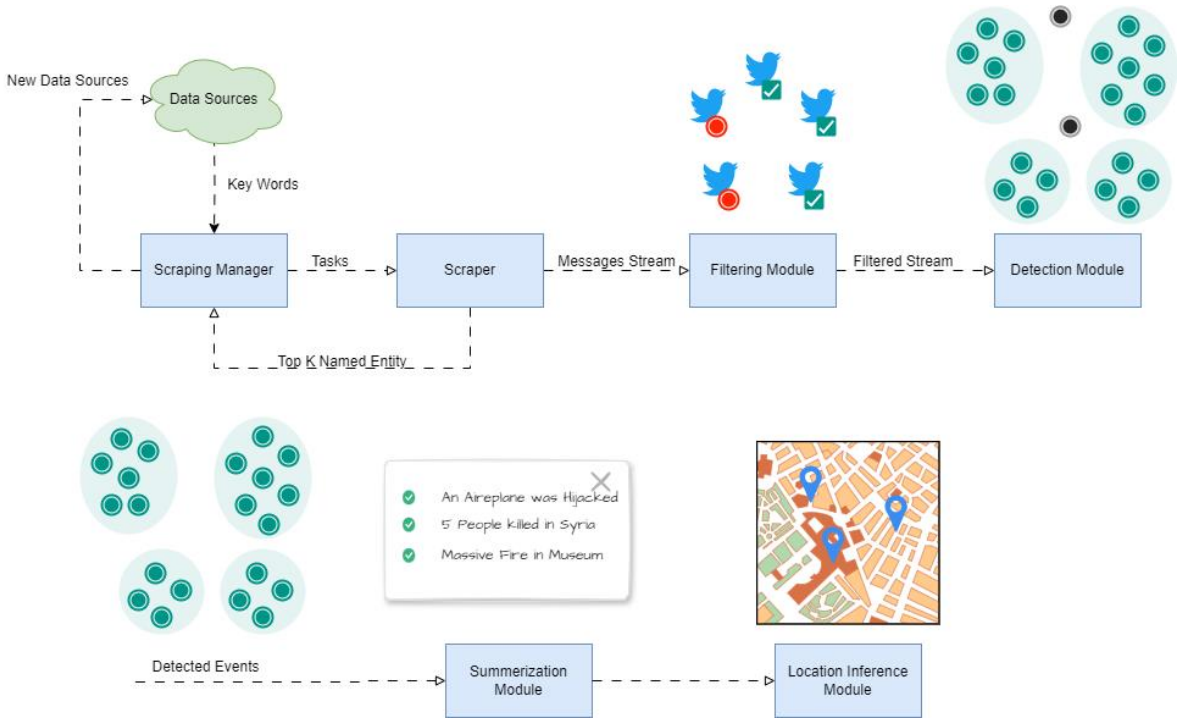
الفصل الخامس

المنهجية المقترحة

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

1.1- مقدمة

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

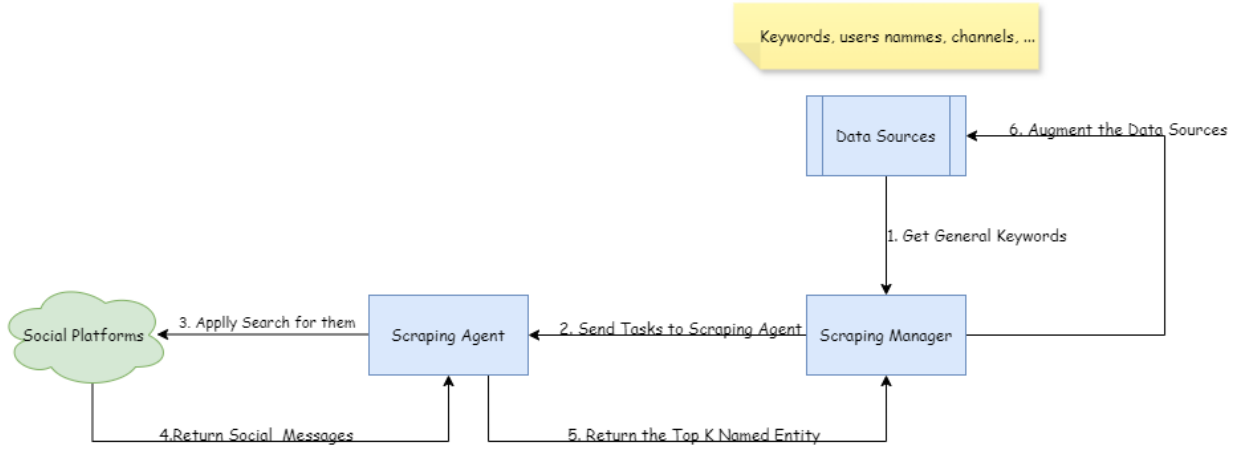


الشكل 1: البينة العامة للمنهجية.

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

وحدة البحث واستخراج دفق الرسائل

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



الشكل 2: 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 ال Scraping Manager قائمة بأكثر K كيان مسمى مكرر في نتائج البحث (مثل، مشفى ابن النفيس، قرية بلابلابلا، الشابة فلانة، ...) التي بدورها تستفيد منهم لتغذية مصادر البحث وتوليد كلمات مفتاحية جديدة وإرسالها من جديد للبحث عنها. حيث بهذا الشكل نكون قد غذينا مصادر البيانات بمصادر رائجة حالياً (بناءً على فرضية أنّه عند وقوع حادث في قرية او مكان ما فإنّ اسمه سيرد بشكل أكثر من بقية الكيانات المسماة.

وحدة فلترة دفق الرسائل

مسؤولية هذا الجزء هو فلترة الرسائل وإزالة الرسائل غير المرغوب بها (مثل الإعلانات والمنشورات الشخصية غير ذات الصلة بالمسألة)، وفي هذا الجزء اعتمدنا على بناء مصنف ثنائي، باستخدام المجموعة المنمطة التي جمعناها، حيث لجأنا لبناء هذا المصنف إلى نماذج تعلم الآلة التقليدية مثل (SVM, KNN, ...) لكون مجموعة البيانات غير كافية لاستخدام نماذج كبيرة ومن أجل بناء مصنف صغير بحيث يتم دمجه مع خدمة سحب البيانات لنحصل على وكيل سحب بيانات ذكي (Scraping Agent).

وحدة كشف الأحداث

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

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

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

وحدة تلخيص الحدث

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

وحدة الاستدلال على الموقع

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

الفصل الثامن

التجارب ومقارنة النماذج

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

1.1- مجموعة البيانات

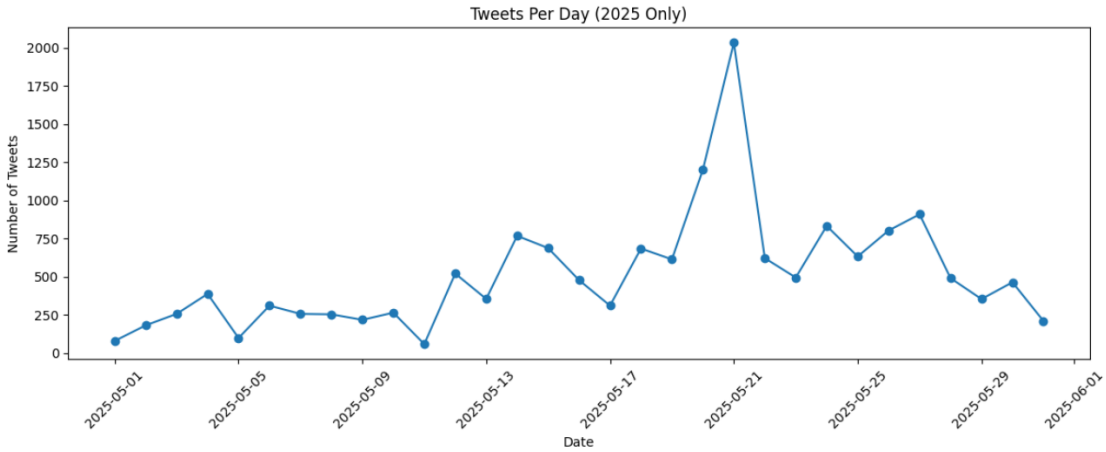
تم جمع مجموعة البيانات من منصة تويتر عبر البحث بكلمات مفتاحية عامة وخاصة (مثل، سوريا، دمشق، اعتصام الميدان، ...)، من تاريخ 1 أيار ولغاية 31 أيار 2025، ومن خلالها جمعنا مجموعة بيانات حجمها 20 ألف تغريدة (23 ألف قبل إزالة التكرارات) بالإضافة على معلوماتها metadata. سميناها SAS (Golden)، وبعد ذلك قمنا بتنميط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

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

جدول 1: مثال على تغريدات ضمن حدث ما.

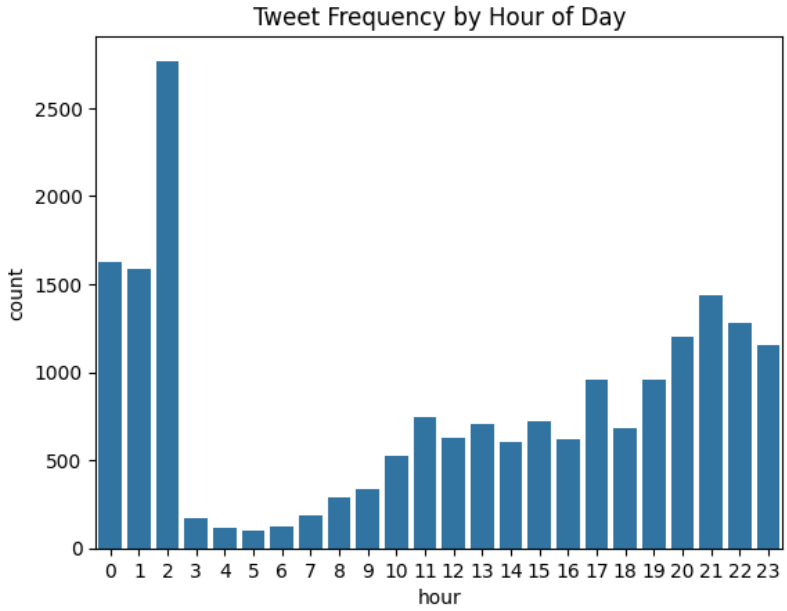
Author	Type	Tweet Text	Tweet ID
Grok	Reply	رامز الخياط هو رجل أعمال قطري، الرئيس التنفيذي لشركة أورباكون القابضة، وهي جزء من مجموعة باور إنترناشيونال القابضة التي يملكها ويديرها عائلة الخياط في قطر. تأسست أورباكون عام 2011، وتتخصص في مشاريع الطاقة والبنية التحتية. الخ	19281218
الإخبارية السورية	Tweet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

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



الشكل 3: 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)، كما أن أغلب المواطنين يبدون رأيهم عما حدث خلال اليوم ليلاً. ويساعد هذا المخطط على تحديد آلية وتوقيت سحب البيانات عند وضع النظام بالخدمة.

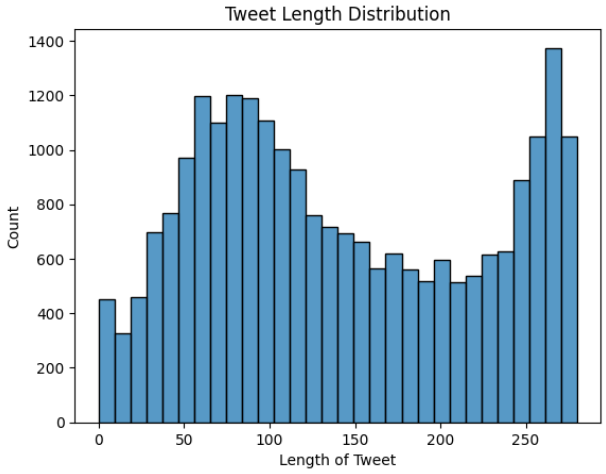


الشكل 4: تواتر التغريجات خلال ساعات اليوم.

