University of Sebha
Faculty of Information Technology
Section computer Science

جامعة سبها
كلية تقنية المعلومات
قسم علوم الحاسب

دولة ليبيا

وزارة التعليم العالي والبحث العلمي

بحث مقدم لاستكمال درجة البكالوريوس بكلية تقنية المعلومات

استخدام تقنية تحديد الكيانات علي بيانات مجمعة من موقع تويتر

دراسة حالة (الهجرة غير الشرعية في ليبيا)

Using Named Entity Recognition technique on data collected from Twitter case study (Illegal immigration in Libya)

إعداد الطالبات:

زمزم محمد علي أوحيده هاجر محمد علي بدر

20180625 ***202180657***

تحث إشراف:

د. منصور الصغير

البريد الإلكتروني:

Zamz.nejam@sebhau.edu.ly

Hage.abdalrhman@sebhau.edu .ly

العام الجامعي 2021-2022

إقرار

إقرار الطالب /الطلاب

أنا الطالبة: هاجر محمد علي **الرقم الدراسي:** 20180625

وأنا الطالبة: زمزم محمد علي **الرقم الدراسي:** 202180657

أقر /نقر بأن ما ورد في هذا البحث هو من مجهودي الشخصي ماعدا الفقرات التي تم إسنادها إلى مرجع.

التاريخ:..... **التوقيع:**.....

التاريخ:..... **التوقيع:**.....

إقرار المشرف

اسم المشرف: د. منصور علي الصغير

أقر بأني اطلعت على مادة البحث، وأن هذا البحث جاهز للمناقشة.

التاريخ:..... **التوقيع:**.....

إقرار بالموافقة على التصحيحات وتسليم النسخة النهائية:

بعد التصحيح والاطلاع على مادة هذا البحث، تمت الموافقة عليها، وتسليم النسخة النهائية

اسم الممتحن الأول:..... **التوقيع:**..... **التاريخ:**.....

اسم الممتحن الثاني:..... **التوقيع:**..... **التاريخ:**.....

الآيـة

بسم الله الرحمن الرحيم

﴿وَلَقَدْ آتَيْنَا دَاوُودَ وَسُلَيْمَانَ عِلْمًاۖ وَقَالَا الْحَمْدُ لِلَّهِ الَّذِي فَضَّلَنَا عَلَىٰ كَثِيرٍ مِّنْ عِبَادِهِ الْمُؤْمِنِينَ ﴾

(سورة النمل: الآية15)

الإهداء

نهدي هذا البحث إلى كل **طالب علم** يسعى لكسب المعرفة وتزويد رصيده المعرفي العلمي و الثقافي.

إلى من تشاركني أفراحي و أساتي، إلى من سهرت الليالي تنير دربي، إلى من ساندتني في صلاتها و دعائها، إلى أروع امرأة في حياتي**..... أمي الغالية.**

إلى من علمني أن الدنيا كفاح وسلاحها العلم والمعرفة، إلى من سعى لأجل راحتي ونجاحي، إلى قدوتي و سندي و مسندي و اتكائي إلى من أحمل أسمه في مسيرتي**..... والدي العزيز**.

إلى من كان لهم بالغ الأثر في كثير من العقبات والصعاب**.....أخوتي**.

إلى كل ..... **الأصدقاء** ومن كانوا برفقتي، والى كل من كان عونا لي.

إلى كل**.....** **معلم** ساهم في تلقيني ولو بحرف في حياتي الدراسية.

شكر وعرفان

قال تعالى ﴿ وَمَنْ يَشْكُرْ فَإِنَّمَا يَشْكُرُ لِنَفْسِهِ ﴾ (سورة لقمان، الآية: 12).

قال النبي صلى الله عليه وسلم: «لا يَشْكُرُ اللَّهَ مَنْ لا يَشْكُرُ النَّاسَ» (رواه أحمد وأبو داود والبخاري).

نحمد الله تعالى حمدا كثيرا طيبا مباركا ملئ السموات والأرض ونشكر الله عز وجل الذي بتوفيق منه وبفضل منه تمكنا من إنجاز هذا البحث ونسأل الله أن يجعله من العلم النافع في الدنيا والآخرة، ونصلي ونسلم على خاتم الأنبياء والرسل حبيبنا وشفيعنا محمد صلى الله عليه وسلم عدد ما ذكره الذاكرون وغفل عن ذكره الغافلون.

وبعد نتقدم بجزيل الشكر والعرفان إلى الدكتور المشرف: منصور علي الصغير على كل ما قدمه لنا من معلومات قيمة و لمجهوداته وتعاونه معنا في هذه المسيرة، وتكرمه بنصحنا وتوجيهنا حتى إتمام هذا العمل.

كما نتقدم بالشكر لجامعة سبها عامة وكلية تقنية المعلومات خاصة والأستاذة رئيسة قسم علوم الحاسب، وكل من بذل جهد بالنصح والإرشاد لنا في كل مراحل المسيرة الجامعية من أعضاء هيئة تدريس وموظفين.

الطالبات:

هاجر محمد علي زمزم محمد علي

فهرس الأشكال

فهرس الجداول

فهرس المصطلحات

اختصار	المصطلح باللغة الإنجليزية	المصطلح باللغة العربية
NER	Named Entity Recognition	تحديد الكيانات المسماة
NLP	Natural Language Processing	معالجة اللغة الطبيعية
_	Tokenization	تقسيم النص
POS Tag	Part-of-speech tagging	وضع علامات على جزء من الكلام
API	Application Programming Interface	واجهة برمجة التطبيقات
_	Pearson Correlation	تحليل ارتباط الخصائص
_	Univariate analysis	تحليل إحصائي لخاصية واحدة
_	Bivariate analysis	تحليل إحصائي لخاصيتين
_	Five Number Summary	تحليل الخمس أرقام

الملخصAbstract

وتعتبر وسائل التواصل الاجتماعي من أهم مصادر المعلومات إذ أصبحت وسيلة لمشاركة البيانات في أي وقت بسهولة، من هنا ظهرت الحاجة إلى استغلال هذه المعلومات والاستفادة منها لاستخراج أنماط مفيدة، قمنا في هذه الدراسة تحديد وتصنيف الكيانات التي تتحدث عن موضوع الهجرة غير الشرعية في ليبيا باستخدام تقنية تحديد الكيانات المسماة على البيانات التي تم تجميعها من موقع التواصل الاجتماعي تويتر، حيث قمنا بتجميع التغريدات لثلاث سنوات كاملة 2012، 2013، 2014 ثم قمنا بدمجها في قاعدة بيانات واحدة وفلترتها بناء على كلمة Libya وعليها أجرينا مرحلة استكشافية لمعرفة طبيعة البيانات التي نتعامل معها، بعد ذلك قمنا بتحليل الخصائص ومعالجة البيانات وتنظيفها وإزالة الشوائب الغير متعلقة بموضوع الدراسة، قمنا في هذه الدراسة باستخدام مكتبة spaCy التي تعتمد على معالجة اللغة الطبيعية حيث تقوم هذه المكتبة ببعض المعالجات قبل تطبيق التقنية المستخدمة، منها تقطيع النص ووضع علامات على جزء من الكلام ، تم تطبيق تحديد الكيانات المسماة من خلال النموذج المدرب مسبقا الذي تقوم به المكتبة spaCy وبينت النتائج مجموعة كيانات من بينها أشخاص مثل: القذافي وعمر، ودول مثل: الولايات المتحدة الأمريكية وأفغانستان، ووكالات أنباء ومنظمات مثل: وكالة الأنباء اليبية ليبيا نيوز ومنظمات الأمم المتحدة المتعلقة بشوؤن الهجرة غير الشرعية، ومجموعات دينية وقومية مثل: الجنسيات التي تتعلق ببعض الدول والديانات الاسلامية واليهودية، والعديد من الكيانات التي سيتم التعرف عليها بشيء من التفصيل فيما بعد، للكيانات المكتسبة أهمية لصانعي القرار في ليبيا من حيث معرفة أبرز الشخصيات السياسية وكذلك وسائل الإعلام المهتمة التي تناولت موضوع الهجرة.

الفصل الأول

المقدمة

1.1 المقدمة Introduction:

تعتبر الهجرة غير الشرعية أحد أبرز المشكلات التي تواجه المجتمع الدولي اقتصادياً وقانونياً وسياسياً حيث شهدت ليبيا ارتفاع في عدد المهاجرين غير الشرعيين خلال السنوات الأخيرة نتيجة لسوء الأوضاع الأمنية في ليبيا وانعدام الأمن والاستقرار، معتبراً أنها من أقرب الدول إلى الموانئ الأوروبية. الهجرة غير الشرعية أصبحت موضوعاً ملحاً بشكل متزايد في الوقت الحاضر، حيث أنه يجري جزء كبير من النقاش السياسي والإعلامي على وسائل التواصل الاجتماعي التي أصبحت المنصة المفضلة للتعبير الصريح، تعد ظاهرة الهجرة غير الشرعية مثيرة للجدل بحيث يتم تسليط الضوء عليها والحديث عنها بكثرة غالباً في بعض القنوات الأخبارية وفي مواقع التواصل الاجتماعي التي أصبحت جزءًا أساسيًا من الحياة اليومية، إذ يبلغ عدد مستخدميها أكثر من ثلاثة مليار مستخدم في العالم(نصر،2020)، وهناك اهتمام متزايد بشأن استخدام المعلومات التي توفرها منصات التواصل الاجتماعي لاستخدامها في أدوات البحث العلمي وذلك في ظل ما توفر تلك المنصات من كم هائل من البيانات التي من السهل الوصول إليها، لذلك فإن هذه المنصات بما فيها التويتر وسيلة للإطلاع والوصول إلى أخبار العالم، فهناك العديد من التغريدات على التويتر عن الهجرة غير الشرعية التي تثير الاهتمام حول جمعها وتحليلها، حيث توفّر هذه الكميات الكبيرة للبيانات المتاحة اليوم مصادر غنيّة للمعلومات التي يُمكن تحويلها إلى معارف جديدة، و مفيدة، بالتالي يوجد اهتمام متزايد في استكشاف هذه البيانات من أجل استخلاص المعارف، مما ادى إلى ظهور أدوات تحليل البيانات وهو العلم الذي يُعنى بتحليل البيانات الخام لاستخلاص معرفة أو أنماط مفيدة، ومن أدوات تحليل البيانات: تنقيب البيانات وهي خطوات غير تقليدية لإيجاد أنماط مفيدة ومفهومة وغير مكتشفة سابقا من بيانات خام، يظهر التنقيب عن البيانات معلومات مفيدة مخفية في مجموعة من البيانات النصية.

يعد التعرف على الكيان المسماة Named Entity Recognition (NER) جانباً مهماً من معالجة اللغة الطبيعية(Natural Language Processing (NLP وهو عملية تسعى إلى تحديد الكيانات المسماة في نص غير مهيكل، مثل: أسماء الأشخاص والمؤسسات والمواقع وتعبيرات الأوقات وغيرها، تتعامل معالجة اللغة الطبيعية مع البيانات النصية، وهذه البيانات إذا تم استخدامها بشكل صحيح يمكن أن تحقق العديد من النتائج المثمرة (jing Li,et.al,2020).

بعض تطبيقات معالجة اللغة الطبيعية الأكثر أهمية هي تحليلات النص وأجزاء من تمييز الكلام وتحليل المشاعر وأحد الجوانب المهمة لتحليل هذه البيانات النصية هو تحديد الكيانات المحددة الذي يساعد كثيرا في حالة استخراج المعلومات من مجموعات البيانات النصية الضخمة.
حيث تهدف الدراسة إلى تحديد الكيانات المسماة (NER) في البيانات المجمعة حول موضوع الهجرة غير الشرعية في ليبيا.

2.1 مشكلة البحث Research Problem:

في السنوات الأخيرة أصبحت الهجرة الغير شرعية تشغل الشعوب كافة وأصبحت مشكلة رأي عام في كل العالم، نتيجة لعدم استقرار بعض الدول أمنيا في الشرق الأوسط ومنها ليبيا، يستخدم الكثير من رواد مواقع التواصل الاجتماعي ومنهم رواد موقع تويتر التغريدات للتعبير، مما تناول مسألة الهجرة الغير شرعية على اختلاف انتماءاتهم ودولهم، كذلك يعتبر موقع تويتر منصة إخبارية للعديد من وكالات الأنباء التي تتناول مواضيع الهجرة الغير شرعية على الدوام.

من هذا المنطلق ومع ازدياد وفرة البيانات جائت الحاجة إلى استخراج معلومات مفيدة منها، عليه تقوم هذه الدراسة استخدام تقنية NER في تحديد الكيانات والأشخاص ووكالات الأنباء التي تناولت موضوع
الهجرة الغير شرعية المنطلقة من ليبيا، تتلخص مشكلة البحث في:

- قلة الدراسات التي تبحث عن موضوع الهجرة غير شرعية في ليبيا من بيانات مجمعة من موقع تويتر باستخدام تقنية NER.
- تحديد الدول والمؤسسات والأشخاص والمنظمات ووكالات الأنباء المهتمة بموضوع الهجرة الغير الشرعية في ليبيا باستخدام تقنية NER.

3.1 أسئلة البحث Research Question:

لحل مشكلة البحث نضع التساؤلات التالية:-

1.ماهي أهم الكيانات التي يمكن استخراجها عن الهجرة غير الشرعية من موقع تويتر باستخدام تقنية NER؟

2.ماهي المنظمات ووكالات الانباء المهتمة بموضوع الهجرة الغير شرعية القادمة من ليبيا؟

3.ماهي الدول والاشخاص والمجموعات القومية والدينية الاكثر تكرار في البيانات المجمعة عن الهجرة الغير الشرعية في ليبيا؟

4.1أهداف البحث Research Objective:

من خلال أسئلة البحث تهدف الدراسة إلى تحقيق الأتي:

- الفهم المعمق للبيانات المجمعة من خلال معرفة أنواع البيانات وخصائصها، وإيجاد علاقات مثيرة للإهتمام بين الخصائص.
- المساهمة في الكشف عن الكيانات في البيانات النصية المجمعة من موقع تويتر التي ذكرت الهجرة غير الشرعية المرتبطة بليبيا.
- تحديد وتصنيف الأشخاص والمنظمات والدول ووكالات الأنباء والمجموعات القومية والدينية التي تناولت موضوع الهجرة غير الشرعية المتعلقة بليبيا عن طريق استخدام تقنية التعرف على الكيان المسماة NER من نموذج مدرب مسبقا.

5.1 نطاق البحث Proposed Scope:

هي الحواجز والحدود التي يجب الوقوف عندها وعدم تخطيها، حدود الدراسة هي:

الحدود الموضوعية: لدراسة هو تطبيق تقنية NER على بيانات الهجرة غير الشرعية المجمعة من تطبيق تويتر.

الحدود الزمانية: سيتم تجميع البيانات من سنة 2012 بداية الاعداد الكبيرة للمهاجرين إلى نهاية 2014 أي ثلاث سنوات كاملة.

الحدود المكانية: تجمع البيانات عن الهجرة غير الشرعية في ليبيا من موقع توتير.

6.1 أهمية البحث Significance of the Research:

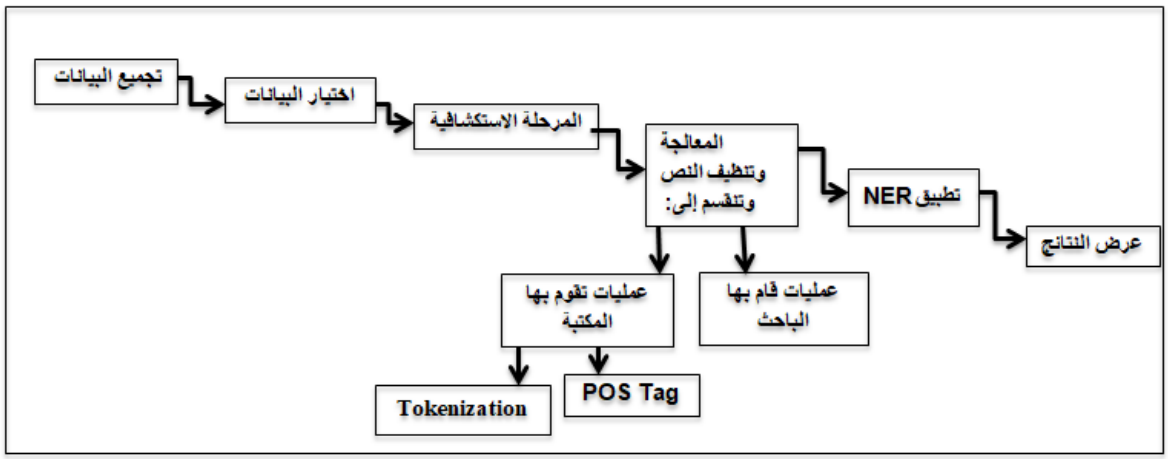
بسبب زيادة ظاهرة الهجرة غير الشرعية في العالم عامة وليبيا خاصة وتعدد الحديث عنها في مواقع التواصل الاجتماعي وبعض القنوات الإخبارية، ظهرت الاهمية إلى تجميع هذه البيانات من منصة التواصل الاجتماعي تويتر واستخدام مهام معالجة اللغة الطبيعية مثل: التعرف الكيانات المسماة NER، حيث تتلخص أهمية الدراسة فيما يلي:

1. تساهم هذه الدراسة في معرفة ارتباط ظاهرة الهجرة غير الشرعية مع كيانات أو جهات داعمة للمهاجرين والتي تفيد الدولة والأشخاص الباحثين والمحللين في تلك المجال.
2. نتائج الدراسة بالإمكان توظيفها في مجال الأمن من خلال معرفة الشخصيات السياسية التي تتحدث عن المهاجرين غير الشرعيين في ليبيا.
3. قد تساهم هذه الدراسة في الحد من ظاهرة الهجرة غير الشرعية عن طريق الفهم العميق لهذه الظاهرة.

7.1 منهجية البحث Research methodology:

يعرف منهج البحث بأنه مجموعة من الخطوات والطرق والتقنيات التي يتم استخدامها بتسلسل منطقي وعلمي منظم وواضح في كتابة البحث العلمي التي توصل الباحث لنتائج معينة أو لحل مشكلة ما.في هذه الدراسة سيتم اتباع منهجية التنقيب عن البيانات القياسية Text Mining الموضحة في الشكل (1) والتي تعد المنهجية الأكثر شيوعاً للتنقيب عن البيانات والتحليلات استخراج معارف وأنماط مفيدة ، ثم دمج الخطوات التي صممت من قبل (Ayan.,et.al.,2014) في منهجية لتحقيق هدف الدراسة.

فيما يلي شكل يوضح خطوات المنهجية المستخدمة في الدراسة:



الشكل رقم (1) يوضح المنهجية المستخدمة في الدراسة

سوف نقوم في هذه الدراسة بتجميع البيانات المتعلقة بموضوع المشكلة، وسيتم اكتشافها ومعالجتها لتصبح جاهزة الاستخراج الكيانات منها تتلخص منهجية Text Mining في الآتي:

1- تجميع البيانات:

في هذه المرحلة يتم بتجميع البيانات من موقع تويتر التي تختص أو تتحدث عن موضوع الهجرة غير الشرعية في ليبيا في مخزن بيانات واحد باستخدام اداة واجهة برمجة التطبيقات Twitter Api تم التجميع التغريدات باللغة الإنجليزية و باستخدام الكلمات الدلالية.

2- اختيار البيانات:

في هذه المرحلة يتم اختيار البيانات التي لها علاقة بالمشكلة وتحديد واسترجاع البيانات الملائمة لعملية التنقيب من مجموعة البيانات.

3- المرحلة الاستكشافية:

المرحلة الاستكشافية تعني فهم قاعدة البيانات وما تحتويه من قيم وحقول واستكشاف كافة المعلومات من قاعدة البيانات التي تم تجميعها، مثل أنواع البيانات والقيم مفقودة وشاذة في مخططات و تمثيلات بيانية.

4- معالجة وتنظيف البيانات ويتكون من قسمين:

عمليات معالجة قام بها الباحث وهي:

قمنا بحذف كل من الروابط والهاشتاق وعلامة @ وعلامات الترقيم وكلمات الشائعة(stop word) ، تحويل الحروف من كبيرة إلى صغيرة لجعل نتائج أكثر دقة

عمليات معالجة تمت بواسطة المكتبة المستخدمة وهي:

- تقطيع النص Tokenization

يقصد بهذه الخطوة تقطيع النص إلى مجموعة كلمات منفصلة، سوف يتم تقطيع كل تغريدة من البيانات المجمعة إلى مجموعة الكلمات.

- وضع علامات على جزء من الكلام POS Tagging

هو نوع من تصنيف الكلمات أو هو عملية ترميز كلمة في النص على أساس تصريفها نحوي مثل :الأسماء والأفعال والصفات والظروف اعتمادا على تعريف الكلمة وسياقها.

5- تطبيق **NER:**

بعد إجراء تحليل أولي للبيانات في المرحلة الاستكشافية وتنظيف البيانات ومعالجتها سيتم في هذه المرحلة تطبيق تقنية NER وتحديد الكيانات.

6- : عرض النتائج:

في هذه الخطوة سيتم عرض النتائج التي ثم وصول إليها وتقييمها إذا كانت النتائج تتوافق مع أهداف البحث.

الفصل الثانـي

الدراسات السابقة

1.2الدراسات السابقة Literature Review:

تقدم بوابات التواصل الاجتماعي منصات عامة قوية بما فيها تويتر، حيث يمكن للناس الأطلاع على كافة أخبار العالم من خلالها، تتعدد وتتنوع المعلومات فيها ومن السهل الوصول إليها، ساعد تويتر في استخراج المعلومات من تغريدات المستخدمين وجمعها وإجراء بعض التقنيات عليها، تقترح هذه الدراسة جمع المعلومات من موقع التواصل الاجتماعي تويتر حول موضوع الهجرة غير الشرعية القادمة من ليبيا باستخدام تقنيات تنقيب البيانات و NER.

في دراسة (مبروكة، 2018)تم دراسة ظاهرة الهجرة غير الشرعية ومعرفة عواملها و أسبابها و أضرارها اإليجابية و السلبية، و تقديم عدة مقترحات وتوصيات للحد من هذه الظاهرة.

تشكل ظاهرة الهجرة غير الشرعية أدى على دولة ليبيا كما في دراسة (محمد إمحمد محمد أبوزيد، 2019)التي تتحدث عن ظاهرة الهجرة غير الشرعية وأثرها على الأمن القومي واعتمدت الدراسة تتبع التسلسل التاريخي لهذه الظاهرة على مستوى ليبيا، وتوصلت لمجموعة من النتائج والتوصيات وأثبتت أن هناك عالقة بين الهجرة غير الشرعية الأثر السلبي على الأمن القومي.

في الدراسة (Sudha,et.al,2012) تم استخدام تقنية NER وهي فرع من الذكاء الاصطناعي والمهمة الفرعية لمعالجة اللغة الطبيعية، الهدف من *NER* هو تصنيف الكلمات إلى بعض الفئات المحددة مثل: اسم شخص، اسم موقع، اسم مؤسسة.

اقترحت دراسة (Alan,et.all,2011) تصنيف الكيانات المسماة NER في التغريدات حيث انها تعتبر مهمة صعبة لسببين، أولا في عدد كبير من أنواع الكيانات المسماة المميزة (شركات، منتجات، فرق، وغيرها) ثانًيا: نظرا لحد أقصى 140 حرف في Twitter غالباً، ما تفتقر التغريدات إلى سياق كاف لتحديد نوع الكيان دون مساعدة الخلفية، تقوم الدراسة بشكل تجريبي بتقييم أدوات معالجة اللغة الطبيعية لتطبيقها على التغريدات في تويتر، تقوم الدراسة بإنشاء ميزة للتعرف على الكيان المسماة تستخدم مجموعة بيانات تتكون من 800 تغريدة تم أخذ عينات منها عشوائيا .

