بسم الله الرحمن الرحيم

جـــامعة الســـودان للعلـــوم والـــتكنولـــوجيا

كليـــة علـــوم الحاسـوب وتقنيـــة المعلومـات

بحـث تكميلـي لنيـل درجـة المـاجستير

العنوان:

تحليل تغريـدات تـويتر بـإستخدام خوارزميات العنقدة لمعرفة أسباب إنتشار المخدرات

Analysis of Twitter's Tweets Using clustering algorithms to recognize the causes of drug proliferation

إعداد:

ماجدة عبد السلام محمد سعيد العوض

إشراف:

د. طلعت محي الدين وهبي

بسم الله الرحمن الرحيم

((رَبِّ أَوْزِعْنِي أَنْ أَشْكُرَ نِعْمَتَكَ الَّتِي أَنْعَمْتَ عَلَيَّ وَعَلَى وَالِدَيَّ وَأَنْ أَعْمَلَ صَالِحاً تَرْضَاهُ وَأَدْخِلْنِي بِرَحْمَتِكَ فِي عِبَادِكَ الصَّالِحِينَ))

صدق الله العظيم

سورة النمل : 19

شكر وتقدير

شكرا لكل من قدّم لي المساعدة عند حاجتي ولكل من وقف بجانبي وأخص بالشكر أمي وأبي علي دعمهما لي في كل مراحل حياتي .

كما أشكر الدكتور هشام علي حسن تعامله وسعة صدره وتوجيهاته وعلي كل جهد بذله من أجلي.

شكر إلي صديقتي أفراح طه التي كانت بجانبي طوال فترة البحث دون ملل وكانت خير عون لي .

الملخص:

تعتبر وسائل التواصل الاجتماعي من اهم مصادر المعلومات حيث تسمح بمشاركة المعلومات بين مجموعات مختلفه من الناس حول العالم وتمكن المستخدم من الوصول الي هذه البيانات و البحث عن اراء الناس حول موضوع او منتج معين دون جهد يذكر مقارنه مع الطرق التقليديه مع ضمان موثوقيه أعلي للنتائج .

في هذا البحث تم استرجاع البيانات من تويتر حول الاستخدام السئ للمخدرات وتم تخزينها في قاعدة بيانات وتحليلها باستخدام كل من خوارزمية الk-means و خوارزمية ward's لمعرفة ماهي أسباب الاستخدام السئ للمخدرات.

بعد تحليل 4000 تغريدة جمعت من تويتر عن إستخدام المخدرات وبحسب آراء مستخدمين تويتر أن 70% من أسباب إنتشار المخدرات البيئة المحيطة والعلاقات تليها أسباب نفسية ثم قلة الوعي .

Abstract

The massive developments of social media sites in the recent years have become an important source of information. Also, it allowed a great opportunity to communicate and share information between people around the world. And you can take advantage of the large quantity of data that are published daily to find out people's views about a particular product, service or personality.

The problem is that the traditional ways to find out public opinion is not effective because it includes a limited number or category of persons in addition to that there is no guarantee their credibility in answer therefore been used clustering algorithms in artificial intelligence.

This project is aimed to retrieving tweets from the social networking site (Twitter) and stored in the database to extract data to know what causes the spread of drugs abuse.

After analysis of 4000 tweets collected from Twitter about the use of drugs and according to the views of Twitter users that 70% of the causes of the spread of drugs surrounding environment and relationships followed by psychological reasons and then lack of awareness.

شرح الاختصارات التي وردت في البحث:

شرح المصطلح	المصطلح	الاختصار
فكرة (API) هي أن تَجلِب البيانات والخدمات بتنسيقات تسمح لنا بإستخدامها مرة أخرى من أي مكان ، وأشهر هذه التنسيقات هي JSON و XML .	Application Program Interface	API
بروتوكول مفتوح يُمكّن من عمل تحقق مُؤمَن بطريقة بسيطة وقياسية	Open Authorization	OAuth
عبارة عن صيغة متسلسلة لنقل البيانات	JavaScript Object Notation	JSON
لغة PHP هي واحدة من أشهر لغات البرمجة التي يتم إستخدامها في إنشاء مواقع الويب و هي من اللغات التي يقوم خادم الويب بتفسير و تنفيذ الكود الخاص بها ثم يرسل النتيجة ليتم عرضها في متصفح المستخدم	Hypertext Preprocessor	PHP
رمز يتم استخدامه من قبل تطبيق تويتر للوصول الى الموارد المحمية نيابة عن المستخدم	Access Token	AT

الفهرس

الفصل الأول

المقدمة

الفصل الأول

المقدمة

التطور الكبير الذي حدث في وسائل التواصل الاجتماعي يعتبر ثروة في عالم البيانات ، فهنالك تغريدات تضاف يوميا من مختلف المستخدمين من أماكن مختلفه حول العالم .

يقوم المستخدمون بالتعبير عن كل مايجول في خواطرهم عن الطريق الكتابة والتصريح تجاه القضايا التي يتناولونها في مواقع الشبكات الاجتماعية. ومن هنا جاء تحليل الآراء كأحد الإجراءات التي يستخدمها الباحثون في مجال الإتصال الرقمي لقراءة مواقف وإتجاهات الرأي العام على تلك المنصه أي أنه إستراتيجية إتصالية للوصول الى مايتداوله المستخدمون على شبكات التواصل.

ومن أشهر شبكات التواصل الأجتماعي التويتر لذلك يعتبر تويتر مصدر غني بالمعلومات لإتخاذ القرارات وتحليل الآراء(Hridoy, Syed Akib Anwar, 2015).

في هذا البحث تم اختيار تويتر كمصدر للبيانات لأنه يحتوي علي عدد كبير من المستخدمين ولتحديده لعدد الاحرف المستخدمه في التغريدة ب140 حرف فقط .

جمعت البيانات من تويتر وبعد تخزينها في قاعدة بيانات بإستخدام MySQL،حللت بإستخدام خوارزميتي k-means و ward's لمعرفة اسباب إنتشار المخدرات .

1.1 مشكلة البحث:

الطرق التقليدية للبحث عن البيانات تتطلب جهدا كبيرا وتستغرق وقتا أطول لجمع المعلومات مقارنة مع إستخدام وسائل التواصل الإجتماعي كمصدر للمعلومات ، كما أن الإستخدام السئ للمخدرات يشكل خطراً علي مستخدمها حيث تضر بصحة كل من عقله وجسده ؛ فمستخدمها يصبح هو نفسه خطر علي مجتمعه.

1.2 اهداف البحث:

إستخدام بيانات (تغريدات) من موقع التواصل الإجتماعي (تويتر) لمعرفة أسباب الإستخدام الخاطئ للمخدرات مما يساعد في محاربتها وتقليل استخدامها.

1.3 حدود البحث:

في هذا البحث جمعت البيانات (التغريدات) حول استخدام المخدرات من الفتره بين 2018-5-12 إلي 2018 -10-15 وتحليلها بإستخدام خوارزميتي k-means وward's.

1.4 منهجية البحث:

جمعت الغريدات من موقع التواصل الإجتماعي تويتر وتخزينها في قاعدة بيانات بإستخدام MYSQLومن ثم معالجتها بحذف الكلمات الأقل تكراراً والأحرف الزائده وإرجاع الكلمات إلي جذورها لتسهيل عملية التحليل وبعد ذلك ،تم تحليلها بايجاد اكثر الكلمات تكرارا

ومن بعدها تم استخدام خوارزميتي k-means و ward's لمعرفة أسباب إنتشار المخدرات.

1.5 هيكل البحث:

يحتوي هذا البحث علي 6 فصول ،حيث يتناول الفصل الأول مقدمة عن مشكلة و أهداف ومنهجية المشروع الذي سنتطرق له فى البحث ونبذة بسيطة عما سيتم تناوله لاحقا في البحث. ويحتوي الفصل الثاني علي المقدمه ، تجميع البيانات و الدراسات السابقة.

يتناول الفصل الثالث منهجية البحث والأدوات التي تم إستخدامها ،كما يوضح كيفية الحصول على البيانات من التويتر وتخزينها في قاعدة البيانات .

أما الفصل الرابع فيتناول خوارزميات التجميع والتصنيف في الذكاء الإصطناعي.

و الفصل الخامس والأخير يحتوي علي النتائج والتوصيات

الفصل الثاني

الخلفية النظرية

الفصل الثاني

الخلفية النظرية

2.1 مقدمة:

هذا الجزء من البحث يحتوي علي خلفيه نظريه عن البحث ، كما يحتوي على دراسات سابقة لها علاقه بهذا البحث .

2.2 تعدين الآراء (Opinion mining):

يعرف أيضاً بإسم تحليل المشاعر (Sentiment analysis) وهدفه تحليل البيانات او النصوص بعد معالجتها للحصول علي معلومات بغرض معرفة ما يحمله النص من آراء سواء إيجابية أو سلبية أو محايدة تجاه موضوع النص (Hridoy, Syed Akib Anwar, 2015) .

2.3 الشبكات الإجتماعية (Social network):

الشبكة الإجتماعية هي موقع على شبكة الإنترنت يجمع الناس معاً للحديث وتبادل الأفكار والمصالح أو تكوين صداقات جديدة (Alessa,2018) وفيما يلي قائمة صغيرة من بعض أكبر الشبكات الإجتماعية المستخدمة اليوم :

- **موقع Facebook :** موقع الشبكات الإجتماعية الأكثر شعبية على شبكة الإنترنت . الفيسبوك هو مقصد للمستخدمين لإعداد صفحات الويب الخاصة بهم، والتواصل مع الأصدقاء، وتبادل الصور، والحديث عن ما يفعلونه ، الخ..

