I.J. Information Technology and Computer Science, **2020, 6, 1-10**
Published Online December 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2020.06.01

Applying Clustering and Topic Modeling to Automatic Analysis of Citizens' Comments in E-Government

Gunay Y. Iskandarli
Institute of Information Technology, Azerbaijan National Academy of Sciences
Address: AZ1141, B. Vahabzade street, 9A, Baku, Azerbaijan
E-mail: gunayniftali@gmail.com

Received: 11 March 2020; Accepted: 25 June 2020; Published: 08 December 2020

Abstract: The paper proposes an approach to analyze citizens' comments in e-government using topic modeling and clustering algorithms. The main purpose of the proposed approach is to determine what topics are the citizens' commentaries about written in the e-government environment and to improve the quality of e-services. One of the methods used to determine this is topic modeling methods. In the proposed approach, first citizens' comments are clustered and then the topics are extracted from each cluster. Thus, we can determine which topics are discussed by citizens. However, in the usage of clustering and topic modeling methods appear some problems. These problems include the size of the vectors and the collection of semantically related of documents in different clusters. Considering this, the semantic similarity of words is used in the approach to reduce measure. Therefore, we only save one of the words that are semantically similar to each other and throw the others away. So, the size of the vector is reduced. Then the documents are clustered and topics are extracted from each cluster. The proposed method can significantly reduce the size of a large set of documents, save time spent on the analysis of this data, and improve the quality of clustering and LDA algorithm.

Index Terms: E-government, text mining, topic modeling, K-Means.

1. Introduction

Currently, e-government is one of the most important social platforms that are in use. Today, any government's aim is to build firm cooperation with its citizens by involving them in the decision-making process. E-government platform has been established to provide the availability of public services and facilitate the delivery of information to citizens [1].

Basing on public interest, E-government services should increase the efficiency of the services, reduce the costs and improve the quality of the services considering their needs as much as possible. The proposed e-services should be based on a demand-driven principle. To achieve this goal, government agencies should be interactively informed about the citizens' information needs and to study the real needs of the public. For this purpose, citizen's (user's) comments may be used. However, such user comments can lead to the creation of big textual information, which makes it difficult to analyze them in terms of their complexity and being unstructured. The advanced computational and analytical tools such as text mining is intended to be used for text mining and detection of correlations [2]. Topic modeling methods have gained popularity to determine which topics are discussed by citizens recently. Topic modeling algorithms are statistical methods that detect the topics by analyzing words in texts [3]. The methods used in machine learning and text mining have been implemented successfully in identifying hidden topics in most documents. These algorithms are described as a "soft" (fuzzy) classification of documents, especially in the classification problem, that is, the document does not belong to one class, but to several classes with different affiliation degrees. The results of these models can also be used to include the document in only one cluster.

Considering this, in the paper, an approach has been proposed to analyze citizen's comments on services in the e-government environment using topic modeling and clustering algorithms. Through the proposed approach, it is possible to find the priorities of services that reflect the needs of citizens, to determine what people think about the use of e-government services, and so on. The main goal here is to solve the problem by reducing the size of the data and saving time and memory. The article is structured as follows. A summary of the related works is presented in the second section. The third section gives information on topic modeling methods and their application fields. The steps of the proposed method are presented in section four, and section five gives information on the experiments and results, and

the sixth section represents the conclusion and future investigations.

2. Related Work

As mentioned above, the interest in modeling methods is increasing. In this regard, research in this field has also increased. Thus, several methods and approaches have been proposed to develop topic modeling methods. In this section, several studies in the field of topic modeling have been analyzed.