نظرا للاستخدام الواسع لتويتر كمصدر للمعلومات، يعتبر الوصول إلى تغريدة شيقة للمستخدم من بين مجموعة تغريدات أمرا صعبا، هدفت دراسة قام بها (Deniz,et.al,2015) بنمذجة ملف تعريف المستخدم إلى التعرف على الكيانات المسماة (NER) لمستخدمي التويتر حيث يستخدم لإنشاء توصيات تغريدة مخصصة استخراج معلومات من هذا الحجم الكبير من التغريدات التي أنشأها ملايين

مستخدمي التويتر يستخدم الباحثون التعرف على الكيان المسماة (NER) ويعرف بتحديد وتصنيف نوع معين من البيانات، في نوع معين من النص.

تعد منصات التواصل الاجتماعي مثل Twitter مصدرًا جذابًا للبيانات لمراقبة الصحة العامة، لأنها تتجنب العقبات القانونية والتقنية للوصول إلى مصادر المعلومات الصحية الأكثر وضوحًا، قد يحتوي جزء ضئيل فقط من التغريدات على معلومات مفيدة للصحة ولكن في تويتر يقابل ذلك الحجم الهائل للتغريدات المنشورة. ركزت دراسة (Antonio,et.al,2015) علي تحديد الكيانات الطبية المسماة في الوقت الفعلي لمشاركات Twitter وتحديد مواقعها الجغرافية.

يتم إخفاء مجموعة متنوعة من البيانات التفصيلية حول الموضوعات الجيولوجية وعلوم الأرض، يوفر التعرف على الكيانات المسماة (NER) كلاً من الفرص والتحديات للاستفادة من هذه الثروة من البيانات في أدبيات علوم الأرض لتحليل البيانات واستخراج المزيد من المعلومات، تعتمد تقنية NER الحالية بشكل أساسي على النهج القائمة على القواعد والتي تخضع للإشراف اقترحت دراسة (Qinjun,et.al،2019) تصميم إطار عمل تدريجي عام لـ NER الخاص بالمجال، بعد هذا الإطار يتم جمع الكيانات الخاصة بالمجال والكلمات العامة للمجال واختيارها كمصطلحات أولية، النموذج المقترح يتعرف بشكل فعال على الكيانات المحددة الجيولوجية ويحددها.

أصبحت المعلومات النصية متوفرة بكثرة على منصات التواصل الاجتماعي، مما أدى إلى متطلبات التقنيات والأدوات لاستخراج المعلومات المفيدة، إحدى هذه مهام استخراج المعلومات التعرف على الكيانات المسماة NER وهي العثور على أعضاء من فئات مختلفة محددة مسبقًا، مثل: الأشخاص والمنظمات والمواقع والتاريخ/الوقت والكميات والأرقام وما إلى ذلك، تم تطوير عدد من التقنيات من قبل العديد من الباحثين لاستخراج تنوع الكيانات من لغات وأنواع نصوص مختلفة، ومع ذلك هناك اهتمام متزايد بين مجتمع البحث لتطوير المزيد من الأساليب الجديدة لاستخراج كيانات مسماة متنوعة مفيدة في تطبيقات اللغة الطبيعية المختلفة. تقدم الدراسة (Archana,et.al,2018) مسحًا للتطورات والتقدم في بحث التعرف على الكيانات المسماة وتصنيفها.

تعد طرق التعرف على الكيانات استراتيجية فعالة لاستخراج المعلومات تهدف الدراسة (Jadon J Jung, 2011) الى توسيع اساليب النصوص الصغيرة (وهي نص قصير على وسائل التواصل

الاجتماعي) حيث تقترح خاصيتين لاكتشاف المجموعات السياقية للنصوص الصغيرة والتي يمكن توقعها لتحسين النتائج التجريبية لأداء مهام استخراج المعلومات وتحديد الكيانات NER من منصات التواصل الاجتماعي.

نظرا للكم الهائل من البيانات في منصات التواصل الاجتماعي وخاصة تويتر، يعد اكتشاف الكيانات مفيدا في تطبيقات التنقيب عن النصوص، هناك الحاجة الى استخراج الكيانات من محتوى تويتر النصي بينما تعمل NER على تصنيف أسماء الأشخاص والمؤسسات والمواقع، حيث ركزت الدراسة التي قام بها (Diana,et.al,2017) على اكتشاف المواقع في التغريدات على منصة التواصل الاجتماعي تويتر، تقترح الدراسة مجموعة من الاستدلالات القادرة على اختيار الموقع الفعلي الصحيح المكون من مرحلتين لنظام مراقبة وسائل التواصل الاجتماعي تصور الأماكن المذكورة في تغريدات على تويتر.

يمكن أن تحتوي التغريدة على 140 حرفا كحد أقصى ويستخدم المستخدمون الاختصارات للتعبير عن أفكارهم، نتيجة لذلك يصبح من الصعب التوصل إلى طرق لتحديد الكيانات المسماة (أسماء الأشخاص، المنظمات، المواقع)، قدمت الدراسة (Ayan,et.al,2014) نهجا من أربع خطوات للتعرف على الكيانات المسماة من المواقع المصغرة.

يعتبر التنقيب عن البيانات مركز اهتمام العديد من الباحثين والعلماء؛ ذلك لأن الناس يشاركون أفكارهم وآرائهم كنصوص في المدونات الصغيرة على وجه التحديد مثل: Facebook و Twitter، حيث أجرت دراسة (Sulan M Altarrazit,et.al,2016) تنقيب عن آراء المغردون، الأشخاص الذين يغردون على تويتر، سبب اختيار موضوع الهجرة على وجه التحديد بالمقارنة مع الموضوعات الهامة الأخرى في السياسة لأنه ظل مشكلة لم يتم حلها لعقود، تم جمع البيانات الخاصة بهذا البحث بعد مناقشة الانتخابات الرئاسية الأمريكية في الشرعيين"، 15 في جامعة كولورادو في بولدر لمدة 10 أيام تقريبًا. الفئات الرئيسية الثلاث للرأي المحددة هي "الإصلاح أو منح الجنسية للمهاجرين غير الشرعيين"، أو "ترحيل جميع المهاجرين غير الشرعيين" ، أو "ترحيل المهاجرين المجرمين غير الشرعيين فقط". وفي استعراض يدوي للبيانات، وُجد أن غالبية الآراء منحازة للفئة الثانية، وهي "ترحيل جميع المهاجرين غير الشرعيين". يتم التصنيف الثنائي لأول رأيين (مؤيد وضد) والتصنيف متعدد الحدود لجميع الآراء الثلاثة.

أصبحت قضية الهجرة بارزة بشكل متزايد في النقاش السياسي والإعلامي، تنظر هذه الدراسة في بروز تغطية المهاجرين من الاتحاد الأوروبي وقضايا الهجرة الى الاتحاد الاوروبي وتستكشف التغطية الاخبارية من حيث المصطلحات (*Paul,et.al,2020*) وتقوم باستخراج المعلومات من وسائل التواصل الاجتماعي لاكتشاف كيف تم النقاش حول الهجرة من قبل أعضاء مشاركين سياسياً من الجمهور على هذه المنصات.

تستكشف الدراسة الجدل السياسي حول الهجرة خلال الحملات الانتخابية في فرنسا وإيطاليا على مدى السنوات الثالث الماضية (Francesca & Alessandro, 2019) تقوم بإجراء تعدين للنص بهدف تحديد بشكل أكثر تحديدا المعلومات المحيطة بالهجرة وكيف يتم تصوير المهاجرين في مناظرةً تويتر عبر الانترنت.

أصبح التنقيب عن النصوص احد المجالات العصرية التي يتم دمجها في مجالات البحث واسترجاع المعلومات حيث اقترحت دراسة (loum,et.al,2017SaidA.Sal) الي وصف كيفية استخدام معلومات في وسائل التواصل الاجتماعي لإجراء تحليلات النصوص وتقنيات التنقيب عن النصوص لتحديد موضوعات الرئيسية في بيانات.

NER هو أحد مجالات معالجة اللغة الطبيعية يستخدم في التعرف على الكيانات المسماة، في دراسة(2019، Eleni) تم استخدام المكتبة spaCy التي تستخدم نموذج مدرب لاستخراج المعلومات المتعلقة بالاضطرابات الكهربائية في أندونيسية من هذه العملية تبين أن أكبر عدد من المواقع المذكورة في التغريدة المتعلقة بفشل التيار الكهربائي جاء من Sleman Regency، وجاء أقل عدد من idul RegencyGunungk، بينما كان الشهر الذي شهد أكبر قدر من انقطاع التيار الكهربائي هو مارس 2020 وأقل كمية كهرباء في يوليو 2020.

يلخص تقرير الفحص الطبي المستقل (IME) الرأي الطبي للطبيب حول الحالة الصحية للمريض بناءً على خبرة الطبيب، تحتوي تقارير محرر أسلوب الإدخال (IME) على معلومات خاصة وحساسة التي يجب إزالتها أو ترميزها بشكل عشوائي قبل إجراء العمل البحثي، قدمت الدراسة (Yuli Vasiliev, 2020) إجراء التعرف على الكيانات المسماة (NER) لتحديد المعلومات الخاصة وإزالتها أو ترميزها لاحقًا من تقارير IME التي أعدها الطبيب، تقوم بتطبيق مجموعة أدوات NER الخاصة بـ

OpenNLP و spaCy، نظامان أساسيان لمعالجة اللغة الطبيعية وتمت مقارنة الدقة والتذكر وقياس الأداء عند تحديد خمس فئات من معلومات تحديد الهوية الشخصية عبر تجارب تقارير محرر أسلوب الإدخال المختارة عشوائيًا باستخدام المعلمات الافتراضية الشائعة لكل نموذج، لقد وجد أن كلا النظامين يحققان أداءً عاليًا (قياس f> 0.9) عند إلغاء تحديد الهوية وأن نموذج spaCy الذي تم تدريبه بتقسيم 70–30 قطار اختبار هو الأكثر أداءً.

تقترح هذه الدراسة (Xavier & Sylvain, 2019) نهجًا للتعلم الآلي لوضع علامات على جزء من الكلام والتعرف على الكيانات المسماة باللغة اليونانية، مع التركيز على استخراج الميزات المورفولوجية وتصنيف الرموز المميزة في مجموعة صغيرة من الفئات للكيانات المسماة تم تقديم نموذج العمارة الذي تم استخدامه وتمت إضافة النسخة اليونانية لمنصة spaCy إلى الكود المصدري، تم استخدامها لبناء النماذج بالإضافة إلى ذلك تم تدريب جزء من أداة تمييز الكلام يمكنه اكتشاف مورفولوجيا الرموز المميزة وأداء أعلى من أحدث النتائج عند تصنيف جزء الكلام فقط للتعرف على الكيانات المسماة باستخدام spaCy، تم إنشاء نموذج يمتد لنوع AMEXEN القياسي (مؤسسة، موقع، شخص).

2.2الدراسات ذات صلة **Related Work**:

NO	الدراسات السابقة	اوجه التشابه
.1	تقوم الدراسة بتحديد الكيانات الطبية المسماة NER في الوقت الفعلي لمشاركات Twitter وتحديد مواقعها الجغرافية. (Antonio,et.al, 2015)	تتوافق الدراسة مع هذه الدراسة من حيث تحديد الكيانات المسماة NER في منصة التواصل الاجتماعي تويتر.
2	تعمل الدراسة على اكتشاف المواقع في التغريدات على منصة التواصل الاجتماعي تويتر، تكون الدراسة مجموعة من الاستدلالات القادرة على اختيار الموقع الفعلي الصحيح المكون من مرحلتين لنظام مراقبة وسائل التواصل الاجتماعي بتصور الأماكن المذكورة في تغريدات على تويتر. (Diana,et.al,2017)	تتبع الدراسة نهج يتماشى مع هذه الدراسة حيث يعمل على اكتشاف الدول والأشخاص والمؤسسات والمنظمات ووكالات الأنباء التي تناولت الحديث عن الهجرة غير الشرعية عن ليبيا، في تغريدات على تويتر باستخدام NER.

.3	تشكل ظاهرة الهجرة غير الشرعية خطرا على دولة ليبيا كما في دراسة(محمد إمحمد محمد أبوزيد، 2019 ،)التي تتحدث عن ظاهرة الهجرة غير الشرعية وأثرها على الأمن القومي واعتمدت الدارسة تتبع التسلسل التاريخي لهذه الظاهرة على مستوى ليبيا، وتوصلت لمجموعة من النتائج والتوصيات وأثبتت أن هناك عالقة بين الهجرة غير الشرعية والأثر السلبي على الأمن القومي.	تناقش الدراسة موضوع الهجرة غير الشرعية في ليبيا وأثرها على الأمن القومي كما في هذه الدراسة تستخدم موضوع الهجرة غير الشرعية في ليبيا.
.4	ner هو أحد مجالات معالجة اللغة الطبيعية يستخدم في التعرف على الكيانات المسماة، في دراسة (2019، Eleni & Eleftherios) تم استخدام المكتبة spaCy التي تستخدم نموذج مدرب لاستخراج المعلومات المتعلقة بالاضطرابات الكهربائية في أندونيسية من هذه العملية تبين أن أكبر عدد من المواقع المذكورة في التغريدة المتعلقة بفشل التيار الكهربائي جاء من Sleman Regency، وجاء أقل عدد من Gunungkidul Regency، بينما كان الشهر الذي شهد أكبر قدر من انقطاع التيار الكهربائي هو مارس 2020 وأقل كمية كهرباء في يوليو 2020.	تستخدم الدراسة مكتبة spaCy التي تستخدم نموذج جاهز مدرب ل تحديد وتصنيف الكيانات المسماة على التغريدات التي تم تجميعها من موقع تويتر عن الهجرة غير شرعية في ليبيا.

جدول (1) يوضح الدراسات ذات الصلة

3.2المناقشة **Discussion:**

بعد استعراض عدد كبير من الدراسات السابقة وأهمها (Antonio,et.al, 2015) ، (محمد إمحمد محمد أبوزيد، 2019) ،(Diana,et.al,2017)، (2019، Eleni) تبين عدم وجود أي دراسة تطرقت لقضية الهجرة الغير شرعية في ليبيا باستخدام تقنيات Data Mining التي تعتبر من أهم المشاكل الاجتماعية في العالم وعدم استخدام التقنيات الحديثة للتنقيب، لاستخراج هذه معلومات من موقع توتير، وقلة استخدام تقنية NER في مجال الهجرة غير الشرعية على بيانات مجمعة من مواقع التواصل الاجتماعي لذلك تقوم هذه الدراسة بتجميع بيانات نصية خام من موقع التواصل الاجتماعي تويتر، وتطبيق تقنية NER لاستخراج معلومات عن اسماء الكيانات التي تتضمن المؤسسات والدول والأشخاص ووكالات الأنباء وغيرها.

الفصل الثالث

التجميع البيانات ومرحلة الاستكشافية

1.3المقدمة:

في السنوات الأخيرة أصبح (Twitter) يلعب دور كبير في السوشيال ميديا ويعتبر منجم ذهب للبيانات شائعا جداً بحيث يقوم المستخدمون بكتابة رسائل قصيرة تسمى تغريدات حول المواضيع المختلفة، فإن تغريدات كل مستخدم تقريبًا عامة تمامًا ويمكن سحبها، إضافة إلى ذلك إذا كنت تحاول الحصول على كمية كبيرة من البيانات لتشغيل التحليلات عليها، تتيح لك واجهة برمجة تطبيقات Twitter إجراء استعلامات متعددة مثل: سحب كل تغريدة حول موضوع معين خلال العشرين دقيقة الماضية، أو سحب تغريدات مستخدم معين غير مُعاد تغريدها، وبالتالي فقد اهتم الكثير من الأشخاص حول كيفية جمع البيانات من تطبيق تويتر وإجراء البحوث العلمية عليها، في هذا الفصل سنقوم بالتطرق لمرحلة تجميع البيانات والأدوات والتقنيات المستخدمة في تجميع البيانات، إضافة إلى ذلك سنقوم بعرض قاعدة البيانات التي تم جمعها والتي هي عبارة عن التغريدات المجمعة عن الهجرة غير الشرعية في ليبيا ثلاث سنوات (2012، 2013، 2014)، وأنواع السمات الأسمية والرقمية في قاعدة البيانات، كما سنقوم بشرح المرحلة الاستكشافية بشيء من التفصيل والتي تعتبر تحليل مبدئي للبيانات.

2.3تجميع البيانات

في هذه الدراسة قمنا بتجميع البيانات من موقع تويتر التي تختص أو تتحدث عن موضوع الهجرة غير الشرعية في ليبيا، باستخدام Twitter Api تم التجميع بناء على التغريدات المغردة باللغة الإنجليزية وذلك باستخدام الكلمات الدلالية التالية(Libya, Illegal Immigration, Refugee) وكانت البيانات المجمعة لثلاث سنوات من سنة 2012 إلى 2014 كل سنة تمثل قاعدة بيانات بحيث يتم دمج ثلاث قواعد بيانات معا لكي تمثل السنوات المجمعة في قاعدة واحدة، تم اختيار هذه السنوات خصيصاً لأن شاهدت ليبيا تغيير للنظام في تلك الفترة وبالتالي تعتبر غير مستقرة أمنياً ومن المتوقع ازدياد ظاهرة الهجرة.

3.3الأدوات والتقنيات المستخدمة لتجميع البيانات:

استخدام واجهة برمجة التطبيقات (API)

لتجميع البيانات حيث سيتم جمع وتنزيل التغريدات وفقا لشروط محدده مثل: استرجاع تغريدات بكلمات رئيسية محددة أو تاريخ محدد أو بلغة معينة، بمعنى اذا تطابقت الشروط المحددة مع تغريدة سيتم تنزيلها.

4.3فلترة البيانات من التغريدات التي خارج النطاق

بعد أن تم جمع البيانات عن موضوع الهجرة غير الشرعية بإستخدام الكلمات الدلالية التي تم ذكرها فيما سبق، سيتم فلترة البيانات بناء على الهدف بحيث يتم تسليط الضوء على كلمة ليبيا، أي أنه سيتم الاحتفاظ بالتغريدات التي لها علاقة بليبيا أو تم ذكر فيها كلمة ليبيا وحذف جميع التغريدات التي لا تتعلق بليبيا منها التغريدات عن المهاجرين السوريين وغيرها ليكون التحليل منطقي وواضح لجعل النتائج أكثر دقة.

5.3وصف لقاعدة البيانات التي قمنا بجمعها:

في الجدول (2) قاعدة البيانات التي قمنا بجمعها حيث تحتوي على 10 خصائص أو سمات منها: التغريدة، التعليقات على التغريدة، إعادة التغريدة، رابط التغريدة، الإيموجي الموجودة داخل التغريدة، اسم المستخدم، الاسم المستعار، الاعجابات على التغريدة، رابط الصورة مع التغريدة، تاريخ ووقت نشر التغريدة، والتي تحتوي على 4742 سجل بعد دمج قواعد البيانات للسنوات الثلاثة التي قمنا بجمعها في قاعدة بيانات واحدة، وفلترة التغريدات التي لم يذكر فيها كلمة ليبيا وإزالتها.

UserScreenName	UserName	Timestamp	Text	Emojis	Comments	Likes	Retweets	Image link	Tweet URL
Nigerian News.Net	nigerian@ewsnet	-01-2012 T16:35:01 Z21.000	Niger, Chad receive 75,000 refugees from Libya http://tf.to/VN6q	NaN	NaN	NaN	NaN	[]	https twi//: tter.com/nigerianewsnet/status/153514660715442177
Zuhair زهيرHussain	zuhair4@7	-01-2012 T10:56:01 Z02.000	Iraqi refugees # -on the Libyan Tunisian border appeal to animal welfare organization,to assist refugees in nhttp://q\Libya# anon302.net/news/news.php?action=view&id=1 cibarA …1171	NaN	NaN	NaN	NaN	[	https twi//: tter.com/zuhairst/47atus/153429265633910784
Zehra Zaidi	Zehra_Z@aidi	-01-2012 T02:50:01 Z48.000	Fact: 2000+ # Gaddafi regime # figures have applied for refugee status in Tunisia. #Libya	NaN	NaN	1	NaN	[]	https twi//: tter.com/Zehra_Zaidi/status/153307154424668160

جدول (2) وصف قاعدة البيانات المجمعة

في قاعدة البيانات المجمعة يوجد بعض السمات الاسمية وبعضها رقمية وتوجد أنواع من السمات في حاجة إلى تغيير نوعها لتتوافق مع تحليل البيانات الذي سيجرى في المرحلة الاستكشافية.

1.5.3أنواع السمات في قاعدة البيانات:

في البداية يجب ان نقوم بإلقاء نظرة أولية على بيانات وأنواعها حيث تتكون قاعدة البيانات التي قمنا بجمعها من 10 سمات، تنقسم الي جزئين: السمات الاسمية، السمات الرقمية وسيتم توضيحها في جدول (3):

السمات الرقمية	أنواع السمات	معني السمة
Comments	float64	عدد التعليقات على التغريدة
Retweets	float64	عدد إعادة التغريدة

جدول رقم (3) يوضح السمات الرقمية

السمات الاسمية	أنواع السمات	معني السمة
Text	object	التغريدة
Tweet URL	object	رابط التغريدة
User ScreenName	object	الاسم المستعار (القابل للتغيير)
UserName	object	اسم المستخدم (الغير قابل للتغيير)
Emojis	object	الإيموجي الموجودة داخل التغريدة
image link	object	رابط الصورة مع التغريدة
Timestamp	object	تاريخ ووقت نشر التغريدة
Likes	object	عدد الاعجابات على التغريدة

جدول رقم (4) السمات الاسمية

هذه سمات الي تحصلنا عليها حول الهجرة غير الشرعية في ليبيا من مرحلة تجميع البيانات سنستخدم أغلب هذه السمات للحصول على مختلف أنواع المعلومات التي نحتاجها وسنقوم بإضافة سمات أخرى للضرورة في المراحل القادمة لتغطية اكثر قدر ممكن من المعلومات.

من خلال الجداول السابقة يتضح أنه يجب التعديل في أنواع البيانات، لانه يوجد خلل ناتج من دمج أكثر من قاعدة بيانات التي كانت مقسمة حسب السنوات، نظرا لان قمنا بدمج 3 قواعد معا، مما جعل النتائج غير دقيقة، لذلك سيتم تعديل انواع بعض الحقول في المراحل القادمة مثل Likes.

6.3المرحلة الاستكشافية:

تحليل البيانات الاستكشافي (Exploratory Data Analysis)، وهي من أهم مراحل البحث تساعد في فهم البيانات بشكل أفضل وتحليلها، وغالباً ما يُشار إليها باسم الفرضية الأساسية حيث يتم التحقيق فيها للإجابة على أسئلة مثل ماذا ولماذا وكيف، قد يكون الباحث على استعداد لتغيير اتجاهه وفقاً للنتائج التي يصل إليها وعادة ما يتم إجراء EDA في مرحلة أولية للبحث، الغرض الرئيسي منه هو المساعدة في الاطلاع على البيانات قبل البدء في عملية معالجة البيانات، يمكن أن يساعد في تحديد الأخطاء الواضحة ، وكذلك فهم الأنماط داخل البيانات بشكل أفضل، واكتشاف القيم المتطرفة أو الشاذة، وإيجاد علاقات مثيرة للاهتمام بين الخصائص، حيث سنقوم بدراسة قاعدة البيانات التي قمنا بجمعها لمعرفة أنواع البيانات وخصائصها، من خلال مخططات بيانية قمنا بتحليلها وعرضها لتحقيق الأهداف المطلوبة.

7.3تحليل الخمسة أرقام (Five Number Summary) بالنسبة لخصائص الرقمية

من خلال هذا التحليل الاحصائي سيتم استعراض المجموع والمتوسط الحسابي و الخ، الجدول (5) يوضح التحليل الإحصائي للخصائص الرقمية.

	المجموع	المتوسط الحسابي	الإنحراف المعياري	أصغر قيمة في سجلات الخاصية	الربع الأول	الربع الثاني	الربع الثالث	أكبر قيمة في سجلات الخاصية
Comments	318.0	3.29	24.73	1.0	1.0	1.0	2.0	420.0
Retweets	368.0	2.54	5.24	1.0	1.0	1.0	2.0	66.0

جدول (5) يوضح تحليل بالنسبة لخصائص الرقمية

من خلال الجدول السابق نلاحظ مايلي:

1- ان قاعدة البيانات تحتوي على خاصيتين رقميتين.