(http://www.facebook.com)

- **MySpace** : الموقع الذي كان الأكثر رواجا ً قبل أن يدخل في منافسة شديدة مع الفيس بوك مؤخرا ً، وهو تطبيق يقدم شبكة تفاعلية بين الأصدقاء المسجّلين في التطبيق، ويمكّن المستخدمين من نشر الصور، وكتابة المدونات، ونشر الموسيقى ومقاطع الفيديو، وإرسال الرسائل (http://www.myspace.com) .
- **تويتر (Twitter):** أحد الشبكات الأسرع نمواً حيث يقوم بتقديم خدمة التدوين المصغر برسالة واحدة لا تتجاوز المائة و الأربعون حرف، المعروفة بإسم التغريدات، كما يمكنك متابعة الأشخاص الذين تعرفهم أو الذين كنت مهتما بهم، وتبادل الرسائل النصية (http://www.Twitter.com).

2.4 واجهة برمجة التطبيقات لتويتر(Twitter API)

هي طريقة للوصول الى واجهات برمجية للتطبيقات لتَستطيع قراءة وكتابة البيانات من حسابك في تويتر، والتعرف على صاحب الحساب ومعلومات عن المتابعين . توىتر لدىه ثلاثه واجهات برمجىه للتطبيقات هي :

- Search API
- REST API
- Streaming API

2.4.1 Search API

يَسمح للمستخدمىن البحث في محتوى توىتر وإسترجاع التغريدات وفقاً لِشروط معينة مثل إسترجاع التغريدات بلغة معينة أو من بلد معين أو تحديد عدد معين .

2.4.2 REST API (Representational State Transfe)

يتكون الRESTمن مبادئ وتوجيهات قابلة للتطوير والمستخدمة لإنشاء خدمات الويب ، وهو مَبني بطريقة تُوافق مبدأ خدمات الويب (Web Services) . فهو يوفر وسيلة لنقل المعلومات بين ال (client) و (server) عن طريق بروتوكلات ال(HTTP).

يمكن المبرمجين من الوصول إلى بيانات موقع التواصل الإجتماعي تويتر(Twitter) لقراءتها مع إمكانية كتابة البيانات. أيضاً يُوفر قراءة الملف الشخصي لمنشئ التغريده المعينه و متابعين البيانات وأكثر من ذلك.

2.4.3 Streaming API

يمكن المستخدمين من البحث عن تغريدات حديثة فقط واسترجاعها .

2.5 العنقدة:

هي عملية تقسيم البيانات الي مجموعات ،حيث تحتوي كل مجموعة علي مجموعة من العناصر متشابهه فيما بينها وتختلف عن العناصر في المجموعات الأخرى ؛ وذلك لتسهيل عملية التعامل مع البيانات وتحليلها (S.-H. Liao,2012)

2.6 طرق تجميع البيانات:

توجد عدة طرق لتجميع البيانات وهي :

- طريقة التقسيم (Partitioning Method)
- الطريقة الهرمية (Hierarchical Method)
- الأسلوب القائم على الكثافة(Density-based Method)

- الأسلوب القائم على الشبكة(Grid-Based Method)
- الأسلوب القائم على نموذج (Model-Based Method)
- الأسلوب القائم على القيد(Constraint-based Method)

2.6.1 طريقة التقسيم(Partitioning Method)

يتم تقسيم البيانات الي عدد من المجموعات بحيث تحتوي كل مجموعة كائن واحد على الأقل وكل كائن يجب أن ينتمي إلى مجموعة واحدة بالضبط(S.-H. Liao,2012)

2.6.2 الطريقة الهرمية(Hierarchical Method)

هذا الأسلوب يقوم بإنشاء التحليل الهرمي لمجموعة معينة من البيانات. وتوجد طريقتان لتقسيم البيانات بالطريقة الهرمية:

i. الطريقة التجميعية (Agglomerative Approach)

ويعرف هذا النهج أيضا بإسم نهج من أسفل إلى أعلى. يتم البدأ مع نقطة من المجموعات الفردية، وفي كل خطوة يتم دمج أقرب زوج من المجموعات (Han and Micheline Kamber,2010).

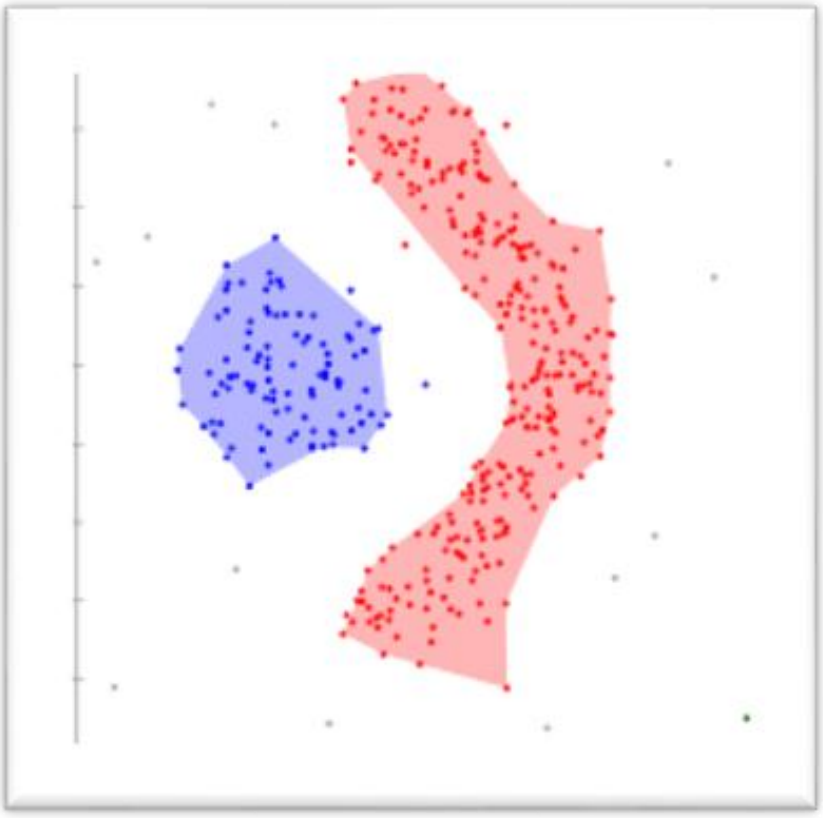
ii. طريقة التقسيم (Divisive Approach)

ويعرف هذا النهج أيضا بإسم نهج من أعلى إلى أسفل . يتم البدء مع كافة الكائنات في نفس المجموعة . يتم تقسيم المجموعة الى أن تصل إلى مجموعات أصغر (Han and Micheline Kamber,2010)

الشكل (2.2) رسم توضيحي يمثل الطريقة الهرميه

2.6.3 الأسلوب القائم على الكثافة(Density-baseMethod)

ويستند هذا الأسلوب على فكرة الكثافة. حيث يستمر النمو في مجموعة بعينها طالما أن الكثافة لم تتجاوز حد معين، أي لكل نقطة من نقاط البيانات داخل مجموعة معينة، في دائرة نصف قطرها من المجموعة المعينة يجب أن تحتوي على ما لا يقل عن الحد الأدنى من النقاط(Han and Micheline Kamber,2010)



الشكل(2.2) رسم توضيحي يمثل يمثل الاسلوب القائم على الكثافة لتجميع البيانات

2.6.4 الأسلوب القائم على الشبكة(Grid-Based Method)

يتم تحديد مجموعة من شبكة الخلايا وتعيين البيانات (الكائنات) إلى خلية الشبكة المناسبة وحساب الكثافة من كل خلية. القضاء على الخلايا، ذات الكثافة الأقل في المجموعة معينة. تشكيل المجموعات من جماعات متجاورة (المجاورة) من خلايا كثيفة (Han and Micheline Kamber,2010).

2.6.5 الأسلوب القائم على نموذج(Model-Based Method)

في هذه الطريقة يتم إفتراض نموذجاً لكل مجموعة للعثور على أفضل تناسب للبيانات لنموذج معين. هذه الطريقة تضع المجموعات عن طريق تجميع دالة الكثافة. وهو يعكس التوزيع المكاني للنقاط للبيانات.

يوفر هذا الأسلوب أيضاً وسيلة للتحديد التلقائي لعدد من المجموعات استناداً إلى الإحصاءات القياسية، بأخذ الضوضاء في الاعتبار. وبالتالي فإنه ينتج أساليب تجميع قوية(Han and Micheline Kamber,2010) .

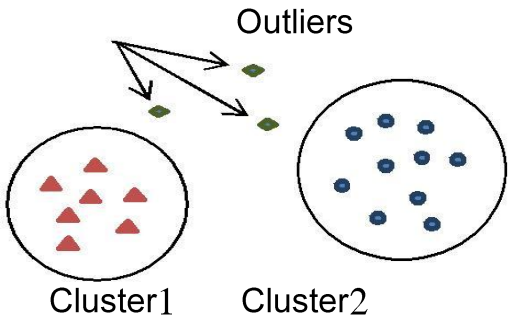
2.6.6 الأسلوب القائم على القيد(Constraint-based Method)

في هذه الطريقة يتم تجميع البيانات عن طريق قيود تحدد من قبل المستخدم (Han and Micheline Kamber,2010)

2.7 خوارزميات العنقدة(Clustering algorithm)

العنقدة أو التجميع هي عملية تجميع للعناصر المتشابهة على شكل عناقيد Clusters حيث يتم تجميع جميع العناصر المتشابهة ضمن عنقود واحد له خصائص معينة يختلف من خلالها عن

باقي العناقيد الأخرى وقد يكون هناك عناصر شاذة لا يمكن أن تنتمي إلى اي تجمع (S.-H. Liao,2012). يوضح الشكل التالي عملية العنقدة أو التجميع:



الشكل (3.2) رسم توضيحي يمثل العنقدة أو التجميع

هناك عدد من الخوارزميات المستخدمة في عملية تجميع البيانات أو تقسيم البيانات، ومن هذه الخوارزميات التي سوف يتم الحديث عنها بشكل مفصل:

- خوارزمية **(K-Medoids Clustering)**
- خوارزمية **(K-means Clustering)**
- خوارزمية **CLARA**

2.7.1 خوارزمية (K-Medoids Clustering)

تستخدم هذة الخوارزمية لعنقدة أو تجميع بعض البيانات اعتماداً على خصائصها إلى K عنقود بإيجاد الوسيط لجميع العناصر، حيث تتم عملية العنقدة من خلال جمع العناصر المتشابهة حول مركز العنقود(الوسيط) (Han and Micheline Kamber,2010).

