Table of Contents

List of Figures

List of Tables

Chapter #1

1. Introduction

This chapter of the report addresses the importance of this study, what are the motivations, how important is this study for the targeted reader. Research questions and the objective of this work are described in this chapter.

1.1 Overview

Dubai is moving towards becoming a smart city, Dubai's vision is to be one of the smartest cities in the world by 2017. The objective of the project launched in 2014 is to make more than 1,000 government services to go smart within two years (Kuzela, 2015). Dubai has already been crowned the smartest city in the region of Middle East and North Africa (Emirates247, 2015).

Being a smart city means that the audience will be able to complete their government services and be more interactive online, social media is the best interactive way nowadays.

A smart government should be able to track the interest and trends of its people. Twitter can provide an excellent medium for discovering such trends and interests. People tweet about Dubai events, expressing their thoughts, sentiments, posting news, advertisements, and much more.

Twitter is not only considered as a social networking tool, it is also an emerging source for information and knowledge sharing. Due to the simple use and informal writing, Twitter has become a very popular media for communication and information sharing. As of May 2015, Twitter is ranked as the 8th top website in the world based on Alexa ranking.

Messages shared by users on Twitter are called tweets; those tweets can contain any form of text, URL, photo, or any combination of them. Because Tweets are considered as a form of micro-blogs, they are limited to 140 characters only, which implies that a message should be short and focused on a topic.

Looking into one specific topic in Dubai from the huge amount of tweets is becoming very difficult task, this study will build a clustering model to group those tweets into a number of clusters where each cluster is contains tweets about a specific topic in Dubai.

Unsupervised learning through clustering techniques is the best approach for achieving the objectives of this study; classifying techniques require supervised intervention through having a prior knowledge about the tweets and label them into classes to predict the testing data. This approach may work well for classifying tweets into topics, but will miss the detection of new and trending events in Dubai, and won't be able to discover new trending topics and events.

The proposed model will help decision makers in Dubai to focus more on their audience on Twitter. For example, a cluster which talks about special event in Dubai will help the events decision makers to know what people are talking about on this event, what are their sentiments, and the volume of tweets generated about this event. Moreover, this model will help decision makers in reaching their audience; they can know the user accounts who are interested in this event so they can reach them through targeted tweets for future similar events.

This work is the first scientific research which uses text mining tools for clustering tweets of Dubai domain, and group those tweets into different clusters based on their similarity measure. It is also the first to examine in such a domain of Twitter the heuristics cluster evaluation techniques for evaluating quality of clusters and helping in determining the right number of clusters to use in the experiments.

1.2 Objectives

Purpose of this thesis is to describe through empirical study how text mining and clustering techniques can be applied on tweets related to Dubai city domain to group these tweets into clusters based on their similarities. This technique will help in identifying key themes, trending topics, and events detection in a particular timeframe on dataset collected on Dubai tweets.

Furthermore, through clustering tweets from different timeframes, the model can discover what are the main top topics discussed on Twitter about Dubai across all timings.

This work will also examine the quantitative internal clustering evaluation techniques such as, SSE and DBI. Empirical tests will answer whether these techniques are enough to rely on for studying such a domain.

1.3 Research questions

Through empirical tests, this research will answer the following questions:

- What is the effect of conducting internal clustering evaluation tests for studying such a domain on Twitter? Are those techniques helpful for evaluating the quality of the clustering model? Would they be helpful in determining the right number of clusters to input before building the model?
- Can we use clustering to group Dubai tweets into clusters of theme-based tweets?
- What are the top trending topics people are tweeting on Twitter about Dubai? Are there any constant topics people are talking about in different timeframes?
- Can this model detect and discover major events happening in Dubai?

1.4 Structure of the report

The rest of the report is organized as the following; chapter 2 provides a background and general information about text mining and clustering, it also surveys literature about related work. Chapter 3 presents the methodology of the research, how the data were collected and processed, it also provides information about the experimental setup. Chapter 4 discusses the experimental analysis and results, how the experiments were conducted and what are the outcomes. Chapter 5 draws closing remarks in conclusion, it also addresses future work which can be built on this research.

Chapter #2

2. Background

This chapter of the report delivers background information about text mining and clustering technique used in this study. It also reviews literature which have used text mining studies on Twitter.

2.1 Twitter

Twitter is an online social network site founded in 2006 which allows its users to use the concept of micro-blogging. Twitter users can write and read micro-blogs messages easily, those messages are called tweets, tweets are very similar to SMS concept. Because it's a micro-blogging platform, each tweet is only limited to 140 characters.

Each user has a twitter page which consists of user's profile and tweets, in order to read others tweets from the user's page, the targeted user should be followed by the follower user.

On Twitter, people talk about specific topics, hashtags are used for this purpose. Hashtag is a phrase followed by "#" which gives a hint about the context or topic of the tweet; hashtags are usually descriptive of the tweet's context (Bruns and Burgess, 2011)

Twitter is unlike other social networks, it is not focused on friendship and relationships, it focuses more on the interest of the user. A follower will follow another user because he/she is interested in tweets published by the followed user even if they don't know each other (Noordhuis, Heijkoop and Lazovik (2010).

2.2 Text mining

Text mining is a form of data mining that deals with text resources. It is the process of discovering by learning machines new or previously unknown hidden information from text resources. The main difference between data and text mining is that text mining finds out the useful and interesting patterns from natural language unstructured text unlike data mining with finds patterns from structured databases (Hearst, 2003).

Text mining is also different from the information retrieval, the latter focuses more on easing and facilitating the access to the text (Jones, 1997), while text mining focuses on analysing and discovering hidden patterns in data.

The key point in text mining is to combine the human linguistics abilities with high speed and accuracy of learning machines to deal with very huge amount of text (Fan et al., 2006).

A lot of technologies and related topics have been emerged from the science of text mining; Aggarwal and Zhai (2012) mentioned those related topics as:

- Information extraction: extract structured information from text inputs
- Text Summarization: extract important sentences and terms from lengthy text to provide users with key information about the text.
- Unsupervised learning from text: learning from text without any prior knowledge about the context of the testing text, main two unsupervised methods are clustering and topic modeling.
- Supervised learning from text: learn a classifier about a training text, and then apply the algorithm on test text. Examples are automatic ranking and categorization of text
- Dimensionality reduction: text data is sparse and highly dimensional, a lot of algorithms have been used to deal with reducing the high dimensions of data such as: Latent Semantic Indexing and Latent Dirichlet allocation.

- Transfer learning with text: the process of transferring knowledge learned from one text domain to another. For example, transfer the text knowledge from one language to another.
- Probabilistic model for text: the use of probabilities to model uncertain text data.
- Text stream mining: mining continuous data, such as news stream or RSS feeds.
- Opinion mining: semantic analysis about opinions of people in a topic or product.
- Multimedia mining: such as mining the caption of photos or titles of videos
- Mining social media: twitter and Facebook are most two popular social media sources. Main problems of mining social media is to deal with short text data, informal language, and unstructured phrases

2.3 Clustering documents

Clustering documents is very popular and important field in text mining, it discovers patterns and hidden information from text documents and performs algorithmic methods to group these documents into groups based on their object similarities.

The main goal of documents clustering is to group unlabelled text documents into groups which may share the same topic of theme. Clustering documents can be used in search engines, documents retrieval, documents organization, categorization, and summarization (AlSumait and Domeniconi, 2008).

Main applications of unsupervised document clustering are information retrieval, automatic topic extraction, and document organization (Zhong and Ghosh, 2003).

2.4 K-means clustering algorithm

K-means algorithm is one of the most used unsupervised learning algorithms for clustering tasks.

K-means uses a pre-defined number of clusters k to partition the introduced data to this number of clusters. It works through the following number of processes (MacKay, 2003):

1- Initialize a number of k centroids
2- Assign each point randomly to a cluster to have all points assigned to cluster
3- Calculate each point to its closest centroid through a distance measurement and assign those points to their closest clusters
4- Recalculate the mean of the points in each cluster to assign new centroids.
5- Repeat the calculation and assignment of points until reaching maximum number of iterations or no change occurs in clusters assignment.

K-means is widely used in data mining tasks, it is by far the most popular clustering technique used in data mining. The main advantages of k-means are the ease of implementation, fast computational cost compared to other techniques, and independence of data ordering (Berkhin, 2006).

It's not only about the low computational cost, Shrestha, Jacquin, and Daille (2012) proved in their study that k-means algorithm performs better clustering results than hierarchal technique which also takes much longer time to generate its output.

2.5 Cosine similarity

Cosine similarity is a popular similarity measurement technique in text mining, it measures the distance between objects in dimensions. Cosine similarity measures the angle between two vectors, a value of 1 which is the cosine of 0 means that the two vectors have the same direction, and a value of 0 which is the cosine of 90 means that the two vectors are in the opposed directions. Cosine similarity measure can be computed as:

$$Sim_c(\overrightarrow{t_a}, \overrightarrow{t_b}) = (\frac{\overrightarrow{t_c} . \overrightarrow{t_b}}{|\overrightarrow{t_a}| * |\overrightarrow{t_b}|}$$

Where $\overrightarrow{t_c}$ *and* $\overrightarrow{t_b}$ are m-dimensional vectors over term set $T= \{t_1,...., t_m\}$

Cosine similarity has been used widely in text mining, Subhashini and Kumar (2010) proved that cosine similarity measure will play an important role in clustering documents, this technique proved to produce best clustering results than the other techniques. Strehl, Ghosh, and Mooney (2000) also proved that cosine similarity and Jaccard correlation are very close and are much better than Euclidean distance when dealing with text. Huang (2008) had a similar study and agreed with the results of the previous study.

2.6 Related Work

Mining Twitter is a challenging task because of the nature of the tweets, each tweets is limited to 140 characters only which means a tweet may consist of very few words. Researchers have done some studies and experiments on twitter data with different findings and outcomes.

Moh and Bhagvat (2012) questioned the effect of the stop words in generating better clustering result using k-means algorithm and cosine similarity to cluster technological tweets. They found that using a custom dictionary of stop words to be removed by text preprocessing will give better results on tf-idf weighting scheme and thus on clustering model than relying on the default stop words dictionary derived from Wikipedia. They stated that some stop words may be important for the tweet topic.

