Using Text Mining and Cluster Analysis to Improve Customers Complaints System

تطبيق تقنيات التنقيب في النص وتحليل المجموعة من أجل تحسين نظام خدمة المتعاملين

by

SALMA HASAN

A dissertation submitted in fulfilment
of the requirements for the degree of
INFORMATION TECHNOLOGY MANAGEMENT

at

The British University in Dubai

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that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.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

Salma Hasan

____

Signature of the student

Abstract

The goal of Customer relationship management in all organizations, regardless of the type of industry and service provided, is to increase customer's satisfaction and achieve retention. Customers are sharing their opinions about products and expectations by multiple communication points, the service centers or via social media platforms. These opinions and feedback shared are valuable data to enlighten organizations about the issues and weakness points requires improvement or development.

The aim of this study is to use text mining and clustering methods to improve customer's complaints system .To this end, the raised research questions to be answered are as follow: Does the generated clusters shows clear patterns that can help to indicate the complaint category? Doses the current complaints subjects matches the complaints contents? Is there a need of creating new complaints Categories or even merging some of the complaints?

The study research question answered through customer's complaints analysis after applying text mining processes and K-means clustering technique. Based on the generated clusters analysis, the results indicated clear patterns that refers to specific complaints category and some clusters had multiple categories in one cluster. Some of the categories patterns are having similarity in keywords so it can merged together and the duplicated can be removed.

The results of the complaints analysis using text mining and clustering techniques will contribute on enhancement of the quality of service provided and weakness points to focus on.

الخلاصة :

الهدف من إدارة علاقات العملاء في جميع المؤسسات باختلاف نوع الصناعة والخدمات المقدمة ،هو زيادة رضا العملاء والحفاظ على ولائهم للخدمة أو المنتج . يشارك العملاء أفكارهم ارائهم حول المنتجات والخدمات سواءً كانت الايجابية او سلبية وتوقعاتهم المستقبلية للمنتج أو الخدمة عبر العديد من قنوات التواصل.يمكن للعميل التواصل بشكل مباشرمع مزود الخدمة عن طريق إجراء مكالمة هاتفية إلى مركز اتصال مراكز الخدمة أو بشكل غير مباشر عبر منصات ومواقع التواصل الاجتماعي. هذه الآراء والملاحظات هي بيانات قيمة لتنوير المنظمات حول رأي العملاء والنقاط السلبية التي يجب العمل على تحسينها لرفع مستوى الخدمة والحفاظ على رضا العميل.

الهدف من هذه الدراسة هو استخدام آليات التنقيب في النص وتحليل المجموعات من أجل تحسين نظام شكاوى العملاء و سيتم الاجابة على عدد من اسئلة البحث التالية من خلال هذة الدراسة :

- هل هناك أنماط واضحة في التجمعات الناتجة عن استحداث المجموعات يمكن أن تخبرنا موضوع الشكوى ؟
- هل مواضييع الشكاوى الحالية مناسبة لمحتوى الشكوى ؟ أم هناك حاجة لإنشاء مواضيع جديدة للشكاوى أو حتى دمج بعض الشكاوى نظرا لتشابه المواضييع والمحتوى؟

ولتحقيق هذه الغاية و الإجابة على الاسئلة البحثية المطروحة بعد تحليل شكاوى العملاء بعد تطبيق عمليات التنقيب عن النص وتقنية المجموعات K-means.حيث أشارت نتائج تحليل المجموعات إلى أنماط واضحة تدل على موضوع معين للشكاوى و أنماط أخرى مماثلة يمكن دمجها أو إزالتها.

ستسهم نتائج تحليل الشكاوى باستخدام تقنيات تنقيب النص والتجميع في تحسين ورفع مستوى و جودة الخدمة المقدمة وذلك بالتركيز على تحسين نقاط الضعف التي ينبغي التركيز عليها حيث ان الاحصائيات الدورية تركز على النسب والارقام فقط ولاتهتم بمحتوى الشكاوى المقدمة التى لاتقل أهمية عن النسب والارقام .

Table Of Contents

Declarations ………………………………………………………II
Abstract………………………………………………………..…III
List of figures ………………………………………………….VII
List of tables ………………………………………..…………VIII

Chapter 1 Introduction……………………………….. Page 1

Aims and Objectives…………………………………..Page 2
Research Questions…………………………………….Page 2
Structure of the Thesis……………………………….… Page 3

Chapter 2. Literature Review…………………………….…Page 3

Data Mining and customer relationship management.….. Page 3
Overview of Data Mining and text mining……….…...… Page 4
K-means clustering Technique…………..……..……. Page 6

Chapter 3 Methodology…………………………………....Page 8

Research Approach and Techniques used……………….. Page 8
Used Tool ……………………………….................. Page 9
Data description and preparation……………….............. Page 9

Chapter 4 Analysis ……………………………….……….. Page 13

Data Processing, execution and results………………..... Page 13

Chapter 5 Conclusion and Future Work……….……... Page 19
Conclusion ……………………………………...…….. Page 19
Future work …………………………………….……... Page 21

References ……………………………………………….. Page 22

List of figures

Figure1	Classification framework for data mining techniques in CRM	…………	Page 9
Figure 2	k-means clustering algorithm workflow	…………	Page 7
Figure 3	research methodology and approach diagram	………….	Page 8
Figure 4	customer’s complaints system work flow	…………	Page 11
Figure 5	Customer’s complaints statistics (2015-2016)	…………	Page 12
Figure 6	Rapid miner text processing	………….	Page 14
Figure 7	Rapid miner clustering process	…………	Page 16

List of Tables

Table 1 Top five Complaints subject …………… Page 12

Table 2 keywords and terms for each complaint subject …………… Page 15

Table 3 clustering results after generating 5 clusters …………… Page 16

Table 4 clustering results after generating 7 clusters …………… Page 18

Table 5 clustering results after generating 10 clusters …………… Page 18

Chapter 1 - Introduction

Customer feedback is becoming an essential factor for customer relationship management and improvement, It reflects customers and users experience either like or unlike of used services and products. Nowadays, users and customers can easily share their feedback via many communication channels such as social media platforms, smart mobile applications and call center operators.