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

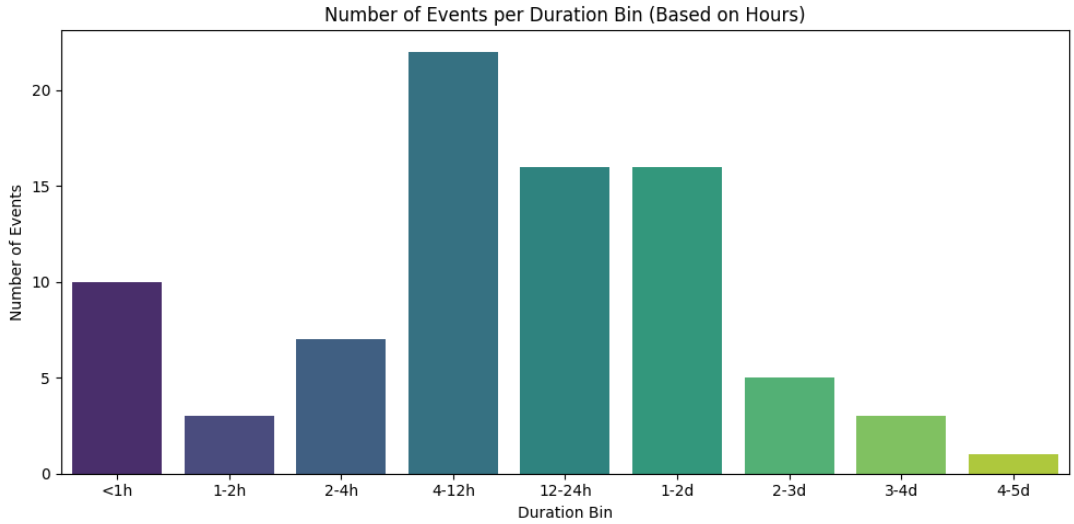
الشكل 5: مجموعة الكلمات المتداولة خلال 8 أيار- الفترة 9AM – 12 AM.

ويوضح الشكل أدناه، توزع طول التغريدات (بالمحرف).



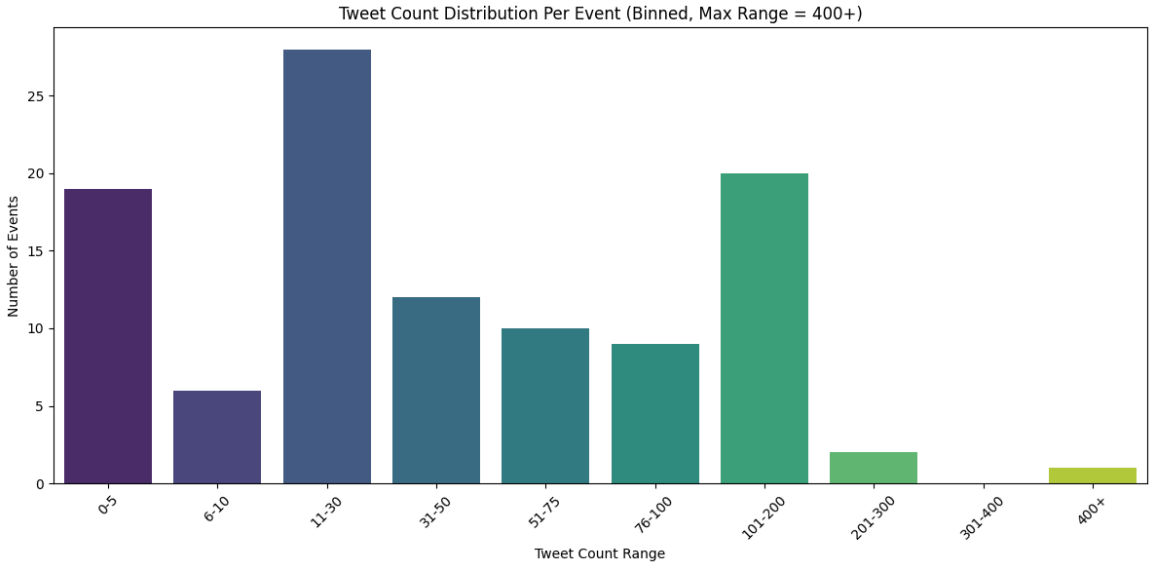
الشكل 6: توزع طول التغريدات.

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



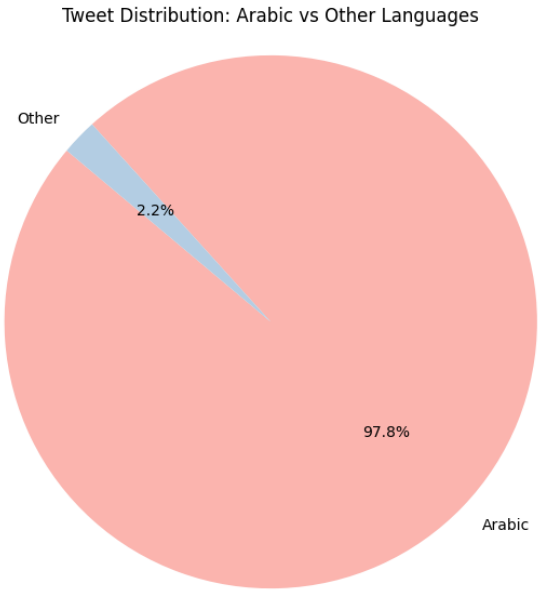
الشكل 7: المدة الزمنية للأحداث.

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.



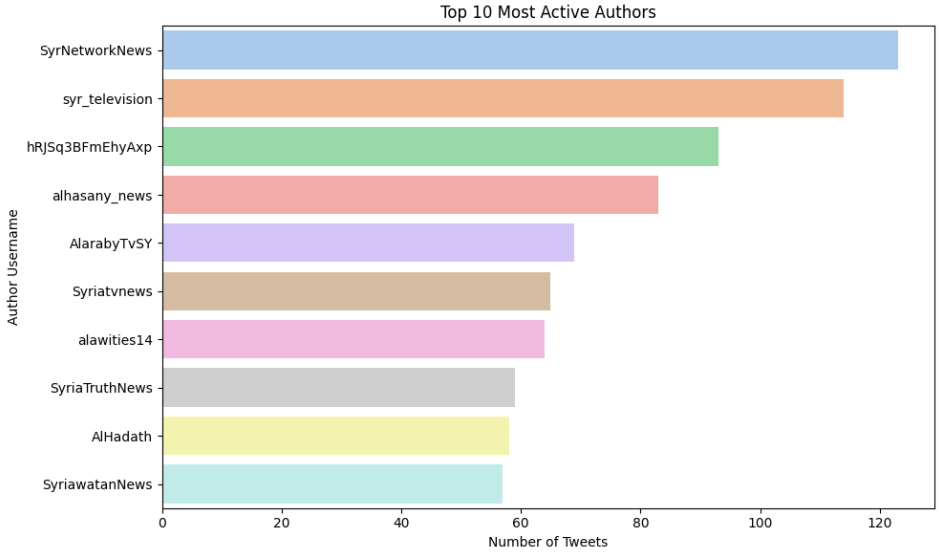
الشكل 8: مجال عدد التغريدات ضمن الحدث.

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



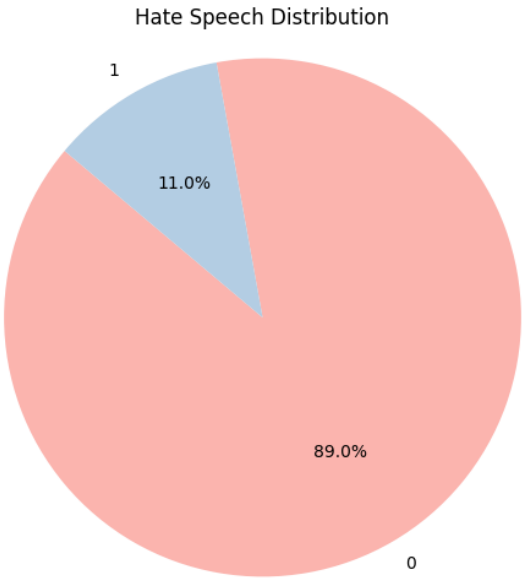
الشكل 9: توزع لغات التغريدات.

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



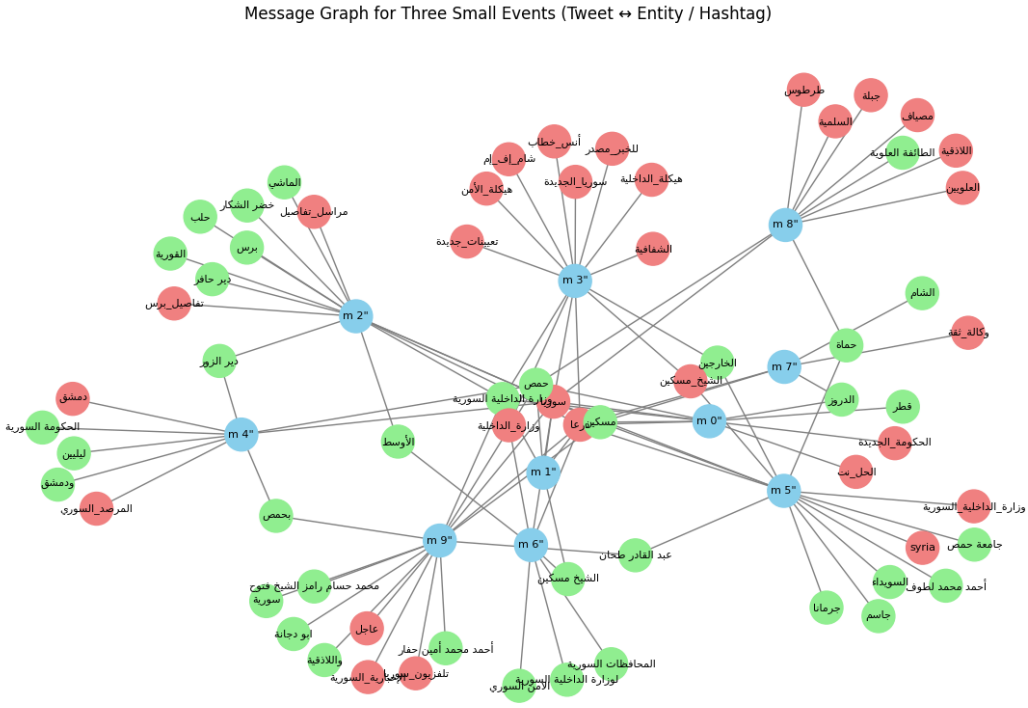
الشكل 10: مصادر البيانات الأكثر نشراُ للأحداث المتداولة.

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



الشكل 11: تحليل خطاب الكراهية والعنف.

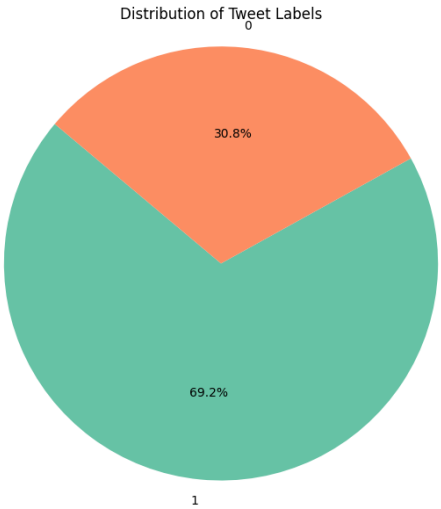
ويبين الشكل أدناه بيان المعلومات، لرسائل تتبع لأحداث مختلفة، ويلاحظ أن المستخدمين يلجؤون لإضافة hashtags ليست ذات صلة بالحدث ذات، مما يقل أهمية الـ hashtag بل وحتى الكيانات المسماة.



الشكل 12: بيان المعلومات للرسائل تتبع لأحداث مختلفة.

2.1-تجارب تصنيف التغريدات

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



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

الطرائق المقترحة، هي

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

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

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

وكانت النتائج وفق الجدول المبين أدناه.

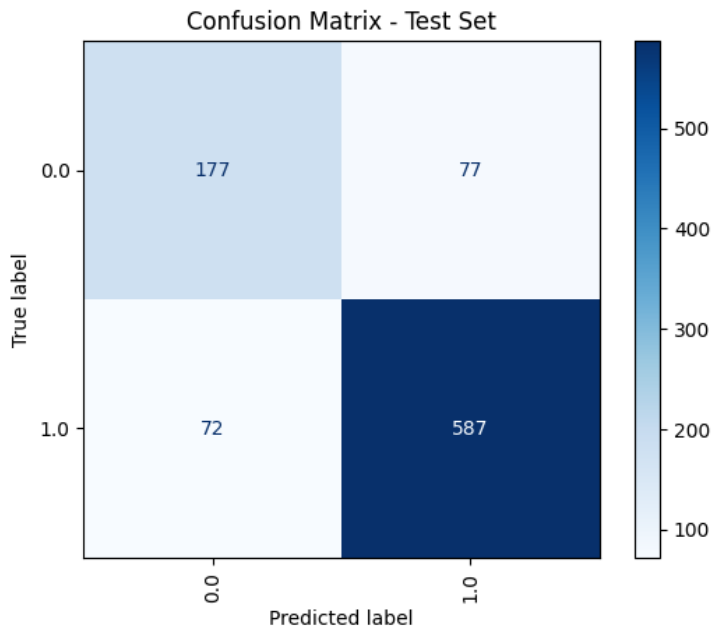
#	Name	Description	Model	Results							
				Accuracy		Recall		Precision		F1 Score	
				Train	Test	Train	Test	Train	Test	Train	Test
Exp. 1	FastText	Embedding Tweets with Fastext Model	Neural Network	0.84	0.82	0.90	0.89	0.86	0.86	0.81	0.77
			KNN	0.86	0.81	0.93	0.89	0.88	0.87	0.81	0.75
			SVM	0.89	0.84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

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

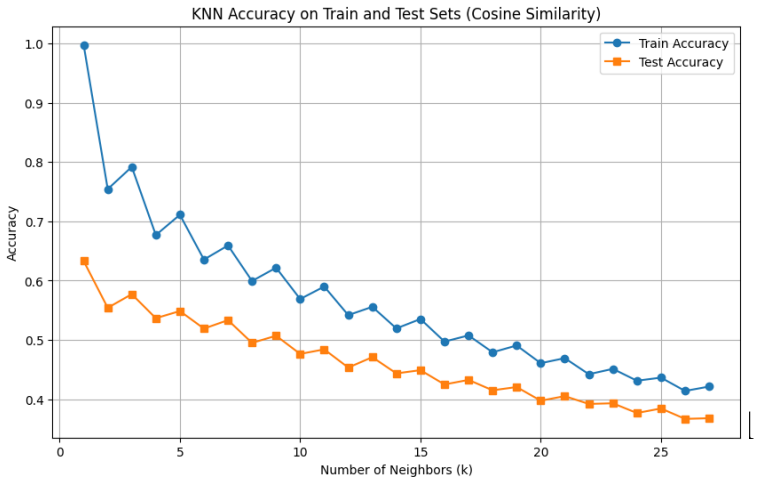
وكان الفرق بين نتائج التدريب والاختبار 3%.

