اكتشاف الموضوعات المهيمنة باستخدام تقنية التنقيب في النصوص في تغريدات الناخبين والمرشحين خلال انتخابات مجلس الأمة الكويتي نموذجًا

د. صلاح راشد الناجم

قسم اللغة العربية

كلية الآداب، جامعة الكويت، دولة الكويت

salah.alnajem@ku.edu.kw

المستخلص:

يقدم هذا البحث دراسة تطبيقية استُخدِمَت فيها تقنية التنقيب في النصوص (Text Mining) للتنقيب في نصوص تغريدات تويتر التي احتوت على كلمات متعلقة بانتخابات مجلس الأمة الكويتي (البرلمان)؛ وذلك لاكتشاف الموضوعات المهيمنة (Topic Extraction) التي تحدث عنها الناخبون والمرشحون خلال فترة انتخابات مجلس الأمة في دولة الكويت في العام ٢٠٢٠ في تلك التغريدات، إضافة إلى اكتشاف العلاقات التي تربط كلمات معينة بموضوعات معينة واكتشاف الكلمات المفتاحية المحورية (Keywords) في إطار موضوعات التغريدات بناء على الوزن الإحصائي لهذه الكلمات. جُمعَت التغريدات من أرشيف تويتر الرسمي (Twitter Full Archive) ؛ حيث جَمَعتُ التغريدات التي احتوت على كلمات بحث متعلقة بانتخابات مجلس الأمة خلال الفترة المذكورة. بعد ذلك استخدمتُ نظام SAS Text Miner للتنقيب في نصوص التغريدات واكتشاف الموضوعات التي تحويها عن طريق استخدام تقنية التحليل الدلالي الكامن (Latent Semantic Analysis (LSA)). حيث تبين أن المزاج العام للناخبين والمرشحين تجاه الإجراءات والمشاريع الحكومية ليس إيجابيا بشكل عام، بل يميل إلى انتقاد تقصيرها في جوانب متعلقة بمحاربة الفساد وتطوير التعليم وتحقيق الإصلاح والتطوير الاقتصادي، كما أن هنالك مزاجًا سلبيًّا تجاه عدم وجود وعي كاف لدى بعض الناخبين عند اختيارهم للمرشحين بشكل يضمن التصويت لمن يستحق من المرشحين للوصول إلى مجلس الأمة، إضافة إلى وجود مزاج سلبي تجاه ظاهرة شراء بعض المرشحين لأصوات بعض الناخبين وذممهم.

الكلمات الدالة: علم اللغة الحاسوبي، المعالجة الحاسوبية للغة العربية، التنقيب في النصوص، علم اللغة التطبيقي، تحليل وسائل التواصل الاجتماعي، المدونات الحاسوبية.

١. مقدمة

١.١. مشكلة البحث وحدوده

انتبهت الحكومات إلى أهمية التعامل مع البيانات الضخمة (Big Data) التي تحويها وسائل التواصل الاجتماعي بشكل عام وتويتر بشكل خاص؛ حيث أدركت أن النقاش الذي يدور على وسائل التواصل الاجتماعي يمثل وسيلة حية لاستطلاع رأي الجمهور ولمعرفة اتجاه الرأي العام. كما يمكن من خلال هذا النقاش معرفة ردود أفعال الجمهور تجاه القضايا السياسية والاجتماعية والاقتصادية. كذلك يعد النقاش الذي يدور على وسائل التواصل الاجتماعي والأنشطة التي ترتبط بهذه الوسائل من المؤشرات الأساسية لقياس

الأداء (Key Performance Indicators)، والتي يستخدمها متخذو القرار والجهات الحكومية للتأكد من تحقيق الأهداف الاستراتيجية لاستراتيجياتهم السياسية والاقتصادية والإعلامية. من أجل ذلك بدأت الحكومات تنتبه إلى أهمية اكتشاف المعلومات المهمة في النصوص المنشورة على وسائل التواصل الاجتماعي بشكل عام وتويتر بشكل خاص والتي تتحدث عن موضوعات / كلمات مفتاحية معينة في فترة زمنية معينة، ويشتمل ذلك - على سبيل المثال - اكتشاف الموضوعات المهيمنة وأبرز الكلمات المفتاحية التي تحدث عنها الجمهور والعلاقات التي تربط كلمات معينة بموضوعات معينة، لكن التحدي الذي يواجه اكتشاف المعلومات المهمة والأنماط السائدة في محتوى هذه النصوص يتمثل في القدرة على اكتشاف وانتزاع معرفة مهمة من نصوص حرة لا تسير وفق بنية منظمة (Unstructured Text) مكتوبة باللغة العربية، كالنصوص المنشورة في وسائل التواصل الاجتماعي؛ حيث يصعب تحقيق ذلك بشكل يدوي، وذلك بسبب الحجم الضخم لمجموعات البيانات المركبة المرتبطة بمشاركات وسائل التواصل الاجتماعي، والتي لا يمكن معالجتها باستخدام الوسائل اليدوية أو باستخدام تطبيقات معالجة البيانات التقليدية. ومن جهة أخرى، هنالك تحديات أخرى تتمثل في أن اكتشاف المعلومات المهمة والأنماط السائدة في محتوى هذه النصوص يتطلب معالجة تلك النصوص (Text Processing) للتمكن من تقسيم النص إلى كلمات (Tokenization) مع استبعاد الكلمات الوظيفية (Function Words) والإبقاء على الكلمات ذات المحتوى (Content Words)؛ لأهميتها في اكتشاف المعلومات المهمة في النص، كما يتطلب ذلك الأمر تجريد الكلمات المذكورة في نصوص وسائل التواصل الاجتماعي إلى صيغتها الصرفية الأساسية (Lemmatization) وهو ما يساهم في تحديد العلاقات بين الكلمات المترابطة صرفيًّا أو دلاليًّا مع اختلافها في البنية السطحية (Surface Structure)، كما يتطلب اكتشاف المعلومات المهمة والأنماط السائدة (Patterns) في محتوى هذه النصوص تمثيل مجموعات الوثائق في النص المراد تحليله بشكل كمي رقمي قابل للتحليل الإحصائي الحاسوبي.

في إطار الحدود الموضوعية لدراستنا، سنستخدم تقنية التنقيب في النصوص (Text Mining) للتنقيب في نصوص تغريدات تويتر التاريخية التي توفرها واجهة برمجة التطبيقات الخاصة بأرشيف تويتر الرسمي (Twitter Full Archive API)[1] من أجل اكتشاف الموضوعات المهيمنة (Topic Extraction) التي تحدث عنها الناخبون والمرشحون خلال فترة انتخابات مجلس الأمة في دولة الكويت في العام ٢٠٢٠، إضافة إلى اكتشاف العلاقات التي تربط كلمات معينة بموضوعات معينة واكتشاف الكلمات الأكثر أهمية في نصوص التغريدات والتي تمثل كلمات مفتاحية محورية (Keywords) في إطار موضوعات تلك التغريدات.

من حيث الحدود الزمانية والمكانية للدراسة، تغطي العينة النصية التي ستُطَبَّق عليها تقنية التنقيب في بيانات التغريدات التي نشرها الناخبون والمرشحون خلال فترة انتخابات مجلس الأمة في دولة الكويت خلال الفترة من بداية يوم ٢٦ أكتوبر ٢٠٢٠ وهو يوم فتح باب الترشح لانتخابات مجلس الأمة إلى نهاية يوم ٦ ديسمبر ٢٠٢٠ وهو يوم إعلان نتائج الانتخابات.

[1] Twitter. "Getting Started with Premium Search Tweets: Full-Archive API." *Developer Platform*, https://developer.twitter.com/en/docs/twitter-api/premium/search-api/quick-start/premium-full-archive. Accessed 20 January 2022.

في هذا السياق، توفر تقنية التنقيب في النصوص أداة فاعلة تمكننا من الوصول إلى اكتشاف وانتزاع معرفة هامة من نصوص حرة لا تسير وفق بنية منظمة (Unstructured Text) مكتوبة باللغة العربية مع التغلب على التحديات المذكورة أعلاه. في سياق حدود البحث، لن يتطرق البحث إلى التفاصيل الرياضية لطريقة احتساب قيم المقاييس الإحصائية المستخدمة في تقنيات التنقيب في النصوص، كما لن يتطرق البحث إلى التفاصيل والأسس الرياضية للمنهجيات المستخدمة في عملية التنقيب في النصوص ومنها منهجية تقسيم القيمة الفردية (Singular Value Decomposition).

٢.١. فرضية البحث

يمكن استخدام تقنية التنقيب في النصوص (Text Mining) لاكتشاف الموضوعات المهيمنة (Topic Extraction) وأبرز الكلمات المفتاحية التي تطرق لها الناخبون والمرشحون في الدوائر الانتخابية الخمس خلال فترة انتخابات مجلس الأمة في دولة الكويت في العام ٢٠٢٠ من خلال تحليل محتوى نصوص تغريداتهم على تويتر خلال تلك الفترة.

٣.١. منهج البحث وأدواته وإجراءاته

لتحديد الموضوعات المهيمنة وأبرز الكلمات المفتاحية التي تطرق لها الناخبون والمرشحون في الدوائر الانتخابية الخمس خلال فترة انتخابات مجلس الأمة في دولة الكويت في العام ٢٠٢٠ سنستخدم تقنية التنقيب في نصوص تغريدات تويتر التاريخية (Historic Tweets) التي احتوت على كلمات مفتاحية متعلقة بانتخابات مجلس الأمة والدوائر الانتخابية، ولتطبيق التنقيب في نصوص التغريدات سنقوم باستخدام نظام SAS Text Miner المتخصص في التنقيب في النصوص. تجدر الإشارة إلى أن نظام SAS Text Miner هو نظام متقدم متخصص في التنقيب في النصوص المكتوبة بلغات عديدة ومنها اللغة العربية وهو نظام مُستخدَم عالميا في جهات حكومية وتجارية عديدة.

خلال عملية التنقيب في النصوص، سنمر بعدد من المراحل وهي:

١- اختيار العينات النصية وجمعها من أرشيف تويتر الرسمي (Twitter Full Archive) عن طريق واجهة برمجة التطبيقات الخاصة بتويتر (Twitter Application Programming Interface (API)) باستخدام برنامج قمنا ببرمجته باستخدام لغة Python للبرمجة وأداة Searchtweets الخاصة بلغة Python وهي أداة تُستخدم في جمع التغريدات من تويتر.

٢- المعالجة الحاسوبية لنصوص التغريدات (Text Processing)

٣- استخدام نظام SAS Text Miner للتنقيب في النصوص المجموعة واكتشاف الموضوعات التي تحويها (Topic Extraction) عن طريق استخدام تقنية التحليل الدلالي الكامن (Latent Semantic Analysis (LSA)) وذلك من خلال تطبيق منهجية رياضية من منهجيات الجبر الخطي (Linear Algebra) تعرف باسم تقسيم القيمة الفردية (Singular Value Decomposition)، وتمكننا هذه المنهجية من فحص أنماط التواجد المشترك (Patterns of Co-occurrence) بين الكلمات الموجودة في مجموعة الوثائق النصية (Corpus of Text Documents).

٤- الوصول إلى النتائج

في سياق مرحلة اختيار العينات النصية، سنقوم بتحديد استراتيجية اختيار البيانات (Data Selection Strategy) والتي يتم من خلالها اختيار مصادر البيانات التي ستُجمَع منها عينة النصوص (Text Sample). ومن حيث نوع العينة، سيكون مصدر بياناتنا النصية هو تغريدات تويتر التاريخية التي توفرها واجهة برمجة التطبيقات الخاصة بأرشيف تويتر الرسمي (Twitter Full Archive API)، حيث ستكون العينة مكونة من التغريدات المكتوبة باللغة العربية والتي تحتوي على كلمات البحث (Search Terms) المتعلقة بانتخابات مجلس الأمة وهي الكلمات التالية:

- مجلس الأمة
- الدائرة الأولى
- الدائرة الثانية
- الدائرة الثالثة
- الدائرة الرابعة
- الدائرة الخامسة

في سياق المعالجة الحاسوبية للنصوص، بعد تحديد الوثائق النصية (وهي تغريدات تويتر التاريخية) وجمع تلك الوثائق من مصدر النصوص وقبل التنقيب فيها، نحتاج إلى معالجة نصوص هذه الوثائق لجعلها قابلة للتنقيب الحاسوبي فيها واستخلاص النتائج منها. تشتمل هذه المعالجة على عمليات أهمها:

١- تقسيم النص إلى كلمات (Tokenization) مع حذف كلمات الإيقاف (Stop Words)

٢- تجريد الكلمات إلى صيغتها الصرفية الأساسية (Lemmatization)

٣- إحصاء عدد مرات ذكر الكلمات المستخرجة من الوثائق النصية وحساب وزنها الإحصائي (Term Weight). وتجدر الإشارة إلى أن الوثائق النصية في حالتنا هنا هي التغريدات؛ حيث تعد كل تغريدة وثيقة نصية (Document) منفصلة.