The rapid development of data analysis tools gave opportunity for organizations and service providers to make an efficient processing of customer feedback which was gathered through various communication channels such as emails, social media platforms and complaints systems.

Marketing departments and customer relationship management have awareness of the importance of textual feedback and the benefits of manual or automatic approaches to analyze this information and text which data analysis and scientists refers to it as text mining.

The study aim is to use text mining and cluster analysis to contribute in customer complaints system improvement. Various text mining methods will be applied to process the customer's complaints text along with k-mean clustering technique to find patterns and trends to evaluate the customers experience and provide guidelines of the weakness and areas to improve to provide a better service and increase customer's satisfaction level.

This paper is organized as follows. Chapter 2 provides a topic background and Literature Review of related researches .in chapter 3, the methodology used is described. Chapter 3 shows how the proposed framework is applied on the chosen dataset. In a last Section, the conclusion and future work of this study are summarized.

1.1 Aims and Objectives

In this study we use text mining and cluster analysis methods to improve customers' complaints system by analyzing the customer's complaints and see if we can find clear patterns from the generated clusters that refers to the current complaints subjects or there is an unknown subject that's need to be added to describe and cover the new found complaint subject. The analysis of complaints text will not only discover the new data patterns but it will highlight the issues and service weaknesses accurately from the keywords and terms found in each generated cluster so Customers relationship management unit can draw a plan with the required improvement to provide a better service and increase customers satisfaction.

1.2 Research Questions

The aim of this study is to use text mining and clusters analysis to improve customer complaints system and answer the below research questions:

- Is there a clear patterns in the generated clusters refers to the current complaints subject?
- Is there any unmatched complaint subjects to be renamed or recreated or even merging it with the together due to similarity in patterns?
- Is there any new patterns in the clusters requires creating new complaints subject to cover this new topic?

1.3 Structure of the Thesis

In chapter 1, we introduce the study topic and highlight the aims, objectives and the research question. In Chapter 2 a literature review about text mining and cluster analysis is given. The methodology used to answer the research question along with the data description and preparation will be explained in chapter 3. Chapter 4 will go through data processing and execution using text mining and clustering techniques and the analysis of the generated results. Finally, the conclusion of the thesis and the future work will be described in chapter 5.

Chapter 2- Literature Review

2.1 Data mining and Customer relationship management

Customer relationship management is a widely known concept for business development field. Service providers /or product suppliers are required to have knowledge about their customers characteristics and nature to be able to attract them and increase level of satisfaction. [1]

Many organization collects and stores massive amount of information about their customers, suppliers and potential customers but they are unable to discover valuable information hidden in their data such as customer objective information and customer service or product evaluation. [12] .Data mining role in CRM will give organization ability to discover trends, customer behavior and characteristics to acquire and retain potential customers and increase customer value.

Figure .1 framework of data mining techniques in CRM, the framework was generated based on the review of the literature on data mining techniques in CRM [19]. This framework is also based on the research conducted by Swift (2001), Parvatiyar and Sheth (2001) and Kracklauer et al. (2004). [9]