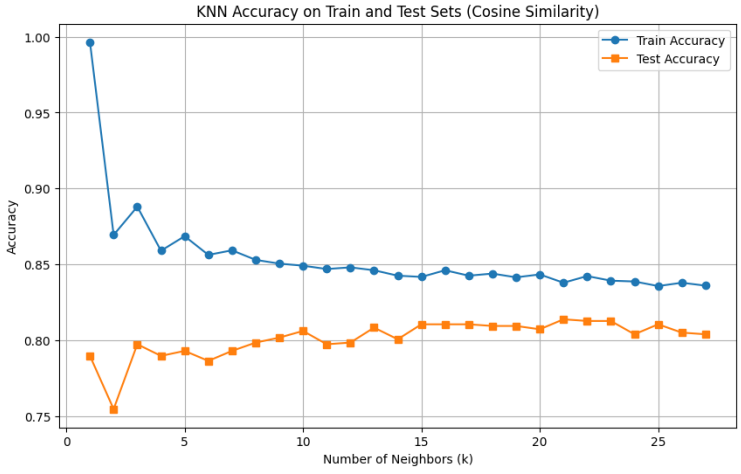
```
Classification Report - Test Set
              precision    recall  f1-score   support

         0.0       0.71      0.70      0.70       254
         1.0       0.88      0.89      0.89       659

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

ويوضح الشكل أدناه مصفوفة الارتباك له.



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





التجربة الثانية: تثقيل شعاع الكيانات المسماة

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

وكان النتائج وفق الجداول التالية.

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

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

التجربة الثالثة: تمثيل الكيانات المسماة فقط

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

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

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

التجربة الرابعة: استبدال الكيانات المسماة بكلمات بدل Placeholders

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

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

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

2.1-تجارب كشف الأحداث (نمذجة المواضيع)

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

1.1.1 تقييم النماذج بنمط عدم الاتصال Offline Evaluation

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

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

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

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

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

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

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

من أجل WMD، أخذ معهFastText كنموذج تضمين.

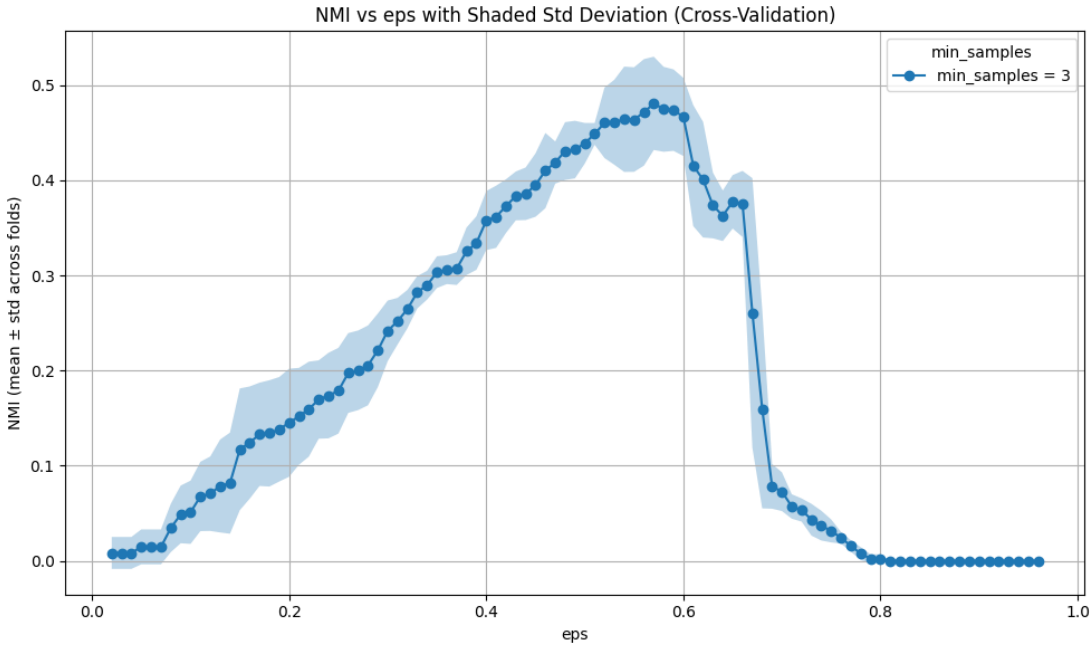
في التجربة الأولى أخذنا النماذج السابقة مع خوارزمية العنقدة DBSCAN.

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

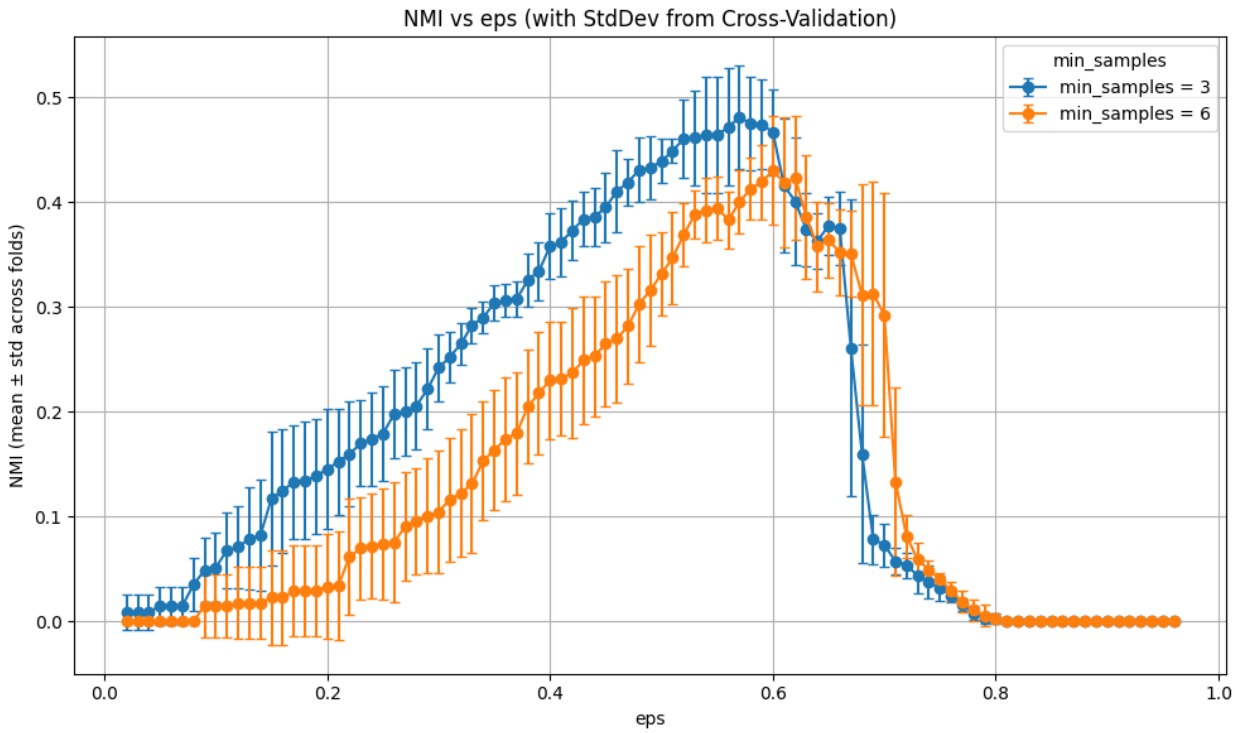
قمنا من أجل كل معامل بعمل K Cross Validation أثناء البحث عن المعامل الأفضل وأخذنا k=5.