2- مجموع عينات أقل من عدد سجلات قاعدة البيانات وهذا يدل على وجود قيم مفقودة في خاصيتين.

3- تقارب المتوسط الحسابي بين خاصتي Retweets, comments.

4- الانحراف المعياري لخاصية comments أكبر من خاصية Retweets.

5- الأرباع الثلاث تحتوي على نفس القيمة، وهذا يدل قلة تفاعل على تغريدات.

6- أكبر قيمة الخاصية Comments 420 و66 لخاصية Retweets.

8.3تحليل بالنسبة لخصائص النصية

يوضح جدول (6) التحليل للخصائص النصية الفريدة والمكرره والأكثر تكرارا في البيانات.

الخاصية	المجموع	الفريدة الغير مكررة	الأكثر تكرارا	تكرار
UserScreenName	4354	1858	DTN Libya	925
UserName	4355	1862	DTNLibya@	925
Timestamp	4356	4319	T12:32:15.000Z04-09-2014	5
Text	4356	4356	UNHCR Provides Tents for Hundreds of Flood Victims: agenc… The UN refugee-UNHCR]Tripoli, Libya] http://bit.ly/13b8t4g #africa #libya	1
Emojis	136	54	⚠🔞	34
Likes	851.0	64.0	1.0	196.0
Image link	4356	85	[]	4258
Tweet URL	4356	4356	https://twitter.com/DTNLibya/status/322722287939891200	1

الجدول (6) يوضح تحليل بالنسبة للخصائص الاسمية

من خلال الجدول السابق نلاحظ ما يلي :

1- وجود ثمانية خصائص غير رقمية.

2- توضيح مجموع السجلات وعدد السجلات الغير مكررة و المكرره و الأعلى ظهور في خصائص الغير رقمية.

3- مجلة DTN Libya @ اكثر من قامت بالتغريد في قاعدة البيانات التي ثم جمعها وتمثل تقريبا ربع البيانات.

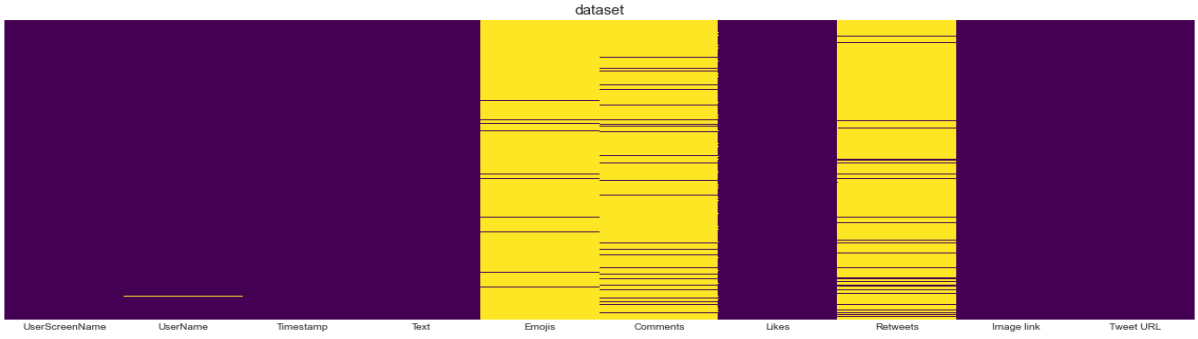
4- أكثر Emojis مستخدم ⚠️🔞 نلاحظ قلة في استخدامه.

5- بلغ عدد الحسابات التي غردت عن موضوع الهجرة 1862.

6- 97٪ من البيانات تغريدات من دون صور.

9.3معرفة القيم المفقودة في قاعدة البيانات:

في هذه الخطوة سيتم استخراج القيم المفقودة في البيانات، لكي تصبح البيانات سليمة وخالية من القيم المفقودة يجب أن نقوم بتحليلها، عند تحليل البيانات دون معرفة القيم المفقودة واستخراجها يكون تمثيل البيانات غير منطقي وواضح، والشكل (2) يوضح القيم المفقودة وتوزيعها في البيانات:



يوضح شكل (2) القيم المفقودة في قاعدة البيانات مجمعة

يوضح الشكل السابق وجود ثلاث أنواع من السمات بها قيم مفقودة تم وصفها في جدول (2) الذي يوضح النسب المئوية للسمات حيث تحتوي الرموز التعبيرية على قيم مفقودة بنسبة 96٪ والتعليقات بنسبة 92٪ وتحتوي إعادة التغريدة على قيم مفقودة بنسبة 91٪ وباقي الحقول لا توجد بها قيم مفقودة، وعليه سنقوم بتبديل القيم المفقودة بصفر لاستخراج بيانات خالية من القيم المفقودة.

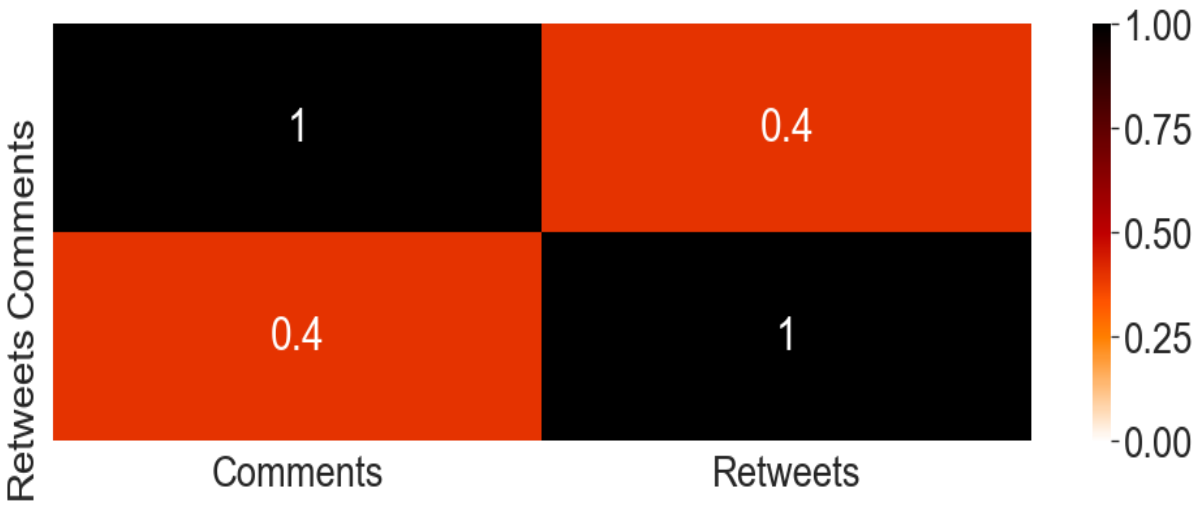
السمات	عدد القيم المفقودة	النسبة المئوية للقيم المفقودة
User ScreenName	4	0 %
UserName	1	0 %
Timestamp	0	%0
Text	0	%0

Emojis	5521	97 %
Likes	0	0 %
Retweets	5224	91 %
Comments	5223	91 %
Image link	0	%0
Tweet URL	0	%0

الجدول رقم (7) يوضح معلومات عن القيم المفقودة

10.3تحليل ارتباط الخصائص عن طريق Pearson Correlation

يوضح تحليل ارتباط السمات أو الخصائص ببعضها، العلاقة بين السمات في حال كانت العلاقة قريبة من 1 تدل على أن العلاقة قوية بين السمات، كما في الشكل (3):



شكل (3) يوضح ارتباط الخصائص قبل اضافة الاعمدة جديد لقاعدة البيانات

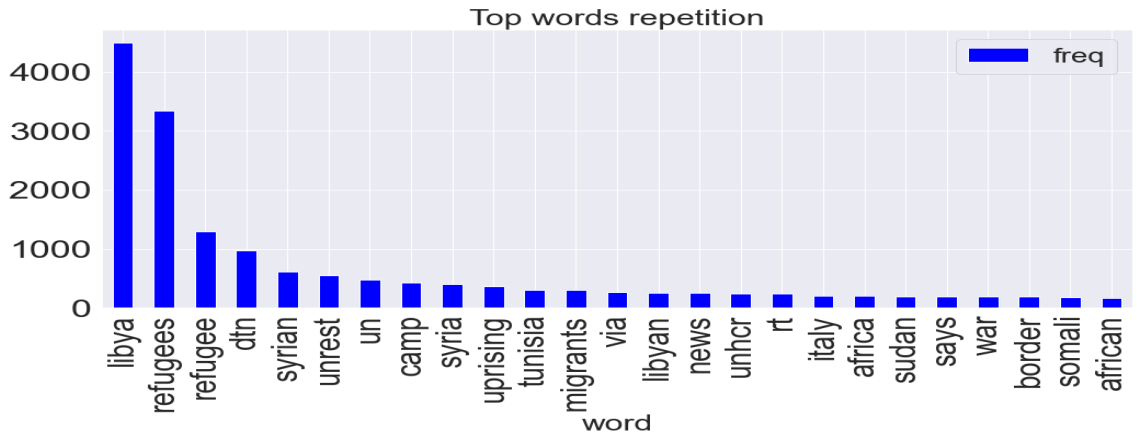
الشكل السابق يوضح تحليل ارتباط السمات، كلما كان رقم اقرب من 1 كلما كانت العلاقة بين الخصائص أقوى، على سبيل المثال في الشكل العلاقة بين Retweet و Comments بنسبة 40% 0.4، أي عندما تكون هناك 10 تعليقات على تغريدة ما توجد 4 إعادة تغريد عليها.

11.3تحليل الاحصائي لخاصية واحدة (Univariate analysis):

التحليل أحادي المتغير هو أبسط شكل من أشكال تحليل البيانات حيث تحتوي البيانات التي يتم تحليلها على متغير واحد فقط، نظرًا لأنه متغير واحد فإنه لا يتعامل مع الأسباب أو العلاقات، الغرض الرئيسي من التحليل أحادي المتغير هو وصف البيانات والعثور على الأنماط الموجودة داخلها.

1.11.3تحليل الكلمات الأكثر تكراراً في نص التغريدة(قبل التوحيد):

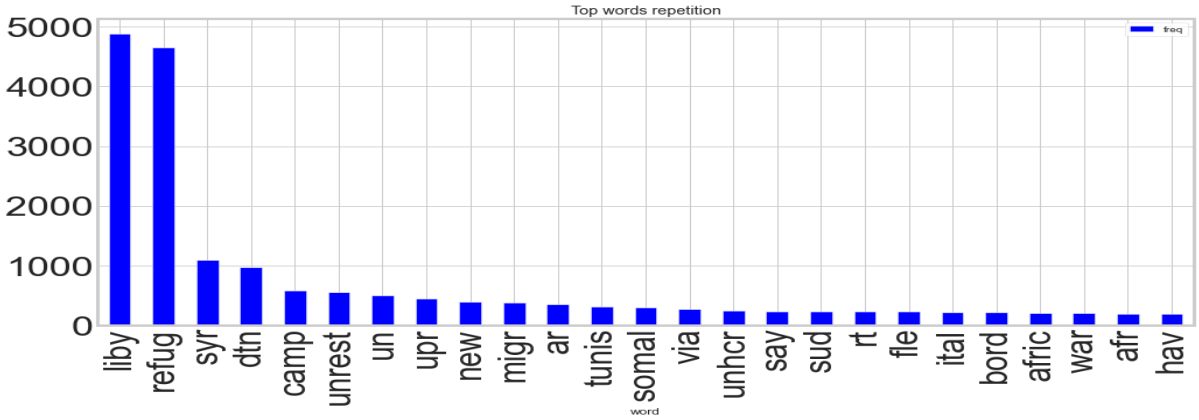
يوضح الشكل (4) أكثر الكلمات تكراراً في نص التغريدة حيث قمنا بإجراء معالجة للنص ومن ثم حذف علامات الترقيم، والروابط، واستبعاد الكلمات الشائعة (stop words) وهي الكلمات التي تتكرر في نصوص مثل (the , in , to , or …..) التي يستحسن تجاهلها لتحسين النتائج.



شكل (4) يوضح اكثر الكلمات تكرارا في نص تغريدة قبل توحيد النص

يبين الشكل (4) أكثر الكلمات تكراراً قبل توحيد النص حيث سيتم التوحيد فيما بعد لمنع التكرار في النتائج، من خلال الشكل نلاحظ تم ذكر كلمة libya في أغلبية البيانات المجمعة وبلغ عدد تكرارها 4000 تقريباً، تليها كلمة refugees بلغ عدد تكرارها 3000 فما فوق، من كلمة syrian إلى كلمة somali كان تكرارها شبه متقارب، أقل كلمة تم ذكرها african.

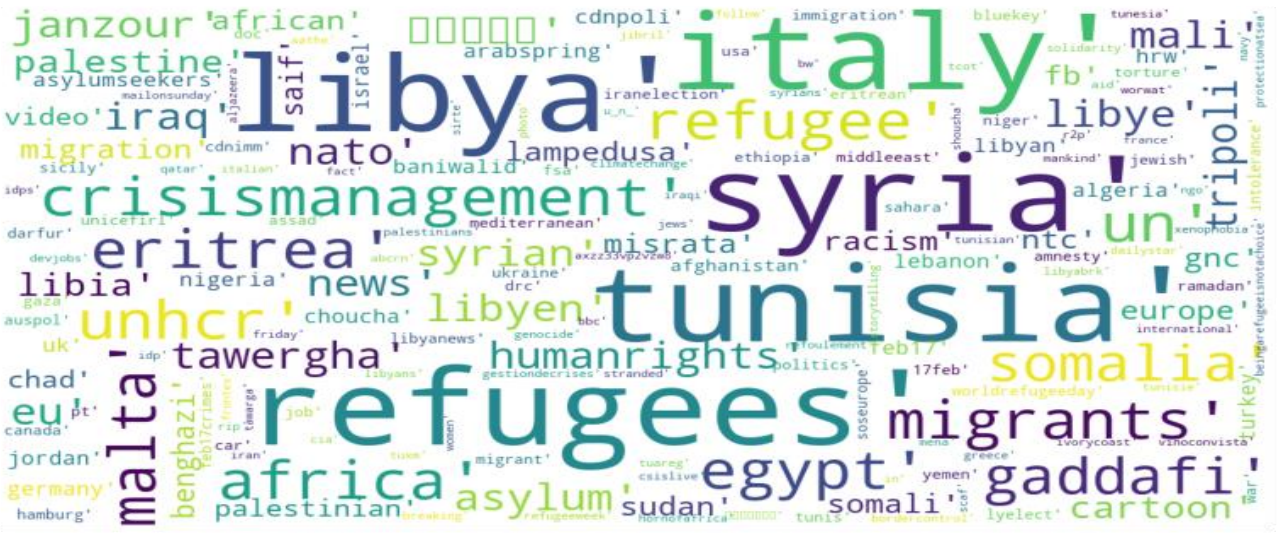
2.11.3تحليل الكلمات الأكثر تكراراً في نص التغريدة(بعد توحيد):



شكل (5) يبين اكثر الكلمات تكرارا في نص تغريدة بعد توحيد النص

يبين الشكل(5) أكثر الكلمات تكراراً في التغريدات، تم توحيد النص عن طريق stemmer، نلاحظ من خلال الرسم تم ذكر كلمة Libya بشكل كبير في التغريدات حوالي 4000 فما فوق، تليها كلمة refuge وهكذا حتى نصل إلى أخر كلمة تم ذكرها بشكل قليل وهي hav، تم توحيد النص لمنع التكرار.

3.11.3تحليل الهاشتاقات الأكثر تكرارا في نص التغريدة:



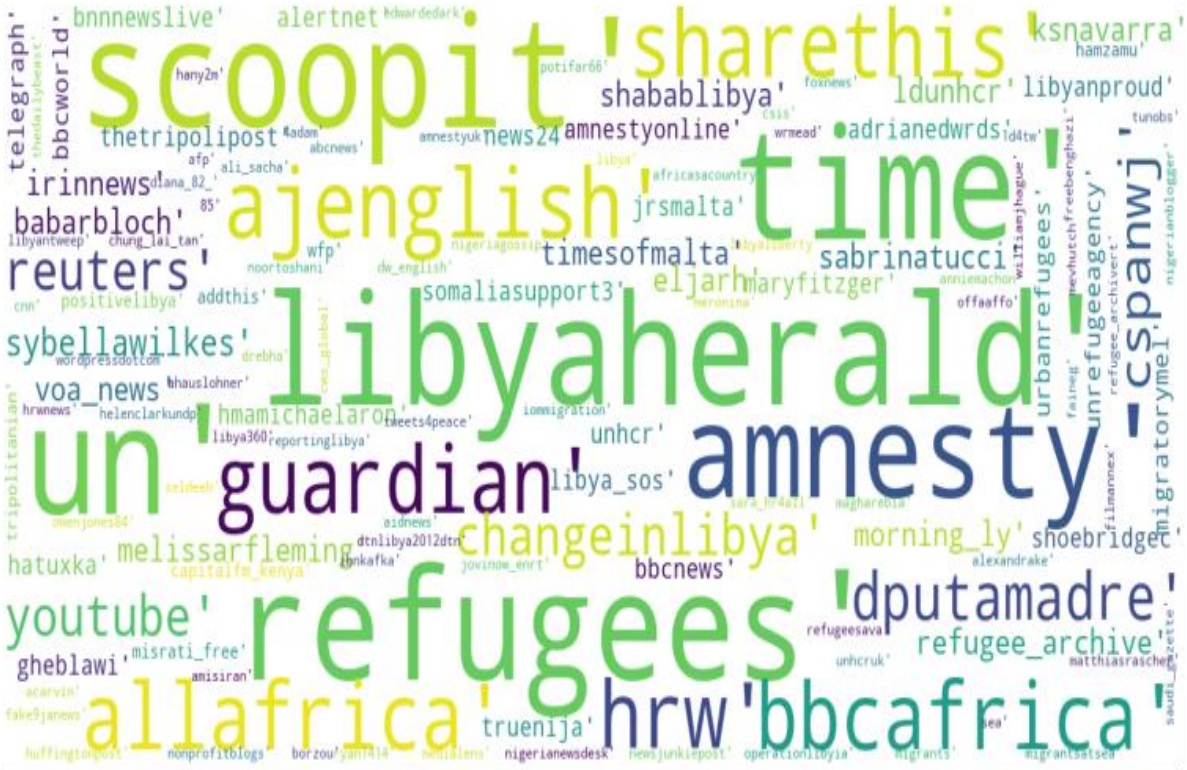
شكل (6) يظهر أكثر الهاشتاقات تكرارا في نص تغريدة

من خلال الشكل السابق الذي يبين أكثر الهاشتاقات تداولاً بين التغريدات حيث كان أكثر هاشتاق تكرارً هو #Libya الذي بلغ معدل تكراره إلى 1490، تلاه #refugees بلغ معدل تكراره 388،

#syria بلغ 208، #tunisia بلغ 106، #italy بلغ 72، حيث يدل ذلك على أن أغلب التغريدات تتحدث عن الهجرة في ليبيا او تم ذكر ليبيا بالاشارة إليها، ويوجد العديد من الهاشتاقات التي تم ذكرها اقلها معدل تكرار 1، بناء على اختلاف حجم الهاشتاقات كما هو موضح في الشكل (6) تم ذكر libya، tunisia، italy، syria، refugees بشكل كبير، ربما هناك علاقة تربط بين ليبيا وتونس وسوريا، على الاغلب هي الدول التي يأتي منها المهاجرين وهناك دول أخرى تم ذكرها منها إيطاليا ومالطا فهي الدول التي يلجأ إليها المهاجرين.

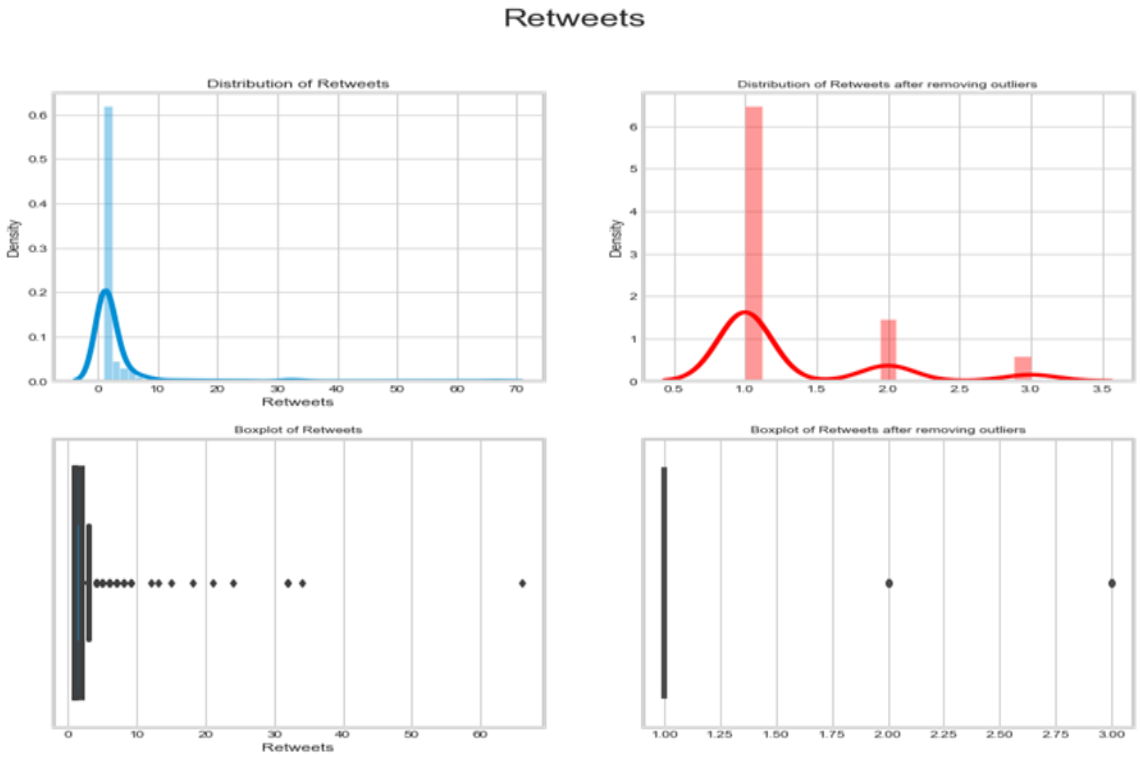
4.11.3تحليل أسماء الحسابات في نص التغريدة

الشكل التالي يبين أكثر الحسابات التي قد تكون (أشخاص، منظمات، مؤسسات، وكالات أنباء، قنوات إخبارية، وغيرها) تمت الإشارة إليهم في التغريدات حول الهجرة في البيانات التي تم جمعها، أكثر حساب تمت الإشارة إليه @refugees والذي تحصل على 254 إشارة @sharethis اقلها حساب تحصل على 14 إشارة.



شكل (7) يوضح اكثر حسابات تكرارا في نص تغريدة

5.11.3تحليل عمود إعادة تغريد (Retweets)



شكل (9) يبين تحليل خاصية إعادة تغريدة

بناء على نظرية Tukeys التي تنطبق في وقت واحد على مجموعة جميع المقارنات الزوجية، يوضح الشكل أكثر القيم التي بها إعادة تغريد مع القيم الشاذة (outlier) وبعد حذف القيم الشاذة، في الرسمة الأولى أعلى اليسار تبدأ القيم من 0 وتنتهي بـ 100 أكثر إعادة تغريد محصور بين الصفر والعشرة مع الـ outlier، في الشكل الثاني بإزالة الـ outlier التوزيع طبيعي حيث تقع القيم بين الـ 0 و 3.5 (الرسم للتوضيح لا يمكن أن يكون لدينا إعادة تغريدة ونص فقط أعداد صحيحة)، في الرسمة الثالثة بعض القيم منحازة إتجاه اليسار وبعضها outlier نلاحظ من الشكل وجود قيمة شاذة عند 100، الشكل الرابع يوضح boxplot أفقي بعد إزالة outlier.

