


Automatic Detection of the Topics in Customer Complaints with Artificial Intelligence

Sevinç İlhan Omurca, Ekin Ekinci, Enes Yakupoğlu, Emirhan Arslan and Berkay Çapar


Abstract— Today, people first make their complaints and compliments on the internet about a product which they use or a company they are a customer of. Therefore, when they are going to buy a new product, they first analyze the complaints made by other users of the product. These complaints play an important role in helping people make decisions of purchasing or not purchasing products. It is impossible to analyze online complaints manually due to the huge data size. However, companies are still losing a lot of time by analyzing and reading thousands of complaints one by one. In this article, online text based customer complaints are analyzed with Latent Dirichlet Allocation (LDA), GenSim LDA, Mallet LDA and Gibbs Sampling for Dirichlet Multinomial Mixture model (GSDMM) and the performances of them are compared. It is observed that GSDMM gives much more successful results than LDA. The obtained topics of the complaints are presented to users with a mobile application developed in React Native. With the developed application not only the customers will be able to see the topics of complaint from the application interface but also the companies will be able to view the distribution and statistics of the topics of complaints.

Index Terms—Topic modelling, latent dirichlet allocation, gibbs sampling, gibbs sampling for dirichlet multinomial mixture, natural language processing.

SEVİNÇ İLHAN OMURCA, is with Department of Computer Engineering Kocaeli University, Kocaeli, Turkey, (e-mail: silhan@kocaeli.edu.tr).

 <https://orcid.org/0000-0003-1214-9235>


EKİN EKİNCİ, is with Department of Computer Engineering Sakarya University of Applied Sciences, Sakarya, Turkey, (e-mail: ekinekinci@subu.edu.tr).

 <https://orcid.org/0000-0003-0658-592X>


ESES YAKUPOĞLU, is with Department of Computer Engineering Kocaeli University, Kocaeli, Turkey, (e-mail: enesyakupoglu61@gmail.com).

 <https://orcid.org/0000-0003-1702-2647>

EMİRHAN ARSLAN, is with Department of Computer Engineering Kocaeli University, Kocaeli, Turkey, (e-mail: emirhan.arslan@outlook.com.tr).

 <https://orcid.org/0000-0002-5978-9590>

BERKAY ÇAPAR, is with Department of Computer Engineering Kocaeli University, Kocaeli, Turkey, (e-mail: berkaycapar@gmail.com).

 <https://orcid.org/0000-0002-3178-0690>

Manuscript received November 27, 2020; accepted July 3, 2021.

DOI: [10.17694/bajece.832274](https://doi.org/10.17694/bajece.832274)

I. INTRODUCTION

OVER RECENT years, online complaint or compliment narratives play a very important role in people's purchasing decisions. Thus, there has been a great rise on the amount of online customer complaints and compliments. However, complaints attract more attention. The frequently preferred online customer review receiving platforms provide a good resource to collect numerous text based complaints on numerous companies or brands. These collected text based data consider main aspects of the products which customers review about. Thus, there is a huge and valuable resource for detecting the main topics that companies and customers are interested in. Determining the main topics provides a good way to summarize and organize these unstructured text data for companies or customers. The misvaluation or misclassification of complaints delays the start of the resolution process, and, what is worse, causes customer dissatisfaction and soon complaint escalation [1]. Hence, it is very important to detect the main topics of the text based complaints among the huge document collections by automating it. Only this way can carry out relationships between customers and companies correctly.

Latent Dirichlet Allocation (LDA) is a successful, generative probabilistic model that has performed well in analysis of customer complaints as in many text mining tasks. It enables unstructured customer complaints to be clustered into a mixture of topics which underlies the main content of the complaints. One of the strengths of the LDA is that it does not require any prior annotations about the customer narratives and it extracts topics by using original unlabeled documents. LDA has been applied in many fields about text mining, there also have been numerous studies on the analysis of text based customer complaints in the literature. Therefore, LDA addresses the challenge about reading and analyzing complaint narratives by human annotators.

LDA has been applied in many fields about text mining, there also have been numerous studies on the analysis of text based customer complaints in the literature. Kalyoncu *et al.* [2] generated a LDA based solution to manage and visualize the complaints for each Mobile Network Operators (MNOs) in Turkey. Liang *et al.* [3] used the LDA as a topic model to identify product aspects that customers frequently mentioned, then identified the related problems of the product. Bastani *et al.* [4] proposed a method based on LDA to extract latent topics in the CFPB complaint narratives, and explores their associated trends over time. Mai *et al.* [5] built a LDA model

to extract the key points of Guangdong Province consumer complaint text accurately and quickly. Atıcı *et al.* [6] used LDA to determine the key features of complaints and dissatisfactions about products, services or companies by using big data taken from Turkey's largest customer complaint website.

In this article, the topic of customer complaints is determined by the topic models. In other words, it is determined which category of service the complaints belong to. For this aim, the Natural Language Processing (NLP) and Deep Learning techniques are used together. Unlike LDA, which is the most frequently used topic modelling method, Gibbs Sampling algorithm for a Dirichlet Multinomial mixing model (GSDMM) is used in this article because it gives more successful results in short texts. With more accurate results, user reviews will be categorized correctly, information pollution will be prevented, and both customers and companies will be able to analyze user comments accurately. Furthermore, thanks to realized study, it would be possible to notice the changes and trends of the main topics of complaint narratives over the time. Supposing that, if it is realized that the trend of some topics is decreasing over time then it indicates that the company must take into account these aspects of its brand or products. The companies can be able to improve if and only if they are aware of the trends of complaints of their customers.

