Clustering Tweets to Discover Trending Topics about دبي (Dubai)

تصنيف تغريدات تويتر في مجموعات لاكتشاف أكثر المواضيع تداولا عن دبي

by

SALAMA KHAMIS SALEM KHAMIS ALYALYALI

A dissertation submitted in fulfilment

of the requirements for the degree of

MSc INFORMATICS
(KNOWLEDGE AND DATA MANAGEMENT)

at

The British University in Dubai

Prof Sherief Abdallah
March 2018

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

Signature of the student

Abstract

Nowadays, a lot of people targeting social networks to learn what are the trending topics and the news alongside the huge flow of texts posted daily in social networks. One of these social networks is Twitter - a microblogging hub and rich environment of data. Scanning tweets online is a hard task and searching effortlessly to find intended topic from huge amount of data is also time consuming. This paper is intended to propose a solution of collecting Twitter of the corpus دبي (Dubai) by using Zapier website and storing them in Google sheet. Then, creating a word vector to the tweets by using TF-IDF methodology. After this, log results into k- mean clustering algorithm with cosine similarity to measure similarity between objects of each cluster. The results demonstrate that internal evaluation techniques failed to evaluate quality of the cluster. In addition to that, interesting topics was found about دبي (Dubai). Moreover, better results achieved by using Filter Tokens (by Region) than without using it. The data were collected for the experiment at several periods to ensure getting the most trending topics about دبي (Dubai). All of the results found in this paper tested with real tweets.

Key words: Twitter, Arabic tweets, K-mean clustering, TF-IDF, cosine similarity.

الخــلاصــة

في هذه الأيام يقوم الكثير من الناس باستهداف مواقع التواصل الاجتماعي وذلك لمعرفة المواضيع و الأخبار الشائعة فهذا يتم بالتزامن مع التدفق الضخم للنصوص اللتي يتم نشرها يوميا وبلغات مختلفة. حيث يعتبر"توتير" أحد أهم مواقع التواصل الاجتماعي واللذي يعتبر مدونة صغيرة و بيئة غنية بالبيانات. وتعتبر عملية البحث خلال التغريدات الجمة مهمة صعبة ، كما أن البحث بعناء لإيجاد الموضوع المراد من الكم الهائل من البيانات هو أيضا مضيعة للوقت. هذه الورقة البحثية تهدف لإيجاد حل لهذه المشكلة عن طريق جمع التغريدات اللتي تخص دبي باستخدام برنامج zapier و من ثم حفظها في google sheet. و بعد ذلك يتم عرض التغريدات باستخدام نظام TF-ID .و بعدها يتم تقسيم النتائج إلى مجموعات بعد إدخالها في خوارزمية التجميع k-mean التي تعمل بالاستناد إلى مقياس تشابه جيب التمام لقياس التشابه بين عناصر كل مجموعة. هذه الدراسة أظهرت أن تقنيات التقييم الداخلي فشلت في تقييم جودة المجموعة. أيضا ، تم العثور على مواضيع مثيرة للاهتمام حول دبي . علاوة على ذلك ، النتائج التي تم تحقيقها باستخدام تصفية الكلمات (حسب النطاق) أفضل من النتائج التي ظهرت بدون استخدام تصفية الكلمات (حسب النطاق). البيانات التي تم تجميعها لإجراء التجربة كانت مجمعة في فترات زمنية مختلفة للتأكد من الحصول على أكثر المواضيع الشائعة حول دبي. تم اختبار جميع النتائج التي تم العثور عليها في هذا البحث باستخدام تغريدات حقيقية.

الكلمات المفتاحية: تويتر ، التغريدات العربية ، TF-IDF ، K-mean clustering ، تشابه جيب التمام .

Dedication

To the late Sheikh Zayed bin Sultan Al Nahyan, may God have mercy upon him, I'm your daughter Salama Alyalyali and I will remain sticking on your path to persevere in the renaissance of this benevolent homeland. You left fingerprints of supporting us on our lives can't be forgotten.

To my wonderful mother who always encourages me to learn

To my dearest father who taught me to patience to achieve supreme

Special thanks to my dear husband who always supported me and insisted me on completing my educational career.

Great thanks to my daughter Mouza and Athari because I was away from you doing my Master's program.

Many thanks to my brothers and sisters with their moral support and for those who took care of my children during my study

Deep thanks to all of you who make prayer for me or encourage me and thanks for my good to be always with me.

Acknowledgements

I appreciate my professor Shereif Abdullah for his encouragement and support during the Master journey.

Also I appreciate all of BUID staff for their support and facilitate many things to me.

Great thank full for Moutaz Hamada, MSc student from BUID, who guided me and helped me to done this research.

Great full thank for ICT for funding me along way of my master journey.

Table of Contents

1. **Introduction** 1

1.1 Overview 1

1.2 Objectives 3

1.3 Research questions 4

1.4 Structure of the report (chapter summary) 5

2. **Background** 6

2.1 Twitter 6

2.2 Text Mining 7

2.2.1 Arabic text mining: 9

2.2.2 Clustering algorithms 10

2.3 Cosine similarity 12

2.4 Related work 13

3. Research Methodology 18

3.1 System Architecture: 18

3.2 Data collection 19

3.3 Data Scrub 20

3.3.1 Removing whole example 21

3.3.2 Removing words from example 24

3.4 Text presentation 27

3.4.1 Text pre-processing 28

3.4.2 Term frequency- Inverse document frequency (TF-IDF) 30

3.4.3 Prune methodology 31

3.4.4 Output 32

3.5 Clustering tweets 32

3.6 Cluster evaluation 33

3.6.1 Internal Clustering evaluation techniques: 33

3.6.2 Evaluation by observing different results 36

3.7 RapidMiner implementation 36

4. Experimental Analysis 41

4.1 Evaluation by internal clustering evaluation techniques.. 41
4.2 Evaluation by observing different results 45
4.2.1 Experiment 1 46
4.2.2 Experiment 2 49
4.2.3 Experiment 3 54
4.2.4 Experiment 4 67
4.2.5 Experiment 5 69
4.3 Discover trending topics in four months about دبي.. 71
5. Conclusion and Future Work.. 74
5.1 Conclusion... 74
5.2 Future Work 76
References:.. 77

List of Figures:

Figure 1: System architecture of extracting topics from tweets 18
Figure 2: Two types of scrubbing data 20
Figure 3: Intial processes in Rabid miner 36
Figure 4: Adding operator of “process document from data” to the preprocessing operators 37
Figure 5: Text mining models inside “process document from data” operator 37
Figure 6: The last process in data scrubbing before entering tweets into “Loop Parameters” 38
Figure 7: Cluster operator and evaluation operators inside “loop parameters” 39
Figure 8: Adding” log to Data” operator 39
Figure 9: Overview of the processes done in rabid miner 40
Figure 10: Sum of Squared results from k=2 to k=25 42
Figure 11: Davies Bouldin Index results from k=2 to k=25 43
Figure 12: Average within centroid results from k=2 to k=25 44
Figure 13: Gini Coefficient results from k=2 to k=25 45
Figure 14: Davies Bouldin Index after adding k from 26 to 31 67

List of Tables:

Table 1 : Attributes name and type in excel sheet 19
Table 2: Examples of unethical Arabic tweets and its translations 23
Table 3 : Words occurances in tweets 25
Table 4: Examples of missed stop words and how many times they occurred in tweets 27
Table 5: Comparisons between different types of conditions from contents types and ranges 29
Table 6: Results of k=2 46
Table 7: Examples of tweets in cluster 0, k=2 47
Table 8: Examples of tweets in cluster 1, k=2 48
Table 9: Results of k=10 49
Table 10: Examples of tweets in cluster 0, k=10 52
Table 11: Examples of tweets in cluster 5, k=10 52
Table 12: Examples of tweets in cluster 2, k=10 53
Table 13: Results of k=25 54
Table 14: Examples of tweets in cluster2, k=25 59
Table 15: Examples of tweets in cluster 3 and cluster 14, k=25 60
Table 16: Examples of tweets in cluster 6, k=25 61
Table 17: Examples of tweets in cluster 8, k=25 62
Table 18 : Examples of tweets in cluster 18, k=25 63
Table 19: Examples of tweets in cluster 20, k=25 64
Table 20:Examples of tweets in cluster 21, k=25 65
Table 21: Examples of tweets in cluster 22, k=25 66
Table 22: Results of k =30 68
Table 23: Examples of tweets in cluster 11, k=30 68
Table 24: Results with using Filter Tokens (by Region) 70
Table 25: Results without using Filter Tokens (by Region) 70
Table 26: size of data before and after scrubbing for each month 71
Table 27: Results of DBI in September and October 72
Table 28: Topics of k=10 for each month 72

Chapter 1

1. Introduction

This chapter is an overview about the significant of this paper and proposes the aims of this study. Moreover, this chapter arises some questions to be investigated and analyzed on the rest chapters and briefly describes the titles of the rest chapters.

1.1 Overview

Dubai aims to be the smartest city in 2021. Dubai's vision is to preamble its services to convoy with the latest technologies and leverage from them to collect data efficiently (Syeeda 2017). Collecting data has diverse sources and massive data could be collected daily. One of the most sustainable source of the data is social networks. With the massive incoming and outgoing data, there are variety types of data collected from these networks based on each network activity.

Twitter is one of these social networks, which considered as microblogging service that pays attention to monitor real-time events (Smith 2014). Twitter is considered a highly active area, with about 6,000 tweets every second and 500 million of tweets every day from Twitter (Internet live stats 2018). In the UAE, the government is encouraging Dubai government entities to create and participate in social networks to improve and ease its services to the people. Approximately 40% of the Twitter users who are following the UAE Government's accounts in the social networks asserting the role of these accounts in discovering the trending news (Nuaim 2017).

Twitter users tweet about Dubai to express their feelings or opinions, post real-time news, advertise something and/or share some information. Messages exchanged via Twitter are called tweets. Tweets can be shared in the form of text, photo, URL or mixed. Twitter as it considered a microblogging platform, has limited tweets, of 140 characters until the end of 2017. By September 2017, Twitter has started testing expansion of the text length into 280 characters (Newton 2017).

Searching about a specific topic in Twitter manually is a hard mission, especially, with the high quantum of tweets. Recent growing in the number of tweets have heightened the need for conducting numerous experiments to extract trending topics in Twitter. Despite that these experiments already done in the field of English tweets, still Arabic tweets have a vacant area for more investigation due to the limited number of researches. Moreover, to shed the light on the models used in the recent studies to discover Arabic trending tweets, following is a brief overview about them. Where these models are unsupervised learning (Al-Rubaiee & Alomar 2017; Abuaiadah, Dileep & Mustafa 2017 ; Sawaf, Zaplo & Ney 2001), agglomerative hierarchical clustering algorithm (Rafea & Mostafa 2013) and online clustering algorithm (Alsaedi, Burnap & Rana 2016; Alsaedi & Burnap 2015).

Many of the software that deals with analyzing data lack of an integrated Arabic system. This mean that these software didn’t provide whole of its functionality in Arabic language as it provide them for the native language of the software manufacturer. Many of the preprocessing steps to clean and prepare tweets is done in this paper to overcome this gap. Many of Twitter users prefer to use their dialects in their tweets. Other users tweet with grammatical and spelling errors, abbreviations, unethical words and spams that contain some of the advertising campaigns (Alsaedi, Burnap & Rana 2016). Consequently, this stands as handicap in front of the language processing in Twitter.

To beat off these difficulties, this paper is proposing unsupervised learning algorithm to cluster similar tweets about the Arabic word دبي (Dubai). Nevertheless, the classification method didn’t use because it provides a predefined list of the events or categories. As a result, this limits its functionality and obscures it to not reveal new events that are out range of the predefined list.

Heretofore, this the first research shed the light on Arabic words “دبي” on Twitter, which will be in conjunction with the big events in Dubai “Expo 2020”. As Expo 2020 is a big upcoming events to Dubai, this action will inform visitors of Expo2020 of changes in the date, time and location of the events. Moreover, this will also inform visitors of Expo 2020 about any changes in the date, time, and location of the events. It worth noting that words “دبي” could be replaced by another Arabic words to find trending topics about this word. For example, find interesting topics about إكسبو (Expo) and find what people says about one of the cities like الشارقة (Al Sharjah).

1.2 Objectives

This research examines the emerging role of discovering trending topics in the context of Arabic tweets about دبي (Dubai). This study aims to investigate ability of data mining and text analysis through empirical test to find out the trending topics about one Arabic key term. This paper doesn't aim to conduct experiment in one key term, however, it can reveal that this experiment could be conducted to find interesting topics about other emirates, by following similar steps. This paper poses how unsupervised clustering could be evaluated by trying internal clustering evaluation and evaluation by observing different results. By adding feature of using filter token by region, this paper hopes to arise quality of clustering algorithm by specifying a range of tokens that could search events on it. Additionally, observing which evaluation gave the most effective results to set number of clusters (k). As the value of k must be identified before clustering the topics, k must be evaluated to see how the clustering method was successfully to set value of k. Since these topic can't being guessed from a huge number of tweets.

1.3 Research questions

This research seeks to address the following questions:

- Whether traditional preprocessing helpful in cleaning data and preparing them to cluster?
- Is there any interesting topics tweeted about دبي (Dubai)? Are these tweets carry meaningful content?
- Is increasing clusters give more interesting topics?
- Can evaluation techniques be able to evaluate clusters effectively? Can these techniques agreed about the best number of cluster?
- Does clusters give better results with using Filter Token by Region or without using it?
- Are there any new topics discovered from each month separately?

1.4 Structure of the report (chapter summary)

The overall structure of this study is divided into 5 chapters, including this introductory chapter. Chapter 2 reviews on some related articles and presents background which provides an overview about Twitter, text mining and clustering. Chapter 3 describes methodology of collecting data and implementing experiments. Chapter 4 discusses, analyzes and evaluates results. Chapter 5 draws the conclusion and mentions future work that could be done based in this study.

Chapter 2

2. Background

This chapter gives a brief overview about some terminologies used in the report. Additionally, this chapter aims to expand knowledge and to benefit from what others did in similar field of this report to introduce best.

2.1 Twitter

Twitter is micro-blogging platform, this means user can send updated short text or micro media like photographs or audio to another user. The establishment of Twitter was in March 2006, but the formal launched was in July 2006 by Jack Dorsery, Biz Stone, Noah Glas and Evan Williams (Mosley 2012). The short text exchanged on Twitter was limited to 140 characters and is called a tweet until it expanded to 280 characters in 2017 (Newton 2017; Baralis et al. 2013). These tweets could be transferred through three different methodologies like mobile application of Twitter, the website of Twitter and the third-parity application, which is supported by API (Application Programming Interface) (Kireyev, Palen & Anderson 2009). Users of Twitter can contact another users by three relationships. Firstly, “follow” relationship, where the user can know what the other parity posts without the need of the “follow” relationship from the other parity, which means the relationship here is one-way connection (Hamadeh 2015). Secondly, recalling the username in Twitter to notify the person to see this posted tweet called “mention”. Thirdly, RT is the abbreviation of “retweet”, where the user can forward the tweets that another user post (Kwak et al. 2010). Moreover, on

Twitter there is a symbol called hashtag (#). Hashtag is usually followed by the phrase which plays the role of the keyword in Twitter. Hashtag uses to identify common topics, attract desired category of the user to see the tweets, initiate a session of conversation and classify the tweet by the type or topic (Adel, ElFakharany & Badr 2014)(Becker 2011).

2.2 Text Mining

Dating foundation of the text mining to 1980s being new emerging field in text analysis, where the text introduced to the computer and the data was manipulated manually (Ignatow & Mihalcea 2016). Text mining is the process of eliciting new and anonymous information from several textual provenances and fastening these elicited information together to form theses or facts. One of the text mining fields is data mining (also known Knowledge Discovery) which is the process of disclosing desired patterns from huge databases (Ertek,Tapucu & Arin 2013). The core difference between data mining and text mining is that the text mining is required to extract information from the raw text, whereas data mining required databases of facts to extract information. One of the field in data mining is computational linguistics or as it called natural language processing. Computational linguistic is in charge with the process of analyzing small text with high efficiency. For example, the task of summarizing text demands on eliminating unnecessary words such as “the”, which is considered as a highly effective work (Hearst 2003).There are variety of sciences and topics which acquired its basics from the principles of text mining as it reported into two researches, one done by Hamadeh (2015) and the other one done by Aggarwal and Zhai (2012):

Information extraction: is the process of extracting structure information from a text, which is not texture or semi texture.

Text summarization: gives an overview of the text with the extraction of the necessary key words.

Unsupervised learning from text: the training data is not set manually (Ko & Seo 2000). As a result, unsupervised learning algorithm will be applied to disclose the new structure of patterns without the need of matching results with the training data as it is not available (Brownlee 2016). There are two mostly unsupervised learning algorithms used called clustering and topic modeling.

Supervised learning from text: The training data is available, so it will be used to evaluate the output after applying the intended algorithms. This process will be repeated until high performance is reached (Brownlee 2016).

Dimensionality reduction: It is the process of squeezing basic data and representing it as index or reference to be used later on as indicative key for many text mining applications.

Transfer learning with text: this is the process of converting data from one field to another if the fields suffer from lack of data and the collecting text is heterogeneous.

Probabilistic model for text: it is basically count the probability of text mining and represent it in model.

Mining text stream: This is the process of mining massive text concurrently with saving data offline as the data keep flowing.

Multimedia mining: It is the process of reinforcement mining process to collect text data concurrently using the same data from another fields.

Mining Text in social media: This is the process of mining text in social media, which contain links and text or either text or links. The social media faces a problem of the quality of the written text, which is sometimes incomprehensible text and sometimes uses slang language.

2.2.1 Arabic text mining:

Arabic language considered one of the popular language as it the language of the Holey Quran. Arabic language is the official language in 26 countries all over the world. It is one of the most spoken language in the world rank (Chepkemoi 2017).There are three classifications of the Arabic language: العربية الفصحى (Classical Arabic), العربية الحديثة (Modern Standard Arabic) and العربية العامية (Colloquial Arabic dialects)(Bekkali & Lachkar 2017). Arabic language consists of 28 alphabet letters:

أ،ب،ت،ث،ج،ح،خ،د،ذ،ر،ز،س،ش،ص،ض،ط،ظ،ع،غ،ف،ق،ك،ل،م،ن،هـ،و،ي

Arabic language when it compared with other languages, considered rich with its synonyms, derivatives, morphological analysis, lexical information and orthography. Therefore, Arabic language is a challenging language in analysis perspective. Additionally, Arabic language has wide range statement and fluent language (Harmain, El Khatib & Lakas 2004; Duwairi 2005; Ayedh et al. 2016).