For example, researches in [4] a method based on topic modeling to analyze texts and the description of objects in Wikipedia and the relationships between them is proposed. In [5], a model has been proposed to identify the clusters between individuals and determine the topics among relevant documents (texts, articles) belonged to them. [6] presents a model based on the most commonly used Latent Dirichlet Allocation (LDA) of topic modeling algorithms. The author-topic model has been developed in [7]. This study proposes a model that identifies the authors and their appropriate topic distributions. The result of the experiment carried out in this study shows that LDA gives a good result when only a few words are used in test documents. In the researches, conducted in [8], the LDA model has been developed and its usage in the micro-blog environment has been studied. In [9], the modeling problem of short texts through LDA has been investigated. In [10], the Probabilistic Latent Semantic Analysis and LDA models have been studied for the clustering of documents. In this study, topic modeling is used to determine the number of topics in order to evaluate the characteristics of documents. Similarly, in [11], models such as Correlated Topic Models, Hierarchical Latent Dirichlet Allocation, and Hierarchical Dirichlet Process have been used for document clustering. Here, scientific articles from different fields have been collected and are the articles from the same field clustered or not, have been analyzed. [12] proposed a method to improve the quality of clustering using the topic and fusion models. In the proposed method, the topic modeling algorithm is first applied to the sets of documents several times, and in each iteration, specific topics are determined for each document. Then, the specific topics obtained in each iteration are combined, the document is represented as a single vector, and finally, the documents are clustered. In the approach proposed in [13], the clustering method is first applied to a set of documents, then the topics are extracted from each cluster. The method consists of four steps. In the first step, similar documents are grouped by the clustering method. In the second step, the algorithm LDA is applied to each cluster to identify cluster topics. In the third step terms with global frequency are derived from extracted topic terms and semantically similar words are identified using WordNet. In the last step, sentences containing these terms and words similar to them are selected and summarized.

As can be seen, a lot of work has been done to extract topics from a large number of documents in different environments through the LDA algorithm. In these studies, the LDA algorithm was expanded and its modified models were used to cluster documents. In the literature, very little attention has been paid to the application of this model in the e-government environment. Considering this, we have focused in this paper on the application of topic modeling and clustering algorithms in e-government. In the paper, the approach has been proposed to extract key topics from citizens' comments in e-government by analyzing them using topic modeling and clustering algorithms. The main goal is to increase the effectiveness of the clustering algorithm and the accuracy of the topic modeling algorithm, reducing the measure of the vectors (using semantic similarity between words). Detailed information on the approach is given in Section Four.

3. Topic Modeling Methods and its Application Fields

In recent years the rapid increase in the number of e-documents requires the use of new methods and tools for their management, search, etc. One of the models created for this purpose is topic modeling algorithms. Topic modeling algorithms is a generative model for documents [14, 15]. Thus, each document is treated as a mixture of topics, and each topic is defined as a probability distribution of the words over them. Several models, for example, Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation, and etc. topic models have been successfully applied in the detection of topics from documents (in the development of classification accuracy) [16,17]. These models are described below.

3.1. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is one of the methods used in Natural Language Processing. In the past, LSA has been called Latent Semantic Indexing but then it has been improved for information search. The documents relevant to the query are selected from sets of documents through this method. The usage of semantically similar words in related segments of the text is considered in this algorithm. Here, a matrix is formed from the big text segments based on the number of words, and the rows of the matrix represent the words and the columns represent the paragraphs. Using a mathematical method called singular value decomposition, the number of rows is reduced keeping the relationship between the columns. Then, the paragraphs are compared with the calculation of the cosine of the angle between the vectors created by both columns. If the evaluation values are close to 1, the paragraphs are considered similar otherwise, they are different. The main application fields of LSA are: the detection of relationship between terms, analysis of word combinations in text sets, search of relevant documents (information retrieval), etc.

3.2. Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (pLSA) is an approach created to reduce some disadvantages of the LSA method. Jan Puzicha and Thomas Hofmann introduced it in 1999 [18]. PLSA is a method produced to automate indexing of documents based on a latent class model. The main purpose of pLSA is to distinguish between different contexts of the word without using of the dictionary or thesaurus. It has two important impacts. First, this method allows you to differentiate polysemantic words. Second, by grouping words with common content defines typical similarities. PLSA has been successfully applied in such areas as image retrieval, automatic question recommendation, etc. [19].

3.3. Latent Dirichlet Allocation

The main purpose of creating the Latent *Dirichlet Allocation* (LDA) is to develop the pLSA and LSA methods. LDA is a text mining algorithm based on a statistical Bayesian topic model. LDA is considered to be one of the standard tools in topic modeling and is one of the most common algorithms for topic modeling [20]. Note that, here only the number of topics is known beforehand, and two main principles are expected: 1) Each document is a mixture of several latent topics; 2) Each topic is a mixture of several words.