٤- تمثيل علاقة الكلمات بالوثائق النصية على شكل مصفوفة تعرف باسم مصفوفة التواجد المشترك للكلمات وفقا للوثائق (Term-by-Document Co-occurrence Matrix) باستخدام تقنية نموذج فراغ المتجِهات (Vector Space Model) المستخدمة في مجال التنقيب في النصوص وذلك لكي نتمكن من تمثيل مجموعة الوثائق النصية بشكل كمي (Quantitative Representation).

في إطار تجريد الكلمات إلى صيغتها الصرفية الأساسية، تُعرَف الصيغة الصرفية الأساسية للكلمة (Lemma) بأنها أصغر صيغة للكلمة مُستخدَمة لغويا أي صيغة الكلمة دون وجود سوابق أو لواحق تصريفية أو اشتقاقية أو ضمائر متصلة بشرط أن تكون هذه الصيغة الصرفية مُستخدَمة لغويا (أي موجودة في معجم اللغة). تقابل هذه الصيغة في اللغة العربية صيغة الماضي المفرد المذكر الغائب للأفعال (مثل كَتَب) وصيغة المفرد المذكر النكرة للأسماء (كاتِب). وعلى سبيل المثال: كَتَب، استكتَب، كاتِب، مكتوب، كِتاب، مَكتَب تمثل صيغًا صرفية أساسية (Lemmas) لأفعال وأسماء مشتقة من جذر واحد (ك ت ب). ويساعد استخلاص الصيغة الصرفية الأساسية للكلمة في تحديد العلاقات بين الكلمات المترابطة صرفيًّا أو دلاليًّا مع اختلافها في البنية السطحية (Surface Structure). وبعد تقسيم النص إلى كلمات واستخلاص الصيغة الصرفية الأساسية للكلمات، نحصل على مجموعة من الكلمات التي ستُطَبَّق عليها عمليات إحصائية

في إطار التنقيب في مجموعة الوثائق النصية من أجل تحديد الموضوعات التي تحويها وأبرز الكلمات المفتاحية المستخدمة فيها.

٤.١. الدراسات السابقة

تعاني ساحة البحث العلمي من ندرة في الأبحاث المكتوبة باللغة العربية والتي تتناول موضوع التنقيب في النصوص العربية المنشورة على وسائل التواصل الاجتماعي، وفي هذا السياق، قدمَت دراسة محاولة لقياس مستوى التعصب الرياضي في التغريدات المنشورة على تويتر من خلال تحليل الآراء المنشورة وذلك باستخدام نموذج حاسوبي (Computational Model) لتحليل المحتوى[1]. كذلك قدم بحث آخر دراسة وصفية تحليلية تهدف إلى التأصيل النظري للتنقيب في النصوص وتحليل المشاعر في وسائل التواصل الاجتماعي والتعريف بأهم المفاهيم المستخدمة في هذا المجال. كما تحدث البحث عن أبرز الطرق والآليات المستخدمة في مجال التنقيب في نصوص وسائل التواصل الاجتماعي[2] . كما تناول بحث آخر إمكانية استخدام مواقع التواصل الاجتماعي كقاعدة بيانات لقياس الرأي العام. حيث أوضح البحث أهمية التنقيب في النصوص لاكتشاف الآراء (Opinion Mining) ودور ذلك في قياس الرأي العام من خلال استخدام المعلومات المتاحة على الإنترنت لمعرفة اتجاهات رأي الأفراد بشأن موضوعات محددة[3]. كما استخدمت دراسة أخرى تقنية تعلم الآلة (Machine Learning) لتحليل مشاعر الجمهور تجاه تفشي فيروس كورونا (Covid-19)، والتعرف على أهم الموضوعات السائدة في المناقشات المتعلقة بالفيروس على تويتر؛ حيث حللت الدراسة تغريدات جُمِعَت من تويتر خلال الفترة من ١ مارس ٢٠٢٠ إلى ٣٠ مايو ٢٠٢٠، وتوصلت الدراسة إلى أنه يمكن تقسيم الموضوعات المتعلقة بفيروس كورونا في التغريدات المجموعة إلى خمسة موضوعات تعبر عن مخاوف الجمهور، وهي: بيئة الرعاية الصحية، الدعم النفسي والعاطفي، اقتصاد الأعمال، التغير الاجتماعي، والتوتر والإجهاد النفسي[4]. كذلك تناولت دراسة أخرى تحليل المزاج العام (Sentiment Analysis) لتغريدات تويتر التي تحدثت عن الأخبار في العالم العربي، حيث استخدمت الدراسة مدونة نصية حاسوبية (Corpus) مكونة من تغريدات جُمعت من تويتر ووُسِمَت (Annotated) كلماتها بمعلومات المزاج العام، ثم استُخدِمَت تقنية تعلم الآلة لتدريب الحاسوب على المدونة النصية الموسومة بمعلومات المزاج العام لتمكينه من تحديد المزاج العام لنصوص أخرى متعلقة بالأخبار على

[1] الخزاعى، محمد رده، و حسن بن عواد السريحى. "تحليل الآراء على شبكات التواصل الاجتماعي: نموذج تطبيقي لقياس مستوى التعصب الرياضي في تويتر." *Cybrarians Journal*، البوابة العربية للمكتبات والمعلومات، ٢٠١٨، العدد ٥٠، الصفحات ١-٢١.

[2] الخليفى، طارق. "تنقيب بيانات وسائل التواصل الاجتماعي واستخداماته في البحوث الإعلامية: تحليل المشاعر نموذجا." *مجلة البحوث والدراسات الإعلامية*، المعهد الدولي العالي للإعلام بالشروق، ٢٠١٩، العدد ٨، الصفحات ٢٧٩ - ٣٥١.

[3] يوسف، ريهام سامي حسين. "مواقع التواصل الاجتماعي كقاعدة بيانات لقياس الرأي العام: الواقع والإشكاليات." *مجلة البحوث والدراسات الإعلامية*، المعهد الدولي العالي للإعلام بالشروق، ٢٠١٨، العدد ٦، الصفحات ١٩٣-٢١٥.

[4] خليل، حمزة السيد حمزة. "توظيف تطبيقات الذكاء الاصطناعي لتحليل مشاعر مستخدمي مواقع التواصل الاجتماعي في الوقت الفعلي لأزمة جائحة فيروس كورونا." *المجلة المصرية لبحوث الرأي العام*، جامعة القاهرة - كلية الإعلام - مركز بحوث الرأي العام، مصر، ٢٠٢١، مجلد ٢٠، العدد ٢، الصفحات ١٤٩-٢٠٢.

وسائل التواصل الاجتماعي[1]. وتناولت دراسة أخرى التعرف على خطاب الكراهية في وسائل التواصل الاجتماعي بناء على علاقته بالدين، العرق، الجنسية، والجنس؛ حيث استخدمت الدراسة مدونة نصية حاسوبية مكونة من تغريدات تشتمل على محتوى له علاقة بخطاب الكراهية جُمعت من تويتر ووُسِمَت كلماتها بمعلومات متعلقة بخطاب الكراهية، ثم استُخدِمَت تقنية تعلم الآلة لتدريب الحاسوب على المدونة النصية الموسومة لتمكينه من التعرف على نصوص أخرى متعلقة بخطاب الكراهية على وسائل التواصل الاجتماعي[2]. من جهة أخرى، قدمت دراسة أخرى نموذجا حاسوبيا للتعرف على المحتوي المنشور على إنستجرام باللغة العربية والذي يحتوي على تهديدات كالتعليقات (Comments) وذلك باستخدام خوارزمية الشبكات العصبية (Neural Networks). لتصميم هذا النموذج الحاسوبي، استَخدَمت الدراسة مدونة نصية حاسوبية مكونة من تعليقات منشورة على إنستجرام صُنِّفَت يدويا إلى تعليقات تحمل تهديدا وأخرى لا تحمل تهديدا، ثم استُخدِمَت تقنية تعلم الآلة لتدريب الحاسوب على المدونة النصية الموسومة لتمكينه من التعرف على مشاركات (Posts) إنستجرام الأخرى التي تحمل تهديدا[3]. كما قدم بحث آخر نظاما للتعرف على الخطاب الذي يحتوي على المحتوى المرتبط بالعنف والذي يُنشَر على تويتر وفيسبوك باللهجة الأردنية باستخدام تقنية تعلم الآلة [4]. كذلك استخدمت دراسة أخرى تقنية تعلم الآلة لتصميم نموذج حاسوبي للتعرف على المحتوى المرتبط بالتنمر والذي يُنشر باللغة العربية على وسائل التواصل الاجتماعي (Cyber-Bullying and Cyber-Harassment)[5]. قدمت دراسة أخرى أيضا نموذجا حاسوبيا للتعرف على خطاب الكراهية الموجه إلى الشخصيات السياسية في العالم العربي على وسائل التواصل الاجتماعي باستخدام تقنية تعلم الآلة[6]. كما استخدمت دراسة أخرى خوارزمية الشبكات العصبية لتصميم نموذج حاسوبي للتعرف على المحتوى المرتبط بالسخرية (Irony) والذي يُنشر باللغة العربية على وسائل التواصل الاجتماعي[7] . إضافة إلى ما سبق، استخدمت دراسة أخرى تقنية التنقيب في النصوص لاكتشاف أهم الموضوعات التي تناولتها تغريدات نُشِرَت باللغة العربية حول موضوع فيروس كورونا (Covid-19)؛ حيث استخدمت الدراسة تقنية التجميع الآلي للنصوص (Text Clustering) للتعرف على أهم

[1] Al-Laith, Ali, and Shahbaz, Muhammad. "Tracking sentiment towards news entities from Arabic news on social media." *Future Generation Computer Systems*, 2021, vol. 118, pp. 467-484.

[2] Alsafari, Safa, Sadaoui, Samira, and Mouhoub, Malek. "Hate and offensive speech detection on arabic social media." *Online Social Networks and Media*, 2020, vol. 19.

[3] AlAjlan, Shatha AbdulAziz, and Saudagar, Abdul Khader Jilani. "Machine learning approach for threat detection on social media posts containing Arabic text." *Evolutionary Intelligence*, 2021, vol. 14, no. 2, pp. 811-822.

[4] Khalafat, Monther, et al. "Violence Detection over Online Social Networks: An Arabic Sentiment Analysis Approach." *International Journal of Interactive Mobile Technologies*, 2021, vol. 15, no. 14.

[5] Kanan, Tarek, Aldaaja, Amal, and Hawashin, Bilal. "Cyber-Bullying and Cyber-Harassment Detection Using Supervised Machine Learning Techniques in Arabic Social Media Contents." *Journal of Internet Technology*, 2020, vol. 21, no. 5, pp.1409-1421.

[6] Guellil, Imane, et al. "Detecting hate speech against politicians in Arabic community on social media." *International Journal of Web Information Systems*, 2020.

[7] Allaith, Ali, Shahbaz, Muhammad, and Alkoli, Mohammed. "Neural Network Approach for Irony Detection from Arabic Text on Social Media." *FIRE (Working Notes)*, 2019.

الموضوعات في تلك النصوص[1]. كذلك استخدمت دراسة أخرى تقنية تعلم الآلة لاكتشاف أهم الموضوعات التي تناولتها تغريدات تويتر التي نشرها المغردون في المملكة العربية السعودية خلال الفترة من ١ فبراير ٢٠٢٠ إلى ١ يونيو ٢٠٢٠ والتي تحدثت عن إجراءات الحكومة السعودية لمواجهة جائحة كورونا (Covid-19)، وتحدثت أيضا عن الاهتمامات التي تشكل قلقا للجمهور خلال هذه الجائحة مثل الاستدامة الاقتصادية والاجتماعية[2].

١.٥. مفاهيم أساسية

يقدم هذا القسم مفاهيم أساسية سيتناولها البحث في مجال علم اللغة الحاسوبي:

علم اللغة الحاسوبي (Computational Linguistics): هو استخدام الحاسوب في دراسة علم اللغة وتطوير برمجيات يمكنها القيام بعمليات لغوية مثل التحليل النحوي الآلي (Miller and Brown, 2013).

المعالجة الحاسوبية للغة الطبيعية (Natural Language Processing): تطلق على الأنظمة التي تستخدم لتحقيق التفاعل بين الحاسوب والبشر (Miller and Brown, 2013).

التنقيب في النصوص (Text Mining): في مجال علم اللغة الحاسوبي، يُعرف التنقيب في النصوص بأنه عملية انتزاع معلومات من نص معين (Miller and Brown, 2013).

