفحة الرئيسية للموقع **rsoe-edis.org**

♣ كيف يعمل الموقع؟

- يقوم بجمع البيانات من جهات رسمية ومنظمات إنسانية ووسائل إعلام موثوقة. حيث يجمع الموقع البيانات من مجموعة كبيرة من المصادر مثل وكالات حكومية، وسائل الإعلام الرسمية، منظمات دولية شبكات المتطوعين
- يوفر تقارير تفصيلية حول أنواع مختلفة من الأحداث مثل الكوارث الطبيعية، التهديدات المتعلقة بالبنية التحتية، الحوادث البيولوجية

5.3- خلاصة الدراسة المرجعية

سنعرض هنا أهم النتائج التي توصلنا إليها من خلال دراستنا المرجعية:

- تعاني الطرائق المعتمدة على السمة Feature-Pivot، من أنها لا تستطيع كشف الأحداث قبل أن تصبح شائعة Trend، كما أنها غير قادرة على كشف الأحداث ذات الأثر الصغير.
- تكمن المشكلة الرئيسية في الطرائق المعتمدة على المستند Document-Pivot، في استخلاص معلومات تمييزية من الرسائل Extract Discriminative Information.
- لا يوجد مجموعة بيانات للتقييم والاختبار باللغة العربية مشابهة لمثيلاتها باللغات الأجنبية، حيث أن المجموعة الوحيدة المتوفرة هي KAWARITH تحوي على سبعة أحداث فقط وهي كوارث طبيعية. على عكس مجموعة البيانات EVENT 2012 باللغة الإنجليزية وتحوي على 506 أحداث ذات أثر متفاوت.
- معظم الأبحاث ركزت على مكون ما من النظام دون غيره، حيث يوجد ورقتين فقط ناقشتا النظام بكل مكوناته.
- استخدمت معظم الأبحاث في المقاربة الثانية Document-Pivot، خوارزمية العنقدة DBSCAN، وSinglePass
- تعاني خوارزمية العنقدة SinglePass، من حساسيتها الشديدة للضجيج، حيث على عكس خوارزميات العنقدة القائمة على الكثافة، فهذه الخوارزمية لا تستطيع إزالة الضجيج.

الفصل الرابع

الدراسة التحليلية

يقدم هذا الفصل تحليلاً للمتطلبات التي أوردناها في الفصل الأول.

1.4- مقدمة

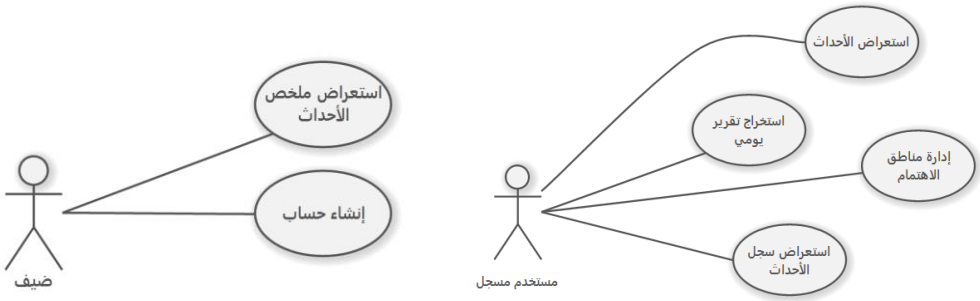
نقدم في هذا الفصل دراسة تحليلية للمسألة المطروحة، ونبين حالات الاستخدام ووصفها السردي وبعض المخططات الداعمة.

إنّ الفاعلين في هذا النظام هم على النحو الآتي:

- الضيف، وهو المشخص الذي يهتم فقط بملخص الأحداث، وليس مسجلاً في النظام.
- المستخدم المسجل، وهو الشخص الذي يهتم بالأحداث وتفاصيلها، وهو مسجل لدى النظام.
- المراقب، وهو الشخص الذي مهمته هي إدارة الأحداث ومراجعتها والتأكد منها.
- مدير النظام، وهو مسؤول عن إدارة مصادر البيانات.

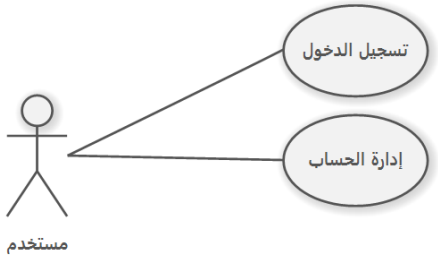
2.4- مخطط حالات الاستخدام

أ- حالات الاستخدام الخاصة بالمستخدم المسجل والضيف



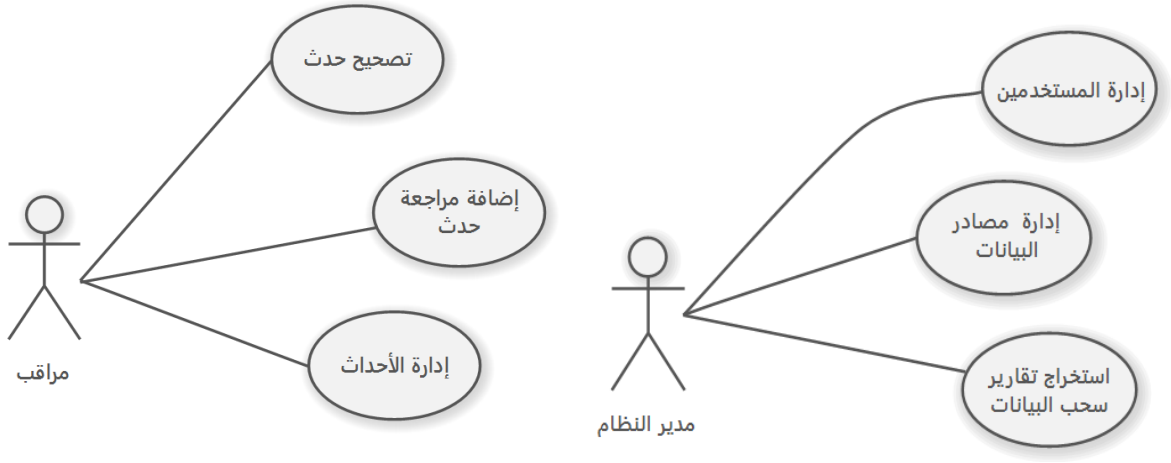
الشكل 12: مخطط حالات الاستخدام للضيف والمستخدم المسجل.

ب- حالات الاستخدام الخاصة بالمستخدم



الشكل 13: مخطط حالات الاستخدام للمستخدم

ت- حالات الاستخدام الخاصة بالمراقب، ومدي النظام



الشكل 14: مخطط حالات الاستخدام للمراقب، ولمدير النظام.

3.4- السرد النصي لحالات الاستخدام

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

أ. 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).	
	3. يختار نمط العرض المراد.
4. يعيد مجموعة الأحداث اليومية وفق النمط 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 (ملخص، عنوان، ...)	

المسارات الخاطئة

لا يوجد.

ب. 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:

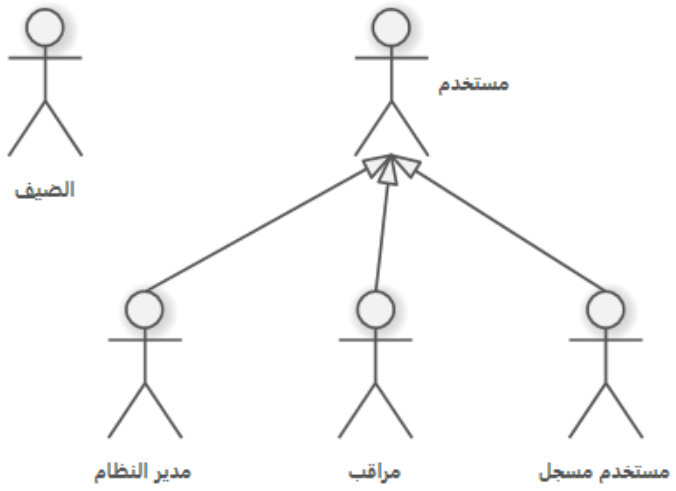
	1. يختار بدء عملية إضافة منطقة اهتمام
2. يعيد خريطة المناطق، والنطاقات الممكنة.	
	3. يختار المنطقة (الإحداثيات، ونصف القطر)
4. يضيف النظام المنطق الجديدة إلى قائمة المناطق.	
	5. ينهي عملية إضافة المنطقة.

A2 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.

4.4- العلاقات بين الفاعلين

يوضح المخطط أدناه العلاقات بين الفاعلين في هذا النظام.



الشكل 15: العلاقات بين الفاعلين.

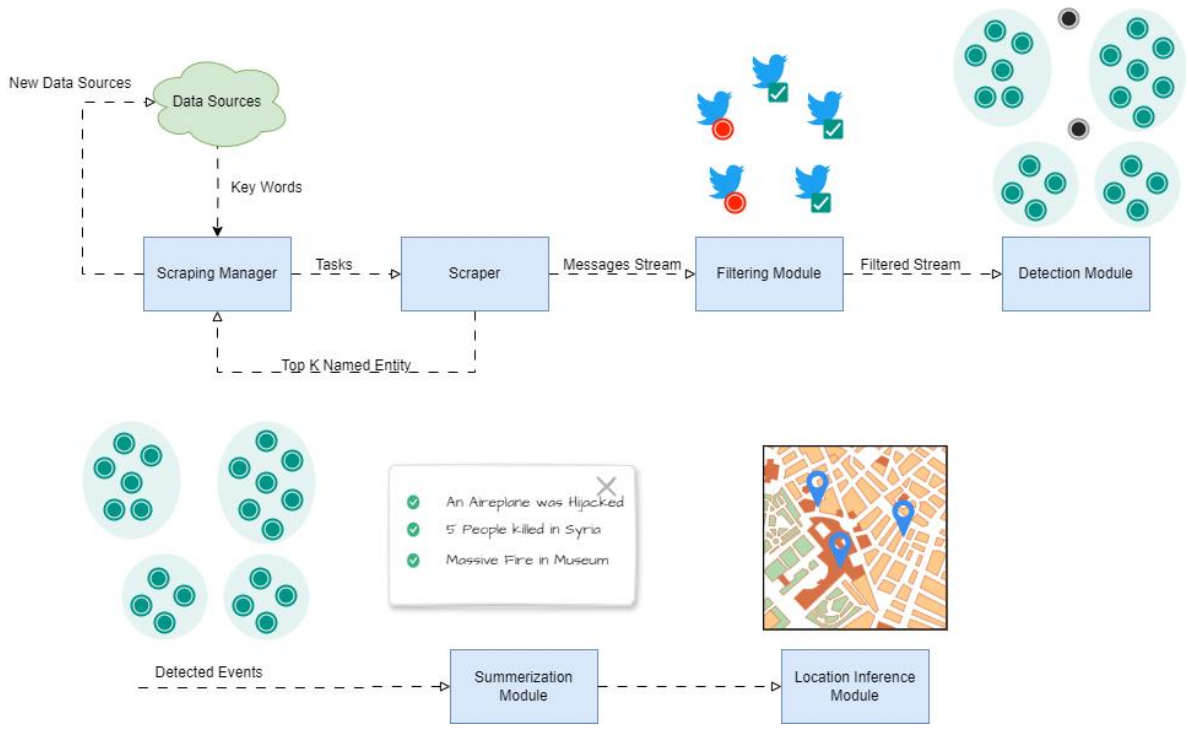
الفصل الخامس

المنهجية المقترحة

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

1.5- النهجية المقترحة

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، 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.

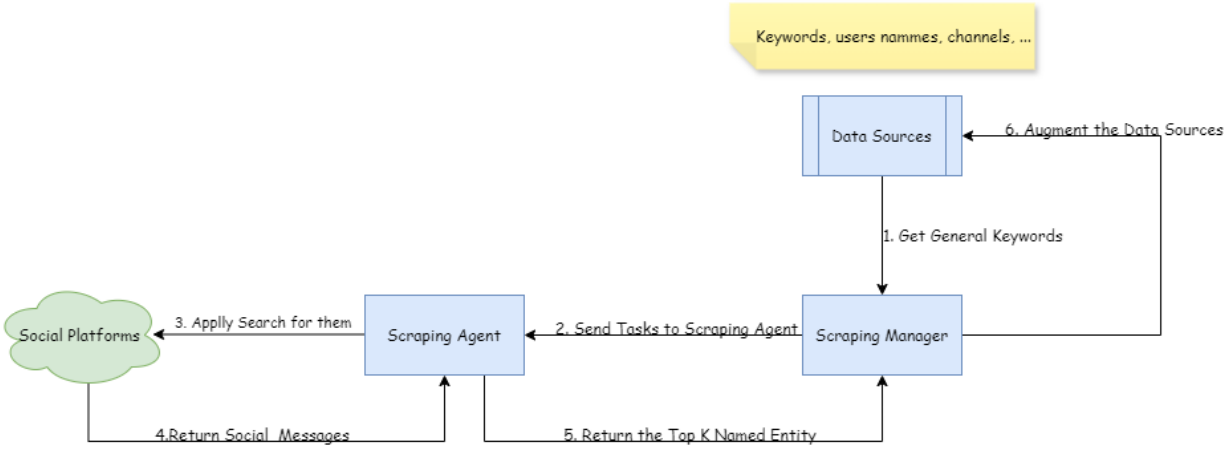


الشكل 16: البينة العامة للمنهجية.

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.

1.1.5 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 17: 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.

2.1.5 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).

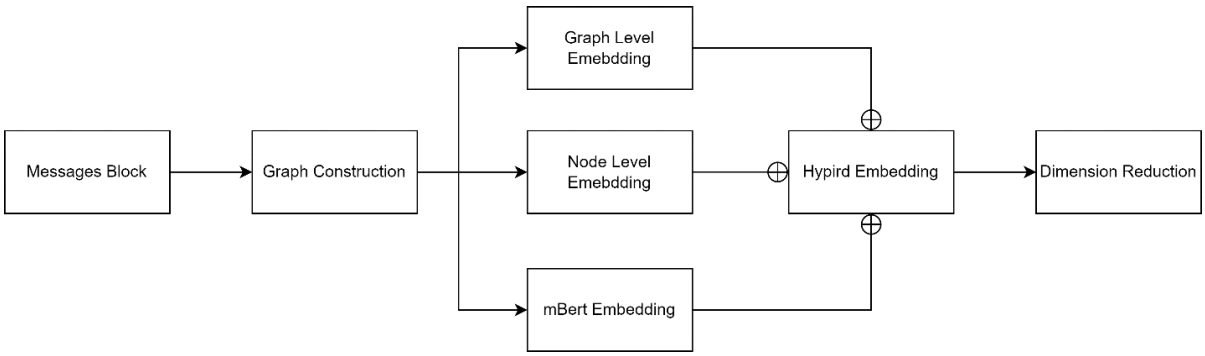
3.1.5 وحدة كشف الأحداث

مهمة هذا المكون هو عنقدة الرسائل، 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).

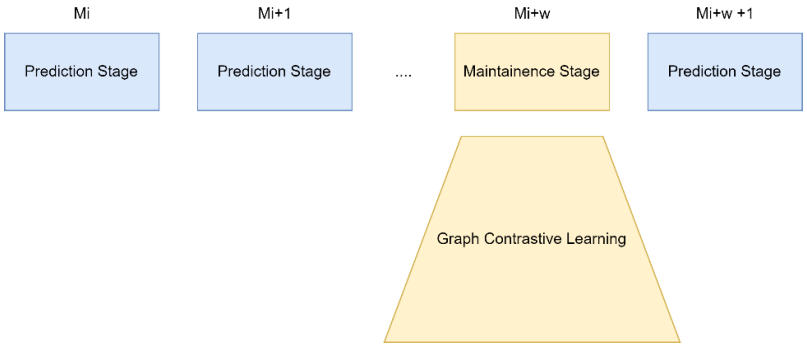
في هذه الوحدة، إن أفضل نموذج حصلنا عليه هو كما يلي:

أخذ تمثيل البيان Graph Level Embeddingوتمثيل العقدة Node Level Embedding مع تمثيل mBert(256) ومن ثم تطبيق خوارزمية تقليص الأبعاد PCA إلى 128 بعداً. كما هو موضح أدناه.



الشكل 18: آلية التضمين المقترحة.

استبدال عملية التدريب المستمر قبل كل كتلة بيانات بعملية صيانة دورية (تدريب كل X يوم على التوازي).



الشكل 19: عملية التدريب الدورية للنموذج.

4.1.5 وحدة تلخيص الحدث

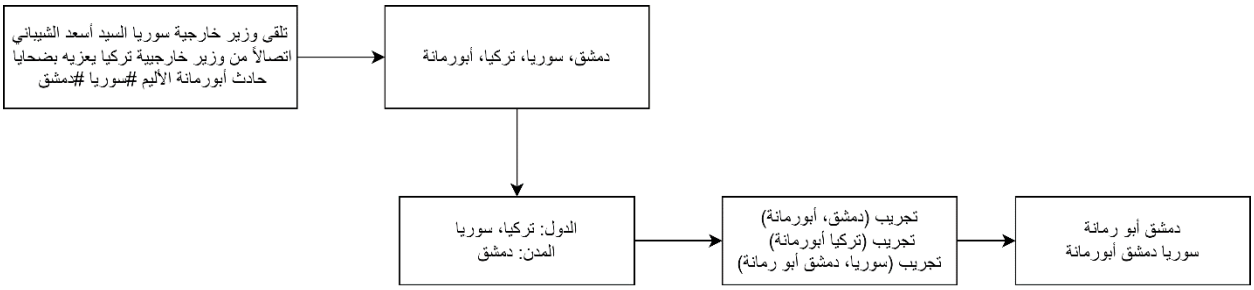
مهمة هذا المكون هو تلخيص مجموعة الرسائل وإعطاء ملخص للحدث، وفي هذا الجزء اعتمدنا على نموذج لغة ضخم لتلخيص الرسائل، كما بالإضافة إلى ذلك ولأنه قد يكون هنالك عدد كبير من التغريدات ضمن الحدث الواحد لذلك ليس من المنطقي أن نمررها كلها سوية، لأن ذلك غير ممكن بسبب محدودية طول الأمر Prompt المرسل إليه، ولأنها 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.

5.1.5 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 ضبط النموذج، فقط ووضع تقييم لجودة النموذج، حيث تم استخدام نموذج تعرف على الكيانات المسماة، وضبطه على مجموعة البيانات ، ومن ثم تعريف مجموعة من التقديرات heuristics، لتحديد الموقع المذكور انطلاقاً من أسماء المواقع المذكورة. حيث قمنا بذلك على النحو التالي:

- استخراج أسماء المواقع المذكورة في نص التغريدة باستخدام محول مدرب على ذلك (CaMel NER – mix)
- من أجل كل اسم موقع استخدام خدمة (أو Gazetteer) لمقابلة كل اسم بإحداثياته ونوعه (اسم مدينة، دولة، قرية)
- من أجل الأسماء التي ليست أسماء مدن أو دول، إعادة الاستعلام عنها بعد الحاقها بأسماء الدول ثم المدن.
- أخذ متوسط الإحداثيات لأقرب 3 نقاط لبعضها

مثال



الشكل 20: مثال على آلية الاستدلال على الموقع.

الفصل السادس

تصميم النظام

يوضح هذا الفصل تصميم خدمات النظام والأنماط التصميمية المرتبطة بها.

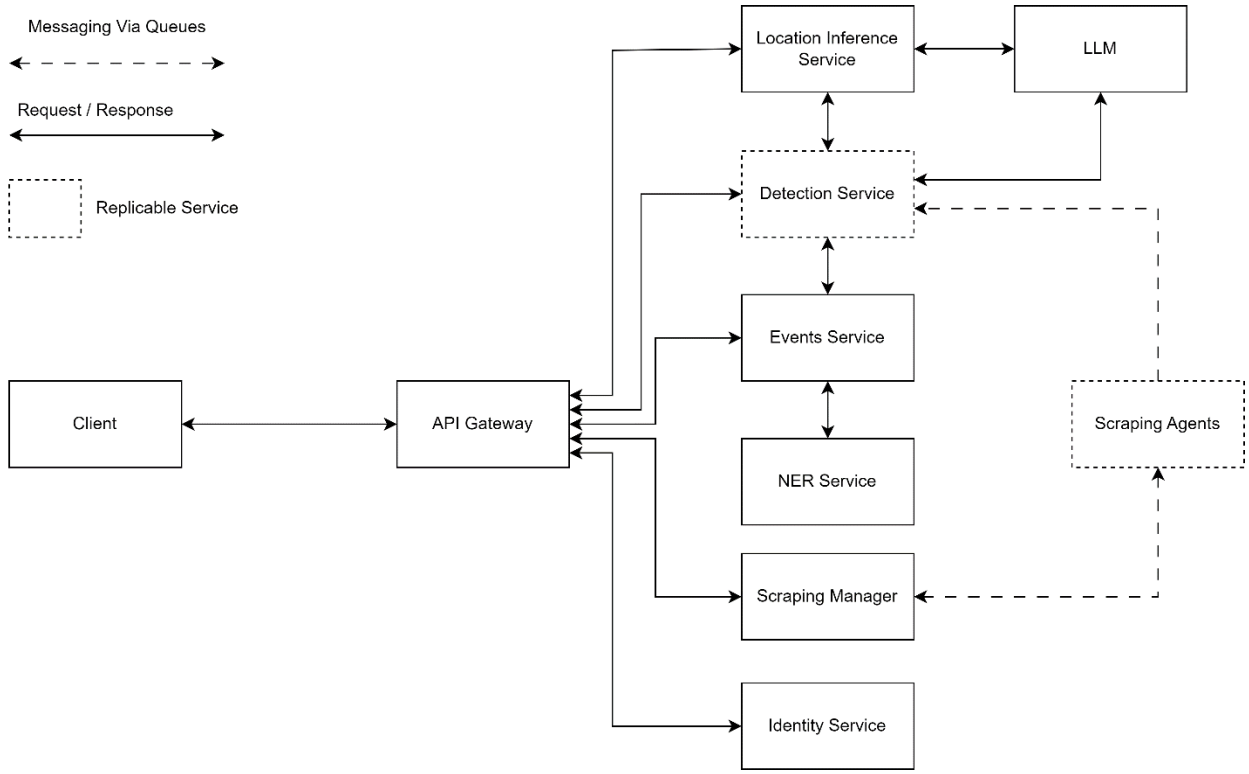
1.6- مقدمة

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 سحب (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.

2.6- مخطط خدمات النظام التصميمي



الشكل 21: 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.

3.6- تصميم خدمات النظام

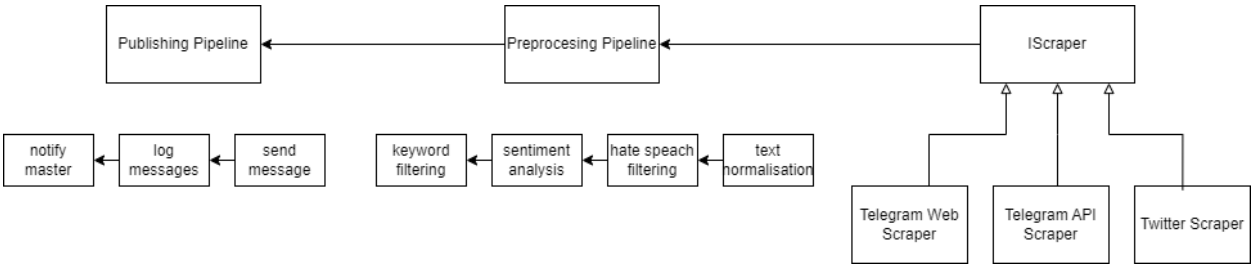
نعرض فيما يلي تصميمنا المقترح لخدمات النظام كل على حدة.

1) خدمة سحب البيانات 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 السحب Scraping Agent الموافق لمنصة التواصل التي تتبع لها المهمة، ومن ثم يبدأ بسحب البيانات.

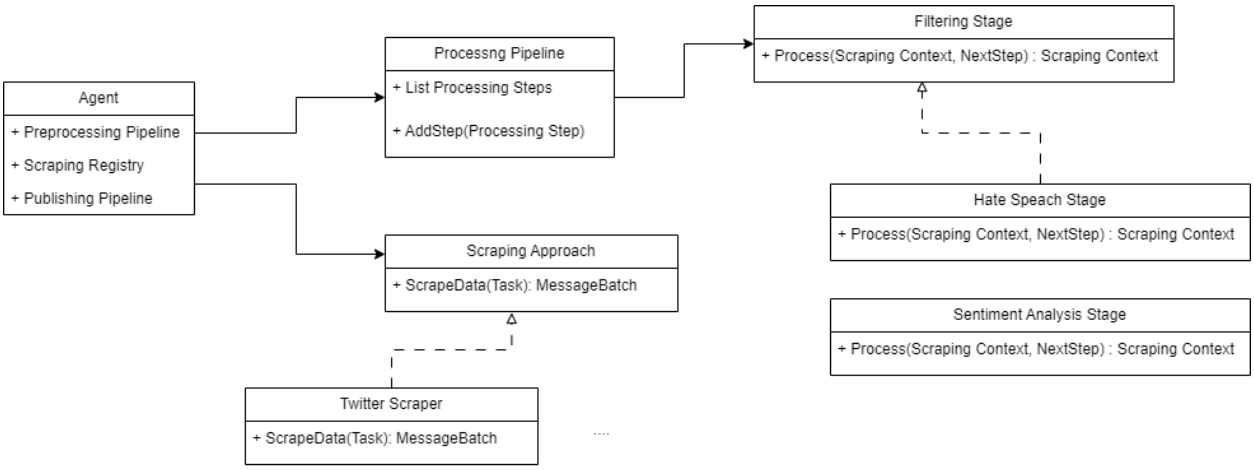
عندما يسحب دفعة من البيانات، ترسل إلى خط أنابيب المعالجة المسبقة ومن ثم إلى خط أنابيب النشر (Pre Processing Pipeline & Publishing Pipelin). ويوضح الشكل أدناه تلك العملية.



الشكل 22: مراحل العمل ضمن سحب تجريف البيانات.

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

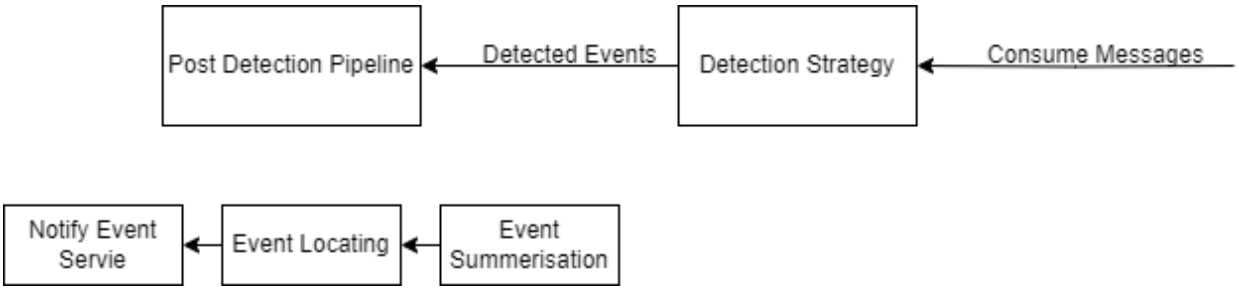
ويوضح الشك أدناه مخطط الصفوف للخدمة.



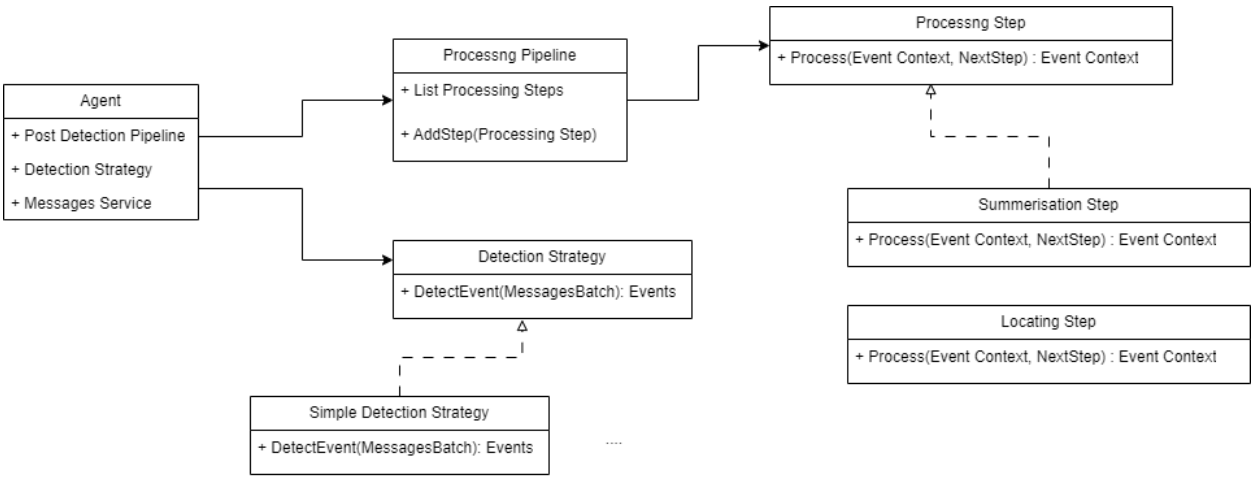
الشكل 23: مخطط الصفوف لخدمة تجريف البيانات.

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.

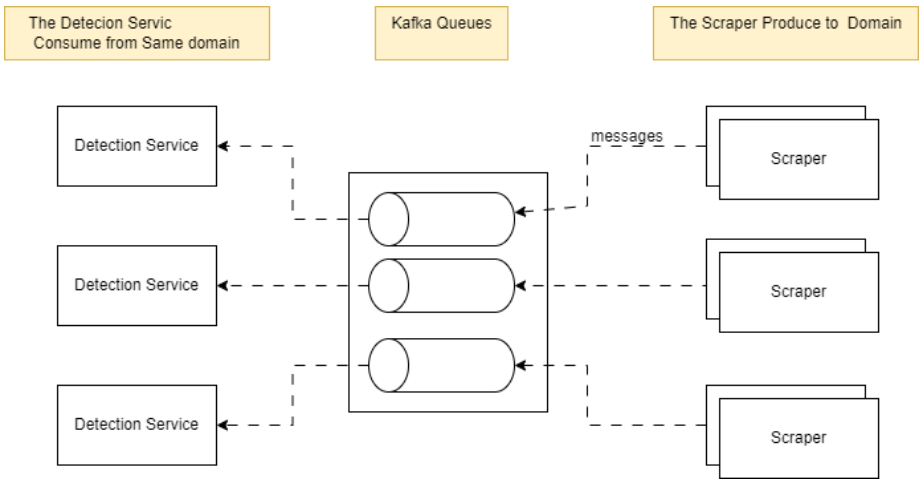


ويوضح الشكل أدناه مخطط الصفوف لها.