The relation between the document and the topic is determined by *Dirichlet distribution* and the relation between the topic and the word by polynomial *distribution.* The generative process of LDA is illustrated in Fig. 1 [21]:

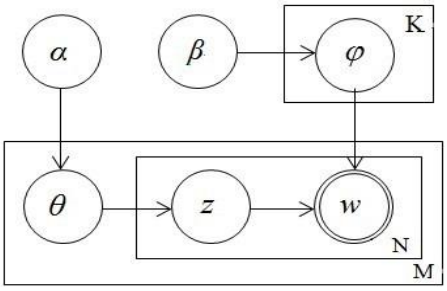

Fig.1. Generative process of LDA

Where, M – is the number of documents, K – is the number of latent topics, N – is the number of words in the document. z – is the topic of the j -th word in the i -th document , w – is the word observed in the document, α, β – are the parameters, α – defines the relative power of latent topics in the sets of documents, β – is a probability distribution of all latent topics. θ – is the topic probability distribution for the current document, φ – word distribution on the current latent topic. The rectangle presents the selection process, the circle shows latent variables, and the binary circles are the obvious variables. The formula for this model is given below:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \quad (1)$$

The main steps of LDA are as follows:

1) For i -th latent topic, the polynomial distribution φ is calculated according to *Dirichlet distribution*;
2) The number of words (N) in the document is obtained according to the Puasson distribution;
3) For each text, topic probability distribution θ is calculated;
4) In the set of documents, the following is considered for each word involved in the document:
a. latent topic z is randomly choosen from probability distribution on topic θ ;
b. the word is randomly choosen from the polynomial distribution of topic z .

The aplication fields of LDA are the following: determination of Emotion topic or topic-based sentiment analysis, Automatic essay grading, Anti-Phishing, etc. [22].

4. The Application of Topic Modeling in E-government

We need regular user-focused evaluations for increasing the accessibility and efficiency of e-government services. With this evaluation, we can improve the quality of resources and services of e-government sites. Here, we can touch on a few issues:

- What services do citizens need most?
- How should we determine the priority of services that reflect the citizens' requirements?
- What do people think about the use of e-government services? etc.

As is known, citizens can show their attitude to any service by commenting on in the e-government environment. Analyzing these comments, it is possible to identify the main issues annoying them [23, 24]. It is known that as the number of comments increases, it becomes more difficult to analyze them. The LDA algorithm is currently being used successfully to extract key topics from a large number of documents. Through this algorithm, it is possible to identify the main concerns of citizens in the e-government environment. Considering this, we propose an approach using LDA and clustering algorithms to determine what topics they are concerned about or interested in through the analysis of citizen comments in the e-government environment. The main steps of the proposed approach are illustrated in Fig. 2.

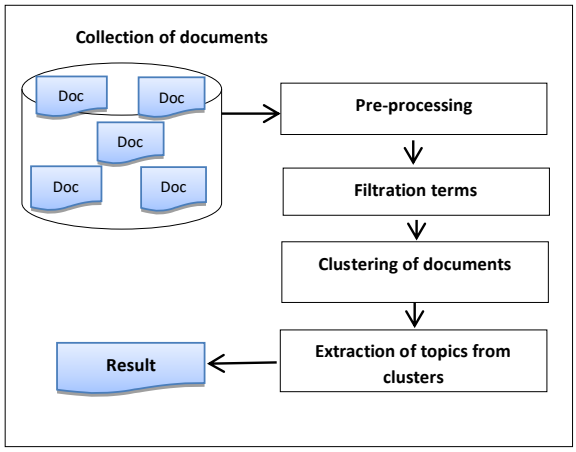

Fig.2. Methodology of the proposed approach

These steps are described in detail below:

Step 1. Firstly, user comments are collected in the e-government environment. For simplicity, these comments are treated as documents and are signified as follows:

$$D = \{d_1, d_2, ..., d_n\} \tag{2}$$

Where n - is the number of documents (comments).

Step 2. The collected comments are pre-processed. In pre-processing, common words, figures and punctuation marks are extracted from the documents. Each word is converted to its original (root) form as they take affixes in their different forms.

Step 3. The terms are extracted from the comments. Then, the sets of documents are described using the TF-IDF[25].

$$d_i = \{w_{i1}, w_{i2}, ..., w_{im}\}, \qquad i = 1,2,..., n \tag{3}$$

Where w_{ij} – is the TF-IDF weight of j -th term in the i -th document. w_{ij} weight is calculated according TF-IDF scheme as follow:

$$w_{ij} = tf_{ij} \times \log(\frac{n}{n_j}) \tag{4}$$

where tf_{ij} – is the term frequency of j -th in the i -th document, n_j – is a number of documents where term j -th is observed.