الكلمة الوظيفية (Function Word): هي الكلمة التي يكون لها دور نحوي (Miller and Brown, 2013).

الكلمة ذات المحتوى (Content Word): هي الكلمة التي يكون لها معنى في المعجم (Matthews, 2014).

التواجد المشترك (Co-occurrence): هو العلاقة بين عنصرين أو أكثر يمكنهما أن يظهرا معا في نفس الوحدة كما هو الحال في ظهور كلمتين في عبارة واحدة (Miller and Brown, 2013).

مدونة (Corpus): مجموعة من النصوص المخزنة أساسا بشكل ورقي، ولكنها حاليا تُخزَّن بشكل رقمي (Miller and Brown, 2013).

بنية سطحية (Surface Structure): هو تمثيل البنية التركيبية للكلمة بعد تطبيق كل التحويلات (التغييرات) عليها (Miller and Brown, 2013).

[1] Jafarian, Hamoon. "Topic Discovery on Farsi, English, French, and Arabic Tweets Related to COVID-19 Using Text Mining Techniques." *Navigating Healthcare Through Challenging Times*, Edited by D. Hayn et al., IT Austrian Institute of Technology and IOS Press, 2021, vol. 26.

[2] Alomari, Ebtesam, et al. "COVID-19: Detecting government pandemic measures and public concerns from Twitter arabic data using distributed machine learning." *International Journal of Environmental Research and Public Health*, 2021,vol. 18, no. 1 , p. 282.

إرجاع الكلمة إلى صيغتها الصرفية الأساسية (Lemmatization): هو إرجاع الكلمة الموجودة في مدونة إلى وحدتها المعجمية الأساسية (Lexeme) (Miller and Brown, 2013).

الوحدة المعجمية الأساسية (Lexeme): هي أصغر وحدة مميَّزة في النظام الدلالي للغة معينة (Crystal, 2011).

التصنيف الدلالي (Ontology/Taxonomy): في مجال علم اللغة الحاسوبي، هو منظومة تُستخدم لتنظيم المعلومات المتعلقة بمجموعة من المفاهيم والعلاقة بين تلك المفاهيم في إطار مجال معين. يُستخدم التصنيف الدلالي في تطبيقات الذكاء الاصطناعي وغيرها من التطبيقات باعتباره شكلا من أشكال التمثيل المعرفي (Miller and Brown, 2013).

تحليل النص (Text Parsing): هو العملية التي تقوم بإعطاء فئة نحوية معينة لكلمة معينة، كما تقوم هذه العملية بإعطاء بنية نحوية للجملة. في مجال علم اللغة الحاسوبي، تُجرى هذه العملية عن طريق برنامج حاسوبي للتحليل يعرف بالمحلل النحوي (Parser) (Matthews, 2014).

نهج التحليل الدلالي الكامن ((Latent Semantic Analysis (LSA): يَستخدم نهج التحليل الدلالي الكامن تقنيات إحصائية متنوعة للتعرف على الأبعاد الضمنية (Underlying Dimensionality) الموجودة في مجموعات البيانات النصية لكي يستنتج المحتوى النصي المشترك الذي يمكن تصنيفه تحت موضوع معين والذي يقود بدوره السلوك المُلاحَظ للنص (DeVille, Barry, and Bawa, Gurpreet Singh, 2021).

٢. التنقيب في النصوص

يُعرف التنقيب في النصوص (Text Mining) بأنه اكتشاف معرفة مهمة وانتزاعها من نصوص حرة، أي نصوص لا تسير وفق بنية منظمة (Unstructured Text) كالنصوص المنشورة في وسائل التواصل الاجتماعي ومواقع الشبكة العنكبوتية (World Wide Web).

ولتحقيق ذلك، تُستَخدَم أنواع عديدة من التمثيل المعرفي (Knowledge Representation) للمعلومات اللغوية نحصل عليها عن طريق استخدام المعجم الحاسوبي (Lexicon) إضافة إلى استخدام القوانين النحوية والمعلومات الدلالية كالتصنيف الدلالي (Ontology/Taxonomy) للكيانات والأحداث فضلا عن استخدام مكانز (Thesaurus) المترادفات والاختصارات.

يمثل التنقيب في النصوص مجالا بينيا (Interdisciplinary) حديثا يدمج أكثر من مجال أكاديمي أهمها علم الحاسوب (Computer Science)، علم اللغة الحاسوبي (Computational Linguistics)، استرجاع المعلومات (Information Retrieval)، التنقيب في البيانات (Data Mining)، تعلُّم الآلة (Machine Learning)، والإحصاء (Statistics).

تُستخدم تقنيات التنقيب في النصوص في المجال التجاري والحكومي والأكاديمي وذلك لأن أغلب المعلومات الرقمية المستخدمة عالميا مخزنة على شكل نصوص لا تسير وفق بنية منظمة مقارنة مع البيانات التي تسير

وفق بنية منظمة مثل البيانات الموجودة في قواعد البيانات[1]. وفي هذا السياق، أدى التطور في مجال البيانات الضخمة (Big Data) إلى إنتاج كميات هائلة من البيانات النصية. كذلك تسبب هذا التطور في إيجاد تطبيقات ومنصات تحليلية عديدة ولغات برمجة وأدوات برمجية وخوارزميات (Algotithms) متخصصة للتعامل مع هذا الكم الهائل من البيانات النصية. وتُعرف البيانات الضخمة بأنها مجموعات البيانات (Data Sets) المركبة الضخمة والتي لا يمكن معالجتها وتحليلها باستخدام الوسائل اليدوية أو باستخدام تطبيقات معالجة البيانات التقليدية. من أمثلة البيانات النصية الضخمة مشاركات وسائل التواصل الاجتماعي كالتغريدات ومشاركات المدونات (Blogs). تجدر الإشارة إلى أن من أبرز الأنظمة الحاسوبية المتخصصة المستخدمة في التنقيب في نصوص البيانات الضخمة نظام SAS Text Miner [2]. ومن جهة أخرى، هنالك لغات للبرمجة مستخدمة في هذا المجال أهمها لغات Python و R. كما أن هنالك أدوات برمجية تُستخدم أيضا في هذا المجال ومنها NLTK ، GATE ، Pandas ، وNumPy.

وفي سياق أهمية البيانات الضخمة، يقول المحلل السياسي جاري كينج: إن الجانب الثوري في البيانات الضخمة ليس حجم مجموعات البيانات، ولكن الجانب الثوري هو ما يستطيع الباحثون عمله الآن باستخدام هذه البيانات عن طريق الخوارزميات والأدوات البرمجية والتطبيقات المتخصصة في تحليل هذا النوع من البيانات، حيث أدى ذلك إلى زيادة في استخدام التحليل الكمي في المجال الأكاديمي والعلمي والصناعي والحكومي[3].

ومن جهة أخرى شهد العالم مؤخرا تطورا سريعا في قدرات الحاسوب والمتمثلة في القدرة على معالجة البيانات (Data Processing)، إضافة إلى التطور في تقنيات الذاكرة الضخمة المبنية على الحوسبة السحابية (Cloud-Based High Capacity Computer Memory). وقد ساعد ذلك في التركيز على المعالجة الحاسوبية للنصوص وتحليلها[4].

كما أدى التطور في البيانات الضخمة إلى توافر مصادر متنوعة من البيانات النصية الضخمة التي استُخدِمَت في أبحاث ومشاريع التنقيب في النصوص. ومن مصادر هذه البيانات المتعلقة بوسائل التواصل الاجتماعي أرشيف تويتر الرسمي (Twitter Full Archive) الذي يوفر أرشيفا كاملا لتغريدات مستخدمي تويتر منذ أول تغريدة نُشِرت على تويتر في العام ٢٠٠٦ إلى الآن[5].

---

[1] Ignatow, Gabe, and Mihalcea, Rada. "*Text Mining: A Guidebook for the Social Sciences.*" SAGE Publications, 2016.

[2] لمعرفة المزيد عن نظام SAS Text Miner يمكنكم الرجوع إلى:

Gouta, Chakraborty, Pagolu, Murali, and Garla, Satish. *Text mining and analysis: practical methods, examples, and case studies using SAS*, SAS Institute, 2014.

[3] Shaw, Jonathan. "Why 'Big Data' is a Big Deal." *Harvard Magazine*, Harvard University, 2014, https://www.harvardmagazine.com/2014/03/why-big-data-is-a-big-deal. Accessed 20 January 2022.

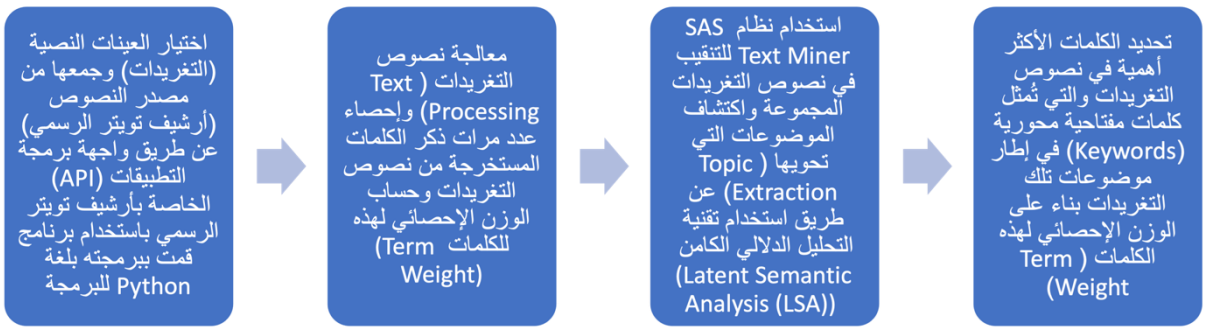
[4] DeVille, Barry, and Bawa, Gurpreet Singh. *Text as Data: Computational Methods of Understanding Written Expression Using SAS*, Wiley, 2021.

[5] Twitter. "Getting Started with Premium Search Tweets: Full-Archive API." *Developer Platform*, https://developer.twitter.com/en/docs/twitter-api/premium/search-api/quick-start/premium-full-archive. Accessed 20 January 2022.

وفي هذا السياق، تحدثَت دراسات عن أهمية استخدام تقنية التنقيب في النصوص المستخرجة من وسائل التواصل الاجتماعي (Social Media Mining)، حيث أشارت هذه الدراسات إلى أن البيانات المستخرجة من منصات وسائل التواصل الاجتماعي يمكن أن تقارن من حيث فائدتها (في تحليل اتجاهات الرأي) مع البيانات التي تقدمها وسائل جمع البيانات الأخرى مثل الاستبانات[1].

٣. تطبيق تقنية التنقيب في النصوص على تغريدات الناخبين والمرشحين خلال انتخابات مجلس الأمة في دولة الكويت في العام ٢٠٢٠

باستخدام تقنية التنقيب في النصوص، قمت بالتنقيب في نصوص تغريدات تويتر التي احتوت على كلمات متعلقة بانتخابات مجلس الأمة الكويتي وذلك لاكتشاف الموضوعات المهيمنة (Topic Extraction) التي تحدث عنها الناخبون والمرشحون خلال فترة انتخابات مجلس الأمة في دولة الكويت في العام ٢٠٢٠، إضافة إلى اكتشاف العلاقات التي تربط كلمات معينة بموضوعات معينة واكتشاف الكلمات الأكثر أهمية في نصوص التغريدات والتي تمثل كلمات مفتاحية محورية (Keywords) في إطار موضوعات تلك التغريدات. ولتحقيق ذلك قمت باتباع الخطوات المبينة في الشكل الآتي:



شكل رقم (١): خطوات التنقيب في نصوص تغريدات تويتر

قمتُ بجمع التغريدات المنشورة خلال الفترة من بداية يوم ٢٦ أكتوبر ٢٠٢٠ وهو يوم فتح باب الترشح لانتخابات مجلس الأمة إلى نهاية يوم ٦ ديسمبر ٢٠٢٠ وهو يوم إعلان نتائج الانتخابات. بلغ عدد التغريدات المجموعة ٤٧٨١٠٣ تغريدة مقسمة إلى ١٣١٧١٣ تغريدة منشورة (Posts) شكلت نسبة ٢٨٪ من المجموع الكلي للتغريدات و ٣٤٦٣٩٠ إعادة تغريد (Retweets) شكلت نسبة ٧٢٪ من المجموع الكلي للتغريدات. تجدر الإشارة إلى أن التنقيب في النصوص اقتصر على التغريدات المنشورة (Post) فقط (١٣١٧١٣ تغريدة منشورة) وذلك لتجنب تحليل إعادة التغريد (Retweet) والذي كثيرا ما يكون صادرا من حسابات آلية (Bots) وهو ما قد يؤثر على دقة وسلامة نتائج التنقيب.