People on Twitter do not always use formal language, and because the tweets are character-limited, people use a lot of abbreviations and sometimes combine two or three words in one. Perez et al. (2011) believed that those abbreviations may be important to the context of the tweets; they used tweets dataset collected about high reputational companies in the world, and enriched those tweets with abbreviation dictionary before

applying k-means clustering. Because they had already previously labelled tweets, they found through f-measure test that enriching the tweets with abbreviation dictionary and adding more formality to its language will improve the clustering model.

In another study, Tripathy et al. (2014) presented a new clustering algorithm called WIKI-kmeans, which is based on k-means algorithm. They mapped tweets to a set of Wikipedia web page topics, and then measured the distance between the tweets and Wikipedia pages through cosine similarity measure. They found that this new algorithm produces better results than the traditional model.

In a similar study, Chen, Shipper and Khan (2010) extended their tweets feature vector with search results for similar topics from Wikipedia. They used Naïve Bayes classifier to learn the algorithm about the training predefined labelled tweets, they found that Naïve Bayes returned poor classifying results; they improved the results by extending the feature vector with search results from Wikipedia.

As stated by Rosa et al. (2011), unsupervised clustering techniques such as k-means, cluster tweets based on their word similarities more than their topical relation. They found in their study on tweets collected about predefined trending topics that using supervised classifying algorithms such as SVM and K-NN performs better than unsupervised clustering algorithms such as k-means. Although the number of clusters to use in their study is already known because of choosing tweets from a pool of predefined topics, the algorithm still did not outperform classifier algorithms.

Tweets about World Cup football event were collected by Godfrey et al. (2014) in order to perform some clustering techniques on these tweets and discover what the most trending topics people are talking about before the launch of World Cup event. They compared the clustering results of k-means algorithm and Non-Negative Matrix Factorization (NMF) method with using cosine similarity measure, which creates two matrices, one for term-topic matrix, and the other for topic-document matrix. They found that both techniques perform very similar clustering results, but using NMF has an advantage of easier interpretation of clusters and faster running time.

Hashtags in Twitter can give an idea about the context of a tweet. Tsur, Littmanand and Rappoport (2013) used hashtags in tweets to label their tweets dataset into a number of 9 classes based in major interest such as music, sport, politic, etc. They used the text representation model of bag of words, and then they extended that model by creating a co-occurrence model of hashtags, they believed that hashtags will occur more if they are related to the cluster they are labelled in. Then, they performed Web-Scale Fast k-means algorithm which is a modified k-means algorithm proposed by Sculley (2010). They first used the value of k=9 based on the 9 topics these tweets came from, but they performed other experiments with different value of k as they believe that tweets can belong to more than one topic or cluster. Using validation measurements such as F-measure and VI-measure, they found that their proposed method performs better results than traditional clustering.

Another study on Twitter hashtags was done by Antenucci et al. (2011). In their experiment, they aimed to classify the tweets' topics based on clustering the hashtags appeared in the tweets dataset. At first, they were interested in collecting tweets which have hashtag terms in it, then, they removed the hashtags from the tweets and collected the most 2000 used hashtags in their dataset. After that, they performed clustering tasks on these hashtags based on their co-occurrence relation in the tweets as they believe that if two or more hashtags occurred in one tweet means that they are similar. They represented the tags as vectors where each tag is tokenized with its letters and used cosine similarity to measure the distance between hashtags. Finally, they assigned labels to the tweets based on the clustered hashtags and performed clustering techniques such as SVM and PCA after splitting the data into training and testing set.

A similar study conducted by Muntean, Morar and Moldovan, D. (2012) aimed to cluster the hashtags of tweets using k-means algorithm. Qualitative evaluation was done on the results because no previous class labels were annotated. They found through the clustering model that hashtags clustered in one group can semantically be representative of the top terms of tweets in that cluster.

Baralis et al. (2013) used density-based DBSCAN clustering algorithm in their study on twitter data. Their research aimed to find similar topics among tweets talking about special events. They collected tweets over a period of one month related to sports and music, and then extracted the top event from each topic based on the most frequent hash tags in each topic. After that, they performed a multiple-level DBSCAN iteration clustering algorithm; they found that in each iteration the algorithm will generate more specific topic-related clusters. For example, advanced iteration level will generate clusters about sentiments and how people are reacting on these events.

A new clustering method for tweets was proposed by Kim et al. (2012) called core-topic-based clustering (CTC). The focused on the re-tweets of famous TV shows on Twitter, they believed that retweets are considered important and may hold important contexts for information retrieval from Twitter. Retweets are preferred tweets from a number of real account, from this perspective, the researchers used retweets also to avoid spam as they believed that spam tweets will not be retweeted by real users. Their CTC method extracted the top 10 topics from the collected dataset, and then clustered the tweets based on the extracted topics from the proposed algorithm. They believe that their proposed method is different from the k-means in focusing on the core topics from the tweets, while the k-means algorithm focuses on the terms occurred in the tweets.

Another clustering method was performed on Twitter data was performed by Adel, ElFakharany and Badr (2014), they used cellular genetic algorithm and generational genetic algorithm on two datasets collected which consisted of 5000 and 1000 tweets from 8 predefined topics. They ran the two algorithms on the collected data for 50 and 40 runs respectively. They found that both algorithms generate similar results based on the fitness quality test. A comparison of these techniques and traditional clustering techniques such as k-means or density-based DBSCAN was not discussed.

Twitter lists is a feature supported by Twitter which allows users to create lists of people they follow, a list created by a user for multiple accounts is mostly used for accounts which share a topic or interest. From this perspective, de Villiers, Hoffmann and Kroon (2012) collected popular topic-related lists from Twitter in order to cluster users' accounts related

to these lists and build clusters of topic-related lists. They found in their results that cosine similarity measure along with extended Jaccard produce better similarity results, also they concluded that the affinity propagation with Latent Dirichlet allocation (DLA) topic modeling, gave better results than using k-means with tf-idf vector space model (VSM).

On the other hand, Rangrej, Kulkarni and Tendulkar (2011) found that using affinity propagation algorithm with cosines similarity performs better results than using Jaccard similarity on Twitter data with tf-idf document representation, while Jaccard similarity gave better results than cosine similarity when performed on same dataset.

Mosley (2012) studied the tweets related to the insurance topic in order to inspect what are people tweeting about the insurance. Through Ward's Minimum-Variance method, the author clustered the collected 68,370 tweets dataset into 47 clusters. Findings were interesting as clusters contained tweets with related specific topics such as sentiments, claims, and complaints.

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 Data collection

Because this research is about mining tweets related to the term “Dubai”, there is no available corpus for such research. Though, collecting tweets was done with the help of the online tool zapier.com. An account was created to sync tweets related to “Dubai” from twitter to a predefined google account where tweets where saved in a newly created google spreadsheet.

Two datasets were collected, the first dataset was collected in Feb and Mar 2015, while the other dataset was collected in May 2015. The first dataset consists of 124,539 tweets; the other dataset consists of 12,051 tweets.

Both of the datasets are needed in this study for different purposes, for the ease of referring to the datasets, the first dataset collected in Feb and Mar is called Dataset A, the other dataset collected in May is called Dataset B.

Each row inside the spreadsheets of both datasets consists of three columns as described in Table 1.

Table 1, columns of dataset

Column	Role	Description
Username	Regular	The owner of the tweet, the user account which posted the tweet
Date	Regular	The date of posting the tweet
Text	Text	The body text of the tweet

For the purpose of this research, only the column of text was used and the other two columns were not introduced in the experiments.

3.2 Data Scrub

Scrubbing or cleansing the data is the process of improving the quality of the data by removing noise, errors or any aspect in the data which can affect the results of the work (Rahm & Do, 2000).

All of the work of cleaning the data in this phase was done manually; the total number of tweets left after scrubbing Dataset A is 28,787 tweets, while Dataset B ended with 5,906 tweets.

The process of scrubbing both tweets datasets can be described in the following tasks:

3.2.1 Removing non-English tweets

This study is focusing on mining English tweets related to Dubai. While collecting the data, a lot of non-English languages where using the hashtag (#Dubai) in their tweets. Removing non-English tweets was done using two dictionaries; the first one consists of the most used letters in non-Latin alphabet, such as Arabic, Russian, and Urdu.

The other dictionary consisted of the most used words in non-English languages that use the Latin alphabet, such as French and Spanish. The removal of tweets was done by searching through them using these two dictionaries.

3.2.2 Removing spam

Because the hashtag (#Dubai) is so famous, a lot of accounts where using this hashtag to promote for their spam tweets. A very huge portion of the tweets were spam which are totally unrelated to the search topic (Dubai), because most of the spam accounts retweet their tweets many times in order to flood the twitter feeds with spam.

Removing spam was done by having a list of spam accounts and delete all tweets generated from these accounts. An example of spam tweets are shown in Table 2.

Table 2, examples of spam tweets

Examples of spam tweets
- Retweet to get more followers, follow me, everyone else that RTs, this for 50+ followers fast #free #tfbjp #dubai - #SEO 1000 Twitter. Retweets, Favorites for $5: Buy 1,000 twitter followers for $5. Safe and delivered fast #dubai - Leonard Nimoy's incredible life in pictures http://t.co/BQDVpCMNVZ #Dubai #UAE - 15 ways you're probably misusing social media http://t.co/9m7cvztLZq #Dubai #UAE - Photo series captures manly men posing with their adorable cats http://t.co/LpH7RIGTZ0 #Dubai #UAE - #Retweet only if you want #followers in #Dubai #followtrick, retweet for fast followers

3.2.3 Removing URLs

A lot of tweets contained URLs for webpages especially if these tweets were shared from a webpage or talking about a topic which is detailed in a webpage. Because tweets are limited with 140 characters only, it is very common to post a title of a topic or a headline and post with it the full story via a URL.

Also, because we are extracting text from twitter, a lot of tweets had photos attached with it which are not extracted, so a URL of these photos as available in the tweets.

URLs are meaningless in our research; they can distract the learning machine, so they are removed from the whole corpus by removing all strings which starts with (http).

3.3 Text representation – Vector Space Model (VSM)

In order to prepare the text data for clustering analysis, the text strings in documents should be presented in a way that the learning machine can understand and process it. In this research, the text documents (tweets) will be presented as vectors. One of the most used models to represent the text is the Bag of Words (BOW). The text is broken down into tokens (words), where each token is considered as feature vector for the text documents.

Each document or tweet in our case is presented by a bag of its own words. Each word or term in feature space is weighted by a weighting scheme.