Euclidean distance is used to calculate the distance between documents. Then, the similarity between the vectors $d_i = \{w_{i1}, w_{i2}, ..., w_{im}\}$ and $d_l = \{w_{l1}, w_{l2}, ..., w_{lm}\}$ is calculated as follows:

$$dist(d_i, d_l) = \sqrt{\sum_{j=1}^{m} (w_{ij} - w_{lj})^2} \quad i, l = 1,..., n \tag{5}$$

Known that, the number of terms in the set of documents is too high and this number is greater than the number of terms found in a single document. Then, most vector elements represented by the TF-IDF of documents will be "0". In other words, vectors will be sparse. This creates two important problems in document clustering:

1) "Cursed" measurement problem;
2) Quality of clustering.

Sparse terms are pre-removed from the vector to overcome these problems. After removing the sparse terms, another factor emerges that affects to the problems represented above. The reason for this is the existence of synonyms in the documents. If the set of documents contains a lot of synonyms, then documents with similar contents may fall into different groups in clustering. This leads to a decreasing in quality of clustering. To overcome such situations, it is suggested to find and extract semantic similar words from the sets of documents. The usage of extended sets of synonyms of each term is suggested to find the semantic similarity of words. For this purpose, we find the set of synonyms of each term using the WordNet and they are signified by $t_i \to synset(t_i)$. Note that WordNet is a network that provides you to determine semantic relationships between words. For example, synonyms, hypernyms, hyponyms, etc. can be easily detected via this network [26, 27].

After finding sets of the extended synonyms of each term, the semantic similarity between words is calculated using the following metric: [28]:

$$sim(t_g, t_s) = \frac{2 * \left| synset(t_g) \cap synset(t_s) \right|}{\left| synset(t_g) \cup synset(t_s) \right|} \geq \alpha$$
$$g, s = 1,2,..., m \tag{6}$$

where $\left| synset(t) \right|$ – is the number of synonyms of the word t, $0 \leq \alpha \leq 1$ – is a managed parameter. If the similarity between words is greater than α, these words are considered as an unique term. Thus, only one of these words is maintained and the others are omitted. So, we reduce the measure of the vector d. In this case, the vector d_i is transformed into the following vector:

$$d_i \to d_i^* = \{\overline{w}_{i1}, \overline{w}_{i2}, ..., \overline{w}_{i0}\}, \quad m_0 \leq m \tag{7}$$

where $\overline{w}_{ij}$ – is TF-IDF weight of j-th word in the i-th document after removing the synonyms.

Step 4. Documents are clustered after displaying as vectors. Various methods exist for document clustering. In this paper, we propose the use of the K-means method for document clustering [29]. K-means is considered to be one of the most popular algorithms in the analysis of big data due to its low-performance time and ease of use. According to this algorithm, clusters are defined by the formula (8):

$$E = \sum_{i=1}^{k} \sum_{d \in C_i} \left| d - O_i \right|^2 \to \min \tag{8}$$

where k – is the number of clusters, C_i – is a i-th cluster, O_i – is the center of i-th cluster.

Step 5. After the clustering of documents, we can find topics for each cluster. For this purpose, the usage of LDA is proposed. The previous sections provide detailed information about LDA. The extraction of main topics from the documents for each cluster via LDA is implemented as follows.

The LDA algorithm is applied to each cluster separately and the most commonly used terms are identified in the documents in the cluster. When we look at these terms, we see that each of them is related to a specific topic. By analyzing the content of these terms, we can identify the topics in the cluster. Note that depending on the accuracy of the clustering, the same topics can be found in different clusters.

Let's assume that clusters $\{C_1, C_2, ..., C_k\}$ are selected. The LDA algorithm is applied to each cluster and for each C_q cluster, $T_q = \{T_{q1}, T_{q2}, ..., T_{qs}\}$ topics are assigned. Where s – is the number of topics. The process is schematically illustrated in Fig. 3.

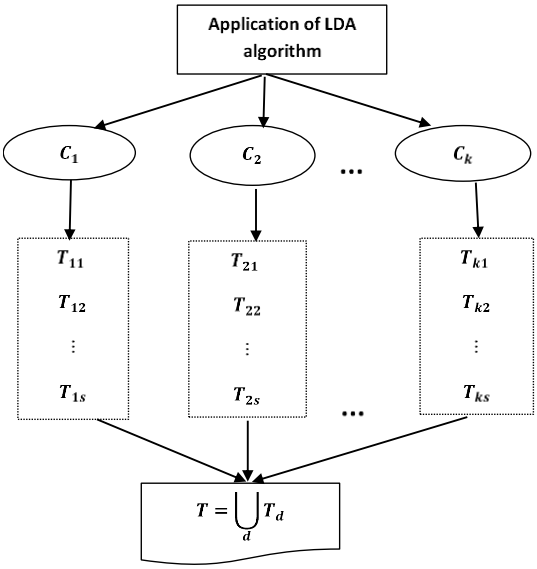

Fig.3. Extraction of topics from the clusters

Thus, we identify the main topic of citizens' comments. Experiments on the proposed method and their results are presented in Section 5.