[1] Brendan, O'Connor, et al. "From tweets to polls: Linking text sentiment to public opinion time series." *Fourth international AAAI conference on weblogs and social media*, 2010.

Ahmed, Hassan, Qazvinian, Vahed, and Radev, Dragomir. "What's with the attitude? identifying sentences with attitude in online discussions." *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.

كما ذكرنا في قسم منهج البحث وأدواته، بعد جمع البيانات النصية من أرشيف تويتر الرسمي، وقبل التنقيب فيها، نحتاج إلى معالجة نصوص هذه الوثائق لجعلها قابلة للتنقيب الحاسوبي فيها واستخلاص النتائج منها. في هذا السياق، استخدَمتُ نظام SAS Text Miner لتطبيق المعالجة الحاسوبية لنصوص التغريدات. وتشتمل هذه المعالجة على عمليات أهمها:

١- تقسيم النص إلى كلمات (Tokenization) مع حذف كلمات الإيقاف (Stop Words)

٢- تجريد الكلمات إلى صيغتها الصرفية الأساسية (Lemmatization)

٣- إحصاء عدد مرات ذكر الكلمات المستخرجة من الوثائق النصية وحساب وزنها الإحصائي (Term Weight)

١.٣. تقسيم النص إلى كلمات وتجريد الكلمات إلى صيغتها الصرفية الأساسية

في عملية تقسيم النص إلى كلمات (Tokenization) يتعرف نظام التنقيب في النصوص على الكلمات في مجموعة الوثائق النصية (التغريدات في حالتنا هنا) باعتبار أن المسافات وعلامات الترقيم هي حدود فاصلة بين الكلمات (تُعرف الكلمة في سياق نظام SAS Text Miner بمصطلح Term). كذلك يقوم النظام في هذه المرحلة بإحصاء عدد مرات ذكر الكلمات في التغريدات التي جمعناها (Frequency of Occurrence) لتحديد عدد مرات ذكر كل كلمة في نصوص تلك التغريدات (Term Frequency) وعدد التغريدات (الوثائق) التي احتوت على كل كلمة (Document Frequency). كما يقوم النظام أيضا في هذه العملية بحذف أدوات الترقيم الملتصقة بالكلمات. على سبيل المثال، في تتابع الكلمات التالي "هذا كتاب مفيد!"، يقوم النظام بتقسيم هذا التتابع إلى الكلمات التالية: هذا، كتاب، مفيد. نلاحظ هنا أن النظام قام بتقطيع النص إلى كلمات منفصلة مع حذف علامة التعجب الملتصقة بالكلمة الأخيرة. وتُنتِج هذه العملية من مجموعة الوثائق النصية مجموعة من الكلمات (Tokens) يمكن استخدامها باعتبارها مُدخلا (Input) لعملية أخرى وهي عملية تنقية مجموعة كلمات الوثائق النصية (Text Filtering) والتي تشتمل على عملية حذف كلمات الإيقاف (Stop Words). وتهدف عملية حذف كلمات الإيقاف إلى تنقية مجموعة الكلمات المستخرجة من مجموعة الوثائق النصية والتي حصلنا عليها بعد تقسيم الوثائق النصية إلى كلمات (Tokenization) عن طريق حذف الكلمات الوظيفية (Function Words) كحروف الجر والضمائر المنفصلة والظروف وغيرها من الكلمات الوظيفية التي تظهر بشكل كبير في النصوص وتُعرف في مجال التنقيب في النصوص بكلمات الإيقاف (Stop Words). ولتحقيق ذلك، استخدَمتُ قائمة من كلمات الإيقاف العربية التي قمت بإعدادها وقمت بإدخال هذه القائمة إلى النظام لكي يرجع النظام إليها من أجل تنقية الكلمات المستخرجة من مجموعة الوثائق النصية. يقوم نظام التنقيب في النصوص بتنقية مجموعة الكلمات المستخرجة عن طريق حذف الكلمات الموجودة في قائمة كلمات الإيقاف من هذه المجموعة لتبقى في المجموعة الكلمات ذات المحتوى (Content Words) فقط وهي الكلمات ذات الأهمية في تحليل النصوص كالأسماء والأفعال. بعد ذلك قمت باستخدام نظام SAS Text Miner لتجريد الكلمات إلى صيغتها الصرفية الأساسية (Lemmatization). الصيغة الصرفية الأساسية للكلمة (Lemma) هي أصغر صيغة للكلمة مُستخدَمة لغويا أي صيغة الكلمة دون وجود سوابق أو لواحق تصريفية أو اشتقاقية أو ضمائر متصلة بشرط أن تكون هذه الصيغة الصرفية مُستخدَمة لغويا (أي موجودة في معجم اللغة). تقابل هذه الصيغة في اللغة العربية صيغة الماضي المفرد المذكر الغائب للأفعال (مثل كَتَب) وصيغة المفرد المذكر النكرة للأسماء (مثل كاتِب).

وتجدر الإشارة إلى أن فائدة تجريد الكلمات إلى صيغتها الصرفية الأساسية تتجلى في تحقيق تسوية (Normalization) للكلمات وذلك لتقليل التعقيد الذي يواجهه الحاسوب عند تنقيبه في النص وذلك من خلال تقليل عدد أشكال الكلمات المميزة (Unique Terms) في الوثائق النصية عن طريق تجريد الكلمات المصرفة والمشتقة إلى صيغتها الصرفية الأساسية المميزة[1]. وتجدر الإشارة هنا إلى أنني قد عملت مع فريق البحث والتطوير (Research and Development) في شركة SAS الأمريكية لتحسين قدرات نظام SAS Text Miner على تحليل نصوص اللغة العربية بشكل عام وقدراته على التعامل مع التحليل الصرفي لبنية الكلمة العربية واستخلاص صيغها الصرفية الأساسية (Lemmatization) بشكل خاص من خلال تطوير المعجم الحاسوبي (Lexicon) الخاص باللغة العربية الذي يستخدمه نظام SAS Text Miner لاستخلاص الصيغة الصرفية الأساسية للكلمات. حيث إن النظام يقوم باستخلاص الصيغة الصرفية الأساسية للكلمات المصرفة والمشتقة باستخدام تقنية مبنية على المعجم (Dictionary-Based Stemmer) لاستخلاص الكلمات المميزة في النص[2].

ومن أجل تقسيم الوثائق النصية إلى كلمات وتجريد الكلمات إلى صيغتها الصرفية الأساسية، قمت باستخدام أداة تحليل النص (Text Parsing) في نظام SAS Text Miner. وفي هذه المرحلة، قمت بتحديد أن النص المراد تحليله هو نص باللغة العربية وذلك لكي يستبعد النظام النصوص المكتوبة باللغة الإنجليزية ويستبعد أسماء حسابات تويتر (Usernames) التي تُكتب بحروف إنجليزية. كذلك حددنا أن الكلمات التي يجب أن يقوم النظام بتحليلها هي كلمات مكونة من حروف (Alphabets) ولا تحوي رموزا مثل # أو @. بعد انتهاء أداة تحليل النص (Text Parsing) من عملها، نحصل من مجموعة الوثائق النصية على مجموعة من الكلمات (Tokens) التي سنستخدمها باعتبارها مُدخَلا (Input) لعملية أخرى وهي عملية تنقية مجموعة كلمات الوثائق النصية باستخدام أداة تنقية النص (Text Filtering) في نظام SAS Text Miner.

٢.٣. تنقية كلمات النص المستخرجة وإحصاء عدد مرات ذكر الكلمات المستخرجة وحساب وزنها الإحصائي

بعد استخدام أداة تحليل النص (Text Parsing)، قمنا باستخدام أداة تنقية النص (Text Filtering) في نظام SAS Text Miner. تُمكننا هذه الوظيفة من تنقية الكلمات المستخرجة من مجموعة الوثائق النصية وذلك لتقليل عدد الكلمات التي سيستخدمها النظام خلال تنقيبه في نصوص الوثائق النصية، حيث يحتفظ النظام فقط بالكلمات التي يحدد أنها كلمات مهمة من خلال ما يعرف بوزن الكلمة الإحصائي (Term Weight). ولتحقيق ذلك، تَستبعد أداة تنقية النص (Text Filtering) الكلمات المستخرَجة التي تحمل وزنا إحصائيا منخفضا (Low Weight Terms)، وبذلك لا يستخدم النظام تلك الكلمات المستبعَدة في عملية التنقيب في نصوص الوثائق.

[1]Gouta, Chakraborty, Pagolu, Murali, and Garla, Satish. *Text mining and analysis: practical methods, examples, and case studies using SAS*, SAS Institute, 2014.

[2] المصدر السابق.

تعرض الشاشة التالية نتائج استخدام أداة تنقية النص (Text Filtering)، حيث تعرض الشاشة إحصاءات بعض الكلمات التي استُخرِجَت من نصوص التغريدات، ويَظهَر لنا في الشاشة عدد مرات ذكر كل كلمة في نصوص التغريدات (FREQ) وعدد التغريدات (الوثائق النصية) التي احتوت على هذه الكلمة (#DOCS). العمود الذي يحمل عنوان "KEEP" يبين ما إذا كانت الكلمة من الكلمات التي سيستخدمها النظام في التنقيب في النصوص كالكلمات ذات المحتوى التي لا تحمل وزنا إحصائيا منخفضا (Low Weight Terms) (وفي هذه الحالة تظهر بجانبها علامة صح). أما إذا كانت الكلمة من كلمات الإيقاف (Stop Words) أو كانت من الكلمات ذات المحتوى التي تحمل وزنا إحصائيا منخفضا فلا تظهر بجانبها علامة صح وبالتالي لن تُستَخدَم في التنقيب في النصوص كما هو الحال في كلمتي "بعد" و "أن" في الشاشة المعروضة:

Terms

TERM	FREQ	# DOCS ▼	KEEP	WEIGHT	...	ATTRIBUTE
انتخاب	5134	4776	✓	0.279		Alpha
مواطن	5283	4722	✓	0.282		Alpha
دكتور	5052	4574	✓	0.284		Alpha
اختيار	5062	4495	✓	0.286		Alpha
مطير	5318	4492	✓	0.288		Alpha
بعد	4713	4486		0.0		Alpha
نتائج	4506	4313	✓	0.287		Alpha
نائب	5469	4295	✓	0.297		Alpha
قال	4878	4247	✓	0.292		Alpha
أن	4646	4211		0.0		Alpha
توفيق	4176	4029	✓	0.293		Alpha
رئيس	4474	3936	✓	0.298		Alpha
فساد	4398	3898	✓	0.298		Alpha
ثان	3967	3822	✓	0.298		Alpha

شكل رقم (٢): شاشة تبين نتائج استخدام أداة تنقية النص (Text Filtering) في نظام SAS Text Miner

وجود علامة "+" أمام الكلمة يعني أن هذه الكلمة هي صيغة صرفية أساسية (Lemma) يمكن أن تكون لها صيغ مصرفة أو مشتقة أخرى مذكورة في نصوص التغريدات كما هو الحال في الصيغة الصرفية الأساسية "مواطن" التي تُظهِر الشاشة التالية الكلمات المصرفة والمشتقة منها والموجودة في نصوص التغريدات:

Terms

	TERM	FREQ	# DOCS ▼	KEEP	WEIGHT	...	ATTRIBUTE
⊞	انتخاب	5134	4776	✓	0.279		Alpha
⊟	مواطن	5283	4722	✓	0.282		Alpha
	مواطني	107	107				Alpha
	مواطنون	141	141				Alpha
	مواطنا	11	11				Alpha
	مواطن	3308	3008				Alpha
	مواطنين	1716	1651				Alpha
⊞	دكتور	5052	4574	✓	0.284		Alpha
⊞	اختيار	5062	4495	✓	0.286		Alpha
⊞	مطير	5318	4492	✓	0.288		Alpha
⊞	بعد	4713	4486		0.0		Alpha
⊞	نتائج	4506	4313	✓	0.287		Alpha
⊞	نائب	5469	4295	✓	0.297		Alpha
⊞	قال	4878	4247	✓	0.292		Alpha
⊞	أن	4646	4211		0.0		Alpha
⊞	توفيق	4176	4029	✓	0.293		Alpha

شكل رقم (٣): شاشة تبين نتائج استخدام أداة تنقية النص (Text Filtering) وتعرض الكلمات المصرفة والمشتقة من إحدى الصيغ الصرفية الأساسية الموجودة في نصوص التغريدات

كذلك يمكننا استخدام شاشة أداة تنقية النص (Text Filtering) لكي نَستبعد يدويا (Drop) الكلمات التي نرى أنها ليست ذات فائدة في التنقيب في النصوص وبالتالي لن يستخدمها النظام في التنقيب في نصوص التغريدات. إضافة إلى ذلك، تبين شاشة أداة تنقية النص أيضا الوزن الإحصائي للكلمة (Term Weight). يُستخدم الوزن الإحصائي للكلمة لمعرفة أهمية الكلمة بناء على عدد مرات ذكرها في كل وثيقة وكيفية توزيع الكلمة على الوثائق الموجودة في مجموعة الوثائق النصية (التغريدات). وفي هذا السياق، يعطي النظام الكلمة التي تتكرر بشكل أكبر في عدد قليل نسبيا من الوثائق وزنا أكبر. بينما يعطي النظام الكلمات التي ترد في عدد أكبر من الوثائق أو في كل وثيقة في مجموعة الوثائق النصية وزنا أقل وذلك لأنها كلمات لا تساعد على التمييز بين الوثائق بشكل واضح (Document Discrimination)[1]. لذلك فالكلمات التي تساعد على تصنيف موضوعات الوثائق هي تلك الكلمات التي تَظهَر في وثائق أقل ولكنها تتكرر بشكل أكبر في تلك الوثائق القليلة[2]. لتحديد وزن الكلمة الإحصائي بالنسبة لمجموعة الوثائق النصية، يستخدم نظام SAS Text Miner مقياسا مشتقا من مقياس الانتروبيا (Entropy Measure)[3] المستخدم في نظرية المعلومات

[1] Ulrich, Reincke. "Profiling and classification of scientific documents with SAS Text Miner." 2003.