12.3هندسة الخصائص

هي الحاجة لتغيير نوع خاصية معينة أو استخراج خاصية جديدة من خاصية موجودة مسبقاً، يتطلب إجراء هندسة الخصائص تغيير انواع بعض الخصائص وإضافة خاصية والعمل عليها.

1.12.3 تغيير أنواع بعض السمات:

الخاصية	قبل التغيير	بعد التغير
Likes	object	int 32
Timestamp	object	datetime64

شكل (8) يوضح الخاصية قبل تغير نوع وبعد تغيير

كما في الجدول (8) تم تغيير خاصيتين لأنها لا تتوافق مع الإجراءات الإحصائية التي ستجرى لاحقاً، Likes من المفترض أن يكون عدد صحيح و Timestamp من المفترض أن يكون تاريخ.

2.12.3إضافة خصائص لقاعدة البيانات

بعد تغيير نوع الخاصية Timestamp من object إلى datetime64 لتتوافق مع أهداف البحث قمنا بتحويل الوقت من التوقيت العالمي GMT إلى توقيت ليبيا (المحلي)، و استخرجنا من هذه الخاصية خاصية الأيام والشهور والساعات والسنوات نظراً لأهميتها في البحث، كما موضح في الجدول (9)

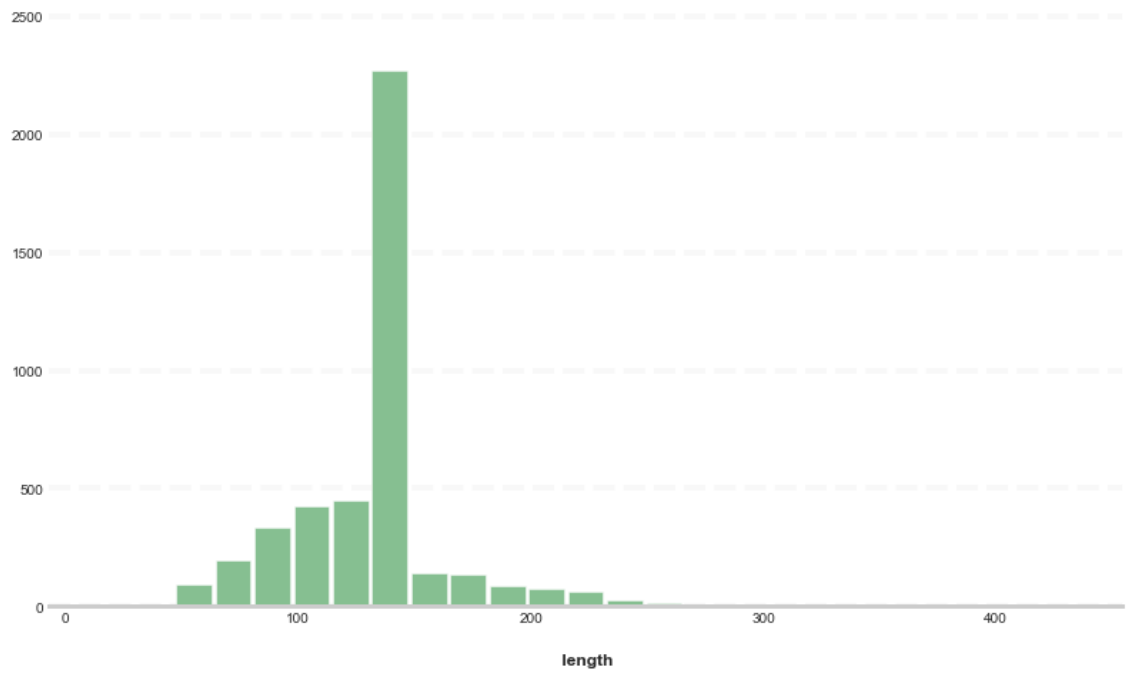
length	years	hours	months	days

يوضح جدول (9) اضافة الاعمدة الزمنية لقاعدة البيانات

كما تم إضافة الخاصية length لمعرفة طول التغريدات، علماً بأن نوع الخصائص الجديدة هو int32.

3.12.3متوسط أطوال التغريدات

قمنا بإضافة خاصية جديدة لقواعد البيانات تمثل طول حروف نص كل تغريدة Length، ثم تحليلها لمعرفة متوسط أطوال التغريدات التي ثم حصول عليها من مرحلة التجميع، الشكل التالي يوضح توزيع أطوال التغريدات في قاعدة البيانات:

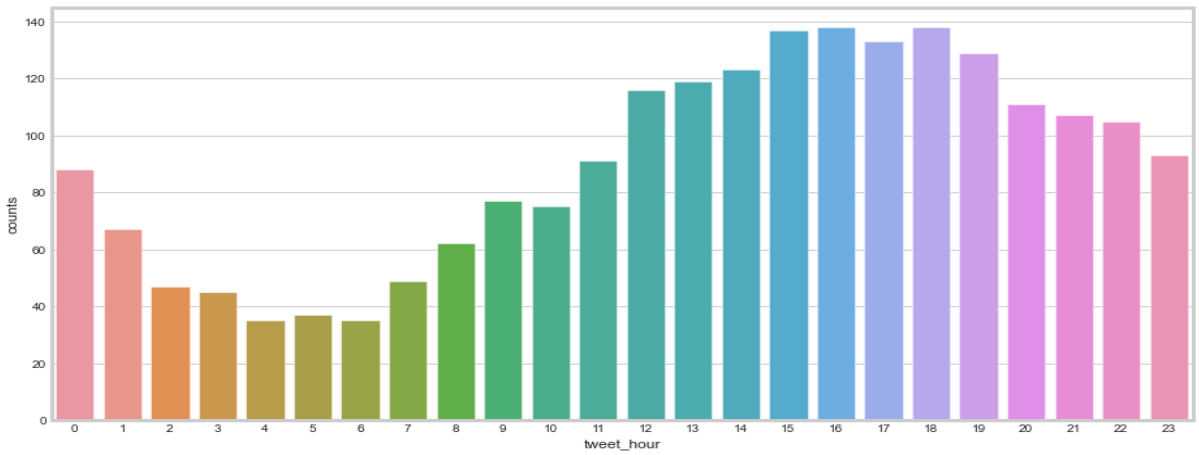


شكل (8) يوضح متوسط أطوال التغريدات

من خلال الشكل (8) الذي يوضح أطوال التغريدات تبدأ من 0 إلى أكثر من 400، يبين أن متوسط أطوال التغريدات يقع بين 100 و150 تقريبا، وباقي القيم بالإمكان اعتبارها قيم شاذة وهذه الأطوال .

4.12.3تحليل خاصية الساعات

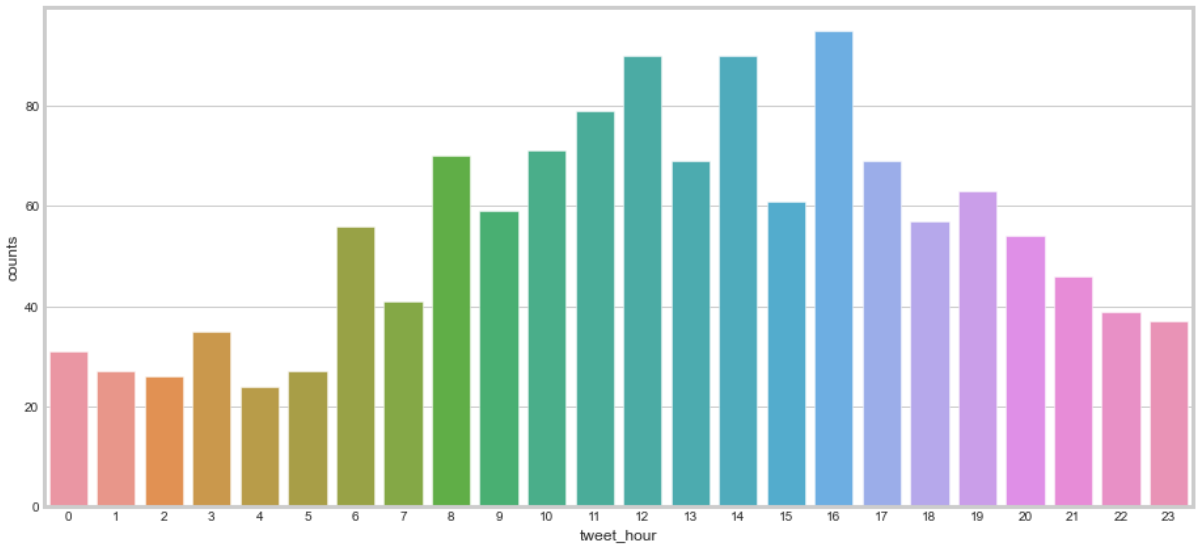
1.4.12.3توزيع التغريدات على الساعات لسنة 2012



يوضح شكل (10) أكثر الساعات تغريد في سنة 2012

الشكل السابق يبين مدرج تكراري لاكثر الأوقات التي يوجد بها تغريدات خلال سنة 2012، في الساعات الأولى نلاحظ قلة في عدد التغريدات، ومن ثم يرتفع عدد التغريدات من منتصف النهار حتى يصل إلى أعلى ما يكون عليه في الساعة 18:00.

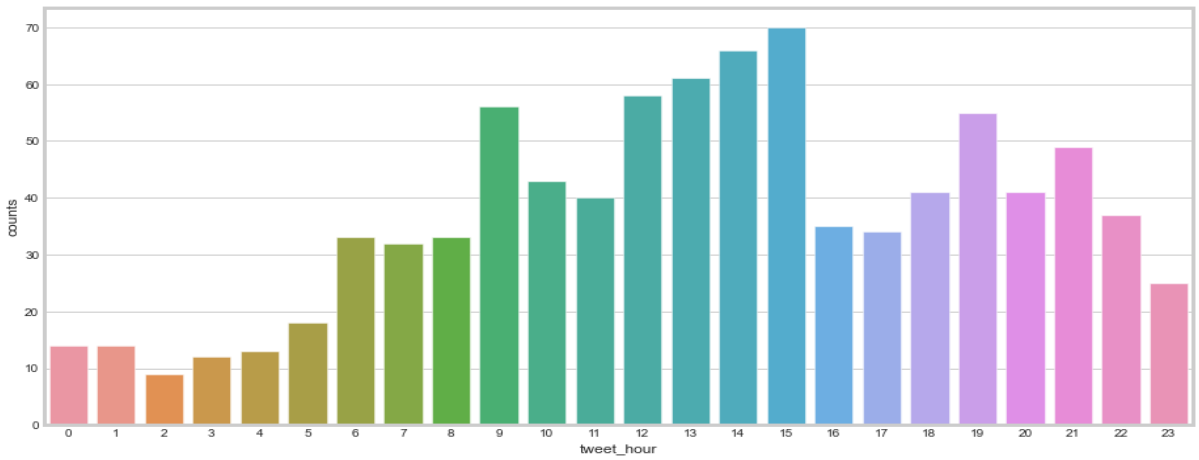
2.4.12.3توزيع التغريدات على الساعات لسنة 2013



يوضح شكل (11) أكثر ساعات تغريدا في سنة 2013

يوضح الشكل عدد التغريدات خلال 24 ساعة في اليوم لسنة 2013 كاملة، نلاحظ من الرسم عدد التغريدات في الساعات متفاوت يقل مرة ويرتفع أخرى، حيث كان أكبر عدد تغريدات في اليوم عند الساعة 16:00 مساءاً.

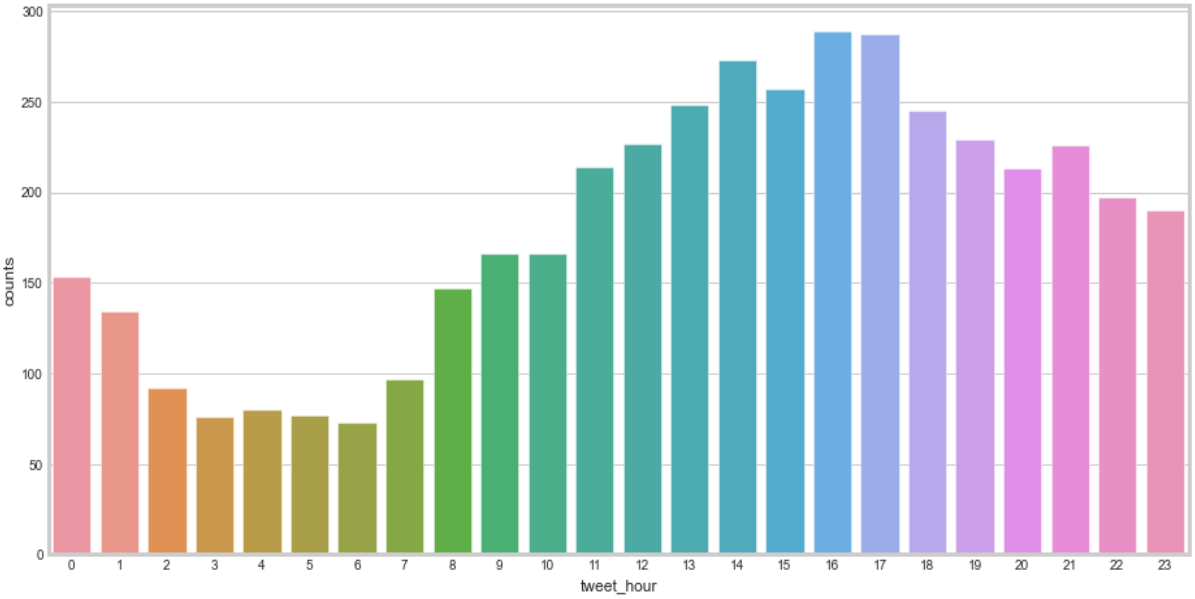
3.4.12.3توزيع التغريدات على الساعات لسنة 2014



يوضح شكل (12) أكثر ساعات تغريدا في سنة 2014

يوضح الشكل مدرج تكراري لتوزيع عدد التغريدات لسنة 2014 على 24 ساعة كل تغريدة حسب الساعة التي تم التغريد بها فيها، نلاحظ من الرسم ارتفاع عدد التغريدات تدريجيا حتى يصل عند الساعة 15:00.

4.4.12.3توزيع التغريدات على الساعات لكل السنوات

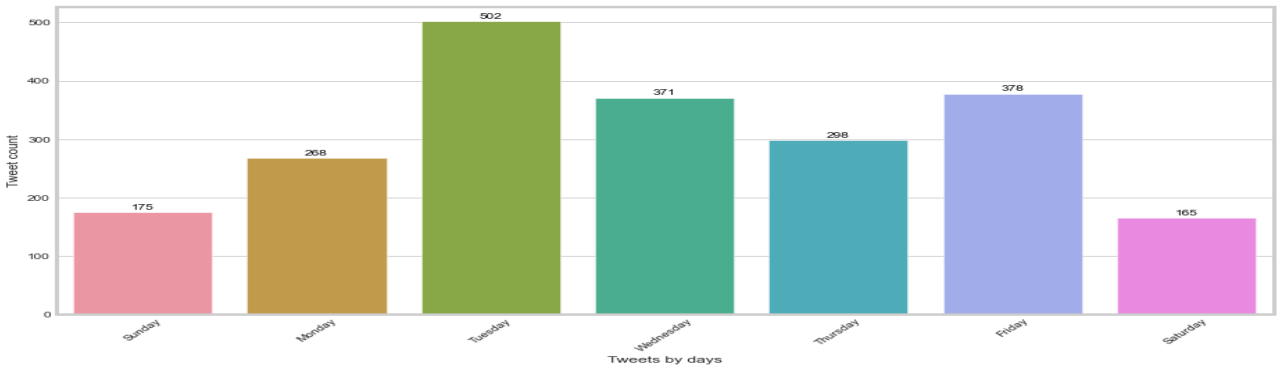


يوضح شكل (13) أكثر الساعات تغريدا لكل سنوات

يبين الشكل توزيع التغريدات على 24 ساعة كل تغريدة حسب الساعة التي تم التغريد بها فيها، خلال سنة 2012 و2013 و2014 مجمعة في الساعات الأولى عدد التغريدات كان قليل نوعاً ما، ويزداد تدريجيا حتى الساعة 16:00 كان أعلى ما يمكن للسنوات الثلاثة، ربما يرجع السبب في ذلك إلى أن الساعة 16:00 مساءاً على الأغلب تعتبر وقت الفراغ بالنسبة لمعظم المغردين، وهذه الأوقات إجابة السؤال العاشر في أسئلة البحث.

5.12.3تحليل خاصية الأيام

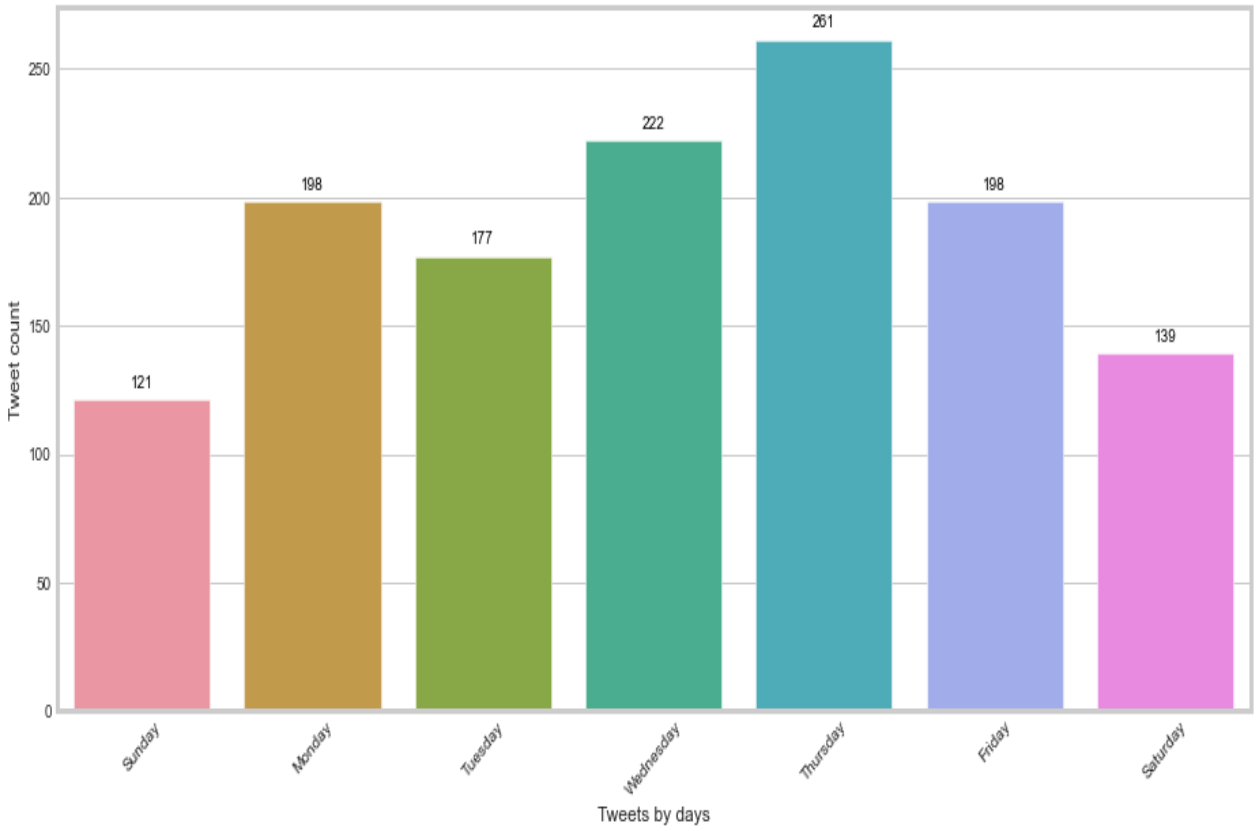
1.5.12.3توزيع التغريدات على الأيام لسنة 2012



يوضح شكل (14) أكثر الأيام تغريدا في سنة 2012

يوضح الشكل توزيع التغريدات على أيام الأسبوع خلال سنة 2012، نلاحظ أن عدد التغريدات خلال يوم الثلاثاء كان 497 تويت وهو العدد الأكبر خلال أيام الأسبوع لهذه السنة، وأقل عدد تغريدات حصل يوم السبت وبلغ 161 تويت، وتقارب عدد التغريدات في يوم الإربعاء والخميس والجمعة باعتبار أنها آخر أيام الأسبوع.

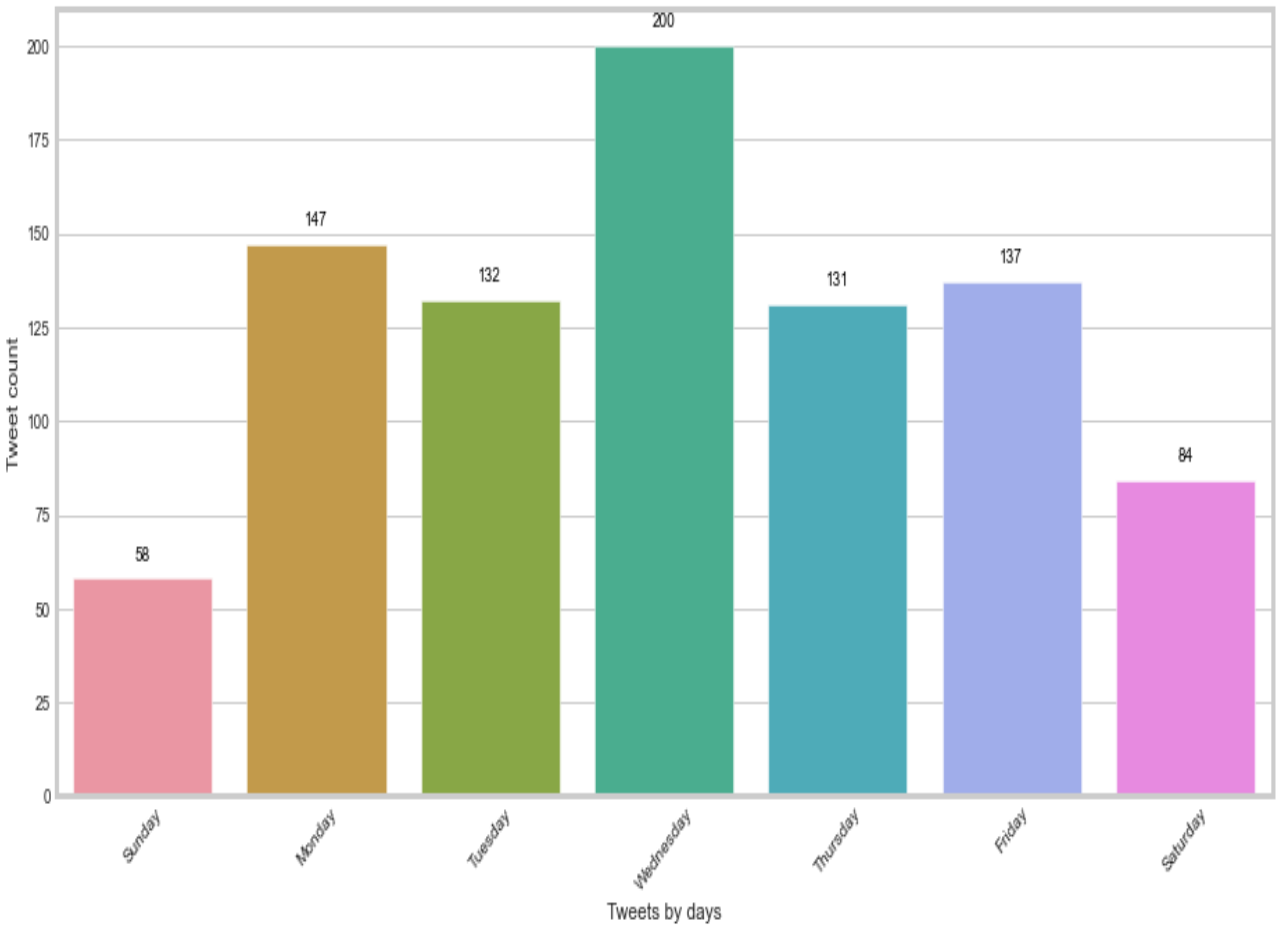
2.5.12.3توزيع التغريدات على الأيام لسنة 2013



يوضح شكل (15) أكثر الأيام تغريدا في سنة 2013

يبين الشكل السابق توزيع التغريدات على أيام الأسبوع خلال سنة 2013، على الأغلب كانت معظم التغريدات خلال هذه السنة متواجدة يوم الخميس حيث بلغت 261 تويته، وأقلها يوم الأحد عددها 121 تويته، نلاحظ ارتفاع عدد التغريدات في نهاية الأسبوع وتقل مع بداية الأسبوع؛ ربما يرجع السبب في ذلك أن يوم الخميس يعتبر يوم الراحة، ويوم الأحد قد يكون يوم الدوام الرسمي لبعض الدول.