5. Experiments and Results

This chapter presents the experiments and their results based on the proposed method. The experiments were conducted in the R programming language. The BBC News database was used for the experiment. This dataset contains 2225 documents collected from the BBC news website in five associated areas: Business, Entertainment, Politics, Sports and Technology covering 2004-2005 [30]. In the experiment, a large number of documents from Business, Entertainment, and Sports were collected and analyzed. The criterion of "purity coefficient" was used to evaluate the clustering results. The purity coefficient is a simple and transparent evaluation criterion related to the concept of entropy. According to this criterion, the purity coefficient of the C_p cluster is defined as follow [31]:

$$purity(C_p) = \frac{1}{|C_p|} \max_{p^+=1,\dots,k^+} |C_p \cap C_{p^+}^+|, \quad p = 1,2,\dots,k, \quad p^+ = 1,2,\dots,k^+ \tag{9}$$

Where k^+ – is the number of classes, $C^+ = (C_1^+,\dots,C_{p^+}^+)$ – is the sets of classes, k – is the number of clusters.

Note that each cluster can contain documents from different classes. The purity coefficient shows the ratio of the dominant class measure to the measure of the cluster. The purity coefficient always gets the value from $\left[\frac{1}{k^+},1\right]$ interval. The high value of this coefficient shows that the cluster is a "purity" sub-class of the dominant class. The purity coefficient of the sets of clusters is accepted as the sum of the coefficients of the different clusters:

$$purity(C) = \sum_{p=1}^{k} \frac{|C_p|}{n} purity(C_p) = \frac{1}{n} \sum_{p=1}^{k} \max_{p^+=1,\dots,k^+} |C_p \cap C_{p^+}^+| \tag{10}$$

where n – is the number of all documents. The high value of the purity coefficient indicates the high quality of clustering. The main purpose of the experiment is to evaluate the effectiveness of the proposed method. The expected result of the experiment is to reduce the size of the vector, to save time and to determine the subject of each cluster.

Pre-processing is one of the key steps in text mining. Considering this, the documents collected during the experiment were pre-processed. In pre-processing, punctuation marks, figures, symbols, common words were removed

from the sets of documents, and they were represented in vector form using the TF-IDF scheme. Then, the documents were cleared from sparse terms. The number of words remaining in the sets of documents before and after the pre-processing is described in Table 1.

Table 1. Number of documents and related words

Number of documents	Number of words	
	Before the pre-processing	After the pre-processing
100	8040	4421
300	18356	8851
500	25490	11766
800	33410	14750
1000	36346	15860

After pre-processing, the proposed method has been applied to the sets of documents. The semantic similarity between words was calculated at different values (0.1, 0.2, 0.3, 0.4, 0.5) of α. The number of terms remaining in the sets of documents after the method is described in Table 2. In the version where the number of documents is 100, we observe that more semantic similar words were found at $\alpha = 0.1$ value and the vector measure decreased significantly (26.42%)compared to the remaining words after removing the sparse terms. Since the α –th value is increased fewer words were omitted. Thus, at the value of $\alpha = 0.5$, we notice that the vector measure gets more less (1.29%). As the number of documents increases, the number of semantically similar words also increases accordingly, and the vector measure significantly decreases. For example, if we consider the number of documents with 800, we observe that the vector measure decreases by 31.67% at $\alpha = 0.1$ value.

Table 2. The number of words remaining in the documents after the removal of sparse and semantically similar words

Number of documents	Number of words					
	After removing the sparse terms	After the removal of semantically similar words (α)				
		0.1	0.2	0.3	0.4	0.5
100	772	568 (26.42%)	691 (10.49%)	733 (5.05%)	756 (2.07%)	762 (1.29%)
300	878	633 (27.90%)	777 (11.50%)	838 (4.56%)	856 (2.51%)	865 (1.48%)
500	827	589 (28.77%)	725 (12.33%)	789 (4.59%)	809 (2.17%)	816 (1.33%)
800	821	561 (31.67%)	716 (12.79%)	778 (5.23%)	802 (2.31%)	810 (1.34%)
1000	801	561 (29.96%)	702 (12.36%)	765 (4.49%)	784 (2.12%)	791 (1.25%)