[2] SAS Institute. "Term Weighting." *SAS Help Center*, https://documentation.sas.com/doc/en/tmref/15.2/p06w2zv74gep4zn135roo9r7xdxk.htm. Accessed 20 Jan. 2022.

[3] لا يتسع المقام هنا للحديث عن تفاصيل مقياس الانتروبيا (Entropy Measure) وأسسه الرياضية. لمعرفة المزيد عن استخدام مقياس الانتروبيا في التنقيب في النصوص، يمكنكم الرجوع إلى:
Gouta, Chakraborty, Pagolu, Murali, and Garla, Satish. *Text mining and analysis: practical methods, examples, and case studies using SAS*, SAS Institute, 2014.

(Information Theory). وفي هذا المقياس يذهب الوزن الإحصائي الأكبر للكلمات التي تَظهَر بشكل أقل في مجموعة الوثائق النصية ككل، إلا أن هذه الكلمات تتكرر بشكل أكبر في وثائق قليلة في هذه المجموعة.

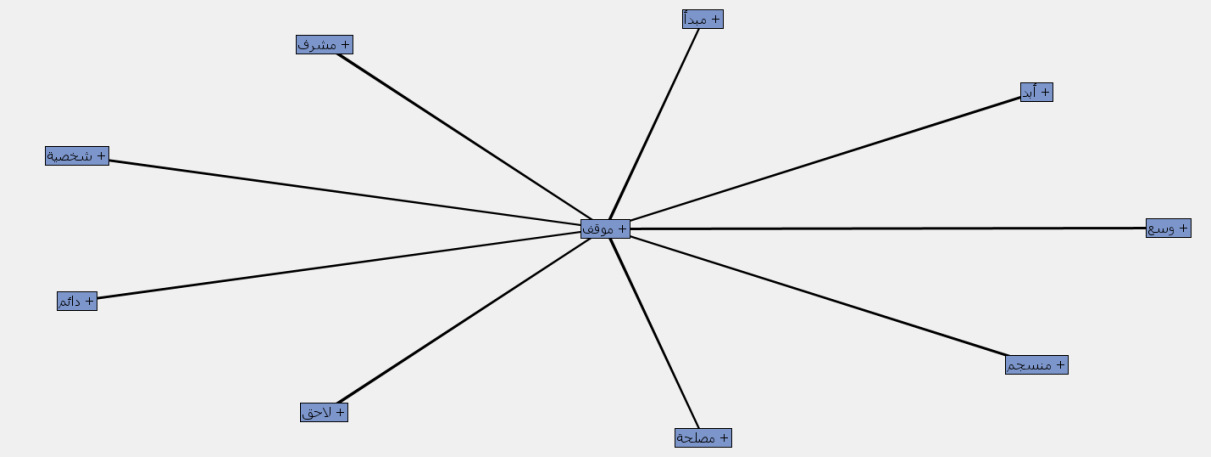
تجدر الإشارة إلى أن نظام SAS Text Miner يقوم من خلال أداة تنقية النص (Text Filtering) بتحديد علاقة كل كلمة من الكلمات الموجودة في مجموعة الكلمات المستخرجة من الوثائق النصية بالكلمات الأخرى في هذه المجموعة وذلك من خلال تقنية ربط المفاهيم (Concept Linking) التي تبين مدى علاقة الكلمة بكلمات أخرى بناء على تواجدهما المشترك (Co-occurrence) في وثيقة واحدة من الوثائق النصية التي يتم التنقيب فيها. يُعبر نظام SAS Text Miner عن هذه العلاقة باستخدام بنية بيانات تعرف باسم Hub-and-Spoke Structure والتي تعرضها شاشة ربط المفاهيم في أداة تنقية النص (Text Filtering) على شكل رسم شجري تفريعي تفاعلي (Interactive Hyperbolic Tree Graph). في هذا الرسم، يمكننا اختيار كلمة من الكلمات الموجودة في مجموعة الكلمات المستخرجة من الوثائق النصية لتَظهَر هذه الكلمة الأساسية في منتصف الرسم متصلة بالكلمات المرتبطة بها بشكل كبير (Most Highly Associated Terms). وفي هذا الرسم، يمكننا نقر الأيقونة التي تمثل أي كلمة من الكلمات المتصلة بالكلمة الأساسية ليتسع الرسم ويعرض تفاصيل الكلمات المرتبطة بتلك الكلمة التي قمنا بالنقر على أيقونتها. يُعبر هذا الرسم عن قوة الترابط (Strength of Association) بين الكلمة الأساسية الموجودة في منتصف الرسم والكلمات المتصلة بها من خلال سماكة الخط الذي يصل بين تلك الكلمة الأساسية والكلمات المرتبطة بها، حيث يمثل الخط الأكثر سماكة ترابطا أكثر قوة بين الكلمتين (أي أن الكلمة الأساسية ظهرت مع الكلمة المرتبطة بها في عدد أكبر من الوثائق). يتم احتساب قوة الترابط (Strength of Association) بين كلمتين باستخدام مقياس التوزيع الثنائي (Binomial Distribution)[1].

يبين الشكل التالي مثالا على شاشة ربط المفاهيم التي تعرض رسما شجريا تفريعيا يمثل بشكل مرئي بنية Hub-and-Spoke Structure التي تعبر عن علاقة كلمة معينة بكلمات أخرى بناء على تواجدهما المشترك (Co-Occurrence) في مجموعة الوثائق النصية:

[1] لا يتسع المقام هنا للحديث عن تفاصيل مقياس التوزيع الثنائي (Binomial Distribution). للتعرف على طريقة احتساب قيمة مقياس التوزيع الثنائي (Binomial Distribution) وصيغته الرياضية يمكنكم الرجوع إلى:

Gouta, Chakraborty, Pagolu, Murali, and Garla, Satish. *Text mining and analysis: practical methods, examples, and case studies using SAS*, SAS Institute, 2014.

SAS Institute. "Strength of Association for Concept Linking." *SAS Help Center*, https://documentation.sas.com/doc/en/tmref/15.2/n0chdwsd64uafcn164shwz4793a2.htm. Accessed 20 January 2022.



شكل رقم (٤): مثال على شاشة ربط المفاهيم (Concept Linking)

تجدر الإشارة إلى أن بيانات ربط المفاهيم (Concept Linking) التي تحدد علاقة كلمة بكلمات أخرى بناء على تواجدهما المشترك في وثيقة واحدة في مجموعة الوثائق النصية تساعد نظام SAS Text Miner على اكتشاف الموضوعات التي تحويها مجموعة الوثائق النصية (Topic Extraction).

٣.٣. التنقيب في نصوص التغريدات

بعد الانتهاء من المعالجة الحاسوبية لنصوص التغريدات (Text Processing)، استخدَمتُ نظام SAS Text Miner للتنقيب في نصوص هذه التغريدات واكتشاف الموضوعات التي تحويها (Topic Extraction) عن طريق استخدام تقنية التحليل الدلالي الكامن (Latent Semantic Analysis (LSA)) وذلك من خلال تطبيق منهجية رياضية من منهجيات الجبر الخطي (Linear Algebra) تُعرف باسم تقسيم القيمة الفردية (Singular Value Decomposition). تُمَكِّنُنا هذه المنهجية من تحديد أنماط التواجد المشترك (Patterns of Co-occurrence) بين الكلمات الموجودة في مجموعة الوثائق النصية (Corpus of Text Documents).

يَستخدم نظام SAS Text Miner نموذج فراغ المُتجِهات (Vector Space Model) للتنقيب في النصوص[1] وذلك لكي يتمكن النظام من تمثيل مجموعات الوثائق في النص المراد تحليله بشكل كمي (Quantitative). في هذا السياق، يُعرف المُتجِه (Vector) في مصطلحات الرياضيات بأنه سهم يتجه من نقطة إلى أخرى. مجموعات الوثائق يمكن أن تَكون مجموعة وثائق نصية مخزنة بصيغة Microsoft Word مثلا، ويمكن أن تكون هذه الوثائق مجموعة تغريدات منشورة على تويتر بحيث تمثل كل تغريدة وثيقة نصية مستقلة. في هذا النموذج، يتم تمثيل الوثائق الموجودة في مجموعة الوثائق النصية على شكل متجِهات (Vectors) بحيث يعبر كل متجِه عن عدد مرات ذكر الكلمات (Terms) التي تمت فهرستها (Indexed) في كل وثيقة من تلك الوثائق. تشير المتجِهات غالبا إلى اتجاهات عديدة متناثرة وذلك لأن عددا

[1] لا يتسع المقام هنا للحديث عن تفاصيل نموذج فراغ المتجِهات (Vector Space Model) وأسسه الرياضية. للمزيد من المعلومات عن نموذج فراغ المتجِهات المُستخدَم بشكل كبير في مجال التنقيب في النصوص واسترجاع المعلومات (Information Retrieval)، يمكنكم الرجوع إلى:
Gerard, Salton, and McGill, Michael J. *Introduction to modern information retrieval*. McGraw Hill, 1983.

قليلا من الكلمات المفهرَسة يمكن أن يوجد في وثيقة واحدة من مجموعة الوثائق[1]. لتوضيح هذه الفكرة، يعرض الشكل التالي رسم فضاء متجِهات مصغر يمثل علاقة كلمتين من الكلمات المفهرسة بثلاثة وثائق ذُكِرت فيها هاتان الكلمتان وفقا لعدد مرات ذكر كل كلمة في كل وثيقة، وكل متجِه في رسم فضاء المتجِهات يمثل عدد مرات ذكر كلمة معينة في وثيقة معينة:

شكل رقم (٥): تمثيل علاقة الكلمات المفهرسة بالوثائق النصية على شكل رسم فضاء متجهات

في هذا الرسم يتضح لنا أن الكلمة ٢ ذُكرت في الوثيقة ١ أكثر من ذكر الكلمة ١ في هذه الوثيقة، وهذا الذي يفسر اقتراب المتجِه الأحمر (الذي يمثل عدد مرات ذكر الكلمات في الوثيقة ١) من الكلمة ٢ وابتعاده عن الكلمة ١. كما يبن لنا هذا الرسم أن الكلمة ١ ذُكرت في الوثيقة ٣ أكثر من ذكر كلمة ٢ في هذه الوثيقة، وهذا الذي يفسر اقتراب المتجِه الأخضر (الذي يمثل عدد مرات ذكر الكلمات في الوثيقة ٣) من الكلمة ١ وابتعاده عن الكلمة ٢.

وفي هذا السياق، يقوم نظام SAS Text Miner بتمثيل رسم فضاء المتجِهات المذكور أعلاه بشكل كمي على هيئة مصفوفة (Matrix)؛ حيث تُمَثَّل هذه المصفوفة على شكل جدول تعرض صفوفه (Rows) الكلمات المُفَهرَسَة وأعمدته (Columns) الوثائق، وتبين الأرقام المذكورة في المصفوفة عدد مرات ذكر كل كلمة من الكلمات المُفَهرَسَة في كل وثيقة من الوثائق الموجودة في مجموعة الوثائق النصية. المصفوفة التالية تمثل رسم فضاء المتجهات الذي عرضناه في الشكل (٥) أعلاه. في هذه المصفوفة نجد أن الكلمة ١ ذُكرت ٥٠٠ مرة في الوثيقة ١ و ٦٥٠٠ مرة في الوثيقة ٢ و ٩٢٠٠ مرة في الوثيقة ٣، بينما الكلمة ٢ ذكرت ٧٠٠٠ مرة في الوثيقة ١ و ٤٠٠٠ مرة في الوثيقة ٢ و ١٠٠٠ مرة في الوثيقة ٣.