- تحليل معامل المسافة min epsilon

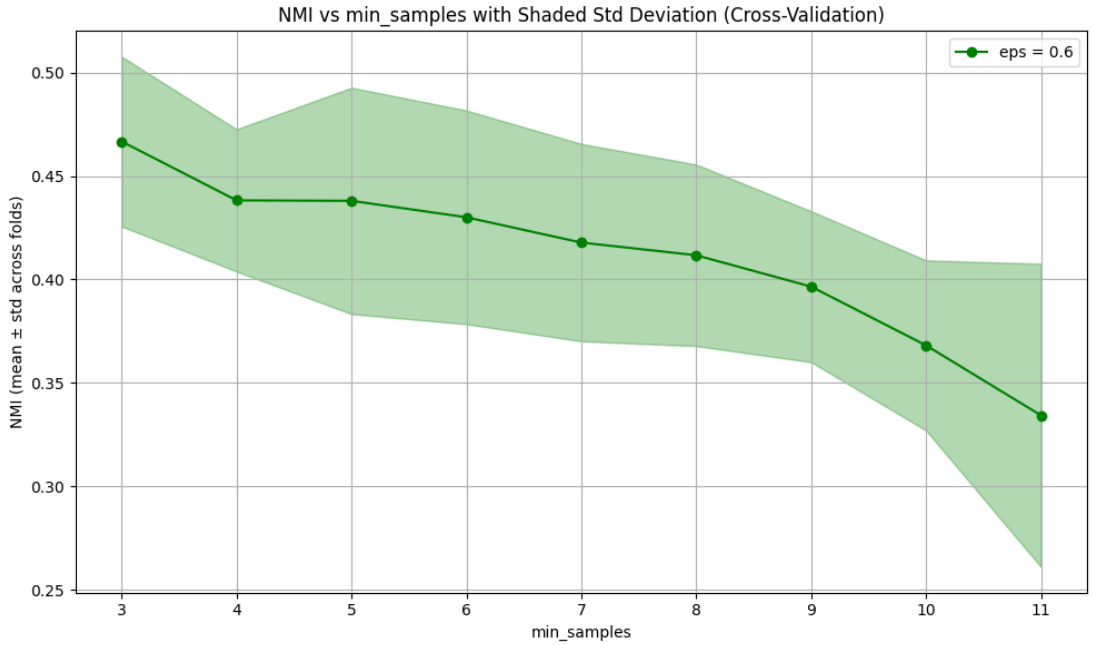
تختلف قيم هذا المعامل حسب طريقة التضمين، ولكن بمعظم التجارب كانت بين 0.1 و0.6، يوضح الشكل أدناه أثر هذا المعامل على نتائج العنقدة، حيث يبين من أجل كل قم له، متوسط نتائج الـ cross validation مع الانحراف المعياري، من أجل طريقة التمثيل mBert، (بقية الطرائق كانت النتائج مشابهة كشكل التابع مع تغير قيمة eps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



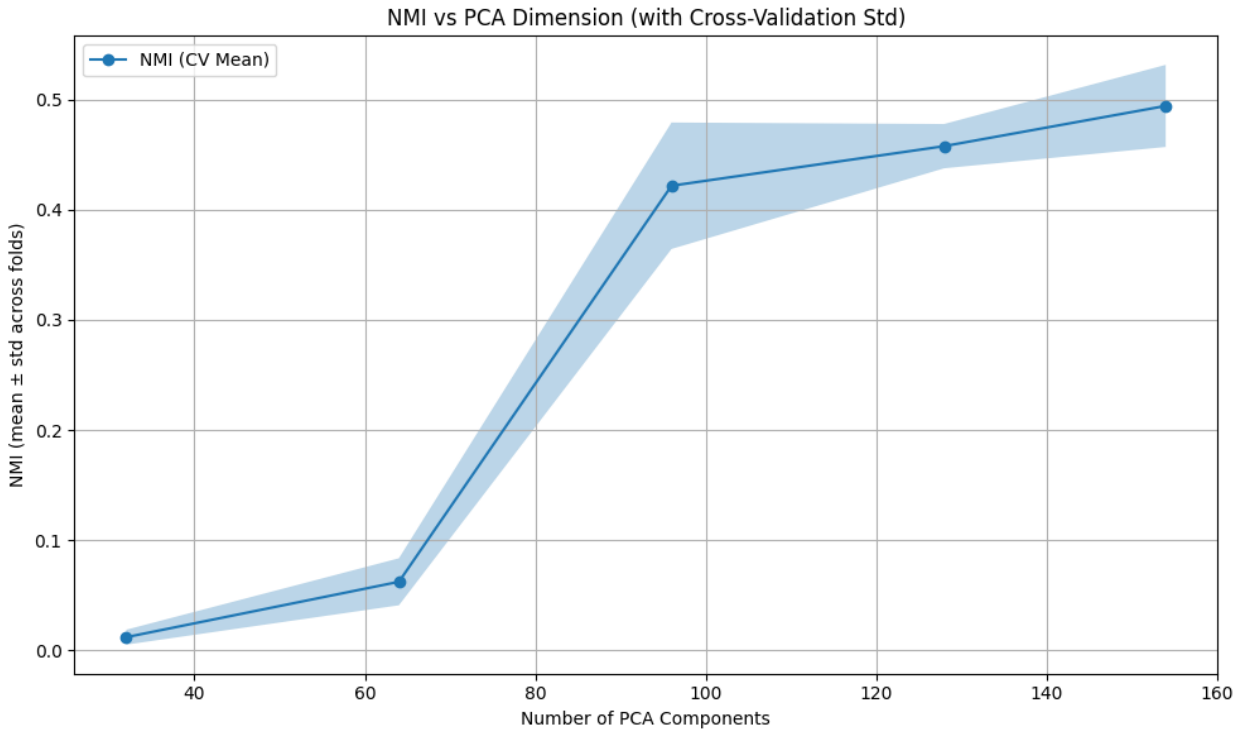
الشكل 13: أثر المعامل **eps** على النتائج باستخدام **mBert** لتمثيل الرسائل.



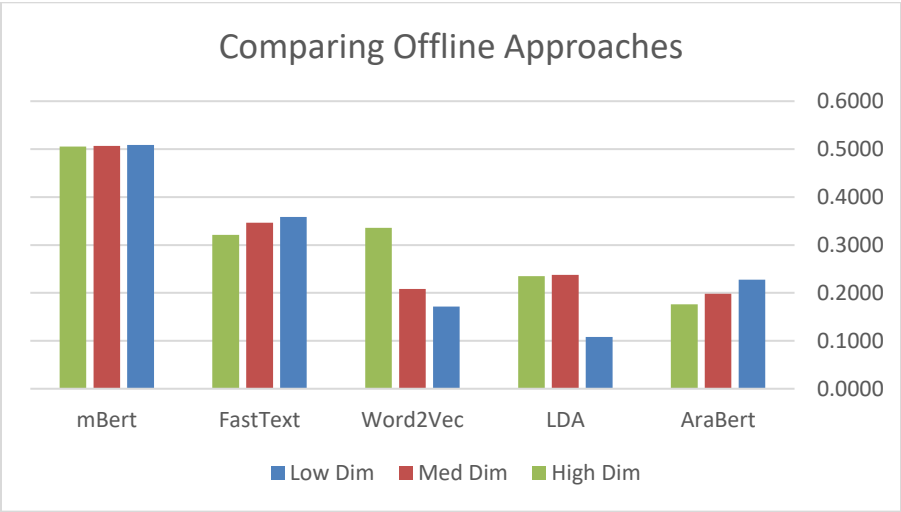
- تحليل معامل الكثافة min Samples، قيم هذا المعامل كانت في معظم الحالات بين 3 و7، القيم الكبيرة كانت تؤدي إلى نتائج غير جيدة، بسبب وجود عدد من الأحداث بعدد قليل من الرسائل، كان الإنحراف المعياري ما يقارب ∓ 0.4، ويشكل تقريباً نسبة $\mp 8\%$، مما يوحي بوجود overfitting، لذلك في التجارب القادمة سنلجأ إلى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



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



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

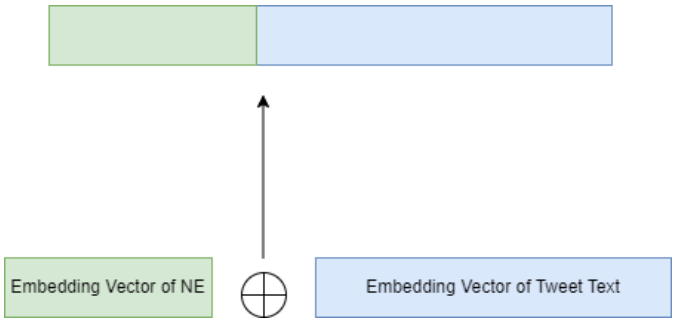


ويوضح الجدول أدناه بقية المقاييس.

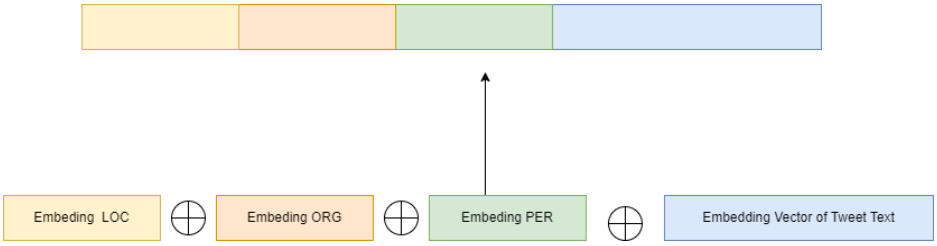
#	Approach	Description	Embedding Model	Results		
				NMI	AMI	ARI
Exp. 1	AraBert	DBSCAN + Arabert	AraBert 768	0.1765	0.0882	0.0039
			AraBert 320	0.1983	0.1030	0.0046
			AraBert 128	0.2278	0.1259	0.0059
			AraBert 32	0.2734	0.1881	0.0104
Exp. 2	mBert	DBSCAN +mBert	mBert 768	0.5056	0.4158	0.0295
			mBert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

التجربة الثانية: إضافة شعاع الكيانات المسماة

في التجربة الثانية قمنا بإضافة شعاع تمثيل الكيانات المسماة،



التجربة الثالثة: تجزئة شعاع الكيانات المسماة



التجربة الرابعة: تعزيز مجموعة البيانات

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

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

1.1.1 تقييم النماذج بنمط الحي Online Evaluation

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