Most popular schemes are term frequency (TF) and term frequency-inverse document frequency (TF-IDF).

3.3.1 Text pre-processing

In order to create the Bag of Words model, a number of pre-processing steps should be completed on the document. Following are the pre-processing steps used on both datasets, they are ordered in sequence as introduced into the RapidMiner:

- ***Tokenization***: it is the process of breaking down each document into number of tokens (terms).
- ***Transform cases***: transform all characters into lower case in order not to have a duplicate of terms having lower and higher cases.
- ***Filter stopwords***: remove stopwords from all documents, because they will affect the tf-idf score as stopwords occur too many times in documents.
- ***Filter tokens by content***: Remove all tokens which contain the string (http), to remove all URLs from the tweets.

- ***Filter tokens by length***: Remove tokens which consist of 2 characters and lower, and remove tokens which have 20 characters and more.
- ***Stem (Porter):*** Stem is the process of transforming the words into their root or as per the Porter stemmer (Porter, 1980).
- ***Create n-gram:*** a value of n=3 was used for creating 2-grams and 3-grams of strings.
- ***Pruning:*** In order to reduce the huge number of features and eliminate unnecessary term, a pruning method was implemented. Any term which has less than 20 appearances in the whole dataset was ignored. This will help in removing un necessary terms such as mentions, wrong spelling, and other rare non-English terms.

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

After doing the previous text pre-processing steps, each term will have a score based on its TF-IDF calculated value.

Term frequency will look into each term in a document and measure how frequent it appears in this document by dividing the number of appearances of each term on the total number of terms in this document.

Because tweets are limited to 140 characters only, which makes each tweet consists of few number of words, using the term frequency will almost results in giving each term in a tweet the same score because mostly each term appears only once.

Thus, inverse document frequency will be the determining factor for scoring a weight for each term. Document frequency is the number of documents which contain a specific term. The Inverse document frequency indicates that if a term is occurring in too many documents it should be have a lower score, it is calculated as:

$$Idf = log\ (N/df)$$

N = the number of documents (tweets) in the dataset.

Finally, TF-IDF is calculated as: $tf\text{-}idf = tf * idf$

This means that TF-IDF will give a high weight for terms which appear frequently in a small number of documents, and a low score for terms which appear in many documents (Manning, Raghavan and Schütze, 2008).

3.3.3 Output

The output of data scrubbing and representation of text of Dataset A is 28,787 tweets; and the output of Dataset B is 5,907 tweets. They were both introduced into the text mining process to represent the tweets in a vector model. After going through all of the processes described above, the number of tokens represented in features for Dataset A were 2734 tokens. Unsurprisingly, the term (Dubai) has the biggest number of occurrences across all tweets with 25,723 term appearance. While the number of tokens generated from Dataset B was 442 tokens.

Figure 1 shows a word cloud of the most frequent terms from the output of text processing of the tweets Dataset A.

3.4 Clustering tweets

The clustering algorithm which will be used in this study is k-means. A description about this algorithm is available in section 2.4.

As for the similarity measure, cosine similarity will be used for measuring the distance of the objects (tweets). A brief description on similarity measure is explained in section 2.5.

Figure 1, output of text pre-processing for dataset A

3.5 Cluster evaluation

Because this is a completely unsupervised clustering task as we don't have any class labels for the tweets, the quality of the clustering task will be evaluated firstly through internal heuristic cluster evaluation techniques, and secondly through observation and analysing the output clusters.

3.5.1 Internal clustering evaluation techniques

To perform the internal clustering evaluation, the following indexes will be used to evaluate the quality of the clusters in this research in order to assist in choosing the right number of k clusters in the experiments:

- ***Sum of Squares Error (SSE)*:** one of the most popular techniques to measure the cohesion of the cluster. Error is the distance between a point and the points in the nearest clusters. The errors then are squared and summed up as the following formula:

$$\sum_{i=1}^{k} \sum_{x \epsilon c_i} (x - m_i)^2$$

Where *x* is a point in cluster c_i, and *m* is the representative point in the cluster c_i. The value of SSE decreases as the number of k increases, until it reaches a point where the value decreases very low with the increase of k. At the point where the value decreases slightly is considered a good value of k clusters.

- ***Davies-Bouldin index:*** Davies-Bouldin index is another important measurement for evaluating the quality of clusters. Its main goal is to reduce the variance inside the cluster, and to maximize the distance between different clusters (Davies and Bouldin, 1979).
The formula of DBI can be represented as:

$$\text{DBI} = \frac{1}{N} \sum_{j=1}^{M} \max_{j \neq k} R_{j,k}$$

$$R_{j,k=} \frac{MSE_{j+ MSE_k}}{M_{j,k}}$$

Where *N*= the number of clusters, *j,k* are clusters, *MSE* is the mean square error for the cluster, *M* is the distance between centroids of clusters *j, k.*

- ***Gini coefficient:*** Gini coefficient evaluates the distribution of the points in the cluster; it is a measure of statistical dispersion. The value of Gini coefficient scales

between 0 to 1. With values tending to 0 indicate equity of distribution among a cluster, and values tending to 1 indicate inequity distribution of points inside a cluster.

- ***Average within centroid distance:*** This method computes the average distance between the centroid of the cluster and all data points in the same cluster. Less value of average within centroid distance indicates a better clustering model. The Elbow method can be used in this technique where the point to choose the number of clusters is where adding more clusters will have no big better modelling of the cluster.

3.5.2 Cluster Evaluation by model analysis

First, we'll look at the results coming from the internal clustering evaluation metrics described in the previous section. If the results of the clusters are not convincing, we'll not rely on numbers as these techniques are heuristics and they don't give an absolute answer to the analyst based on the purpose of the research. It's always been said, that the observation of the analyst expert on the unsupervised clustering results is the dominant factor in determining the quality of the clusters not by relying only on heuristic techniques.

3.6 RapidMiner implementation

The data mining tool which is used for experiments in this work is RapidMiner v5. Previously, this system was called (Yale) and it is built on JAVA environment. It has an easy to use GUI and a set of comprehensive operators for most of the data mining tasks. The friendly process view of the application makes any real word data mining problem easy to build and interact with (Hofmann and Klinkenberg, 2013).

In order to use RapidMiner for text mining problems, a text mining plug-in should be downloaded and installed first, which builds the required operator and environmental setup for text mining problems.

The following steps will describe through steps how the setup for the experiment was built using the RapidMiner tool.

1- The data was imported from an Excel spreadsheet to RapidMiner and stored in the local repository.
2- Retrieve operator was used to call the dataset from the repository.
3- The data set was introduced into "Process documents from data" operator as shown in Figure 2, which creates the term vector from the dataset.

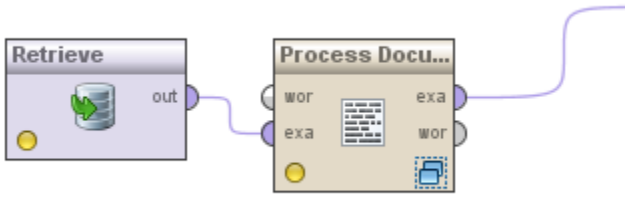

Figure 2, introduction of RapidMiner model

4- "Process documents from data" is a min process which includes a sub operators inside it, the operators are the preprocessing steps used to create the term vector. The steps were described before in section 3.3.1 and are implemented in Rapidminer as shown in Figure 3.

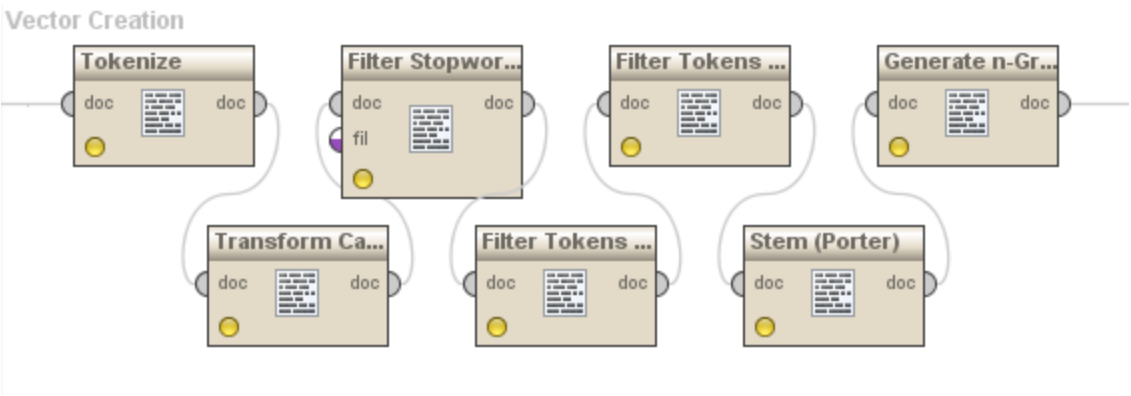

Figure 3, text mining model in Rapidminer

5- The output of the "Process documents from data" operator was introduced into a loop operator as shown in Figure 4.