2.7.2 خوارزمية الـ K-Means:

تقوم هذه الخوارزمية بتجميع البيانات حول مراكز يتم حسابها عن طريق ايجاد الوسط الحسابي لكل العناصر ،حيث أن العنصر الوسط هو المركز الذي تتجمع حوله بقية العناصر المشابهه له (Han and Micheline Kamber,2010).

2.7.3 خوارزمية CLARA

في هذه الخوارزمية يتم أخذ عينة تعبر عن البيانات و من ثم تطبيق خوارزميتي (K-means, K-mediods) لإيجاد المراكز وبعدها يتم تنصيف البيانات إلي عناقيدها المناسبة. تستخدم هذه الخوارزمية في حالة تصنيف حجم كبير جداً من البيانات وتعتمد فعاليتها على حجم العينة التي يتم اخذها من البيانات (Han and Micheline Kamber,2010).

2.7.4 خوارزمية ward's:

خوارزمية ward's إحدى طرق التجميع الهرمي ، حيث تبدأ بمجموعة من العناقيد كل واحد منها يحتوي علي عنصر واحد فقط ،ثم تقوم بدمج العناقيد المجاورة المتشابهة وتستمر بالدمج حتي تصل إلي عدد العناقيد المحدد مسبقاً او أن تصبح جميع العناقيد المتشابهة في نفس المجموعة(Han and Micheline Kamber,2010) .

2.8 الدراسات السابقة :

تعدين النصوص هي عملية القيام بتحليل مجموعة كبيرة من النصوص يصعب معالجتها بالطرق التقليدية ،ولتسهيل عملية التحليل يتم إستخدام خوارزميات مختلفة لإستخراج معلومات مفيدة(K. Sumathy,2013).

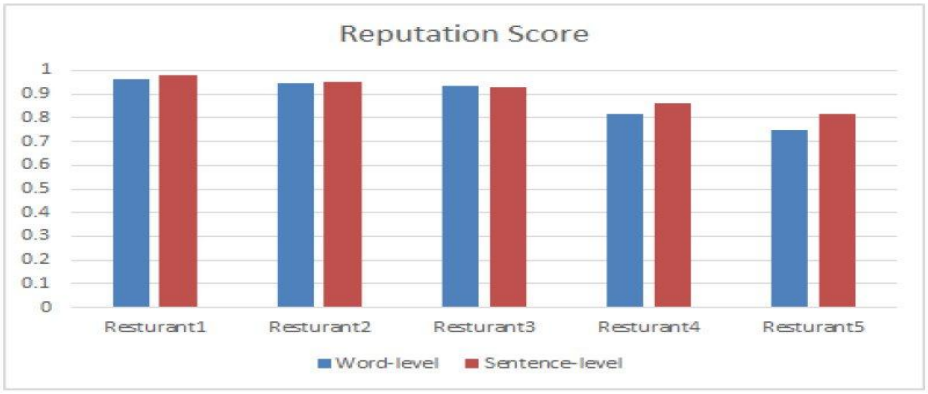
هنالك مجموعة من الدراسات التي إستخدمت تحليل النصوص في مجالات مختلفة منها :

2.8.1 إستخدام تويتر لتقييم المطاعم في السعودية:

Provider Reputation Service A Lexicon-based Approach to Build from Arabic Tweets in Twitter:

هدف هذه الدراسة هو قياس السمعة من خلال تغريدات من تويتر وتم اخذ مجموعة من المطاعم كدراسة حالة (case study) لقياس سمعتهم عن طريق تحليل التغريدات المكتوبة باللهجة السعودية .

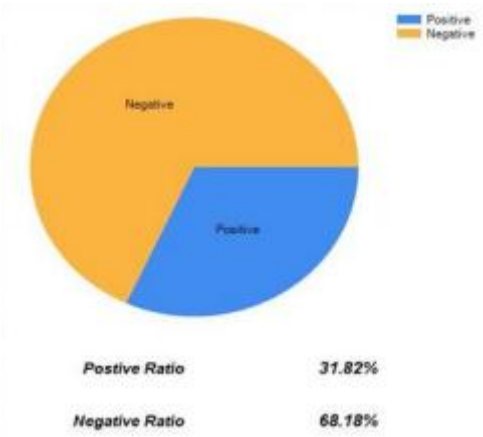
تم إنشاء قاموس بسيط لغرض الدراسة يحتوي علي 1424 كلمة من اللهجة السعودية ،وبعدها تم جمع 550 تغريدة لمجموعة من المطاعم ،وتم حساب السمعة علي مستوي التغريدة وعلي مستوي الكلمة ولإختبار النتائج تمت مقارنتها مع ترتيب المطاعم حسب السمعة من موقع Qaymالمتخصص في تقييم المطاعم(Haifa Al-Hussaini ,2017) .



الشكل(4.2) رسم توضيحي يمثل تقييم سمعة المطاعم مقارنه مع موقع Qaym

2.8.2 دراسة تقوم بتحليل المشاعر المستخدمة في التغريدات :

في هذه الدراسة تم إسترجاع تغريدات باللغة الإنجليزية ، ثم يقوم بتحليل هذه التغريدات لمعرفة نسبة التغريدات الإيجابية والسلبية ، وذلك بإستخدام خوارزمية naïve base (د.محمد مصطفي حجور،2016)



الشكل(5.2) رسم توضيحي يمثل النتائج لتحليل التغريدات

2.8.3التنبؤ بالجرائم عن طريق تغريدات تويتر:

إكتشف أستاذ مساعد في جامعة فرجينيا ماثيو جربر وجود علاقة بين تغريدات والجريمة. من خلال النظر في إحداثيات الموقع الجغرافي المرتبطة بالتغريدات ، يمكن للشرطة التنبؤ بالمكان الذي يحدث فيه حدوث الجريمة على الأرجح (Matthew Gerber,2013).

جمع جربر 1.5 مليون تغريدة في منطقة شيكاغو بين يناير ومارس 2013. كما قام بسحب سجلات الجريمة خلال الفترة نفسها. قام بتقسيم تغريدات على أساس الموقع الجغرافي ونظر في بيانات الجريمة لكل منطقة لمعرفة ما إذا كانت الكثافة السكانية مرتبطة بالجريمة. في النتيجة ، يمكن أن تتنبأ الطريقة بدقة "19 من 25 نوعًا من الجرائم".

وضرب لذلك مثلا بالقول "الناس يغردون حول كيفية تمضية أيامهم، فمثلا إن عرفنا أن فلانا ينوي أن يمضي ليلته في شرب الكحول ثم أعلن أشخاص آخرون أنهم سينضمون إليه، معنى ذلك أننا سنكون أمام احتمال وقوع مخالفات قانونية في ذاك المكان."

وحلل غربر وزملاؤه التغريدات التي كتبها أشخاص في أحياء مدينة شيكاغو، مستندين إلى قاعدة بيانات رسمية حول المخالفات والجرائم في المدينة، وتمكنوا بفضل ذلك من توقع بعض المخالفات التي وقعت فعلا بعد ذلك. وجاء في الدراسة "أن هذه المقاربة تتيح تحديد المناطق حيث احتمالات وقوع جرائم فيها تكون مرتفعة، إذ أن الجرائم والمخالفات القانونية تقع غالبا في أماكن سبق تسجيل جرائم فيها، وبذلك يمكن إعداد خارطة للمناطق سيئة السمعة تكون أداة مفيدة لتوقع الجرائم(Matthew Gerber,2013) "

2.8.4 تحليل التغريدات التي تحتوي علي معلومات طبية :

(Are Health-Related Tweets Evidence Based? Review and Analysis of Health-Related Tweets on Twitter)

تهدف هذه الدراسة إلى النظر في التغريدات التي تحتوي علي معلومات طبية على تويتر للتحقق من صحتها (تستند إلى الأدلة) ولإثارة الوعي في المجتمع حول أهمية التغريدات المرتبطة بالصحة والقائمة على الأدلة. تم جمع 625 من تغريدات باللغة العربية ذات الصلة بالصحة من 8 حسابات لأطباء ، و 10 حسابات لاتنتمي للمؤسسات الصحية ، و 4 حسابات غذائية ، و 3 حسابات حكومية.وبعد مراجعة البيانات كانت النتائج أن هنالك 320 (51.2٪) من التغريدات خاطئة و 305 (48.8٪) من التغريدات صحيحة. وأيضاً أكثر من نصف التغريدات المرتبطة بالصحة (248/169 ، 68.1 ٪) من المعاهد الصحية غير الرسمية

وحسابات التغذية (101/59 ، 58.4 ٪) كانت خاطئة. كانت التغريدات عبر الأطباء في الغالب "صحيحة" مقارنة بالمجموعات الأخرى (Khalid A Alnemer,2015) .

2.8.5 تحليل التغريدات السلبية على تويتر مؤشر للإصابة بأمراض القلب

Psychological Language on Twitter Predicts County-Level Heart Disease Mortality :

هذة الدراسة أعدها فريق من الباحثين في "جامعة بنسلفانيا" ، وحلل فيها عالم النفس ، يوهانس أيكستت ، وزملاؤه ، بيانات تحتوي على 826 مليون تغريدة في «تويتر» ، كتبها أشخاص ينتمون إلى 1400 مقاطعة أميركية ، بين عامي 2009 و 2010 تضم نحو 90% من تعداد الولايات المتحدة وتم ربطها بمعلومات حول الوفاة بسبب أمراض القلب .و من أجل تحليل هذه المعلومات وربطها بِالإصابة بأمراض القلب ، قام الفريق بِتحليل لغة التغريدات للمساعدة في التنبؤ بالوفيات الناجمة عن الإصابة بأمراض القلب إستناداً علي نموذج تعلم الانحدار / الآلة (regression/machine learning model)

حيث شهدت المقاطعات ــ التي تضمنت تغريدات سكانها كلمات تتصل بالخصومة والعدائية والكراهية والإرهاق، مثل "أحمق" و"غيور" و"ضَجِر" ــ معدلات أعلى بشكل ملحوظ من الوفيات الناجمة عن تصلب الشرايين، ومن ذلك النوبات القلبية والسكتات .وعلى النقيض من ذلك ، كانت أمراض القلب أقل شيوعاً في المناطق التي عكست تغريدات سكانها مشاعر أكثر إيجابية ومشاركة(Johannes C. Eichstaedt,2015) .