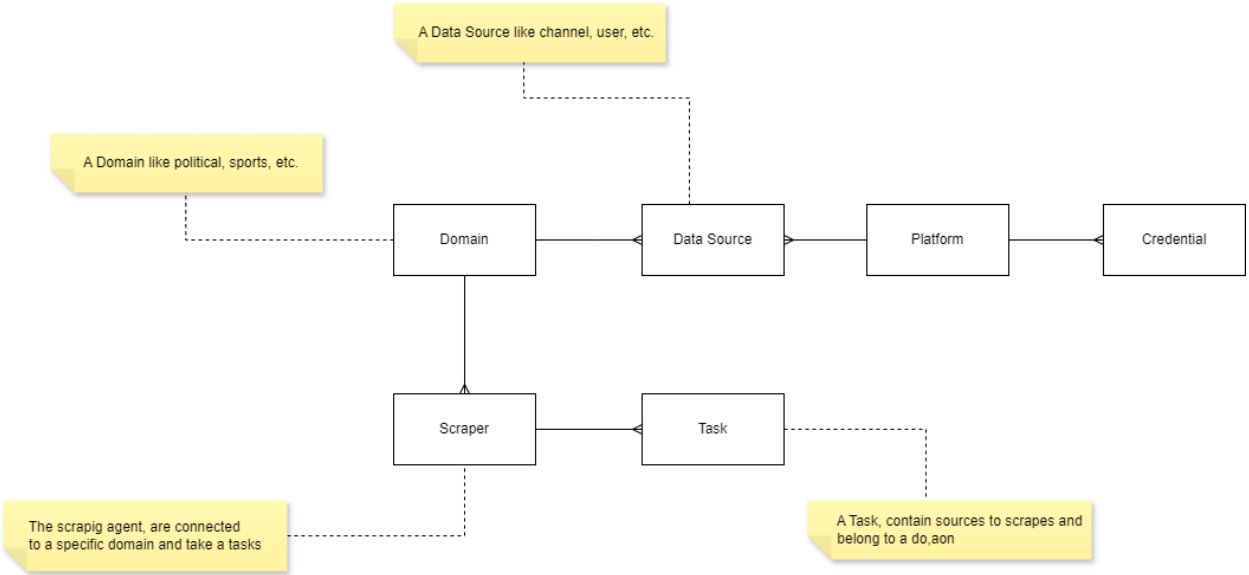
وعند معاينة هيكليتي تلك الخدمتين، نكون قد حققنا النمط Vertical Slice Architecture، أي تقسيم النظام عمودياً إلى خدمات مختلفة مستقلة عن بعضها البعض (فصل المكونات والتداخل بين ملفاتهما) ومتشابهة بمعماريتها الداخلية أي لكلا الخدمتين نفس البينة المعمارية.

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 ما بين الكيانات ERD في هذه الخدمة.

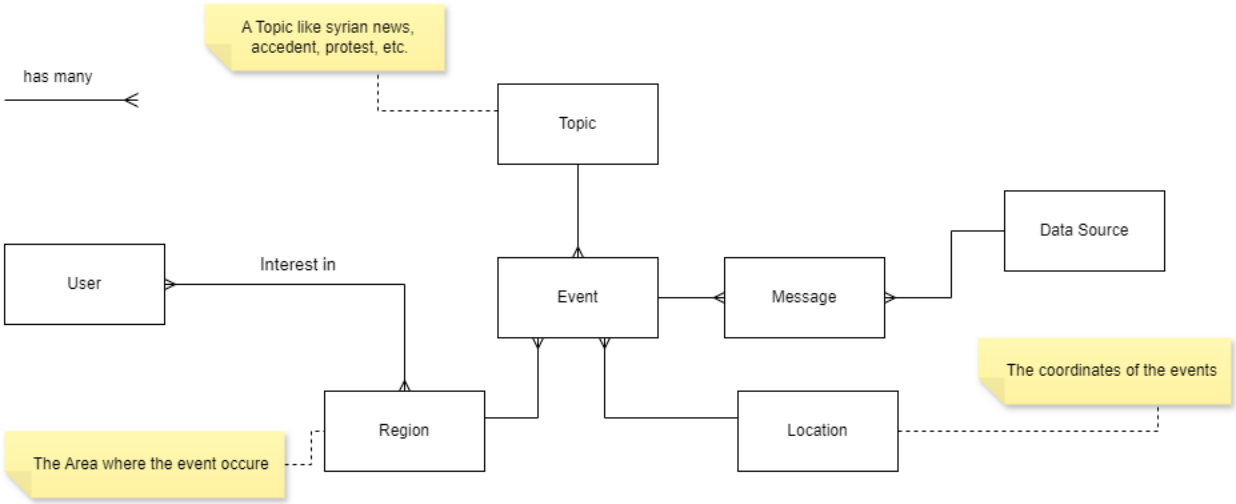


حيث وكما نرى فإنّ لكل مجال بيانات (أخبار سياسية سورية، أخبار رياضية، أحداث محلية ضمن مدينة) مرتبط مع عدد من مجرفي البيانات، الذين يستلمون المهمات. وكل مجال أيضاً مرتبط بمجموعة من مصادر البيانات (كلمات مفتاحية، معرفات قنوات، ... إلخ)، ومصدر البيانات مرتبط مع منصة تواصل اجتماعي ما. ولكل منصة تواصل معلومات 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، بحيث تتولى تسجيل المستخدمين، وإدارة الأدوار، والصلاحيات، وتقوم هذه الخدمة بإعطاء المستخدمين معرفات JWT بحيث يتمكنون من استخدامها للتحقق منهم ومن صلاحياتهم في بقية الخدمات.

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، تم اعتماد معمارية الشرائح العمودية، بحيث يتم تقسيم هذه الخدمة إلى مجموعة وحدات modules، كل وحدة تقابل خدمة في طرف المخدم وتقوم بعرض واجهاتها، وبحيث لكل الوحدات نفس البنية المعمارية.

الفصل السابع

الأدوات المستخدمة

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

Docker -1.7

هي منصة مفتوحة مصممة لأتمتة عملية نشر التطبيقات وتوسيع نطاقها، وإدارتها في حاويات خفيفة الوزن. الحاوية هي بيئة معزولة تجمع بين التطبيق وجميع تبعياته، مما يسمح بتشغيله بشكل متسق عبر بيئات حوسبة مختلفة، مثل أنظمة تشغيل مختلفة أو مزودي الخدمات السحابية.

ASP.Net Web API -2.7

هو إطار عمل طورته شركة Microsoft يتيح إنشاء خدمات تعتمد على http. يمكن استخدام هذه الخدمات بواسطة مجموعة متنوعة من عملاء، بما في ذلك متصفحات الويب وتطبيقات الأجهزة المحمولة وخدمات الويب الأخرى. إنه جزء من إطار عمل ASP.Net. وهو مصمم خصيصاً لإنشاء واجهات برمجة التطبيقات RestFull API.

Jenkins -3.7

هو خادم أتمتة مفتوح المصدر يُستخدم لتحقيق التكامل المستمر (CI) والتسليم المستمر (CD) في تطوير البرمجيات. يساعد Jenkins في أتمتة أجزاء من عملية تطوير البرمجيات من خلال إنشاء خطوط أنابيب (Pipelines) تقوم ببناء واختبار ونشر الكود البرمجي تلقائيًا عند اكتشاف تغييرات في نظام التحكم بالإصدار (Source Control System) مثل Git. كما يوفر Jenkins واجهة ويب تتيح إدارة عمليات الأتمتة بسهولة وفعالية.

YARP -4.7

هي مكتبة عالية الأداء طورتها شركة Microsoft ضمن إطار العمل ASP.NET، وقد صُممت لتسهيل إنشاء خوادم بوابة (Gateways)تعمل على الطبقة السابعة من نموذج (OSI)، مما يتيح لها التعامل مع الطلبات الواردة وتوجيهها إلى الخدمات المناسبة. وعلى عكس خوادم البوابة التي قد تعمل على طبقات أدنى، يوفر YARP مرونة أكبر من خلال تمكين المطورين من إعادة تعيين عناوين الخدمات وإدارة الاتصالات فيما بينها بشكل مستقل عن الطبقات الأدنى. وتُعد هذه القدرة مفيدة بشكل خاص في سيناريوهات مثل موازنة الأحمال (Load Balancing) وتنسيق التواصل في بيئات تعتمد على الخدمات المصغّرة (Microservices).

5.7 قاعدة المعطيات SQL Server

هو نظام إدارة قواعد بيانات علائقية (Relational Database Management System) تم تطويره وتسويقه من قبل شركة Microsoft. يعمل SQL Server على نظامي التشغيل Windows وLinux، ويوفر سعة تخزين عالية مع أداء متميز في

استعادة البيانات. يتميز SQL Server بقدرته على التعامل مع كميات كبيرة من البيانات بكفاءة عالية، مما يجعله مناسبًا للاستخدام في المؤسسات التي تتطلب إدارة بيانات قوية وموثوقة.

6.7 JSON Web Token

هو معيار مفتوح يُستخدم لنقل المعلومات بين طرفين بشكل آمن عبر الويب باستخدام غرض JSON. يتميز JWT بقدرته على تقديم آلية مصادقة لا تعتمد على الجلسات (stateless authentication)، مما يجعله مناسبًا للتطبيقات الحديثة. على عكس ملفات تعريف الارتباط (cookies)، كما يمكن استخدام رموز JWT في المصادقة عبر تطبيقات الويب، تطبيقات الهواتف المحمولة، وتطبيقات سطح المكتب. بفضل هذه المرونة، أصبح JWT خيارًا شائعًا لضمان أمان الاتصال بين المستخدمين والخوادم في بيئات متعددة.

7.7 إطار العمل Angular

هو إطار عمل مفتوح المصدر لتطوير تطبيقات الويب تم إنشاؤه وصيانته من قبل Google. يعتمد على لغة TypeScript ويوفر بنية قائمة على المكونات، حيث يتم تقسيم التطبيق إلى وحدات صغيرة قابلة لإعادة الاستخدام، مما يسهل إدارة وتوسيع التطبيقات. يتيح Angular ربط البيانات ثنائي الاتجاه، مما يضمن التحديث المتزامن بين النموذج وواجهة المستخدم. كما يتضمن نظام توجيه متقدم لتطوير تطبيقات الصفحة الواحدة (Single Page Applications - SPAs) وخدمات حقن التبعية Dependency Injection التي تعزز من قابلية الصيانة واختبار التطبيق. بالإضافة إلى ذلك، توفر أداة Angular CLI دعمًا متكاملاً لإعداد المشروع، بناء التطبيق، وإجراء الاختبارات، مما يعزز من إنتاجية المطورين.

8.7 الأداة Grafana k6

هي أداة اختبار للحمل (Load Testing Tool) مفتوحة المصدر تساعد في اختبار أداء البرمجيات خصوصاً في حالة تعدد المستخدمين. k6 مجاني ومتمحور حول المطورين وقابل للتوسيع. باستخدام k6، يمكن اختبار موثوقية وأداء الأنظمة البرمجية والتعرف على تراجعات الأداء والمشكلات في وقت مبكر.

الفصل الثامن

التجارب ومقارنة النماذج

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

1.8- مجموعة البيانات

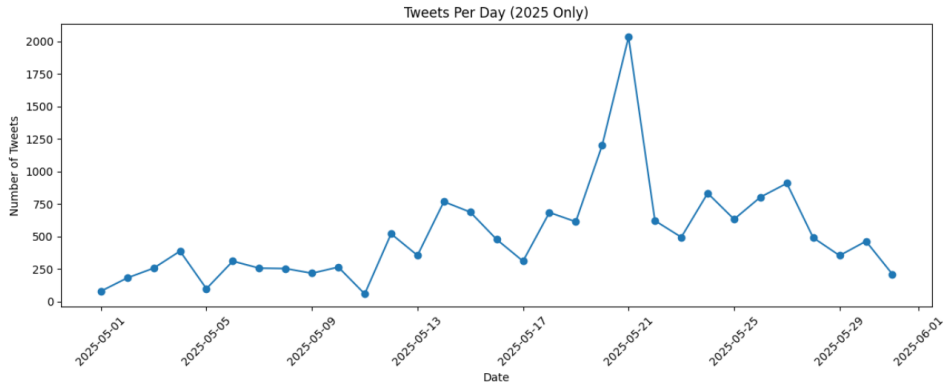
تم جمع مجموعة البيانات من منصة تويتر عبر البحث بكلمات مفتاحية عامة وخاصة (مثل، سوريا، دمشق، اعتصام الميدان، ...)، من تاريخ 1 أيار ولغاية 31 أيار 2025، ومن خلالها جمعنا مجموعة بيانات حجمها 20 ألف تغريدة (23 ألف قبل إزالة التكرارات) بالإضافة على معلوماتها metadata. سميناها SAS (Golden)، وبعد ذلك قمنا بتنميط 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، 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.

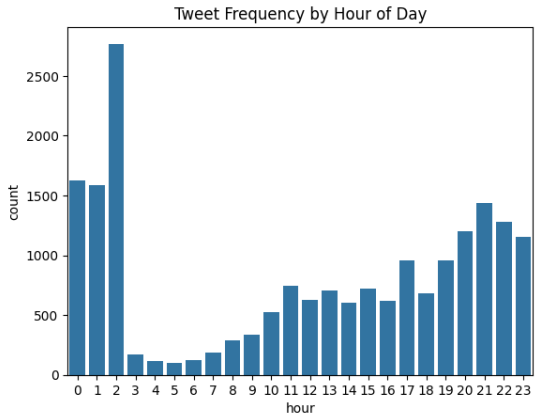
Author	Type	Tweet Text
Grok	Reply	رامز الخياط هو رجل أعمال قطري، الرئيس التنفيذي لشركة أورباكون القابضة، وهي جزء من مجموعة باور إنترناشيونال القابضة التي يملكها ويديرها عائلة الخياط في قطر. تأسست أورباكون عام 2011، وتتخصص في مشاريع الطاقة والبنية التحتية. الخ
الإخبارية السورية	Tweet	مؤتمر صحفي مشترك لوزير الطاقة محمد البشير والرئيس التنفيذي لشركة أورباكون رامز الخياط على هامش توقيع الاتفاقيات الاقتصادية مع شركات الطاقة #سوريا_نحو_الريادة #الإخبارية_السورية
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.
ssy	Tweet	أعجبني برتوكول توقيع الاتفاقيات
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

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 24: 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 مستقراً ظهراً، ويفسر سبب ارتفاعها ليلاً لوجود عدة أحداث مهمة وقعت ليلاً (مثل رفع العقوبات الأميركية). ويساعد هذا المخطط على تحديد آلية وتوقيت سحب البيانات عند وضع النظام بالخدمة.

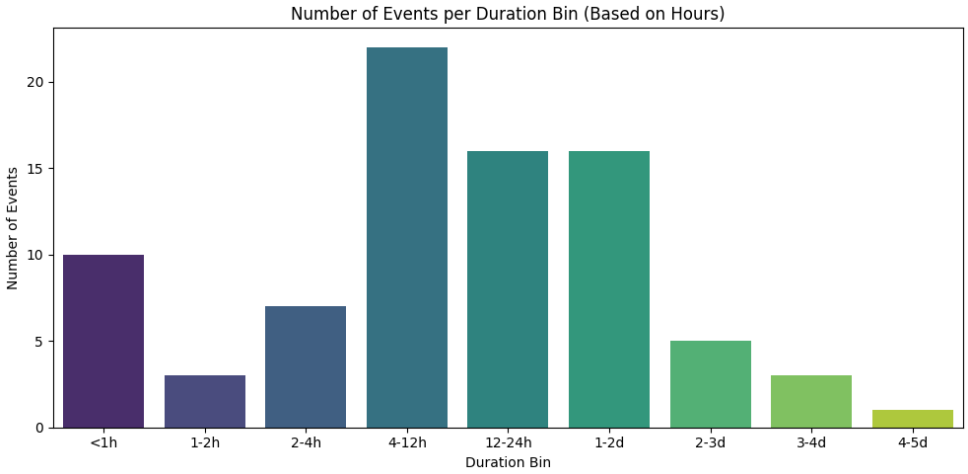


الشكل 25: تواتر التغريجات خلال ساعات اليوم.

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

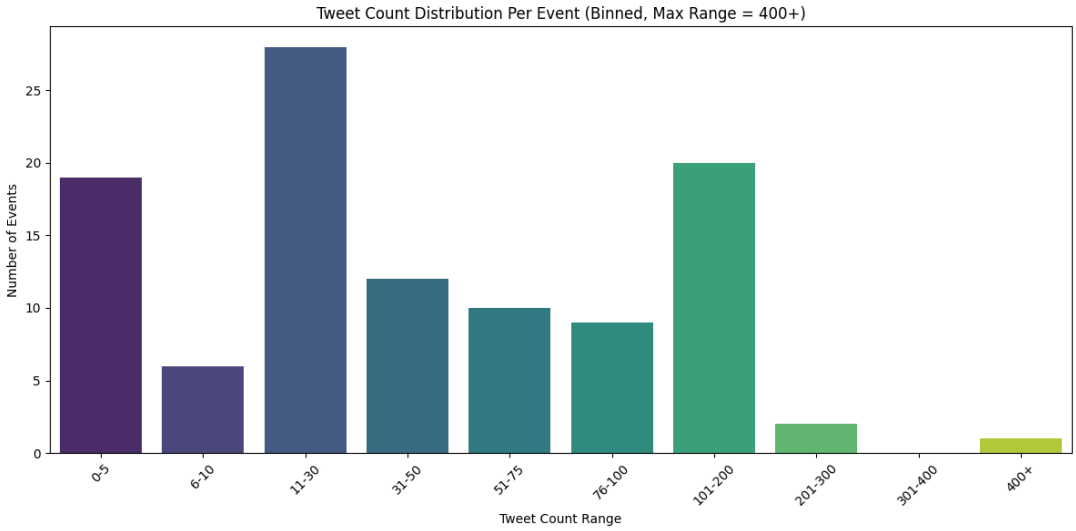
الشكل 26: مجموعة الكلمات المتداولة خلال 8 أيار– الفترة 9AM – 12 AM.

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.



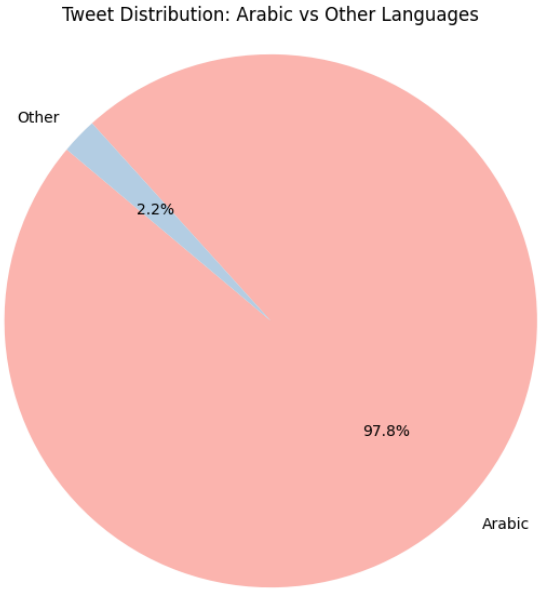
الشكل 27: المدة الزمنية للأحداث.

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.



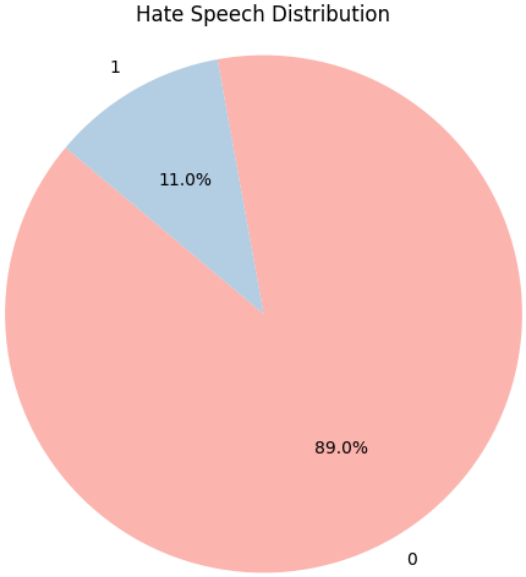
الشكل 28: مجال عدد التغريدات ضمن الحدث.

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



الشكل 29: توزع لغات التغريدات.

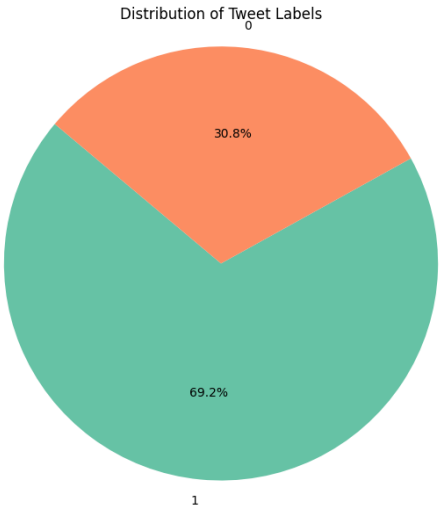
ويبين الشكل أدنا تحليل خطاب الكراهية في التغريدات، ويلاحظ أن 11% بالمئة من التغريدات تحض على العنف والكراهية.



الشكل 30: تحليل خطاب الكراهية والعنف.

2.8- 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.



الشكل 31: توزع نمط التغريدات المصنفة.

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

جدول**8** : نتائج التجربة الأولى في مرحلة تصنيف التغريدات.

#	Embedding Model	Model	Results							
			Accuracy		Recall		Precision		F1 Score	
			Train	Test	Train	Test	Train	Test	Train	Test
Exp. 1	FastText	Neural Network	0.84	0.82	0.90	0.89	0.86	0.86	0.81	0.77
		SVM	0.89	0.84	0.95	0.90	0.91	0.89	0.86	0.79
	AraBert	Neural Network	0.86	0.82	0.88	0.85	0.92	0.90	0.85	0.79
		Logistic Regression	0.87	0.83	0.92	0.89	0.90	0.89	0.85	0.79
	mBert	SVM	0.85	0.81	0.90	0.88	0.89	0.87	0.83	0.77
		KNN	0.86	**0.84**	0.91	0.89	0.88	0.88	0.83	**0.80**
	TF-IDF	SVM	0.87	0.80	0.93	0.89	0.88	0.84	0.84	0.74
		KNN	0.84	0.81	0.88	0.83	0.89	0.90	0.82	0.78
	Distill mBert	Neural Network	0.87	0.79	0.90	0.85	0.91	0.86	0.85	0.74
		SVM	0.83	0.79	0.86	0.82	0.88	0.87	0.80	0.74

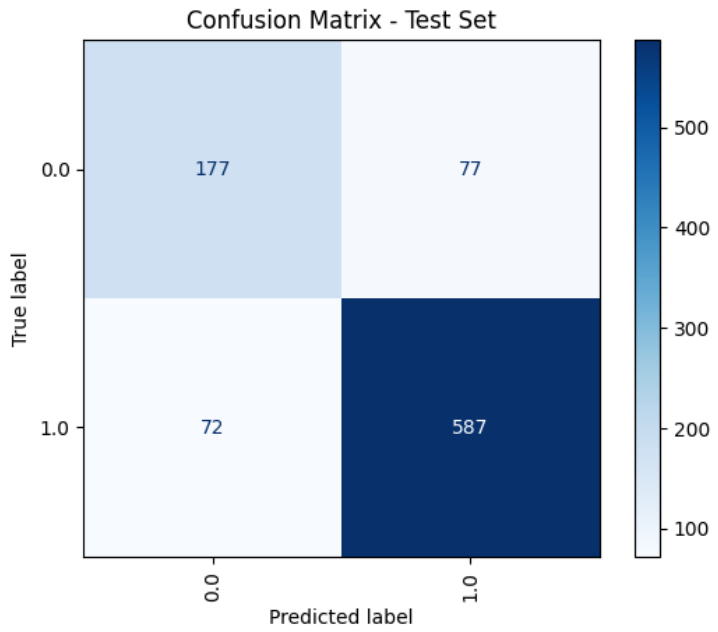
حيث كان النموذج الأفضل للتمثيل هو mBert مع نموذج knn للتعلم، حيث كانت نتائجه لـ F1 macro هي 0.80، وصحة accuracy هي 84%، يليه نموذج FastText بـ F1 macro 0.79، وصحة 84%، ويوح الشكل أدماه نتائج التصنيف له.