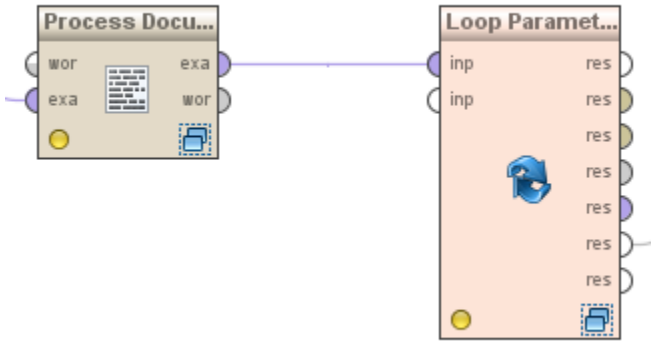

Figure 4, loop operator

6- The loop operator has sub processes inside it, the purpose of the loop operator is to cluster the tweets through k-means operator, then evaluate the clustering quality through other operators, then repeat the process of clustering and evaluating by incrementing the k value in each loop from k=2 to k=25. The sub processes inside the loop operator are shown in Figure 5

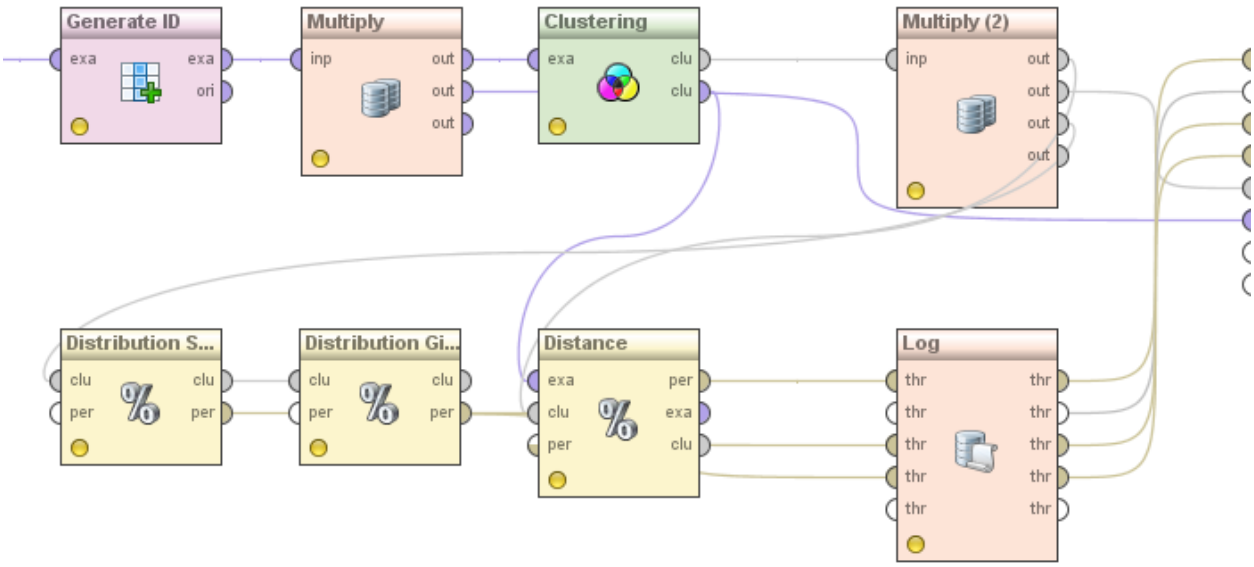

Figure 5, inside the loop model

7- As shown in Figure 5, generate ID operator is used to create a new attribute with an ID for each tweet, this is used for evaluation as evaluation technique will not operate without an ID for each document. Then, clustering the tweets is done through the k-means clustering operator. After that, the output of the clustering model is introduced into multiply operator to have multiple outputs of the clustering model; these outputs are introduced for each evaluation metrics which were explained previously in section 3.5.1. The output of each evaluation metric is then introduced into log operator which will log the result values of the evaluation techniques and save them into a text file stored on the machine.

8- The output of the loop iterations of the loop operator is then introduced into "Log to Data" operator, which is used to log the output of the loop operator and produce and example set of that output which can be shown through the results perspective of the RapidMiner GUI as shown in Figure 6

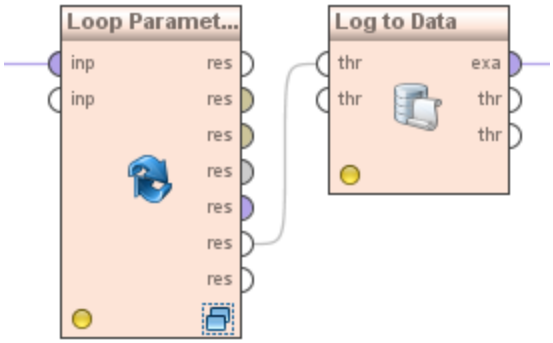

Figure 6, log to data operator

9- The final RapidMiner model explained in the above points is shown in Figure 7

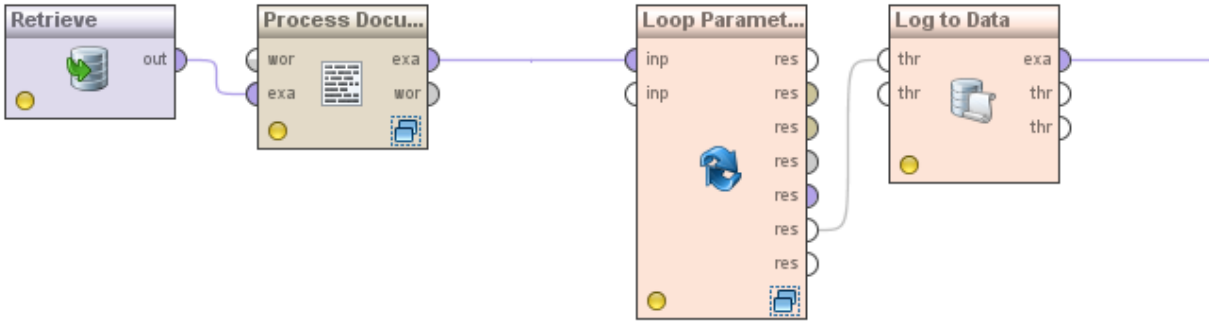

Figure 7, final model of RapidMiner

Chapter #4

4. Experimental Analysis

This chapter of the report delivers a clear discussion and analysis about the results of the experiments on the Dubai tweets datasets, and how are the results addressing the purpose of this research.

4.1 Evaluation by internal clustering evaluation techniques

For the experimental purpose of this section, only Dataset A is used in sections 4.1 & 4.2.

After building the clustering model and running the k-means algorithm on Dataset A, the evaluation metrics used in this research produced its results. The results of each metric produced different clusters validation assumptions. The heuristics of these tools did not give a definite answer about the quality of clusters and what is the best number of k to use in the cluster algorithm.

We'll rely on visual chart plots for studying the quality of clusters and determining the best number of k clusters to use as proposed by the heuristic techniques.

In SSE, we'll use the elbow method from the visual chart to look for the best number of clusters to use in our research. In elbow method, the SSE value will decrease dramatically by incrementing the k value of clusters, until it reaches a point where the value of SSE reaches a plateau while we increment the value of k (Kodinariya and Makwana, 2013).

As shown in Figure 8, the value of SSE decreases rapidly until it reaches the edge of the elbow at k=5, after this point, the decrease of the SSE value is slight and that means that adding more clusters to the model will not have a big effect on the quality of clusters.

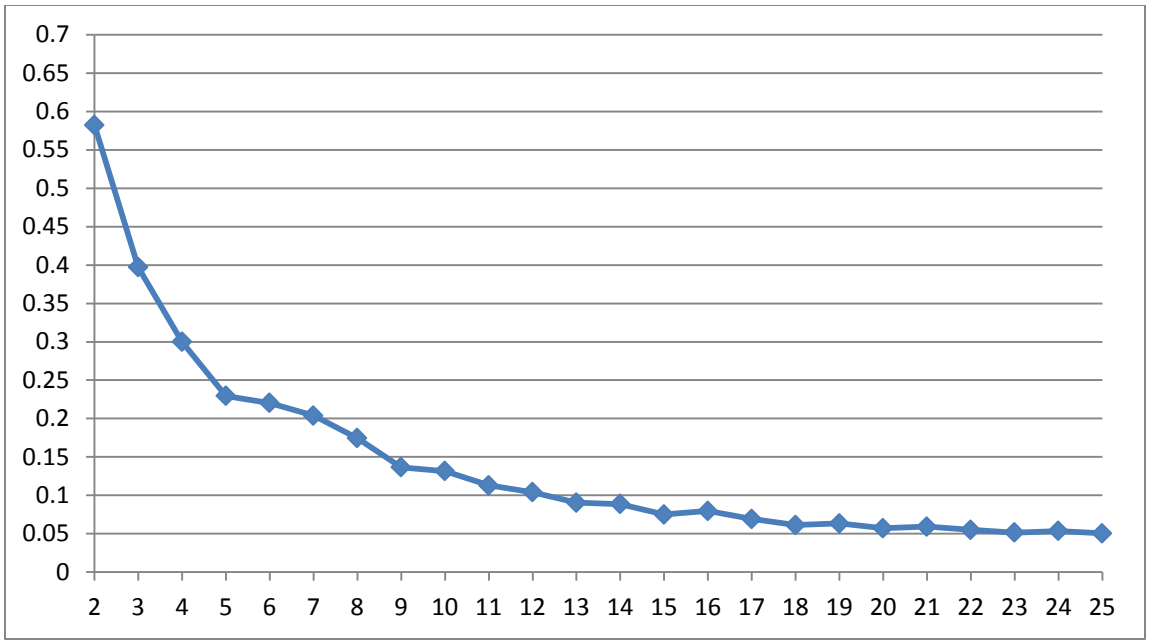

Figure 8, Sum of Squares Error

While the SSE indicates that k=5 is the ideal number of clusters to use in our model. The Davies-Boulding index has another suggestion. The Davies-Bouldin index will look into the intra and inter cluster, it will measure how compact are the points in one cluster and how well they are separated from the other clusters.

The minimum value of DBI is considered the best number of k clusters on use in the model. In our research, DBI indicates that the best number of k=24 as shown in Figure 9. Unlike the SSE, the value of DBI is not necessarily decreasing with the increment of k; it increases sometimes to indicate that a bigger number of clusters is not always a better model. For example, even though the best DBI value among our 25 iterations is the value of k=24, the Figure 9 shows that the value of k=25 produces a worse clustering model than k=24.