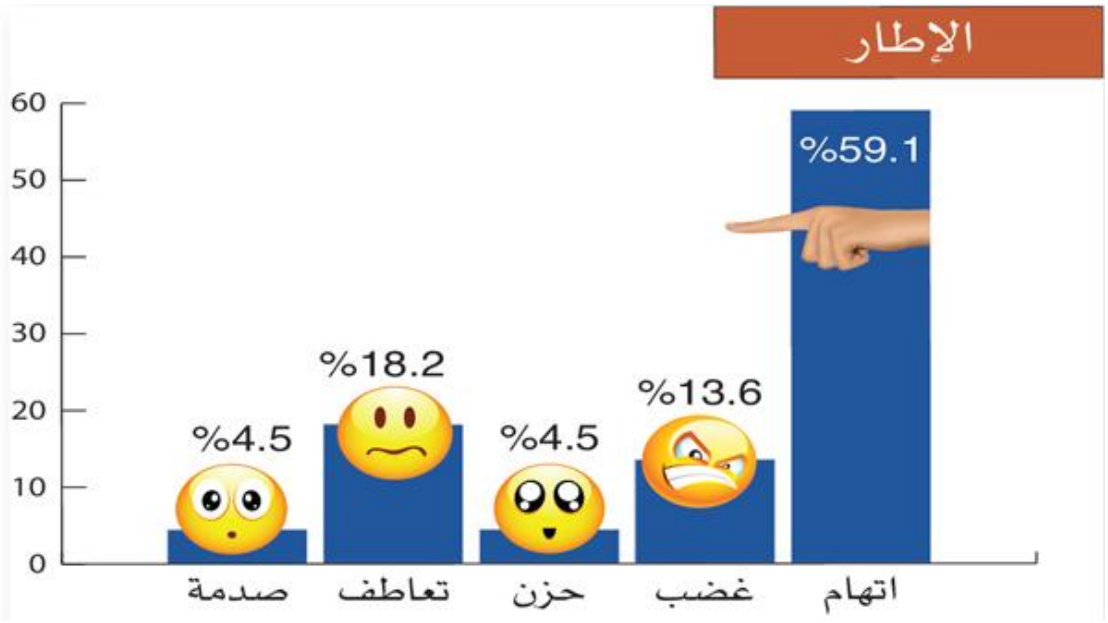
2.8.6 تحليل للرأي العام الإلكتروني عقب إستهداف الأماكن المقدسة

Analyze and monitor the comments users of social networking site (Twitter) about the bombing of the Medina:

تم إعداد هذه الدراسة لتحليل ورصد تعليقات مستخدمي موقع التواصل الاجتماعي (تويتر) حول حادثة تفجير المدينة المنورة لتحليل تعليق الرأي العام العربي على هذا الحادث وكيف تم وصف القائم به.

وبينت النتائج أن أكثر من ثلث المستخدمين (36.4%) وصفوا ما حدث أنه تفجير، بينما ربط ربع المستخدمين 25% الحادث بداعش و20.5% وصفوا الحادث بالإرهاب . وأظهرت الدراسة أن 84.1% من المغردين تحدثوا عن نتائج التفجير فيما غرد 4.5% فقط عن الحادث نفسه .كما بينت أن غالبية المغردين (51.1%) إستخدموا إطار الإتهام في وصف الحادث و18.2% إطار التعاطف .

من جانب آخر، أشارت النتائج أن 29.3% من المغردين يشيرون الى أن سبب التفجير بعض الحكومات العربية بينما أشار 22% أن هناك مؤامرة ما وراء التفجير .غَطت الدراسة نشاط المستخدمين خلال عشر أيام من التفجير وإستَخلصت الآف التغريدات المتعلقة بالحادثة بطريقة تحليل البيانات الكبيرة (big data) وذلك لأخذ عينات بأسلوب أكاديمي لتحليل إتجاهات الرأي العام بمجمله حول آراء المغردين العرب التي تم التعبير عنها الكترونياً . ومن جانب آخر عبر 4.5% عن حزنهم بما حدث و4.5% كانوا تحت تأثير الصدمة خصوصاً أن إستهداف الارهاب للمدينة المنورة والمسجد النبوي وفي الشهر الكريم كان صادماً للمغردين وعملية دموية يصعب تفسيرها(د.فاطمة السالم،2018).



الشكل(6.2) رسم توضيحي يمثل نتيجة تحليل الرأي العام الإلكتروني عقب إستهداف الأماكن المقدسة

ملخص:

البحث	المنهجية	المجال	النيجة
(Haifa Al-Hussaini and Hmood 201)	➢ 1424 كلمة من اللهجة السعودية ➢ تم جمع 550 تغريدة لمجموعة من المطاعم ➢ التصنيف (classification FP growth)	تقييم المطاعم	وتم حساب السمعة علي مستوي التغريدة وعلي مستوي الكلمة ولإختبار النتائج تمت مقارنتها مع ترتيب المطاعم حسب السمعة من موقع Qaymالمتخصص في تقييم المطاعم

د.محمدمصطفي حجور،2016	➢ تم إسترجاع تغريدات باللغة الإنجليزية ➢ لمعرفة نسبة التغريدات الإيجابية والسلبية ➢ خوارزمية naïve base	**تحليل المشاعر المستخدمة في التغريدات**	65%تغريدات إيجابيه
ماثيو جربر 2013	➢ حلل تغريدات "شيكاغو"، مستندين إلى قاعدة بيانات رسمية حول المخالفات والجرائم في المدينة	**العسكري**	**التنبؤ بالجرائم**

خالد النمر ،2015	تم جمع 625 من تغريدات باللغة العربية ذات الصلة بالصحة من 8 حسابات لأطباء ، و 10 حسابات لاتنتمي للمؤسسات الصحية ، و 4 حسابات غذائية ، و 3 حسابات حكومية . تم التقييم من قبل أطباء مختصين وتصنيفها الي صواب وخطأ.	الطب	**التحقق من صحة المعلومات الطبية** هنالك 320 (51.2٪) من التغريدات خاطئة و 305 (48.8٪) من التغريدات صحيحة
JohannesC.Eichstaedt, 2015	التنبؤ بالوفيات الناجمة عن الإصابة بأمراض القلب إستناداً علي نموذج تعلم الانحدار / الآلة (regression/machine learning model)	الطب	**التغريدات السلبية على تويتر مؤشر للإصابة بأمراض القلب** حيث شهدت المقاطعات التي تضمنت تغريدات سكانها كلمات تتصل بالخصومة

			والعدائية والكراهية والإرهاق، مثل "أحمق" و"غيور" و"ضَجِر" – معدلات أعلى بشكل ملحوظ من الوفيات الناجمة عن تصلب الشرايين

إستخدمت الدراسات السابقة بيانات نصية تم الحصول عليها من موقع التواصل الإجتماعي تويتر حول مواضيع مختلفة ، وبعد تحليلها بإستخدام طرق وخوارزميات تختلف من دراسة إلي أخري حسب المعلومات المطلوب الحصول عليها .

يتفق هذا البحث مع الدراسات المذكورة أعلاه في إستخدام بيانات تم جلبها من تويتر وتحليلها لإستخراج معلومات مفيدة ويختلف في المواضيع التي تتناولها تلك البيانات ،ففي هذا البحث تم جلب بيانات متعلقة بالمخدرات والإدمان . وأيضاً يختلف في طريقة التحليل والخوارزميات المتبعة حيث تم إستخدام طريقة العنقدة لتقسم البيانات إلي مجموعات مختلفة .

الفصل الثالث

منهجية البحث

الفصل الثالث

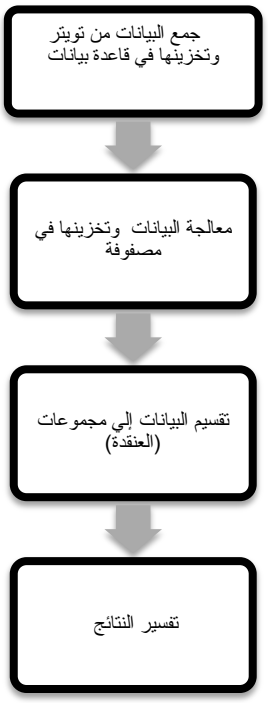
منهجية البحث

3.1 مقدمة:

يحتوي هذا الفصل علي المنهجية التي إستخدمها الباحث في كل مراحل البحث من بداية جمع المعلومات وصولاً إلي إستخراج النتائج.

في هذا البحث تم جمع البيانات المتعلقة بالمخدرات والإدمان المكتوبة باللغة الإنجليزية من موقع التواصل الإجتماعي تويتر بإستخدام كود php ، وتم تخزينها في قاعدة بيانات ومن ثم ، تمت معالجتها عن طريق تقسيم الجملة إلي كلمات منفصلة وحذف المساحات الفارغة الكبيره وعلامات الترقيم ، وأيضاً تحويل الأحرف الكبيره إلي أحرف صغيرة لتقليل مساحة تخزينها في الذاكرة وتسهيل وتسريع عملية تحليلها ثم إزالة الزوائد التي تستخدم للربط بين الكلمات ولاتؤثر علي الرأي عند حذفها مثل الحروف المساعدة وإرجاع كل كلمة إلي اصلها (الجذر) وتخزينها في مصفوفة Term–Matrix.

الخطوة التالية هي تقليل حجم المصفوفة عن طريق حذف الكلمات الأقل تأثيراً (تكراراً) وتخزين الأكثر تكراراً في مصفوفة جديدة . وتم تحليلها وتقسيمها إلي مجموعات بإستخدام خوارزميتي K–means وWard’s لمعرفة أسباب إنتشار المخدرات.