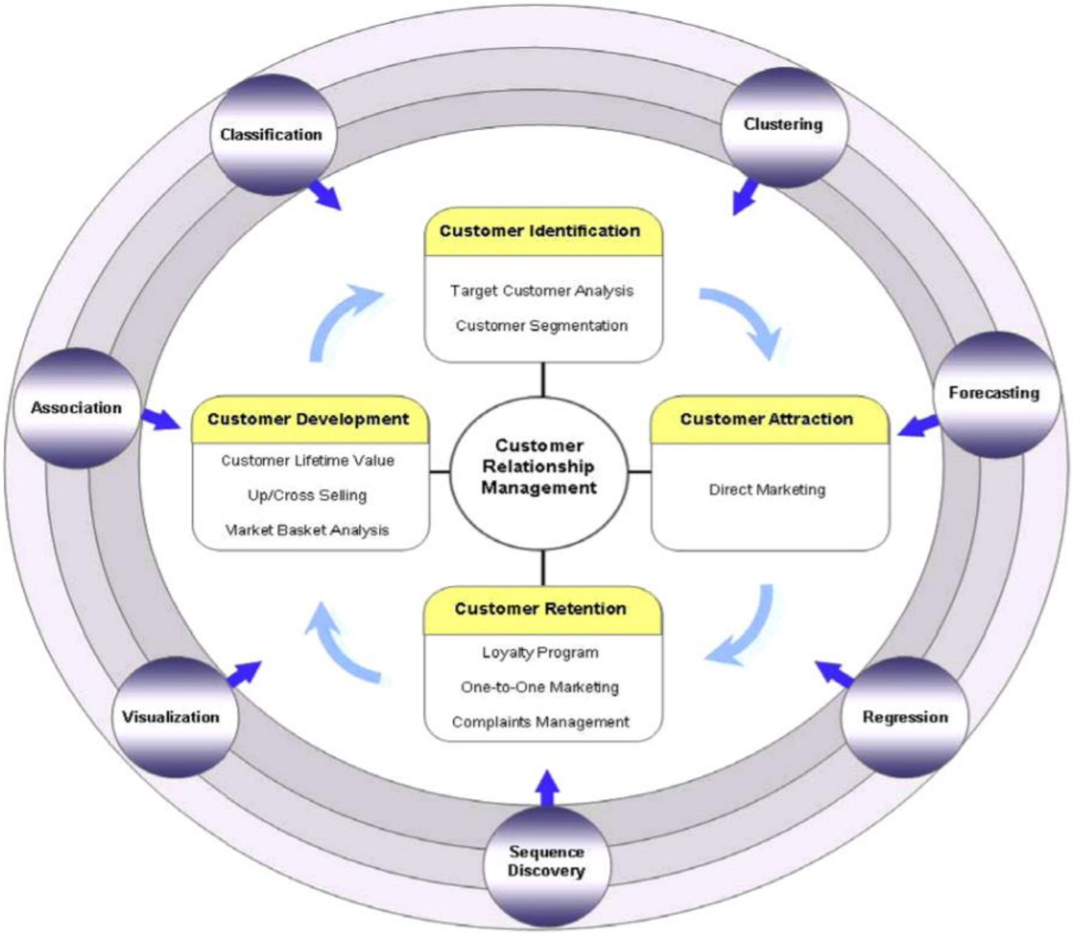

Figure .1 - framework of data mining techniques in CRM

2.2 Overview of Data mining and Text mining

Data Mining

The term “data mining is an identification for coal and gold mining, data mining extracts knowledge buried in databases, or visitors reviews have left on website. Data mining is defined as the process of information extraction and identification to gain knowledge from huge amount of data. [6] This information will help to find / discover patterns and relationships hidden in data [7] Data mining can be also defined as advance data search capable of using statistical algorithms to discover patterns in the data. [12]

Data mining methods made huge contribution in knowledge extraction that was hidden in structures /or unstructured data. [1] The main two objectives of Data mining are prediction and description, these objectives are constructed in six types of models aimed at solving business problems: classification, regression, time series, clustering, association analysis, and sequence discovery [12]. The first two, classification and regression, are used to make predictions, while association and sequence discovery are used to describe behavior. [12]

Data mining algorithms works based on two concepts supervised and unsupervised. The unsupervised concept works in data with predefined patterns and needs to find similar patterns .supervised approaches works with data with no clear patterns or unspecified approaches such as clustering. [4]

Data mining is mainly a section in a bigger process called “knowledge discovery” .it explains steps required to have a meaningful results.[12] Nowadays, in many fields data patterns discovery, trends and similarity is used to contribute in development of business.[4]

Text Mining

It is a difficult and challenging matter to find required knowledge and features in text documents /or data. Text mining contribute in the matter and helps users to find what they are looking for. [14]

The term “text mining “or “text data mining “is a text version of data mining and it is used to describe the information extraction from mountains of document to find interesting information, trends and significant features that might be triggered for useful action and decision making.[9] others defined Text mining acts as another version of what is called data mining and it aims to discover useful patterns and trends from enormous databases. Text mining, also recognized as intelligent analysis of Text.

Text mining also referred to as the detection of interesting knowledge in text documents. [14]

Text mining can add more value to the documents stored in systems/or digital libraries, as Studies indicated that 80% of organizations information is contained in text documents, such as emails, memos, customer reviews and feedback [5].
Many researchers have successfully used text mining techniques to analyze huge amounts of text data in business. [5]

2.3 k-Means Clustering Technique

Clustering concept

The process of grouping similar objects into group is defined as clustering. [1]Another definition of clustering is a representation of number of object grouped based on similarity measures, the similarity between grouped objects is high or similarity low other group.[2] Clustering can be defined also as group /or set of points that is compact and isolated. [2]

The main purpose of clustering is to find the data structure and explore its nature [1] as Grouping together similar objects to each other and dissimilar together. [10] Will help in Cluster analysis and defining statistical classification techniques to discover the individuals or populations groups they fall in. [2]

K-means algorithm

K-Means algorithm was proposed 50 years ago with thousands of other clustering algorithms [2].it is a well-known clustering algorithm that is sensitive to the start point for portioning the items into K clusters (k=number of clusters). [1]

K-means is the simplest unsupervised learning algorithims.it classifies given data into number clusters (assume K clusters).the main idea of the algorithm is choosing the centroid value randomly from the dataset and associate each object to the nearest centroid. [3] .The diagram in figure .x explains the process of K-means algorithm and workflow.[4]

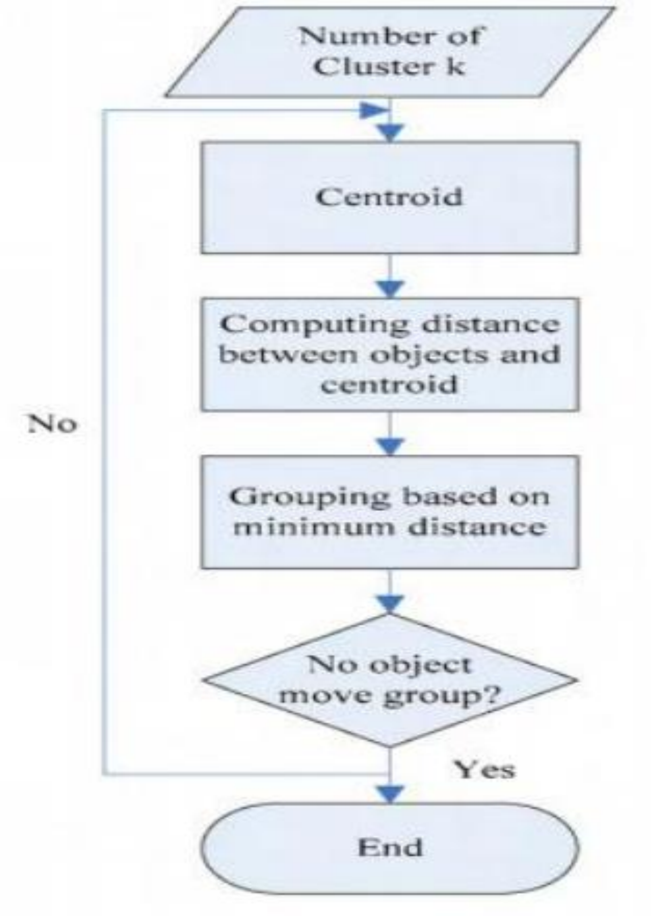
Number of Cluster k
Centroid
Computing distance between objects and centroid
No
Grouping based on minimum distance
No object move group?
Yes
End

Figure. 2 – k-means clustering algorithm workflow

The algorithm inputs K clusters and begins to estimate K initial centroid values to assign each object to the nearest centroid based on the distance between objects and the initial centroids. The process continues for many iterations and the centroids are recomputed until the sum of the distances is minimized, or some maximum number of iterations is reached).