II. TOPIC MODELLING ALGORITHMS

Topic model algorithms convert document collections to low dimensional space for the purpose of modelling hidden thematic structure within. In this study, three different topic models are used. These are LDA with GenSim, LDA with Mallet, GSDMM. The reason for using LDA in this study is because it is a frequently used and generally very successful topic model however may fail on short texts [7]. The GSDMM method was developed to improve topic modelling in short texts [8]. Since working with short texts, the GSDMM method is used.

A. Latent Dirichlet Allocation

LDA is a generative graphical model used to model discrete data such as textual data to reveal the topics that comprise the document. Its use includes topic modelling. LDA does not need any prior knowledge due to a fully unsupervised model. LDA is based on the "bag-of-words" assumption that while word orders in documents are regarded, word co-occurrence in the same document is taken into account. The basic idea behind LDA is this: topics have probability distribution over a fixed vocabulary and documents are composed of random mixture of latent topics. While the input of the model is documents, its outputs are the topics, probabilities of words under these topics, topic for each word, and the topic mixture for each document [9].

LDA is a generative model and this is its most important feature. In the generative process firstly words are sampled from fixed vocabulary for each topic. For every document topic proportions are sampled. The Dirichlet distribution is

used in these two steps. Each topic in the topic distribution is sampled randomly for every word in all documents. At the end, for a related topic a word is sampled. In the sampling steps multinomial distributions are used.

Graphical representation of LDA takes advantage of plate notation. Plate notation is a graphical model of representing repeated variables. A plate or rectangle is used to group variables together into repeating nodes, rather than plotting each repeating variable separately. Plate notation for LDA is given in Fig. 1. Only documents can be observed; the topics, topic distribution for each document and words in the topics are hidden. Therefore, while the observed variables in the graphical model are represented by shaded, those that cannot be observed are represented by non-shaded.

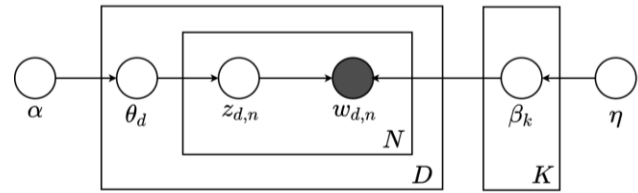


Fig.1. Graphical representation of Bayes [10]

In the graphical model given with Fig. 1 D is the total number of documents in the collection and K is the total number of topics. α and η are Dirichlet parameters. θ shows the probability of the topics which are found in the document, and β indicates the word distributions in the topics. While $z_{d,n}$ represents the topic of the word in the n th position in the d -th document, $w_{d,n}$ represents the word in the n -th position in the d -th document. According to the graphical model, the joint distribution of all hidden and observed random variables $p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})$ is given in Eq. 1.

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \left(\prod_{k=1}^K p(\beta_k / \eta) \right) \left(\prod_{d=1}^D p(\theta_d / \alpha) \right) \left(\prod_{n=1}^N p(z_{d,n} / \theta_d) p(w_{d,n} / z_{d,n} \beta_k) \right) \quad (1)$$

As mentioned earlier, it is intended to obtain model parameters with LDA. For this purpose, the distribution of $p(\beta, \theta, z / w)$ is estimated approximately in the form of $q(\beta, \theta, z)$ using variational inference. Then, model parameters are trained to minimize the Kullback–Leibler Divergence (KL) deviation between q and p . This process is called $KL(q // p)$ Minimization in the literature. q modelling is extremely difficult for a high dimensional distribution. To further reduce complexity, bold independence assumptions similar to the graphical model is made and the common possibility is divided into independent subcomponents. A more detailed explanation for each subcomponent is in Eq. 2.

$$q(\theta_{1:D}, z_{1:D,1:N}, \beta_{1:K}) = \prod_{k=1}^K q(\tilde{\beta}_k / \tilde{\lambda}_k) \prod_{d=1}^D q(\tilde{\theta}_d / \tilde{\gamma}_d) \prod_{n=1}^N q(z_{d,n} / \tilde{\phi}_{d,n}) \quad (2)$$

Here, instead of a common probability model for multiple variables, each variable is modelled independently, each variable v_i is modelled as $q_i(v_i / \rho_i)$ with a given distribution family nominal distribution. This is called Mean field variational inference, where the common distribution is traceable, easy to analyze, separating individual variables into their distributions. Assumptions of independence may be wrong; but observation should be continued as flawed rather than wrong. Subsequent experimental results often produce quality results. To optimize dependent variables, each group is divided into groups with independent variables from each other.