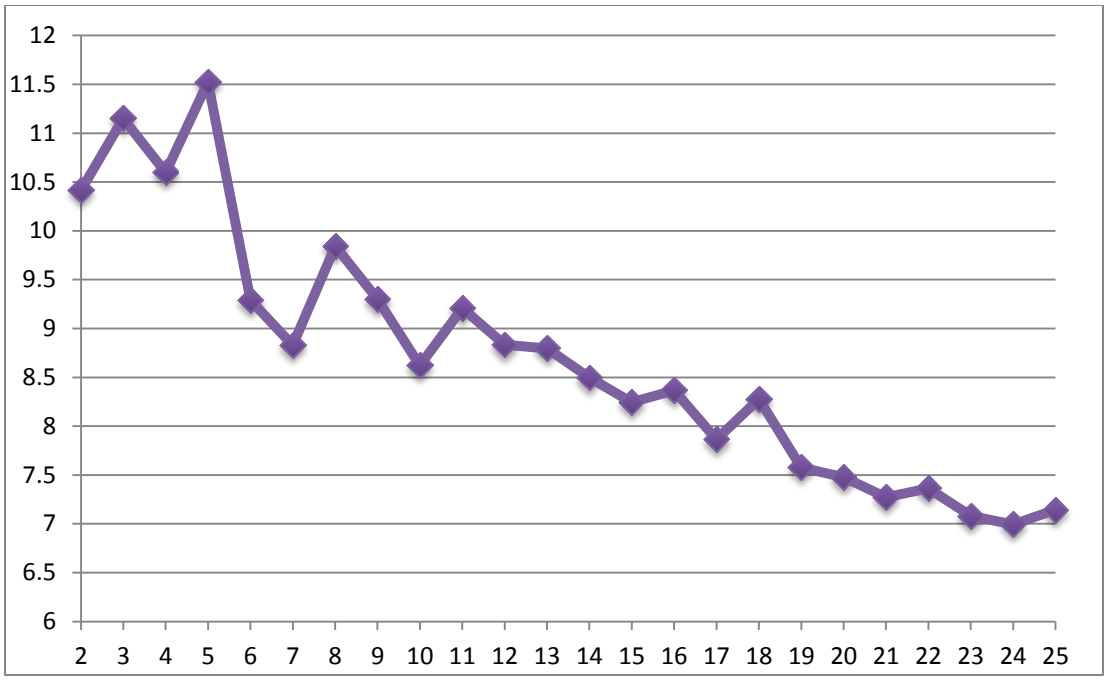

Figure 9, Davie-Bouldin Index

Regarding the average within centroid distance, Figure 10 shows that there is no elbow moment in the figure. The value of the average is close as the number of clusters increases, from k=2 to k=25 the values vary from 0.98 to 0 0.91. This indicates that the average distance of the data points to its cluster distance is still very close even after increasing the number of k many times, which means that there is no indication by this method about the best number of k to use in order to have a better clustering model.

The Gini coefficient also almost produces the same results, as the value of Gini is almost the same for all iterations; it is very close to 1 which indicates that all clusters have inequity distribution of points inside its clusters. The value for Gini coefficient for k=2 is 0.999985906, and for k=5 is 0.999963307, for k=15 is 0.999905822, and for k=25 is 0.999772728. All of the values are almost the same and are close to 1. Thus, the Gini coefficient is not helping in determining the best number of k to use in our clustering model.

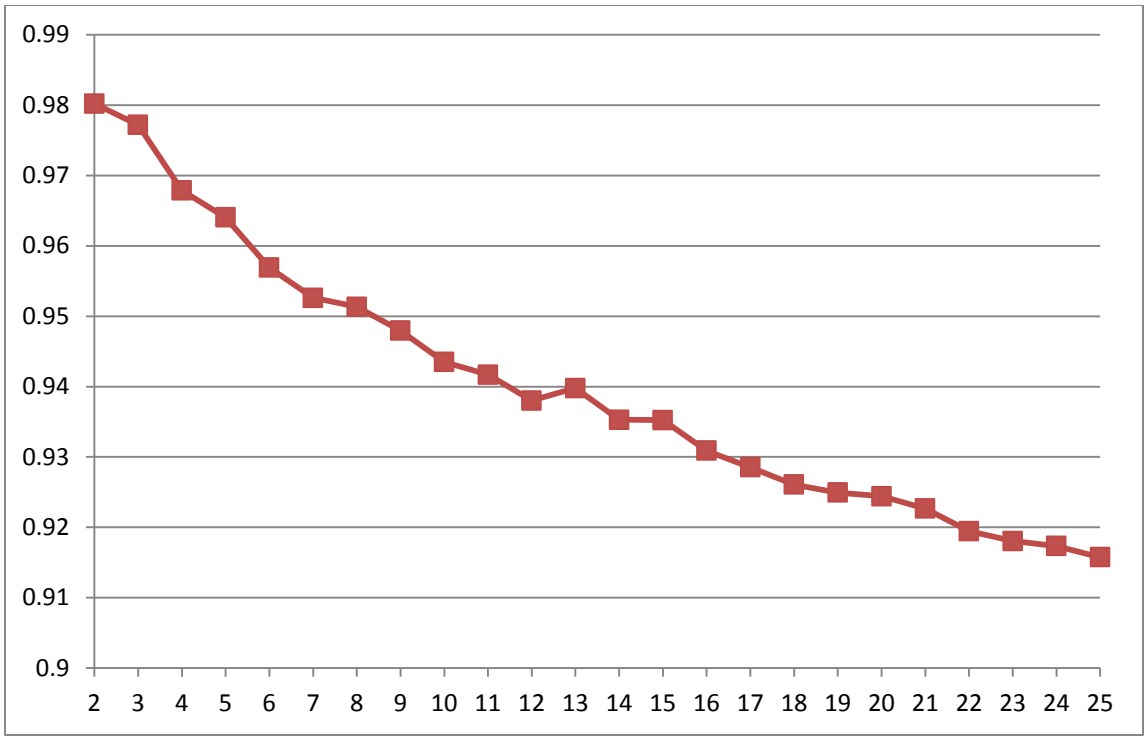

Figure 10, average within centroid distance

4.2 cluster Evaluation by model analysis

The heuristic internal clustering tools did not give convincing results about the quality of the produced clusters and about the best k value to choose on our dataset. The results were different and not helpful, so, relying on the observation and analysis of clustering model and looking into the data will be the dominant factor to produce best clusters out of the dataset we have. A number of tests will be applied on the dataset with different number of k; each outcome of the clustering model will be analysed and discussed.

It's worthy to note that terms in each of the example tables are originally stemmed in the cluster model, but for the ease of presentation the stems are removed when described in the report.

Also, when presenting the clustering results for different models, the redundant clusters from different models will not be explained again. For example, if a clustering model of k=5 produced a cluster about an event in Dubai, and another experiment with k=10 produced a

same cluster topic, the examples will be discussed and presented only once in the first model.

A lot of clustering models will produce unclear clusters that do not indicate a topic about the cluster and the tweets do not share any topic or interest. These clusters will be ignored and examples of them will be presented only once in the first case clustering model.

It's important to note that the key terms in each cluster example explained later are sequenced according to their importance to the cluster, which means that the first term is considered the final centroid of the cluster, and the following terms are sequenced according to their mean in the cluster.

Last important point to mention is that URLs are removed from tweets examples listed below, so some tweets may not be complete as they are titles to a full article available if you click the URL, but those URLs are removed from examples.

4.2.1 Experiment #1

First test is applied with k=5, which is the optimum value of k given by the SSE technique discussed in the previous section. Table 3 shows the results of the clustering model with k=5.

Table 3, results of model k=5

cluster	size	Top terms	Topic
Cluster_0	4,976	dubai, emirates, arab, united, united_arab	Unclear
Cluster_1	2,340	federer, final, djokovic, tennis, roger, rogerfederer	Dubai Tennis Championship
Cluster_2	8,063	mydubai, uae, dubai_mydubai, dubai_uae, mydubai_dubai	Unclear
Cluster_3	2,464	world, cup, dubai_world, world_cup, dubai_world_cup	Dubai World Cup horse racing
Cluster_4	10,944	dubai, mydubai, day, get, come, time, want	Unclear

As shown in Table 3, two of the five clusters produced by the model were informative and the topic of these clusters could be guessed through the keywords. While the other 3 clusters have an unclear topic from the top keywords they contain, these 3 clusters have the top words “dubai” and “mydubai” which are almost available in each tweet in our dataset.

First, we’ll discuss one unclear cluster and present some tweets of it to prove that they don’t share a similar topic. In this model, cluster_4 has an unclear topic based on the keywords produced by the model. It is also the biggest produced cluster with 10,944 tweets in it, which means that 38% of the total dataset are in one cluster. By looking into this cluster, the tweets are totally far from each other and they don’t share any topic, examples of the tweets inside cluster_4 are shown in table 4.

Table 4, examples of tweets in cluster_4, k=5

Examples of tweets in cluster_4, k=4
- A Dubai luxury home developer plans to create a tropical rainforest under a dome for visitors - Exclusive designer collection now also available in Dubai, Limited stocks so book yours fast - Come, live and enjoy this #NYE experience in #Dubai it is a different parts of life world #DXBNYE #MyDubai - FIRST LOOK Al Khazzan Park – the new Safa Park alternative? #dubai - Dubai trip inspired Middle Eastern Lunch at @BKKofta kebabs in pita with Tzatziki sauce - Really wanna live in Dubai or Paris - @DXBMediaOffice Mattar Al Tayer Public transport daily ridership in #Dubai hits 1 million; 470 thousands passengers @RTA_Dubai

As shown in table 4, the examples of tweets show that there is no shared topic among them; it seems that the data points in this cluster are not well close unlike what was suggested by the heuristic techniques. The results of the other unclear clusters (cluster_2) and

(cluster_0) will be the same with no clear shared topic, so the results of the cluster won't be presented.

As for cluster_1, it is clear from the keywords that it is talking about the Dubai Tennis Championship which took place in February while collecting the data. The cluster has the keywords of Federer and Djokovic who are the two tennis players competed in the final match of the tournament. The terms final and tennis also indicate that the cluster is talking about this topic. Example of cluster_1 can be shown in table 5.

Table 5, examples of tweets in cluster_1, k=5

Examples of tweets in cluster_1, k=5
- Congratulations @rogerfederer for winning @DDFTennis championship 7th time! - So far this year, Federer in Dubai was the best form I saw from a player - Finally have a copy of the Dubai 2015 final match! #FedererVsDjokovic - What a game!!! The winning shot video extremely well played @rogerfederer and great effort by @DjokerNole #Dubai - Roger Federer Wins Seventh Dubai Crown - Tennis - ATP World Tour - Djokovic earns Dubai showdown with Federer - Dubai 2015 Novak Djokovic beats Berdych to set up Federer title duel - @tennis_result @DjokerNole beat Andrey Golubev 6-1 6-2 in the 2nd round at Dubai #ATP #tennis - Stay tuned!!! #icymi tennis major upset in Dubai as Murray crash out a case of David vs Goliath

For the final cluster (cluster_3) in the first model where k=5, the cluster is clearly talking about Dubai World Cup event which took place in March, it is the biggest horse racing event in the world. The key terms in this cluster were dubai, world, and cup, with n-gram combination of these term. Examples of tweets of cluster_3 are shown in table 6.