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
```

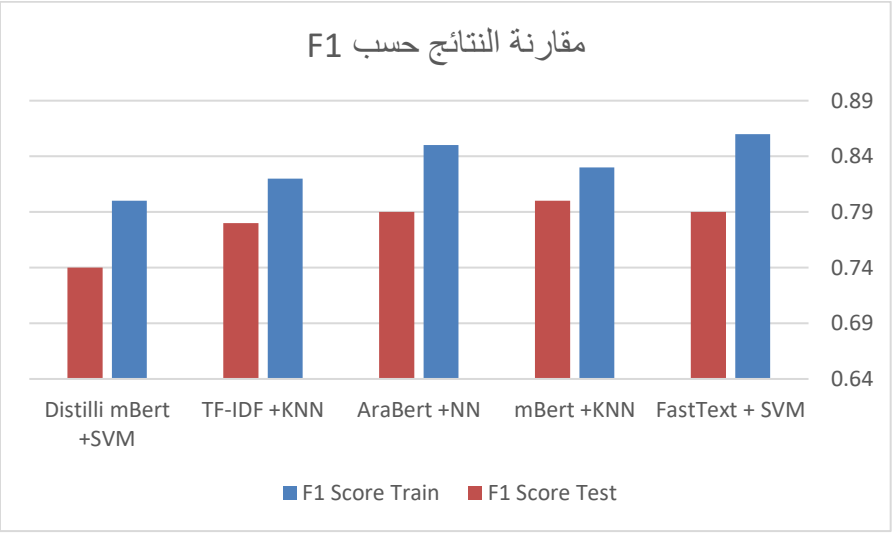
الشكل 32: نتائج التصنيف لنموذج **mBert** في التجربة الأولى.

ويوضح الشكل أدناه مصفوفة الارتباك له.



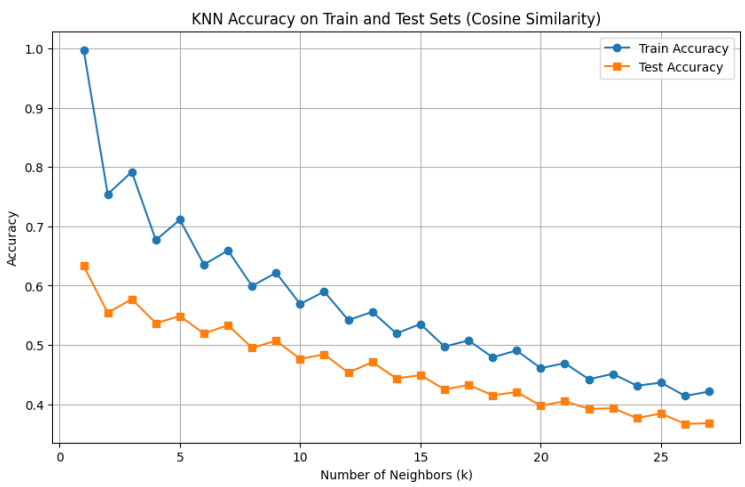
الشكل 33: مصفوفة الارتباك للنموذج **mBert** في التجربة الأولى.

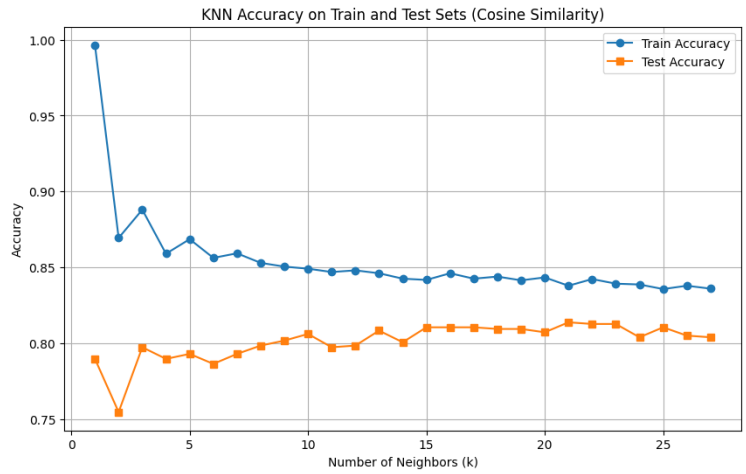
ويبين الشكل أدناه الفرق بين نتاج هذه الطرق، مقارنةً حسب F1-Score.



الشكل 34: مخطط مقارنة النتائج حسب معيار **F1**.

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





المسماة)، لم تضف هذه التجربة قيمة مضافة للنتائج، ولكنها خففت من الفرق بين نتائج التدريب والاختبار.

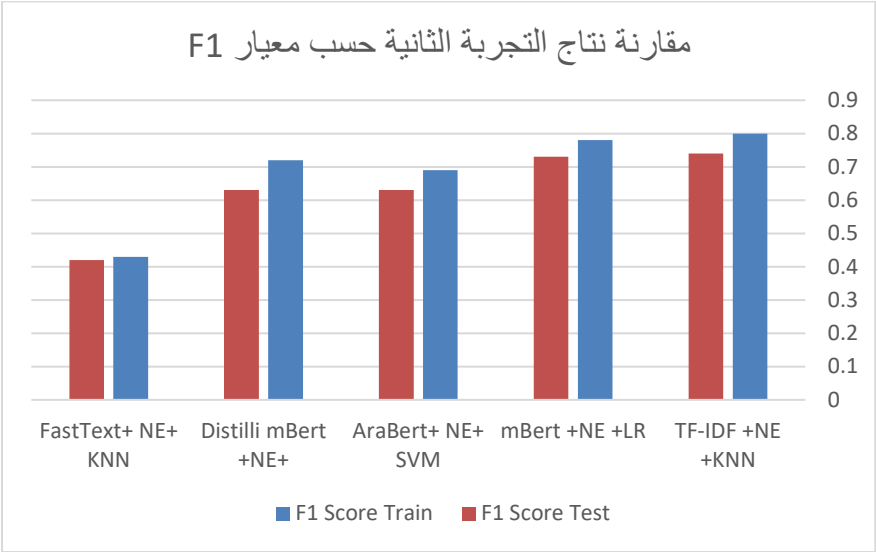
التجربة الثانية: تمثيل الكيانات المسماة فقط

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.

جدول**9** : نتائج التجربة الثانية في مرحلة تصنيف التغريدات.

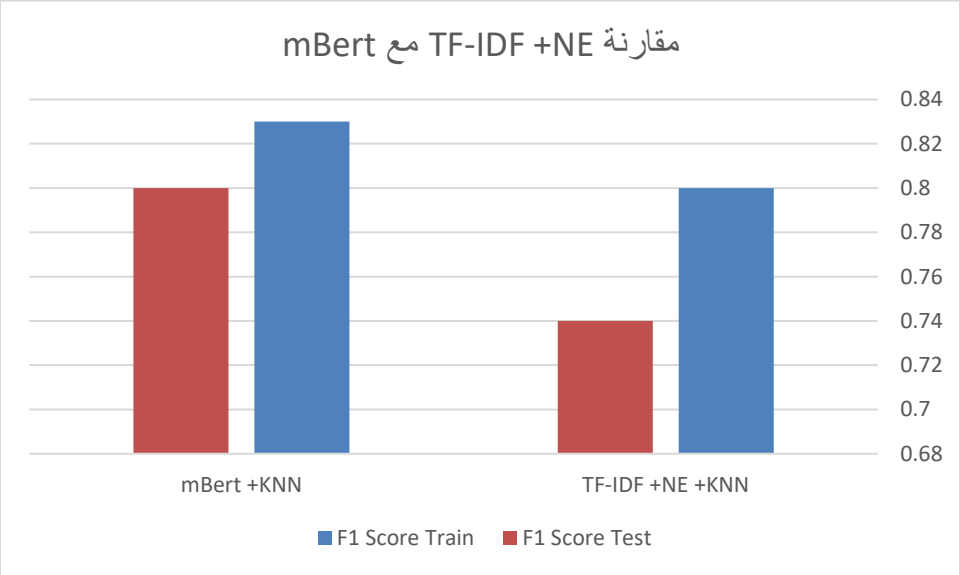
#	Embedding Model	Model	Results							
			Accuracy		Recall		Precision		F1 Score	
			Train	Test	Train	Test	Train	Test	Train	Test
Exp. 2	FastText+ NE	Neural Network	0.68	0.72	1.0	0.50	0.68	0.72	0.41	0.42
		KNN	0.69	0.72	1.0	0.51	0.69	0.72	0.43	0.42
	AraBert+ NE	Neural Network	0.73	0.75	0.95	0.94	0.73	0.77	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
	mBert +NE	Logistic Regression	0.78	0.73	0.94	0.88	0.78	0.78	0.78	0.73
		SVM	0.82	0.74	0.95	0.88	0.82	0.78	0.77	0.64
	TF-IDF+ NE	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**
	Distill + NE	Neural Network	0.78	0.74	0.92	0.89	0.79	0.78	0.72	0.63
		KNN	0.75	0.72	0.88	0.84	0.78	0.78	0.68	0.62

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. ويبين الشكل أدناه، مقارنة هذه النماذج، ومقارنتها مع النموذج الأفضل في التجربة السابقة.



الشكل 35: مخطط مقارنة نتاج التجربة الثانية حسب معيار **F1**.

ويبين الشكل أدناه الفرق بين نتائج TF-IDF + KNN، مع نموذج التجربة السابقة.



الشكل 36: مقارنة **TF-IDF +NE** مع **mBert**.

التجربة الثالثة: تثقيل شعاع الكيانات المسماة

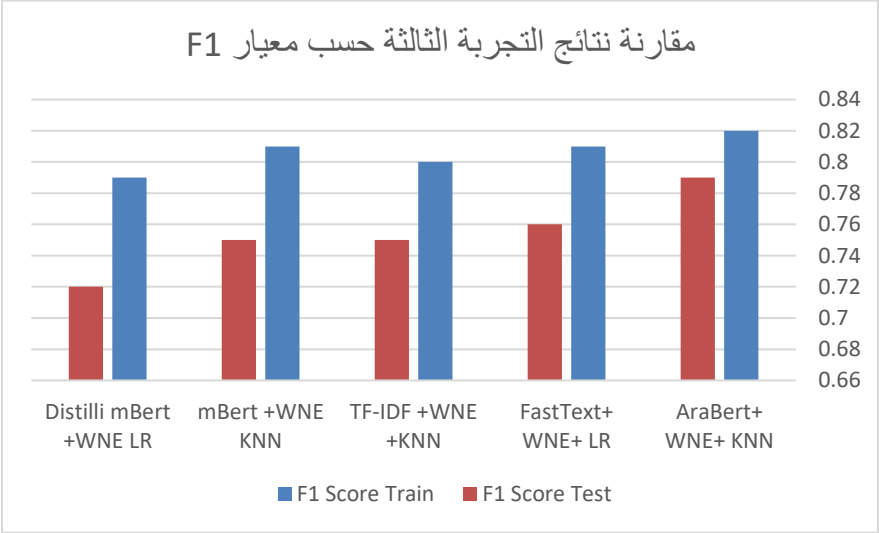
في هذه التجربة أخذنا تثقيلاً لتمثيل الكيانات المسماة (شعاع التمثيل = شعاع التمثيل القديم + w * شعاع تمثيل الكيانات المسماة)، لم تضف هذه التجربة قيمة مضافة للنتائج، ولكنها خففت من الفرق بين نتائج التدريب والاختبار.

وكان النتائج وفق الجداول التالية.

جدول**10** : نتائج التجربة الثالثة.

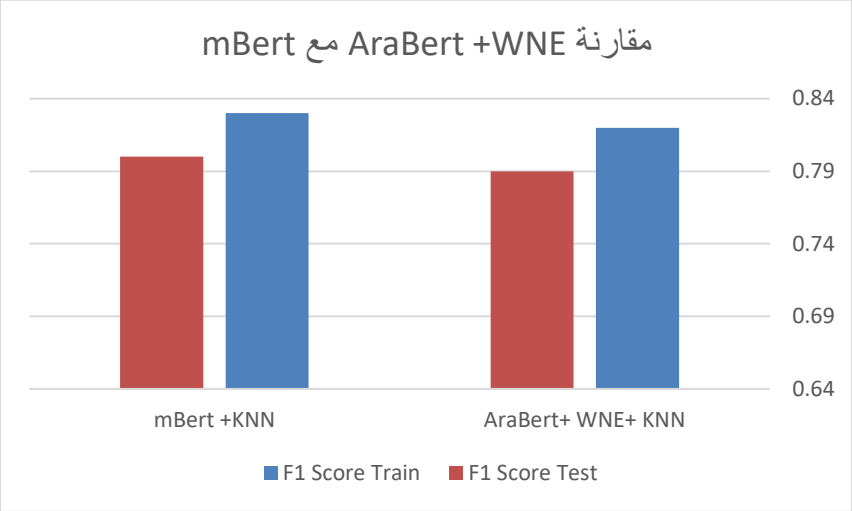
#	Name	Model	results							
			Accuracy		Recall		Precision		F1 Score	
			Train	Test	Train	Test	Train	Test	Train	Test
Exp. 3	FastText	SVM	0.85	0.81	0.91	0.87	0.88	0.87	0.83	0.76
		Logistic Regression	0.84	0.81	0.91	0.89	0.87	0.86	0.81	0.76
	AraBert	Logistic Regression	0.86	0.83	0.91	0.88	0.89	0.89	0.84	0.79
		KNN	0.85	**0.83**	0.90	0.88	0.88	0.88	0.82	**0.79**
	mBert	Logistic Regression	0.85	0.80	0.90	0.87	0.88	0.86	0.82	0.75
		KNN	0.83	0.79	0.86	0.82	0.89	0.88	0.81	0.75
	TF-IDF	Logistic Regression	0.85	0.77	0.92	0.85	0.86	0.83	0.81	0.70
		KNN	0.83	0.75	0.88	0.83	0.87	0.82	0.80	0.75
	Distill mBert	Logistic Regression	0.82	0.77	0.89	0.83	0.86	0.85	0.79	0.72
		SVM	0.83	0.76	0.87	0.80	0.88	0.86	0.80	0.72

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



الشكل 37: مقارنة نتائج التجربة الثالثة حسب معيار **F1**.

ويبين الشكل أدناه الفرق بين نتائج النموذج الأفضل في هذه التجربة مع أفضل نموذج حصلنا عليه.



الشكل 38: أدناه الفرق بين نتائج النموذج الأفضل في هذه التجربة مع **mBert**.

التجربة الرابعة: استبدال الكيانات المسماة بكلمات بدل Placeholders

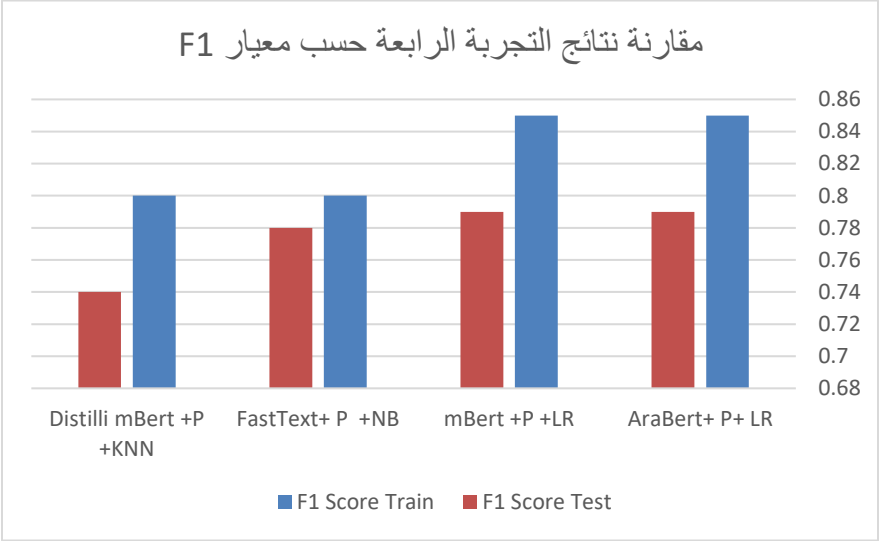
في هذه التجربة قمنا باستبدال الكيانات المسماة بكلمات بدل، لأننا كنا نعتقد أن النموذج يتعلم معلومات مكانية غير دقيقة لا يجب تعلمها (أي أن النموذج رأى مثلاً العبارة "وقع حادث في دمشق" بينما لم يرى العبارة "وقع حادث في حمص" لذلك كنا نعتقد أن النموذج يأخذ في الاعتبار كلمة دمشق أثناء التصنيف مما لا يجعله 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.

جدول**11** : نتائج التجربة الرابعة في مرحلة تصنيف التغريدات.

#	Name	Model	results							
			Accuracy		Recall		Precision		F1 Score	
			Train	Test	Train	Test	Train	Test	Train	Test
Exp. 4	FastText	Neural Network	0.83	0.82	0.89	0.87	0.87	0.87	0.80	0.78
		SVM	0.84	0.82	0.91	0.88	0.87	0.87	0.81	0.77
	AraBert	Neural Network	0.87	0.82	0.91	0.88	0.90	0.88	0.85	0.78
		Logistic Regression	0.87	0.83	0.92	0.89	0.90	0.88	0.85	0.79
	mBert	Neural Network	0.87	0.82	0.91	0.88	0.90	0.88	0.85	0.78
		Logistic Regression	0.87	0.83	0.92	0.89	0.90	0.88	0.85	0.79
	TF-IDF	KNN	0.85	**0.83**	0.89	0.86	0.90	0.90	0.83	**0.80**

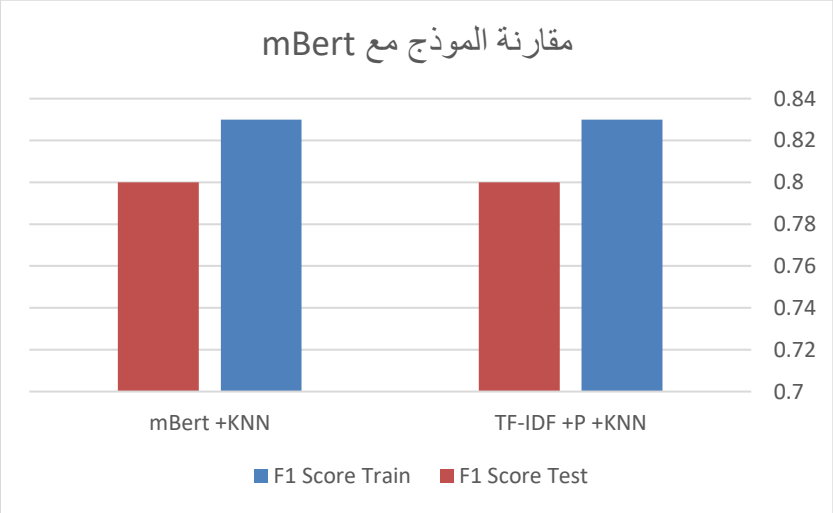
		Logistic Regression	0.84	0.83	0.90	0.89	0.87	0.87	0.81	0.78
	Distill mBert	Neural Network	0.88	0.79	0.90	0.84	0.92	0.86	0.86	0.74
		KNN	0.83	0.80	0.86	0.83	0.88	0.86	0.80	0.74

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%. ويبين الشكل أدناه مقارنة هذه النتائج.



الشكل 39: مقارنة نتائج التجربة الرابعة.

ويبين الشكل أدناه الفرق بين نتاج أفضل نموذج في هذه المرحلة مع النموذج mBert.



الشكل 40: مقارنة نموذج التجربة الرابعة مع **mBert**.

3.8- 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, ...

1.3.8 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، 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. وأجرينا عدداً من التجارب، حيث بدأنا باختبار النماذج المستخدمة في الأبحاث السابقة وهيLDA, Word2Vec, Bert, sBert, FastText.

من أجل Bert، استخدمه الباحثون من خلال أخذ متوسط التضمينات (أي ليس كـ Sentence Transformer) وقمنا باستخدام النسخة العربية منه AraBert. ومن أجل sBert، أي استخدام Bert كـ Sentence Transformer قمنا باستخدام النسخة العربية منه وهي mBert. بينما من أجل FastText لم يستخدمه الباحثون في الأوراق السابقة، ونحن استخدمناه لأنه يعالج مشكلة Out of Vocabularyوهو يعطي تضمين أفضل للكلمة (مع وجود أخطاء إملائية) كونه يستخدم n-gram على مستوى المحارف.

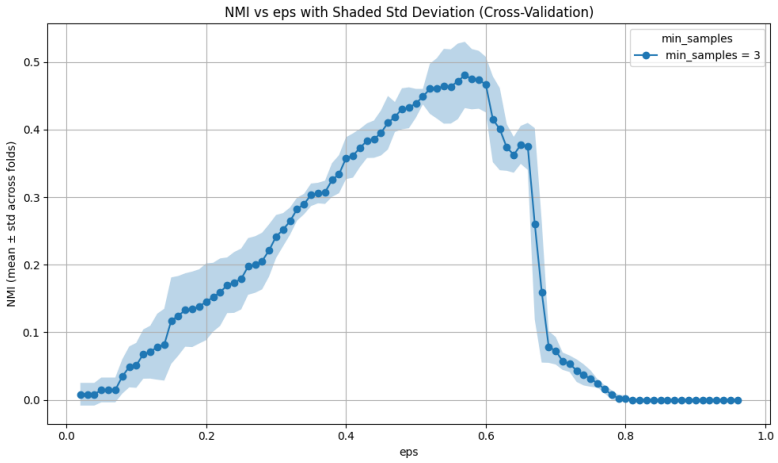
التجربة الأولى: التضمين باستخدام نماذج التضمين المرجعية

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

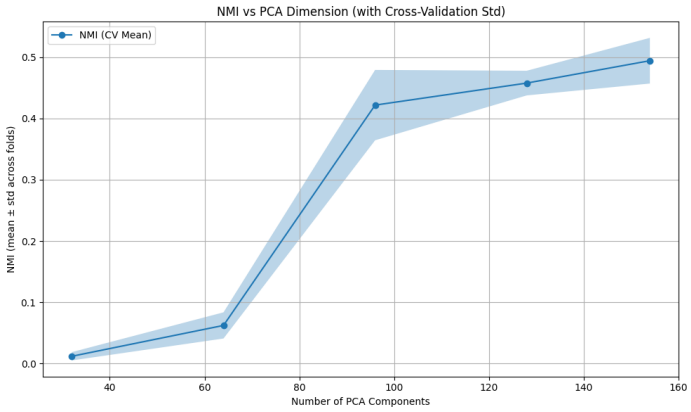
- تحليل معامل المسافة min epsilon

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 قيمة له، متوسط نتائج الـ cross validation مع الانحراف المعياري، من أجل طريقة التمثيل mBert، (بقية الطرائق كانت النتائج مشابهة كشكل التابع)، حيث نلاحظ أنه من أجل القيم الصغير لـ eps، 0.23>، كان الانحراف المعياري كبيراً، عى عكس بقية القيم.



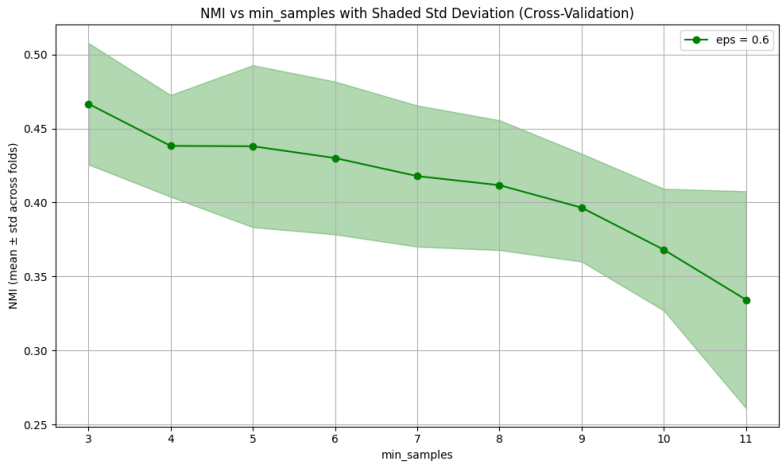
الشكل 41: أثر المعامل **eps** على النتائج باستخدام **mBert** لتمثيل الرسائل.

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



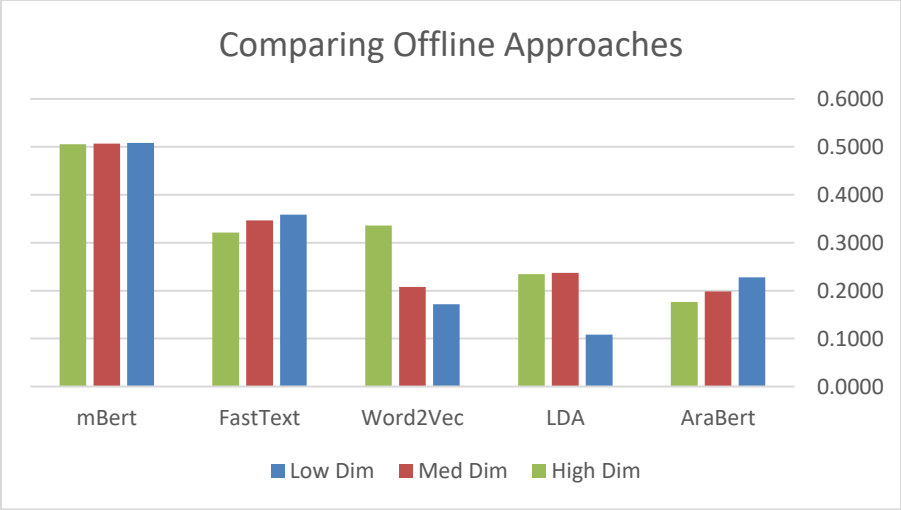
الشكل 42: أثر عدد الأبعاد على نتائج العنقدة.

- 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 8%∓، مما يوحي بوجود overfitting، لذلك في التجارب القادمة سنلجأ إلى تعزيز مجموعة البيانات.



الشكل 43: أثر معامل الكثافة على جودة العنقدة.

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



الشكل 44: نتائج التجربة الأولى.

ويوضح الجدول (الملحق ت – نتائج التجربة الأولى في مرحلة كشف الأحداث) أدناه بقية المقاييس.

التجربة الثانية: خوارزمية العنقدة الهرمية HDBSCAN

عند استخدامنا في التجربة السابقة لخوارزمية العنقدة DBSCAN، كانت تعاني في الوضع الحي من اختلاف كثافة الأحداث مما يؤدي إلى تغير قيمة epsilon المناسبة، وبالتالي كنات النتائج سيئة كما هو موضح أدناه، لذلك لجأنا إلى خوارزمية العنقدة HDBSCAN وهي قادرة على إيجاد عناقيد بكثافات مختلفة.

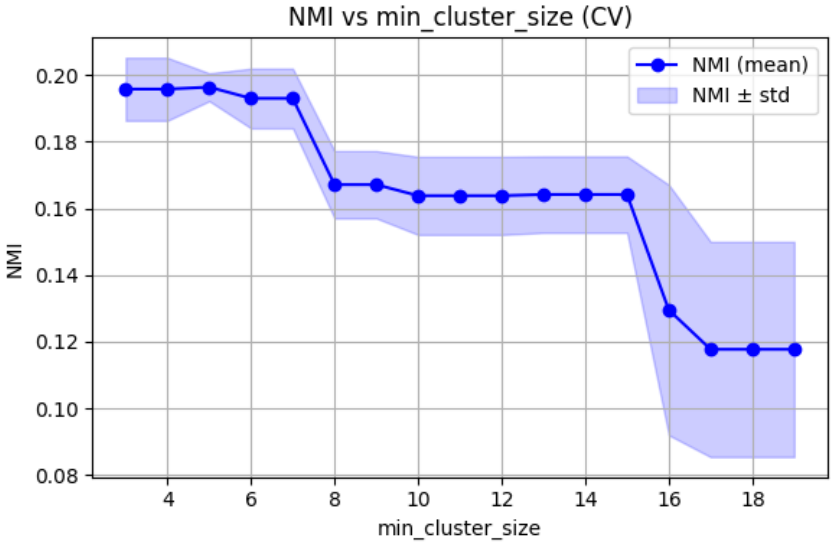
في هذه التجربة أخذنا خوارزمية العنقدة الهرمية HDBSCAN، وأخذنا كنماذج للتمثيل AraBert, mBert, Distilli mBert, LDA, Fasttext, Word2Vec.

وإنَّ موسطات هذه الخوارزمية التي تحتاج إلى ضبط هي العدد الأصغري للعينات min samples، والحجم الأصغري للعنقود min cluster size. وتم تحليل هذه الموسطات كل على حدى كما هو مبين أدناه:

ومن أجل تحليلها تم أخذ K Fold Cross Valdiation، من أجل k=5 وأخذ متوسط النتائج وانحرافها المعياري.

- أثر معامل حجم العنقود الأصغري

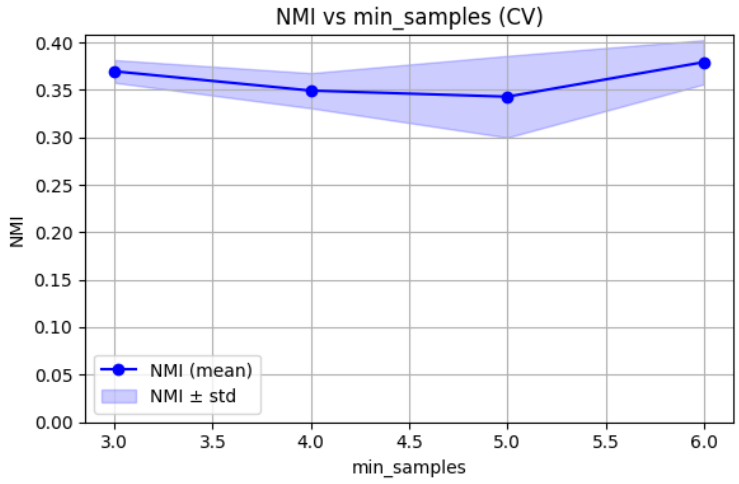
(كانت النتائج متشابهة لمعظم التجارب كشكل المنحني مع اختلاف القيم)، من أجل قيم صغيرة نسبياً لحجم العنقود (بين 3 و6) كانت النتائج متشابهة مع انحراف معياري صغير، ومن أجل قيم متوسطة له (أصغر من 14 كانت النتائج أقل ومع انحراف معياري صغير)، ومن أجل قيم كبيرة كانت النتائج سيئة ومع انحراف معياري كبير مما يدل على وجود ضبط زائد overfitting.



الشكل 45: أثر معامل حجم العنقود الأصغري على نتائج العنقدة.

- أثر معامل حجم العينات

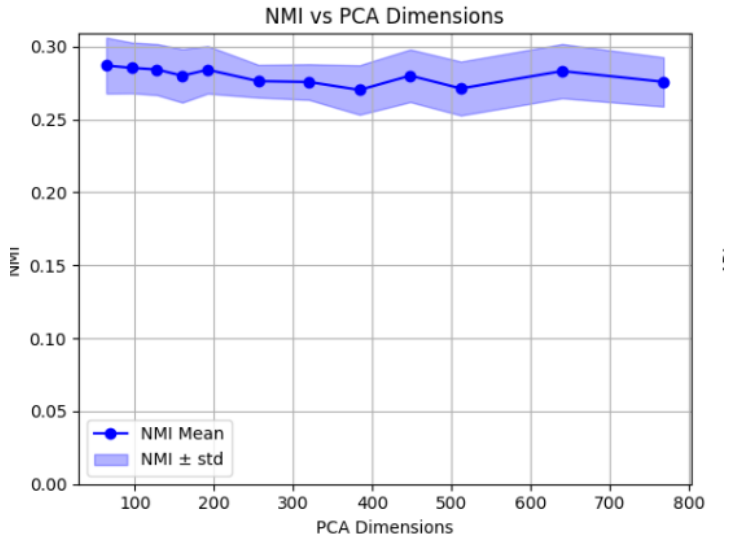
كانت النتائج متشابهة لجميع التجارب، ولم يكن لهذا المعامل أثر كبير على النتائج، حي كانت الناتئج مستقر مع اختلاف قيم هذ المعامل مع انحراف معياري أصغر للقيم الصغير (3,4)



الشكل 46: أثر معامل حجم العينات على النتائج.

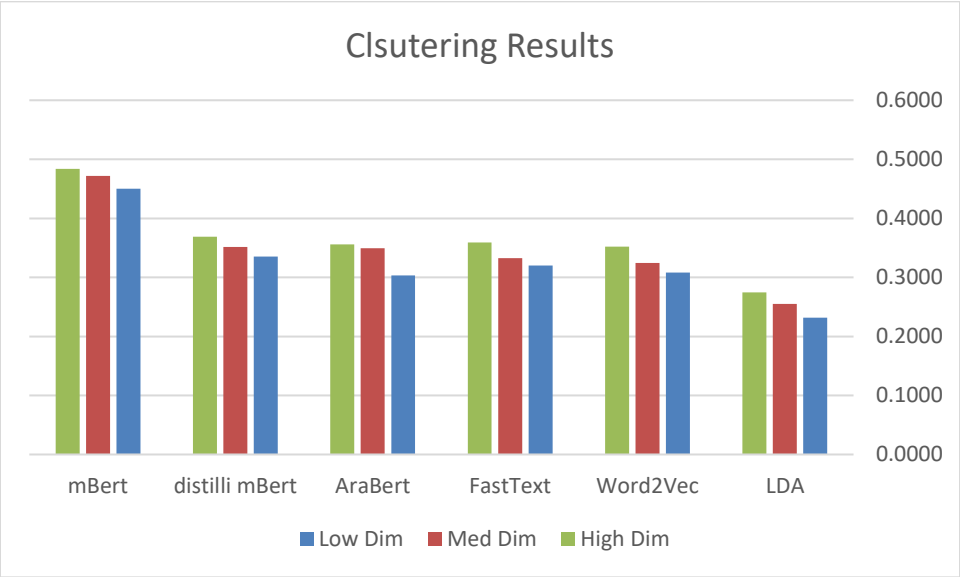
- أثر عدد الأبعاد في التمثيل على النتائج

على عكس التجربة السابقة، في هذه التجربة لم يكن لعدد الأبعاد أثر كبير على النتائج.



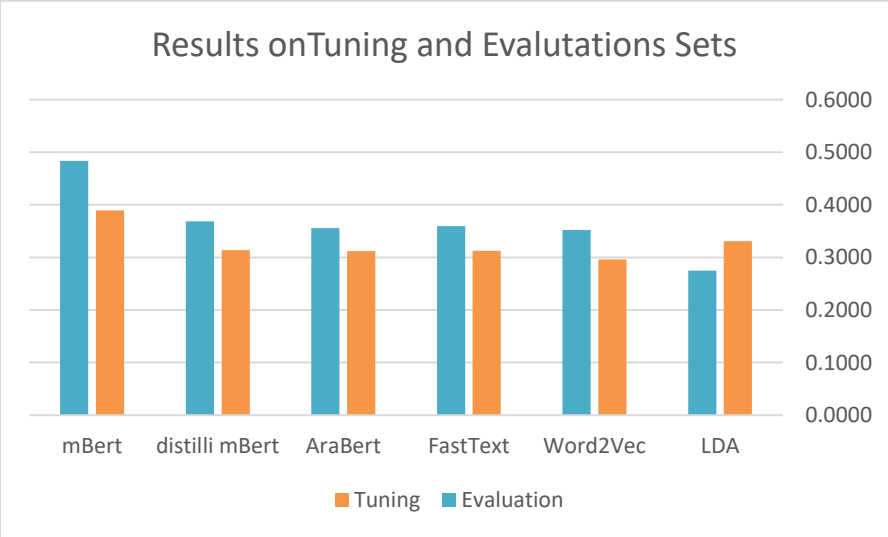