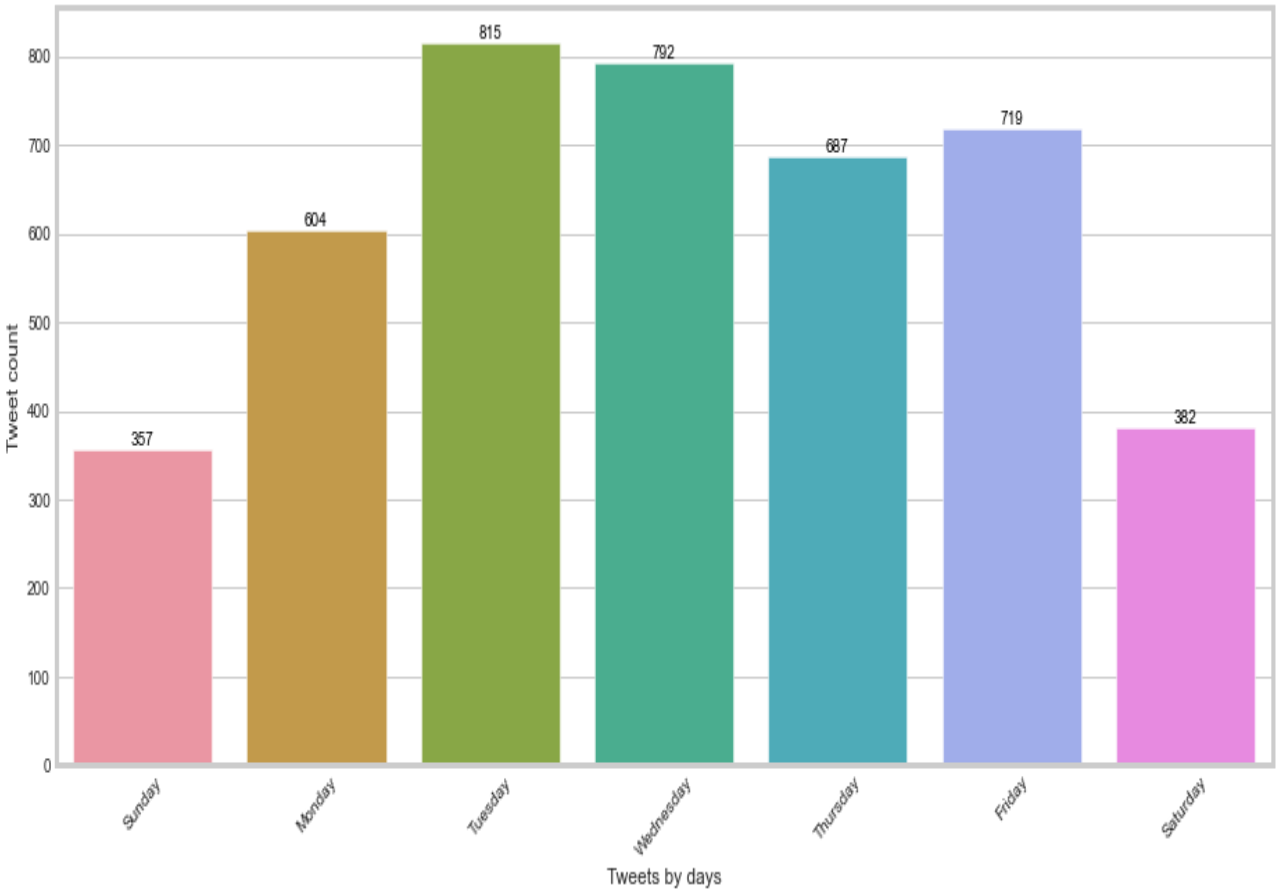
3.5.12.3توزيع التغريدات على الأيام لسنة 2014



يوضح شكل (16) أكثر الأيام تغريدا في سنة 2014

يوضح الشكل مدرج تكراري لتوزيع التغريدات على أيام الأسبوع خلال سنة 2014، نلاحظ ارتفاع لعدد التغريدات يوم الإربعاء خلال سنة 2014 حيث يصل عددها 200 تغريدة وأقلها يوم الأحد بلغ 58 تغريدة أي أن التغريدات تقل نهاية الأسبوع وترتفع مع بدايته.

4.5.12.3توزيع التغريدات على الأيام لكل السنوات

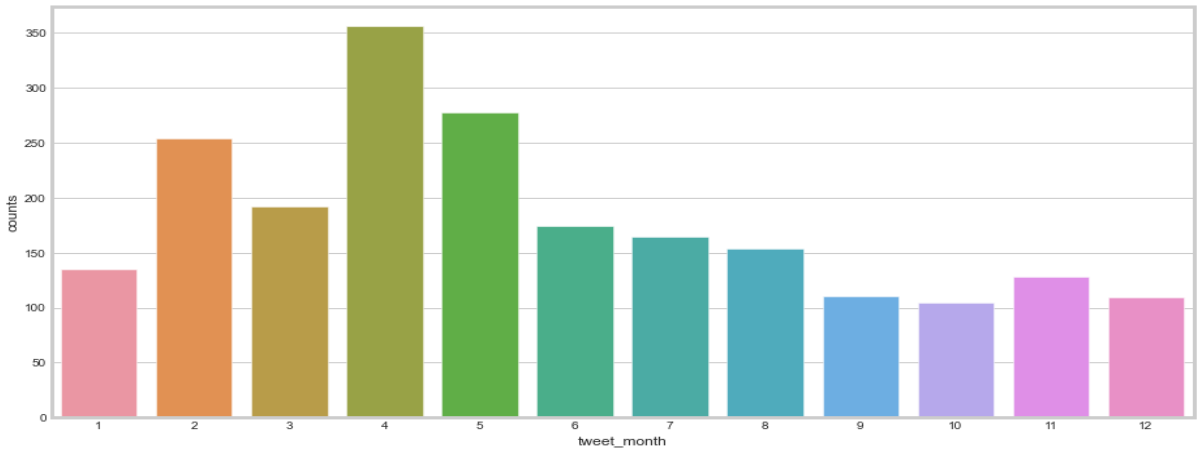


يوضح شكل (17) أكثر الأيام تغريدا في كل السنوات

يبين الشكل التالي أيام الأسبوع من حيث عدد التغريدات التي تم تغريدها في كل يوم خلال السنوات الثلاثة في قاعدة البيانات المجمعة، نلاحظ من خلال الشكل تواجد أكبر عدد تغريدات في يوم الثلاثاء ويبلغ 815 تويت واقل عدد تغريدات حصل يوم الأحد ويبلغ 357 تويت، اعتقد السبب في ذلك انّ يوم الأحد يعتبر يوم الدوام الرسمي على الأغلب لكافة الدول فينشغل الأشخاص بأعمالهم، وبالتالي يفسر ذلك قلة عدد التغريدات في بداية الأسبوع وازدادت في نهاية الأسبوع، وهذه الأيام إجابة السؤال التاسع في أسئلة البحث.

6.12.3تحليل خاصية الأشهر

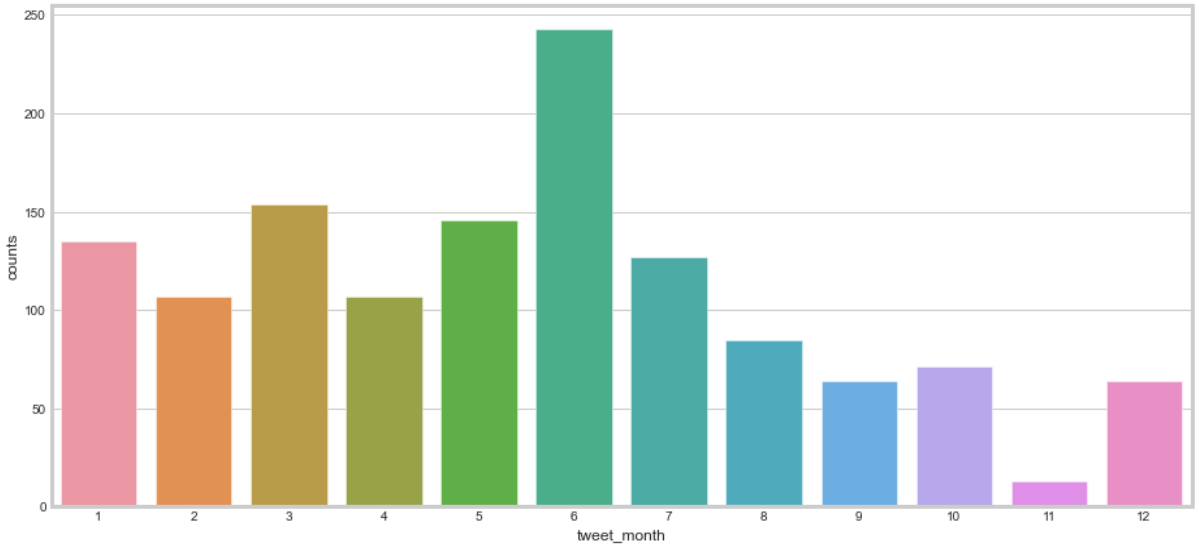
1.6.12.3توزيع التغريدات على الأشهر لسنة 2012



يوضح شكل (18) أكثر الأشهر تغريدا في سنة 2012

يوضح الشكل (18) توزيع تكراري لعدد التغريدات خلال الأشهر لسنة 2012، نلاحظ من خلال الشكل أن عدد تكرار التغريدات في بداية السنة أعلى من نهايتها، في شهر أكتوبر كانت التغريدات أقل ما يمكن خلال السنة، أعتقد أن يرجع السبب في ذلك أن الهجرة كانت في أقل حالاتها، وتكرار التغريدات في شهر أبريل كان الأكثر خلال سنة 2012، إضافة إلى ذلك كانت التغريدات متفاوتة في بداية السنة ومتقاربة في نهايتها.

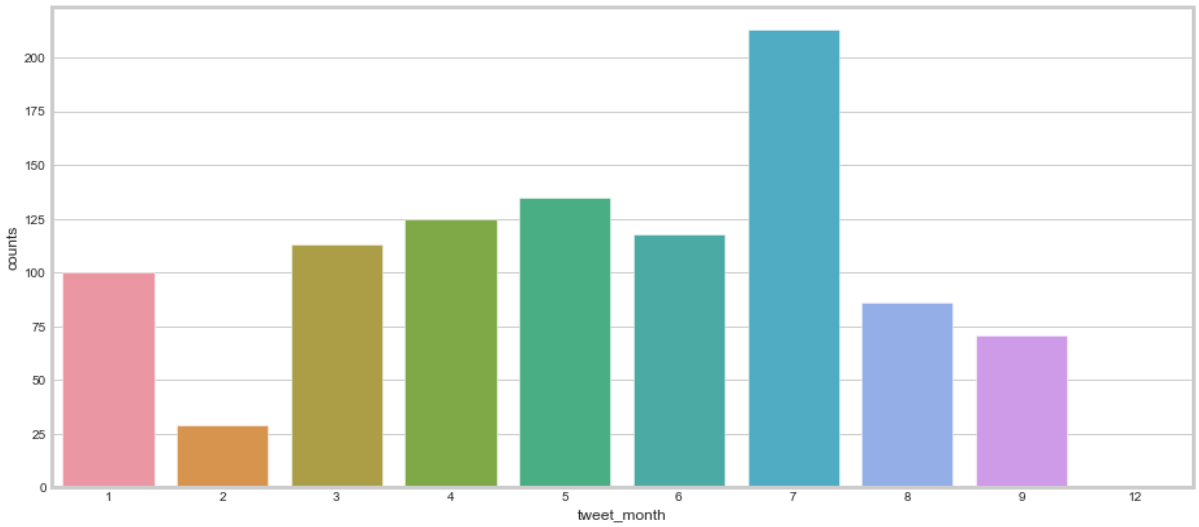
2.6.12.3توزيع التغريدات على الأشهر لسنة 2013



يوضح شكل (19) أكثر الأشهر تغريدا في سنة 2013

يوضح الشكل توزيعاً تكرارياً للتغريدات لسنة 2013 على أشهر السنة، يبين تكرار متفاوت على الأعمدة، نلاحظ أعلى تكراراً كان في شهر يونيو، وأقلها تكراراً في شهر نوفمبر حيث كانت نسبة التغريدات ضئيلة جدا، ربما كان الطقس في هذا الشهر غير ملائم للهجرة.

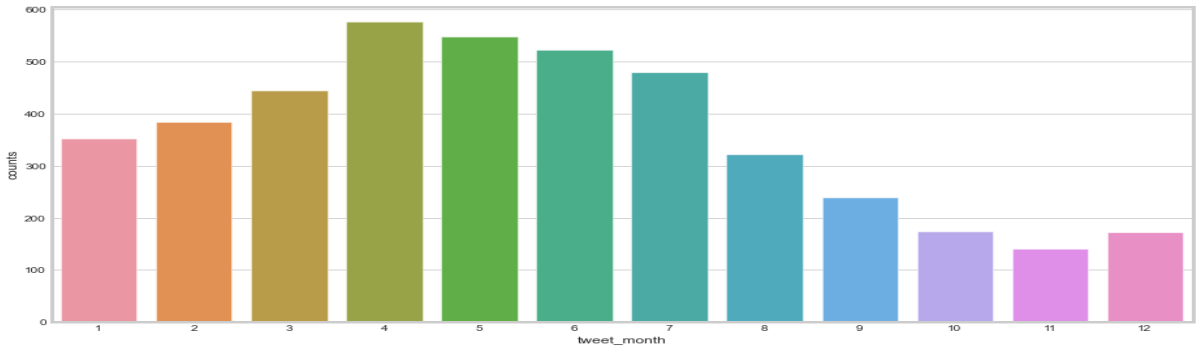
4.6.12.3توزيع التغريدات على الأشهر لسنة 2014



شكل (20) أكثر الأشهر تغريدا في سنة 2014

يوضح الشكل توزيع تكراري لعدد التغريدات خلال الأشهر لسنة 2014، نلاحظ من خلال الشكل أن عدد تكرار التغريدات خلال شهر ديسمبر كان شبه معدوم او على الأرجح لا يوجد تغريدات، وكان تكرار التغريدات في شهر يوليو الأكثر خلال سنة 2014، قد يكون في شهر ديسمبر الأمواج كثيفة وبالتالي هناك خطورة في الهجرة على المهاجرين عبر البحر، قلة الهجرة أو انعدامها يفسر عدم تواجد تغريدات في هذا الشهر.

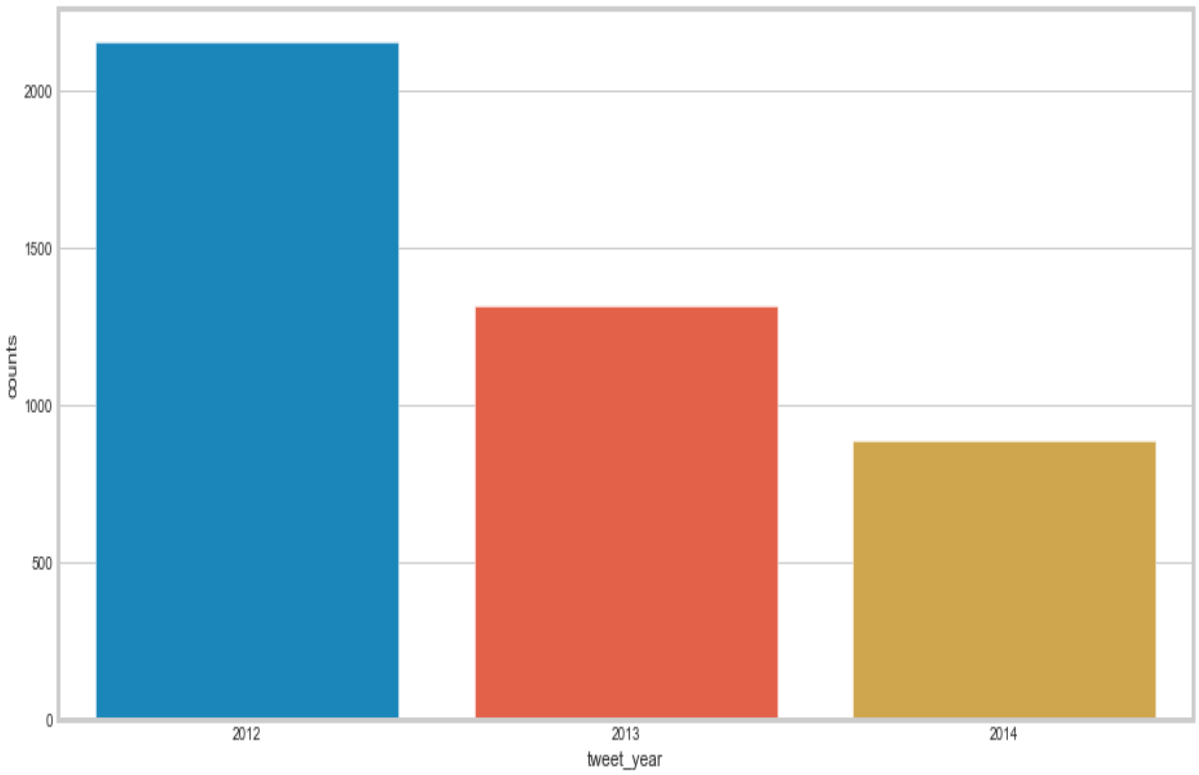
4.6.12.3توزيع التغريدات على الأشهر لكل السنوات



يبين شكل (21) أكثر الأشهر تغريدا في كل سنوات

فـيما سبق تم توزيع التغريدات على 12 شهر في كل سنة من السنوات الثلاثة في قاعدة البيانات على حدى، حيث يبين المدرج التكراري توزيع التغريدات على الأشهر في السنوات الثلاثة مجمعة، على سبيل المثال التغريدات التي تم تغريدها في شهر يناير سنة 2012 وفي شهر يناير 2013 وفي شهر يناير 2014 وهكذا، نلاحظ أن أكبر عدد من التغريدات في شهر أبريل ومايو وأغسطس ويليها بمعدل متوسط شهر يناير وفبراير ومارس وأقلها في أشهر سبتمبر وأكتوبر ونوفمبر وديسمبر.

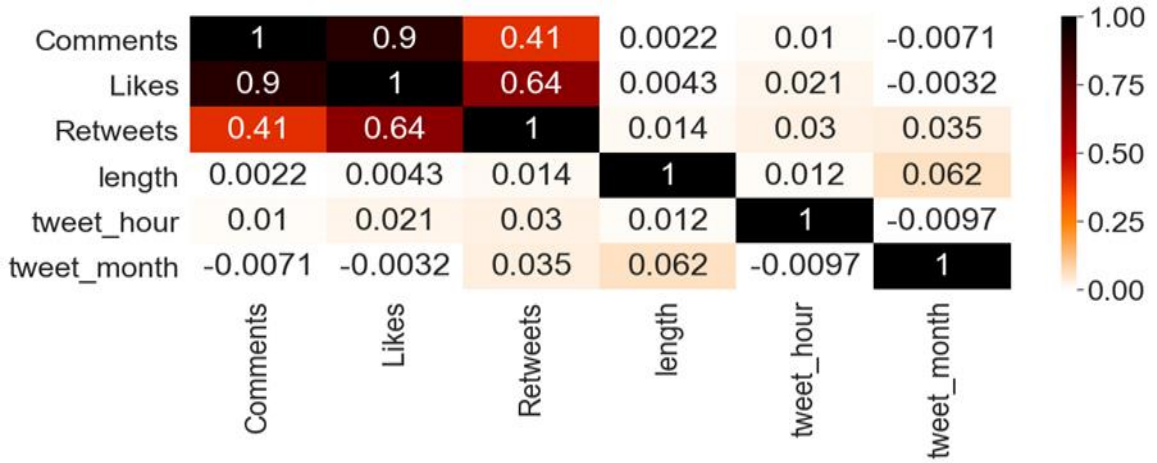
7.12.3تحليل خاصية السنوات



شكل (22) يبين أكثر السنوات تغريد

شكل (22) يوضح أكثر التغريدات على السنوات الثلاثة التي سيتم دراستها في قاعدة البيانات، من خلال الرسم نلاحظ أن هناك تزايد في عدد التغريدات في سنة 2012 مقارنة بسنة 2013 و 2014، وأقل عدد تغريدات كان في سنة 2014، أعتقد أن السبب في ذلك تزايد عدد المهاجرين في سنة 2012 بعد سقوط النظام السابق في سنة 2011 أصبح الحديث عن الهجرة مثير للجدل وازدادت التغريدات في سنة 2012 وقلت مع تقدم السنوات.

تحليل ارتباط الخصائص عن طريق person correlation بعد إضافة السمات الجديدة :



شكل (23) يوضح ارتباط الخصائص بعد اضافة الخصائص جديد لقاعدة البيانات

يوضح الشكل قوة العلاقة بين الخصائص أو السمات، كلما كان اللون أدكن، كلما كانت العلاقة بين خاصيتين أقوى، يفسر ذلك قوة العلاقة بين Comments و Likes و Retweets

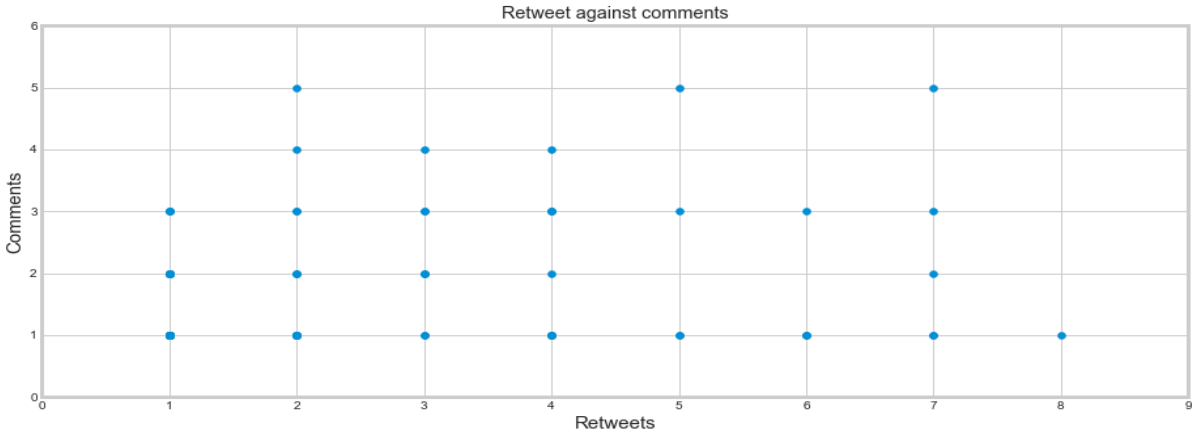
حيث نلاحظ من خلال الرسم هناك علاقة قوية بين تلك السمات الثلاثة، لذلك سيتم رسم مخطط البعثرة يوضح توزيع العلاقة بين تلك السمات.