Table 6, examples of tweets in cluster_3, k=5

Examples of tweets in cluster_3, k=5
- RT @HHShkMohd Congratulations to the UAE Congratulations to Hamdan for Prince Bishop's winning of the 20th Dubai World Cup - Congratulations to Prince Bishop for becoming the oldest horse (age 8) to win the Dubai World Cup! 2nd was California Chrome - Prince Bishop wins Dubai World Cup - My dad sending me pictures from the Dubai World Cup like you could have been there - Prince Bishop Denies US Runners in Dubai. Sheikh Hamdan bin Mohammed Al Maktoum's Prince Bishop overpowered - What a great day of horse racing #Dubai #DubaiWorldCup #Gulfstream #FloridaDerby #FairGrounds #LouisianaDerby #firstsaturdayinmay

4.2.2 Experiment #2

In Experiment #2, we'll increase the number of clusters into k=8, to see if increasing the number of clusters slightly will produce any extra topic oriented clusters. The results of the clustering model of k=8 are shown in table 7.

Table 7, results of model k=5

cluster	size	Top terms	Topic
Cluster_0	2,441	job, emirates, arab, united, united_arab, dubai_job, united_arab_emirates, manage	Jobs
Cluster_1	5,866	mydubai, dubai_mydubai, mydubai_dubai, dxb, morning, beautiful, love	Unclear
Cluster_2	7,175	day, love, come, get, see, dubai, want, time	Unclear
Cluster_3	2,148	federer, djokovic, final, tennis, roger, murray	Dubai Tennis Championship
Cluster_4	2,644	dubai, meanwhile, dubai_meanwhile	Unclear
Cluster_5	2,600	uae, dubai_uae, mydubai, mydubai_uae,	Unclear

		dubai_mydubai, uae_dubai	
Cluster_6	1,510	world, cup, dubai_world, world_cup, dubai_world_cup, prince, chrome	Dubai World Cup horse racing
Cluster_7	4,403	hotel, city, travel, airport	Travel & hotels

The clustering model with k=8 as shown in table 7, produced the same 2 expressive clusters were in model of k=5, which are cluster_3 talking about Dubai Tennis Championship, and cluster_6 which is talking about Dubai World Cup horse racing.

This model did not only retain the same topical clusters, but also produced 2 more important clusters which are cluster_0 and cluster_7.

Cluster_0 is talking about jobs in Dubai and United Arab Emirates as the keywords show. Job is the final centroid of this cluster which means that this cluster is mainly talking about jobs. Table 8 will show examples of tweets in cluster_0.

Table 8, examples of tweets in cluster_0, k=8

Examples of tweets in cluster_0, k=8
- Are you a good fit for this job? Senior Arabic Copywriter in Dubai #job #Arabiccopywriter - Be first to apply – MBBS Doctors Jobs Dubai in Dubai, United Arab Emirates - I'm hiring – Account Director in #dubai #job #mydubai - Job Opportunity - Communication Manager in Dubai #job #mepra #pr - URGENT JOBS IN DUBAI as Business Development Manager / Landscaping, Draftsman, Group Marketing Manager, Hotel manager - Great job opportunity ServiceNow (SNOW) Consultant in Dubai, United Arab Emirates #job - Now hiring for: Digital Marketing Analyst in Dubai, United Arab Emirates

As for cluster_7, this cluster is talking about travelling to Dubai, people tweeting about their flights, hotels in Dubai, recommendations about reservations, and everything related to travel as examples are shown in table 9.

Table 9, examples of tweets in cluster_7, k=8

Examples of tweets in cluster_7, k=8
- Best Dubai hotels; View best prices - RT @e_Travel Dubai has more hotels under construction than any other city in the world - Dubai hotel occupancy drops 19% in February Imbalance in supply-demand equation behind falling room rates - planning to travel to #dubai next month, any hotel recommendations guys? - This was my view of the #PalmIslands from a seaplane my mind was blown! #travel #Dubai #Lp - This time tomorrow I will be sitting in the BA lounge in Heathrow airport waiting for my flight to Dubai ,bring it on lol - @MattPJ1971 Welcome to Dubai Airport flight information service for EK229 DXB-SEA - trip to dubai in May, atlantis hotel stay or saray zabeel? - Time to go home — travelling to Dubai, United Arab Emirates from Suvanabhumi International Airport (Bangkok)

4.2.3 Experiment #3

In this experiment, the value of k has been increased to k=15, to check if this model will discover any extra interesting clusters which were hidden in the previous experiments. The interesting clusters which are expressive through their keyword terms are listed in table10; other clusters which are unclear are ignored and won't be mentioned.

Table 10, results of model k=15

cluster	size	Top terms	Topic
Cluster_2	2,156	federer, djokovic, final, tennis, roger, murray	Dubai Tennis Championship
Cluster_4	1,323	world, cup, dubai_world, world_cup, dubai_world_cup, prince, chrome	Dubai World Cup horse racing
Cluster_5	1,396	hotel, beach, travel, marina, hotel_dubai	Hotels
Cluster_8	981	mall, dubai_mall, drake, khalifa, burj_khalifa	Dubai mall & Burj Khalifa
Cluster_10	1,232	Job, dubai_job, account, manage, director	Jobs

As shown in table 10, the clustering model has produced 5 interesting and topical expressive clusters, while the other 10 clusters where unclear and not interesting to investigate. This cluster model has produced the same 4 clusters in the previous experiment where k=8. What's interesting are the results of cluster_8 which is talking about Dubai mall and Burj Khalifa, this topic was not mentioned in the previous experiments. This cluster has also the term Drake, whom is a singer who performed in Dubai in March, and was spotted at the Dubai mall. Examples of the cluster_8 are shown in table 11.

Table 11, examples of tweets in cluster_8, k=15

Examples of tweets in cluster_8, k=15
- A little bit of Michael Jackson's thriller at the Burj Khalifa fountains #mydubai #MichaelJackson - Official Burj Khalifa, downtown Dubai 2014 New year's eve highlights video - They got a built in aquarium in a mall at dubai?! Bruh - In the heart of Dubai Mall is the world's largest piece of acrylic at 328 metres wide by 83 metres - I SAW DRAKE IN DUBAI MALL!! - RT @Mahaya_Al Drake casually walking in dubai mall

- So it's 8PM and we're at the Dubai Mall, you know the biggest mall in the world Big enough to scuba dive in

4.2.4 Experiment #4

In this experiment, we'll use the k-means based clustering algorithm which is x-means. This new algorithm is an extended version of the k-means which helps by heuristics to determine the best number of k clusters to use in the model without the need for the user to supply the number of k before the experiment. Pelleg and Moore (2000) proposed this new model which works with a minimum number of centroids and minimum number of base clusters provided by the user, then in each iteration the model run the k-means algorithm on the data, if adding more clusters will enhance the quality of clustered data it will do so by splitting clusters based on the Bayesian Information Criteria (BIC). They tested the algorithm on a 4-dimentional dataset, in this experiment the effectiveness of x-means will be questioned on high dimensionality text data.

Before implementing the x-means algorithm, the minimum number of k provided to the algorithm is k=4, and the maximum number of clusters is k=60. The outcome of the clustering model performed by x-means produced 8 clusters, description of the clusters are shown in table 12.

Table 12, results of x-means model

cluster	size	Top terms	Topic
Cluster_0	1,539	world, cup, world_cup, dubai_world, dubai_world_cup	Dubai World Cup horse racing
Cluster_1	2,113	federer, djokovic, final, tennis, roger, win, murray	Dubai Tennis Championship
Cluster_2	6,591	get, dubai, see, time, year, look	Unclear
Cluster_3	5,000	hotel, day, come, travel, city	Travel & hotels
Cluster_4	2,091	job, emirates, arab, united, dubai_jobs, united_arab	Jobs

Cluster_5	2,929	dubai, , meanwhile, dubai_meanwhile, dubai_dubai,	Unclear
Cluster_6	6,127	mydubai, dubai_mydubai, love, mydubai_dubai, morning	Unclear
Cluster_7	2,397	uae, dubai_uae, mydubai_uae, dubai_mydubai	Unclear

As shown in table 12, the x-means algorithm has decided to stop splitting the clusters at k=8. The results of the 8 clusters are the same as the results of the k-means when k=8. There is no additional interesting cluster created by the x-mean. The same 4 interesting clusters are generated by the two algorithms which are: Dubai World Cup, Dubai Tennis Championship, travel & hotels, and jobs.

4.2.5 Experiment #5

Based on the above experiments, it seems that increasing the number of clusters will generate more specific topic-oriented clusters, which are interesting to know on Dubai tweets and what are the people talking about Dubai.

In this experiment, the value of k has been increased to 40; the results of the interesting clusters generated by this clustering model are shown in table 13.

Table 13, results of model k=40

cluster	size	Top terms	Topic
Cluster_1	455	time, time_dubai, dubai_time, great, great_time	Great time in Dubai
Cluster_2	729	job, manage, job_dubai, dubai_job, manage_dubai, vacancy	Jobs
Cluster_4	454	murray, andy, andy_murray, tennis, coric	Dubai Tennis Championship

Cluster_6	739	travel, go, trip, go_trip, trip_dubai, desert, dubai_trip, safari, holiday	Travel and trips
Cluster_9	533	hotel, hotel_dubai, dubai_hotel	Hotels
Cluster_12	571	sale, tower, jumeirah, apartment, house	Properties
Cluster_17	1,132	Federer, djokovic, final, roger, roger_federer	Dubai Tennis Championship
Cluster_18	964	world, cup, dubai_world, dubai_world_cup, prince, bishop	Dubai World Cup horse racing
Cluster_20	880	united, arab, job, emirates, united_arab_emirates,	Jobs
Cluster_28	526	police, dubai_police, luxury, fleet, bmw	Dubai police
Cluster_29	214	miracle, open, miracle, dubai_miracle, dubai_miracle_garden	Dubai Miracle Garden
Cluster_30	322	dhabi, abu, abu_dhabi,	Abu Dhabi
Cluster_31	583	mall, dubai_mall, burj, burj_khalifa, khalifa, drake, drake_dubai	Dubai Mall & Burj Khalifa
Cluster_34	638	airport, flight, dubai_airport	Dubai Airport
Cluster_36	553	festival, jazz, dubai_jazz, art, dubai_festival	Dubai Jazz Festival
Cluster_38	412	marina, dubai_marina, place, view	Dubai Marina

As shown in table 13, this clustering model with k=40 has generated a number of interesting clusters which were hidden in previous experiments. A total of 16 clusters where interesting to know and look at out of the 40. Some of the clusters were mentioned before in previous experiments and some of them are new.