جدول رقم (١): جدول يعرض تمثيل علاقة الكلمات بالوثائق بشكل كمي على هيئة مصفوفة التواجد المشترك للكلمات وفقا للوثائق (Term-by-Document Co-occurrence Matrix)

وثيقة ٣	وثيقة ٢	وثيقة ١	
٩٢٠٠	٦٥٠٠	٥٠٠	كلمة ١
١٠٠٠	٤٠٠٠	٧٠٠٠	كلمة ٢

[1] Ulrich, Reincke. *"Profiling and classification of scientific documents with SAS Text Miner."* 2003.

بعد قيام نظام التنقيب في النصوص بتمثيل علاقة الكلمات بالوثائق بشكل كمي (رقمي)، يمكنه القيام بالتنقيب في نصوص الوثائق الموجودة في مجموعة الوثائق النصية لتحليلها واكتشاف أهم الموضوعات التي تناولتها وتحديد الكلمات الأكثر أهمية في نصوص الوثائق والتي تمثل كلمات مفتاحية محورية (Keywords) في إطار موضوعات تلك الوثائق.

١.٣.٣. تحديد موضوعات النص

يعد مجال تحديد موضوعات النصوص (Text Topic Extraction) ويعرف أيضا بنمذجة الموضوعات (Topic Modeling) أحد أهم مجالات التنقيب في النصوص. وفي هذا المجال تُستخدَم تقنيات لنمذجة الموضوعات أهمها نهج التحليل الدلالي الكامن ((Latent Semantic Analysis (LSA)) [1]. كما أشرنا سابقا، يَستخدم نهج التحليل الدلالي الكامن تقنيات إحصائية متنوعة للتعرف على الأبعاد الضمنية (Underlying Dimensionality) الموجودة في مجموعات البيانات النصية لكي يستنتج المحتوى النصي المشترك الذي يمكن تصنيفه تحت موضوع معين والذي يقود بدوره السلوك المُلاحَظ للنص. في هذا السياق، سنستخدم تقنية التحليل الدلالي الكامن لنمذجة موضوعات النصوص لتحديد أهم الموضوعات التي تناولتها الوثائق النصية (التغريدات) وذلك باستخدام أداة تحديد موضوعات النص (Text Topic) في نظام SAS Text Miner.

كما أشرنا سابقا، يستخدم نظام SAS Text Miner نموذج فراغ المتجِهات (Vector Space Model) لتمثيل النص بشكل رقمي، وفي هذا الإطار، تُعرَف الكلمات المميَّزة (Distinct Terms) في مجموعة الوثائق النصية بمصطلح المتغيرات (Variables)، بينما تُعرَف الوثائق بمصطلح المواد المُلاحَظة (Observations). وفي معظم مجموعات النصوص، يكون عدد المتغيرات (الكلمات المميَّزة) اللازمة لتمثيل كل مادة مُلاحَظة (كل وثيقة) أكبر بكثير مما يمكن نمذجته بسهولة. ونتيجة لذلك، يصبح ما يُعرَف بتقليل الأبعاد (Dimension Reduction) في مجموعات النصوص جانبًا مهمًّا في أنظمة التنقيب في النصوص، ولتحقيق ذلك تُستخدَم تقنية التحليل الدلالي الكامن ((Latent Semantic Analysis (LSA) لتقليل مجموعات النصوص عن طريق استخدام منهجية رياضية من منهجيات الجبر الخطي (Linear Algebra) تُعرف باسم منهجية تقسيم القيمة الفردية (Singular Value Decomposition) [2] وهي منهجية مستخدمة في مجال التنقيب في النصوص وتُطبَّق على المصفوفات التي تمثل مجموعات الوثائق

[1] Deerwester, Scott, et al. "Indexing by latent semantic analysis." *Journal of the American society for information science*, 1990, vol. 41, no. 6, pp. 391-407.

[2] لا يتسع المقام هنا للحديث عن تفاصيل منهجية تقسيم القيمة الفردية (Singular Value Decomposition) وأسسها الرياضية. لمعرفة المزيد من التفاصيل عن هذه المنهجية واستخدامها في ضغط المصفوفات التي تمثل مجموعات الوثائق النصية خلال عملية التنقيب في النصوص، يمكنكم الرجوع إلى:

Albright, Russ. *"Taming Text with the SVD."* SAS Institute, 2004.

Sarma, Kattamuri S. "*Predictive modeling with SAS enterprise miner: Practical solutions for business applications"*. SAS Institute, 2017.

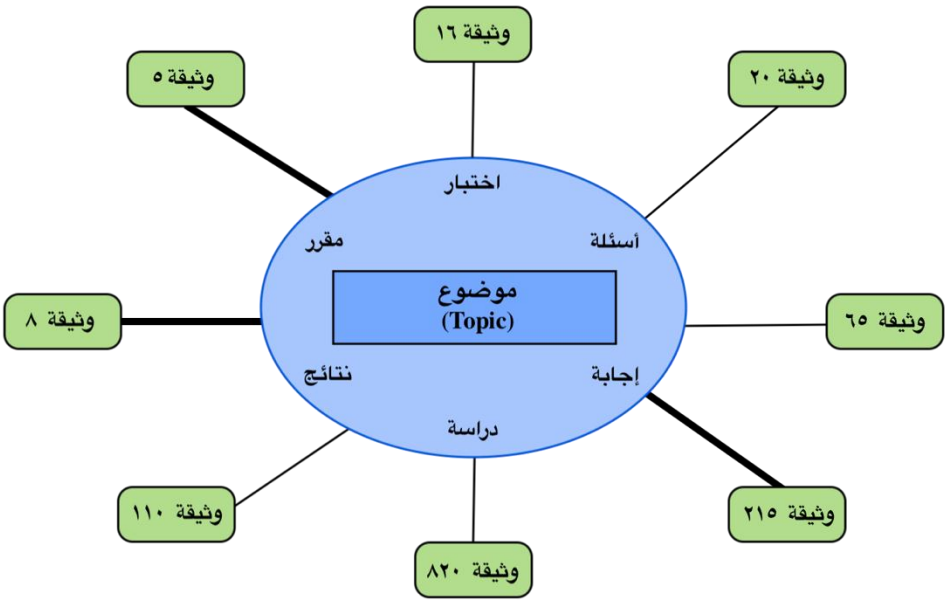
Gouta, Chakraborty, Pagolu, Murali, and Garla, Satish. *Text mining and analysis: practical methods, examples, and case studies using SAS*, SAS Institute, 2014.

النصية وذلك من خلال فحص أنماط التواجد المشترك للكلمات (Patterns of Co-occurrence) الموجودة في مجموعات الوثائق النصية. يتم ذلك عن طريق استخراج معلومات متناثرة من مصفوفة التواجد المشترك بين الكلمات والوثائق (Term-by-Document Co-occurrence Matrix) التي عرضناها سابقا في جدول (١) والتي تحدد علاقة الكلمات بالوثائق في مجموعة الوثائق النصية (Corpus of Text Documents) في إطار تطبيق نموذج فراغ المتجِهات (Vector Space Model) للتنقيب في النصوص والمستخدم لتمثيل مجموعات الوثائق النصية بشكل كمي (Quantitative Representation).

لتحديد موضوعات التغريدات، سنقوم باستخدام أداة تحديد موضوعات النص (Text Topic) في نظام SAS Text Miner. تقوم أداة تحديد موضوعات النص بتحديد الموضوعات المهمة المهيمنة التي تتحدث عنها الوثائق النصية وذلك باستخدام مجموعات الكلمات التي استُخرجت من هذه الوثائق النصية في المراحل السابقة وتم تحليلها وتنقيتها باستخدام أداتي تحليل النص (Text Parsing) وتنقية النص (Text Filtering). وفي هذا السياق، بعد استخدام أداة تنقية النص لتنقية مجموعة الكلمات المستخرجة من التغريدات واستبعاد الكلمات التي تحمل وزنا إحصائيا منخفضا (Low Weight Terms)، استَخدمتُ نتائج أداة تنقية النص باعتبارها مُدخلا (Input) لأداة تحديد موضوعات النص (Text Topic)، وتقوم أداة تحديد موضوعات النص بالتعرف على الكلمات التي تتواجد بشكل متكرر مع كلمات أخرى (Co-occur) في مجموعة الوثائق النصية التي يتم التنقيب فيها للتمكن من تحديد موضوعات هذه الوثائق، وفي سياق التنقيب في النصوص، يُعرَف الموضوع (Topic) بأنه مجموعة الكلمات (Terms) التي تَصف وتُميز فكرة رئيسة (Theme) في نص معين. بكلمات أخرى، يهدف تحديد موضوعات النص إلى التمكن من التعرف على مجموعات الكلمات ذات الأهمية في الوثائق النصية والتي تَصف وتُميز أفكارا رئيسة فيها بحيث يمكن اعتبار هذه الأفكار الرئيسة موضوعات مهيمنة في نصوص تلك الوثائق.

وفي إطار عمل أداة تحديد موضوعات النص (Text Topic)، تقوم الأداة بإعطاء درجة (Score) لكل وثيقة (تغريدة) ولكل كلمة في مجموعة الوثائق النصية (Corpus of Text Documents) التي يتم التنقيب فيها وذلك لبيان قوة الارتباط (Strength of Association) بين تلك الوثيقة أو الكلمة وموضوع معين (Topic). ثم تَستخدم الأداة ما يُعرف بعتبات القَطع (Cutoff Thresholds) لتحديد ما إذا كان الارتباط (Association) قويا بما يكفي لاعتبار أن الوثيقة أو الكلمة تنتمي إلى الموضوع. حيث تُستخدم قيمة Document Cutoff لتحديد أقل وزن للموضوع بالنسبة للوثيقة (Minimum Topic Weight for the Document) يجب أن تحمله هذه الوثيقة لكي يتم اعتبارها منتمية إلى موضوع معين (مرتبطة بهذه الموضوع). كما تُستخدم قيمة Term Cutoff لتحديد أقل وزن للموضوع بالنسبة للكلمة (Minimum Topic Weight for the Term) يجب أن تحمله هذه الكلمة لكي تُستخدم ككلمة تُحدِّد وتُمثل هذا الموضوع، ويتم احتساب وزن الموضوع بالنسبة للوثيقة (Topic Weight for the Document) ووزن الموضوع بالنسبة للكلمة (Topic Weight for the Term) بناء على عدد الموضوعات التي تستخرجها أداة تحديد موضوعات النص. وعلى سبيل المثال، إذا استخرَجَت أداة تحديد موضوعات النص ٥٠ موضوعا للنص، فسيكون هنالك ٥٠ وزنا للموضوع بالنسبة للوثيقة الواحدة و ٥٠ وزنا للموضوع بالنسبة للكلمة الواحدة (كل وزن يقابل موضوعا من الموضوعات المستخرجة). نتيجة لذلك، يمكن أن تنتمي الوثيقة أو الكلمة إلى موضوع واحد أو أكثر أو لا تنتمي على الإطلاق إلى أي موضوع.