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

جدول 2: إحصائيات كتل الرسائل.

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

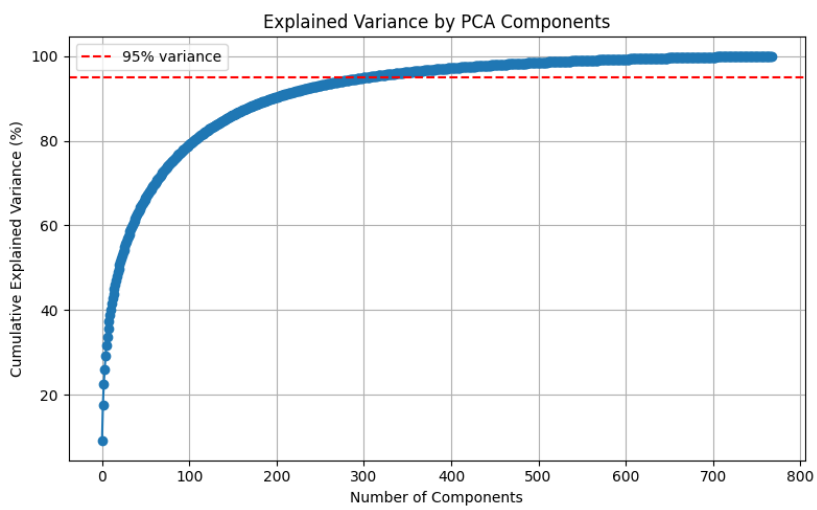
وأخذنا معايير التقييم هي NMI المعلومات المتبادلة المعيرة لمقارنة النماذج وتقييمها، كما قمنا بحساب المقاييس الأخرى AMI المعلومات المتبادلة المضبوطة، ARI مؤشر العشوائية المضبوط، وأرفقناه بالملحق، وهذه المعايير تم استخدامها في معظم الأبحاث المرجعية (المذكورون سابقاً). وأجرينا عدداً من التجارب، حيث بدأنا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

التجربة الأولى

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

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

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



ويوضح الجدول النتائج التي حصلنا عليها.

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

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

جدول 3: نتذاج التجربو الأولى.

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

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

وكانت من الأخطاء التي رأيناها من النمط

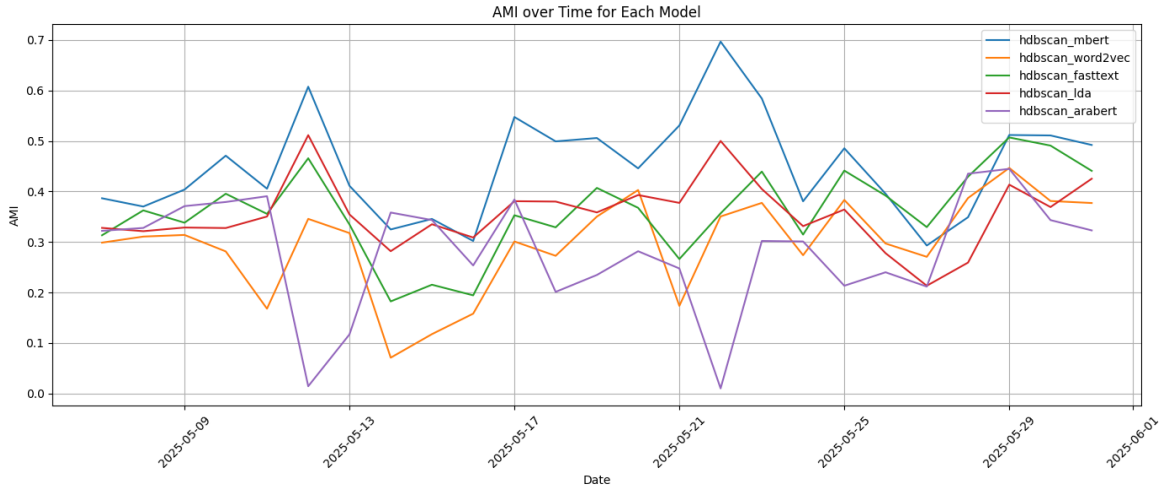
- تجميع الرسائل حسب معناها (أي مثلاً رسائل الحرائق تقع بالقرب من بعضها في فضاء التمثيل على الرغم من أن الحرائق تحدث في أماكن مختلفة، مما يؤدي إلى تداخل في الأحداث)
- كون العنقدة تحدث على كامل المجموعة كان هناك أحداث وقعت في نفس المكان ولكن بزمنين مختلفين (العثور على مستودع *** في ريف دمشق)
- كان هنالك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

وعليه:

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

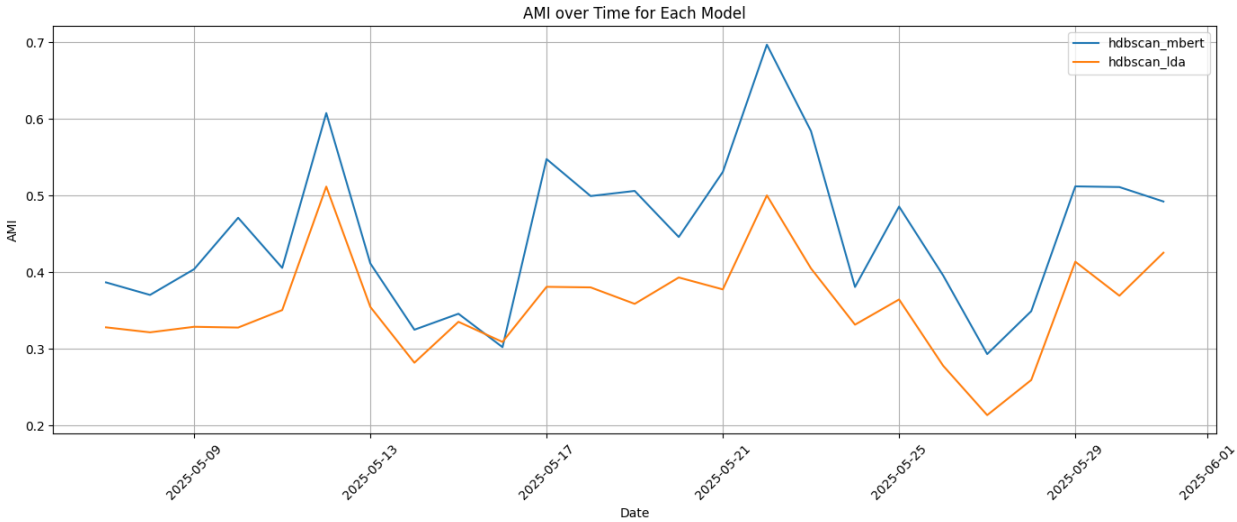
الأخطاء من النمط الثاني، يمكننا تفاديها عند أخذ الزمن بالاعتبار.

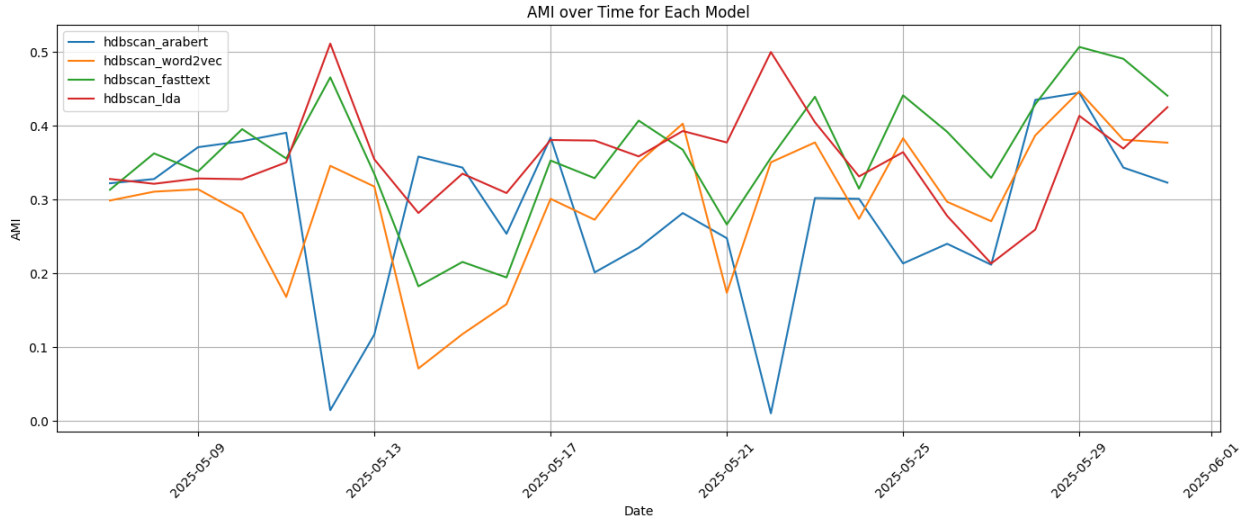
والأخطاء من النمط الرابع هي غير قابلة للتفادي في العنقدة (مما يوجب علينا إضافة مرحة معالجة لاحقة لإزالتها).

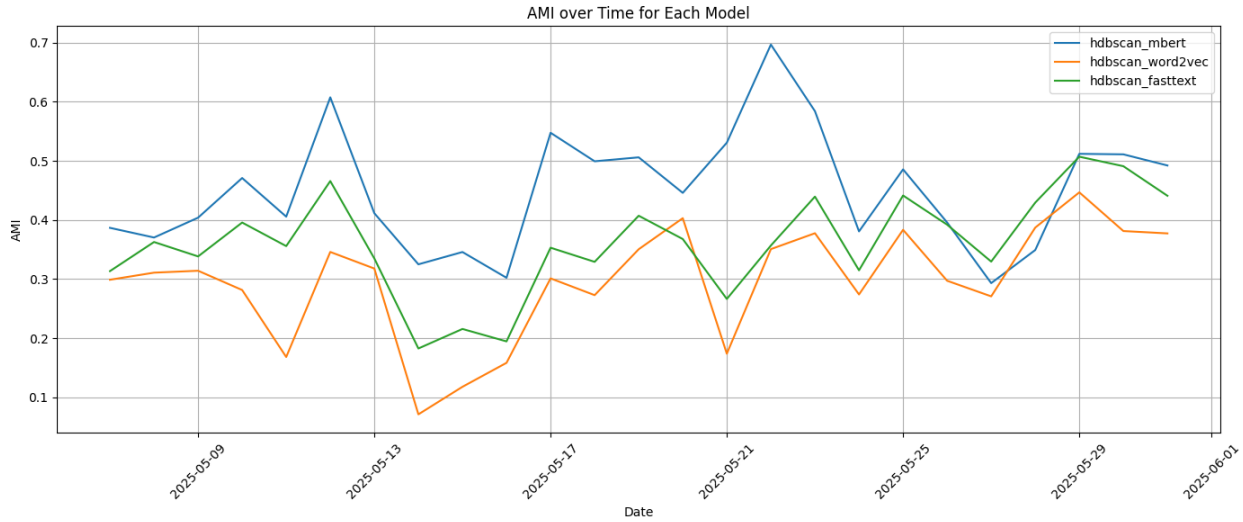


ويوضح الشكل أعلاه مخطط قيم NMI للنماذج المستخدمة، حيث وكما ذكرنا كان mBert هو الأفضل يليه LDA. وكانت نتائج بقية النماذج متقاربة.

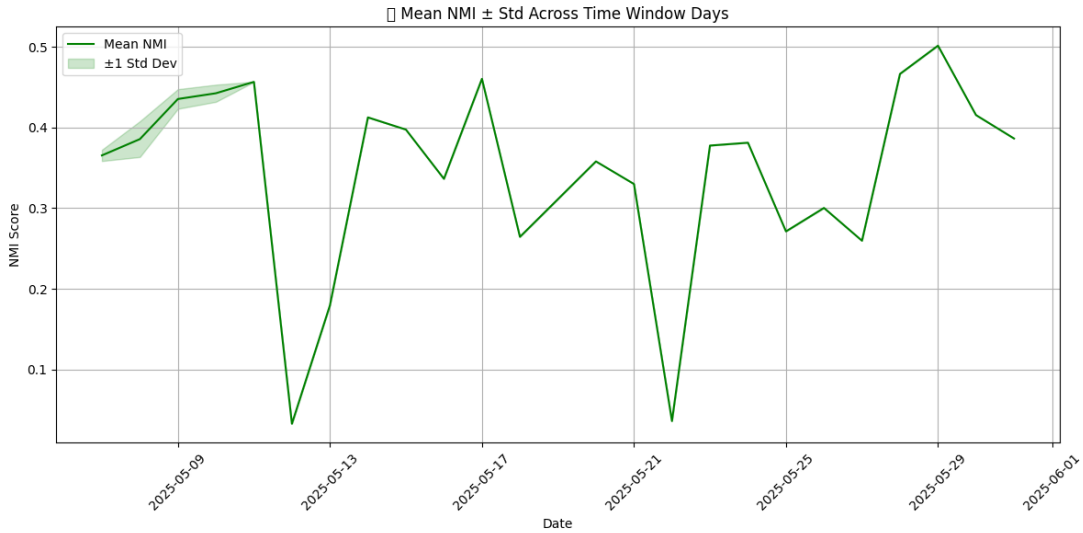