الشكل 47: أثر عدد الأبعاد على نتائج العنقدة.

وبعد تحليلها واختيار المجالات المناسبة للبحث في قيم المعاملات، تم ضبط هذه المعاملات من خلال البحث في تلك المجالات grid search، كان النموذج الأفضل للتمثيل هو mBert بجودة NMI 0.48، بتحسين عن الذي يلي بمقدار 25%، Distilli mBert بجودة NMI 0.36.



الشكل 48: نتائج العنقدة في التجربة الثانية.

ويوضح الشكل أدناه النتائج على مجموعة الضبط ومجموعة التقييم، حيث كانت النتائج متقاربة ولا يوجد تفاوت كبير في النتائج، وكانت النتائج على مجموعة التقييم أدق من مجموعة الضبط.

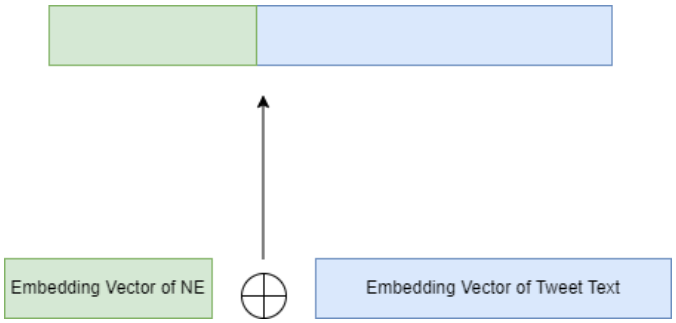


الشكل 49: الفرق بين نتائج الضبط والاختبار.

ويوضح الجدول (الملحق ت – نتائج التجربة الثانية في مرحلة كشف الأحداث) بقية المقاييس.

التجربة الثالثة: إضافة شعاع الكيانات المسماة

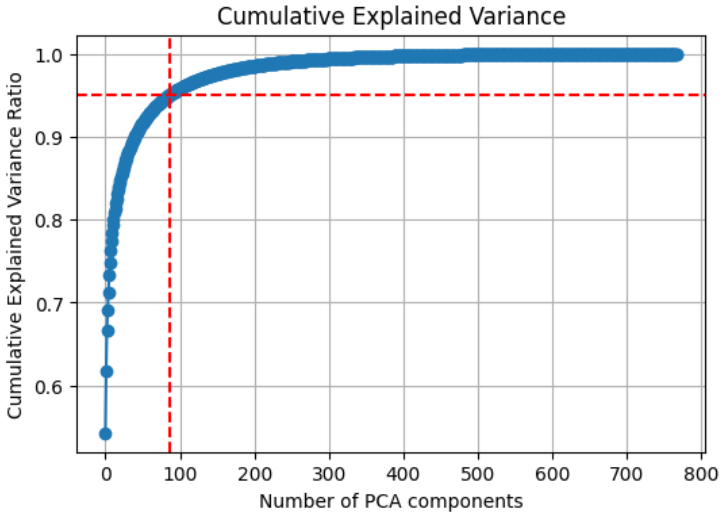
في التجربة الثانية قمنا بإضافة شعاع تمثيل الكيانات المسماة، إلى شعاع التمثيل الأصلي، وذلك بهدف جعل التمثيل يأخذ مكان الحدث والفاعلين فيه أثناء العنقدة، لتفادي مشكلة تجميع أحداث متشابهة في أماكن مختلف.



الشكل 50: مثال عن طريقة دمج شعاع تمثيل الكيانات المسماة.

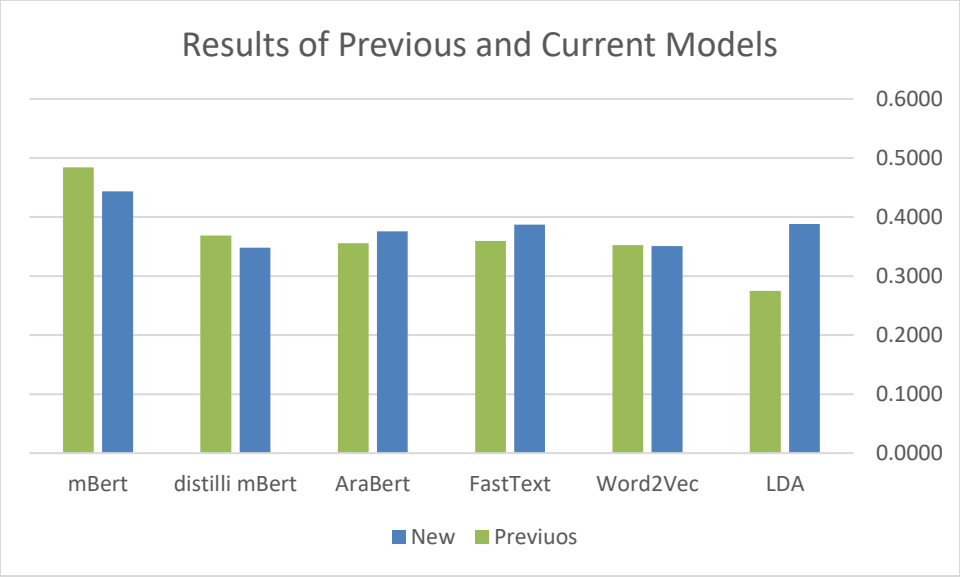
في هذه التجربة أخذنا نماذج التمثيل السابقة، وتم استخدام Stanza NER، لاستخلاص الكيانات المسماة المذكورة ي النص.

وتم استخدام خوارزمية تقليص الأبعاد PCA، لاختزال الأبعاد لشعاع تضمين الكيانات المسماة، وتم اختيار البعد الأمثل باستخدام طريقة الـ Explainable Variance كما هو مبين في الشكل أدناه.



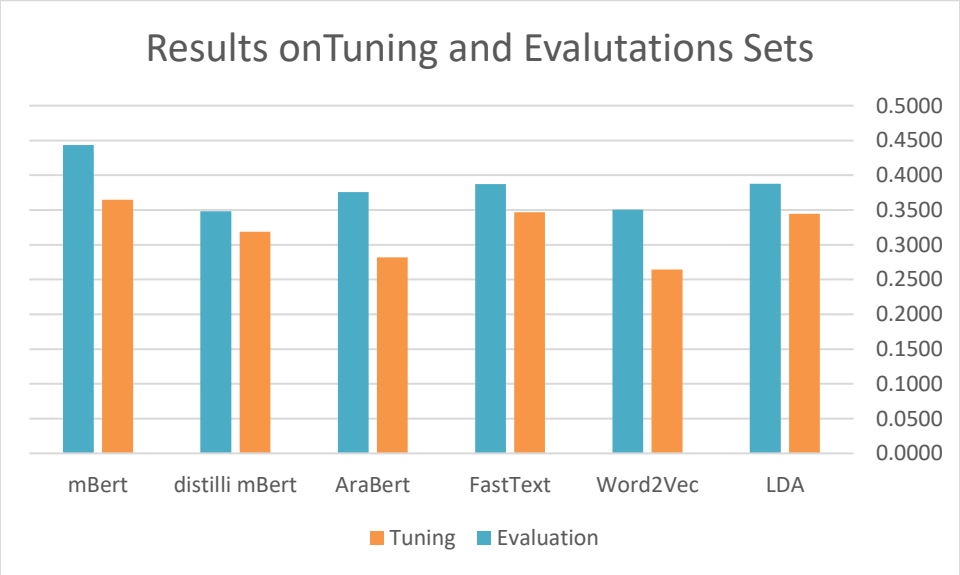
ويلاحظ أن عدد الأبعاد اللازم لتمثيل الكيانات المسماة أصغر بكثير من عدد الأبعاد اللازم لتمثيل الجمل.

في هذه التجربة تحسن النتائج لنماذج LDA Fasttext, Arabert, Word2Vec، حيث تحسنت جودة LDA من 0.26 NMI إلى 0.39 NMI، ولكن نموذج mBert لم يتحسن، وبقي هو أفضلهم.



الشكل 51: نتائج أخذ الكيانات المسماة في التضمين مقارنةً مع النتائج السابقة.

ويوضح الشكل أدناه الفرق بين نتائج الاختبار والتقييم.

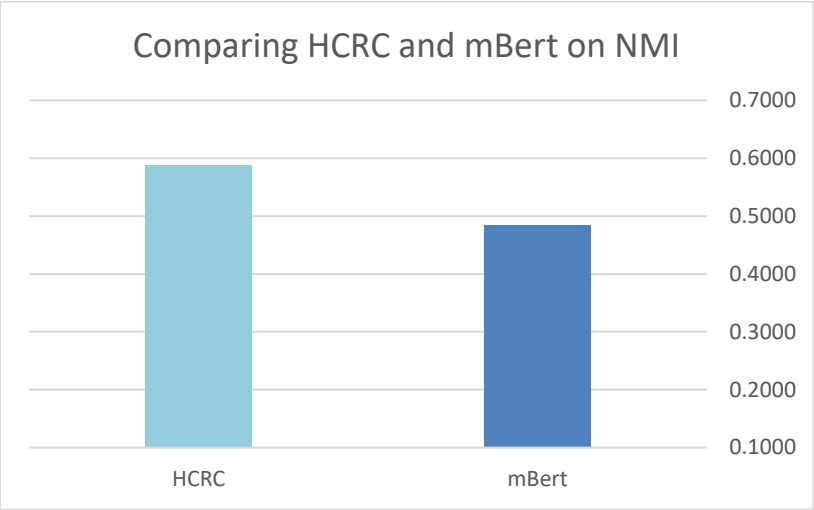


الشكل 52: الفرق بين نتاج الضبط والتقييم، في التجربة الثانية في مرحلة كشف الأحداث.

ويوضح الجدول (الملحق ت – نتائج التجربة الثالثة في مرحلة كشف الأحداث) بقية المقاييس.

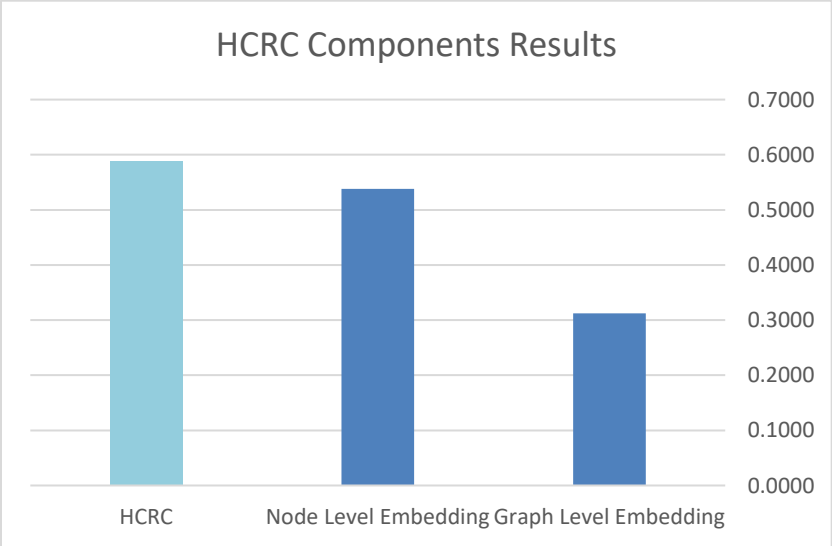
التجربة الرابعة: نموذج HCRC المرجعي

في هذه التجربة تم استخدام النموذج المرجعي HCRC، الذي يستخدم كخوارزمية عنقدة الخوارزمية SinglePass، وله مكونان يعطيان Graph Level Embedding وNode Level Embedding. وفي هذه التجربة قيمنا النموذج المرجعي على مجموعة البيانات خاصتنا، جربنا أيضاً تحليل مكونات هذا النموذج كل على حدى وإضافة مكونات أخرى. ويوضح الشكل أدناه التحسين الذي قدمه هذا النموذج مقابل أفضل نوذج وثلنا إليه، حيث كان مقدار تحسينه هو 17%.



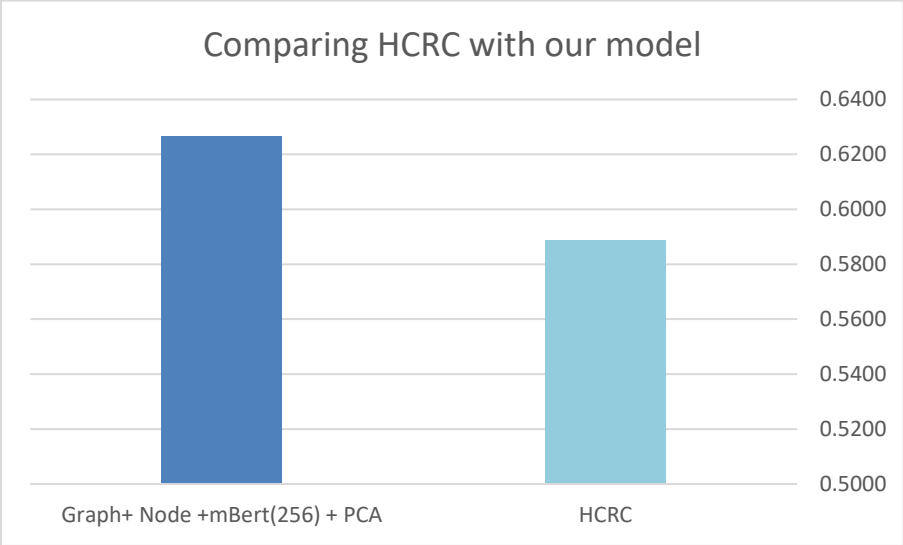
الشكل 53: مقارنة النموذج المرجعي مع النموذج الذي توصلنا إليه في المرحلة السابقة.

ويوضح الشكل أدناه نتائج مكونات كل جزء لوحده.



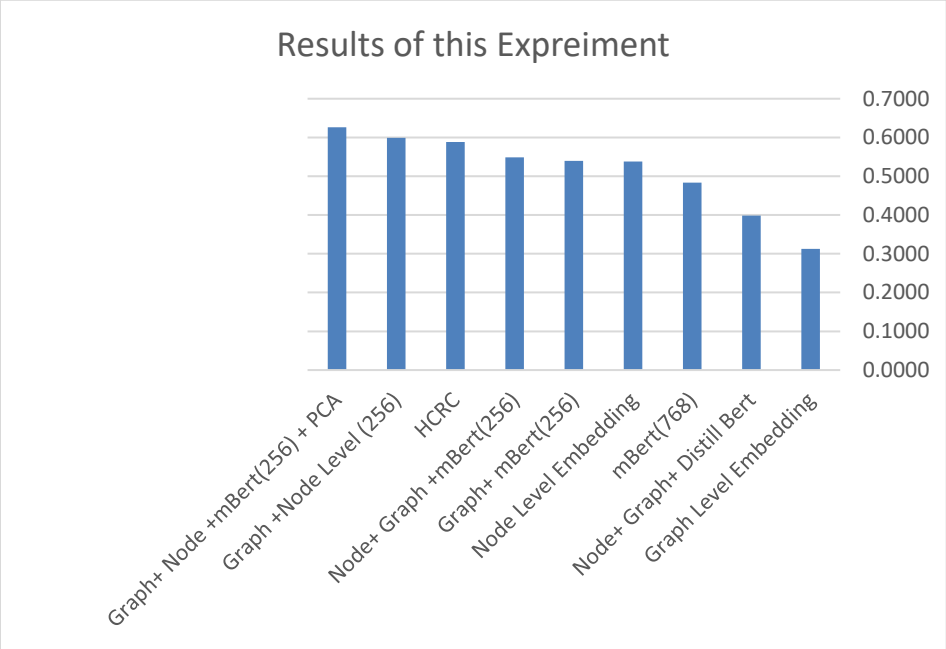
الشكل 54: نتائج تحليل النموذج المرجعي إلى مكوناته.

بعد ذلك تم تجريب أخذ Graph Level Embedding مع Node Level Embedding مع mBert Embeddding ومن ثم تقليص الأبعاد إلى 128باستخدام PCA وكانت النتائج هي 0.62 مقابل 0.58 للنموذج المرجعي.



الشكل 55: مقارنة النموذج المقترح مع النموذج المرجعي في الوضع **Offline**.

ويبين المخطط أدناه مقارنة النتائج التي اختبرناها في هذه التجربة.



الشكل 56: نتاج التجربة الرابعة في مرحلة كشف الأحداث.

ويبين الجدول (الملحق ت – النتائج التفصيلية للتجربة الرابعة في وضع عدم الاتصال، في مرحلة كشف الأحداث) بقية المقاييس.

2.3.8 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.

جدول**12** : إحصائيات كتل الرسائل.

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

Block	Date	Size	Events
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

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.

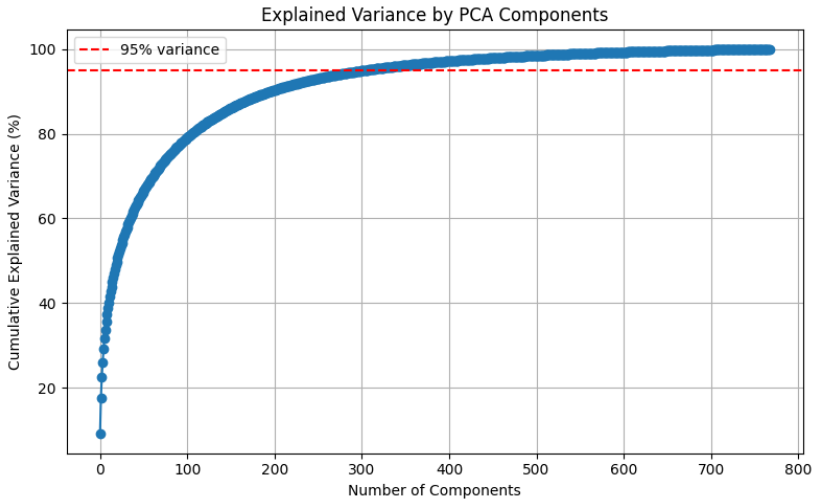
وأخذنا معايير التقييم هي NMI المعلومات المتبادلة المعيرة لمقارنة النماذج وتقييمها، كما قمنا بحساب المقاييس الأخرى AMI المعلومات المتبادلة المضبوطة، ARI مؤشر العشوائية المضبوط، وأرفقناه بالملحق، وهذه المعايير تم استخدامها في معظم الأبحاث المرجعية (المذكورون سابقاً). وأجرينا عدداً من التجارب، حيث بدأنا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

التجربة الأولى: التضمين باستخدام نماذج التضمين المرجعية

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 هو60. 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.

جدول**13** : نتائج التجربة الأولى.

	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%

(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

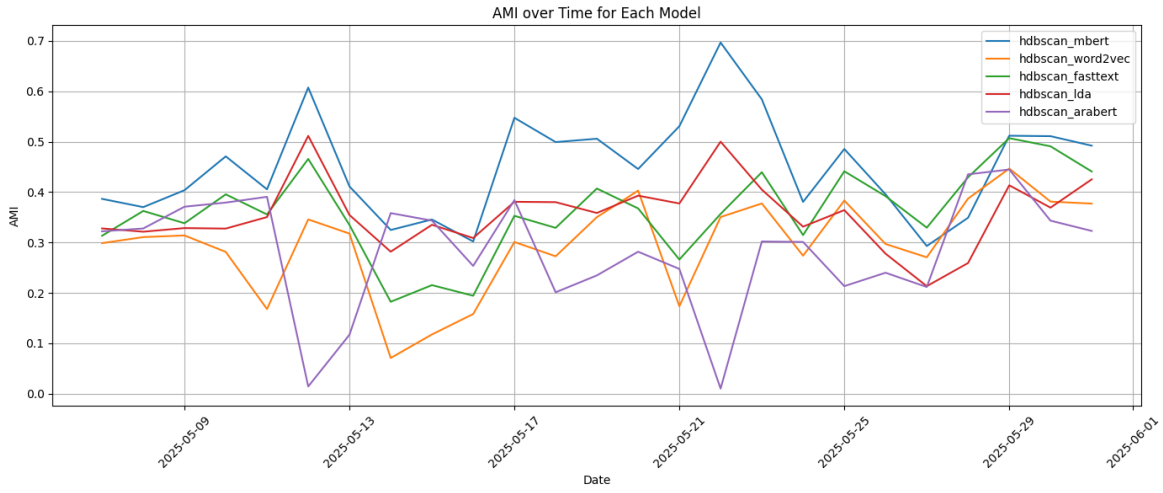
- تجميع الرسائل حسب معناها (أي مثلاً رسائل الحرائق تقع بالقرب من بعضها في فضاء التمثيل على الرغم من أن الحرائق تحدث في أماكن مختلفة، مما يؤدي إلى تداخل في الأحداث)
- كون العنقدة تحدث على كامل المجموعة كان هناك أحداث وقعت في نفس المكان ولكن بزمنين مختلفين (العثور على مستودع *** في ريف دمشق)
- كان هنالك 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 (الرسائل غير ذات الصلة كانت قليلة في الحدث الواحد، ولكن عند تجميع الأحداث كانت كثيرة)

وعليه:

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

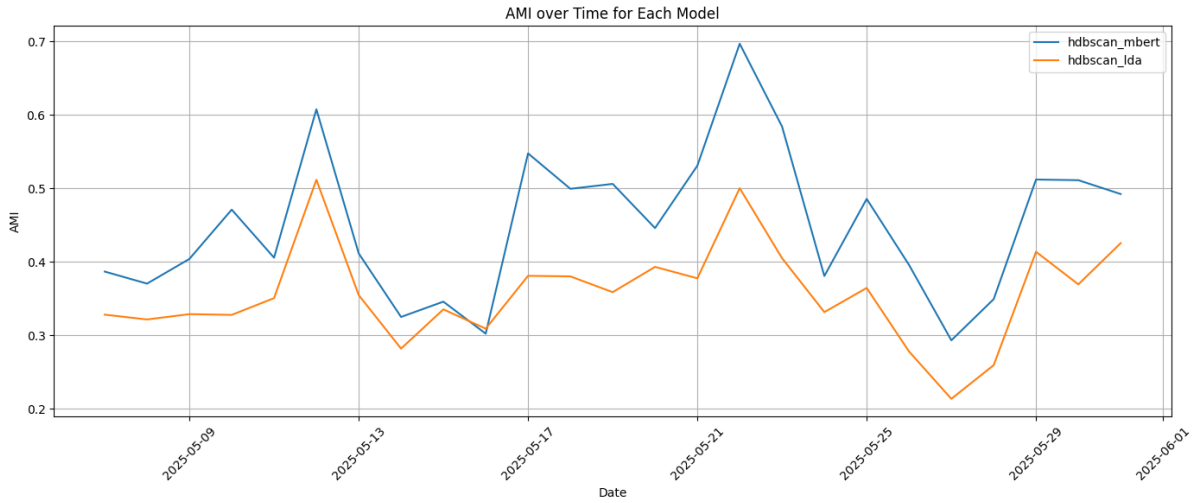
الأخطاء من النمط الثاني، يمكننا تفاديها عند أخذ الزمن بالاعتبار.

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

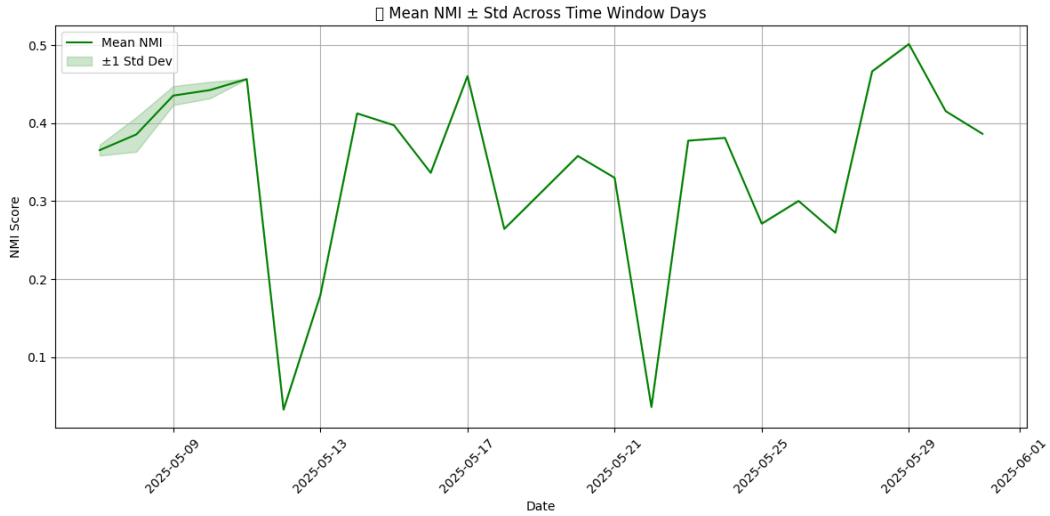


ويوضح الشكل أعلاه مخطط قيم NMI للنماذج المستخدمة، حيث وكما ذكرنا كان mBert هو الأفضل يليه LDA. وكانت نتائج بقية النماذج متقاربة.

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



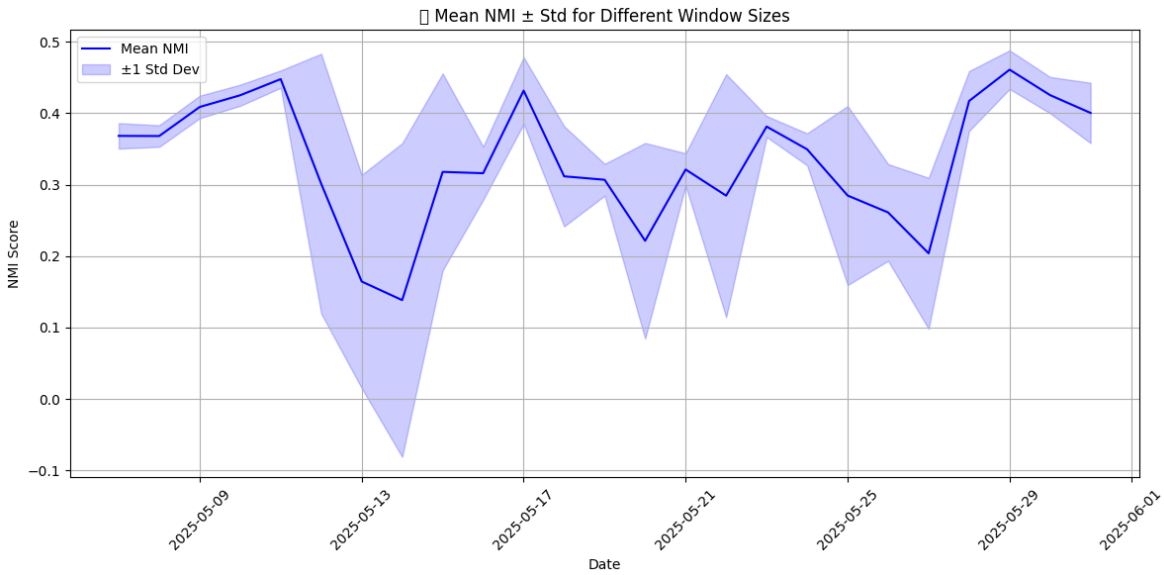
ويوضح الشكل أدناه أثر حجم النافذة الزمنية (معامل النسيان T)، على النتائج، حيث فمنا بالاختبار من أجل القيم (2 إلى خمسة أيام) وأخذنا متوسط القيم وانحرافها المعياري وكانت النتائج متقاربة جداً، حيث يوضح الشكل أدناه المتوسط والانحراف المعياري للنتائج لقيم مختلفة لـ T، حيث كانت قيم الانحراف المعياري صغيرة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



الشكل 57: أثر معامل حجم النافذة الزمنية على نتائج النموذج.

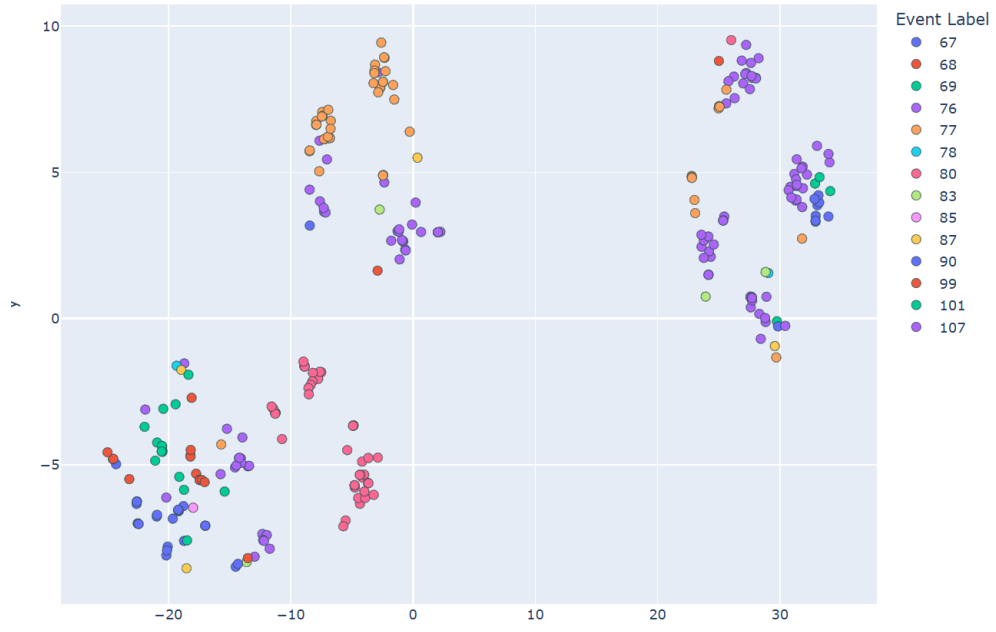
ويوضح الشكل التالي، أثر معامل حجم النافذة المنزلقة Sliding Window على النتائج، وكانت القيم الصغيرة لـ W (-400 200) تؤدي إلى نتائج متفاوتة، أي في الكتل الكبير تكون قيم nmi صغيرة جداً بسبب نسيان الأحداث الحالية بسبب كبر

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 58: أثر معامل حجم النافذة على النتائج.

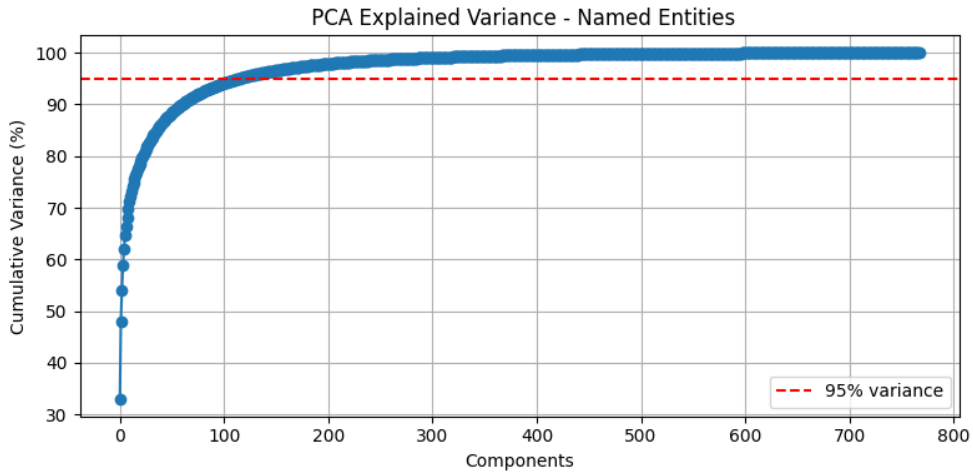
ويوضح الشكل أدناه أسقاط تمثيل الرسائل على فضاء ثنائي الأبعاد.



الشكل 59: 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.

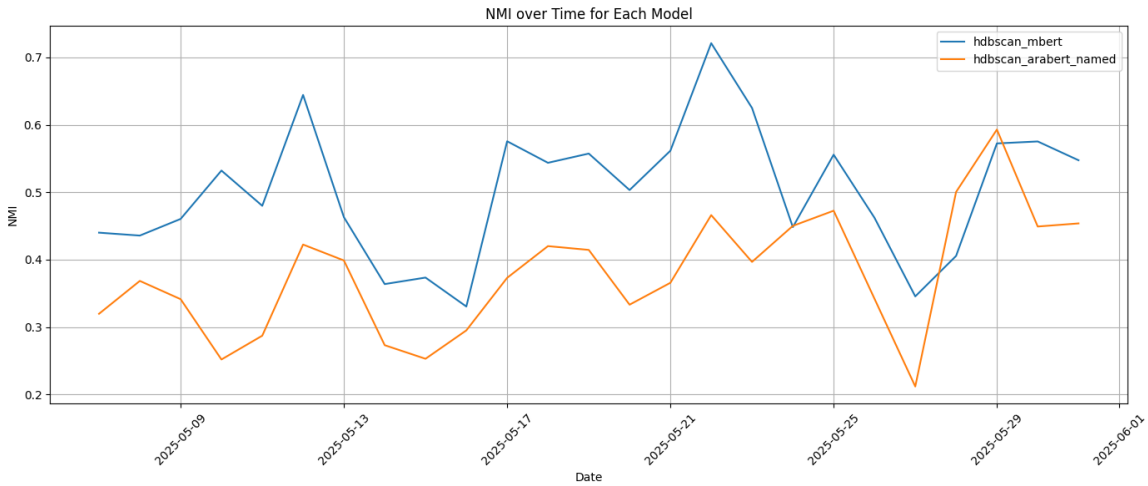


في هذه التجارب لم نحصل على قيمة مضافة على النتائج، حيث بقي mBert أفضل من AraBert + NE بـ 25%.

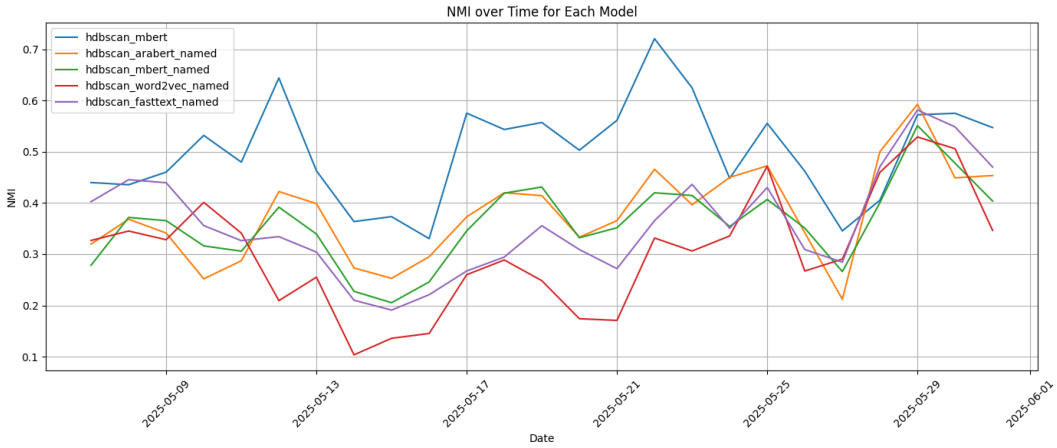
	M_1	M_2	M_3	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
mbert +NE	0.278	0.372	0.366	0.316	0.306	0.392	0.339	0.228	0.205	0.246	0.346	0.419
AraBert +NE	0.320	0.368	0.341	0.252	0.287	0.422	0.399	0.273	0.253	0.295	0.373	0.420
Word2Vec +NE	0.327	0.345	0.328	0.401	0.341	0.209	0.255	0.103	0.136	0.145	0.260	0.289
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%

ويوضح الرسم أدناه الفارق بينهما

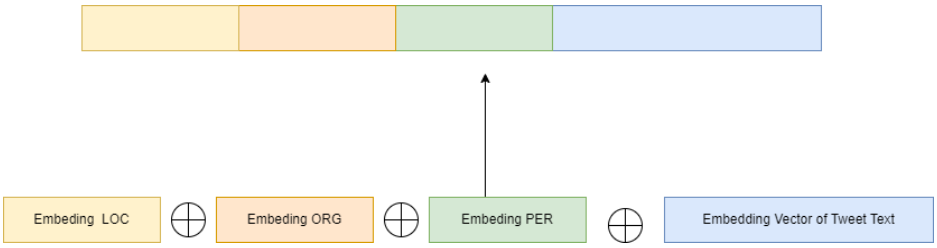


كما ويبين الشكل أدناه مقارنة الطرائق المتبعة في هذه التجربة مع mBert.



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.

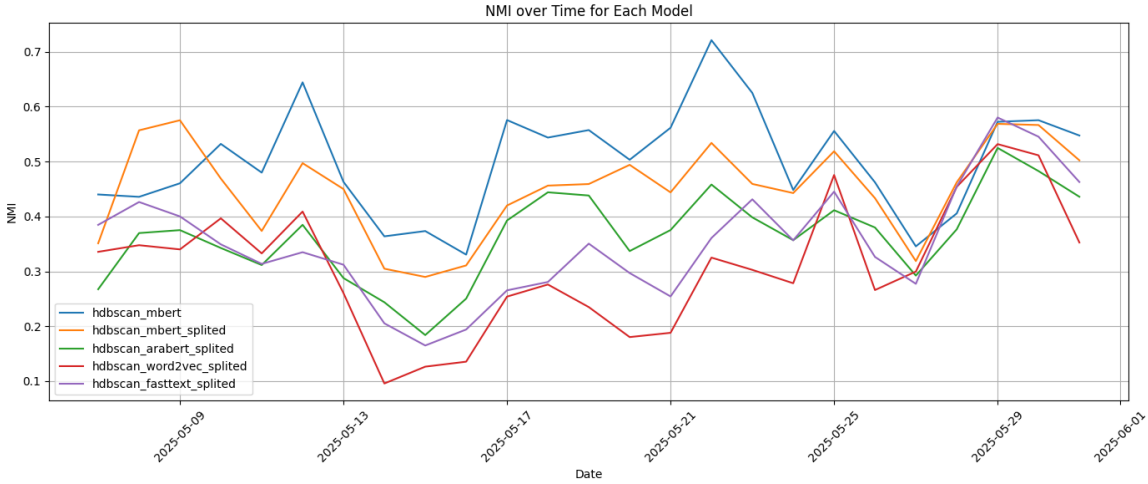
ويوضح الجدول أدناه النتائج التي حصنا عليها.

جدول**14** : نتائج التجربة الثالثة.

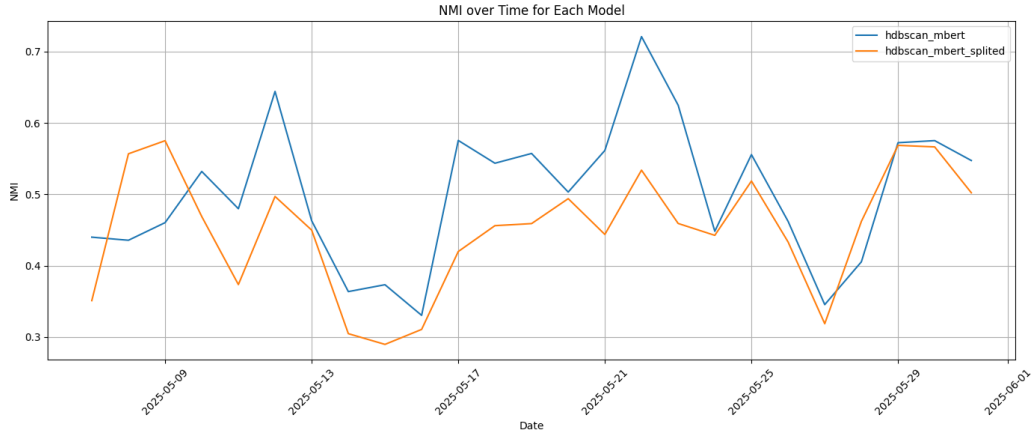
	M_1	M_2	M_3	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
mbert	0.351	0.557	0.575	0.469	0.374	0.497	0.450	0.305	0.290	0.311	0.420	0.456
AraBert	0.267	0.370	0.375	0.343	0.312	0.385	0.288	0.244	0.184	0.250	0.393	0.444
Word2Vec	0.336	0.348	0.340	0.397	0.333	0.409	0.260	0.096	0.126	0.135	0.254	0.276
FastText	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%

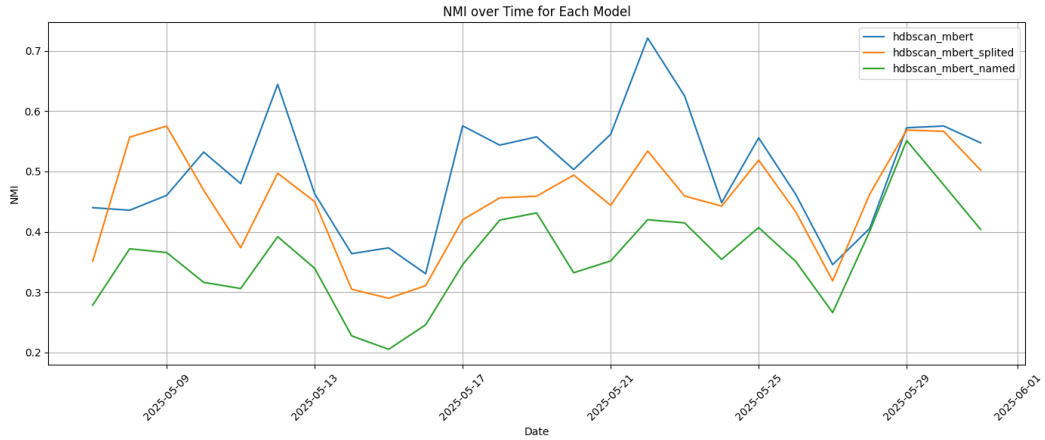
فيما يوضح الرسم البياني نتائج التجربة ومقارنتها مع mBert.



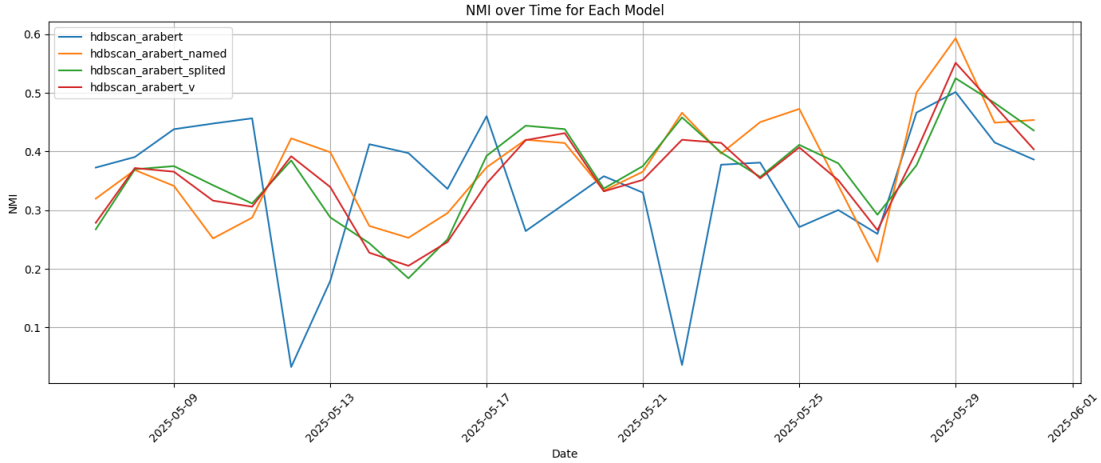
ويبين الرسم أدناه مقارنة mBert مع أفضلهم.



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



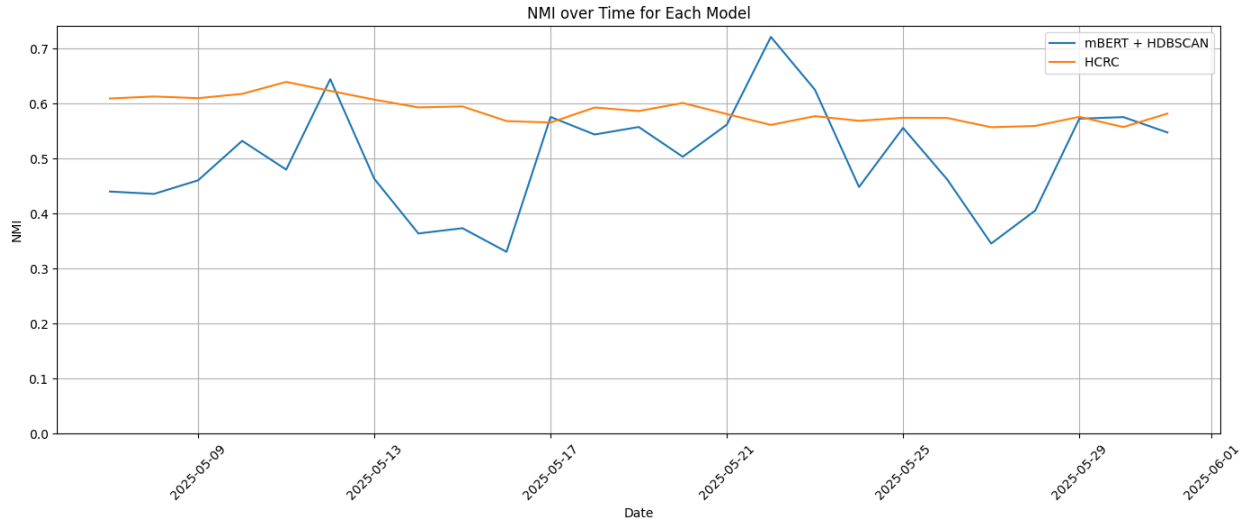
ويبين الشكل أدناه كيف أن الإضافات التي قمنا بها عل التمثيل جعلت التميل باستخدام arabert مستقراً أكثر.



التجربة الرابعة: النموذج المرجعي HCRC

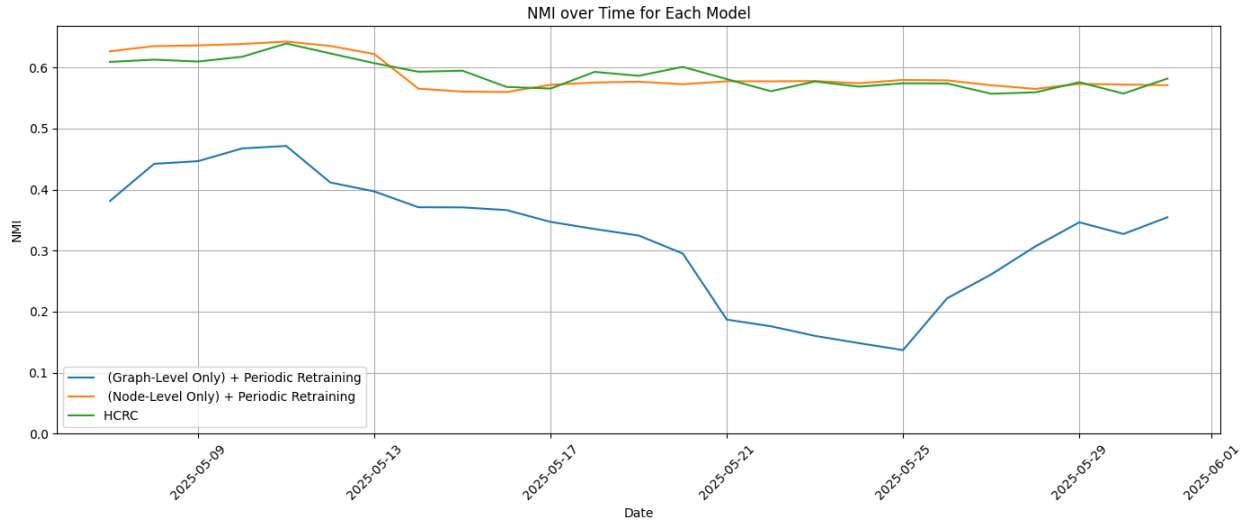
في هذه التجربة تم تقييم النموذج المرجعي ومختلف أجزاءه بالنمط الحي.

يبين الشكل أدناه الفرق بين أداء النموذج المرجعي HCRC مع أفضل نموذج كنا قد حصلنا عليه وهو mBert، حيث يلاحظ استقرار النموذج خلال الأيام، على عكس mBertالذي كانت نتائجه تتفاوت حسب الأيام



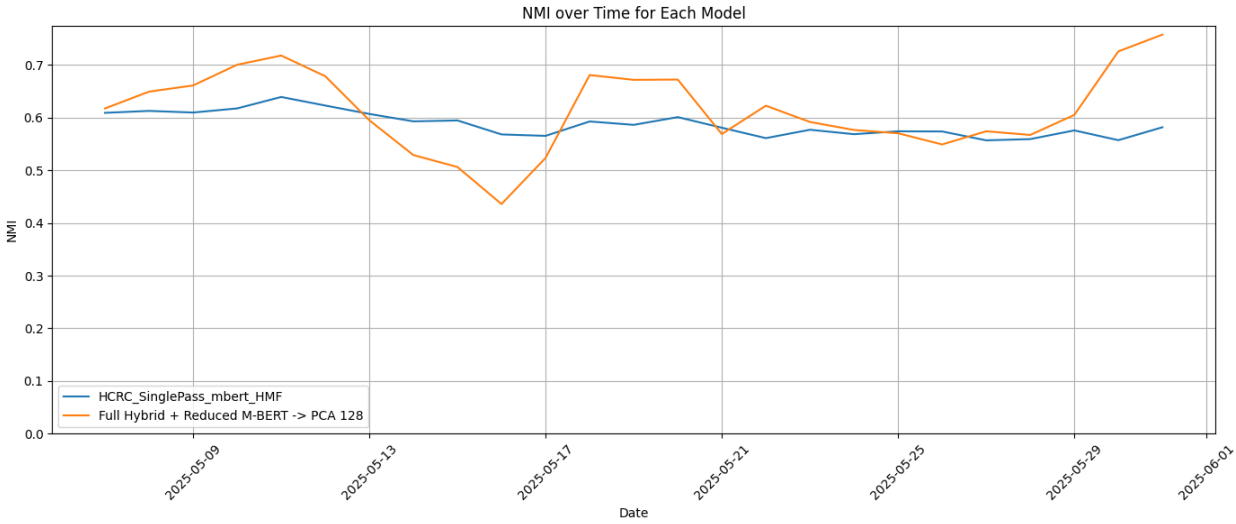
الشكل 60: مقارنة النموذج المرجعي مع أفضل نموذج ي التجارب السابقة.

ويوضح الشكل أدناه الفرق بين نتائج مكوناته

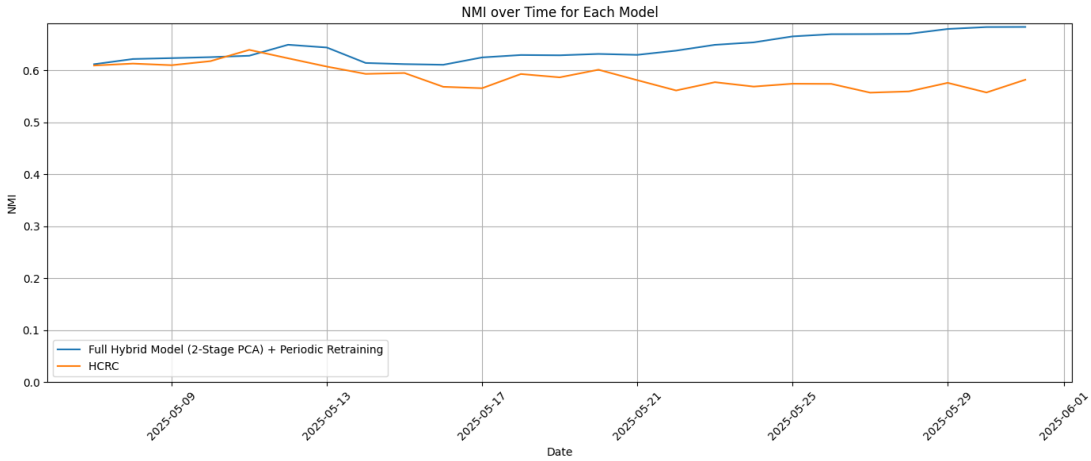


الشكل 61: نتائج تحليل النموذج المرجعي إلى مكوناته.

ويبين الشكل أدناه الفرق بين أداءه بدون القيام بعملية التدريب الدورية.