الشكل(1.3) رسم توضيحي يمثل منهجية البحث

3.2 معالجة البيانات:

عند تحليل أي نوع من البيانات لابد ان تسبق عملية التحليل معالجة للبيانات وذلك لتهيئة البيانات وحذف ماقد يسبب ضوضاء منها، يمكن أن تحدث هذه الضوضاء بسبب ممارسات جمع البيانات من عدة مصادر، وفقد البيانات عن طريق الخطأ ، وبيانات الملوثات مثل التعديلات غير الصحيحة ، والحذف أو الإضافات، أو في كثير من الحالات لمجرد أن البيانات تحتوي على حالات لا علاقة لها بهدف مهمة استخراج البيانات (Han and Micheline Kamber,2010) .

وفي هذا البحث تم اجراء المعالجة بالخطوات الموضحة في الجدول التالي:

الخطوة	الشرح	مثال
جمع البيانات	اذا كان لدينا تغريدة من تويتر عن المخدرات	Bad friends take you to addiction.
التقسيم او التطبيع(tokenization)	تقسيم الجملة الي كلمات منفصلة إعتماداً علي المسافات بين الكلمات لتبسيط عملية المعالجة	Bad Friends Take You to addiction.
التصغير(normalization) وحذف حروف الوقف(stop word)	إزالة الزوائد وإستبدال الأحرف الكبيرة بالصغيرة والحروف.	bad Friend Take Addiction
النابعة(stemming)	إرجاع الكلمة الي جذرها	bad Friend Take Addict

الجدول (3.1) معالجة البيانات

3.3 بناء مصفوفة للمستند (document-term matrix):

هي مصفوفة رياضية تصف تكرار الكلمات في مجموعة من الوثائق(البيانات).حيث تمثل الصفوف المستندات وتمثل الأعمدة المصطلحات ، ويتم تمثليها ب1 في حال ظهور الكلمة في المستند و0 في حالة عدم الظهور ،ومن هنا يمكن تحديد أهمية المصطلح حيث أن المصطلحات الأكثر تكراراً هي الأكثر أهمية. علي سبيل المثال في حال وجود مستندين س،ص

	1	2	3	4	5
abus	1	0	0	0	0
drug	1	1	1	0	1
love	1	0	1	1	0

الجدول(3.2) مصفوفة المستند

في هذا البحث تم إنشاء مصفوفة تحتوي على 21176640 عنصر ولتسهيل عمل الخوارزمية تم حذف الكلمات القليلة التكرار وتكوين خوارزمية جديدة تحتوي علي 343440 عنصر .

الآن بعد أن أصبحت البيانات ممثلة علي صورة مصفوفة جاهزة للتحليل بإستخدام خوارزميات مختلفة ،وفي هذا البحث تم إستخدام خوارزميتي Ward’s وk-means .

3.4 خوارزمية الـــ K-Means:

تقوم هذه الخوارزمية بتجميع البيانات حول مراكز يتم حسابها عن طريق ايجاد الوسط الحسابي لكل العناصر ،حيث أن العنصر الوسط هو المركز الذي تتجمع حوله بقية العناصر المشابهه له [3] .

طريقة عمل خوارزمية الK-means:

1. تحديد عدد التجمعات K.
2. تحديد إحداثيات مراكز التجمعات Centroid عشوائياً لأول مرة ويتم حسابه عن طريق إيجاد (متوسط النقاط التي تنتمي للمركز) لباقي المرات.
3. حساب المسافة بين كل عنصر وبين جميع المراكز، ويتم استخدام البعد الإقليدي.

يعطى البعد الإقليدي d_{ij} بين مثالين i,j بالعلاقة التالية :

$$(1)\ d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2}$$

حيث إن :

n: عدد العناصر المراد تصنيفها الي مجموعات.

$x_{ik} - x_{jk}$: فرق إحداثيات العناصر بعد تمثيلها علي المستوي الديكارتي.

4. تجميع العناصرمع أقرب مركز لها.

وتتكرر الخطوات 2_4 ونقل المراكز حتي نتوصل الي امثل نموذج حيث يتم تصنيف كل عنصر الي مجموعته المناسبة.

3.5 خوارزمية ward's:

خوارزمية ward's إحدى طرق التجميع الهرمي ، حيث تبدأ بمجموعة من العناقيد كل واحد منها يحتوي علي عنصر واحد فقط ،ثم تقوم بدمج العناقيد المجاورة المتشابهة وتستمر بالدمج حتي تصل إلي عدد العناقيد المحدد مسبقاً او أن تصبح جميع العناقيد المتشابهة في نفس المجموعة .

تحدد هذه الخوارزمية العناصر المتشابهة عن طريق حساب الإختلاف بين النقاط ،حيث أن العناصر المتشابهة هي الأقل اختلاف.

3.6 الأدوات التي تم إستخدامها في هذا البحث:

- **wamp serve الإصدار 2.5**
- **MySQL الإصدار 5.6.17**
- **RStudio**

3.6.1 WampServer

هو بيئة لتطوير تطبيقات الويب على نظام التشغيل ويندوز (Windows)، تم تصميمه بواسطة (Romain Bourdon). هذه البيئة متاحة مجاناً بموجب ترخيص البرنامج العالمي (GPML) في نسختين متميزتين : 32 و 64 بت . تسمح البيئة للمطورين بإنشاء تطبيقات الويب بإستخدام (Apache2) ، لغة البرمجة (PHP) وقاعدة البيانات (MySQL) . كما تحتوي البيئة على جزء يُسمى (PhpMyAdmin) الذي يسمح للمطور بسهولة إدارة قواعد البيانات .

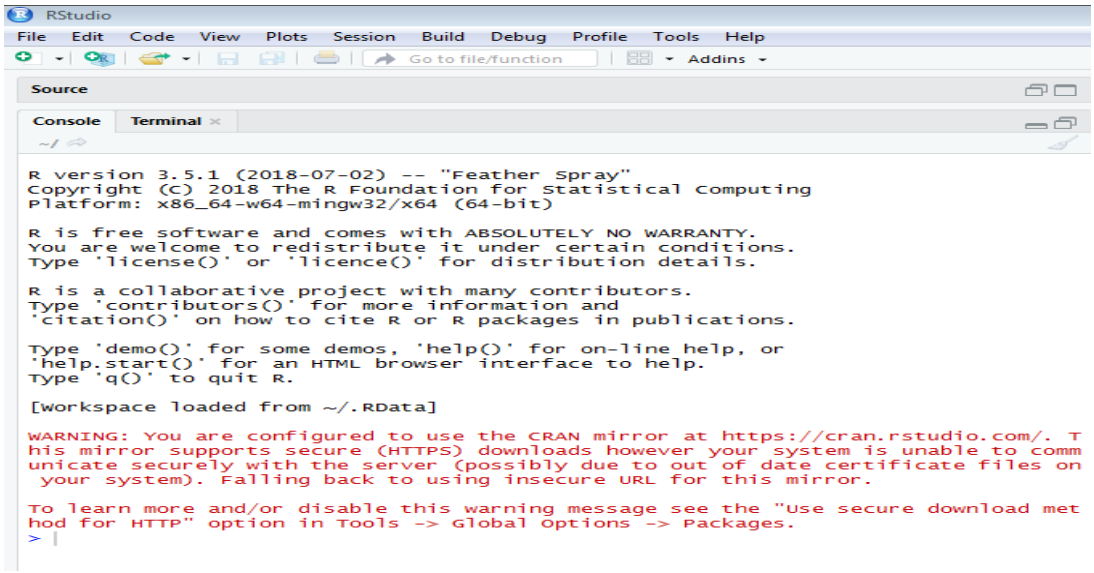
3.6.2MYSQL

تعرف MySQL بأنها قاعدة بيانات، وهي طريقة من طرق الاحتفاظ بالبيانات، وقاعدة البيانات تتكون من جداول والجداول تحتوي على صفوف وأعمدة وخلايا. تتكون أي قاعدة بيانات غالباً من عدة جداول، ويحمل كل جدول إسماً مختلفاً يميزه مثلاً "customers" أو "products"أو"orders" ، وكل جدول يتكون من صفوف تحتوي البيانات.

3.6.3 لغة R:

عبارة عن مجموعة متكاملة من البرمجيات التي تسمح بمعالجة البيانات، القيام بعمليات حسابية و إظهار البيانات الرسومية و يمكن تحميل لغة R من الموقع الرسمي لها على الشبكة والموجود على العنوان http://www.r– project.org

تتكامل RStudio مع الR وتعتبر بيئة تطوير متكاملة (IDE) لتوفير مزيد من الوظائف فهي تجمع بين محرر شفرة المصدر، وبناء أدوات التشغيل الآلي والمصحح.



الشكل(2.3) رسم توضيحي يمثل RStudio

الفصل الرابع

التطبيق

الفصل الرابع

التطبيق

4.1 مقدمة:

في هذا الفصل سيتم توضيح كيفية الحصول علي البيانات من موقع التواصل الإجتماعي تويتر وتخزينها في قاعدة بيانات ومن ثم معالجة البيانات و تطبيق خوارزميتي الK–Means و Ward's لتقسيم البيانات لمجموعات حيث تنتمي العناصر المتشابهة لنفس المجموعة.