ويوضح الشكل أدناه معيار التقييم AMI، حيث كما نلاحظ أيضاً أنّ mBert كان الأفضل أيضاً بناءً عليه.





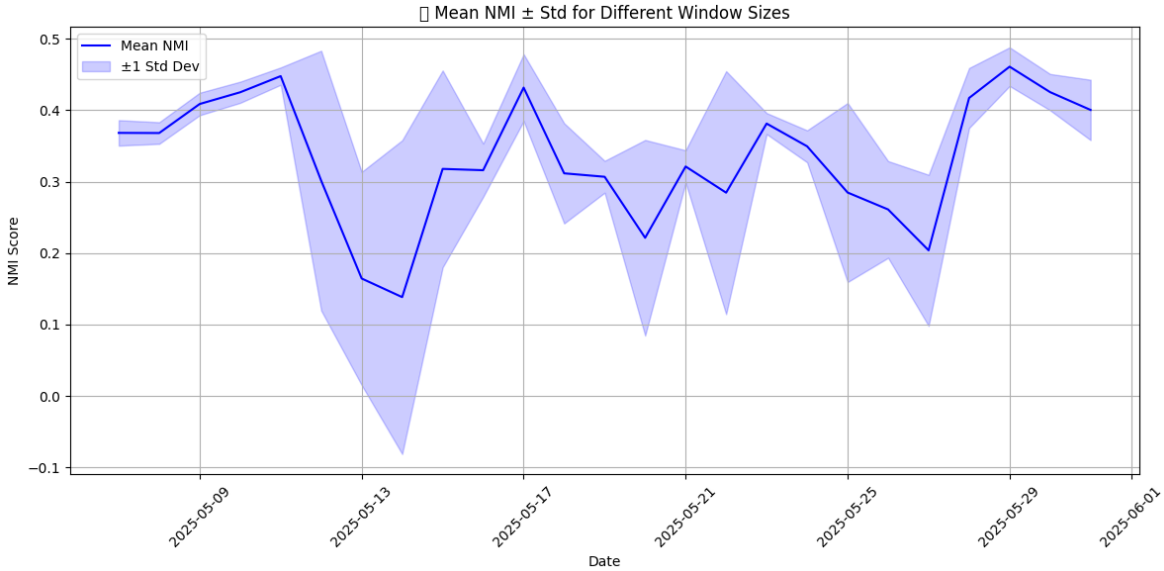


ويوضح الشكل أدناه أثر حجم النافذة الزمنية (معامل النسيان T)، على النتائج، حيث فمنا بالاختبار من أجل القيم (2 إلى خمسة أيام) وأخذنا متوسط القيم وانحرافها المعياري وكانت النتائج متقاربة جداً، حيث يوضح الشكل أدناه المتوسط والانحراف المعياري للنتائج لقيم مختلفة لـ T، حيث كانت قيم الانحراف المعياري صغيرة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



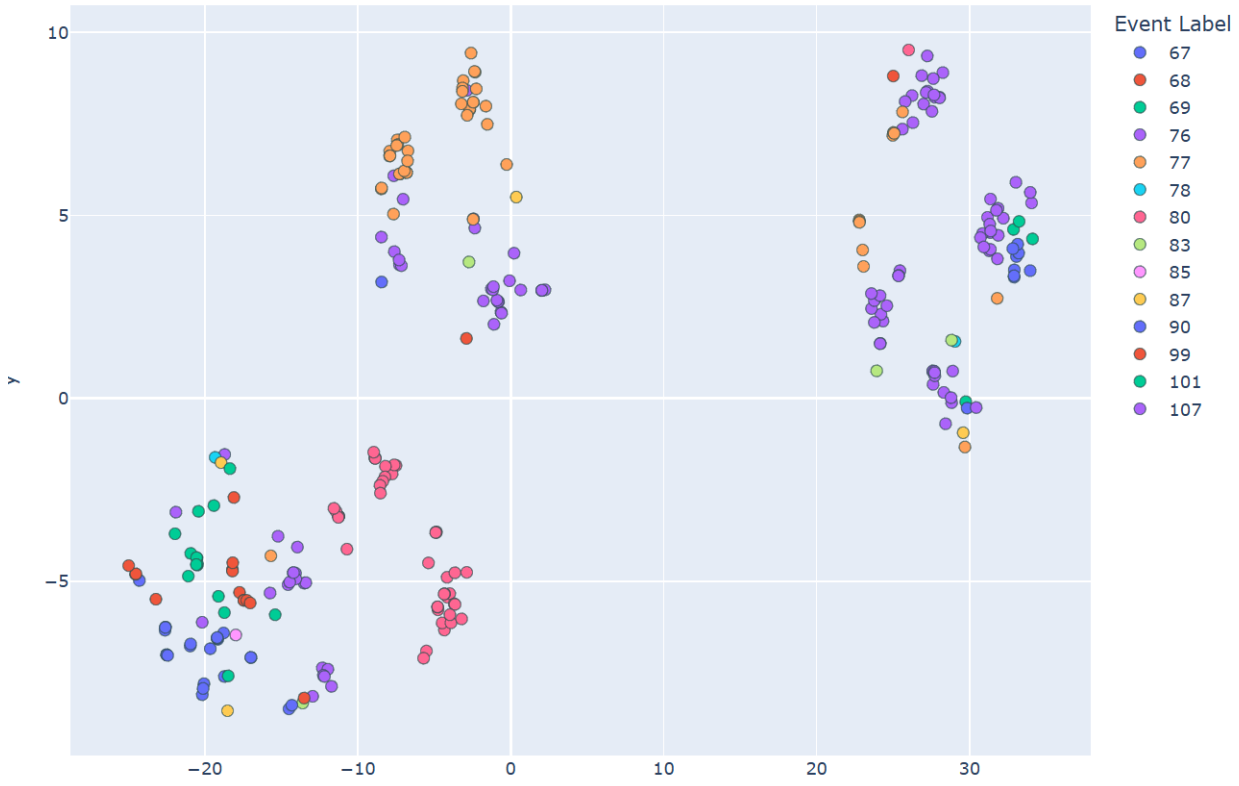
الشكل 14: أثر معامل حجم النافذة الزمنية على نتائج النموذج.

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.



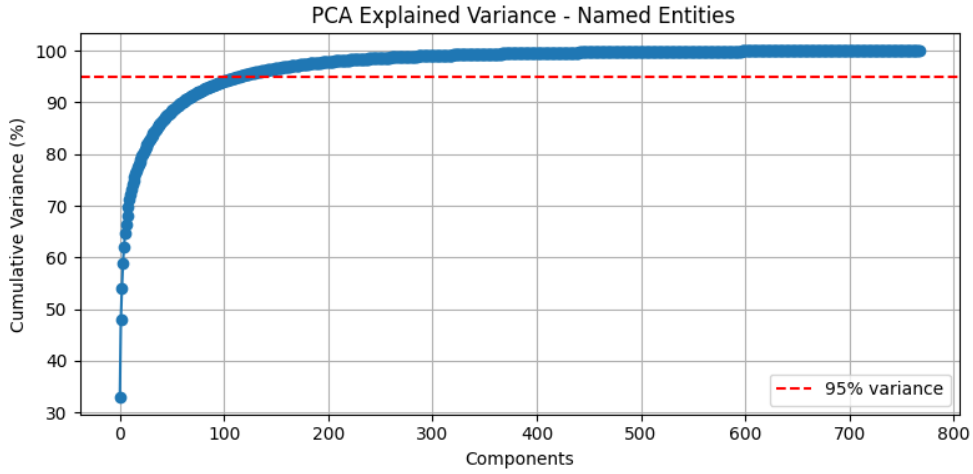
الشكل 15: أثر معامل حجم النافذة على النتائج.

ويوضح الشكل أدناه أسقاط تمثيل الرسائل على فضاء ثنائي الأبعاد.



التجربة الثانية: أضافة شعاع الكيانات المسماة

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

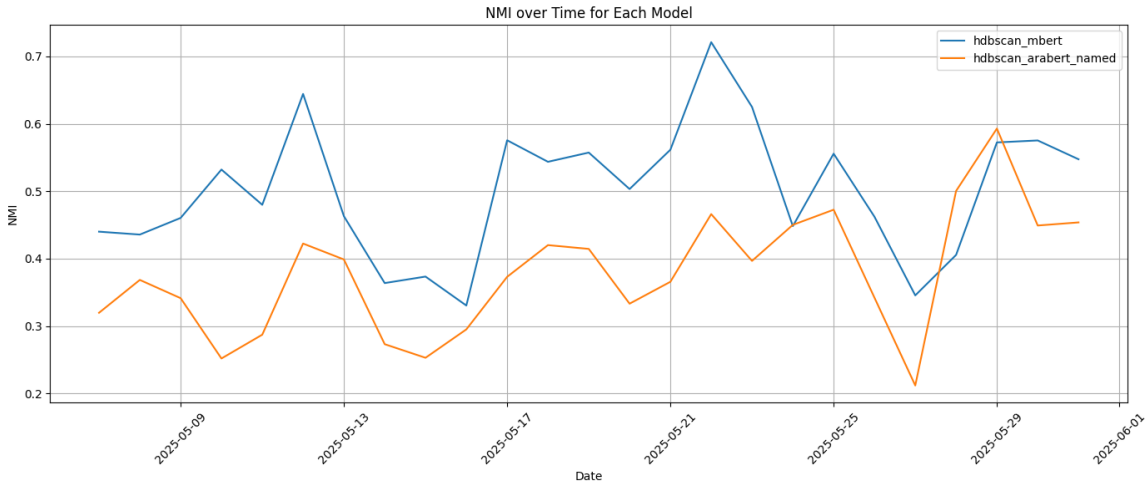


في هذه التجارب لم نحصل على قيمة مضافة على النتائج، حيث بقي mBert أفضل من AraBert + NE بـ 25%.

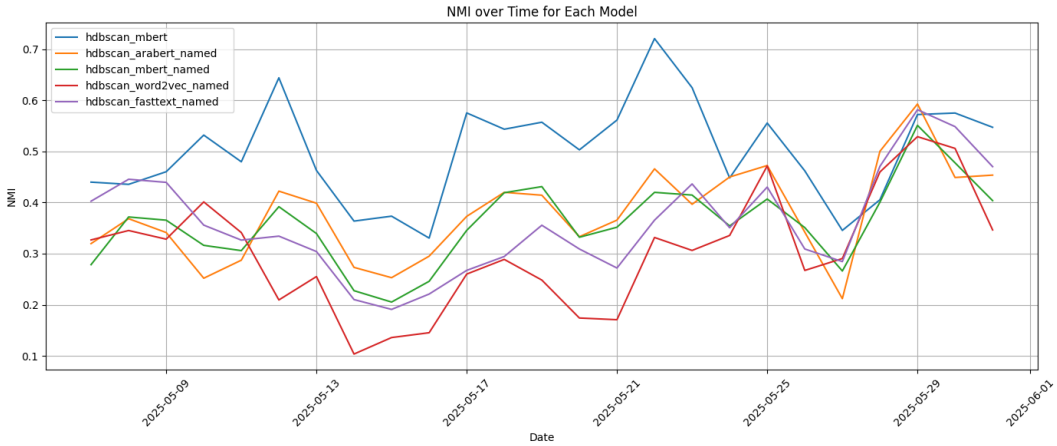
	M_1	M_2	M_3	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
mbert +NE	0.278	0.372	0.366	0.316	0.306	0.392	0.339	0.228	0.205	0.246	0.346	0.419
AraBert +NE	0.320	0.368	0.341	0.252	0.287	0.422	0.399	0.273	0.253	0.295	0.373	0.420
Word2Vec +NE	0.327	0.345	0.328	0.401	0.341	0.209	0.255	0.103	0.136	0.145	0.260	0.289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

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

ويضح الرسم أدناه الفارق بينهما

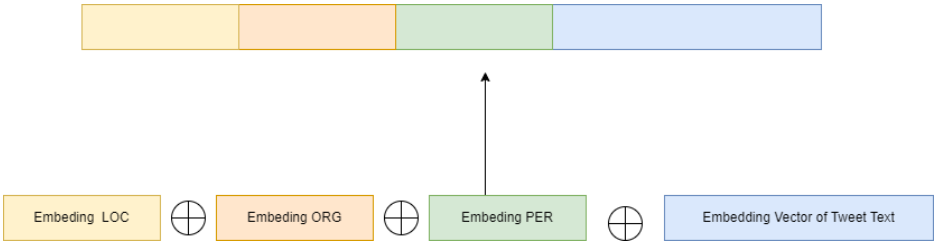


كما ويبين الشكل أدناه مقارنة الطرائق المتبعة في هذه التجربة مع mBert.



التجربة الثالثة: أضافة شعاع الكيانات المسماة بشكل منفصل

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



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

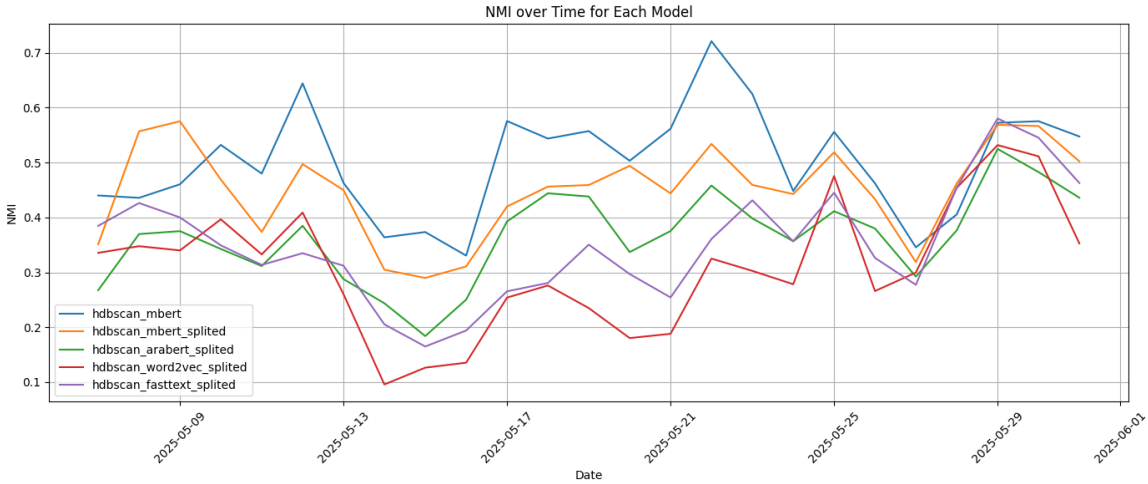
ويوضح الجدول أدناه النتائج التي حصنا عليها.

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