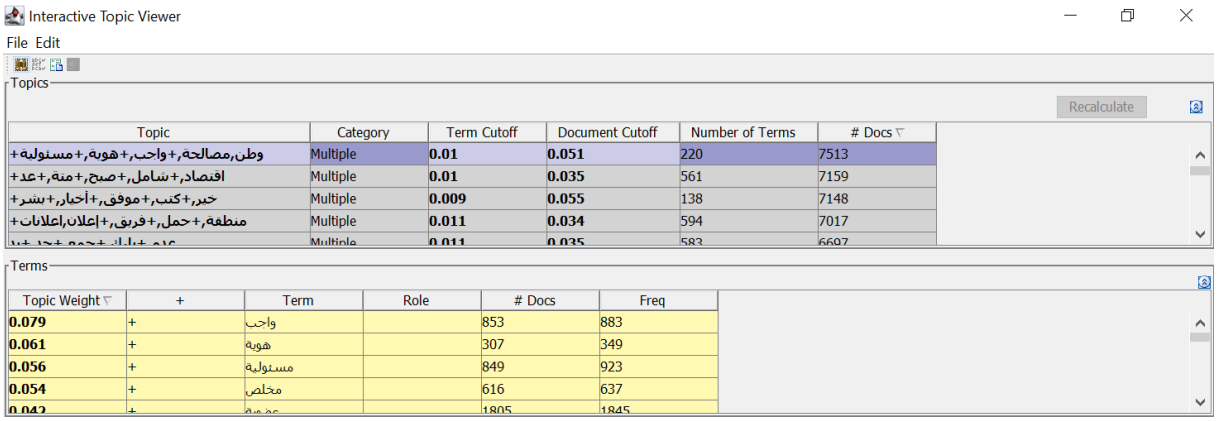
في الشكل رقم (٦) نجد أن الموضوع الذي يظهر في وسط الرسم مرتبط مع الوثائق ١٦ ، ٢٠ ، ٦٥ ، ٢١٥ ، ٨٢٠ ، ١١٠ ، ٨ ، ٥. كما نلاحظ أن الموضوع مرتبط بشكل أقوى (Strong Association) مع الوثائق ٢١٥ ، ٨ ، ٥ مقارنة بقوة ارتباطه مع الوثائق الأخرى في الرسم. ويُعَبِّر الشكل عن قوة الارتباط (Strength of Association) بين الوثيقة والموضوع من خلال سماكة الخط الذي يصل الوثيقة بالموضوع. كما يبين الشكل أن الكلمات التي تمثل الموضوع هي الكلمات اختبار، أسئلة، إجابة، دراسة، نتائج، مقرر. بناء على هذه الكلمات، يمكننا أن نعتبر أن الموضوع يتحدث عن اختبارات المقررات ويمكننا تسمية الموضوع باسم "اختبارات المقررات". وفي هذا السياق، تُحَدَّد أهمية كلمة معينة من الكلمات التي تمثل موضوعا معينا (مثل كلمة "اختبار" في مثالنا هذا) مقارنة مع الكلمات الأخرى التي تمثل هذا الموضوع من خلال الوزن الذي يُعطى لهذه الكلمة مقارنة مع الكلمات الأخرى التي تمثل نفس الموضوع، وتجدر الإشارة إلى أن الكلمة الواحدة قد تكون منتمية إلى موضوعات متعددة، وفي هذه الحالة سوف تحمل هذه الكلمة أكثر من وزن للموضوع بالنسبة للكلمة (Topic Weight for the Term) بناء على أهميتها في كل موضوع.



شكل رقم (٦): رسم توضيحي يعرض مثالا على ارتباط الوثائق والكلمات بموضوع من موضوعات نص معين

كما أشرنا سابقا، من أجل اكتشاف موضوعات النص، تَستخدم أداة تحديد موضوعات النص في نظام SAS Text Miner تقنية التحليل الدلالي الكامن (Latent Semantic Analysis (LSA))، وتجدر الإشارة إلى أن أداة تحديد موضوع النص تحتاج إلى موارد حاسوبية كافية (Computer Intensive Resources) كالذاكرة وذلك لتطبيق منهجية تقسيم القيمة الفردية (Singular Value Decomposition) التي تستخدمها تقنية التحليل الدلالي الكامن لتحديد موضوعات النص.

بعد انتهاء أداة تحديد موضوعات النص من عملها، تَعرض الأداة الموضوعات التي قامت باكتشافها والوثائق والكلمات المرتبطة بكل موضوع، مع عرض عدد الوثائق (التغريدات) في كل موضوع كما هو موضح في الشاشة التالية:



شكل رقم (٧): شاشة تعرض نتائج استخدام أداة تحديد موضوعات النص (Text Topic) وتَظهر فيها الموضوعات التي قامت الأداة باكتشافها

نلاحظ أن الشاشة تعرض أيضا في قسم الكلمات (Terms) وزن كل موضوع بالنسبة للكلمة (Topic Weight for the Term)، وعدد الوثائق (التغريدات) التي ذُكِرَت فيها الكلمة في إطار الموضوع الواحد (# Docs) والعدد الكلي لمرات ذكر الكلمة في الوثائق المنتمية إلى هذا الموضوع (Freq) حيث يمكن أن تُذكَر الكلمة أكثر من مرة في نفس الوثيقة.

٤. نتائج التنقيب في نصوص التغريدات

بعد حديثنا في الأقسام السابقة عن الخطوات التي اتبعناها للتنقيب في نصوص التغريدات باستخدام نظام SAS Text Miner، نعرض في هذا القسم نتائج هذا التنقيب.

١.٤. الموضوعات المهيمنة في التغريدات

بعد قيام أداة تحديد موضوع النص (Text Topic) بتطبيق عملية اكتشاف موضوعات الوثائق النصية (التغريدات) باستخدام تقنية التحليل الدلالي الكامن (Latent Semantic Analysis (LSA)) من خلال منهجية تقسيم القيمة الفردية (Singular Value Decomposition)، تَبَيَّن لنا أن الموضوعات التالية هي أهم الموضوعات (Topics) التي تناولتها التغريدات، وتجدر الإشارة إلى أنه - كما أشرنا سابقا - يمكن أن تنتمي الوثيقة (التغريدة) إلى موضوع واحد أو أكثر أو لا تنتمي إلى أي موضوع على الإطلاق[1]. ويبين الجدول التالي أبرز الموضوعات المهيمنة التي اكتُشِفَت في التغريدات وأبرز الكلمات /الصيغ الصرفية الأساسية (Lemmas) المرتبطة بكل موضوع (بناء على وزن الموضوع بالنسبة للكلمة Topic Weight for the Term).

[1] هذا هو السبب الذي يفسر وجود عدد كبير من التغريدات التي لا يمكن تصنيفها إلى موضوع معين كالتغريدات التي تحوي تعليقات أو تبريكات أو مديحا أو جملا قصيرة، حيث لا يمكن تصنيفها في إطار أي موضوع.

جدول رقم (٢): أبرز الموضوعات المهيمنة التي اكتُشِفَت في التغريدات

الموضوع (Topic)	أبرز الكلمات /الصيغ الصرفية الأساسية (Lemmas) المرتبطة بالموضوع (بناء على وزن الموضوع بالنسبة للكلمة Topic Weight for the Term)
وعي الناخبين	وعي، شعب، ممثل، إرادة، تغيير، حقوق، مصالح، أمانة، أطياف، حسن
أمانة التصويت	صوت، أمانة، استحقاق، إعطاء، مصير، قوي، عضوية، ضمير، طائفة
الاقتصاد	اقتصاد، إصلاح، رفاهية، هيكلة، تنمية، قوة
الفساد	فساد، محاربة، مكافحة، مفسد، فاسد، ملف، مواجهة، قضايا، كشف، هيئة
شراء الأصوات	شراء، صوت، بيع، ضمير، دفع، راشي، مرتشي، ذمة، خيانة، شرف
الشباب	شباب، طموح، مشاكل، رؤية، معالجة، حلول، اهتمام، أولوية، خريج، مستقبل، قضايا
الانتخابات التشاورية (الانتخابات الفرعية)	تشاوري، مُخرَجات، التزام، قبيلة، كرسي، نجاح، فرعي، محفوظ
التعليم	تعليم، معلم، طالب، جامعة، مناهج، دراسة، خريج، شهادة، تطوير، تربوي، بطالة، تصنيف
المصالحة الوطنية	وطن، مصالحة، عفو، شامل

٢.٤. الكلمات المفتاحية المحورية في إطار موضوعات التغريدات

تَبَيَّن من خلال استخدام أداة تنقية النص (Text Filtering) ومن خلال إحصاءات النصوص أن الكلمات التالية هي الأكثر أهمية في نصوص التغريدات حيث تُمثل كلمات مفتاحية محورية (Keywords) في إطار موضوعات تلك التغريدات بناء على الوزن الإحصائي لهذه الكلمات (Term Weight). كما ذكرنا سابقا، يُستخدم الوزن الإحصائي للكلمة لمعرفة أهمية الكلمة بناء على عدد مرات ذكرها في كل وثيقة وكيفية توزيع الكلمة على الوثائق الموجودة في مجموعة الوثائق النصية (التغريدات). يعرض الجدول التالي هذه الكلمات المفتاحية المحورية مرتبة وفقًا للوزن الإحصائي للكلمة:

جدول رقم (٣): جدول يبين الكلمات الأكثر أهمية في نصوص التغريدات والتي تُمثل كلمات مفتاحية محورية

الكلمة /الصيغة الصرفية الأساسية (Lemma)	عدد المرات الكلي لِذِكر الكلمة/الصيغة في التغريدات	عدد التغريدات التي ذُكِرَت فيها الكلمة/الصيغة	وزن الكلمة الإحصائي (Term Weight)
ديمقراطي	1072	989	0.414
شرفاء	1011	975	0.414
مصالح	1094	1020	0.411
التزام	1201	1126	0.404
ملتزم	1169	1112	0.404
مجتمع	1166	1103	0.404
ثقة	1193	1131	0.402
مصلحة	1462	1289	0.393
دين	1339	1269	0.392
قوانين	1411	1307	0.39
صناديق	1348	1319	0.388
مشاركة	1588	1485	0.379
دستور	1794	1610	0.374
تشريع	1682	1603	0.372
قضية	1946	1716	0.369
إعلام	1873	1725	0.367
كرسي	2035	1814	0.366
شباب	1960	1776	0.365
حقوق	1946	1764	0.365
استحق	1951	1793	0.364
مواقف	1895	1785	0.363
رسالة	1861	1797	0.362
عضوية	1845	1805	0.361
خدمة	2188	2029	0.353
سياسي	2278	2078	0.351
مال	2385	2184	0.347
فائز	2213	2142	0.347
قبيلة	2504	2227	0.346
مستقبل	2725	2311	0.344
أصوات	2429	2262	0.343
إصلاح	2677	2388	0.34

الكلمة /الصيغة الصرفية الأساسية (Lemma)	عدد المرات الكلي لِذِكر الكلمة/الصيغة في التغريدات	عدد التغريدات التي ذُكِرَت فيها الكلمة/الصيغة	وزن الكلمة الإحصائي (Term Weight)
اقتراع	2493	2336	0.34
برنامج	2427	2318	0.34
أعضاء	2666	2476	0.336
بلد	2666	2483	0.335
حس	2632	2491	0.335
فاسد	2926	2612	0.333
عمل	2936	2614	0.332
حق	3018	2704	0.33
نواب	3148	2789	0.329
أمانة	2957	2730	0.328
قادم	2852	2694	0.328
وعي	2770	2673	0.328
قانون	3572	2927	0.325
تغيير	3492	3136	0.317
حكومة	3620	3190	0.316
صالح	3369	3146	0.315
حسن	3266	3128	0.315
دعم	3600	3287	0.312
حمل	3424	3288	0.31
فوز	3514	3343	0.309
تصويت	3891	3525	0.306
فساد	4398	3898	0.298
نائب	5469	4295	0.297
توفيق	4176	4029	0.293
نتائج	4506	4313	0.287
اختيار	5062	4495	0.286
مواطن	5283	4722	0.282
انتخاب	5134	4776	0.279
ناخب	5347	4895	0.278
شعب	10386	8660	0.232
مرشح	28217	25744	0.137

٥. نتائج البحث

قدم هذا البحث دراسة تطبيقية استخدَمتُ فيها تقنية التنقيب في النصوص (Text Mining) للتنقيب في نصوص تغريدات تويتر التي احتوت على كلمات متعلقة بانتخابات مجلس الأمة الكويتي وذلك لاكتشاف الموضوعات المهيمنة (Topic Extraction) التي تحدث عنها الناخبون والمرشحون خلال فترة انتخابات مجلس الأمة في دولة الكويت في العام ٢٠٢٠، إضافة إلى اكتشاف العلاقات التي تربط كلمات معينة بموضوعات معينة واكتشاف الكلمات الأكثر أهمية في نصوص التغريدات والتي تمثل كلمات مفتاحية محورية (Keywords) في إطار موضوعات تلك التغريدات. ولتطبيق التنقيب في النصوص استخدمتُ نظام SAS Text Miner المتخصص في التنقيب في النصوص وذلك من خلال الاستفادة من قدراته على معالجة وتحليل النصوص العربية. في هذا السياق، باستخدام تقنية تحديد الموضوعات (Topic Extraction) وإحصاءات عدد مرات ذكر الكلمات في الوثائق النصية (التغريدات) والأوزان الإحصائية لهذه الكلمات (Terms Weights)، توصلتُ إلى أن أبرز الموضوعات المهيمنة التي تناولها الناخبون والمرشحون هي الموضوعات التي عرضتُها في الجدول (٢)، وفيما يلي وصف مختصر لكل موضوع منها:

الموضوع الأول: دور الوعي الشعبي وإرادة للناخبين في تغيير تركيبة مجلس الأمة من خلال اختيار ممثليهم في المجلس عن طريق حسن اختيار المرشحين الذين يمثلون أطياف الشعب ويحرصون على مصالحه والدفاع عن حقوقه.

الموضوع الثاني: ضرورة توخي الناخبين للأمانة وتحكيمهم لضميرهم عند تصويتهم للمرشحين لعضوية مجلس الأمة من خلال إعطاء أصواتهم لمن يستحق ومن يتمتع بالقوة والأمانة مع الحرص على أن يكون الصوت للوطن دون تحيز طائفي وذلك من أجل مستقبل هذا الوطن.