Chapter 3 - Methodology

In this chapter the used methodology and approaches is explained to answer our main research question “how to use text mining and cluster analysis to improve customers’ complaints system?”.

The data set will go through many processes data collection from the customer complaints system, Data filtering and sorting to eliminate any unrequired data. The complaints subject’s key words generation generated using various text mining tools. Then once data is sorted and text is processed then K clusters will be generated .finally, the generated results will be analyzed.

- ## 3.1 Research approach and techniques used

In this study two types of techniques was applied, text processing and clustering technique to see if we can improve customers’ complaints system using text mining and cluster analysis. Figure.1 shows all processes and steps for applying the research approach and techniques.

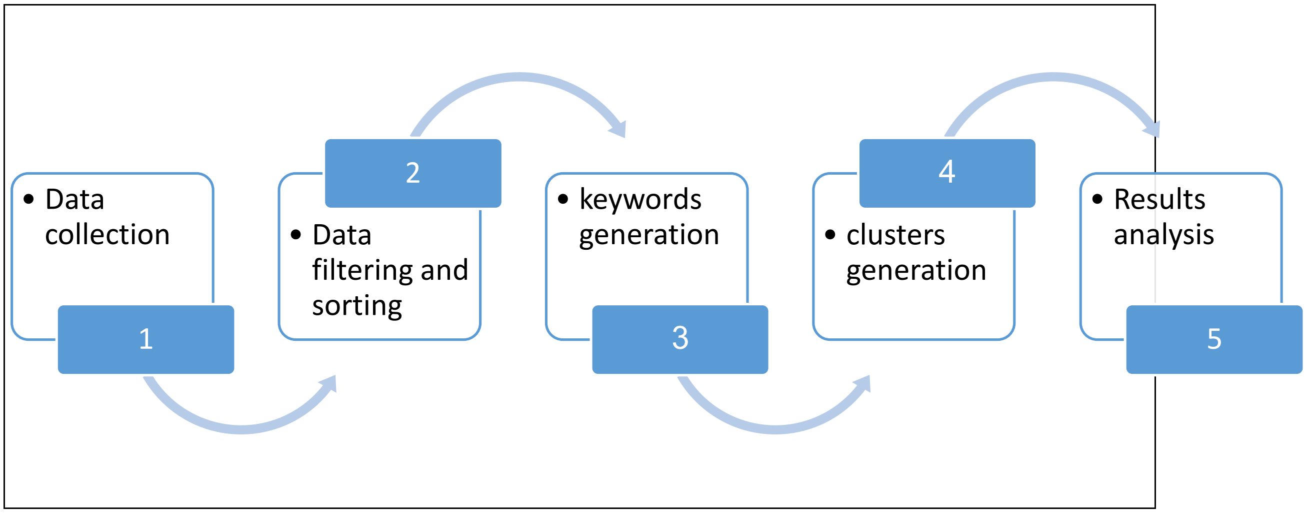

Figure.3 – Research methodology and approach diagram

Step 1, and 2 presents the data collection and preparation process that will be discussed briefly in data description and preparation section, these two steps ensures that the chosen dataset is well prepared and preprocessed by eliminating unrequired data fields and remove any duplication, null values or blank data entries.

Step 3 is considered a millstone step in our methodology as its essential step that cluster analysis relays on. Text processing tools such as tokenization, stemming, n-gram..etc. will be applied and the result of text processing will produce keywords and terms that will guide us to identify on the next stage what complaint subject each cluster represents.

In step 4, the chosen clustering technique K-means was applied to generate K clusters on a range of K values .the results will be compared to identify the best value of K that will show best clusters quality, as the quality of clusters will have impact on the quality of the analysis results .

The final step in applying the research methodology after processing and clusters generation, is the results analysis and it will be discussed briefly in chapter .4.

3.2 Used tool

To apply text mining processes our dataset text and the chosen clustering technique rapid miner software was used. It is one of the data mining tools that allows students or data scientists to analyze and process a chosen dataset to get new and meaningful patterns or results from stuttered or unstructured datasets.

3.3 Data description and preparation

The used dataset is in this study belongs to customer complaints system of a service provider in gulf region that provides community and public services for transportation sector. The service users/customers are from various categories, this variation gives us an interesting combination in dataset as each customer will reflect in his complaint many factors , his/her age, gender ,nationality, educational background, cultural traditions and habits. Majority of service users and customers are from Middle East and gulf region countries, so they report their complaints and remarks in Arabic language.

This data comes from different resources that feeds the customer service system database, Customers/users can get the service by contacting the call center and talk to an agent to report his/her complaint. They can also use mobile app or the official website of the service provider to register their complaints and request for the offered services. The flowchart (figure .1) shows and explains the process of data gathering starting from customer call to the customer service number or using the website and the mobile app until the close of complaint or request raised.

The study targets the customer’s complaints details of the years 2015 and 2016 in Arabic language .the dataset contains 15 fields (complaint subject ,subject type, complaints details, approval date, complaint registrar, driver name, driver ID , plate number ,reply date ,investigation details, company remarks and authority remarks)with16,000 records and above but we are not going to use all the dataset fields/attributes of the dataset ,only two fields will be analyzed , so data was filtered to eliminate the unnecessary fields or attributes (complaint subject and complaint details) .

The Complaints subjects are categorized into 35 subjects to cover the majority of complaints areas but for this study we have chosen the most frequent subjects as per the monthly reports. Figure. (2) and table.(1) Shows 5 subjects that are considered top 5 subjects in the yearly reports of 2015 and 2016.all subjects are linked to drivers penalties list, as shown in figure .1 ,if complaint is proven to be correct after investigation a fine will be issued based on the drivers penalty list .