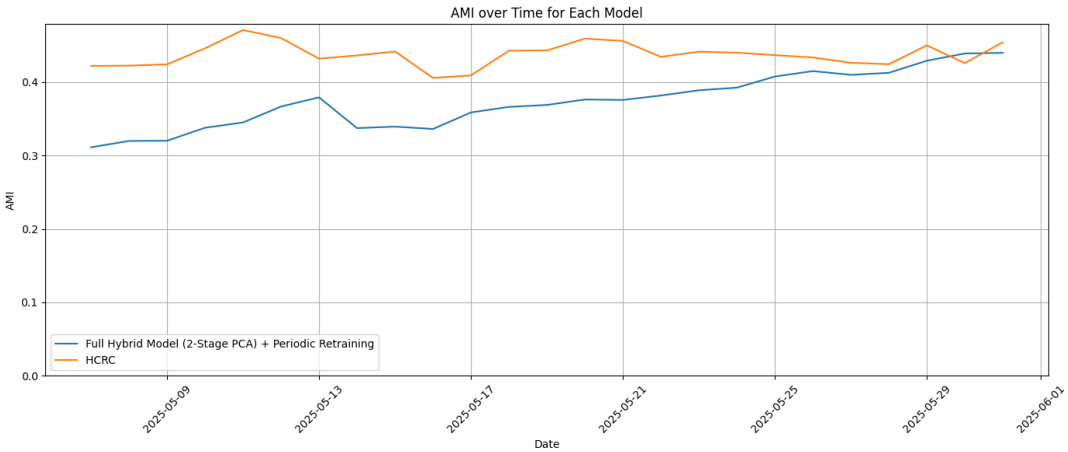


فيما يبين الشكل أدناه النتائج للنموذج Graph + Node + mBert + 2 PCA، مع تعديل عملية التدريب لتصبح دورية عوضاً عن أن تكون قبل كل يوم، مقابل النموذج HCRC، حيث نلاحظ أن النموذج المقترح تفوق بمقدار خمسة بالمئة على النموذج المرجعي وسطياً.



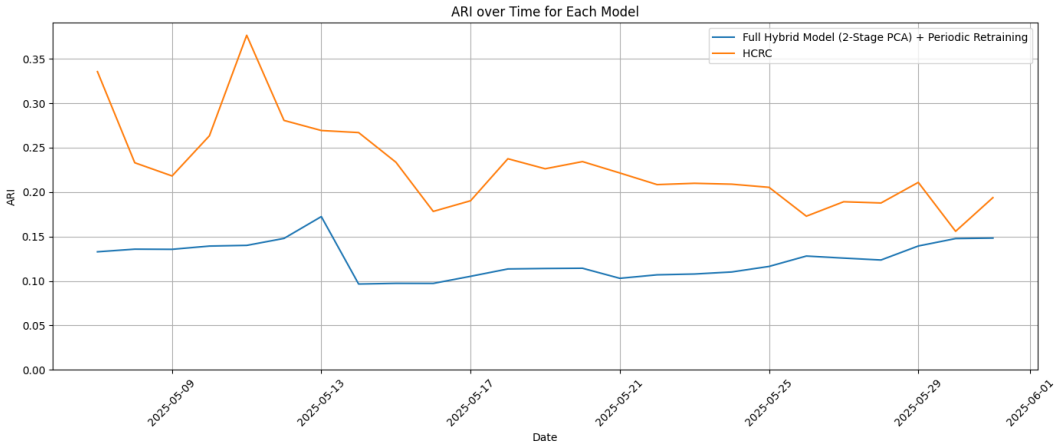
الشكل 62: مقارنة النموذج المقترح مع النموذج المرجعي وفق معيار **NMI**.

ويبين الشكل أدناه الفرق بين نتائج النموذج المقترح والنموذج المرجعي وفق مقياس المعلومات المتبادلة المضبوطة، حيث تفوق النموذج المرجعي بمقدار 6 بالمئة وهي مقاربة لما تفوق عليه في وضع التقييم Offline.



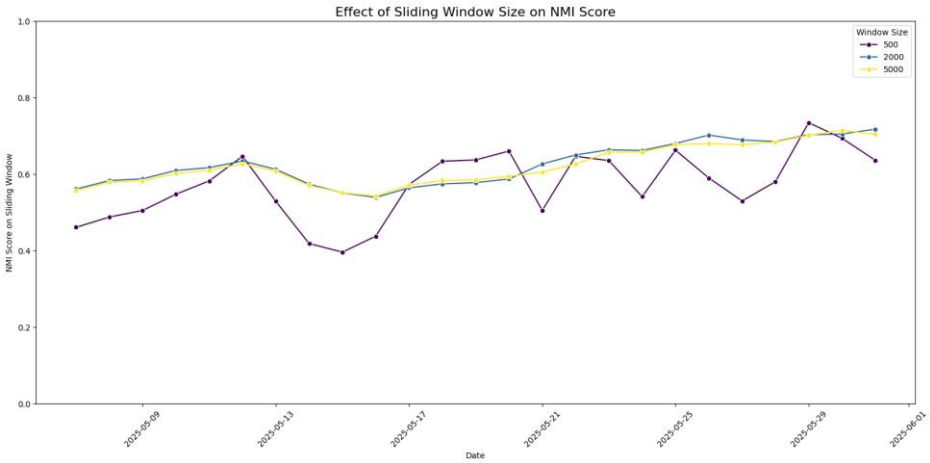
الشكل 63: مقارنة النموذج المقتر مع النموذج المرجعي وفق مقياس **AMI**.

ويبين الشكل أدناه الفرق بين نتائج النموذج المقترح والمرجعي وفق مقياس ARI، حيث تفوق النموذج المرجعي على النموذج المقترح بمقدار 10 بالمئة وسطياً، على الرغم من أنّ نموذجنا قد تفوق عليه في نفس المقياس في وضع التقييم Offline.



الشكل 64: مقارنة النموذج المقترح مع النموذج المرجعي وفق مقياس **ARI**.

ويبين الشكل أدناه أثر معامل حجم النافذ Window Size، على نتائج العنقدة وفق مقياس NMI. حيث نلاحظ أنه مع القيم الصغيرة لحجم النافذة كانت النتائج غير مستقرة، وكانت أكثر استقراراً مع زيادة حجم النافذة.



الشكل 65: أثر حجم النافذة على نتاج العنقدة.

ويبين الشكل أدناه أثر حجم النافذة الزمنية على نتائج العنقدة، ويلاحظ أن النتائج كانت مستقرة مع حجم نافذة سبعة أيام.

الشكل 66: أثر حجم النافذة الزمنية على نتائج العنقدة.

فيما يبين الجدول أدناه النتائج التفصيلية للنموذج المقترح ومقارنتها مع النموذج المرجعي.

جدول**15** : نتائج النموذج المقترح والنموذج المرجعي وفق مقياس **NMI**.

	M_1	M_2	M_3	M_3	M_4	M_5	M_6	M_7	M_8
Our	0.612	0.622	0.623	0.625	**0.628**	**0.649**	**0.644**	**0.614**	**0.612**
HCRC	**0.615**	**0.622**	**0.623**	**0.626**	0.627	0.628	0.613	0.575	0.571
Proportion	-0.6%	-0.1%	0.0%	-0.1%	0.1%	3.3%	4.8%	6.3%	6.6%

	M_9	M_{10}	M_{11}	M_{12}	M_{13}	M_{14}	M_{15}	M_{16}
Our	**0.611**	**0.625**	**0.629**	**0.629**	**0.629**	**0.631**	**0.630**	**0.638**
HCRC	0.569	0.585	0.589	0.589	0.589	0.586	0.582	0.586
Proportion	6.8%	6.4%	6.3%	6.3%	6.4%	7.3%	7.5%	8.1%

	M_{17}	M_{18}	M_{19}	M_{20}	M_{21}	M_{22}	M_{23}	M_{24}
Our	**0.649**	**0.654**	**0.665**	**0.669**	**0.670**	**0.670**	**0.679**	**0.683**
HCRC	0.579	0.570	0.573	0.564	0.557	0.569	0.578	0.577
Proportion	10.7%	12.9%	13.9%	15.8%	16.9%	15.1%	14.9%	16%

4.8 الاستدلال على الموقع الجغرافي

في هذه الجزئية من المسألة استخدمنا المحول المضبوط CaMel NER – MIX المطروح من قبل Suwaileh وآخرون في بحثهم للتعرف على المواقع المذكورة في النص، والذي نتائجه وفق ما هو مبين في الجدول أدناه (Suwaileh et al., 2023).

جدول**16** : نتائج المحول المضبوط امتخدم في التعرف على المواقع المذكورة.

Type	Precision	Recall	F1
Destination	0.914	0.914	0.914
Country	0.744	0.744	0.744
City	0.771	0.771	0.771
Neighborhood	0.856	0.856	0.856
Street	0.93	0.93	0.93
Natural Point of Interest	0.925	0.925	0.925

ومن أجل تحديد الإحداثيات الجغرافية للموقع، تم اتباع التقديرية التالية.

- التعرف على المواقع المذكورة في النص وفق المحول المذكور أعلاه.
- استخدام Gazetteer للتعرف على أسماء المدن والدول المذكورة.
- استخدام مرمز وتمرير الثلاثيات من النمط دولة، مدينة، موقع مذكور من التغريدة،
- أخذ أقرب ثلاث إحداثيات مرجعة.

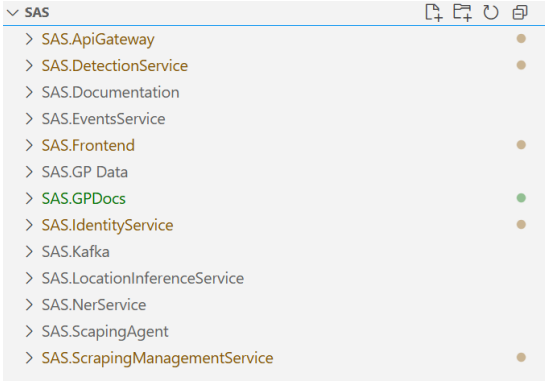
الفصل التاسع

تنجيز النظام

يعرض هذا الفصل كيفية تنجيز النظام.

1.9- 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). ويوضح الشكل أدناه الخدمات المنجزة، حيث كل خدمة منفصلة ولها مستودع خاص بها واختباراتها وخط أنابيب لها.



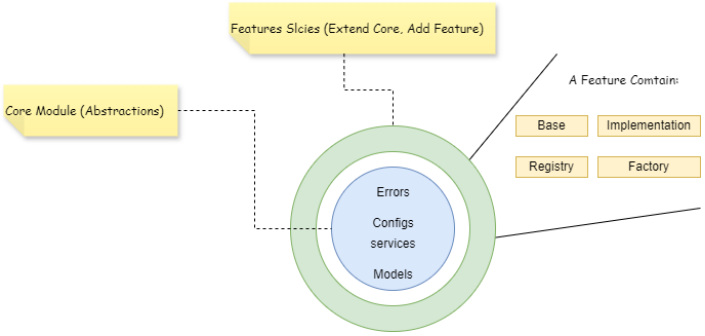
2.9- البنية التصميمية العامة للخدمات

عند تنجيزنا لخدمات النظام، اعتمدنا بنية معمارية موحدة للخدمات المتشابهة، واعتمدنا مجموعة من المبادئ والأنماط التصميمية التي تساعد على جعل الخدمات قابلة للتوسع، وسهلة الصيانة.

حيث عند تنجيز الخدمات المطورة بلغة Python، أملينا على أنفسنا مجموعة من القواعد، لبناء خدمة قابلة للتوسع وإعادة الاستخدام بسهولة، حيث 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. أي عندما نريد إضافة ميزة جديدة أو تعديل ميزة موجودة، فسنجد معظم المكونات، والملفات التي ستتغير بجانب بعضها.

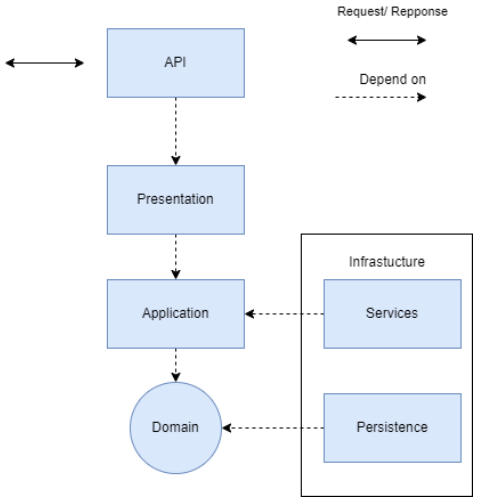
بحيث، في كل مشروع يكون هنالك مكون يحتوي على تجريدات Abstraction، وخدمات عامة، هو نواة المشروع، الذي يحوي على الخدمات العامة التي ستحتاجها بقية الخدمات، والأنماط، وملفات التكوين، رسائل الأخطاء. ويوجد بجانبه مجموعة من المجلدات كل واحد منها عبارة عن ميزة (سمة) متكاملة مترابطة فيما بينها.

وعندما نريد إضافة ميزة ما، فإن كل ما علينا هو إضافة مجلد باسمها، إضافة تجريد لها ضمن مجلدها، وكتابة التنجيزات الممكنة لها، والخدمات التي تقدمها، وتعريف سجلها، بالإضافة إلى تعريف آلية بناءها. ويوضح الشكل أدناه هذه البينية.



الشكل 67: مثال على النبية العامة للخدمات المطورة بلغة **Python**.

وبالنسبة للخدمات المعتمدة على إطار العمل ASP.NET، تم اعتماد البينة المعمارية النظيفة وفق البنية الموضحة بالشكل أدناه. حيث يحوي المشروع على عدد من الطبقات وهي، (1) طبقة العرض التي تعرف نقاط التواصل مع النظام Endpoints، عقود الطلبات والردود، (2) طبقة التطبيق التي تعرف الخدمات التي يقدمها النظام (حالات الاستخدام) بالإضافة إلى عقود الخدمات التي يعتمد عليها النظام، (3) طبقة المجال التي تحتوي منطق العمل، والكيانات، والأخطاء التي تصدر عن التطبيق، وعقود المستودعات. (4) طبقة الخدمات وهي الطبقة التي تنجز الخدمات التي تعتمد عليها طبقة التطبيق (طبقة التطبيق تعتمد على عقود خدمات يحتاجها النظام، تنجزها طبقة التطبيق) (5) طبقة المزامنة (التخزين عل قواعد المعطيات) في هذه الطبقة يتم تنجيز عقود المستودعات، بالإضافة إلى منطق قواعد المعطيات والوصول إليها. ويغلف كل هذه الطبقة طبقة واجهة تخاطب البرمجيات API، التي تحقن الاعتماديات وتعرف الـ middleware's.



الشكل 68: البنية العامة للخدمات المنجزة بإطار العمل **Asp.Net**.

3.9- تنجيز الخدمات

خدمة سحب البيانات

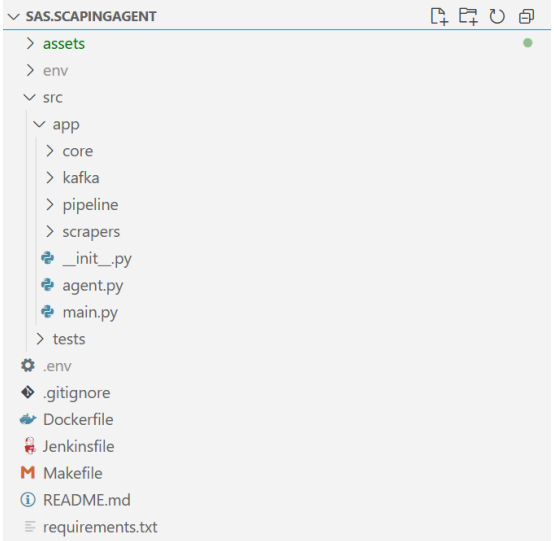
مهمة هذه الخدمة هي، سحب البيانات من منصات وسائل التواصل الاجتماعي بن خلال زواحف (الدخول من خلال حسابات شخصية والبحث في المنصات)، التواصل مع الـ APIs التي تقدمها منصات وسائل التواصل لسحب البيانات. وبعد سحب كل دفعة من البيانات يتم معالجتها وفلترتها ومن ثم إرسالها للخدمة التالية.

تم تنجيز هذه الخدمة باستخدام إطار العمل Flask، باستخدام لغة Python، تم اختيار Python، للأسباب التالية:

- توفر لغة بايثون مجموعة من الأدوات التي تسهل عملية بناء الزواحف، وتحليل محتوى صفحات الوب، بالإضافة إلى أن منصات وسائل التواصل توفر مكتبات عديدة للتواصل مع الـ APIs الخاصة بها مما يسهل عملية التكامل معها.
- ستقوم هذه الخدمة، بعد جمعها للبيانات، بإجراء مراحل من معالجة النصوص، وتقنيات معالجة اللغات الطبيعية NLP، على البيانات التي تجمعها بالإضافة إلى أنها ستحتاج إلى مجموعة من نماذج تعلم الآلة، بما أن بايثون هي البيئة الأساسية لتشغيل تقنيات الذكاء الصنعي (مثل Transformers، Spacy، Scikit-learn)، لذلك كانت هذ الخيار الأقوى هنا.

تفاصيل التنجيز

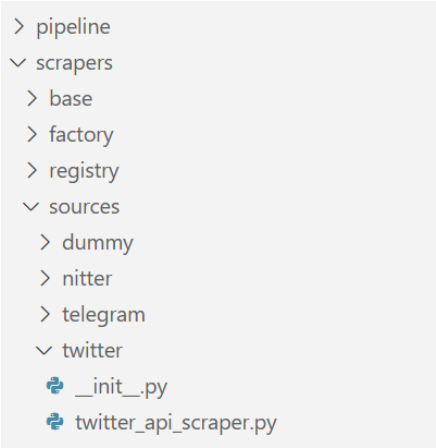
يوضح الشكل أدناه، بنية هذه الخدمة، حيث تحتوي ضمن مجلد الرماز المصدري على أربع مجلدات فرعية ضمنها (أربع شرائح)، أولها مجلد النواة Core، ومجلد الـ Pipelines الذي يعرف مراحل المعالجة اللاحقة للبيانات المسحوبة، بالإضافة إلى مجلد الـ Scrapers الذي يعرف طرائق سحب البيانات، ومجلد Kafka الذي يعرف آلية التراسل مع خدمة إدارة سحب البيانات.



الشكل 69: بنية خدمة سحب البيانات.

وفي مجلد الـ Core والذي يحوي على مجموعة من الخدمات العامة التي تحتاجها بقية الشرائح، مثل سحب المهمات، استخراج الكيانات المسماة، التسجيل logging، تضمين الرسائل. كما يحوي رسائل الخطاء والاستثناءات التي تصدر عن الخدمة.

الشريحة الثانية في هذه الخدمة، هي ساحبي البيانات Scrapers، حيث تحوي على تجريد لعقد سحب البيانات، وعدد من التنجيزات المختلفة له (سحب من تيليغرام، ....)، بالإضافة إلى ذلك فهي تحوي على سجل للتنجيزات المتخلفة لعقد سحب البيانات، وآلية لأخذ تنجيز انطلاقاً من المهمة الموكلة إليه Factory. ويبين الشكل أدناه بنية الشريحة.



الشكل 70: بنية شريحة ساحبي البيانات.

والشريحة الثالثة هنا هي الخاصة بخط أنابيب المعالجة، حيث بعد سحب دفعة من البيانات Messages Batch، يتم تمرير هذه الدفعة إلى خط أنابيب، للمعالجة المسبقة، حيث يوجد فيه عدد من المراحل (يستطيع مدير النظام اختيار ما يشاء) فمثلا يتم فيه تطبيع الرسائل Messages Normalization، فلترة الرسائل التي تحوي على رسائل غير مرغوب بها، فلترة الرسائل التي تحوي على خطاب كراهية، تحليل المشاعر، .... إلخ. وبعدها يتم تفعيل خط أنابيب الرفع، حيث يتم فيه إرسال الرسائل إلى الخدمة التالية، إعلام خدمة الإدارة بالكيانات الأكثر تكراراً. وتتكون هذه الشريحة من تعريف لمعالجة البيانات، ومراحل المعالجة المتاحة، ويتم تسجيل هذه المراحل في السجل، ويتم من خلال الباني، بناء خط أنابيب من مجموعة المراحل.

ويبين الشكل أدناه، رماز يبين آلية استخدام الشرائح السابقة وخدماتها، حيث يتم أخذ المهمات من خدمة الإدارة، ومن ثم إنشاء آلية سحب البيانات، حسب المهمة المسندة (مثل، مهمة سحب بيانات من تويتر => ساحب بيانات من تويتر). بعد ذلك يتم إسناد المهم إلى الوكيل، الذي يأخذ دفع البيانات التي يسحبها الـ scraper ويطبق عليها خطوط الأنابيب.