الموضوع الثالث: أهمية الاهتمام بالتنمية وتحقيق إصلاح وتطوير الاقتصاد وإعادة هيكلته باعتباره ركيزة لاستمرار رفاهية المواطن.

الموضوع الرابع: أهمية تعامُل الحكومة مع قضايا وملفات الفساد ومحاربته ومحاسبة وكشف الفاسدين وتفعيل دور هيئة مكافحة الفساد وضرورة اختيار الناخبين للمرشحين القادرين على مواجهة ذلك الفساد.

الموضوع الخامس: انتقاد ظاهرة شراء بعض المرشحين لأصوات وذمم بعض الناخبين وانتقاد بيع بعض الناخبين لأصواتهم باعتبار أن ذلك خيانة وبيعا للوطن ومخالفة للضمير والشرف وشكلا من أشكال الرشوة.

الموضوع السادس: الحديث عن مدى التزام الناخبين بمُخرَجات الانتخابات التشاورية (الانتخابات الفرعية غير الرسمية) التي تجريها بعض القبائل في دولة الكويت قبل بدء الانتخابات الرسمية لتزكية مرشحي كل قبيلة لخوض انتخابات مجلس الأمة ودور ذلك في زيادة فرص تمثيل القبيلة نيابيا والحفاظ على كراسيها في المجلس.

الموضوع السابع: ضرورة تطوير وإصلاح التعليم ومناهجه بما يساهم في تحسين تصنيف دولة الكويت في مؤشرات التعليم العالمية، والحاجة إلى ربط مُخرَجات التعليم بمتطلبات سوق العمل لتجنب البطالة.

الموضوع الثامن: ضرورة دعم الشباب والاهتمام بقضاياهم مثل قضايا الخريجين والإسكان والتوظيف، وأهمية دور الناخبين الشباب في إحداث التغيير في نتائج الانتخابات لإيصال من يمثلهم من المرشحين الشباب إلى مجلس الأمة.

الموضوع التاسع: ضرورة تحقيق المصالحة الوطنية من خلال العفو الشامل عن المحكومين في قضايا الرأي والمحكومين في قضية دخول مبنى مجلس الأمة التي شملت بعض نواب مجلس الأمة السابقين وبعض المواطنين.

من جهة أخرى، قُمتُ من خلال استخدام أداة تنقية النص (Text Filtering) في نظام SAS Text Miner ومن خلال إحصاءات النصوص بتحديد الكلمات الأكثر أهمية في نصوص التغريدات والتي تُمثل كلمات مفتاحية محورية (Keywords) في إطار موضوعات تلك التغريدات بناء على الوزن الإحصائي لهذه الكلمات (Term Weight). وكما ذكرنا سابقا، يُستخدم الوزن الإحصائي للكلمة لمعرفة أهمية الكلمة بناء على عدد مرات ذكرها في كل وثيقة وكيفية توزيع الكلمة على الوثائق الموجودة في مجموعة الوثائق النصية (التغريدات)؛ حيث عرضتُ هذه الكلمات في الجدول (٣).

وبناء على ما سبق يمكننا أن نستنتج أن المزاج العام للناخبين والمرشحين تجاه الإجراءات والمشاريع الحكومية ليس إيجابيا بشكل عام، بل يميل إلى انتقاد تقصيرها في جوانب متعلقة بمحاربة الفساد وتطوير التعليم وتحقيق الإصلاح والتطوير الاقتصادي. كما أن هنالك مزاجا سلبيا تجاه عدم وجود وعي كاف لدى بعض الناخبين عند اختيارهم للمرشحين بشكل يضمن التصويت لمن يستحق من المرشحين للوصول إلى

مجلس الأمة، إضافة إلى وجود مزاج سلبي تجاه ظاهرة شراء بعض المرشحين لأصوات وذمم بعض الناخبين.

أخيرا، تؤكد نتائج هذه الدراسة أهمية استخدام تقنية التنقيب في النصوص للتعامل مع البيانات الضخمة (Big Data) التي تحويها التغريدات المنشورة على تويتر؛ حيث إن النقاش الذي يُنشَر على تويتر يمثل وسيلة حية لاستطلاع رأي الجمهور ولمعرفة اتجاه الرأي العام. كما يمكن من خلال هذا النقاش معرفة ردود أفعال الجمهور تجاه القضايا السياسية والاجتماعية والاقتصادية. وهذا يؤكد أن على الحكومات الانتباه إلى أهمية اكتشاف المعلومات المهمة في النصوص المنشورة على تويتر والتي تتحدث عن موضوعات / كلمات معينة في فترة زمنية معينة ويشتمل ذلك على اكتشاف الموضوعات المهيمنة في هذه التغريدات وأبرز الكلمات المفتاحية المحورية التي تحدث عنها الجمهور والعلاقات التي تربط كلمات معينة بموضوعات معينة.

المراجع:

الخزاعى، محمد رده، و حسن بن عواد السريحى. "تحليل الآراء على شبكات التواصل الاجتماعي: نموذج تطبيقي لقياس مستوى التعصب الرياضي في تويتر." *Cybrarians Journal*، البوابة العربية للمكتبات والمعلومات، ٢٠١٨، العدد ٥٠، الصفحات ١ - ٢١.

الخليفى، طارق. "تنقيب بيانات وسائل التواصل الاجتماعي واستخداماته في البحوث الإعلامية: تحليل المشاعر نموذجا." *مجلة البحوث والدراسات الإعلامية*، المعهد الدولي العالي للإعلام بالشروق، ٢٠١٩، العدد ٨، الصفحات ٢٧٩ - ٣٥١.

خليل، حمزة السيد حمزة. "توظيف تطبيقات الذكاء الاصطناعي لتحليل مشاعر مستخدمي مواقع التواصل الاجتماعي في الوقت الفعلي لأزمة جائحة فيروس كورونا." *المجلة المصرية لبحوث الرأي العام*، جامعة القاهرة - كلية الإعلام - مركز بحوث الرأي العام، مصر، ٢٠٢١، مجلد ٢٠، العدد ٢، الصفحات ١٤٩- ٢٠٢.

يوسف، ريهام سامي حسين. "مواقع التواصل الاجتماعي كقاعدة بيانات لقياس الرأي العام: الواقع والإشكاليات." *مجلة البحوث والدراسات الإعلامية*، المعهد الدولي العالي للإعلام بالشروق، ٢٠١٨، العدد ٦، الصفحات ١٩٣-٢١٥.

References:

Ahmed, Hassan, Qazvinian, Vahed, and Radev, Dragomir. "What's with the attitude? identifying sentences with attitude in online discussions." *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.

Al-Laith, Ali, and Shahbaz, Muhammad. "Tracking sentiment towards news entities from Arabic news on social media." *Future Generation Computer Systems*, 2021, vol. 118, pp. 467-484.

AlAjlan, Shatha AbdulAziz, and Saudagar, Abdul Khader Jilani. "Machine learning approach for threat detection on social media posts containing Arabic text." *Evolutionary Intelligence*, 2021, vol. 14, no. 2, pp. 811-822.

Albright, Russ. *"Taming Text with the SVD."* SAS Institute, 2004.

Allaith, Ali, Shahbaz, Muhammad, and Alkoli, Mohammed. "Neural Network Approach for Irony Detection from Arabic Text on Social Media." *FIRE (Working Notes)*, 2019.

Alomari, Ebtesam, et al. "COVID-19: Detecting government pandemic measures and public concerns from Twitter arabic data using distributed machine learning." *International Journal of Environmental Research and Public Health*, 2021,vol. 18, no. 1 , p. 282.

Alsafari, Safa, Sadaoui, Samira, and Mouhoub, Malek. "Hate and offensive speech detection on arabic social media." *Online Social Networks and Media*, 2020, vol. 19.

Brendan, O'Connor, et al. "From tweets to polls: Linking text sentiment to public opinion time series." *Fourth international AAAI conference on weblogs and social media*, 2010.

Crystal, D. *A dictionary of linguistics and phonetics*. John Wiley & Sons, 2011.

Deerwester, Scott, et al. "Indexing by latent semantic analysis." *Journal of the American society for information science*, 1990, vol. 41, no. 6, pp. 391-407.

DeVille, Barry, and Bawa, Gurpreet Singh. *Text as Data: Computational Methods of Understanding Written Expression Using SAS*, Wiley, 2021.

Gerard, Salton, and McGill, Michael J. *Introduction to modern information retrieval*. McGraw Hill, 1983..

Gouta, Chakraborty, Pagolu, Murali, and Garla, Satish. *Text mining and analysis: practical methods, examples, and case studies using SAS*, SAS Institute, 2014.

Guellil, Imane, et al. "Detecting hate speech against politicians in Arabic community on social media." *International Journal of Web Information Systems*, 2020.

Ignatow, Gabe, and Mihalcea, Rada. "*Text Mining: A Guidebook for the Social Sciences.*" SAGE Publications, 2016.

Jafarian, Hamoon. "Topic Discovery on Farsi, English, French, and Arabic Tweets Related to COVID-19 Using Text Mining Techniques." *Navigating Healthcare Through Challenging Times*, Edited by D. Hayn et al., IT Austrian Institute of Technology and IOS Press, 2021, vol. 26.

Kanan, Tarek, Aldaaja, Amal, and Hawashin, Bilal. "Cyber-Bullying and Cyber-Harassment Detection Using Supervised Machine Learning Techniques in Arabic Social Media Contents." *Journal of Internet Technology*, 2020, vol. 21, no. 5, pp.1409-1421.

Khalafat, Monther, et al. "Violence Detection over Online Social Networks: An Arabic Sentiment Analysis Approach." *International Journal of Interactive Mobile Technologies*, 2021, vol. 15, no. 14.

Matthews, P. H. "*The concise Oxford dictionary of linguistics*". Oxford University Press, 2014.

Miller, J. E., and Brown, E. K. "*The Cambridge dictionary of linguistics*". Cambridge University Press, 2013.

Sarma, Kattamuri S. "*Predictive modeling with SAS enterprise miner: Practical solutions for business applications"*. SAS Institute, 2017.

SAS Institute. "Strength of Association for Concept Linking." *SAS Help Center*, documentation.sas.com/doc/en/tmref/15.2/n0chdwsd64uafcn164shwz4793a2.htm. Accessed 20 January 2022.

SAS Institute. "Term Weighting." *SAS Help Center*, https://documentation.sas.com/doc/en/tmref/15.2/p06w2zv74gep4zn135roo9r7xdxk.htm. Accessed 20 Jan. 2022.

Shaw, Jonathan. "Why 'Big Data' is a Big Deal." *Harvard Magazine*, Harvard University, 2014, https://www.harvardmagazine.com/2014/03/why-big-data-is-a-big-deal. Accessed 20 January 2022.

Twitter. "Getting Started with Premium Search Tweets: Full-Archive API." *Developer Platform*, https://developer.twitter.com/en/docs/twitter-api/premium/search-api/quick-start/premium-full-archive. Accessed 20 January 2022.

Ulrich, Reincke. "Profiling and classification of scientific documents with SAS Text Miner." 2003.

فهرس الجداول:

Discovering the Dominant Topics using the Text Mining Technique in the Tweets of Voters and Candidates during the Kuwaiti National Assembly Elections as a Model

Dr. Salah Alnajem

Arabic Department

College of Arts, Kuwait University, Kuwait

salah.alnajem@ku.edu.kw

Abstract

This paper presents an applied study in which the Text Mining technique has been used to mine the texts of tweets that contained words related to the elections of the Kuwait National Assembly (Parliament) in order to discover the dominant topics (Topic Extraction) that the voters and candidates talked about during the period of the National Assembly elections in the State of Kuwait in the year 2020. Furthermore, using Text Mining, the relationships that link certain words to certain topics have been discovered, in addition to discovering the important keywords within the context of the topics of the tweets. The tweets were collected from the official Twitter Archive (Twitter Full Archive), where I collected the tweets that contained keywords related to the National Assembly elections during the mentioned period. After that, I used SAS Text Miner system to mine the texts of the tweets and discover the topics they contain by using the Latent Semantic Analysis (LSA) technique. It was found that the general sentiment of voters and candidates towards government measures and projects is not generally positive, but rather tends to criticize the government's shortcomings in aspects related to fighting corruption, developing education, achieving economic reform and development. There is also a negative sentiment towards the lack of sufficient awareness among some voters when choosing candidates in a way that guarantees voting for the eligible candidates to reach the National Assembly, in addition to the negative sentiment towards the phenomenon of buying the votes of some voters by some candidates.

Keywords: Computational Linguistics, Arabic Language Processing, Text Mining, Applied Linguistics, Social Media Analytics, Computational Corpora.