In this way, the parameters of topic assignment modelling and topic ratios are determined in LDA. The most difficult steps here are to assign topics and then update the variational parameters of topic rates. So, it is required to model P with q by minimizing KL Divergence. Because it cannot be directly optimized, operations are continued with Evidence lower bound (ELBO). The ELBO definition is made in Eq. 3.

$$\begin{aligned} ELBO &= -\sum_x q(x) \log \frac{q(x)}{\tilde{p}(z, x)} \\ &= -\sum_x q(x) \log \frac{q(x)}{p(z, x)} + \log Z \\ &= -KL(q \parallel p) + \log Z \\ \log Z &= ELBO + KL(q \parallel p) \geq ELBO \end{aligned} \quad (3)$$

The relationship between KL-Divergence and ELBO is evaluated in Eq. 3. KL-Divergence is always positive, $\log Z$ is therefore always larger or equal to ELBO. ELBO in here is the lower bound of $\log Z$ for any q . From this it can be said that when q is equal to p , $\log Z$ and ELBO are equal to each other. ELBO is wanted to maximize in LDA as in Eq. 4. Here the expectation value is calculated with q .

$$\begin{aligned} L &= \sum_{k=1}^K E[\log p(\tilde{\beta}_k / \tilde{\eta})] + \\ &\sum_{d=1}^D E[\log p(\tilde{\theta}_d / \tilde{\alpha})] + \sum_{d=1}^D \sum_{n=1}^N E[\log p(z_{d,n} / \tilde{\theta}_d)] + \\ &\dots + \sum_{d=1}^D \sum_{n=1}^N E[\log p(w_{d,n} / z_{d,n} \tilde{\beta}_{1:K})] + H(q) \end{aligned} \quad (4)$$

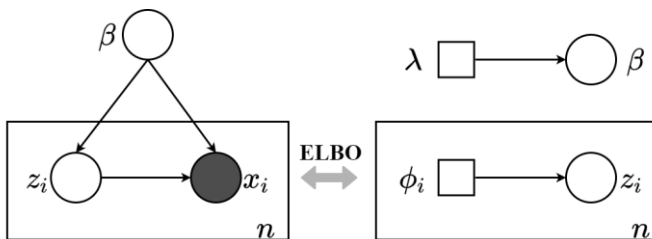


Fig.2. LDA simplified model and ELBO relationship [10]

In Fig. 2, the simplified version of the graphic model of LDA is shown with ELBO. Here x_i is an observation linked to β and z_i . The equation 1 is simplified and becomes as in Eq. 5.

$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^n p(z_i / \beta) p(x_i / z_i, \beta) \quad (5)$$

The right side of the equation (R.H.S) has an exponential distribution, and q will be as in the equation $q(\beta, z) = q(\beta / \lambda) \prod_{i=1}^n q(z_i / \phi_i)$ with Mean field variational inference. β and z_i Eq. 6 it has been removed as said.

$$\begin{aligned} p(\beta / z, x) &= h(\beta) \exp\{n_g(z, x)^T \beta - \alpha(n_g(z, x))\} \\ p(z_i / \beta, x_i) &= h(z_i) \exp\{n_l(\beta, x_i)^T z_i - \alpha(n_l(\beta, x_i))\} \\ q(\beta / \lambda) &= h(\beta) \exp\{\lambda^T \beta - a(\lambda)\} \\ q(z_i / \phi_i) &= h(z_i) \exp\{\phi_i^T z_i - a(\phi_i)\} \end{aligned} \quad (6)$$

Getting the right side of the equation with q and p by optimizing ELBO according to λ and ϕ_i is intended here. The derivative of L according to λ is taken and set to 0. The optimal λ^* subtraction and optimal ϕ_i^* are obtained with Eq. 7 and 8 respectively.

$$\nabla_{\lambda} L = a'(\lambda) (E_{\phi} [n_g(Z, x)] - \lambda) \quad (7)$$

$$\begin{aligned} \lambda^* &= E_{\phi} [n_g(Z, x)] \\ \phi_i^* &= E_{\lambda} [n_l(\beta, x_i)] \end{aligned} \quad (8)$$

B. Gibbs Sampling Dirichlet Multinomial Mixture

GSDMM is a collapsed Gibbs Sampling algorithm for a Dirichlet Multinomial mixing model developed specifically to cope with the challenges of topic extraction from short texts. The functioning of GSDMM is as follows; (I) initially, documents are randomly assigned to clusters, and the information \vec{z} (cluster label), m_z (number of document in cluster z), n_z (number of words cluster z), and n_z^w (occurrence count of word w in cluster z) are recorded, (II) documents are searched in I iteration, (III) in each iteration, a cluster is assigned to each document d based on conditional distribution, respectively: $p(z_d = z / \vec{z}_{-d}, \vec{d})$; where $-d$ means that the label of cluster document d is removed from \vec{z} , (IV) in every iteration \vec{z} , n_z , and n_z^w are updated accordingly, (V) as a result, only some few initial K clusters will not be empty.

- GSDMM can automatically get the cluster count.
- GSDMM can balance the integrity and homogeneity of the results.
- GSDMM is good for convergence.
- GSDMM can overcome sparsity or high dimensionality of short texts.
- LDA and PLSA etc. as with traditional topic models, GSDMM can obtain representative words of each cluster.
- Yin and Wang proposed the Movie Group process (MGP) for instantiation that could help to understand how GSDMM works and the meaning of its parameters [11].

1) *Movie group process*

A film discussion course can be imagined in which the Professor is planning to split students into a number of groups. The professor wants students to write down the movies they are watching. The list won't be very long and students will probably write down the last movies they've watched recently or the ones they've loved so much. Now each student can be modeled as a list of movies. The professor must find a way to split students into a number of groups. The professor's expectation is that while the students in the same group watch common movies so they will have more to discuss, those in different groups also have different movies to discuss.

If only a few clusters have documents and others are empty, if it is explained through the professor, his students and movies, the story described above becomes the following. The professor invites his students to a restaurant where there are K tables. At first he assigns the students to K tables randomly. Then he asks the students to select a table one by one. However, the student should consider the following two rules while selecting the table.

- The student must choose a table which has more students.
- The student must choose the table where the students who have similar interests (watch the common movies more).