```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()
```

الشكل 71: عينة من الرماز، تبين آلية استخدام الأدوات التي نجزناها.

وتين الصورة أدناه، كيف أن الخدمة تقوم بقراءة مراحل المعالجة المطلوبة على البيانات (في وقت التشغيل) ومن ثم الاتصال مع خدمة الإدارة لتسجيل نفسها (لكي تأخذ بعين الاعتبار أثناء توليد المهام)، ومن ثم تبدأ باستقبال المهمات وتنفيذها.

```
(env) PS D:\SAS\SAS.ScapingAgent\src>
(env) PS D:\SAS\SAS.ScapingAgent\src> python -m app.main
Downloading stanza models for language: ar
[INFO] Register Stage : NormalizeTextStage
[INFO] Register Stage : SentimentAnalysisStage
[INFO] Register Stage : HateSpeachFilteringStage
[INFO] Register Stage : NamedEntitiesExtractionStage
[INFO] Register Stage : KeywordFilterStage
[INFO] Register Stage : EmbeddingStage
[INFO] Register Stage : MessagesClasificationStage
[INFO] Register Stage : MessagesPublishingStage
[INFO] Register Stage : MessagesSavingStage
[INFO] Register Stage : MasterNotifyingStage
[INFO] Register Stage : FeedbackStage
[INFO] Connected to master, scraper ID: afc64b78-d74f-4c9e-9b5c-652d7b29717f
[INFO] Assigned scraper afc64b78-d74f-4c9e-9b5c-652d7b29717f to task 8f2ef1ce-a9f4-40c9-aad2-8995c2efe01f successfully.
[INFO] Extracted named entities for the task :36 messages.
[INFO] Loaded 1 normalized blocked keywords from server.
[INFO] Embeddings generated successfully.
[INFO] Starting classification with 5 messages
[INFO] Filtered out 2 messages.
[INFO] Published batch to topic Telegram.Politics: 3 messages
[INFO] First message in batch: 'عم و يرس أ ل د ايت ل يح أت...' with label=positive, score=0.756
[INFO] Task 8f2ef1ce-a9f4-40c9-aad2-8995c2efe01f marked complete successfully.
[INFO] Assigned scraper afc64b78-d74f-4c9e-9b5c-652d7b29717f to task a2ad33b9-58c7-4524-9113-bf66737ef769 successfully.
[INFO] Extracted named entities for the task :36 messages.
[INFO] Embeddings generated successfully.
[INFO] Starting classification with 5 messages
[INFO] Filtered out 2 messages.
[INFO] Published batch to topic Telegram.Politics: 3 messages
```

الشكل 72: عينة من خرج هذه الخدمة.

كما تتواص هذه الخدمة مع قاعدة بيانات من نوع MongoDB، لتقوم بتحميل إعدادات الخدمة منها، بالإضافة إلى قراءة اعتماديات الدخول Credential إلى منصة التواصل (اسم المستخدم وكلمة المرور، API Key، ....)، كما هو مبين أدناه. وتم اختيار MongoDB، لأننا كنا بحاجة إلى قاعدة بيانات غير علائقية لتخزين إعدادات الخدمة، واعتماديات الدخول.

الأنماط التصميمة المستخدمة

في هذه الخدمة تم استخدام النمط التصميمي Pipes & Filters، لإضافة مراحل المعالجة اللاحقة على الرسائل، بحيث يتم إضافة مراحل المعالجة والفلترة المتاحة، ويتاح للمدير النظام في وقت التشغيل تهيئتها وترتيبها واختيار ما يناسبه منها. وتم استخدام نمط السجل Registry، بحيث يتم فيه تسجيل مراحل المعالجة المتاحة، وباستخدام نمط الـ Factory يتم بناء خط أنابيب المعالجة، وكذلك تم استخدام السجل أيضاً لتسجيل تنجيز سحب البيانات المناسب لكل نوع من المهمات، بحيث أيضاً في وقت التشغيل يتم إنشاء آلية السحب بناءً على المهمة.

خدمة إدارة عمليات سحب البيانات

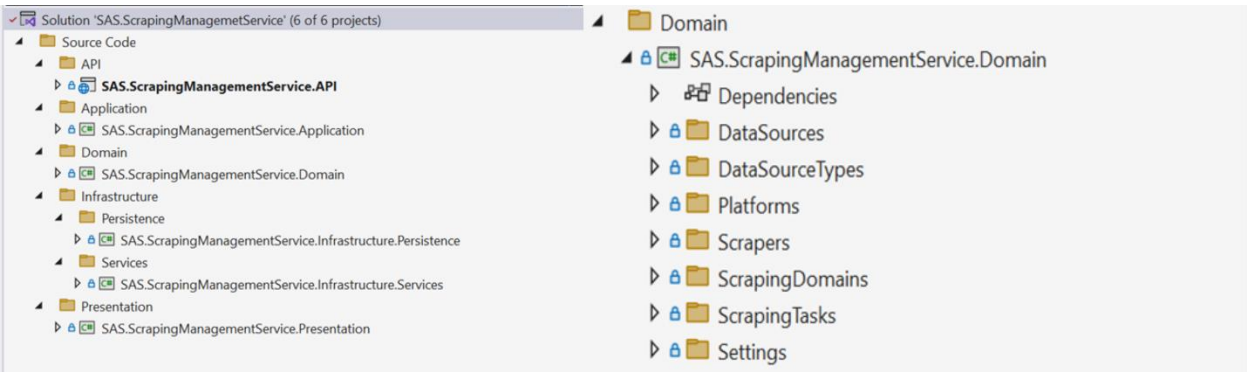
مقدمة

مهمة هذه الخدمة هي إدارة مصادر البيانات التي يتم السحب منها، إدارة مهمات السحب (جدولة المهمات حسب مصادر البيانات المتاحة وخدمات السحب المتصلة)، بالإضافة إلى تلقي الـ feedbacks من خدمات سحب البيانات Scraping Agents وتغذية مصادر البيانات. وتم تنجيز هذه الخدمة باستخدام إطار العمل Asp.Net، الذي يعد من أفضل أطر العمل خصوصاً في المشاريع الكبيرة التي تحتوي على منطق عمل كبير ويجب معالجته بطريقة قابلة للتوسع ومنظمة، بالإضافة إلى دعمه للخدمات التي تعمل بالخلفية (خدمة الجدولة وخدمة تلقي الـ Feedbacks). بالإضافة إلى قاعدة البيانات العلائقية SQL Server، لسهولة تكاملها مع إطار العل هذا، ولحاجتنا إلى قاعدة بيانات علائقية.

نجزت الخدمة باستخدام البنية المعمارية النظيفة، بسبب وجود منطق عمل معقد، وقابل للتغير، لذلك تم استخدام البنية المعمارية النظيفة، التي ساعدتنا في فصل التنجيز الفعلي عن عقوده، بحيث يتم التعامل مع عقود العمليات أثناء كتابة منطق العمل.

تفاصيل التنجيز

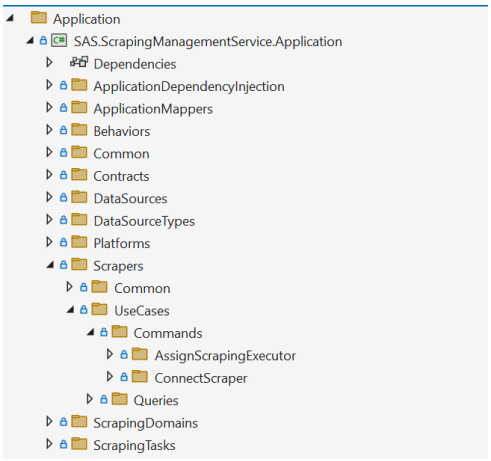
يبين الشكل أدناه المشاريع ضمن هذه الخدمة، وتتكون هذه الخدمة من عدد من المجالات هي: مصادر البيانات وأنواعها، مهمات سحب البيانات، مجالات السحب، والمنصات.



الشكل 73: بنية خدمة إدارة عمليات سحب البيانات.

- تحوي طبقة المجال كما هو موضح أعلاه، على مجالات هذه الخدمة، وفي كل منها، يوجد خمسة مجلدات فرعية، مجلد الكيانات ويحوي على الكيانات الموجودة ضمن هذا المجال، ومجلد المستودعات الذي يعرف عقود تخزين البيانات، ومجلد الأخطاء الذي يعرف الأخطاء التي تصدر عن هذا المجال.

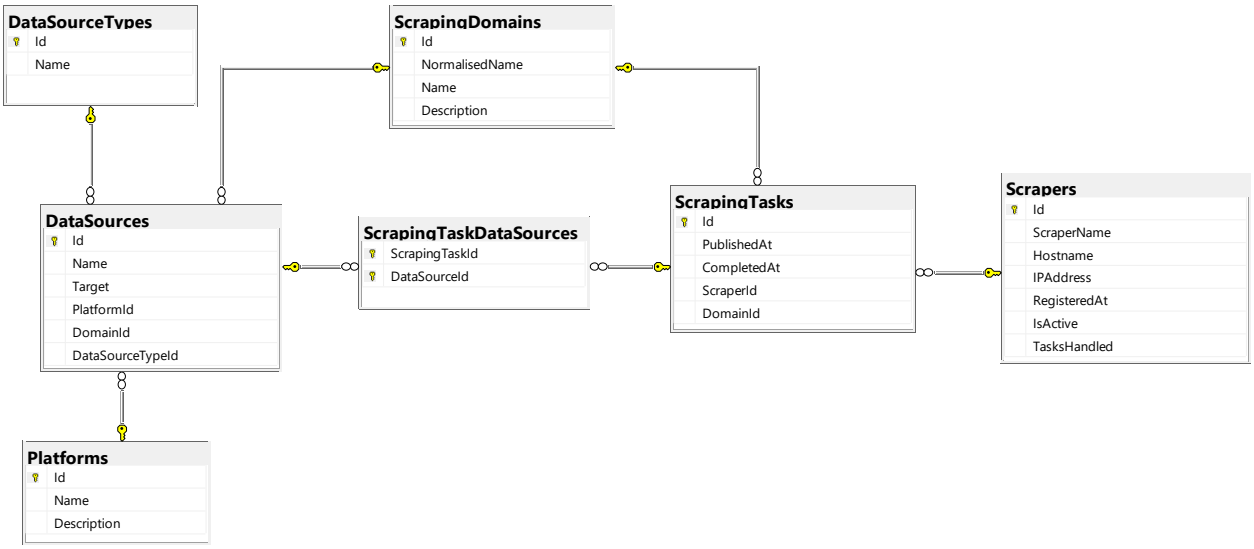
- تحوي طبقة التطبيق، على عقود الخدمات التي تحتاجها، نذكر منها هنا: عقود التراسل (إرسال واستقبال الرسائل عبر رتل التراسل)، عقود جدولة مصادر البيانات (التي تعرف عملية الجدولة بشكل مجرد، ويكون التنجيز الفعلي في طبقة التطبيق حسب نوع مصد البيانات)، كما تحوي هذه الطبقة على حالات الاستخدام التي قسمت إل استعلامات وأوامر.
- يحوي مشروع الخدمات في طبقة البنية التحتية، على التنفيذ الفعلي لخدمة إرسال واستقبال الرسائل (Kafka). وبالإضافة إلى ذلك يحوي على خوارزمية الجدولة الفعلية. ويحوي مشروع المزامنة Persistence على تنجيز عقود المستودعات.



الشكل 74: بنية طبقة التطبيق في خدمة إدارة عمليات سحب البيانات.

تصميم قاعدة المعطيات

يوضح الشكل أدناه تصميم قاعدو المعطيات، حيث لدينا، جدول مجالات السحب الذي يرتبط مع عدد من مصادر البيانات (مثل، أخبار سياسية سورية، يرتبط مع مصادر مثل "الدفاع المدني"، "دمشق"، ... إلخ). ولكل مصدر بيانات نوع (مثل كلمة مفتاحية، اسم مستخدم، قناة عامة، مجموعة) . ومنصة تواصل اجتماعي تابع لها (أي يصبح مصدر البيانات مثلاً "قناة سوريا بلحظ" من نوع قناة، على منصة تيليغرام). ولدينا أيضاً ساحبي البيانات (خدمات السحب، حيث ترتبط كل خدمة سحب مع مجموعة من مهمات السحب، ولكل مهمة مجموعة من مصادر البيانات المرتبطة معها.



الشكل 75: مخطط قاعدة المعطيات الخاص بخدمة إدارة عمليات سحب البيانات.

الأنماط التصميمية المستخدمة

في هذه الخدمة، تم استخدام النمط التصميمي Mediator، ضمن طبقة التطبيق Application، حيث أن جميع الطلبات الداخلة للنظام تمر عبر الوسيط الذي يقوم بمقاطعة الطلب وإجراء بعض العمليات قبل مثل (التسجيل Auditing، التحق من صحة الطلب Validation).

كما تم استخدام النمط التصميمي CQRS، بحيث تم فصل حالات استخدام النظام إلى أوامر تعدل على حالة النظام، واستعلامات تعيد بيانات دون التعديل على حالة النظام.

كما تم استخدام النمط التصميمي النتيجة Result Pattern، لتغليف الأخطاء الصادرة عن التطبيق.

خدمة كشف الأحداث

مقدمة

تأخذ هذه الخدمة دفعات الرسائل Messages Batches القادمة من خدمة سحب البيانات، وتقوم بعنقدة الرسائل وفق ما هو موضح في فصل المنهجية المقترحة، وبعدها تقوم بعمليات المعاجلة اللاحقة Post Event Detection، مثل التواصل مع خدمة الاستدلال على الموقع لتحديد موقع الحدث، والتواصل مع نموذج لغة ضخم من أجل تلخيص الحدث. وبعد ذلك تقوم بإرسال الحدث إلى خدمة إدارة الأحداث لإضافته. كما تم تنجيز هذه الخدمة بلغة Python، بسبب حاجة هذه الخدمة إلى نماذج الذكاء الصنعي وتقنيات معالجة اللغات الطبيعية، وهو ما يميز لغة بايثون. ويبين الشكل أدناه بنية هذه الخدمة.

الشكل 76: بنية خدمة كشف الأحداث.

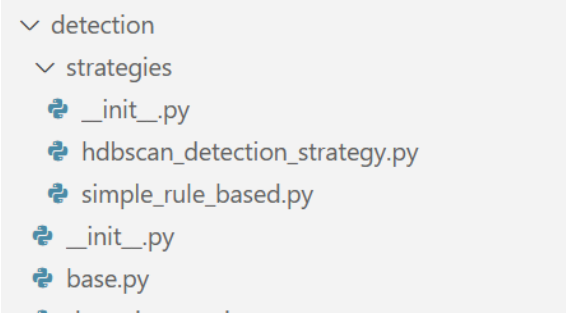
تفاصيل التنجيز

تحوي هذه الخدمة على أربع مجلدات فرعية ضمن مجلد الرماز المصدري، نواة المشروع Core، شريحة الكشف Detection، شريحة المعالجة اللاحقة، شريحة التواصل عبر الأرتال Kafka.

حيث وكما في بقية الخدمات، تحوي هذه الخدمة على مجلد الـ Core الذي يحوي على الأخطاء التي تصدر عن النظام، بالإضافة إلى الخدمات التي تعتمد عليها بقية الشرائح، مثل التواصل مع نموذج لغة ضخم، التواصل مع خدمة الاستدلال على الموقع.

وفي شريحة التراسل Kafka، فتحوي على الصفوف اللازمة لقراءة الرسائل التي ترسلها خدمات سحب البيانات.

أما في شريحة الكشف Detection، فتحوي على تجريد لعملية الكشف (العنقدة) Detection Base، والاستراتيجيات المختلفة للكشف، حيث نجزنا بها الحل المطروح في المنهجية المقترحة.



الشكل 77: بنية شريحة الكشف.

أما في شريحة خط الأنابيب (المعالجة اللاحقة)، فتحوي على تنجيزات مراحل المعالجة (مثل، تصنيف الموضع Topic Classification، تحديد الموقع Event Locating، تلخيص الحدث). وكما ذكرنا في الشريحة المشابهة لهذه في خدمة سحب البيانات فهي تحوي على سجل لتسجيل مراحل المعالجة الموجودة، وآلية بناء خط الأنابيب انطلاقاً من أسماء مراحل المعالجة.

الأنماط التصميمية المستخدمة

في هذه الشريحة تم استخدام النمط التصميمي Pipes & Filters، من أجل تعريف خط الأنابيب، كما تم استخدام النمط التصميمي Strategy Pattern، لتعريف الطرائق المتخلفة لكشف الأحداث.

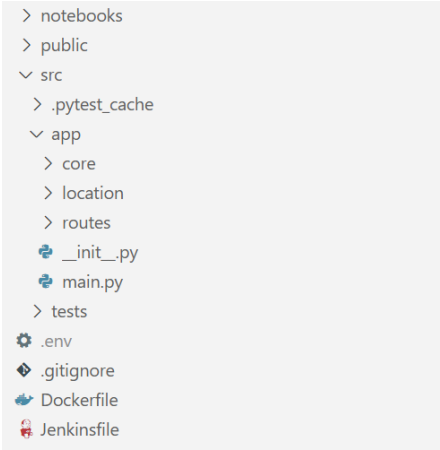
خدمة الاستدلال على الموقع

مقدمة

تتلقى هذه الخدمة طلبات الاستدلال على الموقع (رسالة نصية => الموقع المذكور، موقع مذكور => إحداثيات) من خدمة كشف الأحداث، وتقوم بإعادة الإحداثيات الجغرافية للموقع المذكور. كما وتم تنجيز هذه الخدمة باستخدام إطار العمل Flask وبلغة Python أيضاً بسبب حاجة هذه الخدمة إلى تقنيات معالجة اللغات الطبيعية.

تفاصيل التنجيز

تحتوي هذه الخدمة على ثلاث شرائح، نواة المشروع Core التي تحوي على خدمات للتواصل مع نماذج اللغة الضخمة، صفوف التهيئة. والشريحة الثانية هي شريحة الموقع Location، التي تحوي كما هو مبين بالشكل أدناه، على تجريد لطريقة التعرف على الموقع المذكور Location Mention Recognition، وطريقة تحديد الموقع Location Resolution. وضمن مجلد الـ Recognition يوجد الطرائق المتاحة للتعرف على الموقع، ومجلد الـ resolution يحوي على الطرائق المتاحة لتحديد الموقع، بالإضافة إلى سجل يتم تسجيل الطرائق المتاحة فيه، وآلية بناء انطلاقاً من اسم الطريقة المرادة. ويبين الشكل أدناه بينة الخدمة.



الشكل 78: بنية خدمة الاستدلال على الموقع.

الأنماط التصميمية المستخدمة

تم استخدام النمط Strategy لتعريف الطرائق المتخلفة للاستلال على الموقع، وتم استخدام النمط التصميمي Registry لتسجيل الطرائق المتنوعة للاستدلال والتعرف، واستخدم النمط التصميمي Factory لإنشاء هذه الطرائق بشكل ديناميكي.

خدمة إدارة الأحداث

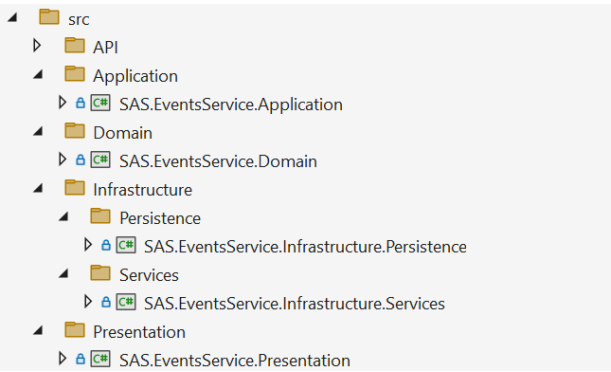
مقدمة

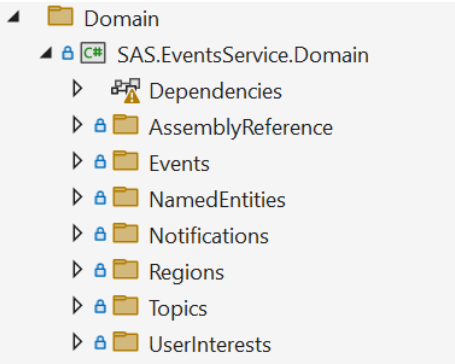
تعد هذه الخدمة الجزء الرئيسي من نظامنا، فهي التي تتعامل مع طلبات المستخدمين لاستعراض الأحداث، وإدارة مناطق اهتمامهم، حيث أنها تخفي كل خط العمل السابق، بحيث فقط يتعامل المستخدم النهائي معها كخدمة تتواصل مع قاعدة بيانات لجلب الأحداث وإدارة اهتماماته. كما تقوم خدمة كشف الأحداث التي ذكرناها سابقاً، بالتواصل مع هذه الخدمة لإضافة الأحداث.

تم تنجيز هذه الخدمة باستخدام إطار العمل Asp.Net، لأنه يعتبر من أفضل أطر العمل للمشاريع الكبيرة والتي تحوي على منطق عمل معقد. وتم استخدام قاعدة البيانات SQL Server بسبب تكاملها السلس مع إطار العمل هذا.

تفاصيل التنجيز

يبين الشكل أدناه بنية هذه الخدمة، حيث تحوي على الطبقات النموذجية للبنية المعمارية النظيفة وفق المبادئ التي ذكرناه في فقرة المبادئ التصميمية. وتحوي هذه الخدمة على عدد من المجالات وهي: مجال الأحداث، مجال اهتمامات المستخدمين، مجال مواضيع الأحداث، مجال الإشعارات، مجال الكيانات المسماة. كما هو مبين بالأشكال التالية.





الشكل 79: بنية خدمة إدارة الأحداث.

في طبقة المجال، يوجد ستة مجلدات تمثال مجالات العمل ضمن الخدمة، التي ذكرناها أعلاه، وفي كل مجال يوجد صفوف الكيانات Entities التابعة له، وعقود المستودعات الخاصة به، والأخطاء والاستثناءات التي تصدر عنه.

في طبقة التطبيق، توجد أيضاً ستة مجالات تمثال مجالات العمل، وفي كل منها يجد حالات الاستخدام الخاصة بهذا المجال، بالإضافة إلى أغراض نقال البيانات الخاصة به Data Transfer Objects (Dtos). كما يوجد في هذه الطبقة عقود الخدمات التي تعتمد عليها هذه الخدمة مثل (عقود التواصل مع نماذج اللغة الضخمة، الإشعارات، استخراج الكيانات المسماة).

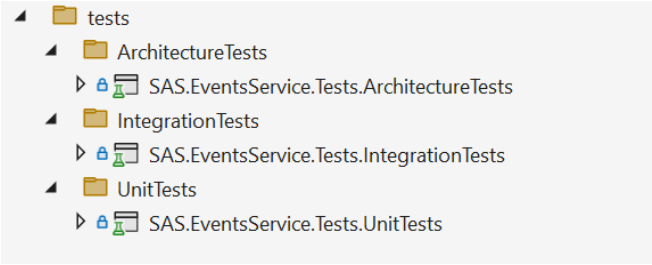
الشكل 80: بنية طبقة التطبيق

في طبقة البنية التحتية، يوجد مشروعين الأول Persistence الذي يتنجز عقود المستودعات والتواصل مع قاعدة البيانات. ومشروع الخدمات الذي ينجز عقود الخدمات التي تعتمد عليها طبقة التطبيق. (فمثلاً ينجز عقود خدمة المستخدم IUserService عبر الاتصال مع خدمة الهويات Identity Service).