14.3تحليل احصائي لخاصيتين (Bivariate analysis):

هو دراسة العلاقة الموجودة بين متغيرين، يساعد في معرفة ما إذا كان هناك ارتباط بين المتغيرات وذلك باستخدام المخطط المبعثر وهو رسم بياني لمتغيرين يتم استخدامه لفهم ما إذا كانت هناك أي علاقة بين متغيرين، يمكن أن تكون العلاقة خطية أو غير خطية، كما أنها تستخدم لتحديد القيم المتطرفة.

المخططات التالية ستوضح العلاقات بين الفترات الأكثر تفضيلاً حسب الساعات والأيام والأشهر والسنة، وفترات إعادة التغريد حسب الساعات والأيام والشهور والسنة، والعلاقة بين التغريدات وإعادة التغريد، والعلاقة بين التغريدات والفترات الأكثر تفضيلاً، حيث تم عرضها ومن ثم تحليلها بعد التخلص من القيم المتطرفة في البيانات، لأن تمثيل البيانات بوجود القيم المتطرفة تأثر نظرا لتمركز البيانات في مساحات ضيقة، وبالتالي ليس بالإمكان تحديد العلاقات بشكل منطقي وواضح بوجود هذه القيم، يمكننا أن نرى النقاط العشوائية ولكن ما العلاقة التي يمكننا رؤيتها؟

1.14.3تحليل العلاقة بين الخاصيتين إعادة التغريد والتعليقات:



شكل (24) يبين العلاقة بين خاصيتين إعادة التغريد والتعليقات

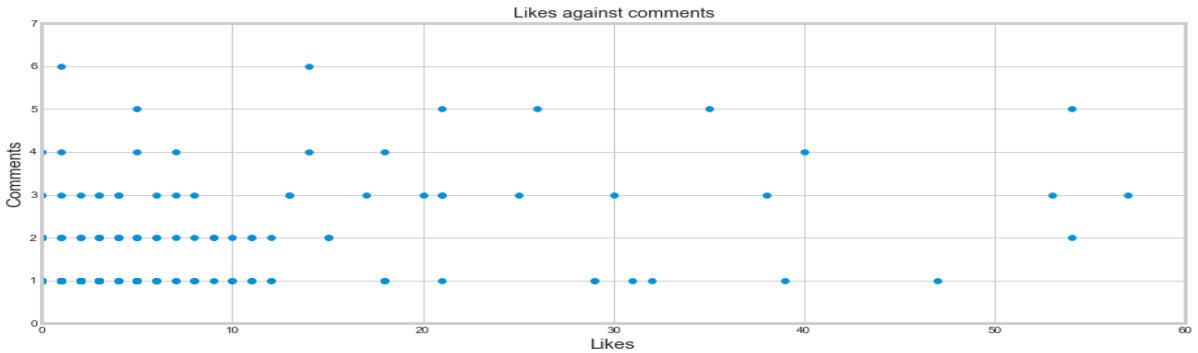
لاحظنا من العلاقة بين إعادة التغريد والتعليق للبيانات المجمعة لكل السنوات كانت كالتالي:

أولا: كانت التعليقات أكثر عندما كان إعادة التغريد أقل من أربعة.

ثانيا: كان التفاعل بالتعليق غير منضبط عندما كانت قيم إعادة التغريد من أربعة إلي ثمانية، أي مرة يمكن وصفه بالزيادة ومرة بالنقصان.

ثالثا: كانت أكبر قيمة لإعادة التغريد (66)، وأكبر قيمة للتعليقات (420)، وذلك حسب الإحصائيات الأولية للبيانات (أو إحصائيات الارقام الخمسة) كما عرضت في الشكل(24).

2.14.3تحليل العلاقة بين الخاصيتين الإعجاب والتعليقات:



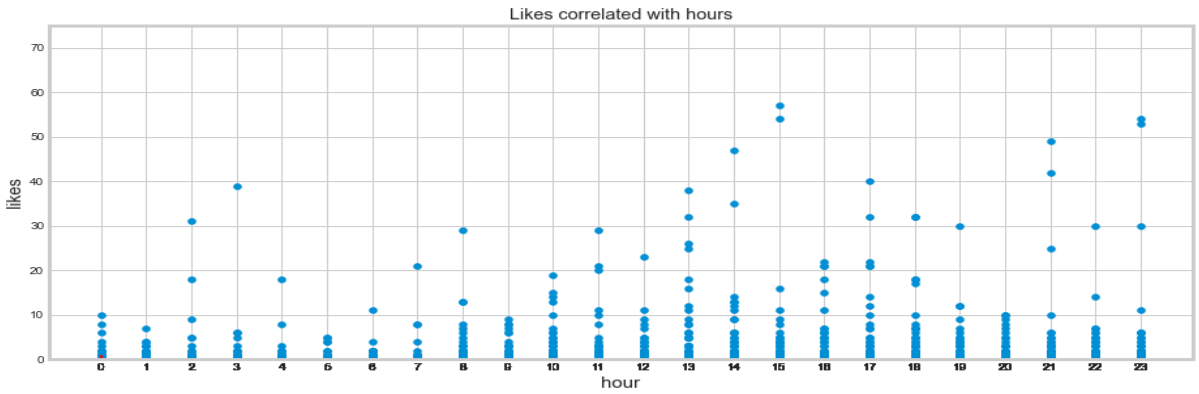
شكل (25) يبين العلاقة بين الخاصيتين الاعجابات والتعليقات

من خلال الرسم نلاحظ العلاقة بين أكثر الفترات تفضيلاً والتعليقات للتغريدات المجمعة لكل السنوات سيتم توضيحها كالتالي:

أولا: التعليقات أكثر عندما كان التفضيل أقل من الثانية عشر.

ثانيا: من الثانية عشر أصبح التفاعل بين التفضيل والتعليق يقل وفي نفس الوقت متذبذب، حيث كانت أكبر قيمة لتعليقات (420) وأكبر قيمة للتفضيل (600).

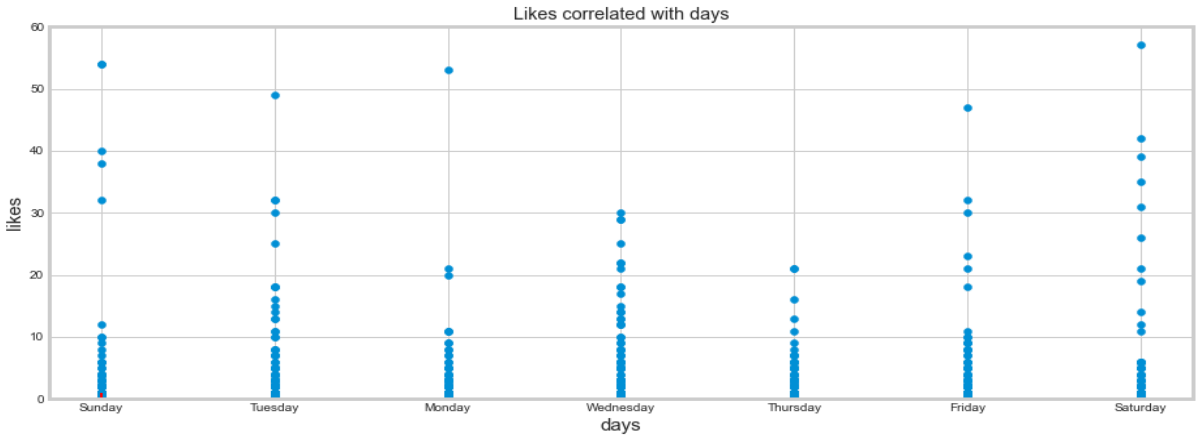
3.14.3تحليل العلاقة بين خاصيتين الإعجاب والساعات



شكل (26) يبين العلاقة بين الخاصيتين الاعجابات والساعات

يوضح الشكل السابق مخطط البعثرة لأكثر الفترات تفضيلاً حسب الساعات للبيانات المجمعة لكل السنوات، نلاحظ من الرسم فترات التفضيل ليست ثابتة أي يمكن وصفها مرة بالزيادة ومرة بالنقصان، كان التفضيل في الفترات الصباحية قليل نوعا ما لا يتعدى 20، من الساعة 13:00 إلى الساعة 23:00 زاد معدل التفضيل حيث كان أعلى تفضيل عند الساعة 15:00 وصل حتى 58 تقريبا، يفسر الشكل أن ساعات المساء كان التفضيل بها أكثر من ساعات الصباح، وهذه الفترات الأكثر تفضيلا إجابة السؤال الثالث عشر في أسئلة البحث.

4.14.3تحليل العلاقة بين الخاصيتين الإعجاب والأيام



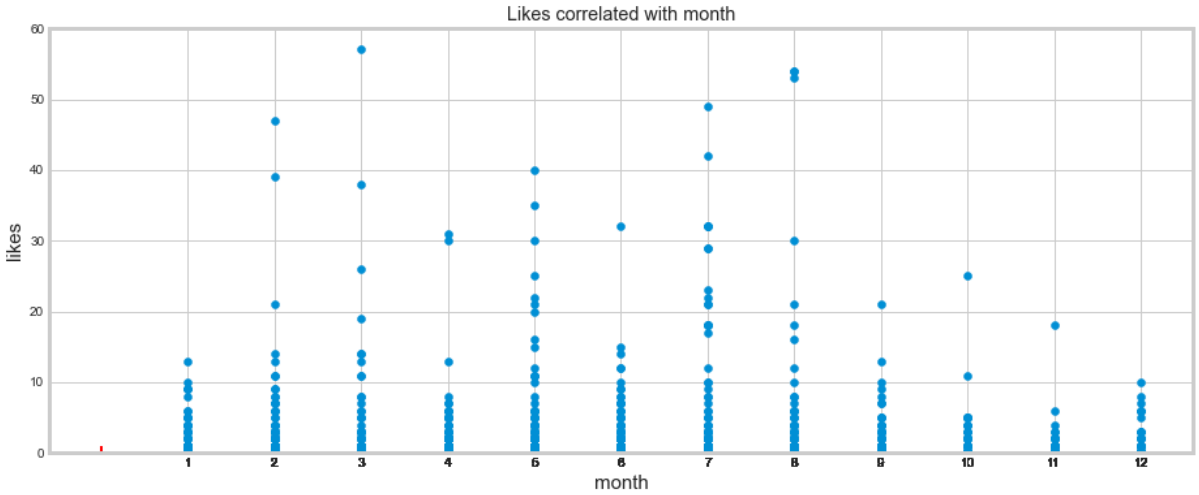
شكل (27) يبين العلاقة بين الخاصيتين الاعجابات والأيام

يوضح الشكل السابق مخطط البعثرة للعلاقة بين أكثر الفترات تفضيلاً والأيام للبيانات المجمعة لكل السنوات:

اولاً: كان التفضيل مرتفع بداية الأسبوع.

ثانياً: وجد أعلى تفضيل خلال أيام الأسبوع في يوم السبت وبلغ 58، وأقل تفضيل يوم الخميس وبلغ 21.

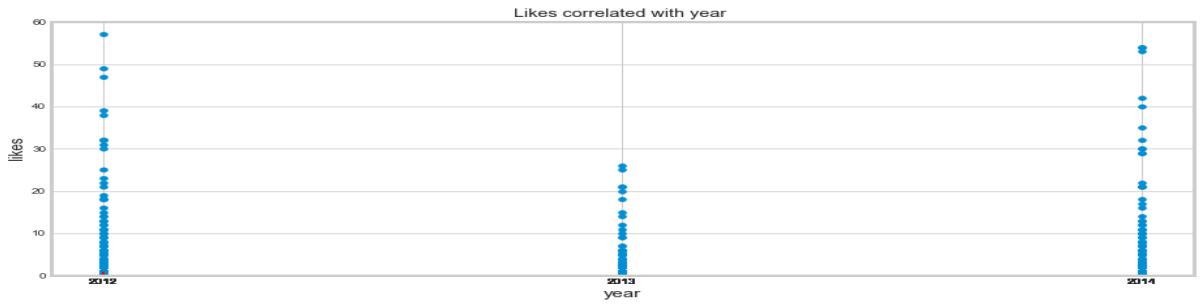
5.14.3 تحليل العلاقة بين الخاصيتين الإعجاب والأشهر



شكل (28) يبين العلاقة بين الخاصيتين الإعجابات والاشهر

يبين الشكل السابق مخطط البعثرة لأكثر الفترات تفضيلاً حسب الأشهر للبيانات المجمعة لكل السنوات، نلاحظ أن قيم التفضيل في الأشهر معظمها يتجاوز 20 باستثناء شهر 1 و11 و12، كانت فترات التفضيل متقطعة خلال الأشهر، أعلى قيمة تفضيل وجدت في شهر 3 وبلغت 58.

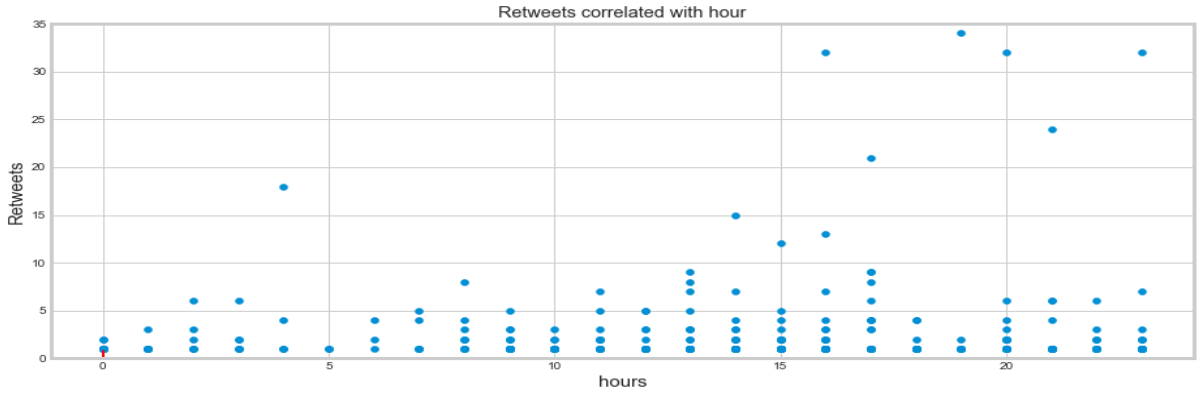
6.14.3 تحليل العلاقة بين الخاصيتين الإعجاب والسنوات



شكل (29) يبين العلاقة بين الخاصيتين الإعجابات والسنوات

يبين الشكل السابق مخطط البعثرة لأكثر الفترات تفضيلاً حسب السنوات للبيانات المجمعة لكل السنوات، أعلى تفضيل خلال السنوات الثلاثة كان في سنة 2012 حيث وصل 58 ، تليها سنة 2014 بلغ التفضيل فيها حتى 54، أقل تفضيل كان في سنة 2013 لا يتجاوز 26.

7.14.3 تحليل العلاقة بين الخاصيتين إعادة التغريد والساعات



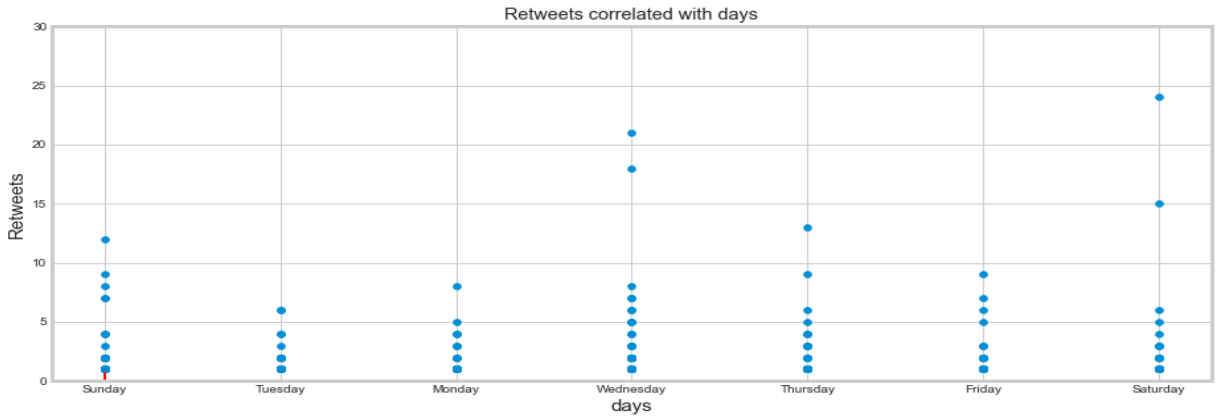
شكل (30) يبين العلاقة بين الخاصيتين إعادة التغريد والساعات

لاحظنا من العلاقة بين إعادة التغريد والساعات للبيانات المجمعة لكل السنوات كانت كالتالي:

أولاً: بشكل عام كان إعادة التغريد أقل من 15 في معظم الساعات.

ثانياً: فقط عند الساعات (4، 16، 17، 19، 20، 21، 23) كان إعادة التغريد فوق 15، عند الساعة 19:00 كان إعادة التغريد أعلى ما يمكن خلال 24 ساعة وبلغ 34،

8.14.3 تحليل العلاقة بين الخاصيتين إعادة التغريد والأيام



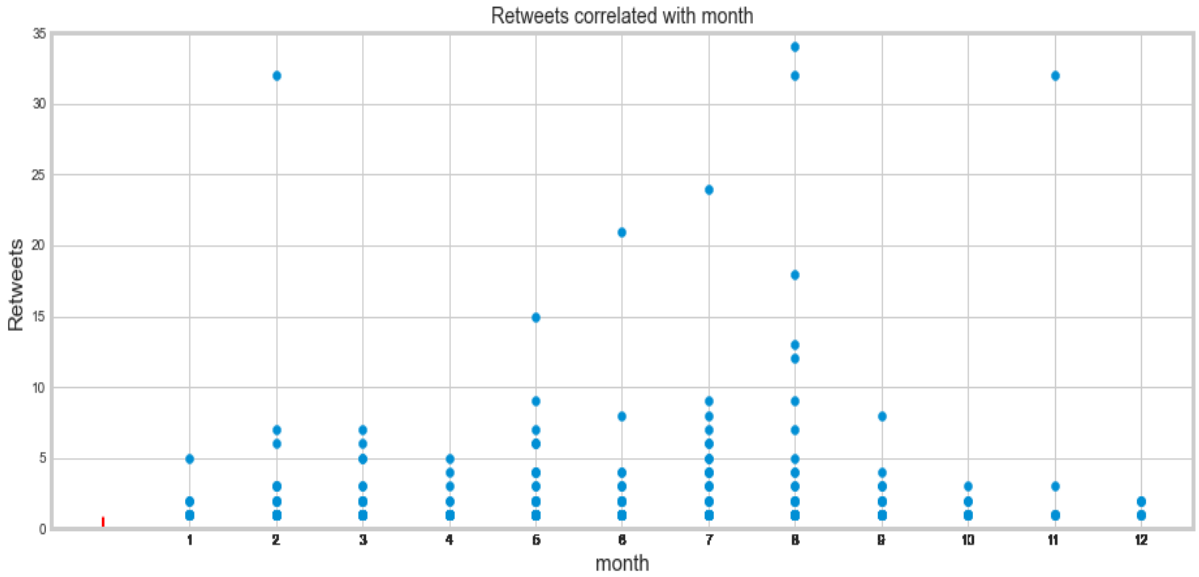
شكل (31) يبين العلاقة بين الخاصيتين اعادة التغريد والايام

يوضح الشكل السابق العلاقة بين إعادة التغريد والأيام للبيانات المجمعة لكل السنوات حيث كانت كالتالي:

أولاً: توزيع قيم إعادة التغريد على الأيام لا تتجاوز 24.

ثانياً: كان إعادة التغريد خلال الأيام كالتالي: يوم السبت وصل 24، يوم الأحد وصل 21، يوم الأثنين بلغ 8، يوم الثلاثاء بلغ 6، يوم الإربعاء بلغ 21، يوم الخميس بلغ 13، الجمعة بلغ 9، وهذا إجابة السؤال الثالث عشر في أسئلة البحث.

9.14.3تحليل العلاقة بين الخاصيتين إعادة التغريد والاشهر



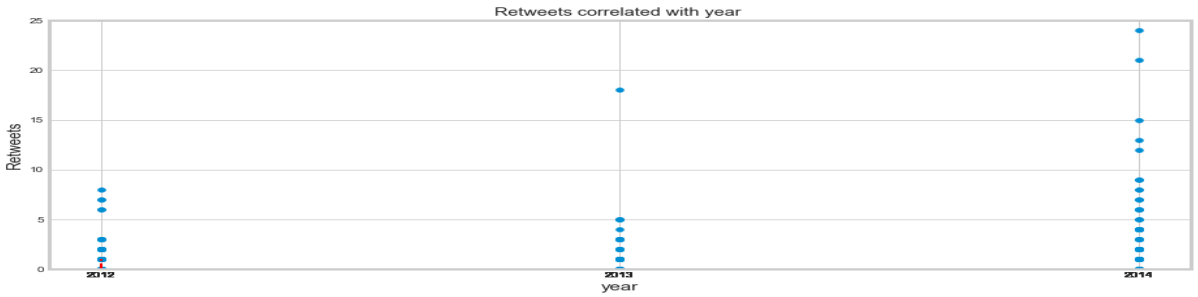
شكل (32) يبين العلاقة بين الخاصيتين اعادة تغريد والاشهر

يوضح الشكل السابق مخطط البعثرة للعلاقة بين إعادة التغريد والأشهر للبيانات المجمعة لكل السنوات، حيث كانت العلاقة كالتالي:

أولاً: كان إعادة التغريد خلال الأشهر الستة (1، 3، 4، 9، 10، 12) قليل نوعاً ما لا يتجاوز 10.

ثانياً: عند الأشهر(2، 5، 6، 7، 8، 11) كان إعادة التغريد مرتفع، حيث بلغ أعلى قيمة له 34 في شهر 8.

10.14.3تحليل العلاقة بين الخاصيتين إعادة التغريد والسنوات



شكل (33) يبين العلاقة بين الخاصيتين اعادة تغريد والسنوات

يبين الشكل السابق مخطط البعثرة للعلاقة بين إعادة التغريد حسب السنوات للبيانات المجمعة لكل السنوات، إعادة التغريد خلال سنة 2012 كان ضعيف نوعا ما لا يتجاوز 8، ووصل إعادة التغريد 18 في سنة 2013، وكان أعلى ما يمكن في سنة 2014 حيث وصل إعادة التغريد حتى 24.

الفصل الرابع

اطار العمل

1.4البيئة البرمجية:

Anaconda Navigator 1.1.4

هذا سطح مكتب واجهة المستخدم الرسومية المضمنة في توزيع Anaconda والتي تتيح لك تشغيل التطبيقات وإدارة حزم Anaconda بسهولة والبيئات والقنوات دون استخدام أوامر سطر الأوامر يستطيع Navigator البحث عن الحزم على سحابة Anaconda أو في مستودع Anaconda المحلي، متاح لأنظمة Windows و macOS و Linux (امبارك أسامة، 2018).

Jupyter Notebook 2.1.4

أحد أشهر وأهم الأدوات المُستخدمة أثناء تحليل البيانات، بالإضافة إلى ذلك هي شيء أساسي لا يستغني عنها عالم البيانات، وذلك لما تُقدمه من مميزات وخصائص تُسهل من التعامل مع البيانات و الشيفرة البرمجية، تُعتبر Jupiter أداة قوية يُمكن استخدامها تفاعليًا في مشاريع علم البيانات وتعليم الآلة.

3.1.4 لغة البرمجة المستخدمة:

قمنا باستخدام لغة بايثون Python وهي منصة مفتوحة المصدر عالية المستوى مُفسَّرة ذات مجالٍ عام، وهي مرنةٌ وتحاول التعبير عن المفاهيم البرمجية بأقل قدر ممكن من الأكواد، تدعم لغة Python البرمجة الكائنية والبرمجة الإجرائية، وفيها مكتبة قياسية كبيرة، تستخدم Python في كل شيء بدءًا من البرامج النصية السريعة وحتى خوادم الويب الكبيرة والقابلة للتطوير، حيث تعتبر لغة البرمجة الأسرع نموا في الاونة الاخيرة؛ وهي لغةٌ مفتوحة المصدر، ومدعومةٌ من أغلبية أنظمة التشغيل.