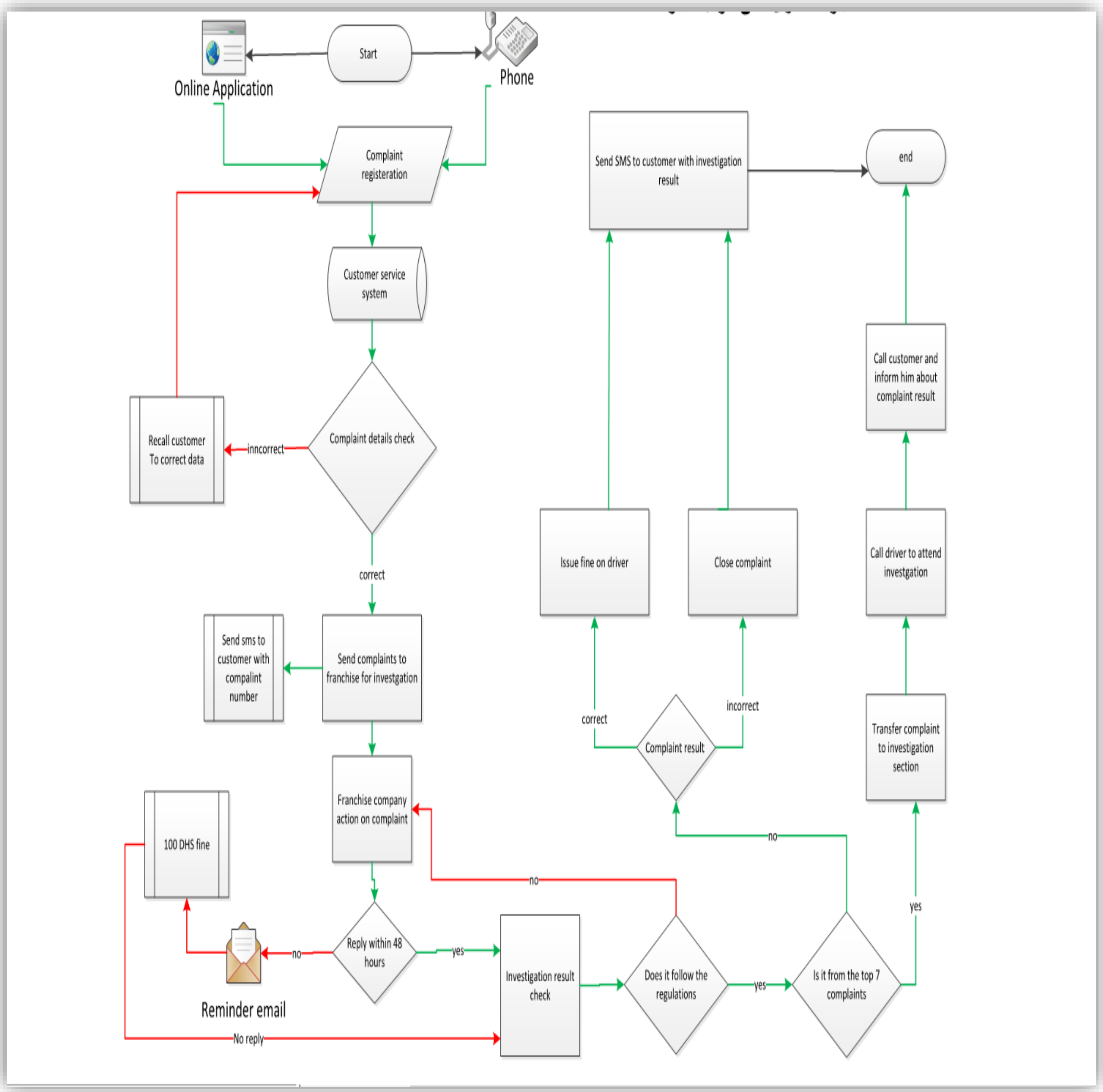
Start
Online Application
Phone
Complaint registeration
Customer service system
Complaint details check
inncorrect
Recall customer To correct data
correct
Send complaints to franchise for investgation
Send sms to customer with compalint number
Franchise company action on complaint
Reply within 48 hours
no
yes
Reminder email
100 DHS fine
No reply
Investigation result check
Does it follow the regulations
no
yes
Is it from the top 7 complaints
no
yes
Complaint result
correct
incorrect
Issue fine on driver
Close complaint
Send SMS to customer with investigation result
end
Transfer complaint to investigation section
Call driver to attend investgation
Call customer and inform him about complaint result