What's noted from this model that there are redundant clusters, for example, cluster_2 and cluster_20 are both talking about jobs. Another redundant clusters are cluster_4 and cluster_17, but by looking at the keywords, we can figure out that each one of these clusters is talking about one match. Cluster_4 is talking about a match between the tennis player

Murray and Coric, while the other cluster_17 is talking about the final match between Federer and Djokovic.

Cluster_1 is a new generated cluster which it talking about people tweeting about their time they are having in Dubai. Examples of tweets in cluster_1 are shown in table 14.

Table 14, examples of tweets in cluster_1, k=40

Examples of tweets in cluster_1, k=40
- spent a great time in dubai, now off to London - Reached Dubai today from London. Thank you to all my friends there who looked after me and I had a great time and got material on #Joker - We had such a great time at yesterday's swoop competition with all the skydivers and staff #SwoopDubai #Dubai - Unreal difference, been here a few times and it changes every time! Incredible journey Dubai #mydubai - Finally adjusted to Dubai time and slept all through the night

Another interesting outcome is that the cluster which was in previous experiment (k=15) talking about travel, hotels, and airport is now separated into three clusters where each cluster is now talking about one topic of these. The clusters now in this model are more topic-specific and each term of these has been assigned to one cluster to talk about one specific topic. These clusters are cluster_6, cluster_9, and cluster_34. Examples about each cluster of them are listed in table 15, table 16, and table 17.

Table 15, examples of tweets in cluster_6, k=40

Examples of tweets in cluster_6, k=40
- Book a #dhowcruise trip in our Travel Desk located in the Lower Plaza level or call 04 439 8888 #cruises #RamadaPlazaJBR #MyDubai #Travel - Playing the part at the desert! #MyDubai #BeAHero Arabian Desert, Dubai - Follow us for best day trips in Dubai - I wish I could travel to Dubai #dreamdestination

- Dubai is definitely one of the first places I wanna travel too - Dubai is the first place on my travel list

Table 16, examples of tweets in cluster_9, k=40

Examples of tweets in cluster_9, k=40
- One of the most expensive hotels in #Dubai Burj Al Arab #Day2 #HalfDayCityTour #LatePost - Get the best price Dubai hotels; See deals now! - Top 10 Most Affordable #Hotels in #Dubai: - RT @e_Travel Dubai has more hotels under construction than any other city in the world httptcosmtmNa8b17 - RT @lorraine_bogan @LeMeridienDXB - see you on the 20th for another week in our favorite hotel with the best staff in Dubai #sunshine

Table 17, examples of tweets in cluster_34, k=40

Examples of tweets in cluster_34, k=40
- Dubai Airport is *big* This is just the baggage claim area - @kamalgro Welcome to Dubai Airport flight information service for EK515 DEL-DXB - Dubai airport should learn from changing and even Mumbai International Airport about how to provide free wifi Such a must for travelers - RT @DXBMediaOffice CNBC In the last year, #UAE Airports welcomed 101 million passengers with #Dubai Airport becoming the busiest international airport

Another interesting cluster which was not generated in previous experiments is cluster_12, which is talking about properties for sale or rent, and real estate in Dubai, especially in Jumeirah district. It is a very important topic in Dubai which is known globally for its real estate. It seems that this model has discovered some important tweets talking about this topic. Examples of cluster_12 are shown in table 18.

Table 18, examples of tweets in cluster_12, k=40

Examples of tweets in cluster_12, k=40
- Palm Jumeirah, Golden Mile 6, Dubai, Dubai 2 bedroom apartment – For sale - Beautiful 5Bedroom villa in Zabeel Saray Jumeirah Cresent For Sale #viila #forsale #sale #palm #jumeirah #dubai #dxb - Business #Bay, #Executive Tower K, Dubai, Dubai House – For Sale - DIFC, Park Tower 1, Dubai, Dubai 1 bedroom apartment – For Sale - Emaar launches Acacia at Park Heights apartments in Dubai - Apartment and villa prices fall in #Dubai at beginning of #UAE #DXB #MyDubai #Property #Realestate - #Apartment for #sale in #AlHamri, Shoreline Apartments, Dubai, Price 3,050,000 AED Area 2,159 SqFt - looking for an apartment in jlt Anyone can advise? #real #estate

Cluster_28 which is talking about Dubai Police was also hidden in previous experiments; this cluster contains tweets about Dubai police news and announcements. Besides to that, people are tweeting with photos about the new luxurious Dubai Police fleet cars. Examples about this cluster are shown in table 19.

Table 19, examples of tweets in cluster_28, k=40

Examples of tweets in cluster_28, k=40
- Car on fire in JLT, try avoiding #mydubai @RTA_Dubai @DubaiPoliceHQ - RT @mrabbouduae What a fast response from Civil Defense and Dubai Police #mydubai @DubaiPoliceHQ @DXBMediaOffice #JLT - Only in Dubai - Dubai Police Cars | Dubai Police Force is known worldwide for its super cars fleet - Dubai Police Will Wear #GoogleGlass With Facial Recognition Software to ID Crooks - The Dubai police fleet includes a Ferrari, Lamborghini and Bentley so they can catch speeders that would be able to outrun other cars - A BMW i8 is the latest exotic car to join the Dubai Police fleet

Another new cluster is cluster_29, which from the keywords is talking about the Dubai Miracle Garden, which is an outdoor flower garden in Dubai. People tweeting about their experience in the garden are shown in table 20.

Table 20, examples of tweets in cluster_29, k=40

Examples of tweets in cluster_29, k=40
- Finally visited the flower place #mydubai #miraclegarden an amazing place A must visit - A majestic day in the #garden | Welcome to Dubai Miracle Garden #visitdubai #mydubai #dubailife - Dubai Miracle Garden The world's biggest natural flower garden - Dubai miracle is a real miracle! - A new extreme world wonder? Miracle Garden makes 45 million flowers bloom in the Dubai desert

Even though the collected data set is related to Dubai city, there is a portion of these tweets talking about Abu Dhabi, the capital of UAE. This cluster (cluster_30) was not discovered in previous experiments. A list of examples of this cluster is listed in table 21.

Table 21, examples of tweets in cluster_30, k=40

Examples of tweets in cluster_30, k=40
- How the experts plans to tackle Tri Yas in Abu Dhabi - Abu Dhabi Police reveal how they caught Reem Island murder suspect - via 7daysuae #MyDubai #Dubai - Visit to sheikh zayed mosque in Abu Dhabi #inabudhabi #sheikhzayedmosque #abudhabi #dubai #shots - Reuters – State-run Abu Dhabi National Oil Company ADNOC has signed a technical agreement

Dubai Jazz Festival was a musical event which took place in February; this clustering model has discovered a new cluster (cluster_36) which contains tweets talking about this festival.

This even was not mentioned in the previous clustering models and tweets about this topic were dispersed as outliers in other clusters. Tweets examples of cluster_36 are shown in table 22.

Table 22, examples of tweets in cluster_36, k=40

Examples of tweets in cluster_36, k=40
- The #magic that James Blunt is! #DubaiJazzFestival #mydubai - Emirates Airline Dubai Jazz Festival 2015 Feat John Legend | 1200 more Regular tickets have been released - RT @dubaijazzfest more than 8,000 fans attended the opening night of the Dubai Jazz Festival Feat James Blunt & Christina Perri - Last night Sting, tonight John Legend I love Dubai Jazz Festival - What happened when we spoke to the talented johnlegend ahead of the dubaijazzfest?

The last interesting cluster to mention in this model is cluster_38, which is talking about the Dubai Marina. This cluster has not been generated by previous models. From the examples listed in table 23, we can see that this cluster is containing some tweets about a fire incident in Dubai Marina where the 5th tallest residential building in the world was on fire.

Table 23, examples of tweets in cluster_38, k=40

Examples of tweets in cluster_38, k=40
- Dubai #Marina 2 #bedroom apartment - For Sale - Huge blaze guts 79-floor Dubai skyscraper in Dubai Marina - EXPOSED! Pics Fire rips through one of the world's tallest residential buildings in Dubai The Dubai Marina - RT @JannahHotels Come to Jannah Place Dubai Marina experience the amazing service - Dubai Marina Torch tower fire latest 20 free hotel nights for residents - Partial Sea View Large Studio in My Tower Dubai Marina , AED 892,095

4.2.6 Experiment #6

In this experiment, the number of clusters is increased to k=50, in order to check if increasing the number of clusters will generate more interesting and exciting clusters, or the model of k=40 was enough for this dataset. The outcomes of the interesting clusters from this model are mentioned in table 24

Table 24, results of model k=50

cluster	size	Top terms	Topic
Cluster_0	1,031	rogerfederer, ddftennis, live, win	Dubai Tennis Championship
Cluster_2	546	burj, khalifa, photo, burj_khalifa	Burj Khalifa
Cluster_7	484	police, car, dubai_police, luxury, fleet	Dubai Police
Cluster_13	738	job, dubai_job, account, job_dubai	Jobs
Cluster_22	478	marina, dubai_marina, look, view	Dubai Marina
Cluster_23	1,005	cup, world, world_cup, dubai_world_cup	Dubai World Cup
Cluster_26	363	mall, dubai_mall, drake	Dubai Mall
Cluster_37	427	travel, dubai_travel, travel_dubai	Travel
Cluster_38	513	hotel, dubai_hotel, hotel_dubai	Hotels
Cluster_48	1,401	feder, djokovic, final, roger, roger_federer	Dubai Tennis Championship

Table 24 shows that results generated from this clustering model are not exciting, the model where k=40 generated much better clustering results. This model did not produce clusters about some important topics which were mentioned in the previous experiment, such as: properties, Abu Dhabi, Dubai Jazz Festival, Dubai airport, and Dubai miracle garden. The tweets which were talking about these topics were dispersed among other meaningless clusters. It seems that increasing the number of clusters above 40 will not produce any better results.

4.3 Constant clusters over time

In this section, Dataset B is called for the first time in the report.

One of the questions raised is to check what the people are mostly tweeting about in Dubai, what are the similar topics about Dubai trending in Twitter over time. In order to check that, we need to build clustering model for datasets collected over different periods of time.