These steps repeat until some grow and some disappear. The expectation is that there will be students at some tables and students who are at these tables will share a similar interest.

2) *Gibbs sampling*

Gibbs Sampling is a special case of the Metropolis and Metropolis-Hastings algorithm used to make interpretations about complex stochastic models. Gibbs Sampling produces sample values for the density distribution for the corresponding result values of all the parameters in the model by sampling from all the distribution values in sequence. Due to its simple logical basis and ease of application, this method is quite widely preferred in NLP, artificial intelligence, and deep learning tasks.

$\theta = (\theta_1, \dots, \theta_k)$ parameter vector, $p(y|\theta)$ possibility and suppose $\Pi(\theta)$ a predecessor distribution. Expression $\Pi(\theta_i|\theta_j, i \neq j, y)$ can be written as $\Pi(\theta_i|\theta_j, i \neq j, y) \propto p(y|\theta)\Pi(\theta)$. The Gibbs sampling algorithm is as follows:

- i. $t = 0$ is taken and an arbitrary $\theta(0) = \{\theta_1(0), \dots, \theta_k(0)\}$ initial value is selected.
- ii. Each component of θ is obtained in the format as below:
- iii. $\Pi(\theta_1|\theta_2(t), \dots, \theta_k(t), y)$ as $\theta_1(t+1)$,
- iv. $\Pi(\theta_2|\theta_1(t+1), \theta_3(t), \dots, \theta_k(t), y)$ as $\theta_2(t+1)$,
- v.
- vi. $\Pi(\theta_k|\theta_1(t+1), \dots, \theta_{k-1}(t+1), y)$ as $\theta_k(t+1)$.
- vii. Take $t = t + 1$ and go to step 2 if $t < T$ (T is the desired sample width). Otherwise, the process is finished.

3) *Dirichlet multinomial mixture*

Unlike normal text clustering, short text clustering poses sparsity problems [12]. Most of the words occur only once in short texts, as a result TF-IDF measure does not fare well in this type of texts. Also, if VSM is used for representation, sparse and high-dimensional feature space results in a waste of both computational time and memory [13] In response to all these short text clustering challenges, the DMM model should be used [14]. DMM is a probabilistic generative model like LDA and is based on two assumptions behind the generative process. The first one is that a mixture model is used in the generative process for documents. The other is there is a one-to-one match between documents and clusters, hence, it is claimed that each document comes from only one topic. When generating document d , initially DMM chooses cluster k that is the mixture of cluster based mixture component $p(z=k)$. The document d is then generated with mixture component drawn from $p(d/z=k)$. The Fig. 3 represents the plate notation of DMM as below.

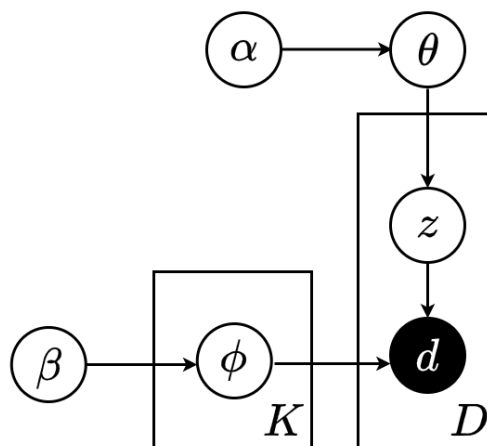


Fig.3. DMM's graphical model [11]

In the Fig. 3 above α is a parameter that affects the shape of the probability distribution and is derived from the probability that a document is grouped in a cluster. β is another shape parameter for distribution. β represents the similarity of words between two documents. ϕ is the omnidirectional distribution of words, $p(w/z=k) = \phi$, where w , words and z are the cluster label. θ is a multi-term distribution considering alpha, so $p(d/z=k) = \theta$; where d is the document.

$$p(d) = \sum_{k=1}^K p(d/z=k) p(z=k) \tag{9}$$

In Eq. 9, K represents the number of clusters, the calculation of $p(d/z=k)$ is the principal problem, where DMM makes the Naive Bayes assumption. Based on this assumption, if the cluster k which the document belongs to is known, words in this document are generated independently and probability of each word is accepted as independent within the document. Then, the probability of the document d generated by cluster k can be obtained as in Eq. 10.

$$p(d/z=k) = \prod_{w \in d} p(w/z=k) \tag{10}$$

III. RESULTS AND DISCUSSIONS

A. Model Design Process

The first stage of the study is data mining using KNIME from sikayetvar.com website. Customer's complaints data taken with KNIME and saved in csv format. Then, the dataset is cleaned for pre-processing with java and python programming. The purpose of pre-processing is to clear the data as much as possible and reduce it to words that will work only when modelling the topic. Three different topic modelling methods are applied to the cleaned data and the results of the models are compared. The study is continued by using GSDMM because it gives more successful results than other methods. After topic modelling with GSDMM, assigning the topic title to the documents is done with java programming on the NetBeans IDE platform. After assigning the topic title to the documents, the complaints of each document are saved. The complaint-topic information is saved and converted to JSON format and is presented to users in the mobile application developed in the React Native environment. The architecture of the model design process is given in Fig. 4.