Figure.4 – Customer’s complaints system

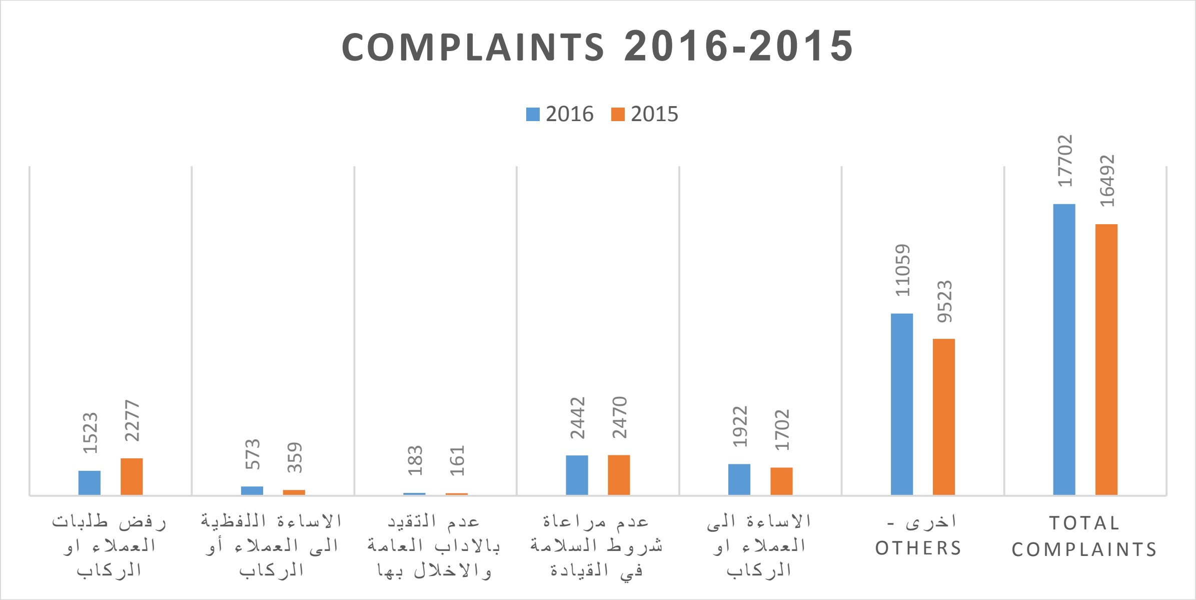

Figure.5 – Customer’s complaints statistics (2015-2016)

Top 5 complaints subjects		
Code	Complaint Subject - English	Complaint Subject - Arabic
C1	refuse of reasonable customers’ requests	رفض طلبات العملاء او الركاب
C2	driving with recklessness	عدم مراعاة شروط السلامة في القيادة
C3	abuse of passengers or customers	الاساءة الى العملاء او الركاب
C4	violation of public morals	عدم التقيد بالاداب العامة والاخلال بها
C5	Verbal abuse by using offensive and improper words and terms	الاساءة اللفظية الى العملاء أو الركاب

Table .1 – Top five Complaints subject

Chapter 4 - Analysis and results

This chapter explains how the research methodology used to answer our study research question and what was applied on our chosen data set ending up with data execution and analysis results and generated from the applying text mining techniques and k-means clustering method.

4.1 Data analysis and execution

This study tries to use text mining and cluster analysis to improve customer complaints system by Appling text mining techniques and clustering methods on the chosen data set to find patterns or results that will help to improve the customer's complaints system. The chosen data set was extracted from customer's complaints system. The type of data is text and the language used in most of the received complaints is Arabic as the service is provided in Gulf countries region. We have 37 complaint subject but we included only 5 topics based on the statistics of the most received complaints subjects or topics as below:

1. refuse of reasonable customers' requests
2. driving with recklessness
3. abuse of passengers or customers
4. violation of public morals
5. Verbal abuse by using offensive and improper words and terms

The assumption made before processing and generating the clusters that every clusters or grouped objects will show me a clear pattern for a specific complaint subject and this assumption is formed in our main research question as below:

- Is there a clear patterns in the generated clusters refers to the current complaints subject?
- Is there any unmatched complaint subjects to be renamed or recreated or even merging it with the together due to similarity in patterns?
- Is there any new patterns in the clusters requires creating new complaints subject to cover this new topic?

As discussed briefly in chapter 3, about the methodology was applied our dataset, the chosen dataset was filtered and cleaned to include only the required fields to be process, The results of this after eliminating some fields will reduce number of attributes and fields and leave us with only two fields of text type, field 1: complaints subject that acts as a label in clusters, and field 2: complaints details that is in type text and eliminate unnecessary or unused data fields.

The Complaints details text was processed by applying many text processing Techniques. As a start, all non-textual type of information will be removed such as punctuation, then stop words also will be detected and removed. Then terms was reduced to their basic stem by using stemming algorithm and last process to be applied in text processing was tokenization as shown in figure .6.

Tokenization process generated words and terms that my dataset contains specifically in the field complaint details. These terms and keywords was generated to be used as a reference to identify what each cluster refer to what complaint subject. Each of the 5 complaints subject's original data was filtered separately and processed on subject wise to identify the keywords which refers to each subject .These keywords generated in (table.2) is our guide   to define what complaint subject the generated clusters refers to.