For this purpose, we will compare clustering models for 3 datasets which are from different time periods (Feb, Mar, and May 2015), the three dataset are:

- February: A number of 3,356 tweets were chosen randomly from Dataset A which are over the period of Feb 2015
- March: A number of 3,688 tweets were chosen randomly from Dataset A over the period of Mar 2015
- May: Dataset B is called for the first time of the report; it consists of 5,906 tweets collected in May 2015.

The 3 datasets were introduced into RapidMiner to build a k-means clustering model. The value of k chosen for the 3 models is k=10. We'll check from the generated clusters what are the top keywords shared in clusters among the 3 models, this will give us a hint about what are the top trending topics in Dubai on Twitter.

Table 25 shows the interesting clusters generated from the clustering model of February tweets.

Table 25, results of clustering model in February

cluster	size	Top terms	Topic
Cluster_1	275	federer, roger, rogerfederer, tennis, djokovic	Dubai Tennis Championship
Cluster_4	225	andy, murray, andy_murray, tennis, championship	Dubai Tennis Championship
Cluster_8	257	hotel, burj	Hotels

Cluster_9	391	job, dubai_job, uae, job_uae	Jobs

Table 26 shows the interesting clusters generated from the clustering model of March tweets.

Table 26, results of clustering model in March

cluster	size	Top terms	Topic
Cluster_1	274	hotel, see, burj	Hotels
Cluster_7	340	job, uae_job, hire	Jobs
Cluster_8	695	World, dubai_world, cup, dubai_world_cup	Dubai World Cup

Table 27 shows the interesting clusters generated from the clustering model of May tweets.

Table 27, results of clustering model in May

cluster	size	Top terms	Topic
Cluster_1	225	flight, diver, threat, media	Flight divert incident
Cluster_2	338	job, dubai_job, manage, account	Jobs
Cluster_3	599	Onedirection, luise, directioncalvin, track	One Direction band concert in Dubai
Cluster_8	397	hotel, hotel_dubai	Hotels

From table 27, we can see that the clustering model generated 4 interesting clusters. In cluster_1, the cluster is talking about a flight which has diverted its landing direction to a military base because of threats from a terrorist social media accounts. And cluster_3 was talking about a concert in Dubai for One Direction musical band.

The other two clusters generated from May model are jobs and hotels. As shown in the above tables, these two topics were repeated in each clustering model for the three experiments conducted about tweets in February, March, and May 2015.

This experiment gives an interesting hint about what people are mostly tweeting about Dubai, the two most trending topics as per the output of the experiment are jobs and hotels.

Chapter #5

5. Conclusion and Future Work

This chapter of the report concludes the research with remarking findings and contributions. It also constructs a solid base for opening future work.

5.1 Conclusion

Twitter is an exciting and emerging field of research, conducting text mining techniques on tweets is a challenging task due to the sparseness, vastness, and informality of text tweets.

In this work, a corpus of tweets talking about Dubai has been collected to conduct totally unsupervised clustering techniques.

By using text mining combined with k-means clustering algorithm and cosine similarity measure, and after conducting a number of experiments, this research showed a process for clustering Dubai tweets into topic-related clusters.

This work has also shown that internal heuristics clustering evaluation techniques such as sum of squares error and Davies Bouldin Index are not helpful in determining the right number of cluster in initiate the experiments. Those techniques can underestimate the required number of clusters to discover in such a domain. Empirical tests have also shown that the best clustering achieved was when the number of clusters = 40. This may propose to start with this number of clusters for clustering tweets of such domain.

The proposed approach has shown that each cluster centroid keyword gives a semantic meaning about the topic of that cluster. It also partitioned tweets into a number of

expressive theme-based clusters. Those clusters may be events, places, attractions, news, main topics, and more.

An interesting finding is the ability of the model to detect and identify major events in Dubai, for example, this model was able to cluster tweets talking about the Dubai World Cup horse racing event in March 2015, and it was also able to detect Dubai Tennis Championship even in February 2015.

Furthermore, the model has shown that jobs and hotels are the two most trending topics in Dubai people are talking about in Twitter. This finding was reached through conducting the clustering model on different datasets; each of the dataset was collected in different period of time.

Such a model will be very helpful in identifying twitter user accounts who share the same interest in a topic; this may help in targeting those user accounts for marketing campaigns.

5.2 Future Work

This work can be expanded by collecting a larger volume of data across multiple months, and see if jobs and hotels are still the two top trending topics about Dubai in Twitter, it is interesting to know if these topics will be still on top in summer, as summer in Dubai is not a high season of hotels and hospitality.

It will be interesting to know what the top topics are about other cities in UAE; it is well-known that Dubai is the top tourism city in UAE and the top attraction for job seekers, so it will be interesting to see what other neighbour cities in UAE and also in the region have as trending topics.

A comparative study could be conducted in comparing the results generated by k-means algorithm and other clustering algorithms. Also, previous studies mentioned in related work have proven that Jaccard similarity measure can generate competitive results on text

mining, so it will be interesting to compare the results obtained by Jaccard with the results of cosine similarity.

Finally, this model can be examined on larger set of data with the use of dimensionality reduction techniques such as Latent Semantic Indexing (LSA) and Principal Component Analysis (PCA). It will be interesting to know if these techniques can enhance the quality of the clustering model.

References

Adel, A., ElFakharany, E., & Badr, A. (2014). CLUSTERING TWEETS USING CELLULAR GENETIC ALGORITHM. *Journal of Computer Science*, *10*(7), 1269-1280.

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

AlSumait, L., & Domeniconi, C. (2008). Text clustering with local semantic kernels. In *Survey of Text Mining II* (pp. 87-105). Springer London.

Antenucci, D., Handy, G. R. E. G. O. R. Y., Modi, A. K. S. H. A. Y., & Tinkerhess, M. I. L. L. E. R. (2011). Classification of tweets via clustering of hashtags.

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

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25-71). Springer Berlin Heidelberg.

Bruns, A., & Burgess, J. E. (2011). The use of Twitter hashtags in the formation of ad hoc publics.

Chen, Q., Shipper, T., & Khan, L. (2010, May). Tweets mining using WIKIPEDIA and impurity cluster measurement. In *Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on* (pp. 141-143). IEEE.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2), 224-227.

de Villiers, F., Hoffmann, M., & Kroon, S. (2012, September). Unsupervised construction of topic-based twitter lists. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* (pp. 283-292). IEEE.

Emirates 24/7. (2015, April 23). Dubai crowned Mena Smart City of the Year1980. Retrieved May 18, 2015, from www.emirates247.com/news/emirates/dubai-crowned-mena-smart-city-of-the-year-2015-04-23-1.588284

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*, *49*(9), 76-82.

Godfrey, D., Johns, C., Meyer, C., Race, S., & Sadek, C. (2014). A Case Study in Text Mining: Interpreting Twitter Data From World Cup Tweets. *arXiv preprint arXiv:1408.5427*.

Hearst, M. (2003). What is text mining. *SIMS, UC Berkeley*.

Hofmann, M., & Klinkenberg, R. (Eds.). (2013). *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.

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

Jones, K. S. (Ed.). (1997). *Readings in information retrieval*. Morgan Kaufmann.

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

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, *1*(6).

Kuzela, C. (Ed.). (2015, April 8). Smart Dubai: Out to be the world's best-connected, happiest city by 2017. Retrieved May 18, 2015, from www.siliconangle.com/blog/2015/04/08/smart-dubai-out-to-be-the-worlds-best-connected-happiest-city-by-2017/

MacKay, D. (2003). An example inference task: clustering. *Information Theory, Inference and Learning Algorithms*, 284-292.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Scoring, term weighting and the vector space model. *Introduction to Information Retrieval*, *100*.

Moh, T. S., & Bhagvat, S. (2012, March). Clustering of technology tweets and the impact of stop words on clusters. In *Proceedings of the 50th Annual Southeast Regional Conference* (pp. 226-231). ACM.

Mosley Jr, R. C. (2012). Social media analytics: Data mining applied to insurance Twitter posts. In *Casualty Actuarial Society E-Forum, Winter 2012 Volume 2* (p. 1).

Muntean, C. I., Morar, G. A., & Moldovan, D. (2012, January). Exploring the meaning behind twitter hashtags through clustering. In *Business Information Systems Workshops* (pp. 231-242). Springer Berlin Heidelberg.

Noordhuis, P., Heijkoop, M., & Lazovik, A. (2010, July). Mining twitter in the cloud: A case study. In Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on (pp. 107-114). IEEE.

Pelleg, D., & Moore, A. W. (2000, June). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *ICML* (pp. 727-734).

Perez-Tellez, F., Pinto, D., Cardiff, J., & Rosso, P. (2011, June). On the difficulty of clustering microblog texts for online reputation management. In*Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 146-152). Association for Computational Linguistics.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130-137.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3-13.

Rangrej, A., Kulkarni, S., & Tendulkar, A. V. (2011, March). Comparative study of clustering techniques for short text documents. In *Proceedings of the 20th international conference companion on World wide web* (pp. 111-112). ACM.

Rosa, K. D., Shah, R., Lin, B., Gershman, A., & Frederking, R. (2011). Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM*.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., & Sperling, J. (2009, November). Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems* (pp. 42-51). ACM.

Sculley, D. (2010, April). Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web* (pp. 1177-1178). ACM.

Shrestha, P., Jacquin, C., & Daille, B. (2012). Clustering short text and its evaluation. In *Computational Linguistics and Intelligent Text Processing* (pp. 169-180). Springer Berlin Heidelberg.

Strehl, A., Ghosh, J., & Mooney, R. (2000, July). Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)* (pp. 58-64).

Subhashini, R., & Kumar, V. J. S. (2010, August). Evaluating the performance of similarity measures used in document clustering and information retrieval. In*Proceedings of 1st International Conference on Integrated Intelligent Computing (ICIIC): August* (pp. 5-7).

Tripathy, R. M., Sharma, S., Joshi, S., Mehta, S., & Bagchi, A. (2014, March). Theme Based Clustering of Tweets. In *Proceedings of the 1st IKDD Conference on Data Sciences* (pp. 1-5). ACM.

Tsur, O., Littman, A., & Rappoport, A. (2013, June). Efficient Clustering of Short Messages into General Domains. In *ICWSM*.

Zhong, S., & Ghosh, J. (2003, April). A comparative study of generative models for document clustering. In *Proceedings of the workshop on Clustering High Dimensional Data and Its Applications in SIAM Data Mining Conference*.