Schema used in mining Arabic text are similar with other languages. Some of the text mining schemas applied as preprocessing steps for Arabic text: tokenization, word normalization, stemming, light stemming, removing stopwords, filter by length and filter by content. All of these schemas are explained in more details in chapter 3 research methodology except normalization. Normalization is the process of normalizing letters of one word that could shape into different formats steinto one standard format. As an example, أ (aleph with Hamza on top),ء (Hamza), ؤ (Hamza on waw), آ (aleph mad), إ (aleph with Hamza at the bottom) and ئ (hamza on ya) could be normalize to ا (aleph) (Ayedh et al. 2016).

2.2.2 Clustering algorithms

Clustering data is one of the unsupervised learning methodologies which is aforementioned. Clustering is the process of splitting the data into a number of splits that defined in advance and each split represent a cluster. There are two disciplines of clustering, one of them is the hard clustering (or partitional clustering) and the other one is the soft clustering (or hierarchical clustering). In one hand, hard clustering contains points of data that each point considers as a part of only one split. On the other hand, each point that represents data in soft clustering has a probability of joining each split. Some of the famous hard clustering are k-mean clustering, bisecting k-mean and k-modes. However, Agglomerative clustering considered as soft clustering algorithm (Ahmad & Dey 2007) (Ghwanmeh 2007) (Banerjee et al. 2005).

2.2.2.1 K-mean clustering algorithm

K-mean is the process of collecting information from a group of objects that are similar to each other in the same cluster and dissimilar from other objects in another groups. K-mean is unsupervised hard clustering algorithm. The similarity of the objects are measured by the distance between these objects. K represents number of the clusters that could be formed from the data. Centroid of the cluster could be assigned by averaging all of the values in the cluster or could be any point in the cluster that sometimes is a fictitious point.

How k-mean algorithm works?

1- Set initial value of k, which is the intended number of clusters

2- Randomly choose centroids of the clusters with k defined in step 1, which could be an imaginary points.

3- Select the objects that could be in each cluster by measuring distance between each object and the centroid. If this distance was smaller than the measured distances of this object and another centroids, then this object will belong to the cluster with shortest distance.

4- Choose again centroids of the clusters by averaging the points in each cluster, so the average will be the new centroid of each cluster.

5- Repeat the process in step 3 to assign objects that will belongs to each clusters.

6- Again repeat the process in step 4 and step 5, until reaching maximum run. Maximum run will stop when all of the objects within the cluster are near to the centroid assigned to them and no object are able to move to another clusters (Schulte Im Walde 2006; Ghwanmeh 2007).

There are two different k-mean measurements- one is for measuring distance and the other for measuring similarity. To measure distance in k-mean, Euclidean distance and average KL divergence are used. However, to measure similarity, cosine similarity, Pearson coefficient and Jaccard coefficient can compute the similarity. Much of the literature on k-mean algorithm pays particular attention to the effectiveness of k-mean. Recent evidence has examined capability of k-mean as partitional clustering to cluster long text and short text (De Villiers 2013).One of the studies declared by (Huang 2008) mentioned that k-mean clustering is effective to deal with big data. Because it had low computation, which reduced time consuming and made k-mean highly efficient to handle big data than hierarchal algorithms. Despite of the study observed by (Huang 2008), more recent study declared by Adel, ElFakharany and Badr (2014) investigated that k-mean algorithm is not reliable to handle big data of real time tweets because k-mean need to predefined number of k (clusters) as social media is an increasing data. Therefore, k required to have flexible k to be able to absorb the increasing big data.

2.3 Cosine similarity

There are different types of distance’s measurements for example mixed measures, nominal measure, numerical measure and Bregman diveregences. Cosine similarity is considere one of the famous numerical measure. The angle between two term’s vectors is the cosine similarity, while the angle is equal to 1, two terms are similar. By contrast, if the angle is 0, two terms are dissimilar. Cosine similarity measures the similarity of short text documents. Following is the formula of how to calculate cosine similarity:

$$\mathrm{cosSim}(m, n) = \frac{\sum_{i=1}^{n}(m_i \cdot n_i)}{\sqrt{\sum_{i=1}^{n}(m_i)^2 \cdot \sum_{i=1}^{n}(n_i)^2}}$$

Where m and n are two document’s vector (Al-Shalab & Obeidat 2008; Adel, ElFakharany & Badr 2014).

As more recent attention is focused on the provision of the cosine similarity in several studies (Huang 2008; Nishida 2016; Τζιωρτζής 2013). Cosine similarity as reported by (Huang 2008) used to calculate distance between centroid point and object (tweets) as it the preferable algorithm in measuring numerical distance for text as it represents terms as vector. Another study limit measurements that could deal with words to cosine similarity (Τζιωρτζής 2013). Third study revealed by (Nishida 2016) emphasis rule of cosine similarity in high dimensional space. This study described as following, if two terms, termA and termB, need to compute angle between them and observe directions of the angles. If the directions were the same, that mean termA and termB are similar to each other otherwise two terms are dissimilar. What about applying all of these comparison between all of the rest terms? This will be considered as high dimensional space and need an effective methodology like cosine similarity to compute angles between vectors. Therefore, cosine similarity success reveal its effectiveness to compute high dimensionality space more than Euclidean distance, which can only measure distance between terms (Nishida 2016). Although research has been carried out on (Nishida 2016), about success of cosine similarity in high dimensionality, recent study conducted by De Villiers (2013), showed that the cosine similarity and Euclidean distance gain similar effectiveness in clustering text documents.

2.4 Related work

Although extensive researches has been carried out on Arabic text mining, few of these researches exists, which adequately on clustering Arabic text. Most of the attention in the Arabic text mining has been focused in classifying Arabic text rather than clustering Arabic text.

Question have been raised about the best features to cluster online financial Arabic tweets in (Rafea & Mostafa 2013). They answered the question by manipulating different features in the experiment. These features are unigram, bigram, trigram and TF-IDF. They used the agglomerative hierarchical clustering algorithm to extract texts into clusters. Moreover, for clustering topics within the texts, bisecting k-mean algorithms was utilized into CLUTO tool. Overall, they found that the unigram had achieved the best results to cluster Arabic tweets. Another experiment conducted in this study was by incrementing number of clusters k that starts with 6 to 12 and 20 clusters. The results indicated that the quality of the clusters increased when the k increased as in this study when k was equaled to 20.

Several questions have been asked by Ghwanmeh (2007) about whether increasing of the clusters would increases the quality of the clustering in Arabic text? Another question has been asked about adding clustering to the text would perform a high precision in information retrieval? They investigated that rising the number of the clustering didn’t affect the precision of the Arabic text and by utilizing clustering technique, better precision acquired. There are two more researches conducted to English text assured that the number of clusters didn’t always gave better results unless the data were small (Bellot & El-Bèze 1999, 2000).

Similar authors Alsaedi, Burnap and Rana (2016); Alsaedi and Burnap (2015) have followed similar steps to find out significant events. Both of the studies conducted to discover real-world events. Similar five steps followed in both studies ex: data were collected, preprocessed, classified, clustered and summarized. Both studies utilized for classification Naïve Bayes algorithm to eliminate irrelevant tweets and for clustering tweets an online clustering algorithm is utilized. After this, summarized clusters done by TF-IDF schema. Both studies agreed that by applying online clustering algorithm after Naïve Bayes algorithm achieved a high-performance level. However, the outcome from the second study showed that TF-IDF achieved bad results in weighting terms. Moreover, another study introduced by Hammad and El-Beltagy (2017) to cluster real time event by using Arabic tweets. They utilized entropy, stream chunking, TF-IDF to disclose bursty features .They found that extracting events was successfully to mention all of the events which were synchronized with the local news except one event. English study in real time world’s event in Twitter reported by Wei, Sankaranarayanan and Samet (2017). They utilized in their study an online Twitter user geotagging to find the location of the users who didn’t use the location feature in their Twitter account by approximating the location of their friends. To extract events form tweets, TF-IDF based online clustering algorithms is utilized. The outcome of this study indicated that the events found from the real time tweets is more than the events collected from constant number of tweets. Although there were a lot of events could be collected from the real time tweets, the author of this study believed that this study failed to detect them. The reason that prevented detecting of some events was that the tweets didn't reach the appropriated TF-IDF weight to be considered as an events.

Another study done by Al-Rubaiee and Alomar (2017) conducted in student's tweet. They first collected data and applied preprocessing steps included TF-IDF and BTO (Binary Term Occurrence) to weight tweets. Secondly, they utilized n-gram to the data and in another experiments n-gram were removed. At the next step, they clustered data by using k-mean clustering. The results they found indicated that BTO achieved a high performance in identifying similarity of the student's tweets. In addition to that, using n-gram shown better results in clustering tweets than the results found from not using n-gram. Also n-gram created higher performance than using single terms in classifying Arabic text in the research done by Al-Shalab and Obeidat (2008).

Abuaiadah, Dileep and Mustafa (2017) conducted a clustering tweets to extract positive and negative tweets. In this experiment as preprocessing steps, the authors believed that removing stopwords can speed up the memory. They clustered the tweets through two clustering algorithms ex: SKM (Standard K-mean algorithm) and BKM (Bisect K-mean algorithm). There were five functions applied to measure similarity between tweets Cosine similarity, Pearson Correlation, Jaccard Coefficient, Euclidean and Average Kullback-Leibler Divergnce (KLD). They evaluated the clustering by purity and entropy measurements. As final result, they found that root-based stemming achieved better results than light stemming. Moreover, by using (KLD) joined with root-based stemming the purity becomes higher. They found in this study that cosine similarity indicated low performance conversely to what the other four measurements performed.

Recently investigators Sawaf, Zaplo and Ney (2001) have examined unsupervised technique with maximum entropy classification and mutual information document clustering on Arabic text. They didn’t use any preprocessing steps. They found bright results for Arabic text by following the same methodologies.

Study conducted by De Roeck and Al-Fares (2000) to manage the infixes in Arabic text to be prepared to the clustering algorithm. They found that by applying light stemming and some improvements to Arabic morphology, the result gain more accurate clusters. Moreover, another study examined by Saad and Ashour (2010), applied stemming and pruning to a classification of an Arabic text. They disclosed that stemming and pruning augmented performance of weighting terms and representing text.

Some advanced methodology used k-mean algorithm and enhanced it to cluster English text examined by Rao and Govardhan (2015). They believed that by using silhouette index, the quality of the clusters will improve. Moreover, they believed that the k-mean algorithm considered as time consuming and gave bad results. The modified K-means clustering algorithm utilized the sum of squares errors between the clusters to certain that the right points are associated to the right clusters. Detailed examination of the modified K-means clustering algorithm showed that the modified K-means clustering algorithm separated the clusters perfectly and performed impressive results to clustering quality.

One of the study conducted in English tweets to discover trending topics about Dubai by Hamadeh (2015). He used K-means algorithm to cluster tweets and cosine similarity to compute similarity between tweets. Stemming and N-gram are some of the text mining methodology used. There were four internal measurements to evaluate quality of the clusters ex: Sum of squares Error(SSE), Davies Bouldin Index (DBI), Gini coefficient and average within centroid distance. Moreover, there was another model to evaluate cluster which was cluster evaluation by model analysis. SSE and DBI indicated two different suggestions about the best number of cluster, which wasn't helpful. Moreover, Gini coefficient and average within centroid distance didn't give a clear results to set best value of k. The topics extracted from the tweets shown that this model had the ability to detect major events in Dubai at the time of collecting tweets.

New methodology has been used to extract events from English tweets conducted by Kim et al. (2012). Kim and his group in this experiment used CTC, which is Core-Topic-based Clustering to cluster tweets. They found that meaningful topic successfully extracted and faster results appeared from CTC rather than results appeared from k-mean algorithm. Another research used new methodology and is certain that there is better algorithm than k-mean, the study treated by Sapul, Aung and Jiamthapthaksin (2017) tried to compare three models to cluster English tweets. They used K-mean, CLOPE clustering and LDA topic modeling algorithms. CLOPE clustering was used for transactional data. However, LDA topic modeling algorithms extracted topics from a document. The results found from this research indicated that CLOPE had the ability to cluster non-transactional data and was successful in extracting more topics than LDA and K-means.

The experiment data are rather controversial, and there is no general agreement about the best features to cluster Arabic text.

Chapter 3

3. Research Methodology

This chapter of the report describes the steps of this research methodology, steps include data collection, data scrubbing, text representation, clustering technique, and evaluation methods. Also, it will design the experimental setup conducted by RapidMiner tool.

3.1 System Architecture:

The system consists of five phases as it reveals in figure 1. First, data collecting from Twitter database. Second, data scrubbing shows in purple and yellow color. Third, distributing tweets into clusters, which shows in light green color. Fourth, evaluating cluster quality, which shows in light blue color. Final phase, extracting topics from the clusters, which shows in dark blue color.

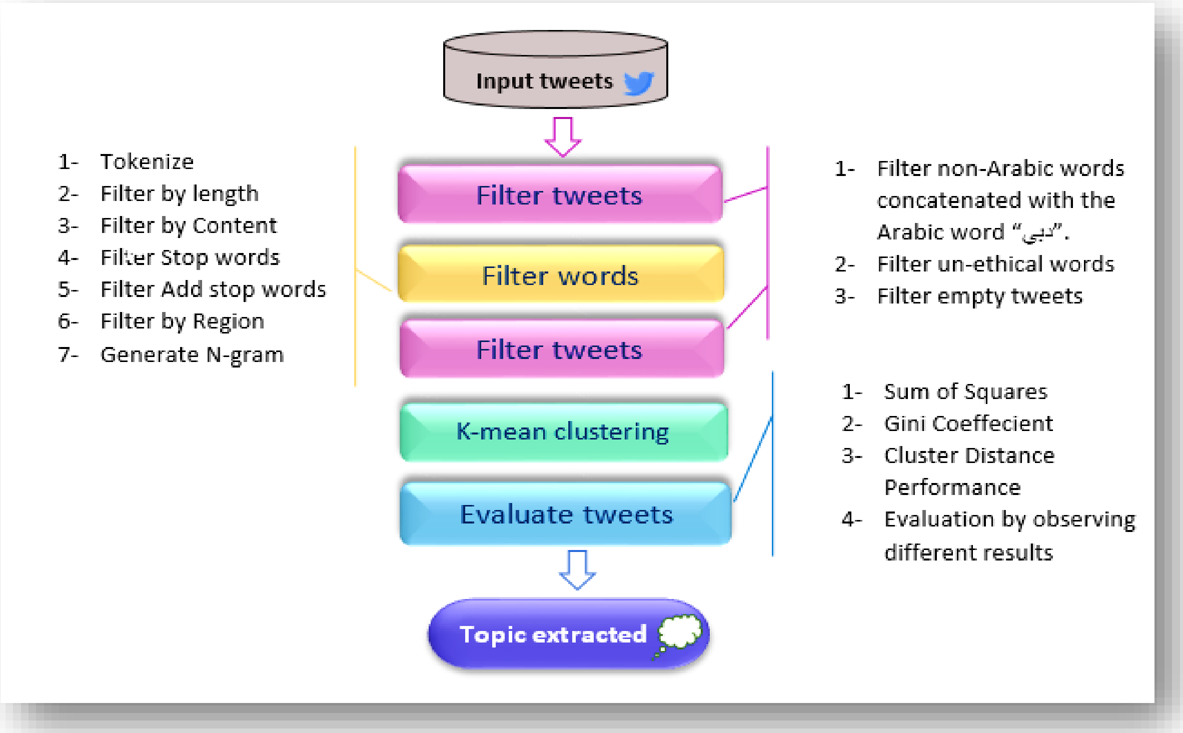

Figure 1: System architecture of extracting topics from tweets

3.2 Data collection

The intended tweets to be collected are about دبي (Dubai), these tweets were collected based on online tool called “Zapier”. This tool works as a platform to connect many applications in several ways. A new account was created in Zapier to collect and store the data from two accounts established in advanced of tweeter and Google Sheet. Zapier tool collected tweets from Twitter and impeded them into a predefined Google Sheet.

Data were collected through three several phases. The first phase was continued from 26 April until 27 April 2017 with 5,787 examples. The second phase of collecting data was started on 22- July-2017 and extended until 9 days and ended with 12,133 examples. The last phase was started on 24-Septemper-2017 and continued to 5- October- 2017 and collected 75,930 examples. The whole data gathered from Twitter about Arabic word دبي (Dubai) were 93,850 examples. Additionally, whole of tweets collected when the limited text length was 140 characters.

The first data set collected without adding the column of date and time, which refers to the data and time of posting tweets, so it only has the two columns one for the user name and the other for tweet. The two rest data sets of 12,133 examples and 75,930 examples have the three column as it shows in table 1 below. Only the column of tweet was used with the type text for this research.

Table 1 : Attributes name and type in excel sheet

Column	Type
User name	Polynomial
Tweet	Text
Date and time	Date_time

3.3 Data Scrub

Data scrubbing is a very important area for preparing the data to be cleaned before clustering them. Scrubbing the data contain removing noisy data that will affect negatively on the final results and eliminating data that won’t add any value to the collected data. The removed data were totally about 36,677 examples .After cleaning the data, the dataset left with 57,173 examples.