Then, the K-means clustering method was applied to sets of documents, and the clustering accuracy is illustrated in Table 3. Note that the value of $\alpha = 0$ indicates that sets of the remaining documents after removing the sparse terms. As seen from the table, the removal of semantically similar words did not negatively affect the quality of the clustering, but rather, the purity coefficient got sufficiently high value. This also shows the high quality of clustering. As seen from the table, the purity coefficient gets significantly high value as the number of documents increases. Thus, if we consider the case $\alpha = 0.1$, the purity coefficient is 0.88 when the number of documents is 100, and when the number of documents is 1000, The purity coefficient increases to 0.97.

Table 3. Evaluation of the purity coefficient of clustering at different values of α

Number of documents	Purity $\alpha =$					
	0.0	0.1	0.2	0.3	0.4	0.5
100	0.950	0.880	0.810	0.820	0.820	0.880
300	0.996	0.980	0.980	0.980	0.980	0.980
500	0.996	0.930	0.940	0.990	0.990	0.990
800	0.980	0.810	0.890	0.990	0.990	0.990
1000	0.982	0.970	0.980	0.970	0.980	0.980

After the clustering of documents, topic modeling method was used to extract topics from each cluster. For this purpose, Latent Dirichlet Allocation was applied to each cluster and the clustering topics were extracted. Table 4 and 5 represent the top 10 words extracted from each topic within the cluster. The top 10 words mean the most commonly used words in sets of documents.

Table 4 describes $\alpha = 0$, the top 10 words extracted from the sets of documents before the removal of semantic similar words. For comparison, Table 5 describes $\alpha = 0.3$, the top 10 words extracted from the documents after the removal of semantic similar words. These two tables demonstrate that the removal of semantically similar words from the sets of documents does not affect the result. That is, the quality of LDA has not decreased. So, in both table, when we look through the cluster 1, we observe the terms related to entertainment, in cluster 2 terms related to sport, and in cluster 3 terms related to business. This shows that cluster 1 contains the Entertainment, in cluster 2 the Sport and in cluster 3 the Business documents. As seen, the topics extracted from the clusters overlap with each other in both tables. This means that the proposed method works well. Note that this result is also true at other values (0.1, 0.2, 0.4, 0.5) of α. That is, in other values of α, the topics extracted from the clusters overlap with Table 5, so only $\alpha = 0.3$ value was given for comparison.

Table 4. Topics and top 10 words on clusters ($\alpha = 0$)

Cluster 1			**Cluster 2**			**Cluster 3**		
Topic 1	**Topic 2**	**Topic 3**	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 1**	**Topic 2**	**Topic 3**
film	music	music	rugbi	ireland	Open	cluster	Cluster	cluster
star	award	band	england	england	Play	govern	Growth	Year
year	people	show	player	cluster	Cluster	compani	Year	Sale
a.m	show	year	play	win	Win	Firm	Rate	profit
role	year	album	cluster	wale	Year	Deal	economi	compani
actor	won	number	year	side	Seed	Tax	Bank	Firm
cluster	radio	cluster	cup	game	Match	Plan	econom	market
director	veto	chart	zealand	Tri	Set	İndia	Oil	Share
includ	years	top	world	nation	Beat	countri	Price	Car
festiv	song	singer	tour	scotland	World	Foreign	Rise	Euro

Table 5. Topics and top 10 words on clusters ($\alpha = 0.3$)

Cluster 1			**Cluster 2**			**Cluster 3**		
Topic 1	**Topic 2**	**Topic 3**	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 1**	**Topic 2**	**Topic 3**
music	film	music	england	england	open	cluster	cluster	cluster
band	star	show	Ireland	play	play	firm	team	govern
album	year	song	cluster	cluster	cluster	year	year	compani
number	award	peopl	wale	year	year	sale	rate	tax
top	role	year	zealand	rugbi	seed	compani	economi	state
year	cluster	award	side	player	match	profit	rise	Foreign
chart	includ	radio	rugbi	game	set	share	price	Countri
singer	director	make	player	season	beat	market	bank	Oil
rock	play	years	tri	cup	world	a.m	econom	İndia
cluster	bbc	british	game	week	australian	Car	month	china

We observe that the topics were accurately extracted, and we win in time. Thus, the following table describes the time spent on clustering and the extraction of topics from each cluster and their comparative analysis (Table 6).