4.2 جمع البيانات:

4.2.1 إنشاء حساب في تويتر:

الخطوة الأولي هي انشاء حساب في تويتر عن طريق الموقع الرسمي لتويتر https://twitter.com

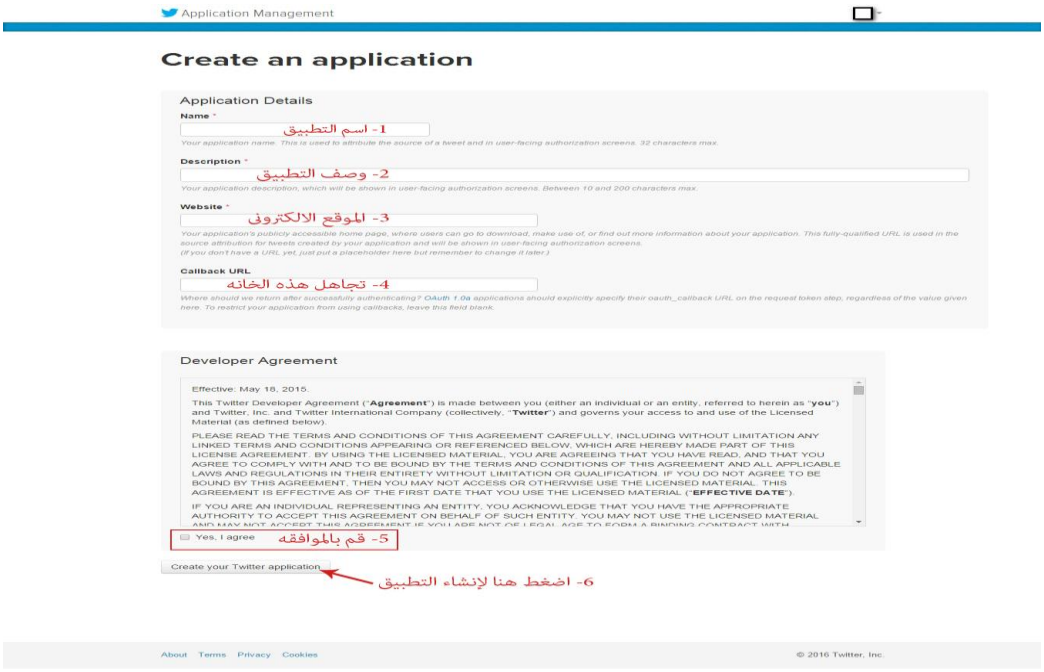
4.2.2 المصادقة و التفويض في تويتر

المصادقة : هي عملية التحقق من هوية المستخدم.

التفويض أو الإذن : هو عملية التحقق من أن المستخدم لديه الحق في تنفيذ بعض الجراءات، مثل قراءة وثيقة أو الوصول إلى حساب البريد الإكتروني.

4.2.3 إنشاء تطبيق تويتر:

يمكنك إنشاء التطبيق من خلال هذا الرابط: https://apps.twitter.com ويتطلب التطبيق إدخال إسم التطبيق ووصف عنه ورابط موقعك .يجب الموافقة على الشروط المكتوبة ثم الضغط على إنشاء تطبيق (Create your twitter application).

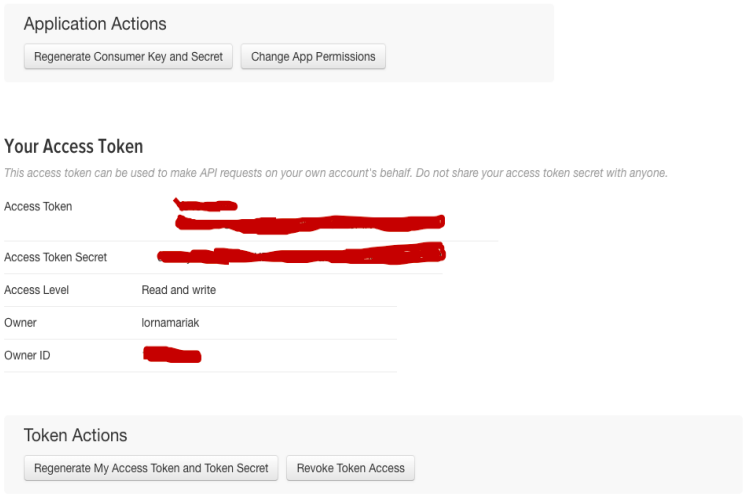


الشكل (1.4) رسم توضيحي يمثل إنشاء تطبيق تويتر

4.2.4 الصلاحيات بعد إنشاء تطبيق تويتر

بعد إنشاء التطبيق يعطي المستخدم الصلاحيات التي تمكنه من الوصول إلى جميع المعلومات وهي :

- Consumer key مفتاح المستهلك.
- Consumer secret الرقم السري للمستهلك.
- Access token رمز الوصول.
- Access secret الرقم السري للوصول.



الشكل (2.4) رسم توضيحي يمثل الصلاحيات التي تعطى للمستخدم بعد إنشاء تطبيق تويتر

4.2.5 الحصول على البيانات

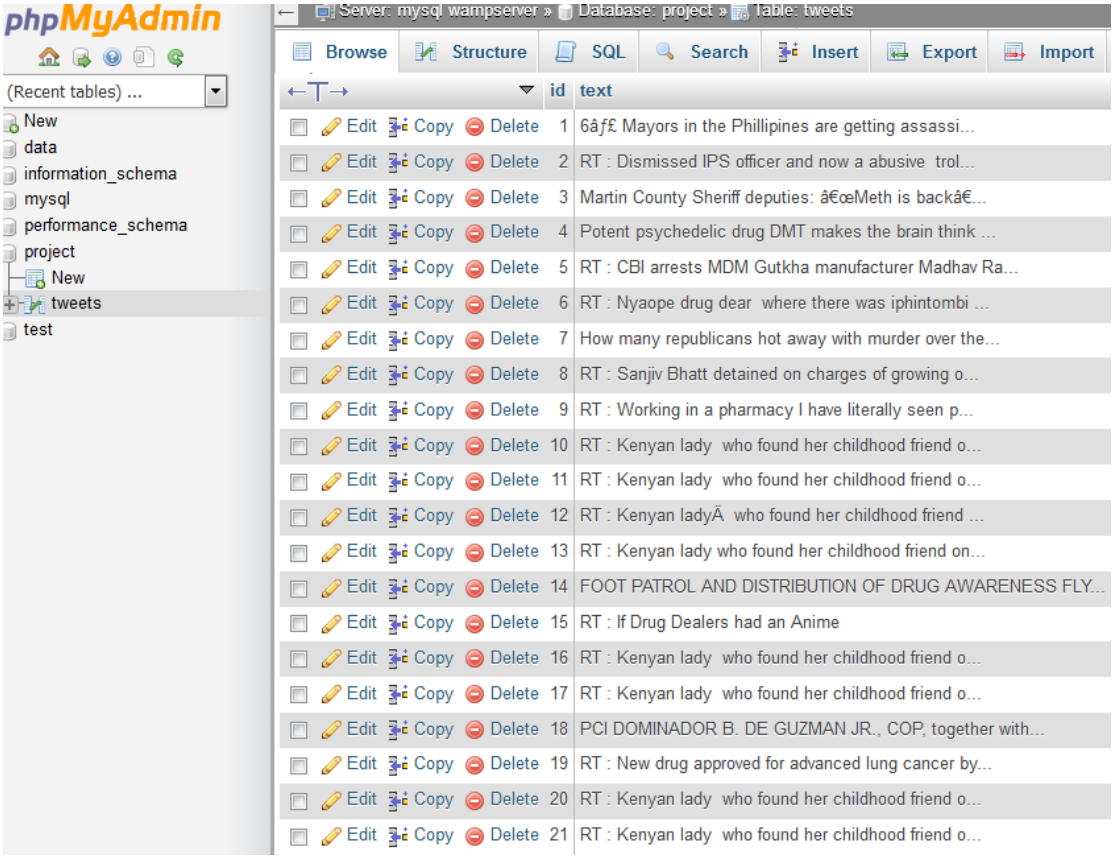
تم إنشاء ملف من نوع PHP وتم إستخدام الدالة **twitteroauth( )** للإستفادة من الصلاحيات التي تم الحصول عليها من الخطوة السابقة.

تم إستخدام search API لإسترجاع مجموعة من التغريدات المكتوبة باللغة الإنجليزية والتي تحتوي علي كلمة مخدرات أو إدمان ، وتم جمع ما يقارب5 الف تغريدة وتخزينها في قاعدة بيانات لتتم معالجتها وتحليلها.

get(https://api.twitter.com/1.1/search/tweets.json?lang=en&count=100&

q=drug&–filter:retweet

الشكل(3.4) رسم توضيحي يمثل الحصول علي البيانات من تويتر وتخزينها



الشكل(4.4) رسم توضيحي يمثل نموذج للبيانات التي تم الحصول عليها من تويتر

4.3 تحليل البيانات :

يحتوي هذا الجزء من البحث علي معالجة البيانات و تطبيق خوارزميتي ال-K Means و Ward's لتقسيم البيانات لمجموعات حيث تنتمي العناصر المتشابهة لنفس المجموعة.

4.3.1 تحميل البيانات إلي الR:

تم تصدير البيانات من قاعدة البيانات إلي ملف نصي (file.txt) ومن ثم تحميله إلي الR بإستخدام الداله(( )readLines(file.Choose ، فهي تمكن المستخدم من استعراض الملفات وتحديد المستند المطلوب تحميلة.

4.3.2 معالجة البيانات:

معالجة البيانات هي عملية تنظيف البيانات وإزالة الزوائد وعلامات الترقيم وارجاع الكلمات إلي جذورها بعد أن تم تقسيم الجمل إلي كلمات (Han and Micheline Kamber,2010) .

في هذا البحث تم إستخدام المكتبة tm التي تحتوي علي مجموعة من الدوال لمعالجة النصوص مثل الدالة ()tm_map التي إستخدمت في هذا البحث لحذف الأحرف الخاصة (",/" "|" , "@") والمساحات الكبيرة بين الكلمات وتحويل الأحرف الكبيرة الي الصغيرة . كما أن الأرقام حذفت بإستخدام الدالة( )**removeNumbers**

1.3.3 بناء مصفوفة للكلمات:

تم بناء المصفوفة بإستخدام الدالة ( )TermDocumentMatrix ولحذف الكلمات الأقل تكرار ()removeSparseTerms ،وهكذا أصبحت البيانات مُعدة لإستخدام الخوارزمية عليها.