Following are two types of scrubbing the data:

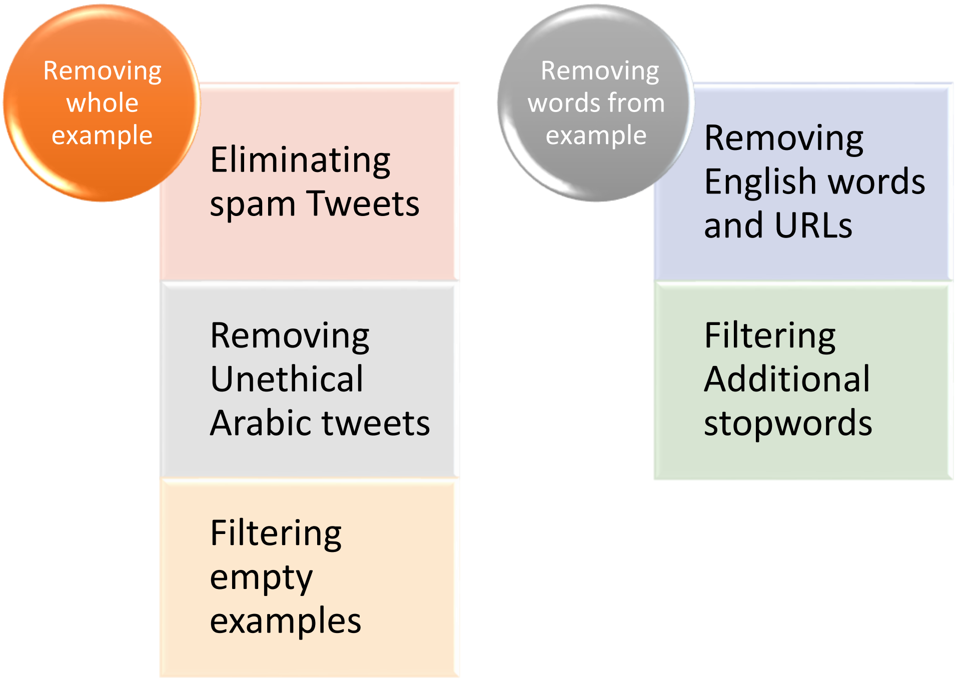
Removing whole example
Eliminating spam Tweets
Removing Unethical Arabic tweets
Filtering empty examples
Removing words from example
Removing English words and URLs
Filtering Additional stopwords

Figure 2: Two types of scrubbing data

Figure 2 shows two categorizations to scrub data. First type is removing whole example from the data set. This type will apply to the following methods removing spam tweets, removing unethical Arabic tweets and filter empty examples. Second type is removing words or characters from example. The second type responsible for removing part of the example which will apply on removing English words and URLs and filtering additional stopwords. All of the methods in first type explained firstly then the second type with it methodology is explained.

3.3.1 Removing whole example

3.3.1.1 Eliminating spam Tweets:

There are a lot of spam accounts in Twitter some of them have a silly purpose and other intended to have some returns. These accounts seek to post same tweets several times to take attention of the user to something. Sometimes these spam accounts post tweet with unrelated hashtag and this hash tag achieves a high reputation that will return with benefits to the publisher of the spam account (Twitter n.d.). Below are the tweets that invest word دبي (Dubai) from several accounts by using hashtag #دبي (# Dubai) into their special interest that mostly spilled to feed the account's owner's advertisements. These spam tweets form exactly 552 tweets from the total 93,850 tweets. Most of these tweets had the hashtag #دبي (# Dubai) with non- Arabic words. As a result, the whole text will be in any language that isn't Arabic language with the hashtag #دبي (# Dubai). For solving this problem, filtering tweets that contain the word دبي (Dubai) concatenated with non-Arabic letters (which means didn't contain 28 Arabic letters) will be applied. The following is formula for illustrating this step and examples:

$$Spam\ tweets\ for\ \text{دبي}\left(st_{\text{دبي}}\right) = \text{دبي} + (l \notin A)$$

$$,where \quad l\ is\ any\ litter\ , A = all\ of\ 28\ Arabic\ litters$$

Examples:

دبي +non-Arabic words

- دبي https://t.co/wx7lt4N5kh
- I'm at The Mall in Dubai دبي https://t.co/aPqrr7DDWN
- Everyone needs a hobby. 🎁 @ دبي& dubai https://t.co/7wYj4cxWkN
- I'm at sunny side in دبي ,دبّي https://t.co/z6vezRRPJd
- Happiness is #Sunday ☀ □E #Dubai #Travel #Traveler #Traveling #UAE #دبي https://t.co/BASBoBwr43

3.3.1.2 Removing unethical Arabic Tweets:

Unethical Arabic tweets are the tweets that contain unethical words, which abuse to the general ethics and laws. Some of these tweets are repeated many times and are considered as spam tweet. Unethical tweets cover 29,388 tweets from the whole tweets. One of the unethical Arabic words achieved high frequency reached to 9012 times (ex: سكرانة ،drunk).These tweets were eliminated by collecting all of the unethical words and removed any tweets that contain any words from these unethical words. Because, if the tweet have one unethical word, the whole sentence will be talked about the same idea, so eliminating whole tweet that contain this word will be the good solution. Examples of unethical Arabic tweets illustrates in table 2. There are more examples that shows unethical tweets, but in the following examples are the simplest one that didn’t contain a lot of unethical words.

Examples:

Table 2: Examples of unethical Arabic tweets and its translations

Tweets	Translation
❖ بعدين ما اعتقد فيه سعودي اذا طفش يأخذ اهله لمرقص بالبحرين ولا دبي !!	Then, I didn’t think that Saudi men if he became sick of will take his family to dance in Bahrain or Dubai !!
❖ يارجال خلهم يرقصون في بلدهم احسن مايروحون دبي والبحرين ومصر ويترقصون هناك ويمسكهم بوليس الاداب	Man , let them dancing in their country better than going to Dubai, Bahrain and Egypt and dancing there and held by Police of seemliness
❖ فلم سكرانه في دبي..https://t.co/vean1Ix0WT sexy	Movie about drunk woman in Dubai https://t.co/vean1Ix0WT sexy

3.3.1.3 Filter Empty example:

“Filter examples” operator filters spaces which results from the preprocessing steps especially after “Process Document form Data” operator. These spaces form a cluster by itself when it inputs to cluster model. As result, this performs a problem when the cluster has an empty tweets. Tweets with spaces have high occurrences with 6737 empty tweets. There two reasons might led to empty tweets. First reason, Some of the sentences consist of Arabic stopwords and (English words or URL). Therefore, stopwords will be removed by Filter Stopwords (Arabic) and (English words and URL) will be removed by Filter Tokens (by Contents) which explained more in the following section 3.3.2. Then the tweet will left empty. This is results because some of the collected data didn’t contain word دبي (Dubai).This is a mistake by Zapier software when it collect the tweet about دبي, some of the tweet didn’t have the word دبي (Dubai), but it have the word Dubai and it talks about Dubai. Another reason that tokens of the sentence didn’t match the condition of weighing tokens which is prune above 50, so these tweets left with empty if all of its content weight less than 50.

3.3.2 Removing words from example

3.3.2.1 Removing English words and URLs:

There are a lot of English words appear in the text, but most of the text are written in Arabic language except for these words that takes a small part from the text, but achieved high frequency. Therefore, high frequency words take into account to form topic related to دبي (Dubai), which is meaningless as this is not the focused in this research. Moreover, some of these tweets contain URLs that have more details about the topic in the tweet because Twitter allow user to write tweets with maximum of 140 characters at the time of collecting data. The solution to this problem is to remove only English words that contain English small or capital letters without removing whole tweet. The word “https” appears with frequency reached to 60154, which considers useless to be one of the key word of the topic related to دبي (Dubai). The filter used in this solution called “Filter tokens (by content)”, more details about this filter in section 3.4.1-iii.

Examples are described in the following table 3:

Table 3 : Words occurances in tweets

Words	Document occurrences	Total occurrences
https	83864.0	60154.0
Dubai	1112.0	1067.0
Vean	993.0	993.0
kAIbB	977.0	977.0
uAyW	972.0	972.0
HtCv	955.0	955.0
ahH	955.0	955.0
EcofAlR	941.0	941.0
hWezZirZtv	940.0	940.0

3.3.2.2 Additional stopwords in Arabic Language:

Arabic **s**topwords have been known as the words that are frequently repeated in the text, but can be dispensed from the text before starting the data preprocessing to enhance the final results like أين (where), أنا (me) and لماذا (why). Although that there are several lists of Arabic stopwords, each list has its own words which make each list differ from the other lists. Therefore, in this part additional stopwords will be added to enhance preprocessing methodology of the lists that acquired from RapidMiner software. The filter used in this solution called “Filter Stopwords (Dictionary)”, more illustrations about this filter in section

3.4.1-v.There are three types of the stopwords that didn’t match the list and explained as following:

3.3.2.2.1 Stopwords from the slang language:

In Arabic language there is also slang language words that is equal to the official language words such as: ولا = وإلا (and what)، وين = أين (where), اللي =الذي (whom or who). Therefore, in Twitter many comments were written in slang language that includes the stopwords written in this slang language. The problem here is that these stopwords achieved highly frequency, which is unacceptable because it will perform the trending topics.

3.3.2.2.2 Stopwords with spelling errors:

Moreover, some of the words were written in official language, but written with some mistakes in spelling like removing Al hamza “ء” above the letter Alef (أ) as in this example اكثر (more), that must be wrote like this”أكثر”. As a result, these aren’t considered as a stopwords even though it is one of them. Here for this problem normalize could be apply by using “replace operator” but this operator will add more errors instead of solving the problem. This solution will make more linguistic mistakes in many words because if the word was المدرسة (the school), by using this replacement operator will be ألمدرسة, which is not correct.

3.3.2.2.3 Missed Stopwords:

In addition to that, some of the stopwords are written correctly and considered logically as stopword, but it doesn’t match the list of stopwords in the RapidMiner list exactly in “Filter stopwords (Arabic) operator “as it predefined list.

The solution provided here is to make a new dictionary with all of these three types of stopwords. Therefore, this will eliminate these words from the tweets because some of these words repeated highly like the word اللي (whom or who), which occurred 1371 times in the whole tweets as it appeared in table 4. This will only eliminate the words not the whole of the tweets.

Table 4: Examples of missed stop words and how many times they occurred in tweets

Words	Document occurrences	Total occurrences
اللي	1284.0	1371.0
ولا	1184.0	1338.0
انا	1202.0	1242.0
خلال	902.0	919.0
يعني	725.0	770.0
شكرا	369.0	763.0

3.4 Text presentation

The words in Twitter are wrote as sentences, so to represent these words to be understood by clustering methodology, numerical vector will be used to represent these words. In RapidMiner, “process documents from data” operator was used to create vector of words and generate weight for each vector. Moreover, this operator transfers sentences into a bag of words (BoW) or it called Vector Space Model (VSM), which means representing each word separately and counted occurrences of the word in the document (Brownlee 2017).Below are several types of schema to weight vector or words as it reported by RapidMiner (2017).

1-Binary term occurrences: it weights if the word is occurred in document or not? Therefore, the value will be either 0 or 1.

2-Term occurrences: this measures how often a term occurred?

3-Term frequency (tf): this schema measures number of times the term appears in all of the documents compared to the total number of terms in the document (El-Fishawy et al. 2014).

4-TF-IDF: this is the abbreviation of term frequency inverse document frequency. This schema measures importance of term based on how many times it’s appear in several documents. If the term appears many times in one document that means it is significant term. Despite that, if the term appears many times in several documents, this conclude that this term isn’t important term (El-Fishawy et al. 2014).

In this research the schema used to create weight is TF-IDF (Term Frequency- Inverse Document Frequency), which will explain in more details in section 3.4.2.

3.4.1 Text pre-processing

Presenting the text into BoW requires some preprocessing steps to extract desire words and remove unwanted words. Following are the preprocessing steps done in the process document from data operator RapidMiner (2017):

i. **Tokenize:** is the process of splitting the text into words, symbols, phrases and useful elements, so each split will consider as a token. In this filter non letters chose to be partitions between tokens that will lead to have one single word as token (Harmain , El Khatib & Lakas 2004; Verma, Renu & Gaur 2014).

ii. **Filter Tokens (by Length):** words that contains characters within the range of minimum 3 characters and maximum of 18 characters, will be extracted to be tokens. This parameter set to be minimum with 3 characters because word دبي (Dubai) consists of three characters. Moreover, this parameter will be maximum of 18 characters because the longest word in Arabic language is 15 characters. Therefore, by adding to 2 characters as spare, in order if there any company name need additional characters, so it will expand to 18 characters (Atika 2017).

iii. **Filter Tokens (by Content):** this means filter tokens that match a specific value. Then by choosing invert condition, the filter will show all of the results that didn't match this value. For example, this filter used "contain matchs" condition to remove all of the English letters even small or capital by setting invert of the regular expression from [a-zA-Z]. The differences between conditions explained in table 5 below:

Table 5: Comparisons between different types of conditions from contents types and ranges

Conditions	**Content type**	**Range of content**
Equals condition	String	Exactly the same characters of string
Contains condition	String	Can add characters before or after string
Matches condition	Regular expression	Exactly the same characters of regular expression
Contain matches condition	Regular expression	Can add characters before or after regular expression

iv. **Filter Stopwords (Arabic):** removing common words in Arabic Language, where this is a predefined list of common word aforementioned in data scrub section.

v. **Filter Stopwords (Dictionary):** this operator will remove additional stopwords that didn't mentioned in" Filter Stopwords (Arabic)" operator. The additional stopwords in format with Unicode UTF-8 encoding.

i. **Filter Tokens (by Region):** this filter will search about tokens in specific range of tokens. This research search for token around the string دبي (Dubai) by 3 tokens after and 3 tokens before.

ii. **Generate n-Grams (Terms):** this means represent number of consecutive tokens of length n. By choosing n to be 3, this means 3 consecutive tokens will be connected and 2 consecutive tokens will be connected and only 1 token will appear.

In this preprocessing steps, stem (Arabic) operator and stem (Arabic light) operator is experienced to revert Arabic words to the base or root and keep simplest character of the word that carry the same meaning (Deshpande 2012) (Verma, Renu & Gaur 2014). As result, some of the outcome obtained indicated that if the word was"دبي" (Dubai), the stem will convert it to be "دب" (Dub). Thus, this operator will be eliminated from the experiment as the word "دبي" is the main issue in this project and the word "دب" (Dub) isn't meaningful to indicate it means the words "دبي" (Dubai).

3.4.2 Term frequency- Inverse document frequency (TF-IDF)

TF-IDF gives weigh for each term. The tweets of this research were collected with limited 140 characters because the officially launch of the expansion to 280 characters was on 8 of November 2017 and all of tweets collected before this date. Term frequency wasn't effective by itself to measure if the term is important as TF-IDF can give evidence about the importance of the term in the documents by adding IDF. Term frequency computed by the following equation:

$$TF = \frac{n_t}{\sum t}$$

n_t is the number of times term appears in all of the documents , $\sum t$ is the total number of terms in the document.

Document frequency (df_{wi}) is the number of documents that have at least one word.

IDF is computed:

$$\text{IDF} = \log \frac{N}{df_{wi}}$$

Where N is the number of documents (tweets examples) and df_{wi} is the number of documents that have at least one word, i is natural number > 0 (Kim et al. 2012).

Thus, TF-IDF is computed by:

$$TF - IDF = TF * IDF$$

$$TF - IDF = \frac{n_t}{\sum t} * \ \log \frac{N}{df_{wi}}$$

TF-IDF will show high weigh for the term when it indicates a high frequency in a document and small frequency in several documents (El-Fishawy et al. 2014) . Despite that, TF-IDF records low weigh of the term if this term appears frequently in several documents. Therefore, TF-IDF will distinguished from other schema to be a good measure for discriminating significance of the term from these equations (De Villiers 2013).

3.4.3 Prune methodology

This process used to reduce number of token appears by using one of the following approaches RapidMiner (2017):

1. Absolute: show words with total occurrences in specific range such as words that total occurrences of it above 50 and below 100,000. This approach used in this research.
2. Perceptual: show words that appear in range of specific percentage from the whole document.
3. By ranking: ordering the frequency of the words and show the words that ranked in a specific percentage range.

Prune methodology facilitates choosing words with high appearance in order to form topic. Moreover, prune methodology expedites running the process as it takes a long time to process these huge number of data in clustering steps.

3.4.4 Output

After data scrubbing and presentation, the data now are ready for modelling in clustering process. Whole data were 93,850 examples, but after scrubbing the data becomes 57,173 examples. Outcome from the previous process will be a wordlist with total occurrences for each word or for n-gram of words. In addition to that, TF-IDF weigh calculated for n-gram of words per each document (tweet).

3.5 Clustering tweets

So far, however, there has been doubt about efficiency of k-mean to collect big data explained in section 2.4.2.1., this study will utilize k-mean. Because of the collected tweets in this report aren't scalable data, so k didn't require to be flexible to handle increasable data.

There several studies done in the field of cosine similarity showed some contradictions about effectiveness of cosine similarity compared to Euclidean distance explained in section 2.4.2.2. However, the majority of the studies emphasis on the capability of cosine similarity to deal with words. For this reason this report will apply cosine similarity to measure distances between terms.

3.6 Cluster evaluation

K-mean clustering is unsupervised hard clustering, so there is no predefined expert knowledge to compare this experiment's result with the expert knowledge. Therefore, to evaluate cluster quality for unsupervised clustering, internal evaluation will be one of the solutions. Internal evaluation baptizes to check how the homogeneity of the objects (tweets) inside each cluster and how the clusters itself are isolated from each other. Another solution is observing results by manipulating with the parameter's setting. Even so, the algorithm of the clustering will be the same. Consequently, observing different results to decide the best outcomes extracted from which parameters as declared by Han (n.d).

3.6.1 Internal Clustering evaluation techniques:

There is a confusion about the proper number of k, so the internal clustering evaluation techniques try to set a proper number of k in order to get better results. There are several clustering evaluation techniques applies such as Sum of Square Errors, Davies Bouldin index, Gini Coefficcient and average within centroid distance.

Sum of Square Errors:

Sum of Squares (Item Distribution Performance): is the sum of all distances between each object inside the cluster and the centroid of the cluster. Then, the sum of all distances, which called errors, squared to get sum of square error. The following is the equation of how to calculate it:

$$\textbf{SSE} = \sum_{i=1}^{k} \sum_{x \in o_i} (x - c_i)^2$$

k is the number of clusters, x is one of the objects in the clusters, o_i is the set of all objects in a cluster and c_i is the centroid of the clusters. Elbow methodology use SSE to evaluate cluster by choosing number of k. Elbow methodology plots linear chart for SSE in range of k. After this, observe line chart when it looks like an arm, the value of k located in the elbow of the arm considers the best value of k. This happen when low value of k has low value of SSE, before value of k increased dramatically and value of SSE decreased gradually to be 0 (Gove 2017).

Gini Coefficcient:

Gini Coefficcient measures equality distributions of the values among the clusters. Gini Coeffecient calculated by the following formula:

$$Gini\ S_X = 1.0 - \sum_{y=1}^{k} (\frac{S_{xy}}{O_x})^2$$

Where S_{xy} is the number of objects included to y^{th} class in cluster x. O_x is the total number of objects in cluster x.