Table 6. Clustering of documents at different values of α and the time spending on the application of LDA

Number of documents	**Time used** $\alpha = 0$					
	0	0.1	0.2	0.3	0.4	0.5
100	11.6	9.3 (19.82%)	9.86 (15%)	10.32 (11.04%)	10.65 (8.19%)	11.06 (4.65%)
300	12.39	9.08 (26.71%)	9.74 (21.39%)	10.68 (13.80%)	11.26 (9.12%)	11.51 (7.10%)
500	18.7	11.35 (39.30%)	12.87 (31.18%)	13.28 (28.98%)	14.21 (24.01%)	14.98 (19.89%)
800	22.28	13.49 (39.45%)	14.59 (34.52%)	15.76 (29.26%)	16.57 (25.63%)	17.85 (20.33%)
1000	21.06	12.78 (39.31%)	14.09 (33.09%)	15.01 (28.72%)	16.01 (23.97%)	17.25 (18.09%)

As seen in Table 6, a significant success was achieved for time through the proposed method. Thus, if the number of documents is 100 then we win 19.82% at a time, we observe the increases in percentage to 39.31% at $\alpha = 0.1$ as the number of documents increases (the number of documents is 1000). This means a significant increase in efficiency.

Thus, the experiments and results show that increasing the measure of large documents the proposed method can significantly save time for analyzing this data, improve the clustering and the quality of the LDA algorithm.

6. Conclusion and Future Works

Evaluation of services should regularly be carried out to improve the availability and effectiveness of e-government services for citizens. Therefore, to precisely understand the current state of e-government services, it is necessary to analyze both the technical quality and the content of these services. By analyzing the comments on these services, we can improve the quality of services and decrease existing dissatisfaction. LDA topic models can be used to analyze large text sets of citizens' opinions and recommendations in identifying their thoughts quickly. Considering this, in the paper, the proposed approach analyzes citizens' comments using the LDA algorithm and K-Means clustering methods. The main purpose of the method is to increase the efficiency and accuracy of clustering and topic modeling methods by reducing vector size. Through this method, the vector size was reduced, the efficiency of the method was calculated at different boundary values, and the comparative analysis was performed. Finally, the comments were analyzed and the main topics were extracted on each cluster. The result of the experiment it was determined that the method works better as the number of documents increases. This method can reduce the size of large data, save time, and improve the quality of clustering. In future studies, the comments in different languages will be analyzed.

Acknowledgment

This work was supported by the Science Development Foundation under the President of the Republic of Azerbaijan – Grant No. EIF-BGM-4-RFTF-1/2017-21/08/1.

References

[1] G. M. P. Gupta, D. Jana, "E-government evaluation: A framework and case study", *Government Information Quarterly*, vol. 20, no.4, pp. 365–387, 2003.
[2] Ch.-Ch. Huang, "User's Segmentation on Continued Knowledge Management System Use in the Public Sector", *Journal of Organizational and End User Computing*, vol.32, no.1, pp. 19-40, 2020.
[3] L. Hong, B. D. Davison, "Empirical Study of Topic Modeling in Twitter", *Proceedings of the First Workshop on Social Media Analytics*, pp.80-88, 2010.
[4] J. Chang, J. Boyd-Graber, D. M. Blei, "Connections between the lines: augmenting social networks with text", *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pp. 169–178, 2009.
[5] A. McCallum, X. Wang, N. Mohanty, "Joint group and topic discovery from relations and text", *Journal of Statistical Network Analysis: Models, Issues and New Directions*, volume 4503 of Lecture Notes in Computer Science, pp. 28–44, 2007.
[6] H. Zhang, C. L. Giles, H. C. Foley, J. Yen, "Probabilistic community discovery using hierarchical latent gaussian mixture model", *Proceedings of the 22nd National Conference on Artificial Intelligence*, pp. 663–668, 2007.
[7] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, "The author-topic model for authors and documents", *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 487–494, 2004.
[8] D. Ramage, D. Hall, R. Nallapati, C. D. Manning, "Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 248–256, 2009.
[9] X.-H. Phan, L.-M. Nguyen, S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections", *Proceedings of the 17th International Conference on World Wide Web*, pp. 91–100, 2008.
[10] L. Yue, M. Qiaozhu, Z. ChengXiang, "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA", *Information Retrieval*, vol. 14, no.2, pp. 178–203, 2011.
[11] C.-K. Yau, A.L. Porter, N.C. Newman, A. Suominen, "Clustering scientific documents with topic modeling", *Scientometrics*, vol. 100, no.3, pp. 767-786, 2014.
[12] M. Pourvali, S. Orlando, H. Omidvarborna, "Topic Models and Fusion Methods: Union to Improve Text Clustering and Cluster Labeling", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 4, pp. 28-34, 2018.
[13] N. K. Nagwani, "Summarizing large text collection using topic modeling and clustering based on MapReduce framework", *Journal of Big Data*, vol.2, no.6, pp.1-18, 2015.
[14] R. Alghamdi, K. Alfalqi, "A Survey of Topic Modeling in Text Mining", *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, pp. 147-153, 2015.
[15] K. E. C. Levy, M. Franklin, "Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry", *Social Science Computer Review*, vol.32, no.2, pp. 182-194, 2013.
[16] D. M. Blei, Introduction to probabilistic topic models. Communications of the ACM, 2011. Retrieved from http://www.cs.princeton.edu/*blei/papers/Blei2011.pdf
[17] S. I. Nikolenko, S. Koltcov, O. Koltsova, "Topic modelling for qualitative studies", *Journal of Information Science*, vol. 43, no. 1, pp.88-102, 2017.
[18] S. Liu, C. Xia, X. Jiang, "Efficient Probabilistic Latent Semantic Analysis with Sparsity Control", *IEEE International Conference on Data Mining*, pp. 905-910, 2010.
[19] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis", *Machine Learning*, vol. 42, no. 1, pp. 177-196, 2001.
[20] D. M. Blei, A.Y. Ng, M.I. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, vol.3, pp.993-1022,