1.3.4 إيجاد الكلمات الأكثر تكراراً في البيانات:

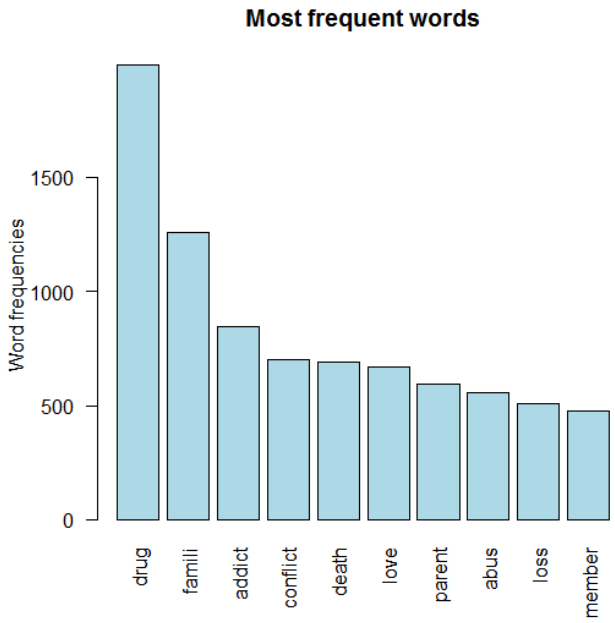
توفر لغة البرمجة R إمكانية تحليل البيانات وإجراء حسابات إحصائية وتحاليل إستكشافية غلى البيانات للحصول على نتائج معينة. قمنا بالإستفادة من هذه الخاصية لإيجاد أكثر الكلمات تداولاً في التغريدات التي تم جمعها سابقاً وذالك بإتباع الخطوات التاية :

- تحميل الحزم المطلوبة وهي (tm,ggplot) وإستدعاء مكتباتها .
- تحديد رقم يمثل أقل ظهور للكلمة وقمنا بإختيار الرقم 150 بحيث يتم جلب كل الكلمات التي تكررت على الأقل 150 مرة وعدد الكلمات هي 150 كلمة بالشكل التالي .

```
                word freq
drug             drug 1480
stress         stress 1297
friend         friend  920
poor             poor  684
cultur         cultur  676
work             work  539
school         school  470
relationship relationship  447
financ         financ  443
addict         addict  313
```

الشكل (5.4) رسم توضيحي يمثل تكرار الكلمات في البيانات

1.3.5 إنشاء المخطط البياني لأكثر الكلمات تكراراَ :



الشكل (6.4) رسم توضيحي يمثل لأكثر الكلمات تكراراً

من المخطط السابق نجد أن أكثر الكلمات هي "drug" وقد تكررت أكثر من 1400 مرة تليها كلمة "stress" بمعدل 1297 مرة.

4.3.6 إنشاء سحابة الكلمات(Word Cloud):

تعد سحابة الكلمات واحدة من أفضل الأدوات التي تسمح لنا تصور معظم الكلمات والمصطلحات الواردة في التغريدات على الرغم من الإستخدام الرئيسي لها هو لأغراض إستكشافية، فهي لديها ميزة أن تكون مفهومة من قبل معظم المستخدمين، وأن تكون جذابة بصرياَ إلى العين البشرية. وقمنا بإنشاء سحابة الكلمات بإتباع الخطوات التالية :

- تحميل الحزم المطلوبة وهي (WordCloud,RColorBrewer) وإستدعاء مكتباتها .
- حساب عدد الظهور لأكثر الكلمات تكراراَ وترتيبها تنازلياَ.
- رسم الWordCloud لأكثر الكلمات تكراراَ بالشكل التالي:

شكل(7.4) رسم توضيحي يمثل سحابة الكلمات

4.4 تطبيق خوارزمية Hierarchical clustering

تعتمد هذه الخوارزمية على تقسيم البيانات بشكل هرمي.

4.4.1 كيفية عمل الخوارزمية

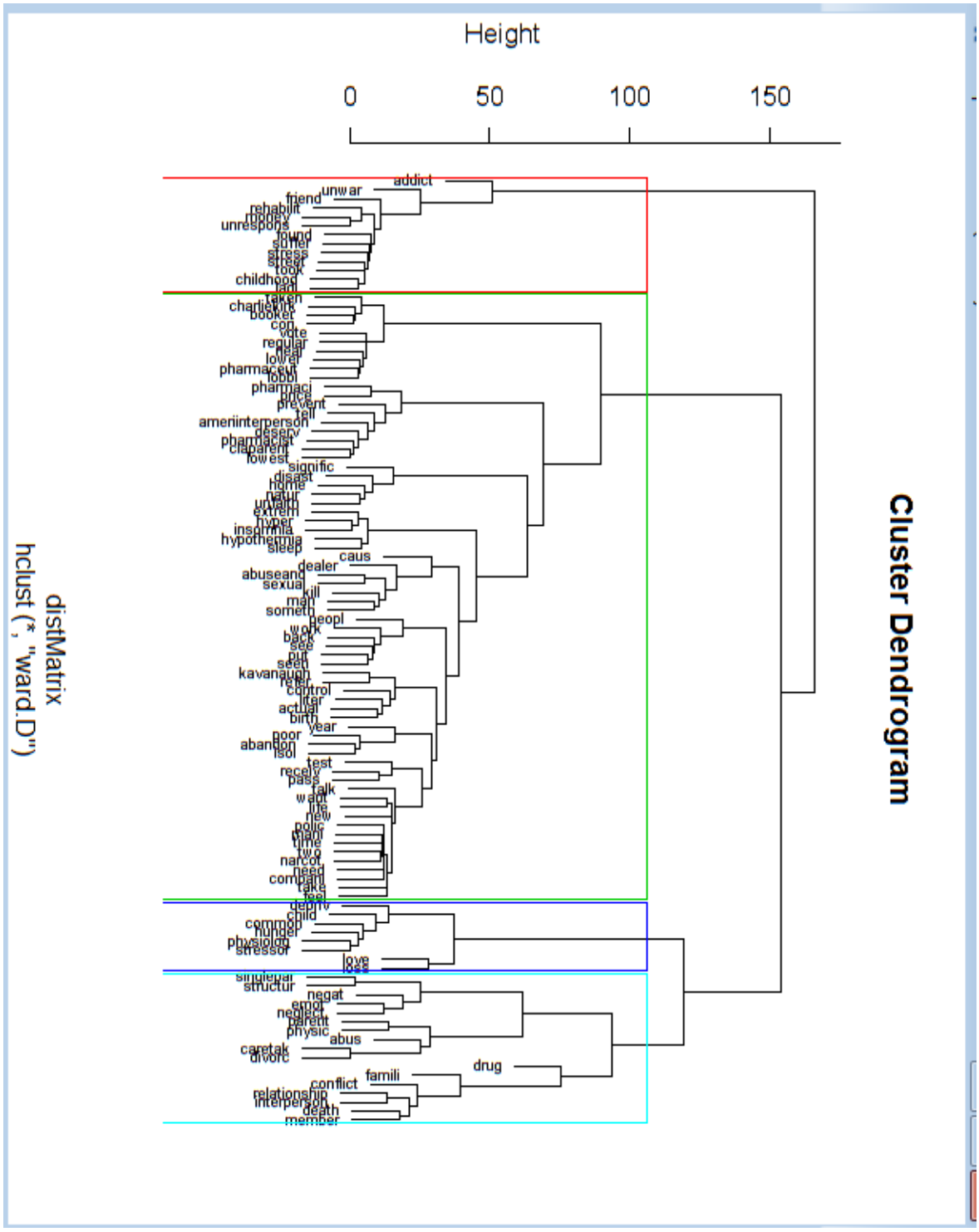
نفترض أنه لدينا مجموعة من البيانات "N"ومصفوفة مسافة N*N (المصفوفة تحتوي على المسافة بين البيانات)

نقوم بإتباع الخطوات التاية :

1. نبدأ بوضع كل البيانات في مجموعة على حدة بالتالي إذا لدينا N من البيانات تصبح لدينا N من المجموعات.
2. كل أقرب زوج من البيانات يتم دمجها في مجموعة واحدة.

3. يتم حساب المسافة للبيانات التي تم دمجها(لحساب المسافة تستخدم طريقة تسمى single linkage clustering والتي تعتمد على حساب المسافة من أقرب نقطة لل clustering الجديد) .

4. تكرار الخطوات 2 و 3 إلى أن تصبح البيانات كلها في مجموعة واحدة.



الشكل(8.4) رسم توضيحي يمثل عناقيد البيانات بإستخدام خوارزمية ward's

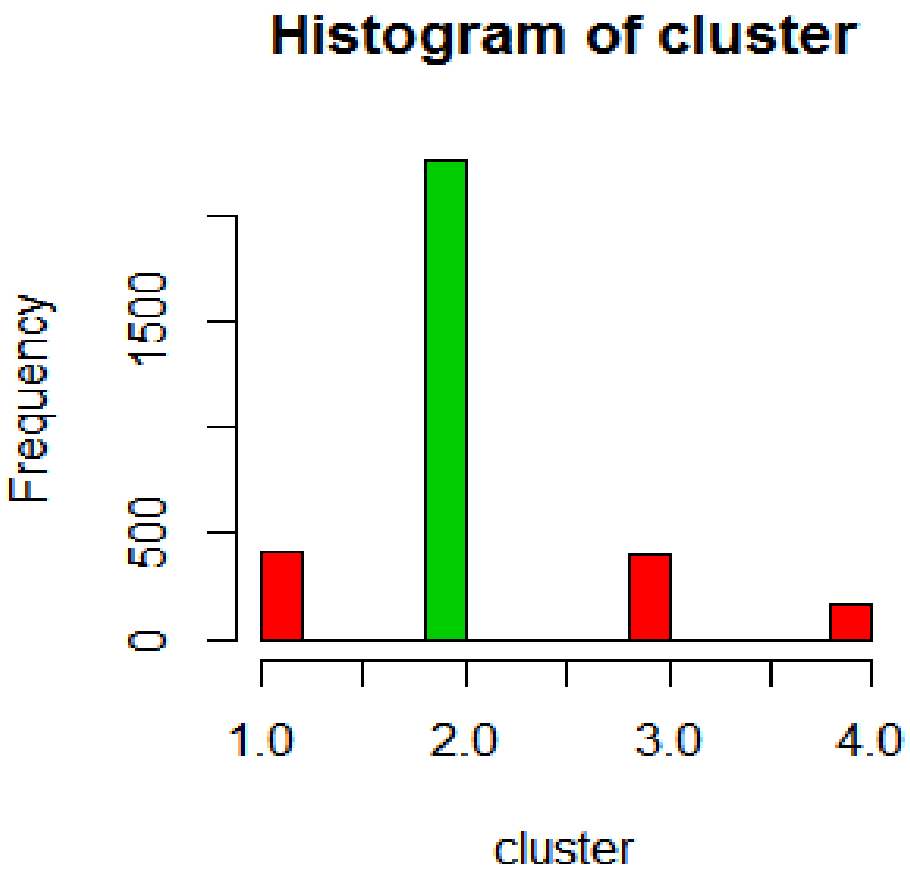
4.4.2 تطبيق خوارزمية (K Means)