B. Dataset

The dataset, which is used in this study, consists of the complaints of customers for the products or services they

receive from the companies. The complaint data written on the sikayetvar.com website is used to create experimental dataset and the KNIME software is preferred to get these complaints. KNIME is an open source and cross platform data analysis platform. It was developed by a company named Konstanz Information Miner in Java and was founded on Eclipse basis [6]. With KNIME it is convenient to process large data limited to the available hard disk space, and Weka, Tableau, the statistics package R project, LIBSVM etc. supports other machine learning and data analysis projects. KNIME contains many nodes for machine learning and data mining needs. These nodes can be associated to process, interpret, visualize and report data. The associated nodes are run in the specified order and produce output. The turns of the nodes can be followed on the console. KNIME offers community extensions for needs. These extensions offer many nodes from various application areas. These extensions can be installed from the application interface. extensions from outside the community as ZIP archives can be also added and used. In realized KNIME project, Palladian extension is added to the application to process the data on sikayetvar.com. Palladian KNIME is an open source Java library that is not included in community extensions. With the web mining node, it offers, data on HTML, RSS feeds and JavaScript supported websites can be parsed and important content can be obtained.

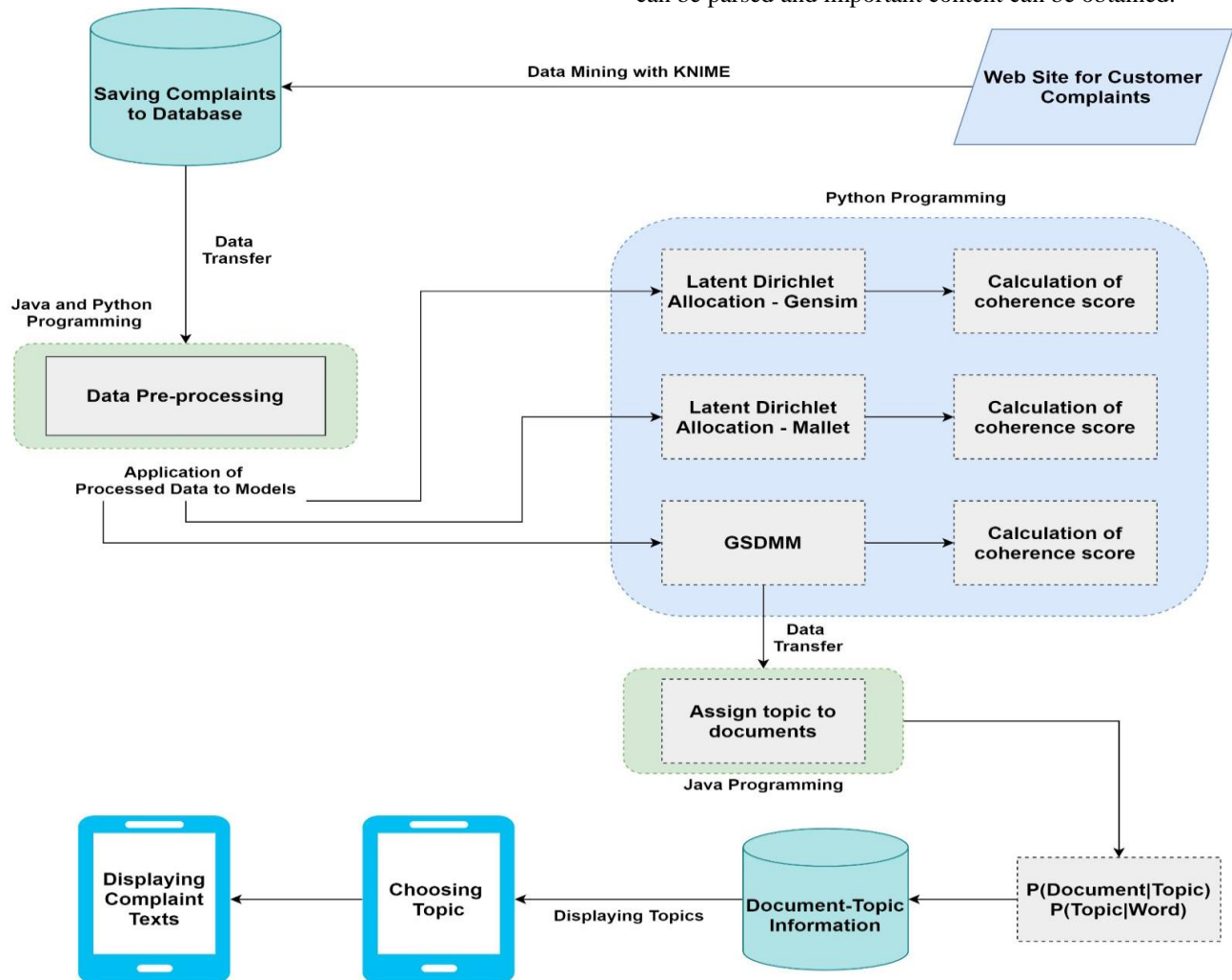


Fig.4. Model design process

C. Comparison of Topic Modelling Algorithms

In this study three topic modelling methods are used and compared in terms of the performance. These 3 methods are: LDA with GenSim, LDA with Mallet and GSDMM. The reason for choosing these models is that LDA is the most popular topic modelling method and GSDMM is successful in short texts. LDA is one of the most used topic modelling methods and the success rate is generally high. GSDMM is developed to improve topic modelling in short texts.

1) Topic modelling with Mallet LDA

Mallet is a Java based library used for performing topic modelling. With Mallet Latent Dirichlet Allocation, Pachinko Allocation and Hierarchical Latent Dirichlet Allocation topic models can be realized. Mallet version 2.0.8 is used in this study. There are two steps to be followed for Mallet topic modelling; topic import from dataset and topic training [15].