This measure have binary values of 0 and 1.With low value of Gini index, the distribution of the values is near from the equality. Whereas, high value of Gini index that is closer from 1 indicates the inequality of distributing the values among the clusters (Demiriz , Bennett & Embrechts 1999).

Davies Bouldin index (DBI):

This schema introduced by David Davies and Donald Bouldin in May 1979 (Davies & Bouldin 1979). DBI calculated by calculating total of (within cluster distances divided by distances between clusters). Measuring distances between centroids of the clusters done by using Euclidean distance. Following is how the DBI is calculated:

$$DBI = \frac{1}{\mathrm{k}}\sum_{x=1}^{\mathrm{k}} R_x$$

$$R_{x=} \max_{y=1\ldots\mathrm{k},x\neq y}(R_{xy}), x = 1 \ldots \mathrm{k}$$

, where

$$R_{xy} = \frac{ms_x + \mathrm{m}s_y}{d_{xy}}$$

$$ms_x = \frac{1}{\|c_x\|}\sum_{l\epsilon c_x} d(l, v_x), ms_y = \frac{1}{\|c_y\|}\sum_{z\epsilon c_y} d(z, v_y)\,, d_{xy} = d\big(v_x, v_y\big)$$

k is number of clusters, x and *y* are clusters, ms is the mean square error ,d_{xy} is the distance between centroid x and y. v_x is the centroid of the cluster x and v_y is the centroid of the cluster y. C is the set of all objects of the clusters. l and z are objects in the clusters. As the distance between clusters increase and the distance between the object inside each cluster decrease, this indicate a good cluster (Maulik & Bandyopadhyay 2002).

Average within centroid distance:

This method calculates distance between centroid and each point in the cluster then average these distances. When the value of average within centroid distance become lower, this is an indicator of good number of clusters (RapidMiner 2017).

3.6.2 Evaluation by observing different results

In this technique the evaluation is based on observing final results after several trials. Therefore, by changing parameters such as increasing number of k and using Filter Tokens (by Region) operator and removing Filter Tokens (by Region) operator, the data will give different results. As a result, by observing different outcomes, the decision will be taken to set suitable parameters to give better outcomes.

3.7 RapidMiner implementation

RapidMiner is a good GUI to implement different model of operators. The following are the setup parameters of the operations used in the experiment:

1- Import data from excel sheet (data imported from google sheet) by using “Read Excel” operator.

2- The data then input into two “Filter Examples” operators used to eliminate some examples from the whole data as a preprocessing steps, which shown in figure 3 and explained bellow:

   a- Filter non-Arabic words that concatenated with Arabic word "دبي".

   b- Filter un-ethical words that written in Arabic language.

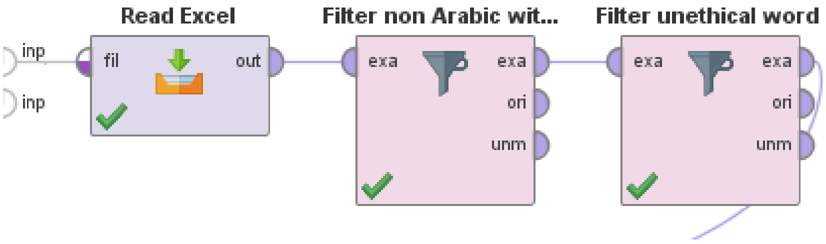

Figure 3: Intial processes in Rabid miner

3- The data rest from the filter examples are input into “Process Documents from Data” operator as it appeared in figure 4:

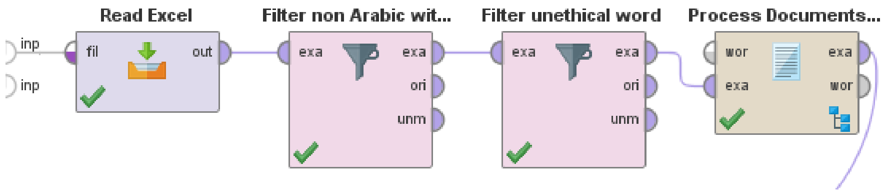

Figure 4: Adding operator of “process document from data” to the preprocessing operators

4- “Process Documents from Data” operator also have a preprocessing steps, which have a sub operators inside it as it illustrated in figure 5:

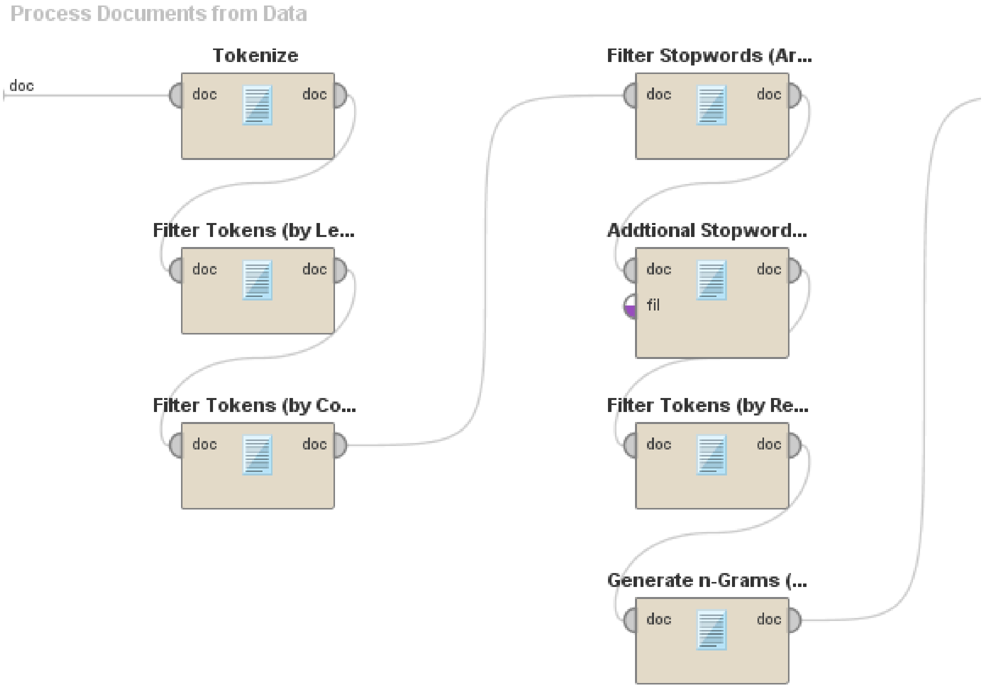

Figure 5: Text mining models inside “process document from data” operator

1- Some of the tweets after “Process Documents from Data” operator will be empty, so these tweets will be eliminates by inputting them into “Filter Examples” operator.

After this, the data will be ready for clustering and introduces into "Loop Parameters" operator as it displayed in figure 6:

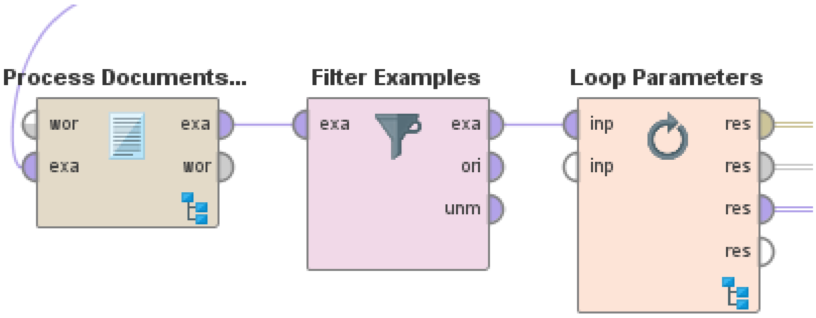

Figure 6: The last process in data scrubbing before entering tweets into "Loop Parameters"

2- Loop operator has 7 sub operators inside it described below, presented in figure 7:

i. Generate ID: add ID for each tweet attribute in order to evaluate ID of the tweet against clusters, which ID belongs to it to notice best value of k.

ii. K-mean Clustering: cluster the tweets into a number of clusters (k), numerical measure to measure distance is cosine similarity. The main differences between max run and Max optimization steps that max run indicates how many times to repeat the process of choosing initial k? Whereas, max optimization step shows how many iterations performed for one run of k-Means?

iii. Multiply: copy output of the cluster and distribute it into two evaluation operators (Item Distribution Of Performance and Cluster Distance Performance).

iv. Sum of Squares "Item Distribution Of Performance":out put of this evaluation input into "Gini Coeffecient" operator.

v. Gini Coeffecient "Item Distribution Of Performance": the result of this operator introduced into log operator.

vi. Cluster Distance Performance: also result of this operator introduced into "log" operator. This operator set to be maximize to show positive value.

vii. Log: This operator will save results into log table into a text file.

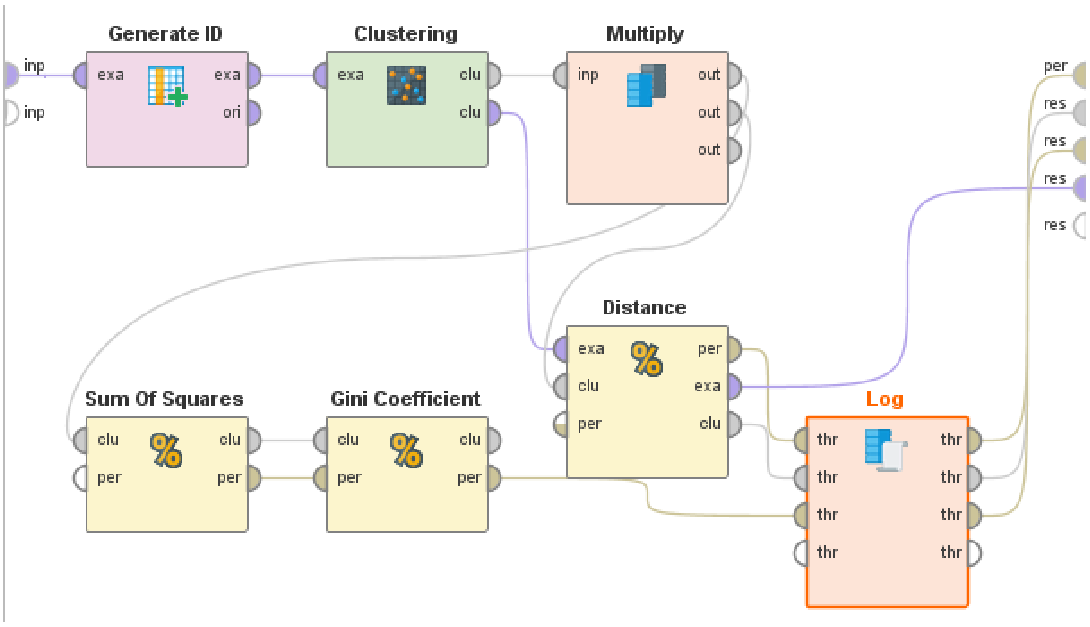

Figure 7: Cluster operator and evaluation operators inside "loop parameters"

3- The final step was introducing loop parameters into "log to data "operator. This log to data will create an example set of the data as it appeared in figure 8.

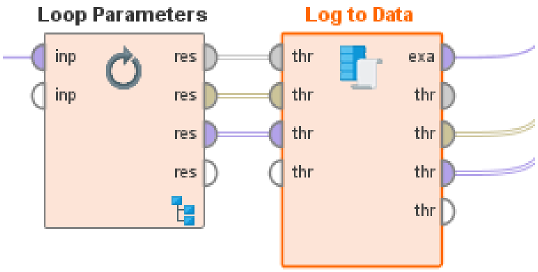

Figure 8: Adding" log to Data" operator

4- Figure 9 shows overall operators done in RapidMiner.

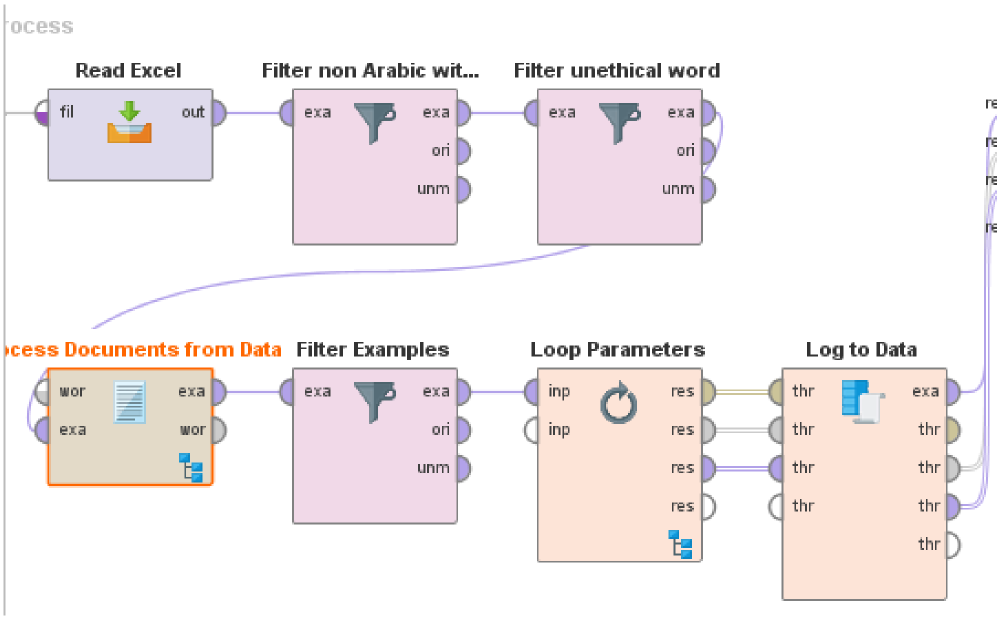

Figure 9: Overview of the processes done in rabid miner

Chapter 4

4. Experimental Analysis

In this chapter all of the investigations, analyses and outcomes will be discussed about دبي (Dubai) tweets.

4.1 Evaluation by internal clustering evaluation techniques

After getting the results of the clusters, 4 internal evaluation used to find optimal value of k. By contrast the results got from internal clustering are not helpful to determine best value of k. Because only two k found from 4 evaluation techniques. These two k were different and very far from each other.

The data used in this section is whole of the data which is after preprocessing 93,850 examples become 57,173 examples. Internal evaluation are Gini Coefficient, average within centroid, Davies Bouldin index (DBI) and sum of square errors. To observe results from internal cluster evaluation techniques, flow charts are used to illustrate results.

For Sum of Squares error elbow methodology will use to find out best value of k. When value of SSE decreased gradually before increasing value of k, elbow of the chart will be best value of k.

From figure10, it is clear that the value of k is vague and difficult to recognize from the chart because there is no elbow. Moreover, there is no need to add more clusters to the figure to find out elbow. Because value of elbow possible to be when k= 10, where SSE values declined gradually before value of k grew, but with going up moderately of SSE in k= 11 desolate elbow.

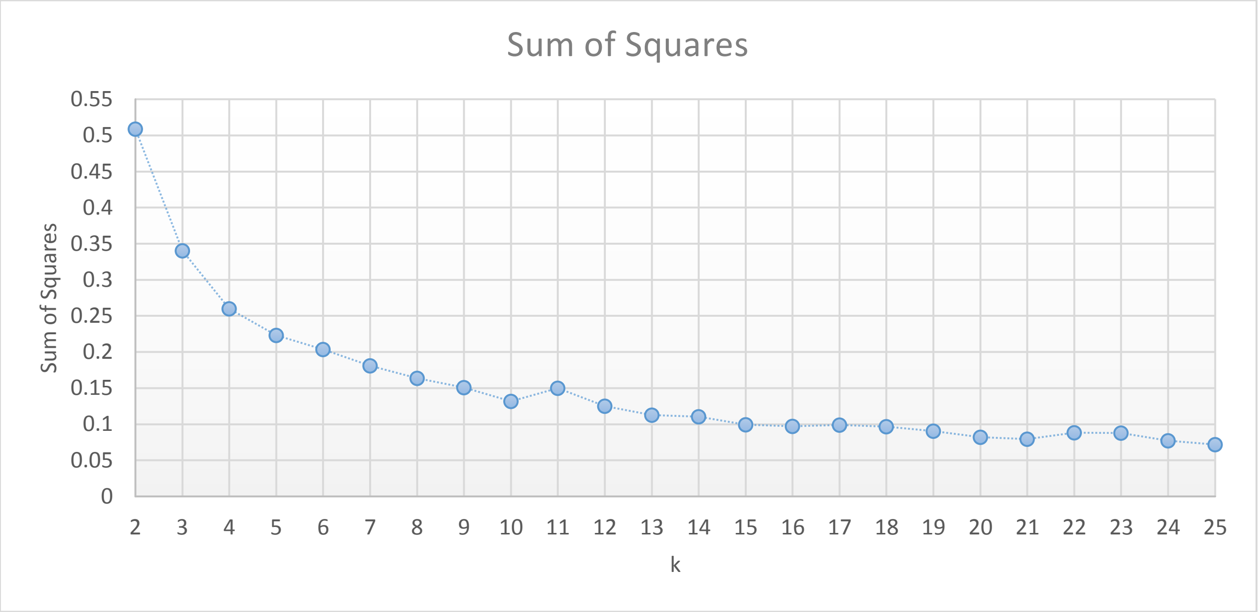

Figure 10: Sum of Squared results from k=2 to k=25

Another evaluation technique is Davies Bouldin Index (DBI). DBI measures homogeneity between intra cluster’s tweets and heterogeneity between inter cluster’s tweets.

When the value of DBI is the lowest, the best cluster is obtained. As it appeared in figure 11, the lowest value of DBI when k= 2. However, logically for thousands number of tweets two topics will be extracted considered unbelievable, but the topics of k=2 will be explained in section 4.2.1. For this problem number of k is expanded to be 30 to see if the value of DBI will decrease to be the lowest. The value of DBI when k= 2 is 4.9, whereas the value of DBI in k= 25 is 5.3, so it is possible to went down.