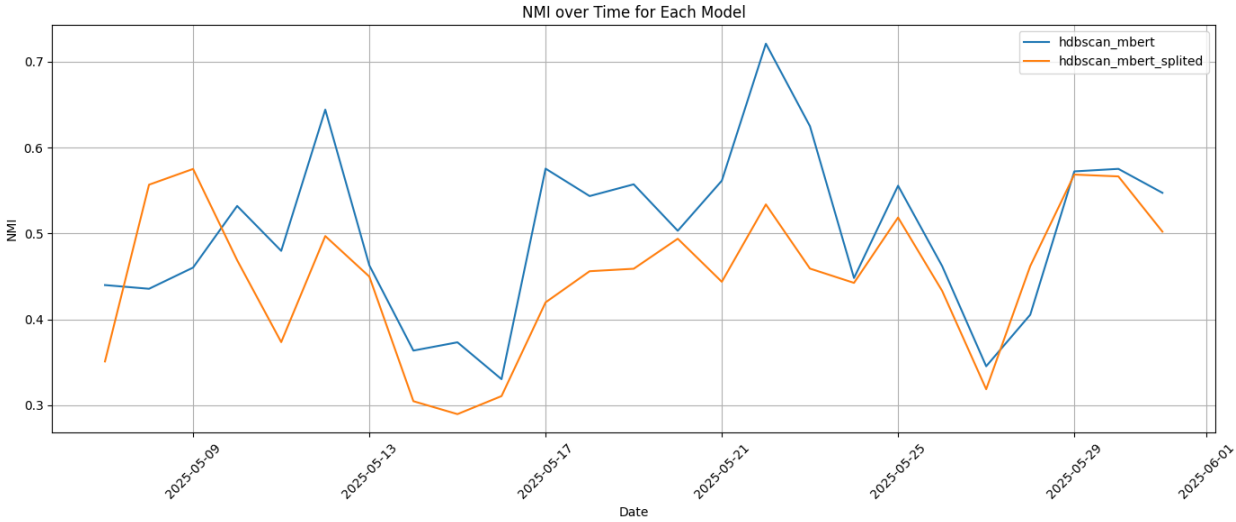
	M_{12}	M_{13}	M_{14}	M_{15}	M_{16}	M_{17}	M_{18}	M_{19}	M_{20}	M_{21}	M_{22}	M_{23}	M_{24}
mbert	0.459	0.494	0.444	0.534	0.459	0.443	0.519	0.433	0.319	0.462	0.569	0.567	0.502
AraBert	0.438	0.337	0.375	0.458	0.398	0.357	0.411	0.380	0.292	0.377	0.525	0.483	0.436
Word2Vec	0.235	0.180	0.188	0.325	0.303	0.278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

جدول 4: نتائج التجربة الثالثة.

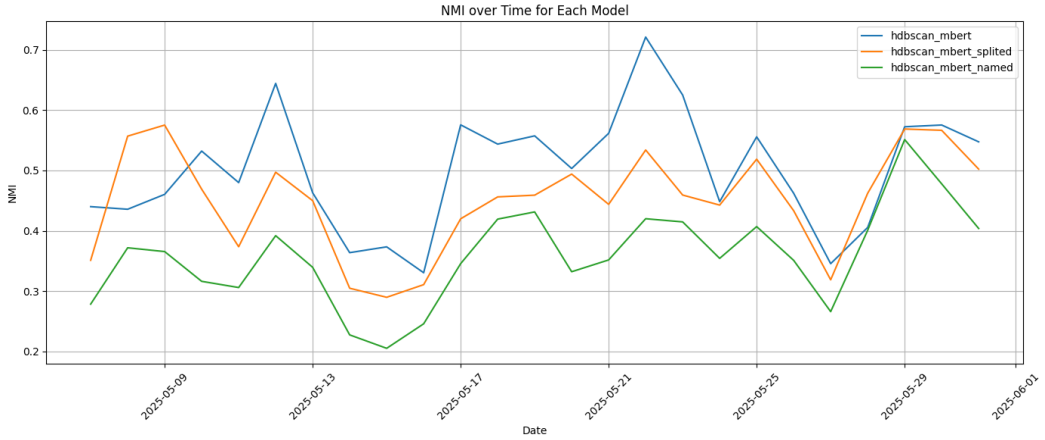
فيما يوضح الرسم البياني نتائج التجربة ومقارنتها مع mBert.

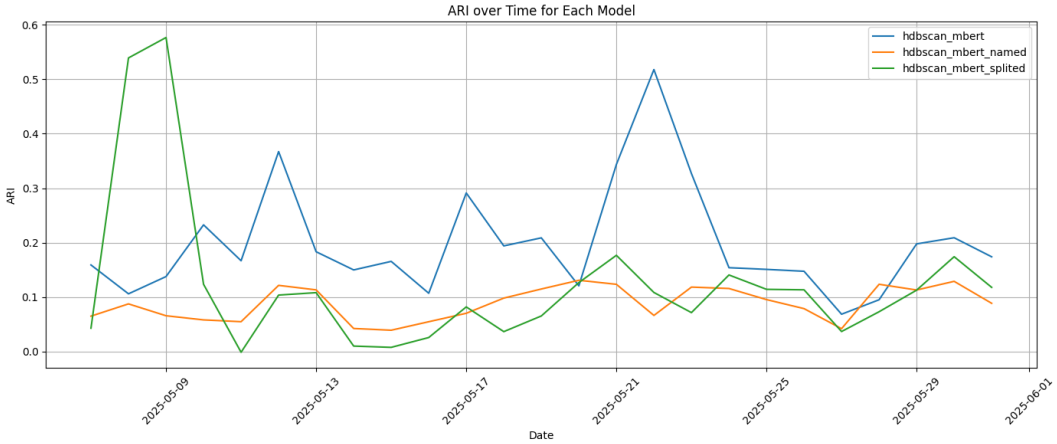


ويبين الرسم أدناه مقارنة mBert مع أفضلهم.

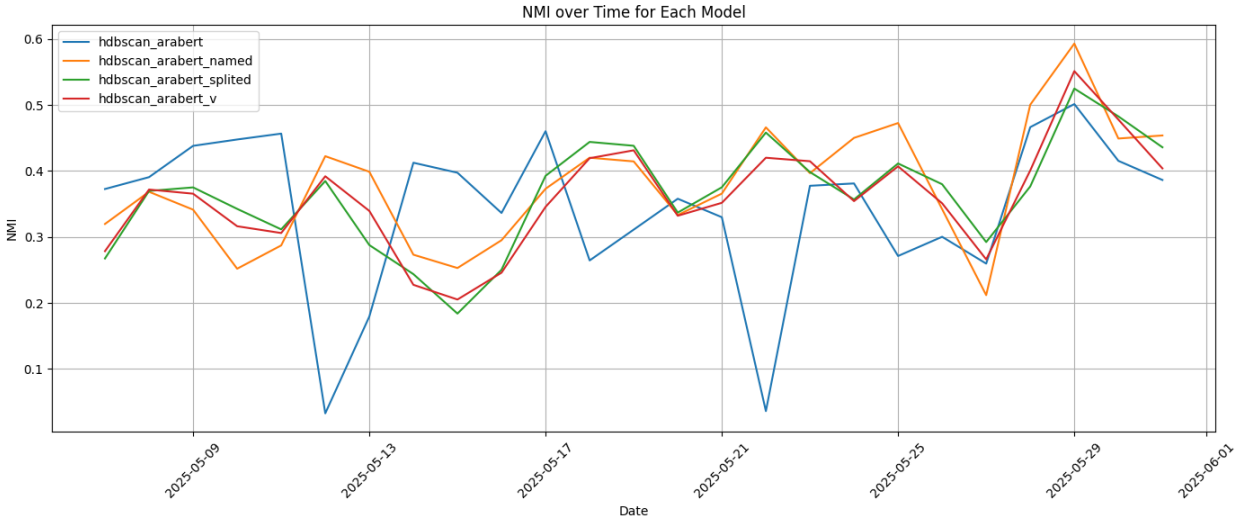


فيما يبين الرسم أدناه مقارنة نتائج النموذج مع إضافة الاشعة إليه.

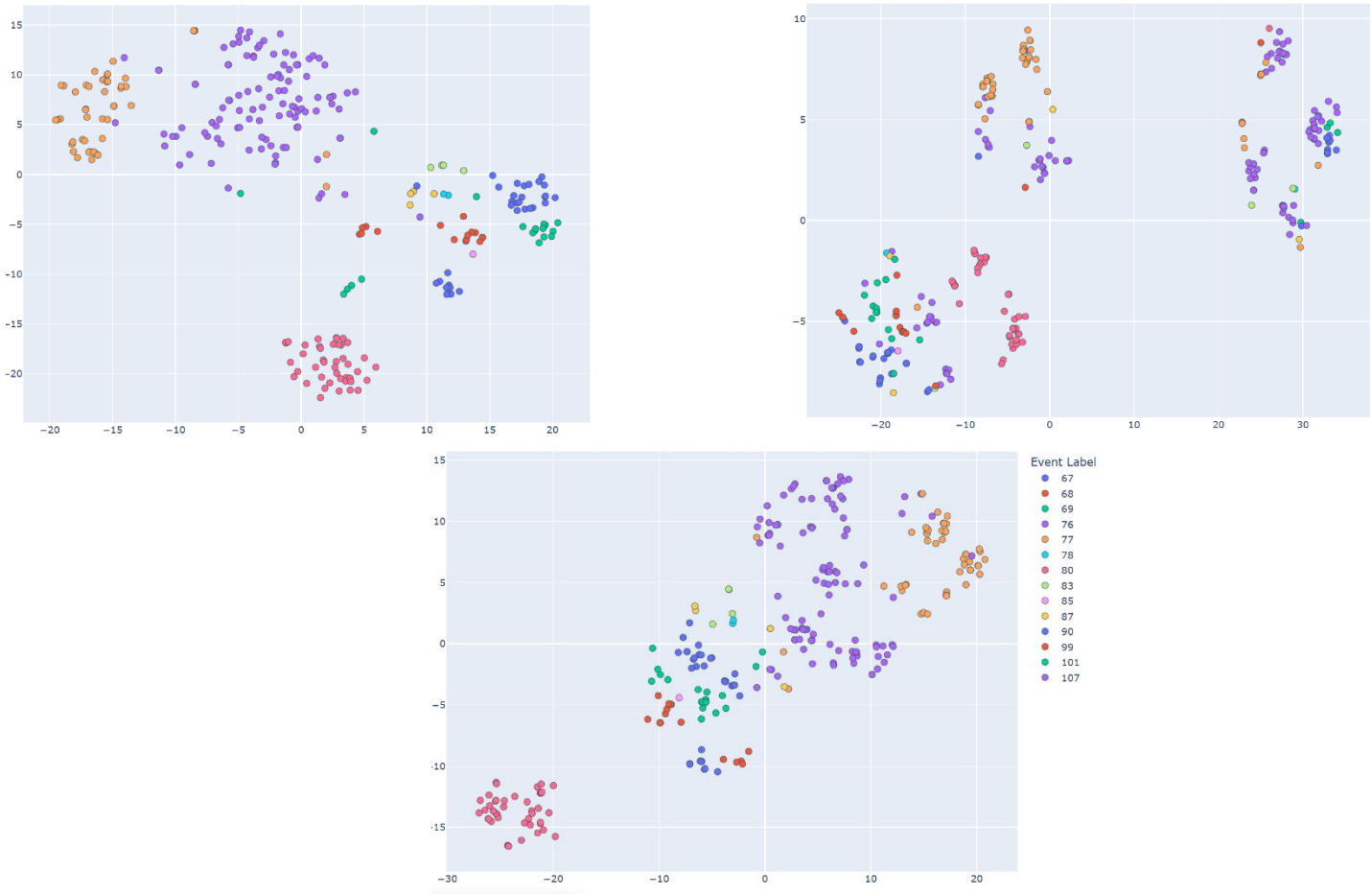




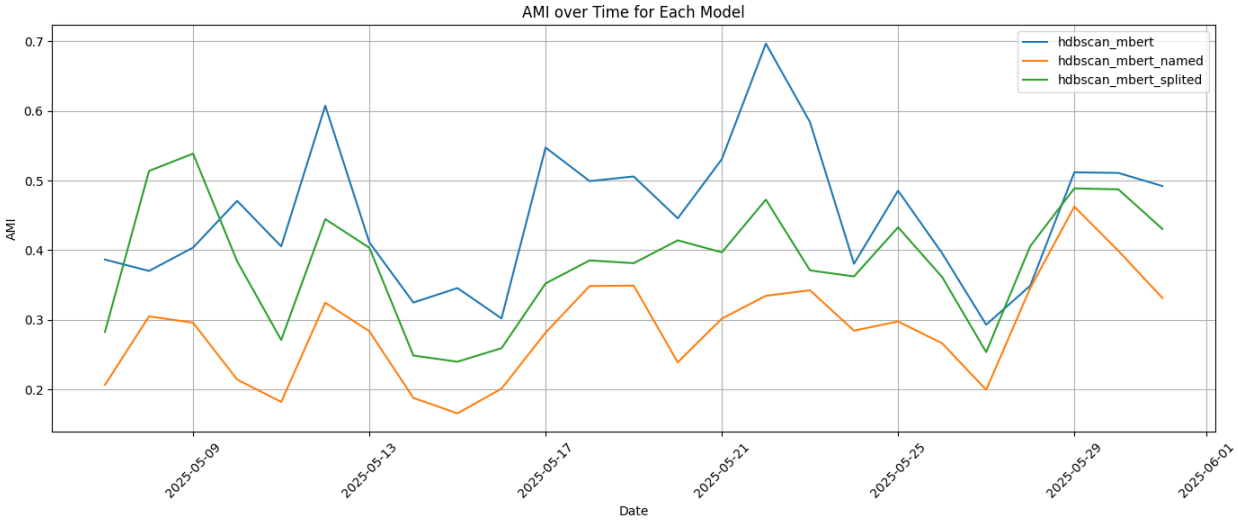
ويبين الشكل أدناه كيف أن الإضافات التي قمنا بها عل التمثيل جعلت التميل باستخدام arabert مستقراً أكثر.

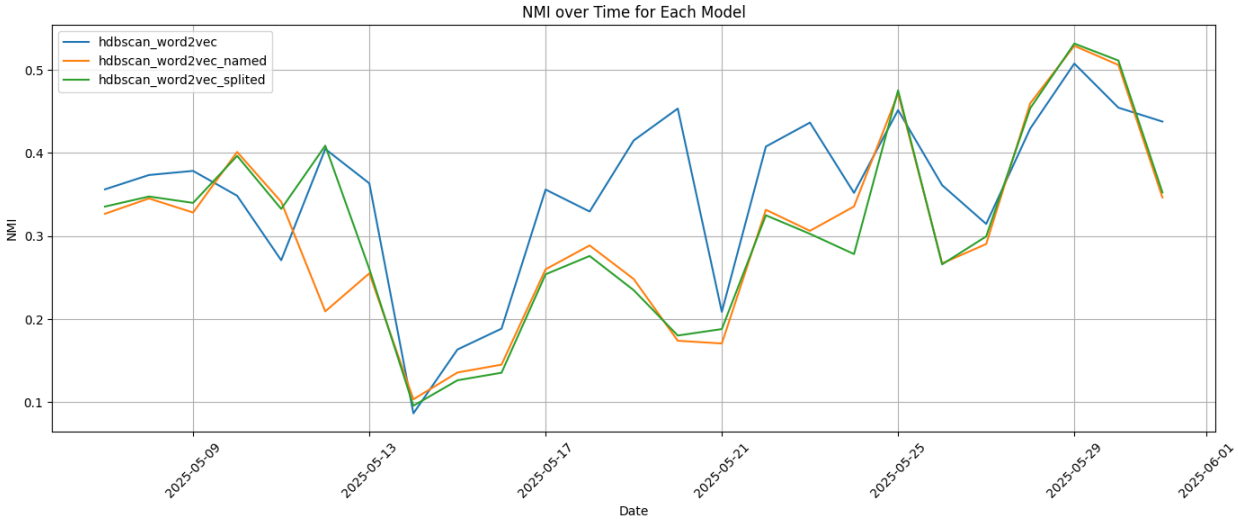


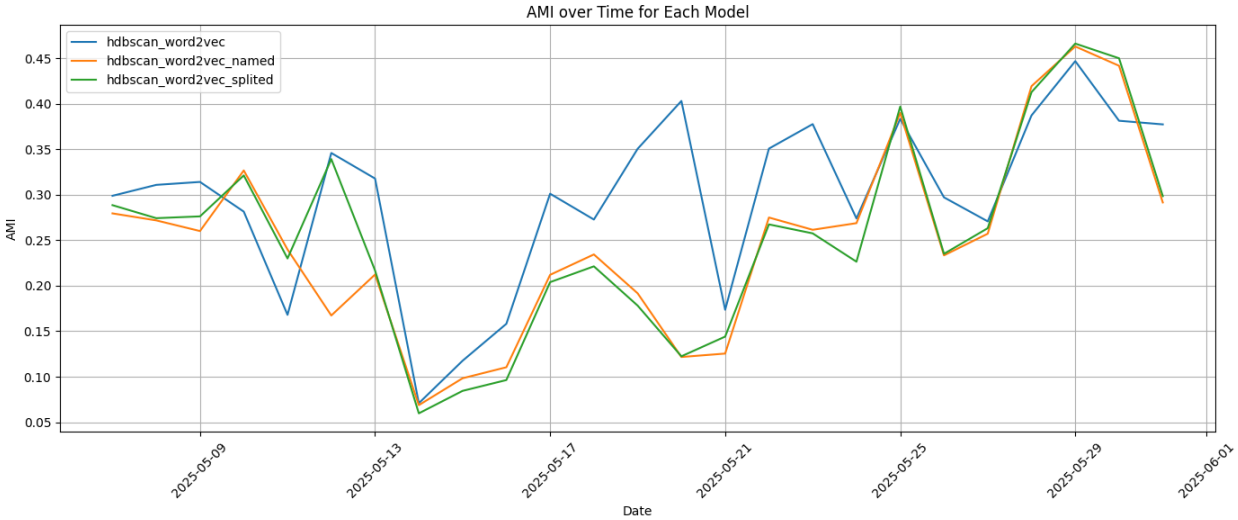
ويبين الشكل أدناه نتائج العنقدة على كتلة رسائل، في التجارب السابقة.

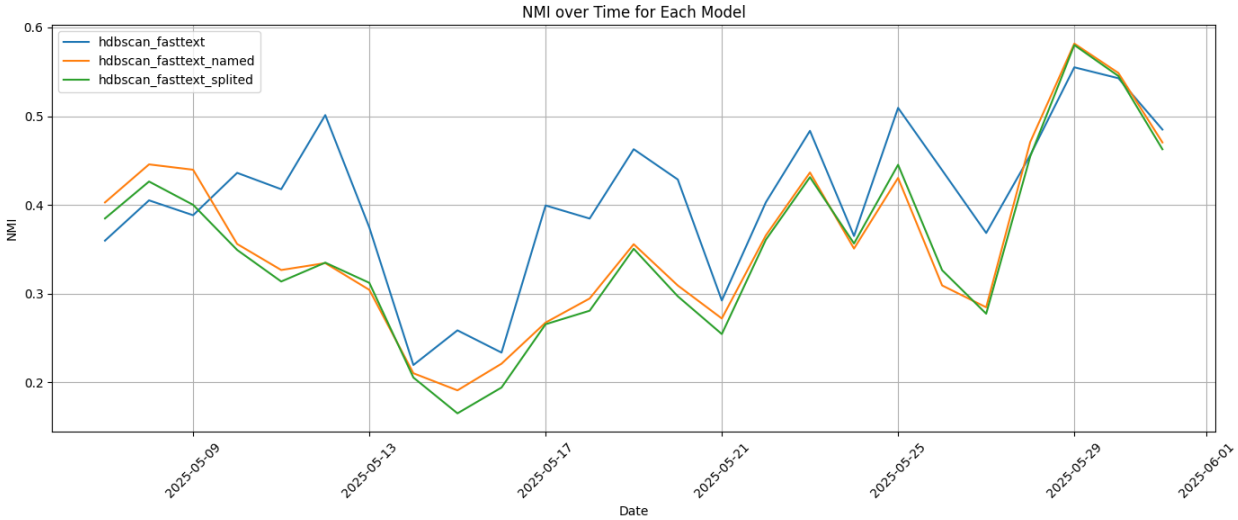


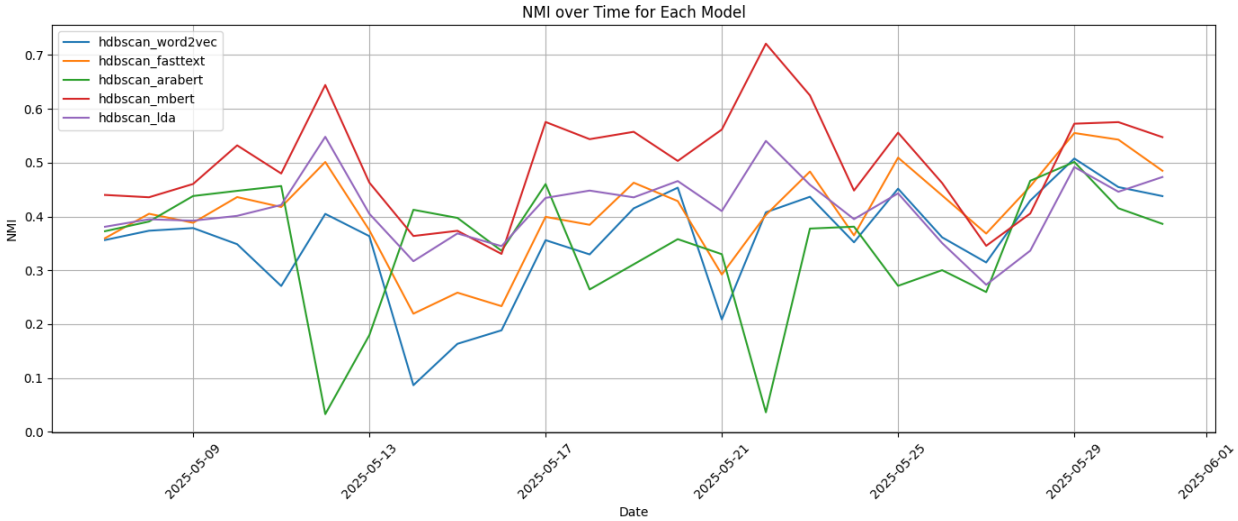
التجربة الرابعة: تعزيز مجموع بيانات الضبط

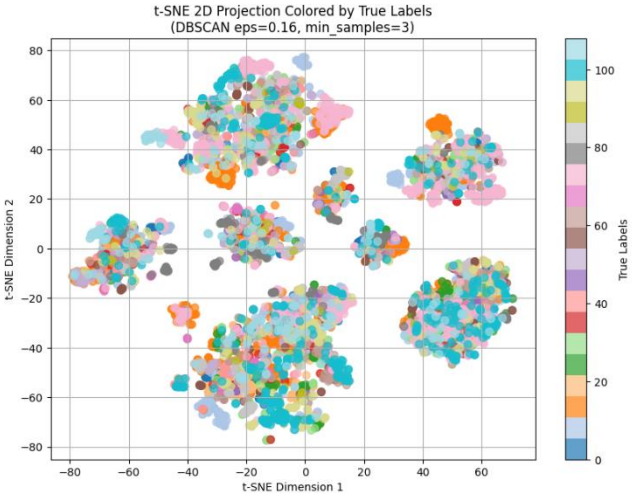

AMI over Time for Each Model
hdbscan_mbert
hdbscan_mbert_named
hdbscan_mbert_splited
AMI
Date
0.7
0.6
0.5
0.4
0.3
0.2
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01

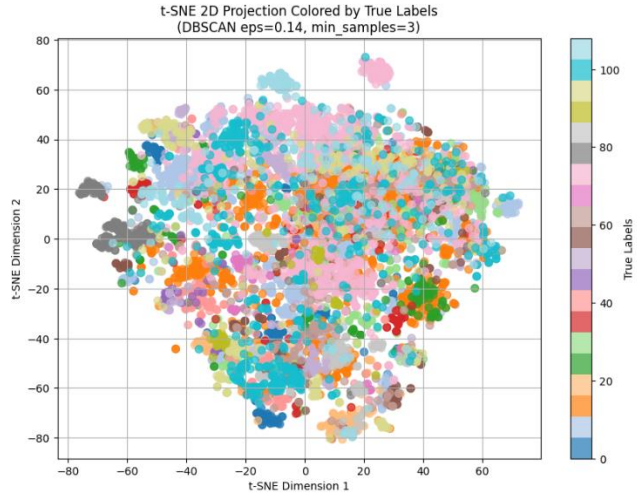

NMI over Time for Each Model
hdbscan_word2vec
hdbscan_word2vec_named
hdbscan_word2vec_splited
NMI
Date
0.5
0.4
0.3
0.2
0.1
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01

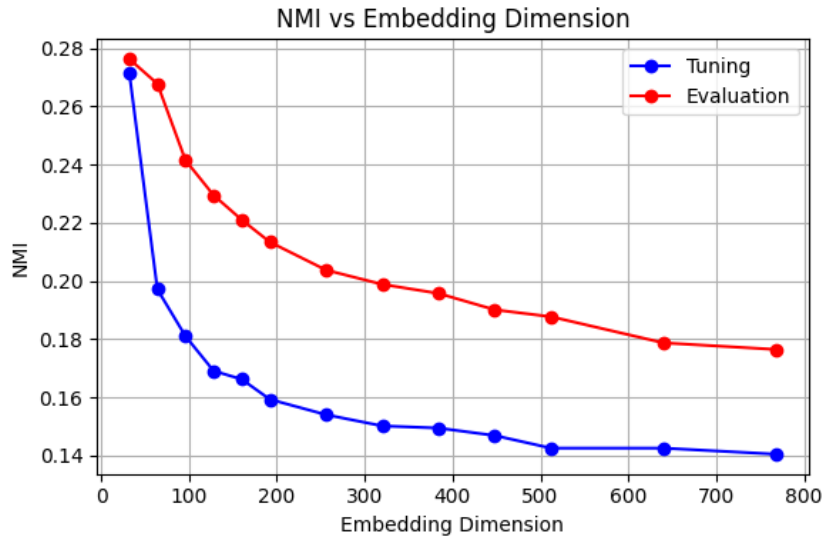

AMI over Time for Each Model
hdbscan_word2vec
hdbscan_word2vec_named
hdbscan_word2vec_splited
AMI
Date
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01

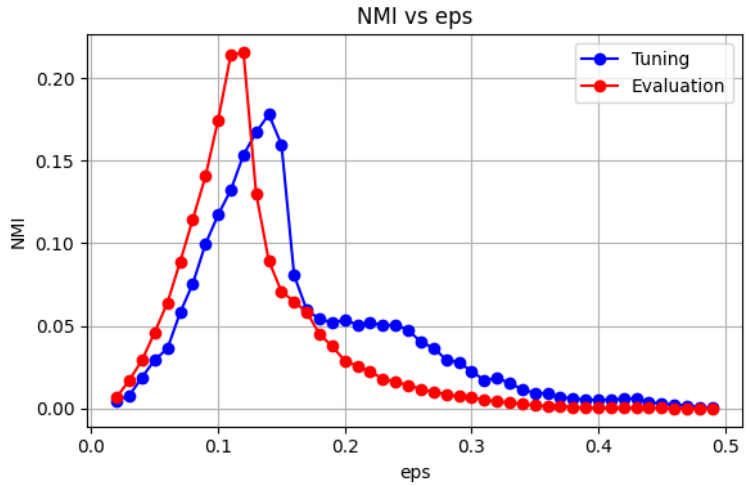

NMI over Time for Each Model
hdbscan_fasttext
hdbscan_fasttext_named