4.1.4المكتبات والأدوات المستخدمة مع لغة البرمجة:

Pandes

الباندا هي مكتبة برمجية مُطورة بلغة البرمجة بايثون لمعالجة البيانات وتحليلها، وبالتحديد فهي تقدم هياكل بيانات وعمليات التلاعب بالجداول الرقمية و السلاسل الزمنية، الهدف الأساسي لمكتبة Pandas هو إجراء ما يسمى بـ Data Munging، والمقصود به هو إجراء تغييرات على بيانات

أساسية خام Raw data بحيث ينتج عن هذا التغيير تحويل البيانات الى شكل أخر يُمكن فهمه والتعامل معه.

numpy

هي إضافة على لغة البرمجة بايثون، تُستخدم للتعامل مع المصفوفات الكبيرة والحقول متعددة المستوى، وكذلك توفر مكتبة كبيرة من الاقترانات الرياضية عالية المستوى للعمل على هذه الحقول والمصفوفات، متخصصة في الحوسبة العلمية بلغة البايثون، وتحتوي على تشكيلة متنوعة من الأدوات والتقنيات التي من الممكن ان تستخدمها لحل مشاكل رياضية.

seaborn

هي مكتبة متخصصة بتحليل البيانات و دراستها بحيث تزود واجهة ذو مستوى عال لرسم رسومات تختص بالإحصاء، جذابة و غنية بالمعلومات تم بنائها بأستخدام matplotlib حيث تعتبر أساس لها، و لها القدرة على التعامل مع الهياكل التي تدعمها pandas و بذلك تعتبر من أهم المكتبات التي تستخدم في (data visualization)، بحيث توفر المرونة اللازمة للتفاعل مع البيانات الموجودة و كانت seaborn داعم و مكمل لمهام matplotlib.

Matplotlib

هي مكتبة مكتوبة ب بايثون مهمتها عمل الرسومات البيانية ذات البعدين D2، وتعتمد تلك المكتبة على NumPy للتعامل مع المصفوفات الكبيرة لضمان أداء أفضل، Matplotlib تعتبر مشابهة لبيئة MatLab و منها إشتقت matplotlib إسمها وتستطيع مع matplotlib إستخدام الأوامر السطرية مثل MatLab كما يمكنك دمج matplotlib في برامج المبنية على البرمجة كائنية التوجه OOP.

SpaCy

هي مكتبة مفتوحة المصدر لمعالجة اللغة الطبيعية (NLP) في بايثون تتيح هذه المكتبة العديد من المهام التي تساعد في معالجة اللغة الطبيعية، وتدعم العديد من لغات مثل: الإنجليزية والألمانية واليونانية والإسبانية والفرنسية والإيطالية واليونانية والنرويجية البوكمالية والهولندية والبرتغالية، توجد العديد من المهام التي تقوم بها المكتبة SpaCy فيما يلي سيتم عرضها، يوضح جدول (10) مهام مكتبة spacy مع شرح مبسط لها.

مهام	معناها	الوصف
Tokenization	الترميز	تقسيم النص إلى كلمات وعلامات الترقيم وما إلى ذلك
Part of speech Tagging (pos)	وضع علامات على جزء من الكلام	هو نوع من تصنيف الكلمات يعين أجزاء من الكلام لكل كلمة مثل الاسم والفعل والصفة
Dependency Parsing	الإعراب التبعية	تحليل التبعية هو عملية إنشاء العلاقة بين الكلمات المختلفة للجملة ووصف أدوارها النحوية.
Lemmatization	توحيد	تحديد الصيغ الأساسية للكلمات
Sentence Boundary Detection (SBD)	كشف حدود الجملة	تحديد مكان بداية الجملة ونهايتها
Named Entity Recognition (NER)	التعرف على الكيان	تحديد الكيانات وتصنيفها مثل :–أشخاصه ، مواقع وغيرها
Entity Linking (EL)	ربط الكيانات	إزالة الغموض عن الكيانات النصية إلى معرفات فريدة في قاعدة المعرفة
Similarity	تشابه	مقارنة الكلمات والجمل والمستندات النصية ومدى تشابهها مع بعضها البعض
Text Classification	تصنيف النص	تصنيف النص هو تصنيف المستندات النصية تلقائيًا إلى فئة محددة واحدة أو أكثر مثل :–فهم مشاعر الجمهور من وسائل التواصل الاجتماعي
based Matching-Rule	المطابقة المستندة إلى القواعد	استخراج أنواع معينة من النصوص من خلال قواعد دون الحاجة إلى تدريب النماذج الإحصائية
Training	تمرين	تمرين وتحسين تنبؤات النموذج الإحصائي
Serialization	التسلسل	حفظ الكائنات في ملفات أو سلاسل بايت

جدول (10) يوضح مهام مكتبة spaCy

بعد التعرف على مكتبة spaCy ومهامها بشكل عام سيتم التركيز على هدف البحث وهو استخراج الكيانات المسماة (NER)، تعمل مكتبة spaCy على بعض المهام لديها بشكل مستقل بينما تحتاج المهام الأخرى بإجراء مهمة أو عدة مهام قبلها كما هو الحال في NER حيث سيتم شرحها لاحقا، بالإمكان تدريب بيانات واستخدام المكتبة spaCy لتحديد الكيانات المسماة NER، وكذلك توجد نماذج مدربة جاهزة لهذا الغرض.

2.4النماذج المدربة مسبقا الجاهزة في المكتبة:

للمكتبة spaCy العديد من النماذج المدربة الجاهزة تحتوي على المعرفة التي تم جمعها حول لغة معينة، تتوفر العديد من النماذج المدربة للغات مختلفة، في هذه الدراسة سنقوم بتحميل نموذج مدرب جاهز للتعرف على الكيانات المسماة (NER).

en_core_wed_sm 1.2.4

هو نموذج مدرب باللغة الانجليزية تم تدريبه على بيانات من ويب، يبلغ حجمه 11 MB وهناك اصدارات أكبر لكنها تحتاج إلى اجهزة ذات ميزات قوية مقارنة بهذا الإصدار الذي يشتغل علي أي جهاز بميزات متوسطة، وهذا النموذج المدرب الذي توفره المكتبة spaCy سيتم استخدامه لتحديد وتصنيف الكيانات المسماة NER.

Named Entity Recognition(NER)3.4

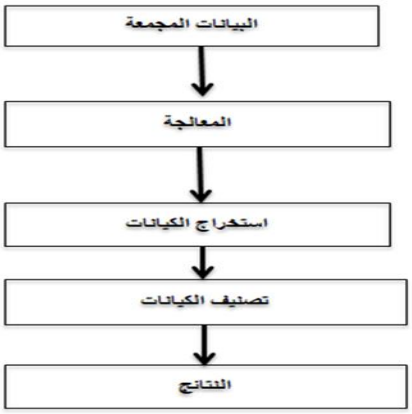
أول ظهور NER في مؤتمر فهم الرسائل (MUC) في التسعينيات (جريشمان، 2014)، كان هذا المؤتمر إجماعًا على أن NER مهمة فرعية مهمة جدًا لاستخراج المعلومات التي تهدف إلى العثور على الاسم و تصنيفه في نص غير منظم.

التعرف على الكيان المسماة (NER) هو مهمة فرعية لاستخراج المعلومات التي تسعى إلى تحديد وتصنيف الكيانات المسماة المذكورة في نص غير منظم إلى فئات محددة مسبقًا مثل الأشخاص والأسماء والمؤسسات والمواقع وتعبيرات الوقت والكميات والقيم النقدية والنسب المئوية وما إلى ذلك، وهو حقل فرعي من الذكاء الاصطناعي، و يعد التعرف على الكيانات المسماة خطوة مهمة في خط

الإنتاج للعديد من تطبيقات معالجة اللغة الطبيعية التي تتضمن استخراج المعلومات، تم تصميم مكتبة spaCy لتوفير حلول جاهزة للإنتاج لمهام معالجة اللغة الطبيعية.

سنستخدم مكتبة spaCy للتعرف على الكيانات المسماة (NER) حيث تتمثل المهمة في تحديد ما يسمى بالكيانات المسماة في محتوى نصي وتصنيف نوع الكيان الخاص بها، يتعرف spaCy على مجموعة من الفئات من الكيانات المحددة بما في ذلك الأشخاص والدول والمنظمات ووكالات الأنباء ومجموعة قومية ودينية وغيرها.

المخطط الانسيابي NER



الشكل (34)يوضح مخطط انسياب عمليات NER

من خلال المخطط السابق الذي يوضح انسياب العمليات في هذا البحث، تم تجميع البيانات من خلال واجهة برمجة التطبيقات ثم إجراء عملية معالجة والتي تم التحدث عنها في الباب الرابع والتي تنقسم إلى قسمين بعضها قام بها الباحث، والبعض الأخر تم القيام بها عن طريق المكتبة spaCy وذلك لاستخراج الكيانات التي تم التحصل عليها فيما بعد، وتصنيفها وفق جدول الاختصارات الخاص بالكيانات المسماة على سبيل المثال شخص، وكالة، منظمة، وسنقوم بعرض النتائج كاملة وشرحها وتفسيرها.

4.4جدول اختصارات NER

حيث توجد العديد من الاختصارات في NER والتي تشير كلا منها إلى كيان محدد بحيث تسهل أو تساعد هذه الاختصارات في التعرف على الكيانات و تصنيفها حسب الاشخاص والتواريخ والدول والمواقع وغيرها، والجدول (18) يوضح اختصارات NER كل نوع ومعناه:

Type Named Entity	Description
DATE	التواريخ أو الفترات المطلقة أو النسبية
PERSON	أسماء الأشخاص
GPE	البلدان والمدن والدول
LOC	المواقع التي لا تتبع GPU مثل:- السلاسل الجبال ،المسطحات المائية.
MONEY	القيم النقدية ، بما في ذلك الوحدة
NORP	الجماعات الدينية أو القومية
CARDINAL	الأعداد التي لا تندرج تحت نوع آخر
EVENT	أعاصير ، معارك ، حروب ، أحداث رياضية
FAC	المباني والمطارات والطرق السريعة والجسور
ORDINAL	"الأول"، "الثاني، إلخ.
PERCENT	النسبة المئوية ،بما في ذلك "%"
PRODUCT	الأشياء والمركبات والأطعمة وما إلى ذلك (ليست خدمات)
WORK_OF_ART	عناوين الكتب والأغاني وما إلى ذلك.
LANGUAGE	أي لغة مسماة.
LAW	مايتعلق بالقوانين
QUANTITY	القياسات من حيث الوزن أو المسافة.
TIME	الوقت
ORG	الوكالات والمؤسسات

جدول (18) يوضح اختصارات NER ومعناها

في الجدول السابق تم ذكر اختصارات NER لكل كيان، على سبيل المثال عند تطبيق التقنية على قطعة نصية تحتوي على جمل عند وجود كلمة تعبر عن التاريخ والفترات النسبية يتم تحديدها وتصنيفها على أنها DATE، وعند التعرف على كلمة تعني لغة معينة يتم تصنيفها على أنها LANGUAGE.

الفصل الخامس

المعالجة وتنظيف البيانات

1.5 المقدمة:

في هذا الفصل سنقوم بتوضيح عملية المعالجة التي قمنا بها على البيانات المجمعة، كان لدينا مجموعة من البيانات حوالي 5700 تغريدة، تحتوي التغريدات على بيانات حول الهجرة غير الشرعية في ليبيا التي تم جمعها في قاعدة البيانات، هذه التغريدات الخام دون معالجة مسبقة غير منظمة إلى حد كبير وتحتوي على معلومات زائدة عن الحاجة للتغلب على هذه المشكلات تتم المعالجة المسبقة للتغريدات من خلال اتخاذ خطوات متعددة.

2.5 معالجة البيانات المجمعة:

غالبا النصوص في مواقع التواصل الاجتماعي تحتاج إلى معالجة، يمكن أن تحتوي التغريدات على الكثير من الأشياء من النص العادي والإشارات وعلامات التصنيف والروابط وعلامات الترقيم والعديد من الأشياء الأخرى، عندما تعمل في مشروع علم البيانات أو التعلم الآلي، قد ترغب في إزالة هذه الأشياء للحصول على نتائج أفضل.

1.2.5حذف الروابط

بعد تجميع البيانات من موقع تويتر كانت معظم التغريدات مجمعة بروابط وبالتالي سيتم إزالتها لأنها ليست ضرورية عادةً لعملية لمعالجة النص، لذا من الأفضل إزالتها من النص الخاص بك.

يوضح جدول (11) عملية حذف الروابط من التغريدة:

التغريدة قبل تعديل	التغريدة بعد تعديل
Niger, Chad receive 75,000 refugees from Libya http://tf.to/VN6q	Niger, Chad receive 75,000 refugees from Libya

جدول(11) حذف الروابط من التغريدة

2.2.5 حذف الهاشتاق Hashtags

تحتوي النصوص عادة على هاشتاق (#) يستخدم بشكل كبير في المواقع، قبل الكلمة التي تريد أن تحدث ضجة في مواقع التواصل الاجتماعي هناك حالات تريد إزالته بحيث تحصل فقط على المحتوى النظيف للتغريدة؛ لأنه يؤثر في نتائج الدراسة فيما بعد.

التغريدة بعد تعديل	التغريدة قبل تعديل
Over 2 thousand former jamahiriya officials apply for refugee status UN Libya	Over 2 thousand former jamahiriya officials apply for refugee status UN# Libya

يوضح الجدول (12) عملية حذف وسم الهاشتاق #

3.2.5 حذف العلامة @:

غالبا تحتوي بيانات النصوص في مواقع التواصل الاجتماعي على الكثير من العلامات المختلفة التي تشكل فرق في النتائج المرجوة منها العلامة @، التي تستخدم في كثير من الحالات مثل الإشارة إلى الأشخاص في تغريدة ما أو في كتابة البريد الأكتروني ،وبذلك سيتم حذف هذه العلامة التي لا علاقة لها في الناتج النهائي بحيث يتم الحصول على نص تغريدة نضيف، والجدول (13) يوضح التغريدة قبل بعد عملية حذف العلامة @:

التغريدة بعد التعديل	التغريدة قبل تعديل
BBC News - Illegal immigrants detained in Libya refugees	BBC News - Illegal immigrants detained in Libya @refugees

والجدول (13) حذف العلامة @ من التغريدة

4.2.5 حذف علامات الترقيم:

في جميع اللغات تحتوي النصوص على علامات ترقيم بحيث تسهل أو توضح للقارئ القراءة في النص أو المحتوى الكامل، ولكن في بعض الحالات لا نحتاج إليها على سبيل المثال في عملية التنقيب عن النصوص أو البيانات هنا في حاجة إلى إيجاد أو استخراج معرفة كما في هذه الدراسة وبالتالي سيتم حذف هذه العلامات لأنها غير ذات جدوى في النتائج، والجدول (14) يوضح العملية قبل وبعد حذف علامات الترقيم.

التغريدة بعد التعديل	التغريدة قبل التعديل
Niger Chad receive 75000 refugees from Libya	Niger, Chad receive 75,000 refugees from Libya

جدول (14) حذف علامات الترقيم من التغريدة

5.2.5 حذف الكلمات الشائعة Stop Words:

هي الكلمات المستبعدة أو كلمات الوقف، تحتوي النصوص عادة على كلمات مستبعدة يمكن أن تؤثر في نتائج الدراسة وبالتالي فهي بحاجة إلى معالجة و التي يستحسن إزالتها لتنظيف البيانات النصية للحصول علي نتائج افضل كما أكدت العديد من دراسات في مجال NER، والجدول (15) يوضح التغريدة قبل وبعد عملية المعالجة.

التغريدة قبل التعديل	التغريدة بعد التعديل
Iraqi refugees on the LibyanTunisian border appeal to animal welfare organizationto assist refugees in Libya Arabic	Iraqi refugees LibyanTunisian border appeal animal welfare organizationto assist refugees Libya Arabic

جدول(15) حذف الكلمات الشائعة من التغريدة

6.2.5 تحويل الحروف من كبيرة إلى صغيرة:

بعد تجميع التغريدات باللغة الإنجليزية توجد بعض الكلمات في بعض التغريدات تم كتابتها بأحرف إنجليزية كبيرة (كابيتل) وبالتالي يجب معالجتها وتوحيد الحروف الإنجليزية وجعلها كلها حروف صغيرة (سمول).

الجدول (16) يوضح عملية المعالجة قبل وبعد تحويل الأحرف الإنجليزية من كابيتل إلى سمول.

التغريدة قبل التعديل	التغريدة بعد التعديل
Niger Chad receive 75000 refugees Libya	chad receive 75000 refugees libya niger

الجدول (16) تحويل الحروف من كبيرة الي صغيرة في تغريدة

3.5 تقوم المكتبة spaCy في الخلفية ببعض المعالجات:

توجد العديد من المهام التي تقوم بها المكتبة spaCy، حيث سنقوم بالتطرق إلى المعالجات أو الوظائف التي تعتبر من اساسيات المعالجة في هذا البحث وهي تقسيم النص Tokenization، ووضع علامات على جزء من الكلام POS، سيتم الحديث عنها بشيء من التفصيل فيما يالي:

1.3.5 تقطيع النص Tokenization

يقصد بهذه الخطوة تقطيع الكلمات المكونة للجملة إلى كلمات أبسط، وتتم عملية التقطيع بناءً على الفاصل بين الكلمات المكونة للجملة، سيتم تقطيع كل تغريدة من البيانات المجمعة إلى مجموعة الكلمات المكونة لها.

2.3.5 وضع علامات على جزء من الكلام(Tagging POS)

هو عملية ترميز كلمة في النص لجزء معين من الكلام أي تحديد الكلمات على أنها أسماء وأفعال وصفات وغيرها ، اعتماداً على تعريف الكلمة وسياقها، عملية تصنيف علامات وتمييزها للكلمات Pos Tag التي تسمى أجزاء من علامات الكلام، تُستخدم لفهم السياق الذي تُستخدم فيه الكلمة في الجملة، تعد مفيدة في تحليل جملة واسترجاع المعلومات وتحليل المشاعر وما إلى ذلك، حيث سيتم وضع Tag بجانب كل كلمة في النص مثل الاسم (PROPN) وفعل (VERB)، يوضح الجدول (17) كل Pos Tag في spaCy ومعناها ومثال عليها.

Pos Tag	وصف	مثال
ADJ	adjective	first
ADP	adposition	in, to
ADV	adverb	very, tomorrow
AUX	auxiliary	(has (done
CONJ	conjunction	and, or, but
DET	determiner	a, an, the
INTJ	interjection	hello
NOUN	noun	girl, cat
NUM	numeral	seven,7-one, seventy
PART	particle	ton ،s’
PRON	pronoun	you, he, she
PROPN	proper noun	mohamed, John, London

PUNCT	punctuation	؟ ،(،) ،.
SCONJ	subordinating conjunction	if, while, that
SYM	symbol	$، %، §، ©، +، −، ×، ÷، =، :
VERB	verb	run, runs, running, eat, ate, eating

الجدول (17) يوضح اختصارات posTag ومعناها وأمثلة عليها

4.5الخلاصة:

في هذا الفصل قمنا بشرح مرحلة معالجة البيانات التي تم تجميعها من موقع تويتر، حيث تم تنظيف النص وحذف العالمات والرموز التي ال عالقة لها بالنتائج أو قد تؤثر في الناتج النهائي بما فيها الروابط والهاشتاقات وغيرها من عالمات الترقيم وحذف العالمة @ وحذف الكلمات الشائعة وتحويل الحروف من كبيرة إلى صغيرة.

الفصل السادس

النتائج

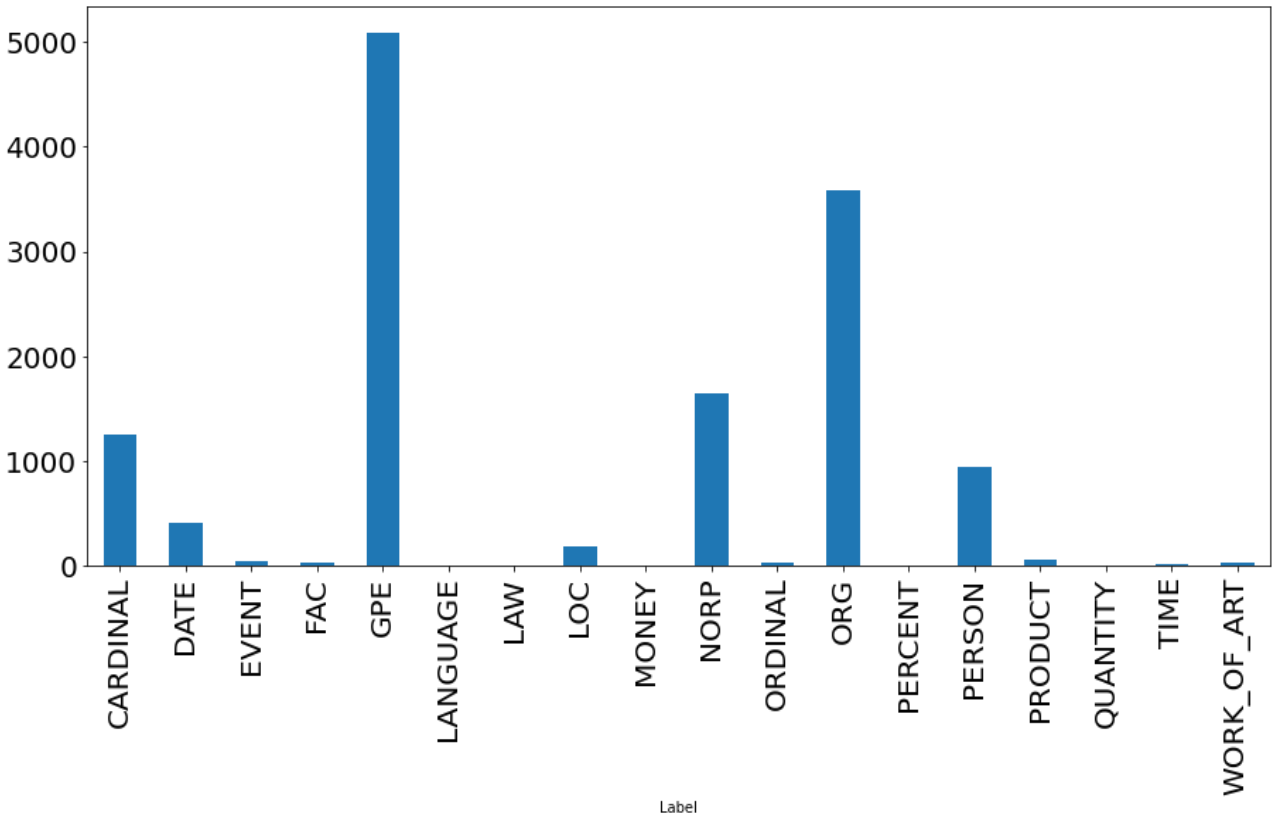
1.6المقدمة :

في هذا الفصل سنقوم بعرض النتائج التي تحصلنا عليها من خلال تطبيق تقنية GS على البيانات التي تم جمعها من موقع التواصل الاجتماعي تويتر عن موضوع الهجرة غير الشرعية في ليبيا والتي سبق لها المعالجة في الباب السابق، وبعد عرض النتائج في حال وجود أي نتائج تحتاج إلى معالجة سنعمل على معالجتها وتفسيرها

2.6النتائج:

بعد أن تم معالجة البيانات الأولية وتنظيف البيانات في الفصل السابق، هنا أصبحت البيانات جاهزة لتطبيق تقنية NER حيث قمنا بإضافة عمودين جديدين وهي Lable يمثل نوع الكيان و Entity الكيان نفسه، وقمنا بتطبيق عروض مرئية للمكتبة Matplotlib لعرض النتائج بشكل أسهل وأوضح، عليه سنقوم ببرمجة البيانات لتحديد الكيانات التي تتحدث عن الهجرة غير الشرعية في ليبيا وتصنيفها، وتوحيد بعض الكلمات التي قد تسبب تكرار في النتائج بما فيها الاختصارات على سبيل المثال: توحيد كلمة libia إلى Libya، وتوحيد الاختصار uk إلي united kingdom، ثم عرض النتائج للكيانات التي تفيدنا وتختص بالدراسة وإهمال باقي الكيانات لأنها ليست ذات أهمية، فيما يلي سيتم عرض كيانات الأشخاص والمنظمات ووكالات الأنباء والدول والمجموعات الدينية والقومية التي تم التحصل عليها من البيانات التي قمنا بجمعها.