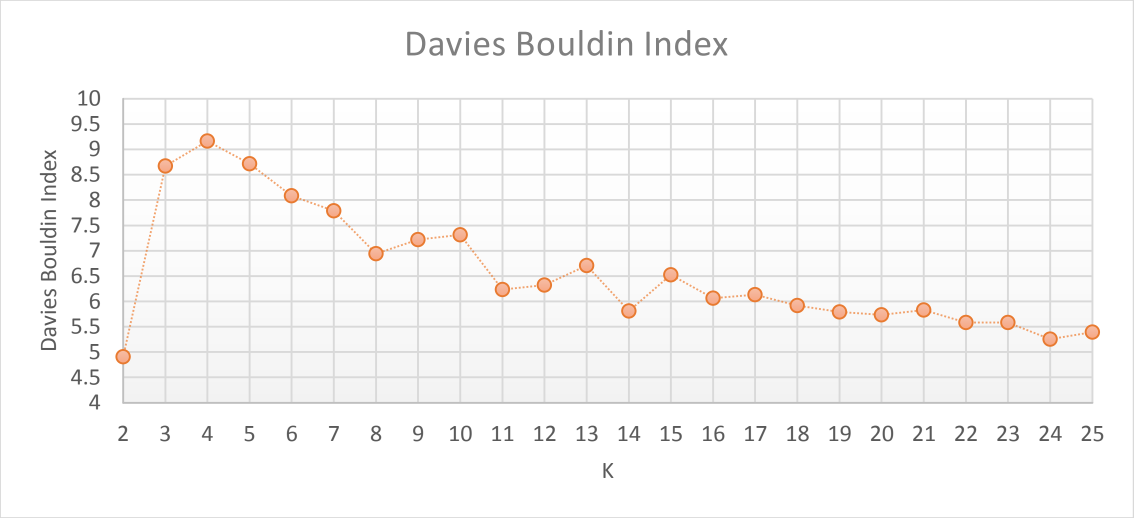

Figure 11: Davies Bouldin Index results from k=2 to k=25

For average within centroid, elbow methodology also will be used to check optimal value of k in elbow. As it shown in figure 12, values of average within centroid are fluctuated over cluster from k=2 to k=25. As a result, there is no elbow for average within centroid. The values of average within centroid are very close where it vary from 0.92981 to 0.788545. Furthermore, I didn’t think that if the number of cluster expand the values of average within centroid will form elbow.

There is another discussion about best value of average within centroid, where smallest value of average within centroid gave more compatible cluster reported by RapidMiner (2017). From figure 12 it is clear that the minimum values of average within centroid is 0.788545 when k= 25. Therefore best value for k indicated by this method is k=25. Value of average within centroid could be declined as the k increased as it was 0.92981 when k= 2 and 0.788545 when k=25.

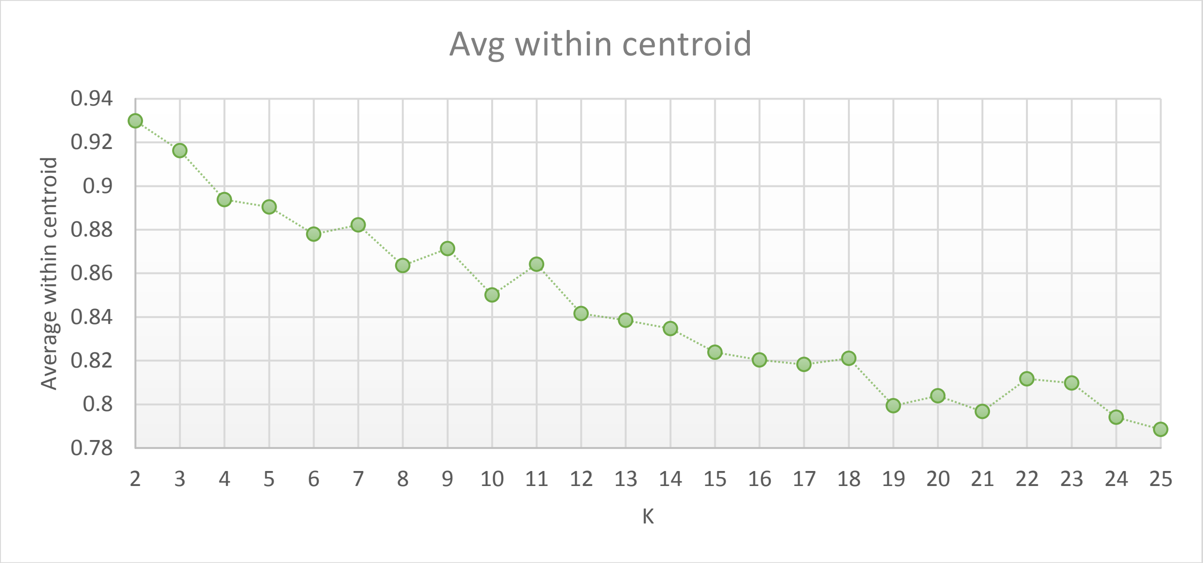

Figure 12: Average within centroid results from k=2 to k=25

Regarding the last internal evaluation technique, which is Gini index, this technique measures equality distribution of the values in the clusters. When the value of Gini index is close to 1, this represents inequality of distributing tweets inside the cluster, whereas when Gini index is close to 0, tweets inside cluster are distributing equally. As it appeared in figure 13, all of the values of Gini index are close to1, which indicates inequality of distributing tweets among clusters. Gini index varies from 0.99999768 to 0.99986596 along the period of k from k=2 to k=25.

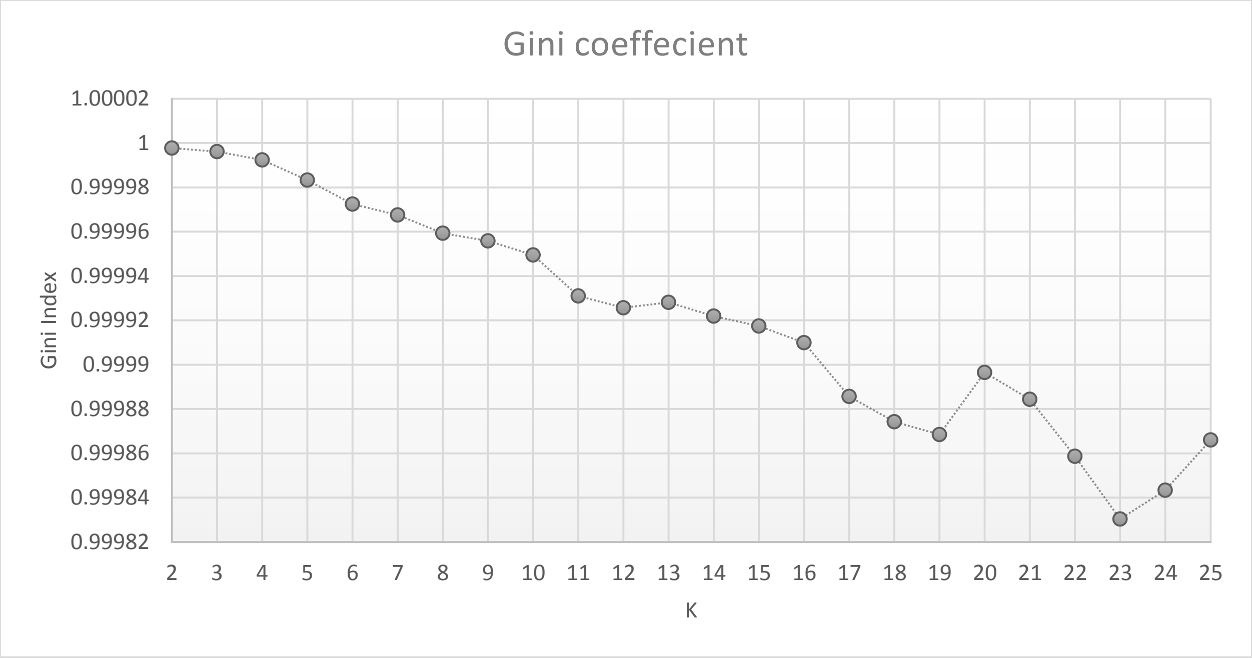

Figure 13: Gini Coefficient results from k=2 to k=25

4.2 Evaluation by observing different results

Internal evaluation failed to agree on the best number of k, so this evaluation methodology will introduce here, which is observing different results. By observing different k and figuring out the topics inside several clusters, will notice best k which shows trending topics about دبي (Dubai). Additionally, by manipulating with the parameters used in Rabid Miner the results will be different. Manipulating with the parameter such as using “Filter Tokens (by Region)” operator, will gave better results than not using this operator.

There are several notes will be followed for the following experiments:

1- “@ username” and some mobile numbers were removed from the tweets only in the examples not in the experiments for privacy issue.

2- Tweets that shared the same topic will be explained only once.

3- Only one example of unclear tweets will be explained.

4.2.1 Experiment 1

In this experiment k =2 will illustrated in more details as it is the best value of k extracted from Davies Bouldin Index. Table 6 reveals the results got from k=2

Table 6: Results of k=2

K	Size of data	Top terms	Topic extracted
Cluster 0	24,795	دبي ، للدعم ، الله، محاكم، محاكم_دبي، أمل ،أمل_دبي، أمل_دبي_للدعم	مبادرة أمل دبي للدعم
Cluster 1	32,381	الإمارات ،شباب، الامارات،السعودية ، شرطة ، شرطة_دبي،قطر ، الرياض، الأهلي، الأهلي_دبي	مباراة الإمارات و السعودية
Translation			
Cluster 0	24,795	Dubai, support, Allah, courts, Dubai_courts, amal, Dubai_amal, amal_Dubai _support	Amal Initiative in Dubai
Cluster 1	32,381	UAE, Shabab, UAE, Saudi Arabia, Police, , Dubai_Police, Qatar, Riyadh, Ahli, Dubai_Ahli	UAE and Saudi Arabia match

As it shown, topic extracted from cluster 0 and cluster 1 are clear and could be guessed easily. Cluster 0 will be explained firstly. Cluster 0 consists of 24,795 tweets which is a big number. Some of the tweets inside this cluster are shown in table 7:

Table 7: Examples of tweets in cluster 0, k=2

Examples of tweets in cluster 0 in Arabic	Examples of tweets in cluster 0 in English
#قروب_سيمو_للدعم #امل_دبي_للدعم https://t.co/XLrEz06MmW	# Group_semo_for support # Amal_Dubai _support https://t.co/XLrEz06MmW
🎉🎉🎉🎉🎉🎉🎉🎉 إبداع 🌹💫🌟 #دعم_خاص ╭━🌟━╯ تميّز 🌹💫🌟 ╭━🌟━╯ تألق 🌹⭐💫🌟 https://t.c… امل_دبي_للدعم# _ حساب ـ يستحق_المتابعة#	🎉🎉🎉🎉🎉🎉🎉🎉 #Special_support 🌟💫🌹creativity ╭━🌟━╯ 🌟💫🌹Distinguish ╭━🌟━╯ 🌟💫⭐🌹shine on # Account_worth_following # Amal_Dubai _support https://t.c…
الإنسان الذي لايفهم صمتك # لن يفهمك حين تتحدث #انين_الصمت #امل_دبي_للدعم #نلتقي_لنرتقي_للدعم …قروب_ابوب#اصدقاء_امل#	#Human who doesn’t understand your silence , will not understand you when you speak # Silence_whine # Amal_Dubai _support # We_meet_to_get_up_support # Friends_of_amal_support
محاكم #دبي تطلق مبادرة #أبش https://t.co/hZFH5xHuOf	Dubai_courts# launch Abshar initiative https://t.co/hZFH5xHuOf

As it displayed in table 7, tweets are close to each other, except the last tweet, which was similar to the other tweets, but have different topic. Although the last tweet contains token محاكم (courts), which shared with the sup-topic in token مبادرة(initiative), this is not the main topic. In spite of that, the topic could be guessed from the rest of tokens.

Table 8 presents examples of cluster 1, first three tweets shows that the tweets are related to the UAE and Saudi Arabia match. The two last tweets were talking about spirited ideas, where tweets number 4 was talking about Dubai police and Saudi Arabia, where it was inclined to Dubai police not to the UAE and Saudi Arabia match. Furthermore, tweets number 5 the major idea of it was about beauty in GCC countries. Despite that, cluster 1 was easy for estimating topic about the UAE and Saudi Arabia match.

Table 8: Examples of tweets in cluster 1, k=2

Examples of tweets in cluster 1 in Arabic	Examples of tweets in cluster 1 in English
: القنوات الناقلة لمباراة #الشباب_الوطني الرياضية السعودية-الرياضية التونسية- الاردنية -دبي الرياضية- …DMCالقناة المصرية	Ducts to match # youth_national: Sports Saudi Arabia - Sports Tunisian - Jordan - Dubai Sports - Egyptian channel Damak ...
شباب الاهلي - دبي "ديوب ولوفانور ، القوة الضاربة لخط هجوم https://t.co/ZJR7rPMYQJ . في الموسم الجديد "	Diop and Louvain, the سفقهؤن force of attacking line of Al Ahli youth -Dubai in the new season. https://t.co/ZJR7rPMYQJ
: القنوات الناقلة لبطولة #دورة_تبوك الرياضية السعودية المصرية DMC قناة دبي الرياضية الرياضية الأردنية الرياضية التونسية	Tabuk_course # Ducts for the championship Saudi Sports Egyptian DMC channel Dubai Sports Jordan Sports Tunisian sports.
تقدم بلاغ لشرطة دبي بالايميل وانت في السعوديه ويتم التواصل معك في صباح اليوم التالي وحل مشكلتك..هذا جنون# دبي	Submit a report to Dubai Police by email and you are in Saudi Arabia and you will be communicated on the next morning and your problem will be solved this madness # Dubai
السعودية #تجميل #جلدية #معارض #مناسبات #سناب # الامارات #الكويت #قطر #الخليج ##عرض #معرض #دبي https://t.co/TIT5wifhv8 …#خليجي #اخبار #مشاهير	Saudi Arabia # Beauty # Leather # Exhibitions # Occasions # Snap # Show # Exhibition # Dubai # UAE # Kuwait # Qatar # Gulf # Gulf # News # Famous ... https://t.co/TIT5wifhv8

4.2.2 Experiment 2

Experiment 2 will discuss k=10, which was supposed to be the optimal k extracted from SSE. The purpose from this experiment to investigate if increasing k will give more topics as it displayed in table 9.

Table 9: Results of k=10

K	Size of data	Top terms	Topic extracted
Cluster 0	2582	للبيع ، شقق، عقارات،للبيع _دبي، دبي_شقق، شقق_وفلل عقارات_للبيع، عقارات_للبيع_دبي، شقق_وفلل_للتمليك,	عقارات للبيع و التمليك في دبي
Cluster 1	5932	شباب ،الامارات، الأهلي، الأهلي_دبي، شباب_الأهلي، شباب_الأهلي_دبي، العين، الإمارات_دبي، فريق، السعودية	مباراة الإمارات و السعودية
Cluster 2	4434	الإمارات، مول، دبي_مول، فلاي،فلاي_دبي، الإمارات_دبي، أبوظبي ، العربية، أخبار، الوطني	غير واضح
Cluster 3	7395	مدينة ، قناة ، محمد، العالم، فنادق، قناة_دبي، بلدية_دبي، مدينة _دبي، برج ، حاكم، مشروع، طرق	غير واضح
Cluster 4	4090	مطار ،مطار_دبي ، الدولي، دبي_الدولي، الخير ، مدينه، مون، دبي_مون، تعيش	غير واضح
Cluster 5	1819	سيارات، تأجير، تاجير_سيارات، دبي_دبي، فاخرة، السيارات، تأجير_سيارات_دبي	تاجير السيارات الفاخرة في دبي
Cluster 6	4354	شرطة ، شرطة_دبي ، محاكم، تاكسي، مركز، تكلف ، دبي _تطلق ، اقتصادية _دبي، اقتصادية ، التاكسي	غير واضح

Cluster 7	5154	السعودية ، الرياض ، قطر ،للدعم، الكويت ، امل ، دبي، دبي_للدعم، امل _دبي _للدعم	غير واضح
Cluster 8	13788	دبي	غير واضح
Cluster 9	7628	الله، سوق ، والبحرين، دبي_ والبحرين، سوق_دبي	غير واضح
Translation			
Cluster 0	2582	sale, apartments, real estate, sale _Dubai, Dubai_, apartments ,apartments_and_ villas, real estate_sale, real estate_sale_ Dubai ,apartments_and_villas-owenership	Real estates for sale and ownership in Dubai
Cluster 1	5932	UAE, Al Ahly, Al Ahly_Dubai, Shabab _Al Ahli, Shabab_Al Ahly_Dubai, Al Ain, Emirates, Dubai, Team, Saudi Arabia	UAE and Saudi Arabia match
Cluster 2	4434	Emirates, Mall, Dubai_Mall, Fly, Fly_Dubai, Emirates_Dubai, Abu Dhabi, Arabic, News, National	Unclear topic
Cluster 3	7395	City, canal, mohammed, world, hotels, channel_Dubai, municipality_Dubai, Dubai_city, tower, ruler, project, roads	Unclear topic
Cluster 4	4090	Airport, Dubai Airport, International, Dubai International, benevolence, City, Moon, Dubai_Moon, Live	Unclear topic
Cluster 5	1819	Cars, Rental, Rent_Car, Dubai_Dubai, Luxury, cars, Rent-Cars_Dubai	Rent luxury cars in Dubai

Cluster 6	4354	Police, Police, Courts, Taxi, Center, Cost, Dubai_lunch, Economy, Economy_Dubai, Taxi	Unclear topic
Cluster 7	5154	Saudi Arabia, Riyadh, Qatar, support, Kuwait, Amal, Dubai, Dubai _support, Amal_Dubai support	Unclear topic
Cluster 8	13788	Dubai	Unclear topic
Cluster 9	7628	Allah, Shop, Bahrain, Dubai_ and _Bahrain, Shop_Dubai	Unclear topic

From table 9 , three topics successfully extracted from k=10, topics where about عقارات للبيع و التمليك في دبي، مباراة السعودية و الإمارات،تأجير السيارت الفاخرة في دبي ،(Real estate for sale and ownership in Dubai, UAE and Saudi Arabia match, Rent luxury cars in Dubai). Two only new topics added differ than the one extracted from k=2 they are Real estates for sale and ownership in Dubai and Rent luxury cars in Dubai. Cluster 7 shows that the general topic about Amal Initiative in Dubai although there were some tokens such as السعودية ، الرياض ، قطر (Saudi Arabia, Riyadh, Qatar), that didn't indicate that are related to the same topic, so the topic didn't extracted again in this cluster. Cluster 8 despite that number of tokens inside it was 13788, only one token extracted from it which was دبي(Dubai). Because only دبي(Dubai) achieved high weigh reached to 0.676 and the remaining tokens start with weigh of 0.005, which is too far from weight of Dubai. Cluster 0 and cluster 9 shows that filter stopwords operator didn't filter character و (and). Therefore, Filter stopwords operator wasn't affective enough to remove all of the stopwords.