hdbscan_fasttext_splited
NMI
Date
0.2
0.3
0.4
0.5
0.6
2025-05-09
2025-05-13
2025-05-17
2025-05-21
2025-05-25
2025-05-29
2025-06-01


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


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


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


NMI vs Embedding Dimension
Tuning
Evaluation
NMI
Embedding Dimension


NMI vs eps
Tuning
Evaluation
NMI
eps

ملحق أ

الدراسة الكميّة

نبين هنا الأرقام والقياسات التي يجب أن يعمل ضمنها النظام

مقدمة

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

عدد الأحداث اليومي

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

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

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

كمية البيانات الواجب سحبها

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

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

معد دفق التغريدات

في الأيام الاعتيادية، ومع مراقبة كلمات مفتاحية محددة مرتبطة بالخدمات أو الأمن أو النشاطات المجتمعية، قد يصل معدل التدفق إلى 50 – 200 تغريدة في الدقيقة الواحدة، أي ما يعادل تقريبًا 1 إلى 3 تغريدات في الثانية.

وعند 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 الداخليون (المسجلون، المراقبون):

هؤلاء هم من المستخدمين الراغبين بتلقي التنبيهات، المراسلين، الإعلاميين، ومحللي البيانات، المهتمين بالشأن العام.

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

المستخدمون الخارجيّون (الجمهور غير المسجلين):

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

ملحق ب

دليل تنميط مجموعة البيانات

نرفق هنا الدليل الذي تم وضعه لتنميط مجوعة البيانات.

دليل تعليمات تصنيف التغريدات

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

الغرض

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

معايير التصنيف

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

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

أمثلة على تصنيف التغريدات

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

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