وفي طبقة العرض، يوجد المتحكمات Controllers، التي نعرف فيها واجهة التخاطب Endpoints API.

ويأتي مع هذه الخدمة مجموعة من الاختبارات، وهي:

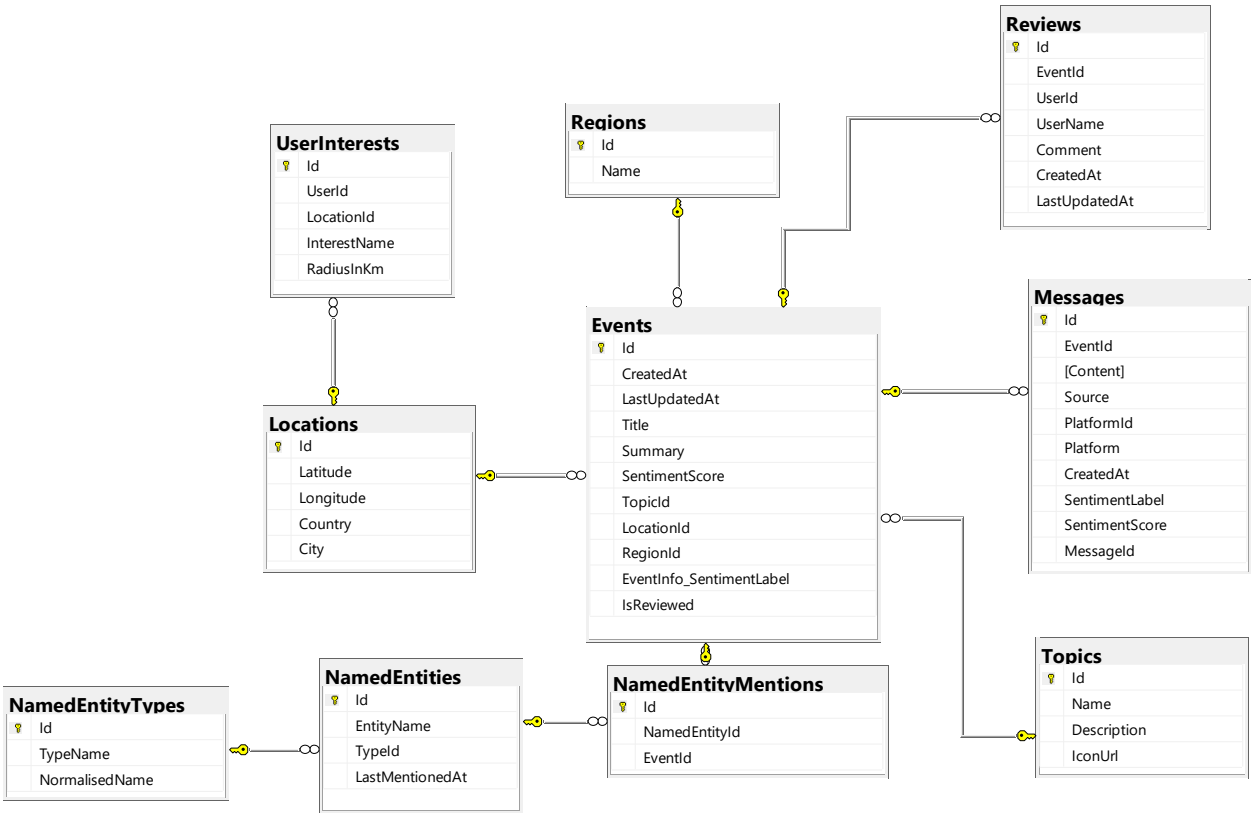
- اختبار البنية المعمارية للتأكد من أن قواع البنية المعمارية لم تخرق.
- اختبار الوحدات بحيث يتم اختبار كل وحدة ضمن النظام واختبار التكامل، بحيث يتم اختبار حالة استخدام كاملة للنظام.



الشكل 81: بنية مجلدات الاختبارات.

تصميم قاعدة المعطيات

يوضح الشكل أدناه تصميم قاعدة المعطيات، حيث نجد جدول الأحداث وهو الأساس (يحوي على ملخص الحدث وعنوانه، تاريخ إنشاءه وهل تمت مراجعته) ويرتبط الحدث بمكان، ومنطقة، ومجموعة من الرسائل، والكيانات المذكورة فيه. وموضوع الحدث. ولدينا جدول الكيانات المذكورة (مثل دمشق، منظمة الصحة العالمية) الذي له نوع Named Entity Type مثل اسم مكان، اسم منظمة، اسم شخص.



الشكل 82: تصميم قاعدة المعطيات لخدمة إدارة الأحداث.

الأنماط التصميمية المستخدمة

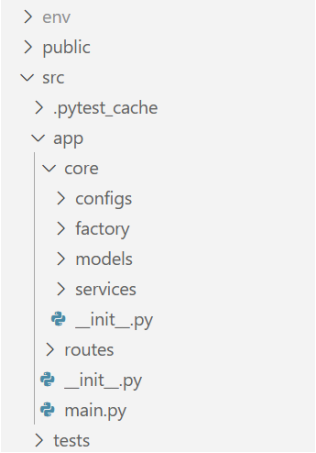
في هذه الخدمة، تم استخدام النمط التصميمي Mediator، ضمن طبقة التطبيق، حيث أن جميع الطلبات الداخلة للنظام تمر عبر الوسيط الذي يقوم بمقاطعة الطلب وإجراء بعض العمليات عليها. كما تم استخدام النمط التصميمي CQRS، بحيث تم فصل حالات استخدام النظام إلى أوامر تعدل على حالة النظام، واستعلامات تعيد بيانات دون التعديل على حالة النظام.

خدمة البوابة

تم تنجيز خدمة البوابة باستخدام إطار العمل ASP.NET، حيث تتولى هذه الخدمة الدور الرئيسي في استقبال وتوجيه طلبات المستخدمين إلى الخدمات الأخرى ضمن النظام. بحيث يتعامل المستخدم مع النظام ككتلة واحدة ويخفي ذلك كل تعقيدات النظام وخدماته. وقد تم اعتماد مكتبة YARP (Yet Another Reverse Proxy) لتنفيذ آلية توجيه الطلبات (Request Routing) بطريقة مرنة وقابلة للتوسّع، بما ينسجم مع طبيعة النظام القائم على الخدمات المصغّرة (Microservices).

خدمة استخراج الكيانات المسماة

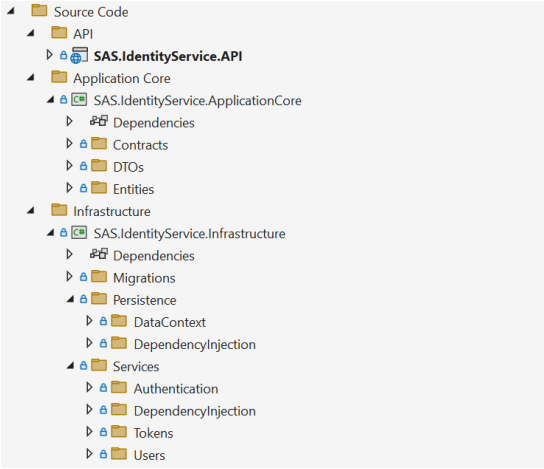
لقد كانت خدمة إدارة الأحداث بحاجة إلى استخراج الكيانات المسماة المذكورة في نص الحدث، ولأن لغة #C كانت تفتقر إلى أدوات معالجة اللغات الطبيعة واستخراج الكيانات المسماة، كنا بحاجة إلى خدمة تقوم بذلك، لذلك بنجزنا هذه الخدمة، التي مهمتها هي تلقي طلبات الاستخراج (الرسائل) وإعادة الكيانات المسماة المذكورة في النص. وتم تنجيز هذه الخدمة بلغة Python لما توفره من دعم لتقنيات معالجة اللغات الطبيعية. ويبين الشكل أدناه بنية هذه الخدمة.



الشكل 83: بنية خدمة استخراج الكيانات المسماة.

خدمة الهوية

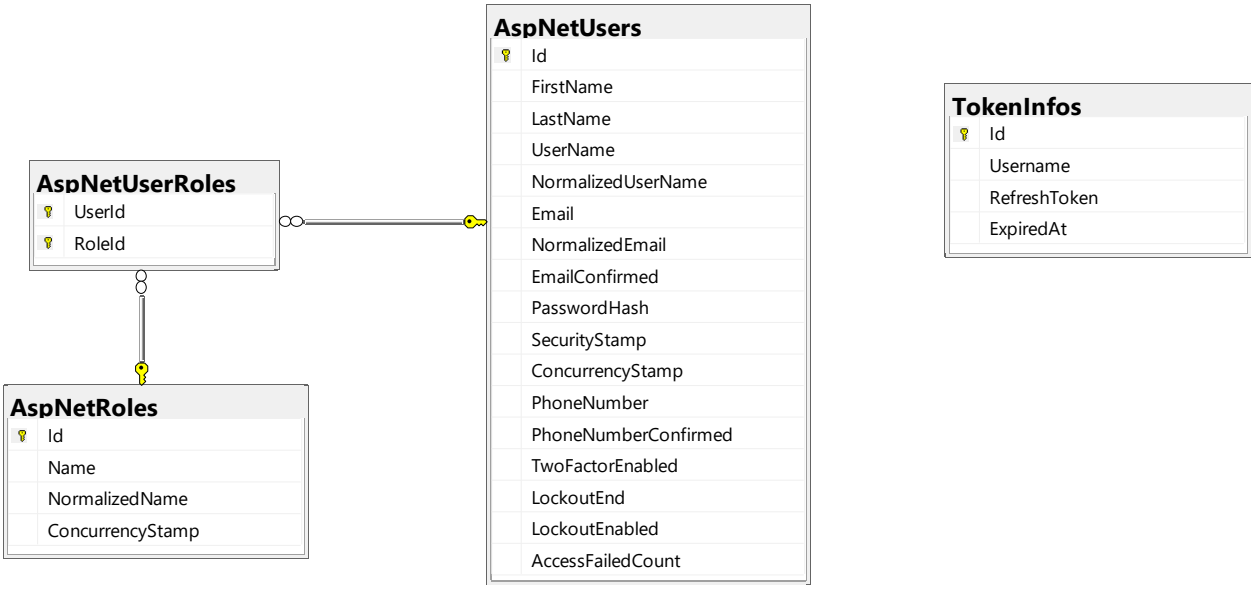
مهمة هذه الخدمة إدارة المستخدمين وصلاحياتهم، وأدوارهم. وتم تنجيز هذه الخدمة باستخدام إطار العمل Asp.Net، حيث قمنا بتغليف مكتبة الهوية التي يقدمها إطار العمل، ويوضح الشكل أدناه بنية هذه الخدمة. وفي هذه الخدمة، اعتمدنا على بنية نظيفة مصغرة، بحيث لدينا طبقة نواة التطبيق Application Core، التي تحوي على الكائنات، وعقود الخدمات، وأغراض نقل البيانات مع الطبقات الأعلى. ولدينا طبقة البنية التحتية، التي تحوي على تنجيز عقود الخدمات باستخدام المكتبة Asp.Net Identity. وفي طبقة الـ API يوجد المتحكمات التي تحوي واجهات التخاطبAPI Endpoints .



الشكل 84: بنية خدمة الهوية.

تصميم قاعدة المعطيات

يبين الشكل أدناه، مخطط قاعدة المعطيات، حيث نلاحظ أنه لدينا المستخدم، الذي له على مجموعة من الأدوار.

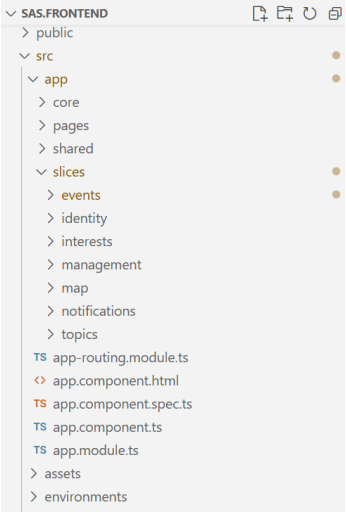


الشكل 85: خطط قاعدة المعطيات لخدمة الهوية.

خدمة واجهة المستخدم

تم تنجيز هذه الخدمة باستخدام إطار العمل Angular، تم تقسيم هذه الخدمة إلى النواة Core التي تحوي على الخدمات التي ستحتاجها بقية الشرائح، ومعترضات لإضافة الـ Tokens، التأكد من صلاحيات المستخدم. ويوجد مجلد الـ Pages الذي يحوي بعض الصفحات الثابتة مثل Landing Page، ومن ثم يأتي المجلد الأساسي وهو مجلد الشرائح Slices، حيث يحوي

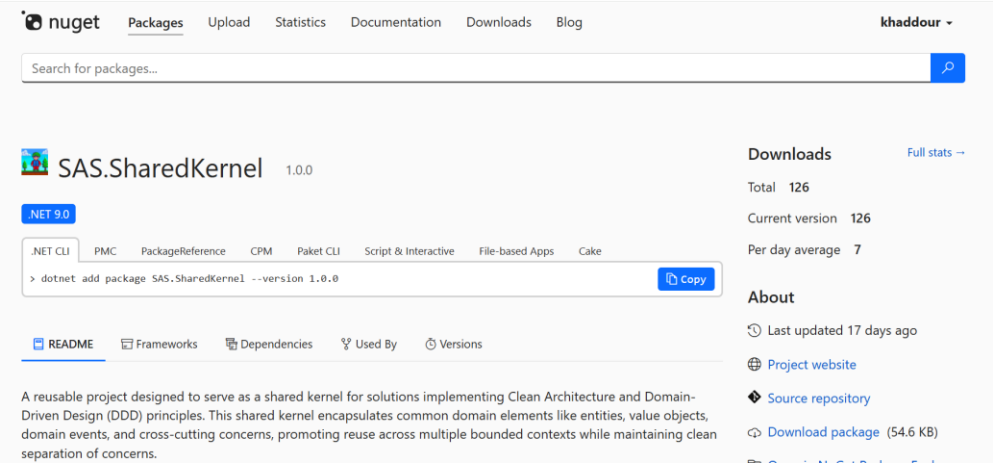
عل مجموعة من الـ Modules المستقلين التي تقابل الخدمات التي يقدمها النظام، وتم تقسيمهم إلى Modules، بحيث يتم تحميل الـ Module التي يحتاجها المستخدم عند الطلب فقط.



الشكل 86: بنية خدمة واجهة المستخدم.

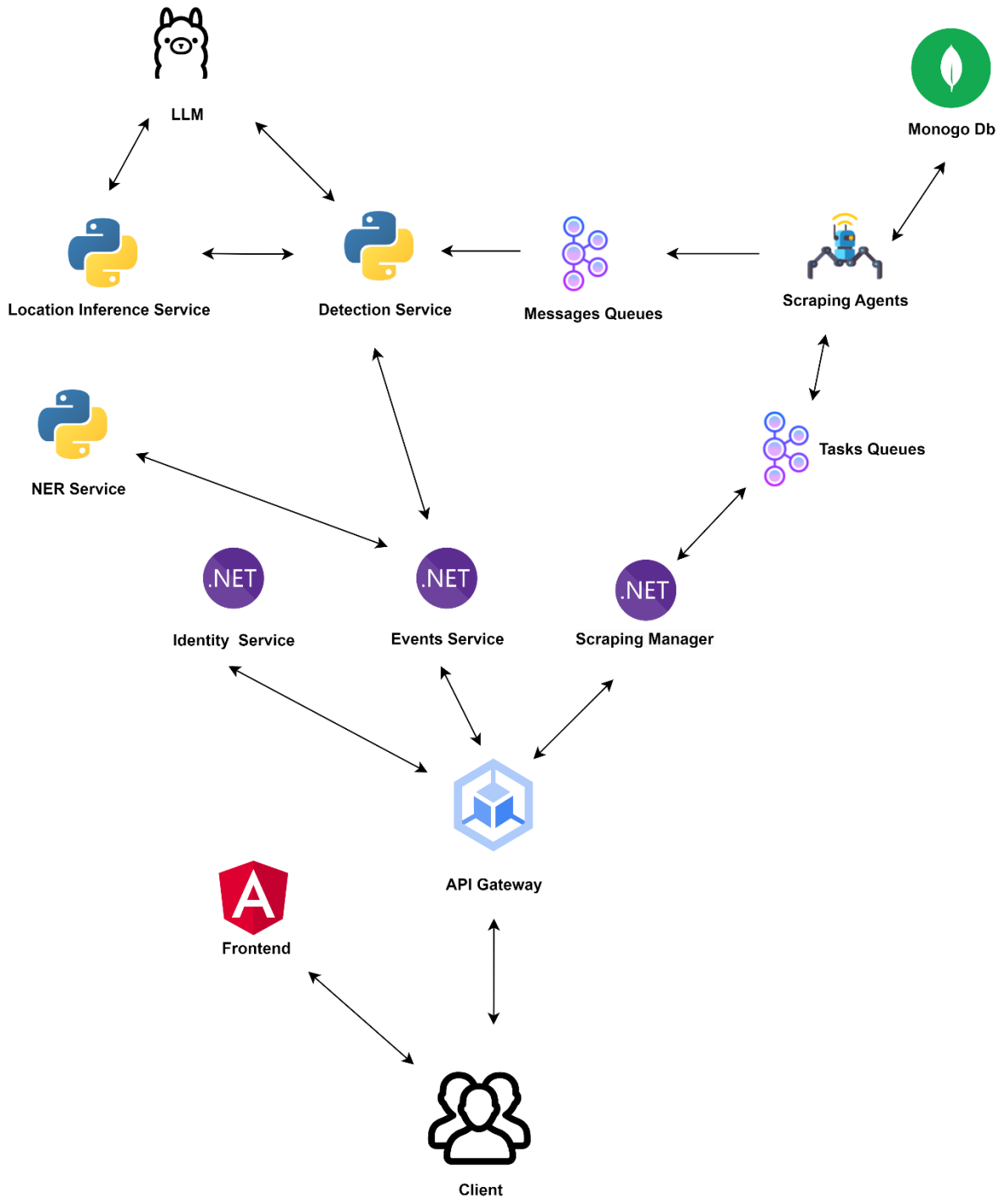
مشروع النواة المشتركة Shared Kernel

إن الخدمات المنجزة باستخدام البنية المعمارية النظيفة، تعتمد على طبقة تسمى النواة المشتركة، وهي تحوي على تجريد للخدمات والكائنات، والأوامر والاستعلامات، ... إلخ. وعند تصميمنا وتنجيزنا للخدمات كنا بحاجة إلى تنجزي النواة المشتركة في جميع الخدمات التي تبنى بالبنية المعمارية النظيفة (لدينا ثلاث خدمات)، لذلك وتحقيقاً للمبدأ DRY, Don't Repeat Yourself، تم تنجيز طبقة النواة، ورفعها كمكتبة Package على مستودع مكتبات Net. وهو Nuget.org، بحيث يتم استخدام هذه المكتبة التي نجزناه في كل الخدمات من خلال التعامل معها كأي مكتبة موجودة في إطار العمل. ويبين الشكل أدناه بنية الطبقة.



الشكل 87: مكتبة **Shared Kernel** التي قمنا برفعها على **Nuget.org**.

4.9- مخطط النظام التنجيزي



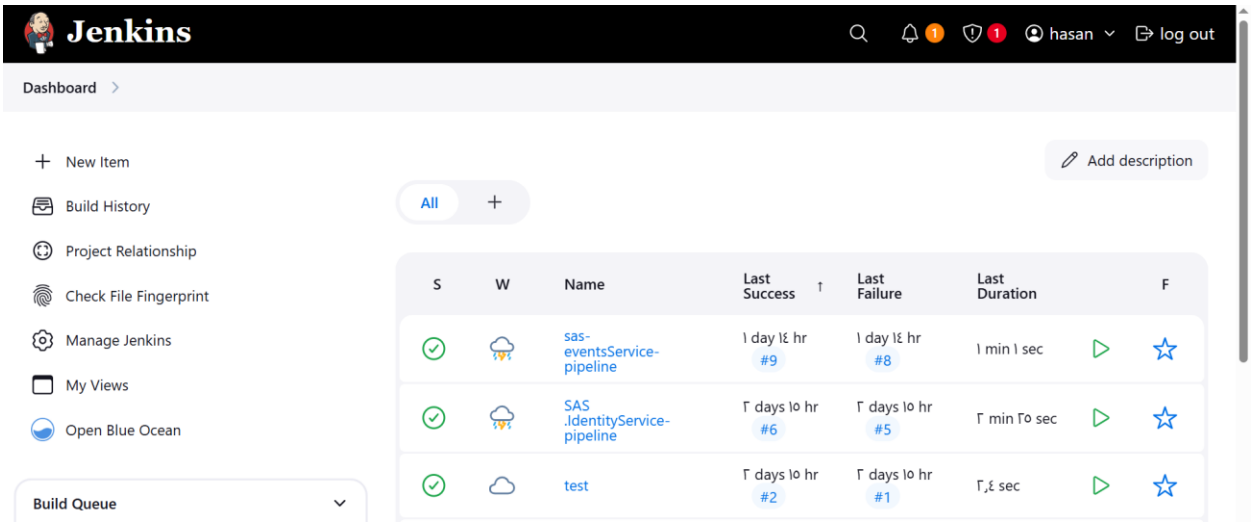
الشكل 88: مخطط النظام التنجيزي.

5.9- التكامل المستمر

تم أتمتة عملية اختبار الخدمات، وبناء الحاويات، من خلال إنشاء خط أنابيب Continuous Integration (CI) لكل خدمة، ويتم التشغيل باستخدام المخدم Jenkins، حيث عند حدوث أي تعديل على الفرع الـ master الخاص بالخدمة على مستودع الرماز المصدري، حيث يتم جلب الرماز المصدري، تنزيل المكتبات والاعتماديات، ويتم بناء الحاوية، وتشغيل الاختبارات، فيتم بذلك التحقق من أن الكود الذي تم تعديله مختبر ويعمل.

من خلال اعتماد هذه المنهجية في نشر التطبيق، تم تحقيق مجموعة من الميزات، أهمها:

- الحفاظ على الرمز البرمجي من الضياع وإمكانية مشاركته ضمن فريق العمل بسهولة، حيث تم تحقيق ذلك من خلال استخدام الأداة Git والموقع GitHub.
- عدم الحاجة إلى إعادة بناء الخدمات بشكل يدوي عند تعديلها ضمن المستودعات، إذ يعمل Jenkins على أتمتة هذه العملية، مما يسرع عملية التطوير والنشر ويحقق هذا مبدأ (Continuous Delivery).
- عدم الحاجة إلى إدارة الاعتماديات الخاصة بكل خدمة ودراسة تأثيرها على الخدمات الأخرى، حيث أن استخدام Docker يساعد على إنشاء حاوية معزولة وخاصة بكل خدمة تحتوي على الخدمة مع جميع تبعياتها.



الشكل 89: مثال على خطوط الأنابيب.

الفصل العاشر

تحليل ومناقشة النتائج

يوضح هذا الفصل النتائج التي حصلنا عليها ومقارنتها مع النموذج المرجعي.

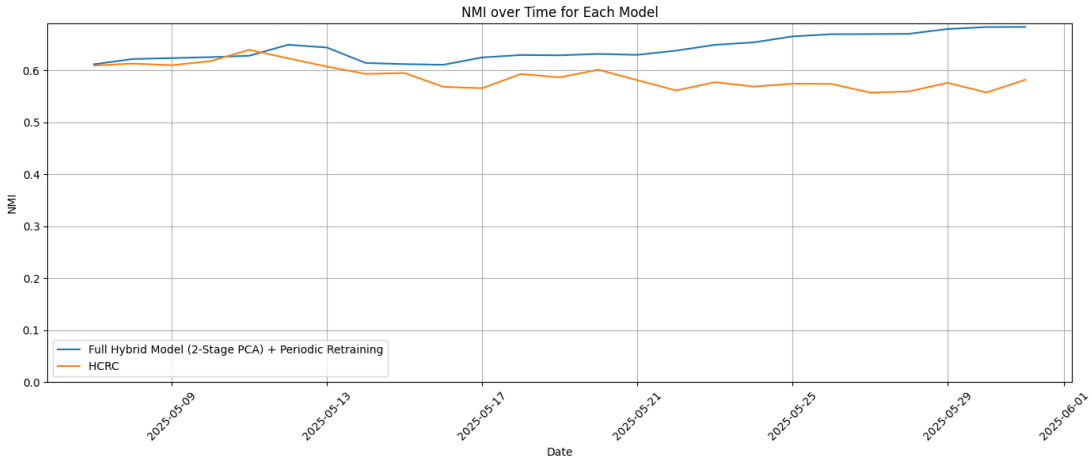
1.10- مناقشة النموذج

يبين الجدول أدناه خلاصة النتائج التي حصلنا عليها ومقارنتها مع النموذج المرجعي في التقييم بوضع عدم الاتصال. حيث تفوق النموذج المقترح على النموذج المرجعي بمقدار NMI 0.38، بما يعادل زيادة بنسبة 4 بالمئة وفق معيار التقييم NMI، بكما وتفوق بمقدار ARI 0.12 بما يعادل زيادة بنسبة 13%. بينما وفق مقياس AMI كان النموذج المرجعي أفضل بمقدار 0.07 AMI بما يعادل نسبة 7%.

جدول**17** : مقاررنة النموذج المقترح مع النوذج المرجعي.

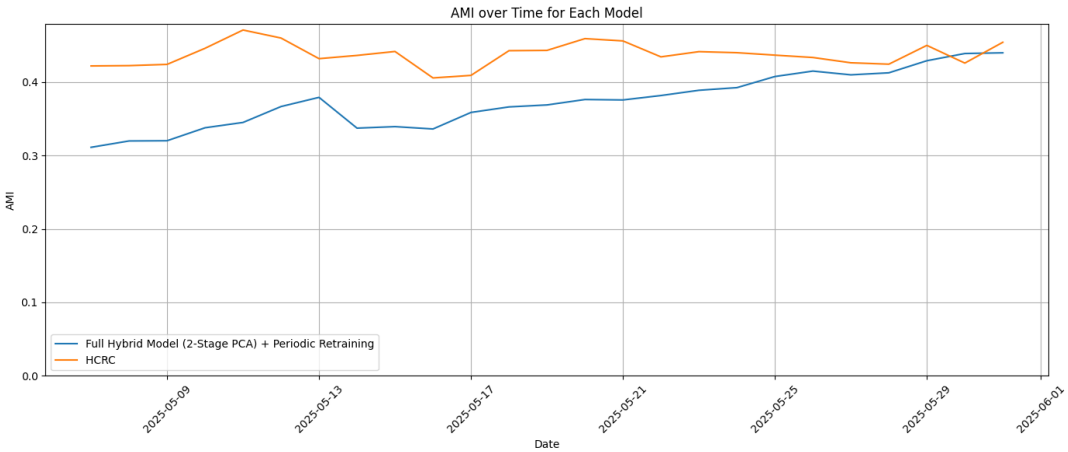
	Model	NMI	AMI	ARI
1	HCRC (Guo et al., 2024)	0.5885	**0.4803**	0.2320
2	Our Model	**0.6265**	0.4118	**0.3576**
		4%	-7%	13%

بينما في التقييم بالنمط الحي، كما هو مبين في المخطط أدناه، كان النموذج المقترح أفضل بنسبة خمسة بالمئة على النموذج المرجعي وسطياً.



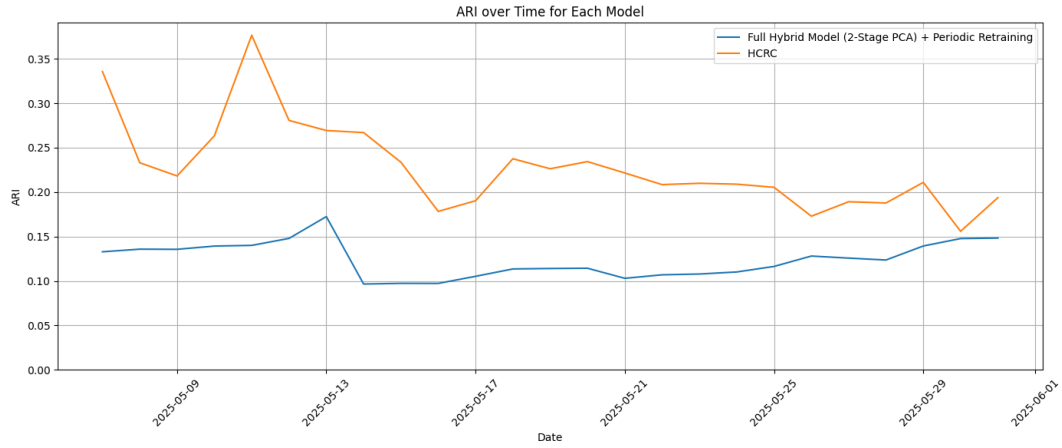
الشكل 90: مقارنة النموذج المقترح مع النموذج المرجعي وفق معيار **NMI**.

بينما في المقارنة وفق مقياس AMI، كان الفرق بنسبة 6 بالمئة لصالح النموذج المرجعي وهي مقاربة لما تفوق عليه في وضع التقييم Offline.



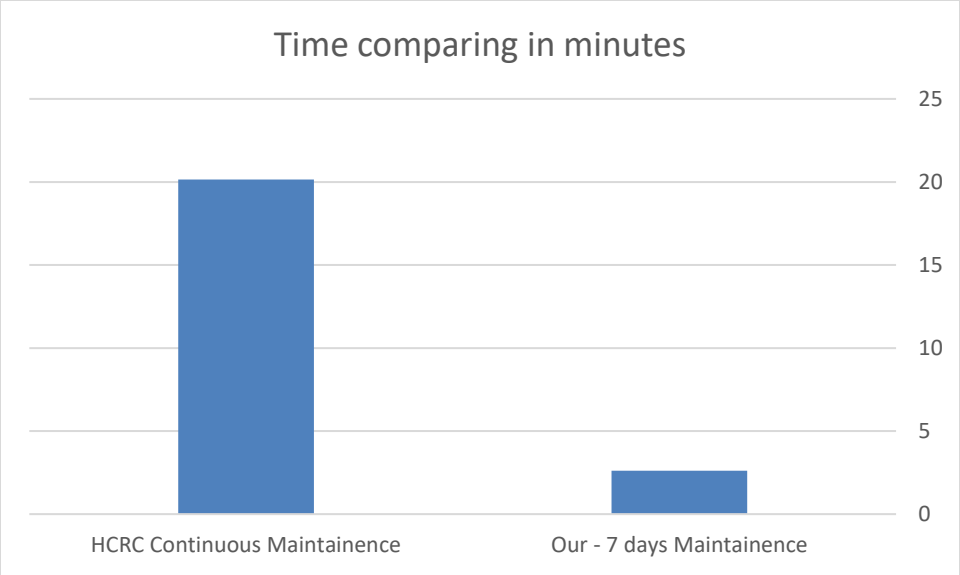
الشكل 91: مقارنة النموذج المقتر مع النموذج المرجعي وفق مقياس **AMI**.

وعند المقارنة وفق مقياس ARI، تفوق النموذج المرجعي على النموذج المقترح بمقدار 10 بالمئة وسطياً، على الرغم من أنّ نموذجنا قد تفوق عليه في نفس المقياس في وضع التقييم Offline. ويعود ذلك إلى بسبب حساسية ARI للتوزيع الداخلي للعناقيد، إذ أنّ وجود ضجيج في بيانات التقييم، يؤثر بشكل كبير على دقة تصنيف الأزواج ضمن نفس العناقيد أو خارجها، والتي تكون مشتتة أكثر في حالة التقييم في اللوذع الحي مما هي عليه في الوضع Offline.



الشكل 92: مقارنة النموذج المقترح مع النموذج المرجعي وفق مقياس **ARI**.

وعند مقارنة الأداء، أي الزمن اللازم لمعالجة كافة البيانات (إجراء عملية التعلم المقارنة والعنقدة) على كامل البيانات خلال الشهر، كان النموذج المقترح الذي يقوم بعملية التعلم المقارن كل سبعة أيام، أسرع من النموذج المرجعي بـ 7.7 ضعفاً حيث كان قادراً على معالجة البيانات كافة في دقيقتين فقط وستة ثوان، بينما كان النموذج المرجعي بحاجة إلى عشرين دقيقة. كما هو مبين في المخطط أدناه.



الشكل 93: مقارنة الزمن اللازم لمعالجة كامل البيانات والتعلم عليها، وفق النموذج المقتحر والنموذج المرجعي.

الفصل الحادي عشر

اختبارات النظام

يوضح هذا الفصل الاختبارات التي تم إجراؤها على النظام.

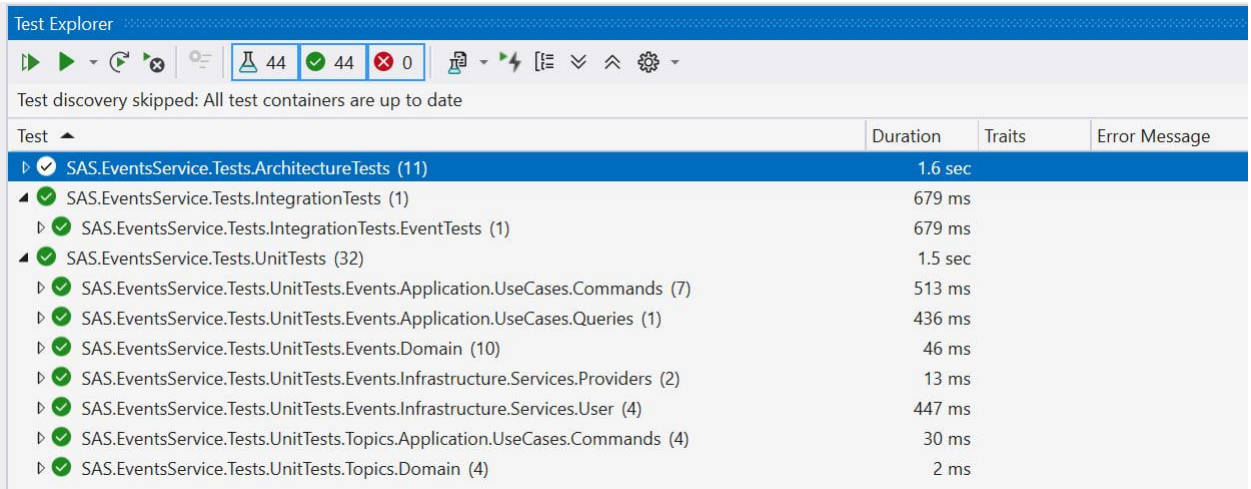
1.11- اختبار خدمات النظام

تُعدّ الاختبارات من الركائز الأساسية لضمان جودة الأنظمة البرمجية، إذ تُستخدم للتحقق من أن جميع مكونات النظام تعمل كما هو متوقع منها. وتتنوّع أنواع الاختبارات بين اختبارات الوحدة (Unit Testing)، واختبارات التكامل (Integration Testing)، واختبارات الأداء (Performance Testing). سنعرض في هذا الفصل الاختبارات التي خضع لها النظام.

اختبار خدمة إدارة الأحداث

تم إنشاء مجموعة من اختبارات الوحدة واختبارات التكامل، بهدف التحقق من مدى التزام الخدمة بالمتطلبات الوظيفية الموكلة إليها. شملت اختبارات الوحدة الحالات الأساسية لاستخدام الخدمة (مجال إدارة الحدث)، حيث جرى التأكد من صحة منطق تنفيذ الوظائف المعزولة داخل كل مكون برمجي. كما تم اختبار حالات الاستخدام كسيناريو كامل في اختبارات التكامل.

يتم تفعيل هذه الاختبارات بشكل أوتوماتيكي ضمن خط أنابيب التكامل المستمر (CI Pipeline)، وذلك عند دمج الرماز المصدري في الفرع المخصص للاختبار (Test Branch) ضمن مستودع الخدمة. يساهم هذا الأسلوب في الكشف المبكر عن الأخطاء، وضمان جودة الكود واستقراره قبل الانتقال إلى مراحل النشر



الشكل 94: اختبارات الوحدة والتكامل.

بالإضافة إلى ذلك تم إجراء مجموعة من الاختبارات للتحقق من عدم خرق قواعد البنية المعمارية. حيث تم إجراء نوعين من اختبارات البنية المعمارية: الأول للتحقق من الاعتمادية الصحيحة بين الطبقات وفق مبادئ البنية النظيفة، والثاني للتأكد من الالتزام بقواعد التسمية الموحدة في الأوامر والاستعلامات والمعالجات، والشكل أدناه يبين النتائج.

Test Explorer

44 44 0

Test discovery skipped: All test containers are up to date

Test	Duration	Traits	Error Message
SAS.EventsService.Tests.ArchitectureTests (11)	1.6 sec		
SAS.EventsService.Tests.ArchitectureTests (11)	1.6 sec		
ArchitectureApplicationTests (4)	449 ms		
CommandHandlers_ShouldHave_NameEndingWith_CommandHandler	367 ms		
Commands_ShouldHave_NameEndingWith_Command	11 ms		
Queries_ShouldHave_NameEndingWith_Query	26 ms		
QueryHandlers_ShouldHave_NameEndingWith_QueryHandler	45 ms		
ArchitectureDomainTests (2)	394 ms		
Event_ShouldHave_NameEndingWith_Event	367 ms		
Repositories_ShouldHave_NameEndingWith_Repository	27 ms		
DependencyTests (5)	762 ms		
Application_Should_Not_DependOnOtherProjectExceptDmain	93 ms		
Domain_Should_Not_DependOnOtherProjectExceptSharedKernel	6 ms		
Persistence_Should_Not_DependOnOtherProjectExceptDomain	467 ms		
Presentation_Should_Not_DependOnOtherProjectExceptApplicationAndContracts	121 ms		
Services_Should_Not_DependOnOtherProjectExceptApplication	75 ms		

الشكل 95: اختبارات البنية المعمارية.

ويوضح الشكل أدناه تنفيذ خط الأنابيب لهذه الخدمة، ويبين كيف تم تحميل الرماز المصدري، وتنزيل الاعتماديات، وتشغيل الاختبارات وبناء الحاوية.

sas-eventsService-pipeline < 9 | Pipeline | Changes | Tests | Artifacts | Logout

Branch: – | 1m 1s | Changes by hasan.b.khaddour

Commit: – | a few seconds ago | Started by user hasan

Start — Checkout — Restore & Build — Unit Tests — Integration Tests — Architecture Validation — Docker Build — Docker Compose Up — End

Integration Tests - 7s | Restart Integration Tests

dotnet test D:\SAS\SAS.EventsService\tests\SAS.EventsService.Tests.IntegrationTests --no-build --no-restore --verbosity normal — Windows Batch Script 7s

الشكل 96: تشغيل خط الأنابيب الخاص بخدمة إدارة الأحداث.

اختبار خدمة سحب البيانات

يوضح الشكل أدناه نتائج اختبار خدمة سحب البيانات حث كانت قادرة على معالجة 75 ألف رسالة بمتوسط عشر رسائل كل ثانية.