Cluster 0 talks about عقارات للبيع و التمليك في دبي(Real estates for sale and ownership in Dubai) , table 10 explains examples of cluster 0. As it sown, all of the tweets are related to the main topic.

Table 10: Examples of tweets in cluster 0, k=10

Examples of tweets in cluster 0 in Arabic	Examples of tweets in cluster 0 in English
استثمر وتملك في قطاع الفنادق عقارات_دبي اسعار تبدأ من 735 الف درهم بالتقسيط https://t.co/IW3sCGRB3x	Invest and own in the hotel sector Real Estate Prices start from 735 thousand dirhams installment https://t.co/IW3sCGRB3x
فيلا_للايجار_في_دبي من 4 غرف نوم + غرفة خادمة مع # اطلالة على برج العرب في #النخلة بسعر 450.000 درهم #عقارات_دبي… https://t.co/LYPRTvuoPc	Villa_for_rent_in_Dubai 4 bedrooms + maid room with a view of Burj Al Arab in the Palm # at a price of 450,000 dirhams # Real Estate_Dubai... https://t.co/LYPRTvuoPc
يقع فندق سيتي بريميير للشقق الفندقية ع شارع الشيخ زايد،ويطل على برج خليفة،بالقرب من مول دبي ويبعد دقيقتان سيرا من محطة ا…	City Premier Hotel Apartments is located in Sheikh Zayed Road, overview of Burj Khalifa, next to Dubai Mall and 2-minute walk from ...

The main topic in cluster 5 was about تأجير السيارات الفاخرة في دبي (Rent luxury cars in Dubai). Table 11 displays some examples of the tweets about topic. All of the tweets in the example talks about the same topic and there are many company for renting luxury Cars Rental in Dubai.

Table 11: Examples of tweets in cluster 5, k=10

Examples of tweets in cluster 5 in Arabic	Examples of tweets in cluster 5 in English
شركة #تأجير_سيارات فاخرة في #دبي للحجز والاستعلام عن عروض والاسعار	# Luxury Car Rental in Dubai # To book and inquire about offers and prices # Kuwait and Saudi Arabia _ one heart

#الكويت_والسعوديه_قلب_واحد #كايلي_حا...	# Kaili_h ..
شركة #تأجير_سيارات فخمه في #دبي أسعار تنافسية تامين شامل توصيل للمطار للحجز	# Luxury Car Rental in Dubai # Competitive prices full insurance Connecting to the airport For reservations
يوجد لدينا جميع انواع السيارات للايجار في دبي سيارات☜اقتصاديه☜فخمه☜عائليه☜فارهه☜رياضيه☜يوجد... https://t.co/8O4P9AxLtB	We have all kinds of cars for rent in Dubai, cars, luxury, fancy , family, sports, economy https://t.co/8O4P9AxLtB

One of the unclear topic was cluster 2 with 4434. Table 12 illustrates some examples about this cluster. Each tweet has its own topic and they are unlike other topics.

Table 12: Examples of tweets in cluster 2, k=10

Examples of tweets in cluster 2 in Arabic	Examples of tweets in cluster 2 in English
خصم حتى 50% مع #فلاي_دبي لدرجتي رجال الأعمال والسياحية الحجز ⬅ حتى 26/9 السفر ⬅ حتى 27/10/2018 https://t.co/lIWK13Qys7	Up to 50% discount with # fly_Dubai For both business and tourism Booking ⬅ until 9/26 Travel ⬅ until 2018/10/27 https://t.co/lIWK13Qys7
أبرز عناوين نشرة أخبار الثامنة والنصف مساء من #مركز_الأخبار التابع لـ #مؤسسة_دبي_للإعلام https://t.co/ToGj26GsNQ	Highlights of the newscast at 8:30 pm from the #News_Center of the Dubai_Media_Foundation https://t.co/ToGj26GsNQ
14: #اليوم_الوطني٨٧_ احتفال بيوم عظيم وشكرا من شاركنا فرحتنا	14: # National _Day#87

الكويت - المنامة - دبي 🖤🖤🖤🖤🖤 https://t.co/a7K6RgZYss	Celebrate a great day and thank you for sharing our joy- Kuwait- Manama- Dubai 🖤🖤🖤🖤🖤 https://t.co/a7K6RgZYss
زحمة مول دبي https://t.co/xexEllfqEd	Dubai Mall Traffic: https://t.co/xexEllfqEd

Same idea for cluster 3, cluster 4, cluster 6, cluster 8 and cluster 9. These all present unclear topic, so they will not be explained.

4.2.3 Experiment 3

Experiment 3 will explain k=25, which is the best k figured out from average within centroid. The intended aim here to see if the increase of k will extract more topics as it shows in table 13.

Table 13: Results of k=25

K	Size of data	Top terms	Topic extracted
Cluster 0	1929	سيارات، تأجير، تاجير_سيارات، دبي_دبي، فاخرة، السيارات، تأجير_سيارات_دبي	تاجير السيارات الفاخرة في دبي
Cluster 1	2264	مطار ،مطار_دبي ، الدولي، دبي_الدولي، الخير ، عام ،مطار_دبي_الدولي ،اهلي_دبي، الخير_دبي	مطار دبي الدولي
Cluster 2	11816	دبي	غير واضح
Cluster 3	962	التاكسي، قريبا ، الطائر، و عجمان ، رأس ، التاكسي_الطائر	خدمة التاكسي الطائر في دبي
Cluster 4	2754	فنادق ، فنادق_دبي، العالمي، مركز، لندن، تستضيف، تستضيف_دبي	غير واضح

Cluster 5	1316	محاكم ، محاكم_دبي ،اقتصادية، اقتصادية_دبي، الذكية ، تطلق،جيتكس، جيتكس_دبي	غير واضح
Cluster 6	1075	سوق ، سوق_دبي، مليار ، مشاريع، المالي، دبي _المالي	سوق دبي المالي
Cluster 7	1539	للبيع ، شقق، عقارات،للبيع _دبي، دبي_شقق، شقق_وفلل ,عقارات_للبيع، عقارات_للبيع_دبي، شقق_وفلل_للتمليك	عقارات للبيع و التمليك في دبي
Cluster 8	1256	قناة ،قناة_دبي، زمان، دبي_زمان، سما ، سما_دبي، تلفزيون، برنامج ، تلفزيون_دبي، نور ، نور_دبي	قنوات دبي وبرامج قنوات دبي
Cluster 9	2429	الامارات، الامارات_دبي، السعوية، دبي_الامارات،ابوظبي، دبي_ابوظبي،السعودية_دبي	غير واضح
Cluster 10	2203	والله، خلاص،مكان ، اروح	غير واضح
Cluster 11	3015	السعودية ، الرياض ، قطر،قطر_دبي، الكويت،السعودية_قطر	غير واضح
Cluster 12	1603	شركة ، الشباب، ولي، عهد، ولي_عهد_دبي، الشباب_دبي ، مجلس _دبي،شهر ، غرفة	غير واضح
Cluster 13	2245	البحرين ، دبي_و البحرين	غير واضح
Cluster 14	1406	مدينة، تاكسي،مدينة_دبي، رحلة،الجوي،طائر، تاكسي_دبي، تختبر، دبي_تختبر، تاكسي_دبي _الجوي	خدمة التاكسي الطائر في دبي
Cluster 15	1194	الله، حكومة، حكومة_دبي، دبي، الله ،يارب	غير واضح
Cluster 16	1993	امل ، دبي، دبي_للدعم، امل _دبي _للدعم	مبادرة أمل دبي للدعم
Cluster 17	2373	الإمارات	غير واضح
Cluster 18	2425	محمد ،مدينة ، تعيش، حاكم ،دار، حاكم _دبي، راشد	الشيخ محمد بن راشد حاكم دبي

Cluster 19	2182	البحرين	غير واضح
Cluster 20	2195	شرطة ، شرطة_دبي، تحدي،تحدي_دبي، دبي_للياقة، تحدي_دبي للياقة	تحدي دبي للياقة
Cluster 21	2045	بلدية ، بلدية _دبي،طرق،طرق_دبي، مشروع ، العالمية، مليون ، مؤسسة	بلدية دبي
Cluster 22	1597	مول ، دبي _مول ، فندق، برج ، جدة،مول _دبي ، خليفة	مباني دبي الشهيرة
Cluster 23	2729	شباب، الأهلي، الاهلي_دبي، شباب_الأهلي_دبي، الاهلي،	مباراة الإمارات و السعودية
Cluster 24	631	بدون، الفيديو، بفندق_دبي ،بفندق_دبي_الفيديو	غير واضح
Translation			
Cluster 0	1929	Cars, Rental, Rent_Car, Dubai_Dubai, Luxury, cars, Rent-Cars_Dubai	Rent luxury cars in Dubai
Cluster 1	2264	Airport, Dubai _Airport, International, Dubai_ International, benevolence, year, International_ Dubai _Airport ,My family _Dubai, benevolence_Dubai	Dubai International Airport
Cluster 2	11816	Dubai	Unclear
Cluster 3	962	taxi, soon, the bird, and Ajman, head, Flying_taxi	Flying taxi service in Dubai
Cluster 4	2754	Hotels, Dubai_hotels, world, center, london, hosts, Dubai_ hosts	Unclear

Cluster 5	1316	Courts, Duba_Courts, Economic, Dubai_Economic, , Smart, Launches, GITEX, GITEX_Dubai	Unclear
Cluster 6	1075	Market, market_Dubai, billion, projects, financial, Dubai _ Mali	Dubai Financial Market
Cluster 7	1539	sale, apartments, real estate, sale _Dubai, Dubai_, apartments ,apartments_and_ villas, real estate_sale, real estate_sale_ Dubai ,apartments_and_villas_owenership	Real estates for sale and ownership in Dubai
Cluster 8	1256	Channel, Dubai_channel, zaman,dubai, zaman, , sama, Sama_Dubai, TV, program, TV_Dubai,K Noor, Noor_Dubai	Dubai channels and its programs
Cluster 9	2429	UAE, UAE_Dubai, Suadi Arabia, Dubai_UAE, Abu Dhabi, Dubai- Abu Dhabi, Suadi Arabia_ Dubai	Unclear
Cluster 10	2203	Allah, finish, place, go	Unclear
Cluster 11	3015	Saudi Arabia, Riyadh, Qatar, Qatar_Dubai, Kuwait ,Saudi Arabia_Qatar	Unclear
Cluster 12	1603	Company, Youth, Crown ,Prince, Dubai_Crown_ Prince, Youth, Dubai,Dubai _council, month , room	Unclear
Cluster 13	2245	Bahrain, Dubai _and _Bahrain	Unclear

Cluster 14	1406	City, Taxi,Duba_ City, Flight, Air, Bird, Dubai_Taxi, Experiment, Dubai_Test, Duba_Air_Taxi	Flying taxi service in Dubai
Cluster 15	1194	Allah, Government, Dubai_Government, Dubai, Allah, Ya Allah	Unclear
Cluster 16	1993	Amal, Dubai, Dubai _support, Amal_Dubai support	Amal Initiative in Dubai
Cluster 17	2373	Emirates	Unclear
Cluster 18	2425	Mohammed, city, living, ruler, home, Dubai_ruler, Rashid	Sheikh Mohammed bin Rashid Ruler of Dubai
Cluster 19	2182	Bahrain	Unclear
Cluster 20	2195	Police, Dubai _Police, Challenge, Dubai_Challenge, Dubai, Fitness, Dubai_ Fitness_ Challenge	Dubai Fitness Challenge
Cluster 21	2045	Municipality, Dubai _Municipality, Roads, Dubai _Roads, Project, International, Million, Institution	Dubai Municipality
Cluster 22	1597	Mall, Dubai _ Mall, Hotel, Burj, Jeddah, Mall_Dubai, Khalifa	Famous buildings in Dubai
Cluster 23	2729	Shabab, Al Ahli, Al Ahly_Dubai, Shabab_Alahly_Dubai, Al Ahli	UAE and Saudi Arabia match
Cluster 24	631	Without, video, Dubai_ hotel, Dubai_ hotel _video	Unclear

From table 13, there are 13 topics extracted, 2 of the topics are extracted twice, so 12 various topics are extracted. There are four topics extracted in experiment 1 and 2. Generally, there are eight new topics extracted in experiment 3 when increasing number of clusters from 10 to 25. The following are examples of 8 topics which are about مطار دبي الدولي، سوق دبي المالي ،قنوات دبي و برامج قنوات دبي ،خدمة التاكسي الطائر في دبي ،الشيخ محمد بن راشد حاكم دبي ، تحدي دبي للياقة ،بلدية دبي و معالم دبي الشهيرة(Flying taxi service in Dubai, Dubai International Airport, Dubai Financial Market, Dubai channels and its programs, Sheikh Mohammed bin Rashid Ruler of Dubai, Dubai Fitness Challenge, Dubai Municipality and Famous Dubai landmarks).

There are some problems in the efficiency of the methods used, but didn’t affect too much in the results such as the same word written twice by using N-gram ex: دبي_دبي (Dubai_Dubai). Another problem that some of the N-gram tokens are the same, but they are in reverse order such as الامارات_دبي، دبي_الامارات (UAE_Dubai, Dubai_UAE). These two tokens must be written only once, they are carrying the same meaning, but they are written twice with different weights 0.119 and 0.103. Furthermore, one of the most important events missed to be extracted as a topic, which was جيتكس في دبي (GITIX in Dubai). Despite that, جيتكس (GITIX) mentioned in cluster 5, it was missed to be a topic due to several tokens that talks about different topics.

Table 14 reveals some examples about مطار دبي الدولي (Dubai International Airport). Although there are some examples didn’t talk about the same topic, they didn’t affect in guessing that the main topic, which about مطار دبي الدولي (Dubai International Airport).

Table 14: Examples of tweets in cluster2, k=25

Examples of tweets in cluster 2 in Arabic	Examples of tweets in cluster 2 in English
اليوم_الوطني #ختم مطار دبي الدولي اليوم بالتزامن مع https://t.co/Jdv54jzSXS شكرا إمارات العطاء #معآ_أبدآ	Dubai International Airport seal today in conjunction with # National Day Thanks Emirates Al-Bid # together_forever https://t.co/Jdv54jzS
ويلومونا في محبتنا للإماراتيين حكومة وشعب ♥□ شاهد استقبالهم لرحلات الخطوط السعوديه في مطار دبي بمناسبة اليوم الوطني💓…	They blame us in our love for the Emirati government and people □ ♥ See their reception for Saudi Airlines flights at Dubai Airport on the occasion of National Day 💓 ...
اخبار الامارات : دبي للثقافة تعلن إطلاق مسابقة أفضل نص مسرحي لعام الخير 2017 https://t.co/FMr9Duxpbl https://t.co/GHZpAnNU37	Dubai Culture announces the launch of the best script competition for the year of good 2017 https://t.co/FMr9Duxpbl https://t.co/GHZpAnNU37
أهلي دبي ♥ https://t.co/lSqyuRUKkT	My family Dubai ♥ https://t.co/ISqyuRUKkT

Table 15 reveals examples about the topic خدمة التاكسي الطائر في دبي (Flying taxi service in Dubai). This topic extracted twice under the same topic in cluster 3 and cluster 14.

Table 15: Examples of tweets in cluster 3 and cluster 14, k=25

Examples of tweets in cluster 3 and cluster 14 in Arabic	Examples of tweets in cluster 3 and cluster 14 in English
التاكسي الطائر X2 بـ 18 مروحية في دبي قريبا .. فيديو وصور https://t.co/ecMinKbOY0	Flying Taxi X2 by 18 Helicopter in Dubai Soon .. Video and Photos https://t.co/ecMinKbOY0

بحضور حمدان بن محمد .. طرق دبي تجري أول رحلة تجريبية لـ«التاكسي الجوي ذاتي القيادة» بالقرب من حديقة شاطئ جميرا في #دبي	In the presence of Hamdan bin Mohammed .. Dubai Roads conducts the first pilot flight for the «Air Taxi self-driving» near Jumeirah Beach Park in # Dubai
دبي : حمدان بن محمد يشهد أول رحلة تجريبية للتاكسي الجوي # ذاتي القيادة وهو الأول من نوعه في العالم https://t.co/… التاكسي_الجوي#	#Dubai: Hamdan bin Mohammed attends the first pilot flight for the self-driving air taxi, the first of its kind in the world #Account Taxi https://t.co/...
بالفيديو.. دبي تختبر أول تاكسي طائر في العالم بدون قائد عبر AlArabiya https://t.co/	Video: Dubai tests the world's first taxis without a captain via AlArabiya https://t.co/

Majority of the examples are talking about the same topic in both cluster 3 and cluster 14. These two clusters must be one cluster, but they are separated, when they are joined into one cluster, they will form 1406+962= 2,368 tweets.

For table 16, the main topic about سوق دبي المالي(Dubai Financial Market). As it shown all of the examples are related to the main topic.

Table 16: Examples of tweets in cluster 6, k=25

Examples of tweets in cluster 6 in Arabic	Examples of tweets in cluster 6 in English
تقنيات متطورة تقدمها لك الإمارات دبي الوطني للأوراق المالية عبر نظام eBrokerPLUS للتداول تعرف على المزيد.... https://t.co/k8yunjUA2m	Advanced technologies offered by Emirates NBD Securities via the eBrokerPLUS Trading System Learn more ... https://t.co/k8yunjUA2m
وقفة مع التحليل المالي والرصد اليومي لتداولات أسواق المال في الإمارات يأتيكم يومياً على #سما_دبي في #تداول https://t.co/D96tzPZptj	A stand with financial analysis and daily monitoring of the UAE's stock market trading comes daily to # Sama_dpi #Trading https://t.co/D96tzPZptj

الشركات الأكثر إرتفاعا لليوم في سوق دبي المالي: AMLAK ارتفع 3.85%، ALRAMZ ارتفع 3.70%، DSI ارتفع 2.43%	The most active companies in the Dubai Financial Market today were: AMLAK gained 3.85%, ALRAMZ gained 3.70%, DSI gained 2.43%
ما هي #المشاريع الكبرى التي تنعش قطاع #العقارات في #دبي؟ https://t.co/zbtoh4gJgL https://t.co/Wr9eYnOZoT	What are the # major projects that revitalize the # Real Estate sector in Dubai #? https://t.co/zbtoh4gJgL https://t.co/Wr9eYnOZoT

To learn more about examples about the topic قنوات دبي و برامج قنوات دبي (Dubai channels and its programs) table 17 illustrates some examples. The topic was easy to extract as the TF-IDF gave the right weight for the tokens and K-mean successfully to cluster these tokens in one cluster.