- Data import: In this step, the dataset used for topic extraction is imported and pre-processing is applied to remove noise and improve quality. After this step, ".mallet" binary file is created by using a pre-processed dataset.
- Topic training: In this step, topic training is realized by using pre-processed data that comes from the data import step. In this step, parameters to smooth document-topic distributions can be decided and adjusted.

After these steps, three files which include topic words, matrix of the most appropriate document-topic pairs, and Gibbs sampling results after each iteration is obtained. With Gibbs sampling it can be learnt that which word belongs to which topic, the source of each word, and the alpha and beta values. While applying Mallet LDA to the dataset, random_seed is set to 100, num_topics is set to 100, workers is set to 3, α is set to 0.5 and β is set to 0.01.

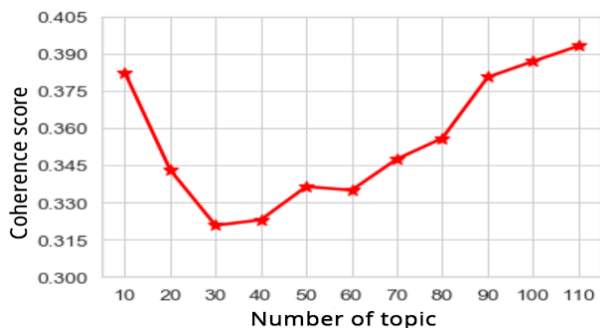


Fig.5. Coherence score graph by topic number with Mallet LDA

Fig. 5 shows the change of coherence score with topic number in Mallet LDA. Here, while calculating the coherence score, the CoherenceModel (c_v) function available in Mallet and GenSim are used. As seen in the graph, it gives the lowest coherence score at $K = 30$, while it gives the highest coherence score at $K = 110$.

2) Topic modelling with GenSim LDA

GenSim is a python python for unsupervised topic modelling implementations such as LDA, LSA, Hierarchical Dirichlet Process (HDP). Its biggest advantage is that it is memory

independent. That means there is no need to load all training data to RAM at any given time. While extracting topics with GenSim, at first preprocessing is applied to the dataset to reduce negative effects of noise and then training is carried out. In this study, the parameters m_state is set to 100, chunksize is set to 10000, passes is set to 20, num_topics I set to 100, α is set to 0.5 and β is set to 0.01.

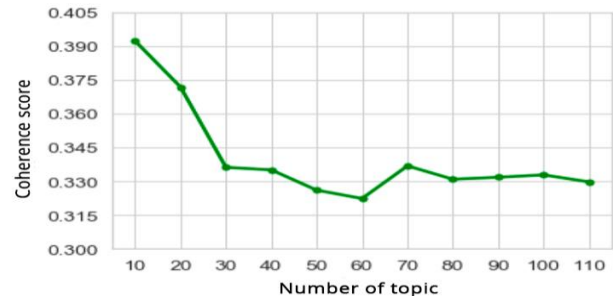


Fig.6. Coherence score graph by topic number with GenSim LDA

Fig. 6 shows the change of coherence score according to the number of topics in GenSim LDA. Here, while calculating the coherence score, the CoherenceModel (c_v) function available in Mallet and GenSim is used. It gives the lowest coherence score at $K = 60$, while it gives the highest coherence score at $K = 10$.

3) Topic modelling with GSDMM

In this model, the assumption that each document is related to one topic at most and the words in these documents depend on the related topic. As mentioned earlier, GSDMM can produce successful results in short texts.

When applying GSDMM in the dataset, K is set to 100, α is set to 0.5 and β is set to 0.01. Due to the algorithm that GSDMM uses MGP, some clusters are empty.

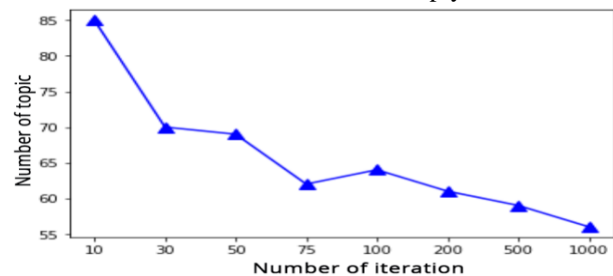


Fig.7. Change in the number of topics with increasing iteration in GSDMM

Fig. 7 shows the number of topics given by GSDMM in different iterations for $K = 100$. It gives the closest value to 100 in 10 iterations and the farthest value to 100 in 1000 iterations. It gives 69 topics in 50 iterations, it means that 31 clusters are empty.

According to Fig. 7, it can be said that some of the initial clusters may be empty after a few iterations so at the end number of clusters will be between 1 and K [16].

In this study, 3 different topic modelling methods are used with same parameters and same dataset and results compared. In all 3 models, K (topic number) = 100, $\alpha = 50 / K = 0.5$, $\beta = 0.01$ are taken. All models are run with 50, 100, 200, 500 and 1000 iteration values. The coherence score formula in Eq. 11 is used to compare the results of the topic modelling.

$$C(k;V^{(k)}) = \sum_{n=2}^N \sum_{l=1}^{n-1} \log \frac{D(V_n^{(k)}, V_l^{(k)}) + 1}{D(V_l^{(k)})} \quad (11)$$

In Eq. 11, the value of N indicates how many words represent each topic. In this study, $N = 10$ is taken and the most popular 10 words of each topic are used. V_n and V_l indicates the number of documents in the number of documents together. The denominator section $D(V_l)$ gives the number of how many documents the word V_l contains.