1. تحديد عدد التجمعاتK، وهي تعتبر خطوة تهيئة أولية. وفي هذه الحالة تصبح عدد التجمعات تساوي 6 لاننا نريد تقسيم التغريدات الى 4 فئات (قلة الوعي والجهل، البيئه المحيطه والعلاقات ، أسباب نفسيه ، اخرى) ، k = 4.
2. تحديد إحداثيات مراكز التجمعات Centroid عشوائياً لكل كلمة لأكثر الكلمات تكرارَا. وذلك لكل فئة من الفئات للمرة الأولى فقط.
3. –حساب المسافة بين كل البيانات ومراكز التجميع. قبل حساب المسافة يتم ترميز البيانات الى أرقام على اساس هذه الأرقام يتم حساب المسافة بين مركز التجمع والبيانات.
4. تجميع البيانات وتنظيمها في 4 مجموعات بناءَ على أقل المسافات بين المركز ونقاط البيانات.

عاده يتم تنفيذ الخطوات من 2– 3 حتى الوصول إلى حاله الثبات.

```
> kmeansResult[["centers"]]
       abus      drug      love        back        two    abuseand    addict
1 0.79326923 0.5985577 0.1514423 0.021634615 0.01201923 0.007211538 0.2283654
2 0.03636364 0.6128603 0.1458980 0.038580931 0.02217295 0.030155211 0.2820399
3 0.30198020 0.6262376 0.1410891 0.014851485 0.01732673 0.027227723 0.1955446
4 0.13939394 0.6484848 1.3454545 0.006060606 0.02424242 0.036363636 0.2121212
       caus      child      common      death      depriv      hunger
1 0.1105769 0.01442308 0.007211538 0.04567308 0.002403846 0.002403846
2 0.1157428 0.01019956 0.002660754 0.06917960 0.027937916 0.001773836
3 0.1237624 0.01485149 0.000000000 1.19801980 0.034653465 0.002475248
4 0.1090909 1.17575758 1.169696970 0.17575758 1.115151515 1.163636364
       loss   physiolog      sexual    stressor   conflict    famili
1 0.36538462 0.002403846 0.009615385 0.002403846 0.33653846 0.7860577
2 0.05144124 0.000000000 0.035476718 0.000000000 0.00000000 0.1135255
3 0.08663366 0.000000000 0.027227723 0.000000000 1.35396040 1.5173267
```

الشكل (9.4) تحديد مراكز التجمعات عشوائياً لكل فئة بإستخدام ال R



الشكل (10.4) رسم توضيحي يمثل عناقيد البيانات (cluster) بإستخدام خوارزمية k-means

الفصل الخامس

النتائج والتوصيات

الفصل الخامس

النتائج والتوصيات

5.1 مقدمة:

يتناول هذا الفصل النتائج التي تم التوصل اليها والتوصيات التي تخص المشروع.

5.2 النتائج:

تم إنشاء مصفوفة تحتوي علي 343440 كلمة و تطبيق خوارزميتي ward's و k_means لتحليلها .

5.2.1 التحليل بإستخدام خوارزمية ward's:

في هذا البحث تم إستخدام هذه الطريقة لتقسيم البيانات إلي 4 مجموعات لمعرفة الأسباب الرئيسية لإنتشار المخدرات وكانت النتائج كالآتي:

العنقود(cluster)	Hierarchical
قلة الوعي أو الجهل	%16
البيئة المحيطة والعلاقات	%65
أسباب نفسية	%12
أسباب أخرى	%7

الشكل (1.5) رسم توضيحي يمثل عناقيد البيانات (cluster) بإستخدام خوارزمية ward's

5.2.2 التحليل بإستخدام خوارزمية K-Means:

تم إستخدام الدالة kmeans لتقسيم البيانات المخزنة في مصفوفة الكلمات (البيانات التي جمعت من تغريدات تويتر حول المخدرات والإدمان) إلي 4 مجموعات أو عناقيد بإتباع خوارزمية الK-Means .وكانت النتائج كالآتي:

العنقود(cluster)	K-means
قلة الوعي أو الجهل	%13
البيئةالمحيطة والعلاقات	%70
أسباب نفسية	%12
أسباب أخرى	%5

الشكل (2.5) رسم توضيحي يمثل عناقيد البيانات (cluster) بإستخدام خوارزمية k-means

5.2.3 مقارنة لنتيجة الخوارزميتين اعلاه :

تم إنشاء مصفوفة تحتوي علي 343440 كلمة و تطبيق الخوارزميتين أعلاه لتحليلها وتم التوصل إلي أسباب إنتشار المخدرات بحسب آراء مستخدمين تويتر :

العنقود(cluster)	Hierarchical	K-means
قلة الوعي أو الجهل	%16	%13
البيئةالمحيطة والعلاقات	%65	%70
أسباب نفسية	%12	%12
أسباب أخرى	%7	%5

الجدول (3.5)يوضح نتيجة تحليل التغريدات المتعلقة بإستخدام المخدرات

نجد ان الخوارزميتين اتفقتا علي أن السبب الرئيسي لإنتشار المخدرات بحسب رأي مستخدمي تويتر يرجع إلى البيئه والعلاقات المحيطه تليها قلة الوعي والجهل بأضرار المخدرات ومن ثم الأسباب النفسيه لمستخدم المخدرات بنسبة 12% للخوارزميتين وأخيراً تأتي أسباب أخرى كالفراغ وعدم المسؤوليه .

5.3 التوصيات :

i. يمكن تطبيق خوارزميات التصنيف Classification algorithms وخوارزميات تجميع اخرى و مقارنة النتائج بنتائج المتحصلة عليها.

ii. إضافة بيانات من وسائل تواصل إجتماعية اخرى مثل الفيسبوك .

الخاتمه

تم بحمد الله ما اردنا جمعه وكتابته عن تحليل التغريدات حول أسباب إنتشار المخدرات وتصنيفها الى الفئات المحددة سلفاً حيث قمنا بإستخدام خوارميات التجميع في الذكاء الإصطناعي وتطبيقها على التغريدات المخزنة في قاعدة البيانات.

وكانت نتيجة التحليل أن هنالك 5 أسباب أتفقت عليها كثير من أراء المغردين أهمها البيئة المحيطة لذلك أرجو من الأباء الإهتمام بالبيئة التي ينمو فيها أبناءهم وتوعيتهم بأضرار المخدرات وتخصيص وقت لمشاركتهم الحديث والفكار ومعرفة مشاكلهم ومساعدتهم في إيجاد الحلول.

المراجع:

1. Anwar Hridoy, S., Ekram, M., Islam, M., Ahmed, F. and Rahman, R. (2015). Localized twitter opinion mining using sentiment analysis. *Decision Analytics*, 2(1).
2. McCormick, T., Lee, H., Cesare, N., Shojaie, A. and Spiro, E. (2015). Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. *Sociological Methods & Research*, 46(3), pp.390-421.
3. Han, J., Kamber, M. and Pei, J. (2011). *Data Mining*. Burlington: Elsevier Science.
4. Alessa, A., & Faezipour, M. (2018). A review of influenza detection and prediction through social networking sites. *Theoretical biology & medical modelling*, *15*(1), 2. doi:10.1186/s12976-017-0074-5.
5. Developer.twitter.com. (2019). *Standard search API*. [online] Available at: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html [Accessed 12 Mar. 2019].
6. Kooi, B. (2013). Assessing the correlation between bus stop densities and residential crime typologies. *Crime Prevention and Community Safety*, 15(2), pp.81-105.
7. Khalid A Alnemer, F. (2019). *Are Health-Related Tweets Evidence Based? Review and Analysis of Health-Related Tweets on Twitter*. [online] PubMed Central (PMC). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4642373/ [Accessed 12 Mar. 2019].
8. Eichstaedt, J., Schwartz, H., Kern, M., Park, G., Labarthe, D., Merchant, R., Jha, S., Agrawal, M., Dziurzynski, L., Sap, M., Weeg, C., Larson, E., Ungar, L. and Seligman, M. (2015). Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological Science*, 26(2), pp.159-169.
9. Kim, D. and Kim, J. (2014). Public Opinion Sensing and Trend Analysis on Social Media: A Study on Nuclear Power on Twitter. *International Journal of Multimedia and Ubiquitous Engineering*, 9(11), pp.373-384.

Liao, S., Chu, P. and Hsiao, P. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12), pp.11303-11311..

10. Al-Hussaini, H. and Al-Dossari, H. (2017). A Lexicon-based Approach to Build Service Provider Reputation from Arabic Tweets in Twitter. *International Journal of Advanced Computer Science and Applications*, 8(4).

11. L.Sumathy, K. and Chidambaram, M. (2013). Text Mining: Concepts, Applications, Tools and Issues An Overview. *International Journal of Computer Applications*, 80(4), pp.29-32.

12. د.محمد مصطفى حجور، تحليل الأراء في تويتر. (2016). حمص.pp ,مجلة جامعة البعث ـــــ المجلد83 العدد 41.

13. القبس الإلكتروني. (2019)، تحليل للرأي العام الإلكتروني عقب استهداف الأماكن المقدسة: مؤامرة ضد السعودية.. http://alqabas.com/163631/ تاريخ الإطلاع 20نوفمبر2018