Table 17: Examples of tweets in cluster 8, k=25

Examples of tweets in cluster 8 in Arabic	Examples of tweets in cluster 8 in English
التحضيرات الأخيرة ما قبل الهوا، دقايق و نكون معاكم على قناة دبي الأولى و برنامج هاش مال… https://t.co/f0t9PjpQOC	The most recent before live broadcast, minutes and we will be together on Dubai first channel on program called HASH MAL ... https://t.co/f0t9PjpQOC
برنامج #كيف_يفعلون_ذلك ما سر الوسادة الهوائية في السيارة ؟ على قناة #نور_دبي https://t.co/mzyqXNetRY	#program how to do this What is the secret of the airbag in the car? On the channel # Noor_Dubai https://t.co/mzyqXNetRY
كنت اتفرج علي برنامج الاسره في قناة سما دبي شدني مقابله مع سيده تحكي عن امها و هي ام المصورين شيخه السويد و كيف دعم بناتها و أحفادها لها	I was watching the program of the family in the Sama Dubai channel caught my attention a woman talk about her mother and she is the mother of

	photographers Sheikh Al Swaid and how his daughters and grandchildren support her
المسلسل التاريخي #مجالس_العرب وأجمل الحكايا من قلب مجلس #هارون_الرشيد يأتيكم على قناة الزمن الجميل #دبي_زمان https://t.co/fAqJpUQ7Fk	The historical series # The Arab Councils and the most beautiful stories from the heart of the board # Haroon_Al Rasheed comes to you on the channel of beautiful time # Dubai_zaman https://t.co/fAqJpUQ7Fk

The next topic is about الشيخ محمد بن راشد حاكم دبي (Sheikh Mohammed bin Rashid Ruler of Dubai). Examples of this topic shown in table 18. All of the examples talks about H H Sheikh Mohammed bin Rashid Ruler of Dubai and his achievements in Dubai.

Table 18 : Examples of tweets in cluster 18, k=25

Examples of tweets in cluster 18 in Arabic	Examples of tweets in cluster 18 in English
لحظة الإعلان لاختيار مدينة دبي لاستضافة المؤتمر الدولي للفضاء في عام 2020م، وكلمة ترحيبية لمركز محمد بن راشد للمشار… https://t.co/hQXdDuKFRb	The announcement of the selection of Dubai city to host International Space Conference in 2020, and a welcome speech to the Mohammed bin Rashid Center for ... https://t.co/hQXdDuKFRb
محمد بن راشد يتفقد سير العمل بمشروع البرج الجديد في خور دبي https://t.co/d07sLD78F7	Mohammed bin Rashid reviews the progress of the new tower project in Dubai Creek https://t.co/d07sLD78F7
محمد بن راشد: #الإمارات تصنع الفرح وتعزز قيم التسامح https://t.co/fmsgC6ShXh #أخبار_دبي #دبي https://t.co/kwpeEnFdX3	Mohammed bin Rashid: # UAE makes joy and promotes tolerance values https://t.co/fmsgC6ShXh # News_dpi # Dubai https://t.co/kwpeEnFdX3

قال سمو الشيخ محمد بن راشد حاكم دبي ـ حفظه الله، ورعاه ـ في أبيات جميلة يصور حاله مع رب العالمين حين لبس الإحرام،... https://t.co/JWTAVXvPk3	His Highness Sheikh Mohammed bin Rashid Al Maktoum, Ruler of Dubai May God protect him, and sponsored him said in beautiful verses depicting his situation with the Allah when he was wearing Ihram https://t.co/JWTAVXvPk3

One of the topic elicited from k =25 was تحدي دبي للياقة (Dubai Fitness Challenge) shown in table 19. This topic collected as last phase of collecting data on 24-Septemper-2017 and continued to 5- October- 2017. As a result, these tweets collected in October where Dubai Fitness Challenge occurred in 20 –October until 18- November.

Table 19: Examples of tweets in cluster 20, k=25

Examples of tweets in cluster 20 in Arabic	Examples of tweets in cluster 20 in English
شاركوا في تحدي دبي للياقة من ٢٠ أكتوبر إلى ١٨ نوفمبر https://t.co/Cktn14W37n #Dubai30x30 https://t.co/YOnq66HU9T	Participate in the Dubai Fitness Challenge from 20 October to 18 November https://t.co/Cktn14W37n # Dubai30x30 https://t.co/YNNq66HU9T
سمو الشيخ حمدان بن محمد آل مكتوم يطلق الآن تحدي دبي للياقة Dubai30x30 https://t.co/jJtH7veruI#	HH Sheikh Hamdan Bin Mohammed Al Maktoum launches Dubai Fitness Challenge # Dubai30x30 https://t.co/jJtH7veruI
لجعل #دبي المدينة الأكثر نشاطا.. الشيخ #حمدان_بن_محمد يطلق مبادرة "تحدي دبي للياقة" 30 دقيقة لمدة 30 يوما شاهد الفي… https://t.co/jhZNS14FdL	To make # Dubai the most active city .. Sheikh # Hamdan_Ben_Mahamad launches Dubai Fitness Challenge 30 minutes for 30 days Watch the ... https://t.co/jhZNS14FdL

حمدان بن محمد: اتحدى شرطة #دبي بجميع اداراتها للمشاركة في الذي يبدأ اعتباراً من 20 تحدي دبي للياقة البدنية…	Hamdan Bin Mohammed: I challenge Dubai Police with all its departments to participate in Dubai Fitness Challenge starting from 20 ...

The topic which talks about بلدية دبي (Dubai Municipality), explains with more examples in table 20. One of the examples shown in the table below that the main topic of it about Sheikh Mohammed Bin Zayed Road, which is also consider one of the Dubai Municipality responsibility. Therefore, all of the examples are talks about Dubai Municipality.

Table 20:Examples of tweets in cluster 21, k=25

Examples of tweets in cluster 21 in Arabic	Examples of tweets in cluster 21 in English
خريطتي هو نظام إلكتروني لإصدار وتجديد الخرائط من بلدية : دبي والذي يتيح لحائزي الأراضي الحصول على الخارطة الموقعية. https…	My map is an electronic system for the issuance and renewal of maps from Dubai Municipality, which allows land holders to obtain the map of the site. https ...
طرق دبي' تطلق مبادرة 'يوم القيادي الشاب' https://t.co/ug4FFXB4T4 #وام	Dubai Roads' launches the 'Young Leadership Day' initiative https://t.co/ug4FFXB4T4 # WAM
شرطة_دبي: ضباب على شارع #الإمارات بعد جسر القدرة # وعلى شارع الشيخ محمد بن زايد بعدالقرية العالمية بالاتجاه إلى #ابوظبي يرجى توخي الحيطة.	# Dubai Police: fog on the street # Emirates after the power bridge and on Sheikh Mohammed Bin Zayed Road after the international village in the direction to # Abu Dhabi Please be careful.
دبي الجنوب تعيّن "القبضة" شريكاً مقاولاً لمشروع "النبض" السكني بعقد قيمته 600 مليون درهم (2/1) https://t.co/6rT5Ajsv0a	Dubai Al-Qabida appoints "Al Qabida" as a contracting partner for the "Pulse" residential project with a contract worth 600 million dirhams (1/2) https://t.co/6rT5Ajsv0a

One of the trending topics about Dubai talks about مباني دبي الشهيرة (Famous buildings in Dubai). Cluster 22 provide this topic, which explained in more details in table 21. Although there are some examples are not related to the main topic directly, still the main idea about the Famous buildings in Dubai Overwhelmed over the other topics.

Table 21: Examples of tweets in cluster 22, k=25

Examples of tweets in cluster 22 in Arabic	Examples of tweets in cluster 22 in English
فلل فاخرة بدبي هارتلاند🌟 تبدأ من 5 مليون درهم شقق فاخرة 🌟 تبدأ من ٧٥٠ الف درهم كلم3 ...برج خليفة ودبي مول و قناة دبي المائية	Luxury villas in Dubai Heartland Starting from AED 5 million Luxury apartments Starting from 750 thousand dirhams 3 km Burj Khalifa, Dubai Mall and Dubai Water Channel
□□😂 !!المعلومة جميلة لكن شكل برج دبي تغير علي https://t.co/VxMw3yQxo3	The information is beautiful but the shape of Burj Dubai has changed on me !! 😂□□ https://t.co/VxMw3yQxo3
صورة ليلية لبرج خليفة بعدسة جالكسي نوت 8 https://t.co/mIpagvp1wm تقنية #برج_خليفة #دبي#	Night Image of Burj Khalifa Tower # Technology # Burj Khalifa # Dubai https://t.co/mIpagvp1wm
الحريصون على اقتراح عناوين قابلة للتطبيق وذات مراجع متوفرة. #جامعة_الشارقة #جامعة_دبي #جدة… https://t.co/b5yYuCfDgF	Who are keen to suggest applicable and available titles. # University of Sharjah # University of Dubai # Jeddah ... https://t.co/b5yYuCfDgF

4.2.4 Experiment 4

In this experiment, additional k will be added to the experiment to discover if there any lowest point for Davies Bouldin Index than k=2 .Therefore, in figure14, k is expanded from k=25 to k=31. As it shown from the chart the lowest point was k=2 with DBI=4.908333. However, after adding k=26, 27,28,29,30 and 31, the outcome indicates that k=30 has the lowest DBI with 4.858839 and it is set to be the optimal k. Then, DBI return to increase when k=31. When k= 29 , DBI = -∞ , which will not consider as the best value of k. DBI gave this result because 28 clusters in k=29 are empty , which means cluster didn't have any tweets, only one cluster had all of the tweets which equal to 57,173. Therefore, results of DBI gave infinity value, but the main reason of the strange behavior that makes this tweets distribute over one cluster is still vague.

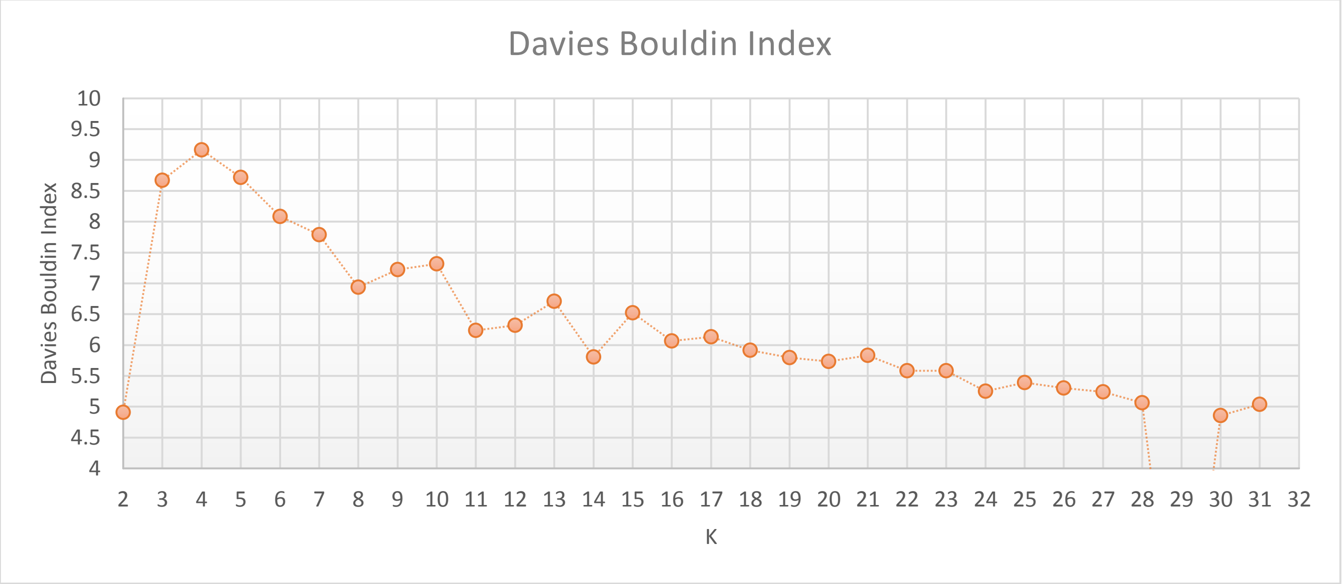

Figure 14: Davies Bouldin Index after adding k from 26 to 31

The following table will illustrate k= 30 with more details as it now carry the best value for k. Unclear topics and redundant topics will not be displayed in table 22, so only new topics will be displayed and explained.

Table 22: Results of k =30

K	Size of data	Top terms	Topic extracted
Cluster 11	728	مدينه ، تعيش، دار،اختار، مدينه_تعيش،الحي، اختار_مدينه ، دار_الحي، دبي_دار، اختار_مدينة_تعيش، دبي_دار_الحي، دانة الدنيا	اختر مدينة دبي للعيش
Translation			
Cluster 11	728	City, live, home, choose, city_live, neighborhoods, select _city, house_alhay, dubai_house, select_city_live, dubai_house_alhay, world pearl	Select Dubai to live

From table 22, the only one topic elicited from k= 30 which is اختر مدينة دبي للعيش (Select Dubai to live). Some of the words such as تعيش و مدينة(live and city) where mentioned in k=25, cluster 18, but there are more words that carry different topics in the same cluster, which make it difficult to be elicited this topic. There were two topics clustered twice which was about عقارات للبيع و التمليك في دبي و مباراة الإمارات و السعودية (Real estates for sale and ownership in Dubai and match of the UAE and Saudi Arabia).

Table 23: Examples of tweets in cluster 11, k=30

Examples of tweets in cluster 11 in Arabic	Examples of tweets in cluster 11 in English

□s♥قدمتم لنا وطن نعيش فيه ؛ نقدم لكم قلوبنا لتعيشوا فيها #فرحة_دبي_بيوم_الوطن□s https://t.co/ROpiwIMHf0	You have given us a home to live in; we offer you our hearts to live in #happy_Dubai_National_Day (SA) https://t.co/ROpiwIMHf0
مزيدا من التقدم والازدهار لدارنا دار الحي دبي	Further progress and prosperity for our Dubai
دبي دانة الدنيا و عشق ماله اخر	Dubai Pearl of the world and the love without end
المقدمه سألت عن دول جديده يتمنون يسوون فيها جوله ، فهو اختار دبي والسعوديه♥ https://t.co/eUjSVu7rSR	Programmer asked about new countries wishing to settle the tour, so he chose Dubai and Saudi Arabia ♥ https://t.co/eUjSVu7rSR

From table 23, it seems that increasing k more than 30 , will not generate more new topics because it only produce one topic in k=30. There are 13 topics extracted from both k=25 and k=30. One of the 13 topics extracted twice from k=25 is, however, 2 topics from 13 topics of k=30 are repeated twice. Because k=30 generated only one new topic and it was optimal k extracted from Davies Bouldin Index and when k=40, DBI again shows result -∞ like results gain from k=29, k will not be incremented more.

4.2.5 Experiment 5

This experiment intend to highlight the significance of “Filter Tokens (by Region)” operator, which used as a preprocessing operator to lookup for the token دبي (Dubai) in specific range. Experiment 5 conducted without using “Filter Tokens (by Region)” operator which is a sup-operator in “Process Documents from Data” operator , but all of the rest operators remain the same as in section 3.7 RabidMiner Implementation. This experiment intended to notice if this operator makes differences to the results collected from internal evaluation. Results shown in table 25 only 3 internal evaluation techniques because SSE required to have set of sequential number of clusters to find Elbow, but in this experiment only three clusters chosen due to the time limitation. DBI, average within centroids and Gini Coeffecient are giving better results when their values are the lowest. As a result, from table 24 results with using "Filter Tokens (by Region)" operator was better than result gained from without using "Filter Tokens (by Region)" operator as in table 25. DBI results from using "Filter Tokens (by Region)" operator shows lower results in all k with high difference such as when k= 10 in in table 24, DBI was 7.3 which is approximately equal result gain from without using "Filter Tokens (by Region)" operator in k= 25 with DBI=7.8. Furthermore, average within centroid in table 24 when k=10 achieved lower results with 0.85 than in table 25 which was 0.94. Additionally, Gini Coefficient when it goes to 0 objects are distributed equally in the cluster. From table 24, the results collected indicated lower value of Gini Coefficient than results collected from table 25.

Table 24: Results with using Filter Tokens (by Region)

Davies Bouldin Index	Avg within centroid	Gini coefficient	k
7.316437871431515	0.8501007585224688	0.9999495197027088	10
5.392540137252953	0.788544693558449	0.9998659574997879	25
4.858839448798699	0.7832210844553535	0.999807819903459	30

Table 25: Results without using Filter Tokens (by Region)

Davies Bouldin Index	Avg within centroid	Gini coefficient	k
9.799170603766306	0.9498353111258042	0.9999595668838744	10
7.858870342420824	0.9091595630346907	0.9999219515050107	25
6.884614864072552	0.8954044861597665	0.9998797134696992	30

4.3 Discover trending topics in four months about دبي

This part focus to read about what people interesting to tweet about دبي (Dubai) in four months that the data collected on. Four months are April, July, September and October. These months wasn't the intended month to conduct experiment Although, intended period to collect data was each two months, but due to fail in computer system while collecting data, new personal laptop bought to collect data. Table 26 summarize size of data before and after scrubbing.

Table 26: size of data before and after scrubbing for each month

Month	Original data size	Size after scrubbing
April	5787	3516
July	12136	7906
September	44534	27438
October	31390	18313

Prune methodology applies for these data starts from 20 to 50000 because these data is smaller than it is as one set. All of the methodology applies to the data described in section 3.7 also applies to these data. Number of clusters conducted on this experiment was k= 2, k= 10 and k=25. Because of the time limitation only these 3 different clusters tested this experiment and because when testing large number of k from k=2 to k=25, it tooks about one month to show the final results of evaluations and topic extracted. In this experiment, four internal evaluation applied to each data. The results from internal evaluation shown are not clear for April and July. However, September and October illustrated informative and clear results by using evaluation technique Davies Bouldin Index as it shown in table 27.