```
PROBLEMS 12   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS                                   powershell

[INFO] Filtered out 0 messages.
[INFO] Published batch to topic Telegram.Politics: 20 messages
[INFO] First message in batch: 'رقم م الع إ لتاصو لج اع...' with label=positive, score=0.963
[INFO] Embeddings generated successfully.
[INFO] Starting classification with 8 messages
[INFO] Filtered out 0 messages.
[INFO] Published batch to topic Telegram.Politics: 8 messages
[INFO] First message in batch: 'رقم م الع إ لتاصو لج اع...' with label=positive, score=0.693
..' with label=positive, score=0.693                    c
Processed 28224 messages in 2641.80s -> 10.68 msgs/se
c

✅ Total messages processed: 75865
.

========================================== 1 passed in 7087.29s (1:58:07) ==========================================
(env) PS D:\SAS\SAS.ScapingAgent>
```

الشكل 97: اختبار قدرة خدمة سحب البيانات على معالجة الرسائل.

2.11- اختبارات الأداء

من أجل اجراء اختبارات الأداء على النظام، تم تحميل 59 ألف حدث تم توزيعها على كامل الشهر السابع، بمتوسط 2000 حدث في اليوم الواحد (حسب الدراسة الكمية).

اختبار الحمل Load Test

أجرينا اختبار الحمل الأول للتأكد من صحة المتطلب غير الوظيفي الذي ينص على النظام يجب أن يكون قادراً على الاستجابة لمعظم طلبات المستخدمين في غضون ثانيتين، حيث تم في هذا الاختبار اختيار مجموعة من حالات الاستخدام (استعراض الأحداث اليومية، استعراض الأحداث خلال مدة، خلال يوم، ... إلخ). حيث بدأنا ب 10 مستخدمين لثلاث ثوان ثم صعدنا إلى مئة مستخدم خلال 10 ثوان، ثم 500 مستخدم لمدة دقيقتين (عدد المستخدمين هنا هو وفق الدراسة الكمية)، وبعد ذلك هبطنا تدريجياً في عدد المستخدمين. ويوضح الشكل أدناه نتائج الاختبار، حيث تم إرسال خلاها 178,000 طلب، تم استقبالها جميعا والرد عليها بنجاح. وكان متوسط زمن الاستجابة هو 150 ms.

```
powershell Tests ×   JS load-test.js U

  TOTAL RESULTS

    checks_total.......................: 178820  584.110412/s
    checks_succeeded...................: 100.00% 178820 out of 178820
    checks_failed......................: 0.00%   0 out of 178820

    ✓ status is 200

    HTTP
    http_req_duration..................................................: avg=158.97ms min=0s    med=106.47ms max=3.44s p(9
0)=414.47ms p(95)=508.95ms
      { expected_response:true }.......................................: avg=158.97ms min=0s    med=106.47ms max=3.44s p(9
0)=414.47ms p(95)=508.95ms
    http_req_failed....................................................: 0.00%  0 out of 178820
    http_reqs..........................................................: 178820 584.110412/s

    EXECUTION
    iteration_duration.................................................: avg=463.04ms min=300ms med=411.5ms  max=3.75s p(9
0)=719.22ms p(95)=813.1ms
    iterations.........................................................: 178820 584.110412/s
    vus................................................................: 12     min=3          max=499
    vus_max............................................................: 500    min=500        max=500

    NETWORK
    data_received......................................................: 1.1 GB 3.5 MB/s
    data_sent..........................................................: 25 MB  81 kB/s



running (5m06.1s), 000/500 VUs, 178820 complete and 0 interrupted iterations
default ✓ [======================================] 000/500 VUs  5m6s
PS D:\SAS\SAS.GPDocs\Tests>
```

الشكل 98: نتائج اختبار الحمل.

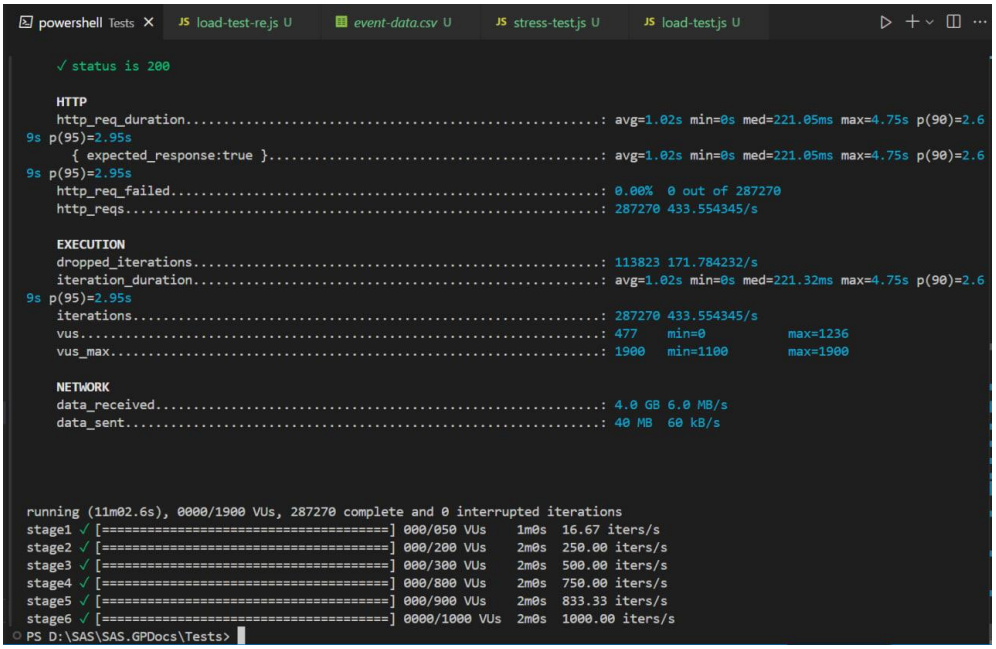
في الجزء الثاني من الاختبار، أجرينا اختبار حمل للتحقق من المتطلب غير الوظيفي الثاني، وهو قدرة النظام على تحمل 20,000 طلب تصفح في الدقيقة والاستجابة لها في غضون ثانيتين، حيث تم إرسال 100,000 طلب خلال خمس دقائق بمتوسط 20,000 طلب في الدقيقة. وكانت النتائج وفق الشكل المبين أدناه، بمتوسط زمن استجابة قدره 10 ms.

```
powershell Tests ×   JS load-test-re.js U   JS load-test.js U

    checks_succeeded...................: 100.00% 81840 out of 81840
    checks_failed......................: 0.00%   0 out of 81840

    ✓ status is 200

    HTTP
    http_req_duration..................................................: avg=10.93ms min=0s med=4ms    max=1.67s p(90)=15.
33ms p(95)=32.78ms
      { expected_response:true }.......................................: avg=10.93ms min=0s med=4ms    max=1.67s p(90)=15.
33ms p(95)=32.78ms
    http_req_failed....................................................: 0.00%  0 out of 81840
    http_reqs..........................................................: 81840  340.994908/s

    EXECUTION
    dropped_iterations.................................................: 161    0.670823/s
    iteration_duration.................................................: avg=11.19ms min=0s med=4.08ms max=1.67s p(90)=15.
81ms p(95)=33.24ms
    iterations.........................................................: 81840  340.994908/s
    vus................................................................: 2      min=0          max=128
    vus_max............................................................: 178    min=140        max=178

    NETWORK
    data_received......................................................: 1.1 GB 4.7 MB/s
    data_sent..........................................................: 11 MB  47 kB/s



running (4m00.0s), 0000/0178 VUs, 81840 complete and 0 interrupted iterations
stage1 ✓ [======================================] 000/020 VUs    1m0s  16.67 iters/s
stage2 ✓ [======================================] 0000/0138 VUs  4m0s  333.33 iters/s
stage3 ✓ [======================================] 000/020 VUs    1m0s  16.67 iters/s
PS D:\SAS\SAS.GPDocs\Tests>
```

الشكل 99: نتائج اختبار الحمل الثاني.

اختبار الضغط Stress Test

في هذا الاختبار، تم اختبار قدرة النظام، على الارتفاع عن الحمل المتوقع للطلبات، حيث بدأنا الاختبار 15000 طلب في الدقيقة، ثم صعدنا إلى 30,000 طلب في الدقيقة، ثم إلى 45,000 طلب ثم إلى 50,000 طلب ثم إلى 60,000 طلب في الدقيقة (أربع أضعاف الحمل المتوقع) حتى وصل زمن الاستجابة إلى 2.3 ثانية للطلب فأوقفنا الاختبار.



الشكل 100: نتائج اختبار الضغط.

اختبار الارتفاع المفاجئ Spike Test

قمنا في هذا الاختبار، باختبار سلوك النظام، عند الارتفاع في عدد الطلبات لفتر قصيرة. حيث قمنا برفع عدد الطلبات إلى 45,000 طلب في الدقيقة لمدة 40 ثانية. ويبين الشكل أدناه نتائج الاختبار، حيث تم هذا الاختبار بنجاح.

```
powershell Tests ×   event-data.csv U   JS spike-test.js U   JS load-test-re.js U   JS stress-test.js U   JS load-test.js

     checks_total.......................: 25526   614.243858/s
     checks_succeeded...................: 100.00% 25526 out of 25526
     checks_failed......................: 0.00%   0 out of 25526

     ✓ status is 200

     HTTP
     http_req_duration..................................................: avg=943.6ms  min=0s med=359.92ms max=3.04s p(90)=
2.36s p(95)=2.49s
       { expected_response:true }.......................................: avg=943.6ms  min=0s med=359.92ms max=3.04s p(90)=
2.36s p(95)=2.49s
     http_req_failed....................................................: 0.00%  0 out of 25526
     http_reqs..........................................................: 25526  614.243858/s

     EXECUTION
     dropped_iterations.................................................: 4474   107.659916/s
     iteration_duration.................................................: avg=944.02ms min=0s med=360.56ms max=3.04s p(90)=
2.37s p(95)=2.49s
     iterations.........................................................: 25526  614.243858/s
     vus................................................................: 389    min=2          max=1000
     vus_max............................................................: 1000   min=101        max=1000

     NETWORK
     data_received......................................................: 351 MB 8.4 MB/s
     data_sent..........................................................: 3.5 MB 85 kB/s



running (0m41.6s), 0000/1000 VUs, 25526 complete and 0 interrupted iterations
spike_test ✓ [======================================] 0000/1000 VUs  40s  750.00 iters/s
PS D:\SAS\SAS.GPDocs\Tests>
```

الشكل 101: نتائج اختبار الارتفاع المفاجئ في الطلبات.

الخاتمة والآفاق المستقبلية

قدمنا في هذا العمل منهجية هجينة لكشف الأحداث في وسائل التواصل الاجتماعية، تجمع بين البحث بكلمات مفتاحية عامة وتعزيزها من خلال البيانات المرجعة، وتعلم تمثيل جيد لها بالاستعانة بشبكات البيان العصبونية، والتعلم المقارن. كما أغنينا محتوى هذا المجال من البحث (مجال كشف الأحداث) بمجموعة بيانات باللغة العربية قمنا بجمعها خلال هذا العمل تحوي أحداث وقعت في شهر أيار من عام 2025. واستفدنا من المنهجية المطروحة، للوصول إلى نظام متكامل يقدم للمستخدم والإعلاميين والجهات صاحبة القرار معلومات عما يحدث في بلدنا سوريا، مما يساعد في دعم اتخاذ القرارات، ومكافحة الشائعات، وتقديم رؤى عما يناقش على وسائل التواصل الاجتماعي.

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.] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching Word Vectors with Subword Information* (No. arXiv:1607.04606). arXiv. https://doi.org/10.48550/arXiv.1607.04606

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

[10.] 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

[11.] 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

[12.] 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

[13.] 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

[14.] 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

[15.] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (No. arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

[16.] 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

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

[18.] 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

[19.] 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

[20.] 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

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

[22.] 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

[23.] 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

[24.] 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

[25.] 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

[26.] 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

[27.] 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

[28.] 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

[29.] 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

[30.] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (No. arXiv:1301.3781). arXiv. https://doi.org/10.48550/arXiv.1301.3781

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

[32.] 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

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

[34.] 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

[35.] 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

[36.] 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

[37.] 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

[38.] 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

[39.] 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

[40.] 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

[41.] 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

[42.] 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

[43.] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (No. arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

[44.] 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

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

[46.] 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

[47.] 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

[48.] 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

[49.] 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

[50.] 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 كلمات 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).

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	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

ملحق ت

نتائج التجارب

نرفق هنا النتائج التفصيلية التي حصلنا عليها في التجارب التي قمنا بها.

جدول**18** : نتائج التجربة الأول في مرحلة تصنيف التغريدات.

#	Embedding Model	Model	Results							
			Accuracy		Recall		Precision		F1 Score	
			Train	Test	Train	Test	Train	Test	Train	Test
Exp. 1	FastText	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	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	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	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
	Distill mBert	Neural Network	0.87	0.79	0.90	0.85	0.91	0.86	0.85	0.74
		KNN	0.83	0.77	0.85	0.79	0.90	0.88	0.81	0.74
		SVM	0.83	0.79	0.86	0.82	0.88	0.87	0.80	0.74
		Naïve Bayes	0.67	0.64	0.68	0.65	0.81	0.81	0.64	0.60
		Logistic Regression	0.82	0.78	0.88	0.84	0.86	0.86	0.79	0.74

جدول**19** : نتائج التجربة الثانية في مرحلة تصنيف التغريدات

#	Embedding Model	Model	Results							
			Accuracy		Recall		Precision		F1 Score	
			Train	Test	Train	Test	Train	Test	Train	Test
Exp. 2	FastText+ NE	Neural Network	0.68	0.72	1.0	0.50	0.68	0.72	0.41	0.42
		KNN	0.69	0.72	1.0	0.51	0.69	0.72	0.43	0.42
		SVM	0.68	0.72	1.0	0.50	0.68	0.72	0.41	0.42
		Naïve Bayes	0.68	0.72	1.0	0.50	0.68	0.72	0.41	0.42
		Logistic Regression	0.32	0.28	0.1	0.09	0.09	0.1	0.24	0.22
	AraBert+ NE	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 +NE	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+ NE	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
	Distill + NE	Neural Network	0.78	0.74	0.92	0.89	0.79	0.78	0.72	0.63
		KNN	0.75	0.72	0.88	0.84	0.78	0.78	0.68	0.62
		SVM	0.78	0.73	0.92	0.87	0.80	0.78	0.72	0.62
		Naïve Bayes	0.48	0.47	0.37	0.40	0.74	0.75	0.48	0.47
		Logistic Regression	0.74	0.75	0.94	0.93	0.75	0.77	0.63	0.62

جدول**20** : النتائج التفصيلية للتجربة الثالثة في مرحلة تصنيف التغريدات.

#	Name	Model	results							
			Accuracy		Recall		Precision		F1 Score	
			Train	Test	Train	Test	Train	Test	Train	Test
Exp. 3	FastText	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	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	Neural Network	0.93	0.80	0.93	0.82	0.97	0.89	0.92	076
		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	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
	Distill mBert	Neural Network	0.87	0.74	0.88	0.78	0.92	0.84	0.85	0.9
		KNN	0.78	0.70	0.79	0.73	0.87	0.83	0.75	0.66
		SVM	0.83	0.76	0.87	0.80	0.88	0.86	0.80	0.72
		Naïve Bayes	0.57	0.56	0.59	0.53	0.79	0.79	0.56	0.54
		Logistic Regression	0.82	0.77	0.89	0.83	0.86	0.85	0.79	0.72

جدول**21** : النتائج التفصيلية للتجربة الرابعة، في مرحلة تصنيف التغريدات.

#	Name	Model	results							
			Accuracy		Recall		Precision		F1 Score	
			Train	Test	Train	Test	Train	Test	Train	Test
Exp. 3	FastText	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	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	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	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
	Distill mBert	Neural Network	0.88	0.79	0.90	0.84	0.92	0.86	0.86	0.74
		KNN	0.83	0.80	0.86	0.83	0.88	0.86	0.80	0.74
		SVM	0.82	0.78	0.86	0.84	0.87	0.86	0.79	0.73
		Naïve Bayes	0.73	0.70	0.77	0.71	0.86	0.86	0.73	0.67
		Logistic Regression	0.81	0.78	0.88	0.85	0.85	0.85	0.78	0.73

جدول**22** : النتائج التفصيلية للتجربة الأولى في وضع عدم الاتصال، في مرحلة كشف الأحداث.

#	Approach	Embedding Model	Results		
			NMI	AMI	ARI
Exp. 1	AraBert + DBSCAN	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 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 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 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	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

جدول**23** : النتائج التفصيلية للتجربة الثانية في وضع عدم الاتصال، في مرحلة كشف الأحداث.

#	Approach	Embedding Model	Results		
			NMI	AMI	ARI
Exp. 1	HDBSCAN + Arabert	AraBert 768	0.3559	0.2408	0.0096
		AraBert 320	0.3496	0.2349	0.0094
		AraBert 128	0.3304	0.2204	0.0094
		AraBert 64	0.3031	0.1930	0.0073
Exp. 2	HDBSCAN +mBert	mBert 768	0.4815	0.3711	0.0249
		mBert 512	**0.4838**	**0.3731**	**0.0251**
		mBert 256	0.4826	0.3715	0.0249
		mBert 128	0.4717	0.3600	0.0219
		mBert 64	0.4500	0.3347	0.0198
Exp. 3	HDBSCAN +FastText	FastText 300	0.3593	0.1902	0.0083
		FastText 160	0.3482	0.2008	0.0089
		FastText 128	0.3324	0.2188	0.0106
		FastText 64	0.3199	0.2261	0.0106
Exp. 4	HDBSCAN+Word2Vec	Word2Vec 300	0.3522	0.2269	0.0104
		Word2Vec 160	0.3295	0.2062	0.0079
		Word2Vec 128	0.3246	0.2007	0.0075
		Word2Vec 64	0.3080	0.1884	0.0068
Exp. 5	HDBSCAN+ Distilli mBert	Distilli mBert 384	0.3686	0.2537	0.0116
		Distilli mBert 256	0.3672	0.2536	0.0116
		Distilli mBert 128	0.3517	0.2490	0.0090
		Distilli mBert 64	0.3352	0.2138	0.0077
Exp. 5	HDBSCAN + LDA	Topics 125	0.2747	0.1495	0.0038
		Topics 100	0.2552	0.1184	0.0038
		Topics 75	0.2365	0.1099	0.0039
		Topics 50	0.2315	0.2315	0.0680

جدول**24** : النتائج التفصيلية للتجربة الثالثة في وضع عدم الاتصال، في مرحلة كشف الأحداث.

#	Approach	Embedding Model	Results		
			NMI	AMI	ARI
Exp. 1	HDBSCAN + Arabert	AraBert 768	0.3759	0.2665	0.0160
		AraBert 320	0.3591	0.2434	0.0129
		AraBert 128	0.3413	0.2340	0.0128
		AraBert 64	0.3108	0.2063	0.0104
Exp. 2	HDBSCAN +mBert	mBert 768	0.4432	0.3194	0.0169
		mBert 512	0.4431	0.3188	0.0169
		mBert 256	0.4424	0.3170	0.0165
		mBert 128	0.4400	0.3157	0.0161
		mBert 64	0.4328	0.3052	0.0147
Exp. 3	HDBSCAN +FastText	FastText 300	0.3873	0.2582	0.0146
		FastText 160	0.3780	0.2536	0.0141
		FastText 128	0.3763	0.2502	0.0138
		FastText 64	0.3634	0.2314	0.0101
Exp. 4	HDBSCAN+Word2Vec	Word2Vec 300	0.3505	0.0087	0.2154
		Word2Vec 160	0.3511	0.0075	0.2183
		Word2Vec 128	0.3500	0.0083	0.2186
		Word2Vec 64	0.3421	0.0090	0.2171
Exp. 5	HDBSCAN + Distilli mBert	Distilli mBert 384	0.3479	0.2139	0.0084
		Distilli mBert 256	0.3438	0.2092	0.0074
		Distilli mBert 128	0.3468	0.2116	0.0076
		Distilli mBert 64	0.3350	0.2006	0.0062
Exp. 5	HDBSCAN + LDA	Topics 125	0.3879	0.2513	0.0133
		Topics 100	0.3774	0.2508	0.0154
		Topics 75	0.3665	0.2253	0.0106
		Topics 50	0.3600	0.2253	0.0139

جدول**25** : النتائج التفصيلية للتجربة الرابعة في ووضع عدم الاتصال، في مرحلة كشف الأحداث.

#	Approach	Embedding Model	Results		
			NMI	AMI	ARI
Exp. 1	SinglePass+ mBert	Graph +Node (512)	0.5885	0.4803	0.2320
		Graph +Node (128)	0.5993	0.4066	0.2409
		Graph +Node +mBert + PCA	0.6265	0.4118	0.3576
		Node level	0.5380	0.4362	0.1645
		Graph level	0.3125	0.3074	0.0428
		Graph + mBert(256)	0.5395	0.4668	0.1957
		Graph+ Node +mBer(256)	0.5488	0.4448	0.1838
Exp. 3	HDBSCAN+ AraBert	Node(128)	0.2831	0.1963	0.0067
		Graph + Arabert(128)	0.2790	0.2221	0.0093
		Node + Graph	0.3001	0.2137	0.0091
Exp. 4	HDBSCAN+ mBert	Node	0.4684	0.3612	0.0229
		Graph +mBer(256)	0.5145	0.4367	0.0327
		Graph+ Node +mBert(256)	0.4770	0.3953	0.0244
Exp. 5	HDBSCAN+ Distill Bert	Node	0.3736	0.2644	0.0131
		G + Distill Bert(128)	0.4015	0.2924	0.0149
		Graph+ Node+ Distill Bert	0.3985	0.2327	0.0102
		Graph + Node	0.3664	0.2655	0.0118