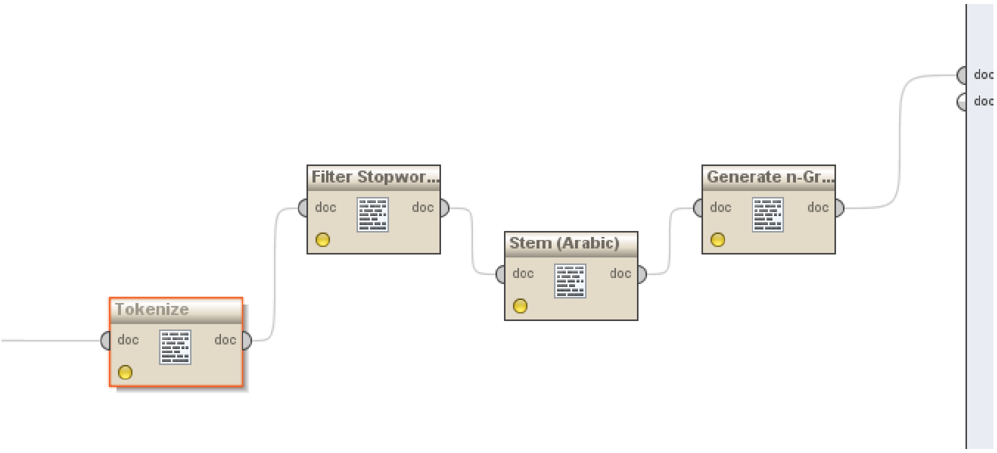

Figure .6 – rapid miner text process

Subject Code	Complaint Subject - English	Complaint Subject - Arabic	Keyword-Arabic	Keyword-English
C1	refuse of reasonable customers' requests	رفض طلبات العملاء او الركاب	رفض	Refused
			رفض ايصاله	He refused to deliver it
			رفض_توصيله	Refused to give a ride
C2	driving with recklessness	عدم مراعاة شروط السلامة في القيادة	متهورة	Reckless
			بسرعة	Fast/speedy
			بحادث	An accident
			بطريقة_متهورة	Reckless way
			بحادث_مروري	accident
			كاد يصطدم	Almost collided
			يشكل خطر	Cause danger
			جنونية	Crazy
			يقود بطيش	He leads with a beast
			بسرعة كبيرة	Very speedy
			تجاوزه بتهور	Overcome it recklessly
			كاد يصطدم	Almost collided
			التفاف مفاجئ	Sudden turn
			بسرعة عالية	High speed
			يتسبب_بالضرر	Causing damage
			يتسبب_بدهس	Causes indecision
C3	abuse of passengers or customers	الاساءة الى العملاء او الركاب	صرخ	Shouted
			بصورة سيئة	Badly
			غير مهذبة	Not polite
			وكاد_يتهجم	And almost attacked
			وقلل_الإحترام	Disrespected them
			غير لائق	Inappropriate
C4	violation of public morals	عدم التقيد بالاداب العامة والاخلال بها	اشار بيده	Inappropriate hand moves
			بذيئة	Expletive
			بنظرات_مخلة	Expletive looks
			يعاكسهم	Flirted them
C5	Verbal abuse by using offensive and improper words and terms	الاساءة اللفظية الى العملاء أو الركاب	بالصراخ	Screaming
			وتلفظ_بألفاظ غير مهذبة	Inappropriate words

Table .2 - keywords and terms for each complaint subject

Clustering is a widely used technique, clustering can be used for many purposes such as underlying structure to gain more insight in data and discover salient features. Also it can be useful for classification to identify level of similarity and compare data summarized in clusters prototypes. In this study K-means clustering technique was applied. K-means work process as a clustering algorithm is to group similar objects based on measures and characteristics of similarity.

K clusters (K = number of clusters) was be generated and the number of clusters is (5=>K<=10) compare results and see if we will find different results in case number of clusters changes. The initial start for clusters generation, K= 5, five clusters were generated and analyzed using the keywords in table.2. each clusters pattern was analyzed as results shows in table .3, Clear patterns were showing in the generated clusters and each pattern with occurrence of specific words and terms refers to a complaints subject .

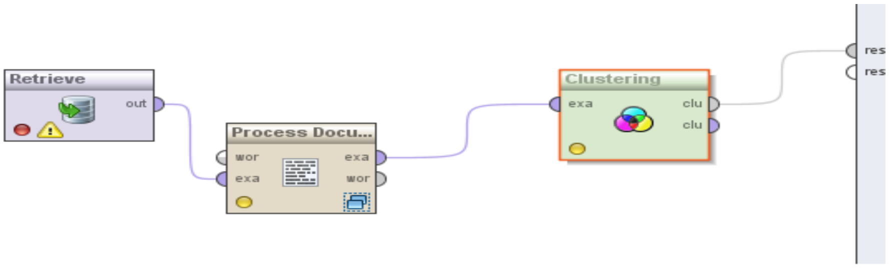

Figure .7 – rapid miner clustering process

cluster	Complaint subject
cluster 1	(C2) driving with recklessness
cluster 2	(C1) refuse of reasonable customers' requests
cluster 3	(C1) refuse of reasonable customers' requests
cluster 4	(C2) driving with recklessness
	(C1) refuse of reasonable customers' requests
cluster 5	(C3) abuse of passengers or customers
	(C5) Verbal abuse by using offensive and improper words and terms
	(C4) violation of public morals

Table.3 - clustering results after generating 5 clusters

each clusters pattern was analyzed as results shows in table .3, Clear patterns were showing in the generated clusters and each pattern with occurrence of specific words and terms refers to a complaints subject .

As demonstrated in the table .3 some clusters are showing clearly the subject refers, as Cluster 1, 2 and 3 to and other clusters are having a combination of two and more different compliant subjects due to similarity in the keywords .cluster 4 refers to two complaints subjects, first one is (C2) driving with recklessness and second one is (C1) refuse of reasonable customers' requests. Also cluster 5 shows a combination of three subjects, (C3) abuse of passengers or customers, (C5) Verbal abuse by using offensive and improper words and terms and (C4) violation of public morals.

Many iteration of clustering using K-means was generated with increasing the value of K (K-number of clusters). The second of clustering process K=7, the generated 7 clusters were analyzed using the same method applied in the 5 clusters analysis by referring back to the keywords and terms table .Table.4 to identify the clusters patterns and what complaint subject each cluster refers to. as shown In table.4, clear patters was discovered as in cluster 2 that shows clearly from the grouped words subject (C2) driving with recklessness.in the other hand, many clusters are sharing several subjects such as clusters 1 and 7, they got a combination of the subjects (C3) abuse of passengers, (C5) Verbal abuse by using offensive and improper words and terms and (C4) violation of public morals or customers. C3, C4, and C5 shares the same keywords and from the subject title we can see that they fall in the same group.
The third iteration of clusters generating process was applied to generate 10 clusters (K=10) and the results of clusters analysis as in table .5 shows similarity in behavior in all iterations of cluster generation the results are so similar to the 1st and 2nd round of clusters generating a number of 5 and 7 clusters.

Cluster	Complaint subject
cluster 1	(C3) abuse of passengers or customers
	(C5) Verbal abuse by using offensive and improper words and terms
	(C4) violation of public morals
cluster 2	(C2) driving with recklessness
cluster 3	(C3) abuse of passengers or customers
cluster 4	(C2) driving with recklessness
cluster 5	(C1) refuse of reasonable customers' requests
cluster 6	(C2) driving with recklessness
cluster 7	(C3) abuse of passengers or customers
	(C5) Verbal abuse by using offensive and improper words and terms
	(C4) violation of public morals

Table.4 - clustering results after generating 7 clusters

Cluster	Complaint Subject
cluster 1	(C2) driving with recklessness
cluster 2	(C1) refuse of reasonable customers' requests
	(C3) abuse of passengers or customers
cluster 3	(C1) refuse of reasonable customers' requests
cluster 4	(C2) driving with recklessness
cluster 5	(C5) Verbal abuse by using offensive and improper words and terms
	(C3) abuse of passengers or customers
cluster 6	(C1) refuse of reasonable customers' requests
cluster 7	(C2) driving with recklessness
cluster 8	(C1) refuse of reasonable customers' requests
cluster 9	(C3) abuse of passengers or customers
cluster 10	(C1) refuse of reasonable customers' requests
	(C5) Verbal abuse by using offensive and improper words and terms
	(C3) abuse of passengers or customers
	(C2) driving with recklessness

Table.5- clustering results after generating 10 clusters

Chapter 5 - Conclusion and Future Work

5.1 Conclusion:

Organizations data bases and documents contains hidden knowledge that is not discovered using the routine and regular statistical reports. Data mining and text mining highly contributed in data analysis and discovery of hidden trends and patterns that can be helpful to enhance the work process and service level.