Table 27: Results of DBI in September and October

September	October	
Davies Bouldin	**Davies Bouldin**	**K**
5.516363547918373	4.02107926284097	2
7.42934466493562	6.773408170494578	10
5.6972029670779705	5.122710234849795	25

The results summarized that DBI in September and October agreed about that best k=2 when DBI is the lowest value. As an overall, When the data was one set, the best value for DBI was k= 2, which is similar to the results collected from September and October. Although the result shows that the optimal value of k is equal to 2, but this cluster will not give us more topics about these months. Therefore, k=10 will be selected to form more trending topics about دبي (Dubai). Moreover, k= 25 not chosen because some of the data set such as data of April and July when k=25, DBI= infinity and average within centroid without results. As a result, while increasing data more some of the clusters shows empty tweets, so k= 10 will be explained to show topics.

Table 28: Topics of k=10 for each month

K	Month	Size of data	Top terms	Topic extracted
Cluster 2	April	53	تستضيف ، دبي _تستضيف، بائعة ، الشاورما، بائعة_الشاورما، تستضيف_بائعة ، تستضيف_بائعة _الشاورما، دبي_تستضيف_بائعة	دبي تستضيف بائعة الشاورما

Cluster 5	April	74	العالمي _صور، روعة ، الأعلى، الأعلى_و التفرد،الأعلى_و التفرد_العالمي، تظهر ، تظهر_روعة،تظهر_روعة _دبي، دبي_ الأعلى	صورة تظهر روعة دبي من الأعلى
Cluster 8	July	681	النص، تستقبل، للاستثمار، دبي_للاستثمار، مليار، دبي_ تستقبل	استثمار دبي
Cluster 9	September	1334	شرطة_دبي، تطلق، دبي_تطلق، مراقب، ذكي،شرطة_دبي ، الشارع،الذكي، المراقب	شرطة دبي تطلق المراقب الذكي
	Translation			
Cluster 2	April	53	Host, Dubai _ Host, Seller, Shawarma, Alshawarma_seller, Hosts_seller, Hosts_ Alshawarma _seller, Dubai_hosts_seller	Dubai hosts shawarma seller
Cluster 5	April	74	Global_picture, splendor,top, top_and _singularity, top_and _singularity_global, shows, shows_ splendor , shows_ splendor _Dubai.	A picture showing the splendor of Dubai from the top
Cluster 8	July	681	half, Receiving, Investment, Dubai_Investment, billion, Dubai_ Receives	Dubai Investment
Cluster 9	September	1334	Dubai Police, Launching, Dubai_English, Monitor, Smart, Police_dpi, Street, Smart, Observer	Dubai Police Launches Intelligent Controller

Table 28 identifies four interesting topics that wasn't discover before when the data was one set. The new interesting topics discovered in 3 months except in October were all of the topics extracted from October are explained before. The new topics are دبي تستضيف بائعة الشاورما، صورة تظهر روعة دبي من الأعلى، استثمار دبي و شرطة دبي تطلق المراقب الذكي (Dubai hosts shawarma seller, a picture showing the splendor of Dubai from the top, Dubai Investment and Dubai Police Launches Intelligent Controller. Unclear topics and redundant topics wasn't covered in the table above.

Chapter 5

5. Conclusion and Future Work

Final decision and comments will be summarized in this chapter. In addition to that, future works will be addressed.

5.1 Conclusion

Twitter is a huge data bank that establishing researches on it, conducting variety of text mining techniques is a complicated task due to the challenges in analyzing Arabic language. In this paper, collecting tokens about دبي (Dubai) to elicit interesting topics by clustering tokens with k-mean algorithm and measuring distance between each tweet with cosine similarity. After this the quality of the collected cluster evaluated by internal evaluation and evaluation done by conducting several experiments to observe different results.

Although in this paper additional stopwords applied to remove stopwords that are not mentioned in Filter stopwords operator, some of these stopwords were not removed and considered one of the highest frequency tokens such as و (and). Therefore, traditional preprocessing didn’t perform their tasks to the fullest.

Moreover, K-mean algorithm successfully extracted tweets into different topics, however it was time consuming where for 93,850 tweets it took 1 month to extract topics for k from 2 to k=25. This model was successfully discovered trending topics about دبي (Dubai) where there were 13 various topics extracted from 4 experiments. Some of the main topics were مباراة الإمارات و السعودية ، تحدي دبي للياقة ، مبادرة أمل دبي للدعم و خدمة التاكسي الطائر في دبي (UAE and Saudi Arabia match, Dubai fitness challenge, Amal Initiative in Dubai and Flying taxi service in Dubai).

Results shows that while increasing number of k, more topic extracted , but when k reached to 30, where it set to be the optimal number of k because only one new topic extracted from k=30, incrementing number of k stopped here.

Furthermore, internal heuristic evaluation failed to agree about the optimal number of k where only Davies Bouldin Index shows that best k=2, but other internal heuristics evaluation failed to set best value of k. Empirical tests based on observing indicates that the best number of k=30.

There is an experiment conducted to check affective of Filter Tokens (by Region) operator in the quality of the clusters extracted by observing results from four internal evaluation techniques. Results from this experiment assure the effectiveness of using Filter Tokens (by

Region) operator in increasing quality of the clusters. Where DBI, average within centroids and Gini Coeffecient gave better results when their values are the lower than the results gained from without using Filter Tokens (by Region) operator. However, SSE didn't show clear results because it need sequential number of cluster to find Elbow and this experiment conducted only in 3 number of clusters.

Internal evaluation techniques also applies to the dataset classified by month of collecting data. The outcome shows results obtained on September and October by using DBI, where best k = 2, which is similar to the results of DBI when the data was 93,850 examples as one set. There were four new topics extracted from April, July and September except October there is no new topic found in it. One of the most interesting topics elicited on September was شرطة دبي تطلق المراقب الذكي (Dubai Police Launches Intelligent Controller).

5.2 Future Work

This study could be expanded to find trending topics through the other social networks by following similar steps such as Instagram.

Further research could be done by using one of detect outlier operators in RabidMiner as preprocessing step. Detect outlier operators are Detect Outlier (Distances), Detect Outlier (Dinsities), Detect Outlier (LOF) and Detect Outlier (COF). These operators intended to find outliers (tweets) from the cluster to filter topics that aren't similar to other. I was studying these operators to include them in the experiment, but due to the time limitations, I aim to do it in the future to find if this operator will arise quality of the cluster or not.

This experiment could be conducted in another filed such as digital forensics, so Arabic words "دبي" could be replaced by any other words that police want to search about. This could help police to view some crimes in digital forensics. For example, if there is any abuse to somebody by using unethical words, police could find the first person who initiate this word about other. The methodology will find all of the tweets which talk about this issue, then maybe by clustering the tweets intended user will be found, who first wrote the word. Following up with an issue of terrorists and attempting to sow strife among public, police could use same methodology to filter out these tweets. Besides, police could know people from their accounts, could use location feature in Twitter to locate them, monitor people's movements and could recognize first person who lunched the tweet. Accordingly, police can reprimand those people, setting fines or catching them.

References:

Abuaiadah, D., Dileep, R. & Mustafa, J. (2017). Clustering Arabic Tweets for Sentiment Analysis:.*IEEE/ACS 14th International Conference on Computer Systems and Applications*.

Adel, A., ElFakharany, E., & Badr, A. (2014). Clustering tweets using cellular genetic algorithm. *Journal of Computer Science*, vol. 10 (7), pp. 1269-1280.

Aggarwal, C.C. &Zhai, C. (eds). (2012*). Mining text data*. Boston, MA: Springer Science & Business Media

Ahmad, A. &Dey, L., 2007. A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, vol. 63 (2), pp.503-527.

AlRubaiee, H. & Alomar, K. (2017). Clustering Students Arabic tweets using different schemes. *International Journal of Advanced Computer Science and Applications*, vol. 8(4), pp.276-280.

Alsaedi, N. &Burnap, P. (2015). Arabic event detection in social media. *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, Cham, pp. 384-401.

Alsaedi, N., Burnap, P. & Rana, O.F. (2016). Sensing real-world events using Arabic twitter posts. *Tenth International AAAI Conference on Web and Social Media*, pp 515–518.

Atika. (2017). *Longest word in Arabic.*[Online]. [Accessed 7 Jan. 2018]. Available: https://preply.com/en/question/longest-word-in-arabic.

Ayedh, A., Tan, G., Alwesabi, K. & Rajeh, H. (2016). The effect of preprocessing on Arabic document categorization. *Algorithms*, vol. 9(2), p.27.

Banerjee, A., Merugu, S., Dhillon, I.S. & Ghosh, J. (2005). Clustering with Bregman divergences. *Journal of machine learning research*, vol. 6 , pp.1705-1749.

Baralis, E., Cerquitelli, T., Chiusano, S., Grimaudo, L. & Xiao, X. (2013). Analysis of twitter data using a multiple-level clustering strategy. *International Conference on Model and Data Engineering.* Springer. Berlin, Heidelberg. September, pp. 13-24

Becker, H., (2011). *Identification and characterization of events in social media*. Columbia University.

Bekkali, M. &Lachkar, A. (2017). 'Web Search Engine-Based Representation for Arabic Tweets Categorization', in From Social Data Mining and Analysis to Prediction and Community Detection (pp. 79-101). Springer, Cham.

Bellot, P. & El-Bèze, M. (1999). Query length, number of classes and routes through clusters: Experiments with a clustering method for information retrieval: *International computer science conference*. Springer, Berlin, Heidelberg. December, pp. 196-205.

Bellot, P. &El-Bèze, M. (2000). Clustering by means of Unsupervised Decision Trees or Hierarchical and K-means-like Algorithm. *Content-Based Multimedia Information Access-. le centre de hautes etudes internationales d'informatique documentaire. April*, vol. 1, pp. 344-363.

Brownlee .J. (2016). *Supervised and Unsupervised Machine Learning Algorithms* [Online]. [Accessed 7 Jan. 2018]. Available at: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms.

Brownlee .J. (2017). *A gentle introduction to the bag-of-words model* [Online]. [Accessed 3 Feb. 2018]. Available at: https://machinelearningmastery.com/gentle-introduction-bag-words-model/.

Chepkemoi, J. (2017). *Countries Where Arabic is an official language* [online] [Accessed 16 Feb. 2018]. Available at: https://www.worldatlas.com/articles/countries-where-arabic-is-an-official-language.html.

Davies, D.L. & Bouldin, D.W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2), pp.224-227.

De Roeck, A.N. & Al-Fares, W. (2000). October. A morphologically sensitive clustering algorithm for identifying Arabic roots. *The 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 199-206.

De Villiers, F. (2013). *Constructing topic-based Twitter lists* . Ph.D. Thesis. Stellenbosch University.

Demiriz, A., Bennett, K.P. & Embrechts, M.J. (1999). Semi-supervised clustering using genetic algorithms. *Artificial neural networks in engineering (ANNIE-99)*, pp.809-814.

Deshpande .B. (2012). *3 ways to use text mining with RapidMiner to juice up your job search*. [Online]. [Accessed 7 Jan. 2018]. Available at: http://www.simafore.com/blog/bid/111839/3-ways-to-use-text-mining-with-RapidMiner-to-juice-up-your-job-search.] [Accessed 5 Jan. 2018].

Duwairi, R.M., 2005, June. A Distance-based Classifier for Arabic Text Categorization. *DMIN*, pp. 187-192.

El-Fishawy, N., Hamouda, A., Attiya, G.M. & Atef, M. (2014). Arabic summarization in twitter social network. *Ain Shams Engineering Journal*, vol. 5(2), pp.411-420.

Ertek, G., Tapucu, D., and Arın, I., 2013. Text Mining with RapidMiner. In: Markus Hofmann, Ralf Klinkenberg (Eds.) RapidMiner: Data Mining Use Cases and Business Analytics Applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC.

Ghwanmeh, S.H. (2007). Applying Clustering of hierarchical K-means-like Algorithm on Arabic Language. *International Journal of Information Technology IJIT*, vol. 3(3), pp.168-172.

Gove, R. (2017). *Using the elbow method to determine the optimal number of clusters for k-means clustering* [online] [Accessed 29 Jan. 2018]. Available at: https://bl.ocks.org/rpgove/0060ff3b656618e9136b

Hamadeh, M.W. (2015). *Using Text Mining and Clustering Techniques on Tweets to Discover Trending Topics in Dubai*. Ph.D. Thesis. The British University in Dubai (BUiD).

Hammad, M. & El-Beltagy, S.R. (2017). Towards Efficient Online Topic Detection through Automated Bursty Feature Detection from Arabic Twitter Streams. *Proceeding Computer Science*, vol. 117, pp.248-255.

Han, J. (n.d.). Clustering Evaluation Measuring Clustering Quality , video [online]. [Accessed 7 Jan. 2018]. Available at: https://www.coursera.org/learn/cluster-analysis/lecture/RJJfM/6-2-clustering-evaluation-measuring-clustering-quality.

Harmain, H.M., El Khatib, H. & Lakas, A. (2004). Arabic text mining. *IADIS International Conference Applied Computing*, pp. 23-27.

Hearst, M. (2003). *What is text mining*. SIMS, UC Berkeley [online]. [Accessed 20 January 2018]. Available at: http://people.ischool.berkeley.edu/~hearst/text-mining.html

Huang, A. (2008). Similarity measures for text document clustering, *the sixth New Zealand computer science research student conference (NZCSRSC2008).* Christchurch, New Zealand. April, pp. 49-56.

Ignatow, G. &Mihalcea, R., 2016. Text Mining: *A Guidebook for the Social Sciences*. Sage Publications.

Internet live stats. (2018). *Twitter usage statistics*. [online]. [Accessed 8 Feb. 2018]. Available at: http://www.internetlivestats.com/twitter-statistics/.

Kim, S., Jeon, S., Kim, J., Park, Y.H. &Yu, H. (2012). Finding core topics: Topic extraction with clustering on tweet*: 2012 Second International Conference on Cloud and Green Computing (CGC).* IEEE. November, pp. 777-782.

Kireyev, K., Palen, L. &Anderson, K., 2009, December. Applications of topics models to analysis of disaster-related twitter data. *NIPS Workshop on Applications for Topic Models: Text and Beyond* , vol. 1. Canada: Whistler.

Ko, Y. & Seo, J. (2000). Automatic text categorization by unsupervised learning. *The 18th conference on Computational linguistics*. Association for Computational Linguistics. 31 July - 4 August, vol. 1, pp. 453-459.

Kwak, H., Lee, C., Park, H. &Moon, S., (2010). What is Twitter, a social network or a news media? *19th international conference on World wide web*. ACM. April, pp. 591-600.

Maulik, U. & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24 (12), pp.1650-1654.

Mosley Jr, R.C. (2012). Social media analytics: data mining applied to insurance Twitter posts. *Casualty Actuarial Society E-Forum*, vol. 2 (winter), pp. 1-36.

Newton .C. (2017). *Twitter just doubled the character limit for tweets to 280.* [Online]. [Accessed 7 Jan. 2018]. Available at: https://www.theverge.com/2017/9/26/16363912/twitter-character-limit-increase-280-test

Nishida, K. (2016). *Demystifying text analytics part 3—finding similar documents with cosine similarity in R*. learn data science [online]. [Accessed 11 Feb. 2018]. Available at:

https://blog.exploratory.io/demystifying-text-analytics-finding-similar-documents-with-cosine-similarity-e7b9e5b8e515.

Nuaim, w. (2017). *The accounts of the "contact" government's department need to be activated more. albayan* [online]. [Accessed 8 Feb. 2018]. Available at: http://www.albayan.ae/across-the-uae/news-and-reports/2017-11-19-1.3104916.

Rafea, A. &Mostafa, N.A. (2013).Topic extraction in social media. *International Conference on Collaboration Technologies and Systems (CTS)*. May. IEEE, pp. 94-98.

Rao, S.G. & Govardhan, A. (2015). Performance validation of the modified k-means clustering algorithm clusters data. *International Journal of Scientific & Engineering Research*, vol. 6 (10), pp.726-730.

RapidMiner. (2017). *Cluster Distance Performance* [online]. [Accessed 29 Jan. 2018]. Available at: https://docs.rapidminer.com/latest/studio/operators/validation/performance/segmentation/cluster_distance_performance.html.

Saad, M.K. &Ashour, W. (2010). Arabic text classification using decision trees. *Proceedings of the 12th international workshop on computer science and information technologies CSIT* . November. vol. 2, pp. 75-79.

Sapul, M.S.C., Aung, T.H. & Jiamthapthaksin, R. (2017). Trending topic discovery of Twitter Tweets using clustering and topic modeling algorithms. *Conference on Computer Science and Software Engineering (JCSSE), 2017 14th International Joint Conference*. July.IEEE, pp. 1-6.

Sawaf, H., Zaplo, J. &Ney, H. (2001). Statistical classification methods for Arabic news articles. *Arabic Natural Language Processing*, workshop on the ACL2001, Toulouse, France.

Schulte Im Walde, S.S. (2006). Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, vol. 32 (2), pp.159-194.

Smith, C. (2014). *Social big data: the user data collected by each of the world's largest social networks and what it means. Blog, p. 2* [online]. [Accessed 19 February 2018]. Available at: http://www.businessinsider.com/social-big-data-the-type-of-data-collected-by-social-networks-2-2014-1.

Syeeda, T. (2017). *Rapid advances in smart mobility. Dubai set to be smartest city : Dubai plan 2021* [online]. [Accessed 8 Feb. 2018]. Available at: http://www.itssa.org/wp-content/uploads/2017/05/DUBAI-Smart-City-.pdf

Twitter. (n.d.). Report spam on Twitter [online]. [Accessed 11 Dec. 2017]. Available at: https://help.twitter.com/en/safety-and-security/report-spam.

Verma, T., Renu, R. & Gaur, D., 2014. Tokenization and filtering process in RapidMiner. *International Journal of Applied Information Systems*, vol. 7(2), pp.16-18.

Wei, H., Sankaranarayanan, J. & Samet, H. (2017). *Finding and Tracking Local Twitter Users for News Detection*

Τζιωρτζής, Σ.Μ. (2013). Sentiment analysis by emoticons and unsupervised comment summarization in Greek e-Government data.