2003.
[21] M. Shao, L. Qin, "Text Similarity Computing Based on LDA Topic Model and Word Co-occurrence", *Proceedings of the 2nd International Conference on Software Engineering, Knowledge Engineering and Information Engineering*, pp.199-203, 2014.
[22] T. Kakkonen, N. Myller, E.Sutinen, "Applying latent Dirichlet allocation to automatic essay grading*", Lecture Notes in Computer Science,* vol.4139, pp.110-120, 2006.
[23] G.Y.Iskandarli, "Using Hotspot Information to Evaluate Citizen Satisfaction in E-Government: Hotspot Information", *International Journal of Public Administration in the Digital Age*, vol.7, no. 1, pp. 47-62, 2020.
[24] R. Iftikhar, M. S. Khan, "Social Media Big Data Analytics for Demand Forecasting: Development and Case Implementation of an Innovative Framework", *Global Information Management*, vol.28, no.1, pp.103-120, 2020.
[25] S.W. Kim, J.M.Gil, "Research paper classifcation systems based on TF-IDF and LDA schemes", *Human-centric Computing and Information Sciences*, vol.9, no. 30, pp. 1-21, 2019.
[26] R.M.Aliguliyev, G.Y.Niftaliyeva, "Detecting terrorism-related articles on the e-government using text-mining techniques", *Problems of Information Technology*, vol. 6, no.2, pp. 36-46, 2015.
[27] R.M.Alguliyev, R.M.Aliguliyev, G.Y.Niftaliyeva, "Filtration of Terrorism-Related Texts in the E-Government Environment", *International Journal of Cyber Warfare and Terrorism*, vol. 8, no. 4, pp.35-48, 2018.
[28] Sh.A. Takale, S. S. Nandgaonkar, "Measuring Semantic Similarity between Words Using Web Documents", *International Journal of Advanced Computer Science and Applications*, vol. 1, no.4, pp.78-85, 2010.
[29] R. Feldman, J. Sanger, The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data, 2007.
[30] D. Greene, P.Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering", *Proceedings of the 23rd International Conference on Machine Learning*, pp. 377-384, 2006.
[31] R. M. Aliguliyev, "Performance evaluation of density-based clustering methods", *Information Sciences*, vol.179, pp.3583-3602, 2009.

Authors' Profiles

Gunay Y. Iskandarli received her BSc and MSc in applied mathematics from the Baku State University, Azerbaijan in 2012 and 2014, respectively. She is currently a PhD student at the Institute of Information Technology of ANAS. Her research interests include: e-government, online social networks, text mining, data mining, and Big Data analytics. She is the author of 15 papers and 1 book.

How to cite this paper: Gunay Y. Iskandarli, "Applying Clustering and Topic Modeling to Automatic Analysis of Citizens' Comments in E-Government", International Journal of Information Technology and Computer Science(IJITCS), Vol.12, No.6, pp.1-10, 2020. DOI: 10.5815/ijitcs.2020.06.01