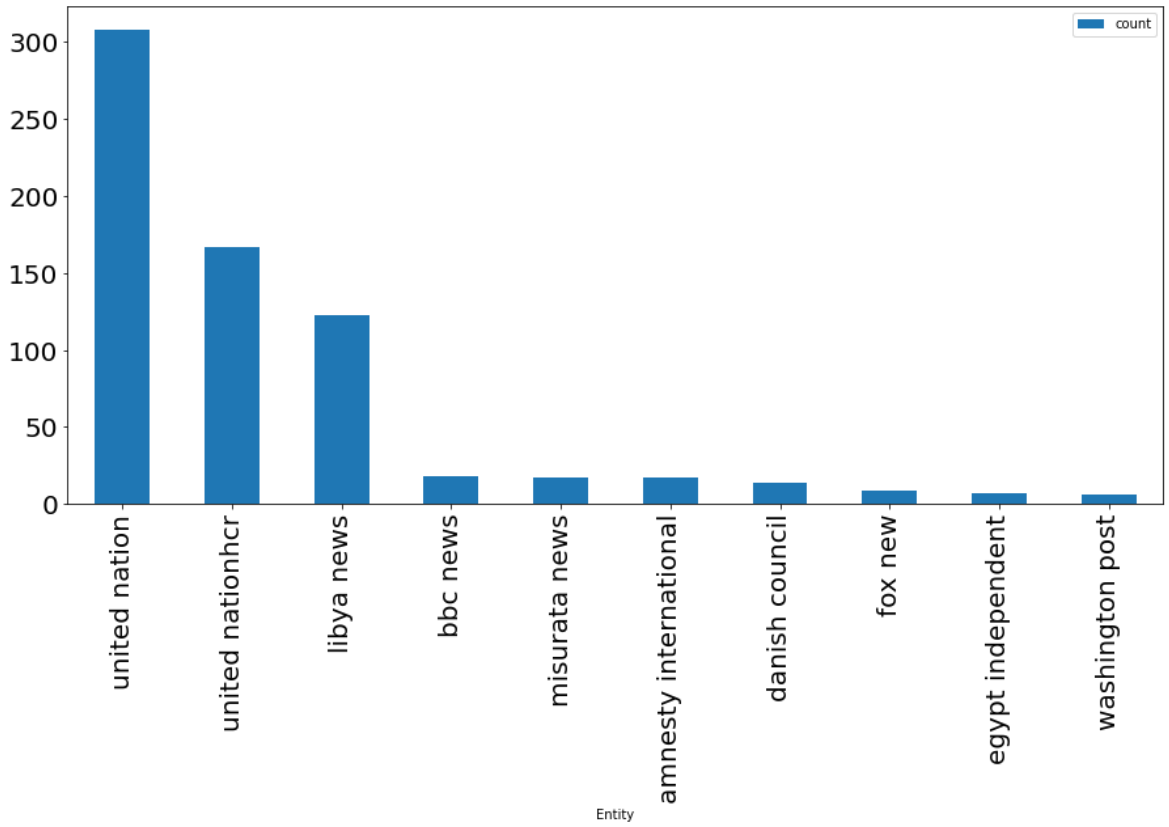
1.2.6 استكشاف الكيانات المسماة الأكثر شيوعا في مجموعة تغريدات لدينا:



يوضح شكل (34) الكيانات المسماة المكتشفة وتكرارها في البيانات

في رسم البياني أعلاه نعرض نتائج أنواع جميع الكيانات التي تم الحصول عليها من التغريدات التي قمنا بجمعها في دراستنا باستخدام المدرج التكراري، من خلال النتائج يمكننا أن نرى أن الأشخاص والأماكن والمنظمات هم أكثر الكيانات المذكورة على الرغم من أن لدينا أيضًا العديد من الكيانات الأخرى، وبالتالي الكيانات التي لن يتم ذكرها بشكل كبير تعتبر غير ذات أهمية في الدراسة، فيما بعد سنقوم بعرض نتائج للكيانات التي أجرينا من أجلها الدراسة وهي الأشخاص والمنظمات ووكالات الأنباء والدول والمجموعات الدينية والقومية التي تتحدث عن الهجرة غير الشرعية في ليبيا والتي تم الحصول عليها من التغريدات التي قمنا بجمعها من موقع التواصل الاجتماعي تويتر في قاعدة بيانات خلال ثلاث سنوات 2012، 2013، 2014، وهذا إجابة السؤال الأول في أسئلة البحث.

2.2.6كيانات المنظمات ووكالات الأنباء الأكثر تكرارا في مجموعة البيانات (ORG):



الشكل (35) يبين وكالات الأنباء والمنظمات وتكرارها في البيانات (ORG)

يوضح الشكل السابق نتائج المنظمات ووكالات الأنباء التي تم استنتاجها من مجموعة التغريدات المجمعة، من خلال الرسم الموضح أعلاه نلاحظ الآتي:

1. بلغ معدل تكرار united nation منظمة الأمم المتحدة حوالي 300 مرة.
2. تلاها تكرار united nationhcr منظمة الأمم المتحدة لشؤون الهجرة بحوالي 150 مرة.
3. نلاحظ بعض وكالات الأنباء مثل libya news وهي الأكثر تكرارا في الوكالات تليها bbc news و misurata news, fox news.
4. في المنظمات المذكورة amnesty international (منظمة العفو الدولية) و danish council (مجلس اللاجئين الدنماركي)، egypt independent، washington post التكرار كان متقاربا تقريبا وهو أقل من 30.

5. من الملاحظ أن المنظمات التي تتبع الأمم المتحدة ذكرت أكثر من غيرها، ولذلك دلالة معينة بأن هذه المنظمات هي أول جهة يلجأ إليها المهاجرين في طلب المساعدات، united nation هى الأكثر تكرارا في البيانات المجمعة.

مما يوضح أنها أكثر جهة مهتمة بموضوع اللاجئين والهجرة غير الشرعية، libya news ذكرت أكثر من وكالات الأنباء الأخرى باعتبارها جهة إعلامية مهتمة بالواقع اليبي وهي جريدة داخل ليبيا وأقرب للأحداث، تلاها في ذلك ذكرت بصفة أقل وكالات الأنباء الأخرى لأنها بنفس الأهمية في نقل الأخبار عن المهاجرين، amnesty international لها علاقة مباشرة بقضايا اللاجئين ولم يتم ذكرها أكثر من غيرها من تفسير الباحث أن اللاجئين لن يلجأو لهذه المنظمة أكثر من غيرها في المنظمات الإنسانية العالمية، جاء ذكر بعض وكالات الأنباء العالمية مثل fox news و washing post بعيد جدا عن موضع الهجرة يعزى الباحث وجود هذه الوكالات إلى تقديم المساعدات أو ذكر الأخبار عن تقديم مساعدات للاجئين في هذه المنطقة الجغرافية، وهذا إجابة السؤال الثاني والثالث في أسئلة البحث.

3.2.6 كيانات الدول والمدن والبلدان الأكثر تكرارا في مجموعة البيانات (GPE):

شكل (36) يبين كيانات الدول وتكرارها في البيانات (GPE)

يوضح الشكل (36) نتائج كيانات الدول التي تم الحصول عليها من خلال التغريدات التي قمنا بجمعها، تم عرضها في سحابة كلمات حيث يدل أختلاف حجم الكلمات على عدد تكرارها، كلما كانت الكلمة حجمها كبير كان عدد تكرارها كبير والعكس، توجد العديد من الدراسات التي صنفت الدول

حسب المصدر والعبور و المستضيفة، من بينها دراسة (محمد إمحمد، 2019) وبالتالي سنقوم بتقسيم كيانات الدول إلى أربعة فئات وهيا:

1. دول المصدر: وهي الدول التي يخرج منها المهاجرين، على سبيل المثال الدول الفقيرة التي أصيبت بالمجاعة في قارة أفريقيا و الدول التي تحدث بها الحروب، هنا يلجأ بعض سكان هذه الدول للهجرة غير الشرعية سواء كانت عن طريق البر أو البحر، فيما يلي الدول التي تعتبر مصب للمهاجرين غير الشرعيين في ليبيا:

سوريا، الصومال، نيجيريا، العراق، مصر، النيجر، تشاد، أفغانستان، اليمن، كما يمكننا اعتبار ليبيا دولة مصدرة بعد الأحداث التي شهدتها 2011.

2. دول العبور: هي الدول التي يمر بها المهاجرين، أي الدول التي توجد في الطريق بين الدولة المصدر والدولة المستضيفة على سبيل المثال السودان تعتبر دولة عبور للمهاجرين غير الشرعيين من دولة المصدر الصومال، فيما يلي دول العبور والتي تعتبر أغلبها دول جوار إلى ليبيا:

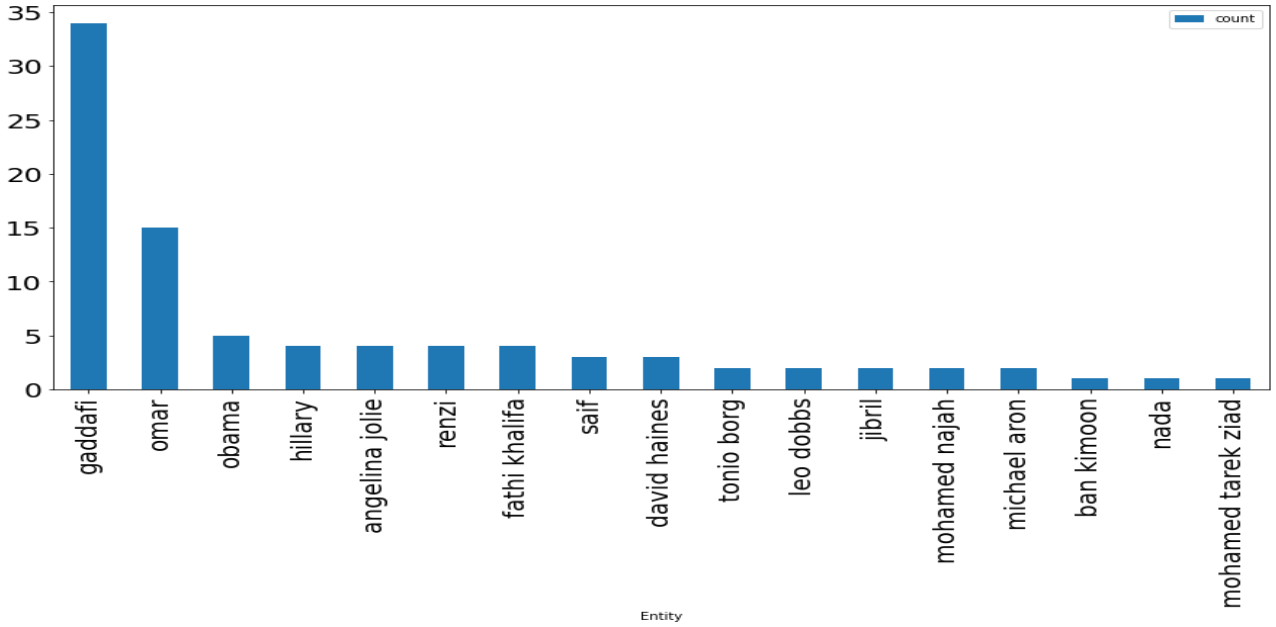
الجزائر، تونس، السودان، وتعتبر ليبيا أحيانا دولة عبور لبعض المهاجرين غير الشرعيين.

3. دول مستضيفة: هي الدول التي يذهب إليها المهاجرين، أي الدول التي تستقبل المهاجرين غير الشرعيين من ليبيا من أشهر هذه الدول إيطاليا والتي تعتبر الوجهة الأساس للمهاجرين ، في ما يلي الدول المستضيفة للمهاجرين غير الشرعيين:

لبنان، تركيا، ألمانيا، إيطاليا، والجدير بالذكر أن ليبيا تعتبر دولة مستضيفة للعديد من المهاجرين من الدول الأفريقية وغيرها.

4. دول أخرى لا تصنف ضمن الفئات الثلاثة السابقة إلا أن تم ذكرها بشكل متكرر في التغريدات على سبيل المثال وجود دولة أمريكا بعد عرض النتائج تم ذكرها في سحابة الكلمات بشكل كبير، حيث قمنا بقراءة التغريدات التي ذكرت فيها لمعرفة العلاقة بينها وبين الهجرة غير الشرعية في ليبيا وتبين لنا أن الولايات المتحدة الأمريكية قامت بتقديم المساعدات للمهاجرين النازحين بعد الأحداث التي شهدتها ليبيا، كما في دولة أستراليا صرح رئيس الوزراء الأسترالي عن غرق قارب لاجئين ليبيين، وأثناء الحروب في ليبيا تم القبض على آلاف اللاجئين الفلسطينيين ووعدت إسرائيل بالسماح ل400 فقط بالعودة، وهذه النتائج إجابة للسؤال الرابع في أسئلة البحث.

4.2.6 كيانات الأشخاص الأكثر تكرارا في مجموعة البيانات (PERSON):



الشكل (37) يبين كيانات الأشخاص وتكرارها في البيانات المجمعة (PERSON)

الشكل (37) السابق يوضح نتائج كيانات الأشخاص PERSON التي تم تحديدها وتصنيفها من البيانات المجمعة، من خلال الرسم نلاحظ ما يلي:

1. أن القذافي تحصل على أكثر عدد تكرار في التغريدات وهو ما يقارب 35.
2. تلاه في التكرار عمر وهو طفل مهاجر من الصومال إلى ليبيا في عمر 13 عاما، عمل في ليبيا حتى اندلاع الصراع في 2011، فر عبر الحدود إلى مخيم شوشة للاجئين في تونس تم إنتاج فيلم وثائقي قصير لحياته.
3. أيضا تم ذكر أوباما (الرئيس الرابع والأربعون للولايات المتحدة الأمريكية) وهيلاري كلينتون (وزيرة الخارجية الأمريكية) و أنجلينا جولي (ممثلة وصانعة أفلام أمريكية) رينزي (رئيس الوزراء الإيطالي) وسيف القذافي وفتحي خليفة (سياسي ليبي) بتكرار لا يتجاوز 5.
4. إضافة إلى ذلك ذكر كلاًّ من ديفيد هينز (ناشط بريطاني) عمل على مساعدة المهاجرين والمعاقين في سوريا وليبيا، تونيو بورج (وزير خارجية مالطا)، محمد فايز جبريل (سفير ليبيا في مصر)، محمد نجاح (صحفي)، مايكل آرون (السفير البريطاني)، بان كيمون (سياسي ودبلوماسي من كوريا الجنوبية وأمين عام الأمم المتحدة سابقًا)، ندى (مصورة)، محمد طارق (ناشط سياسي) تم ذكرهم بمعدل تكرار ضئيل جدا، وهذه النتائج إجابة للسؤال الخامس في أسئلة البحث.

5.2.6 كيانات المجموعات القومية والدينية والجنسية التي تم التحصل عليها من البيانات المجمعة (NORP)

شكل (38) يوضح كيانات المجموعات السياسية الدينية والقومية وتكرارها في البيانات (NORP)

يوضح الشكل (38) سحابة كلمات المجموعات القومية والدينية والجنسيات الأكثر تكرارا في التغريدات، من خلال الرسم السابق نلاحظ ما يلي:

1. وجود بعض المجموعات الدينية ذكرت في البيانات بشكل متكرر مثل: المسلمين (muslims) واليهود (jews).
2. كما هناك تواجد لبعض المجموعات القومية تم ذكرها في التغريدات مثل: العرب (arab)، والأفارقة (african)، ساهلين (sehalian).
3. وتم ذكر العديد من الجنسيات في التغريدات المجمعة، على سبيل المثال:

mchadian, afghan, eritrean, german, tunisian, syrian, italian, egyptian, malian، libyan, plasetinian, somali, swiss.

4. يبين الشكل العديد من الجنسيات من عدة دول، حيث كانت الجنسيات التالية الأكثر ذكر في البيانات المجمعة: الليبية، التونسية، الإيطالية، السورية، الفلسطينية، التشادية، الصومالية، من الملاحظ أيضا وجود دول هذه الجنسيات في نتائج الكيانات الخاصة بالدول والتي تم تصنيفها إلى دول مصدر وعبور ومستضيفه يفسر ذلك أن هذه الجنسيات للأشخاص المهاجرين الذين تم تصنيفهم سابقا، مع وجود مجموعات دينية قلة وأخرى قومية بنفس الأهمية، وهذه الكيانات إجابة للسؤال السادس في أسئلة البحث.

الفصل السابع

الخاتمة والآفاق المستقبلية

1.7 الخاتمة:

تم في هذا البحث تطبيق تقنية تحديد الكيانات المسماة على بيانات مجمعة من موقع التواصل الاجتماعي تويتر لثلاث سنوات (2012، 2013، 2014) تخص الهجرة غير الشرعية في ليبيا، حيث قمنا بتجميع 4742 تغريده وتم إجراء عمليات معالجة عليها بهدف الحصول على الكيانات المسماة المكونة من 18 نوع، تم التركيز على كلا من: الأشخاص، ودول، والمجموعات القومية والسياسية والدينية، والمنظمات ووكالات الأنباء، وإهمال باقي الكيانات لعدم أهميتها في الدراسة.

من النتائج المثيرة للاهتمام أننا وجدنا ارتباط بين الهجرة غير الشرعية من وإلى ليبيا ببعض المجموعات العرقية مثل السوريين واليهود والأفارقة والتونسيين وغيرها، وكذلك وجدنا ارتباط لشخصيات معينة مثل الطفل عمر الذي كان له الأثر في موضوع الهجرة على ليبيا في عام 2011 حيث قام بعض المنتجين بإنتاج فيلم وثائقي يبين معاناة الطفل عمر، وتمكنا من إيجاد ارتباط عدد من الشخصيات كممثلين مثل انجلينا جولي وبعض الشخصيات السياسية مثل أوباما وهيلاري والقذافي وبان كيمون، كما تحصلنا كذلك على دول عدة كانت بعضها دول المصدر ولبعض دول عبور والبعض دول مستضيفة، ومن الجدير بالذكر أنه توجد بعض الدول التي وجدنا لها ارتباط مع ليبيا والهجرة غير الشرعية بشكل غير مباشر مثل أمريكا حيث كان وجودها في هذه الدول بتقديم مساعدات فقط لا غير.

برهنا في هذا البحث أن موضوع الهجرة غير الشرعية من وإلى ليبيا موضوع مهم و بالإمكان استخلاص نتائج مهمة تنفع الأمن الليبي، أيضا بالإمكان الاستفادة من تقنية التعرف على الكيانات المسماة من زاويا مختلفة من خلال تحليل منشورات المستخدمين في مواقع التواصل الاجتماعي الأخرى، من حيث الكيانات التي تم استخلاصها ثم إعطاء صورة غنية عن موضوع الهجرة التي قد تكن قاصرة لأنها لم تغطي كافة السنوات ولكنها أعطت معلومات مخفية لموضوع الهجرة غير الشرعية.

2.7 الصعوبات و العراقيل:

لا شك أن أي عمل يواجه العديد من الصعوبات والعراقيل ولعل أبرز هذه الصعوبات والعراقيل التي واجهتنا في الدراسة ما يلي:

1. قلة الدراسات إلى تبحث في موضوع الهجرة غير الشرعية في ليبيا باستخدام تقنيات تنقيب البيانات.
2. لم يتضح لنا وجود دراسة تربط بين الهجرة غير الشرعية و NER.
3. صعوبة عرض النتائج النصية وتفسيرها على العكس عندما تكون النتائج رقمية.
4. وجود ندرة في الدراسات التي تقوم بعرض كامل للنتائج NER، أغلب الدراسات تعرض مثال على الكيانات.

3.7 التوصيات والآفاق المستقبلية:

بناء على هذه الدراسة، يوصى الباحث بما يلي:

1. جمع التغريدات باللغة العربية أو باللهجة الليبية.
2. زيادة النطاق الزمني للبحث لسنوات ما بين 2015 – 2022.
3. استخدام الفيس بوك كمصدر للبيانات.
4. استخدام مكتبات أخرى لتطبيق تقنية NER مثل: NLTK.
5. التنوع في المواضيع لاستخراج معلومات مفيدة منها باستخدام تقنية NER.
6. الاستفادة من الوظائف المتعددة الأخرى لمكتبة spaCy.

References المراجع

أبو زيد، & محمد إمحمد محمد. (2019). الهجرة غير الشرعية وأثرها على الأمن القومي الليبي (2017−2011)، جامعة الشرق الاوسط.

مروة عبد الرحيم عويدات عبدالرحيم. (2018). ظاهرة الهجرة غير الشرعية وسبل مكافحتها (ليبيا نموذج).

نصر. (2020). الاتجاهات الحديثة في بحوث استخدامات مواقع التواصل الاجتماعي وتأثيراتها النفسية والاجتماعية. المجلة العربية لبحوث الإعلام والاتصال، 2020(30)، 4−105.

Qiu, Q., Xie, Z., Wu, L., & Tao, L. (2019). GNER: A generative model for geological named entity recognition without labeled data using deep learning. Earth and Space .science, 6(6), 931-946

Yepes, A. J., MacKinlay, A .,&Han, B. (2015, July). Investigating public health surveillance using twitter. In Proceedings of BioNLP 15 (pp. 164-170).

Karatay, D., & Karagoz, P. (2015, January). User Interest Modeling in Twitter with Named Entity Recognition. In # MSM (pp. 17-20).

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. Computer Science Review, 29, 21-43.

Ritter, A., & Clark, M. (2011). Sam, and Oren Etzioni. Named entity recognition in tweets: an experimental study. In Empirical Methods in Natural Language Processing.

Morwal, S., Jahan, N., & Chopra, D. (2012). Named entity recognition using hidden Markov model (HMM). International Journal on Natural Language Computing (IJNLC) Vol, 1.

Jung, J. J. (2011, July). Towards named entity recognition method for microtexts in online social networks: a case study of Twitter. In 2011 International Conference on Advances in Social Networks Analysis and Mining (pp. 563-564). IEEE.

Inkpen, D., Liu, J., Farzindar, A., Kazemi, F .,&Ghazi, D. (2017). Location detection and disambiguation from twitter messages. Journal of Intelligent Information Systems, 49(2), 237-253

Altarrazi, S. M., & Sasi, S. (2016, March). Tweeple's microblogs on illegal immigration in USA. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 2011-2018). IEEE.

Nulty, P .,&Poletti, M. (2014). The Immigration Issue in the European Electoral Campaign .in the UK: Text-Mining Public Debate from Newspapers and Social Media

.

Yanti, R. M., Santoso, I., & Suadaa, L. H. (2021). Application of Named Entity Recognition via Twitter on SpaCy in Indonesian (Case Study: Power Failure in the Special Region of Yogyakarta). Indonesian Journal of Information Systems, 4(1), 76-86.

Pearson, C., Seliya, N., & Dave, R. (2021, December). Named entity recognition in unstructured medical text documents. In 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET) (pp. 1-6). IEEE.

Suresh, Y., & Manusha Reddy, A. (2021). A contextual model for information extraction in resume analytics using NLP's spacy. In Inventive Computation and Information Technologies (pp. 395-404). Springer, Singapore.

Vasiliev, Y. (2020). Natural Language Processing with Python and SpaCy: A Practical Introduction. No Starch Press.

Partalidou, E., Spyromitros-Xioufis, E., Doropoulos, S., Vologiannidis, S., & Diamantaras, K. I. (2019, October). Design and implementation of an open source Greek POS Tagger and Entity Recognizer using spaCy. In 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 337-341). IEEE.

Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1), 50-70.