In Fig. 8 and Table II, the coherence scores obtained with different iteration numbers are shown. LDA of the Mallet library, gives the best result in 50 iterations. The LDA of the

GenSim library gives the best coherence score in 500 and 1000 iterations. In addition, it is observed that the results do not change in 500 and 1000 iteration values in GenSim LDA. GSDMM gives the best coherence score with 50 iterations. When the 3 methods are compared, it is seen that the most successful method in each iteration number is the GSDMM method. The biggest reason for this is that the documents contain short texts. Due to the low number of words in short texts, GSDMM gives more successful results than other methods. The next stages of the study is continued with the GSDMM method as it gives the most successful method.

TABLE II
COMPARISON OF SUCCESS OF THE MODELS

Method	Iteration Count				
	50	100	200	500	1000
LDA Gensim	-90.0322	-89.0056	-87.1602	-86.5808	-86.5808
LDA Mallet	-121.2727	-130.1922	-131.4759	-129.2213	-128.8657
GSDMM	-77.6599	-85.4374	-82.5348	-80.6974	-79.6104

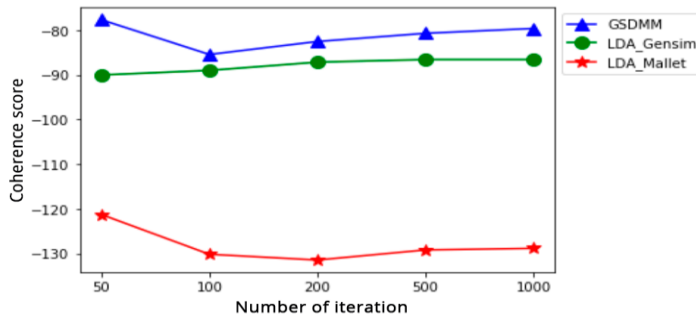


Fig.8. Comparing the success of the model

4) Clustering and topic assignment

GSDMM is used for clustering and topic assignment because it gives the most successful coherence score in experimental results. Topic number (K) = 100, $\alpha = 0.5$, $\beta = 0.01$ and iteration number = 50 values are taken. Due to the algorithm that GSDMM uses, some of the clusters are empty. For this reason, 69 clusters have been obtained, not 100. So 31 clusters are empty clusters. The 69 clusters obtained are reduced to 20 clusters manually, according to their semantic similarities. Thus, 20 semantically most logical, similar and successful topics are obtained. The 10 most popular keywords are used in each topic. The 20 clusters are named manually by us, according to the meanings and similarities of the keywords.

TABLE III
TOPICS AND KEYWORDS

Campaign	Account Operations	Damaged Product	Shoes	Delivery
tl(turkish lira)	password	product	product	cargo
coupon	member	cargo	shoe	delivery
discount	active	broken	brand	product
product	email	return	original	branch
return	login	damage	quality	order
shopping	account	order	fake	deploy
campaign	address	delivery	return	house
customer	support	company	shopping	company
code	issue	tv	sticker	address
card	mail	box	sport	shopping
Accessory	Electronic Household Stuffs	Pets	Housewares	Home Textile
product	machine	cat	product	product
order	product	food	broken	cover
silver	hair	dealer	karaca	english
necklace	brand	insect	return	quilt
color	coffee	brand	cargo	home
set	order	alive	piece	order
earring	dryer	fair	order	linens
chain	arzum	registry	delivery	image
wristband	return	product	plate	brand
pink	house	goody	set	picture

Table 3 shows the 10 topic titles and 10 keywords for each topic. Keywords that are not related to the topic, that is, clustered incorrectly, are indicated with red.

In order to assign the topic to the complaint documents, a java code on the NetBeans IDE platform is written. For the text of the complaint, it is counted how many times the keywords of each topic are mentioned. A score is kept for

each topic and a complaint is assigned to the topic with the highest score. In addition, a threshold value is determined for the score, complaints that do not exceed this threshold value are assigned to another topic called "Others". In other words, if a complaint is not sufficiently similar to the topics obtained, it is assigned to the topic title containing other complaints called "Others".