The study aimed to improve customer complaint system by using text mining and cluster analysis. Interesting patterns were discovered From complaints text processing and the cluster analysis of the generated clusters for different values of K 5,k=7,k=10, it was concluded that no much changes observed in clusters objects grouping as the 3 iterations in clusters generations using k-means.

The three cluster generation iterations showed clear patterns for some Clusters and other clusters contained a combination of more than one pattern in the cluster due to similarity in keywords and terms that refers to subjects.

It was also conclude from the results that the below three subjects contains almost the same key words in all clusters:

- Abuse of passengers or customers.
- Violation of public morals.
- Verbal abuse by using offensive and improper words and terms.

Foe this case, it’s suggested to combine these topics together with one subject since they got the same clusters patterns.

The text mining and clusters analysis of the customer’s complaints answers our research questions, as there is a clear pattern for some of the generated clusters, also there are 3 complaints subjects required to be merged as one subject.

5.2 Future Work:

The process of choosing the complaints subjects which us now done by call taker or call center agent who spends time to decide what complaint subject he/or she is receiving ,so there are chances for human errors and wrong decisions .

As future work can be done to enhance the process of customers complaints system subject selection , we can systemize this process by building a prediction system that uses the discovered clustering patterns along with the terms and key words to identify the complaint subject with accurate match between complaint content and subject .

References

[1] Hosseini, S.M.S., Maleki, A. and Gholamian, M.R., 2010. Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. Expert Systems with Applications, 37(7), pp.5259-5264.

[2] Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), pp.651-666.

[3] Ferdous, R., 2009, November. An efficient k-means algorithm integrated with Jaccard distance measure for document clustering. In Internet, 2009. AH-ICI 2009. First Asian Himalayas International Conference on (pp. 1-6). IEEE.

[4] Vishwakarma, S., Nair, P.S. and Rao, D.S., 2017. A Comparative Study of K-means and K-medoid Clustering for Social Media Text Mining. INTERNATIONAL JOURNAL, 2(11).

[5] He, W., Zha, S. and Li, L., 2013. Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33(3), pp.464-472.

[6] Ngai, E.W., Xiu, L. and Chau, D.C., 2009. Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), pp.2592-2602.

[7] Rygielski, C., Wang, J.C. and Yen, D.C., 2002. Data mining techniques for customer relationship management. Technology in society, 24(4), pp.483-502.

[8] Tan, A.H., 1999, April. Text mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases (Vol. 8, pp. 65-70). Sn

[9] Nasukawa, T. and Nagano, T., 2001. Text analysis and knowledge mining system. IBM systems journal, 40(4), pp.967-984.

[10] Silwattananusarn, T. and Tuamsuk, K., 2012. Data mining and its applications for knowledge management: a literature review from 2007 to 2012. arXiv preprint arXiv:1210.2872.

[11]Tan, A.H., 1999, April. Text mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases (Vol. 8, pp. 65-70). sn.

[12] Rygielski, C., Wang, J.C. and Yen, D.C., 2002. Data mining techniques for customer relationship management. Technology in society, 24(4), pp.483-502.9

[13] Beil, F., Ester, M. and Xu, X., 2002, July. Frequent term-based text clustering. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 436-442). ACM.

[14] Zhong, N., Li, Y. and Wu, S.T., 2012. Effective pattern discovery for text mining. IEEE transactions on knowledge and data engineering, 24(1), pp.30-44.

[15] Gupta, V. and Lehal, G.S., 2009. A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence, 1(1), pp.60-76.

[16] Steinbach, M., Karypis, G. and Kumar, V., 2000, August. A comparison of document clustering techniques. In KDD workshop on text mining (Vol. 400, No. 1, pp. 525-526).

[17] Fung, B.C., Wang, K. and Ester, M., 2003, May. Hierarchical document clustering using frequent itemsets. In Proceedings of the 2003 SIAM International Conference on Data Mining (pp. 59-70). Society for Industrial and Applied Mathematics.

[18] Chen, M.S., Han, J. and Yu, P.S., 1996. Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and data Engineering, 8(6), pp.866-883.

[19] Ngai, E.W., Xiu, L. and Chau, D.C., 2009. Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), pp.2592-2602.

[20] Chen, I.J. and Popovich, K., 2003. Understanding customer relationship management (CRM) People, process and technology. Business process management journal, 9(5), pp.672-688.

[21] Shaw, M.J., Subramaniam, C., Tan, G.W. and Welge, M.E., 2001. Knowledge management and data mining for marketing. Decision support systems, 31(1), pp.127-137.

[22] Gupta, V. and Lehal, G.S., 2009. A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence, 1(1), pp.60-76.

[23] Berkhin, P., 2006. A survey of clustering data mining techniques. In grouping multidimensional data (pp. 25-71). Springer Berlin Heidelberg

[24] Hui, S.C. and Jha, G., 2000. Data mining for customer service support. Information & Management, 38(1), pp.1-13.

[25] Ngai, E.W., Xiu, L. and Chau, D.C., 2009. Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), pp.2592-2602

[26] Aggarwal, C.C. and Wang, H., 2011. Text mining in social networks. In Social network data analytics (pp. 353-378). Springer, Boston, MA.