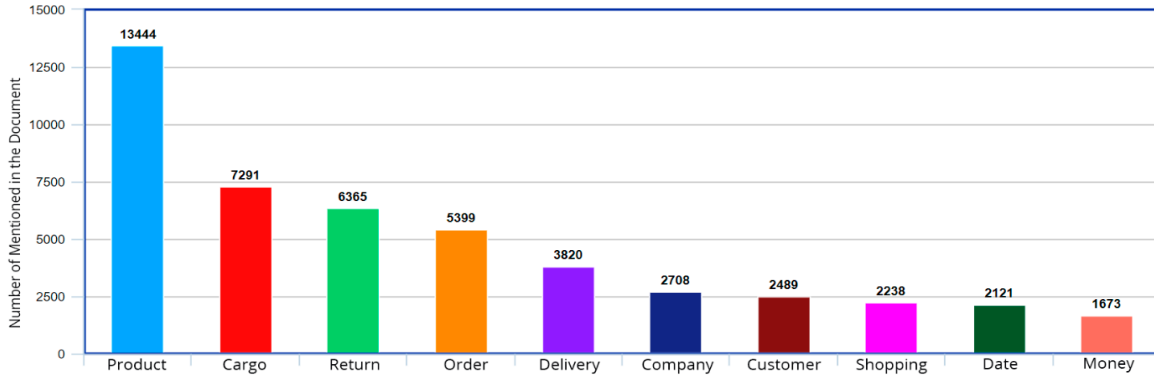


Fig.9. The most popular keywords from the complaints of Trendyol Company

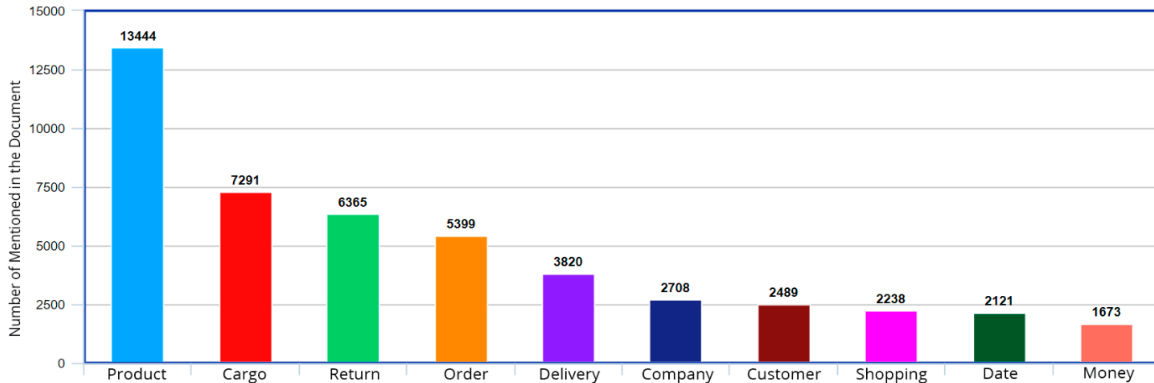


Fig.9. The most popular keywords from the complaints of Trendyol Company

Fig. 9 shows the most popular keywords from 6963 complaints of Trendyol company. Fig. 10 shows the distribution of the topic titles from the 6963 complaints of Trendyol company. In addition, 2477 complaints that do not exceed the threshold value are included in the "Others" topic.

D. Mobile Application

After the topic modelling and assignment of complaints, all complaints and statistics are displayed on a mobile application. The mobile application is developed for android with React Native. The person who logs into the application must choose the company first. Since only Trendyol complaints are used in this study, there is Trendyol company as the only choice in the section of companies. After the company selection, the user can choose 3 different options.

These are: complaints, topic distribution statistics and keyword distribution statistics.

The user who chooses the keyword distribution option shows the top 10 keywords in 6963 complaints of Trendyol company as pie chart and column chart. In the 6963 complaints of the Trendyol company, the user who chose the topic distribution option shows the 10 most popular topic titles as pie chart and column chart. The user, who chooses the complaints option, has 22 different options, including 20 topics, all complaints and other complaints. The user can view all the complaints if he wishes, or, if he wishes, the complaints belonging to a specific topic. From the "Others" option, the topic can view and read the complaints that could not be assigned.

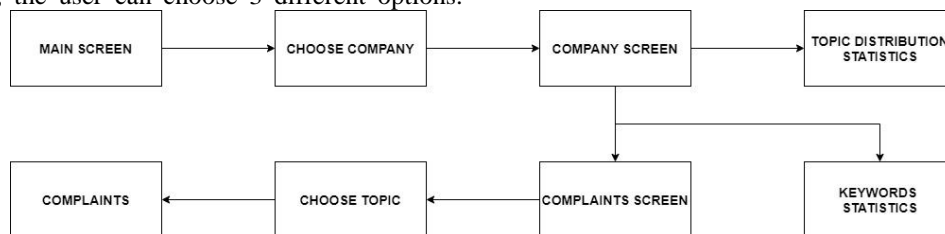


Fig.10. Process of the mobile application

IV. CONCLUSION

As a result of this study, it is realized that how important user complaints are for both customers and companies. Afterwards, the importance of the categorized process, which can be considered as a summary of the complaints, is understood and researches are made on how this structure could be handled.

For the topic modelling, 6963 complaint data and data set are created from the website of sikayetvar.com. In order to facilitate the machine learning process and achieve more consistent results, the data set is pre-processed. Then 2 different methods for topic modelling is used: LDA and GSDMM. In this study, LDA is applied using two different libraries, GenSim and Mallet. Thus, three different models are tested in the complaint data and the results are compared. Since document data consists of short texts, it is determined that the GSDMM method gives more successful results. Therefore, GSDMM method is used in the continuation of the study.

Topic modelling is applied to the document with GSDMM and clusters are obtained. Then, in order to assign the topic of the complaint data, the ratio of how often the keyword is mentioned in the complaint and document is examined. Scores are kept for each topic and each complaint is assigned to the topic that got the highest score. An average threshold value is determined and complaints that do not exceed the threshold value are assigned to a topic called "Other". The complaints, which are assigned to the topic, are presented to the users with the mobile application developed with React Native. Each topic and all complaints of the title can be displayed on a different screen. At the same time, the most frequently mentioned keywords and the most popular topics in all complaints are presented to the users as pie and column charts. The successes of the thesis study can be listed as follows:

- Customers and companies will be able to conduct complaints analysis more consistently, by categorizing their complaints according to their titles,
- Thanks to the process of categorizing complaints by artificial intelligence, companies will use less time and manpower when they analyze complaints,
- Three different methods from topic modelling methods are compared together. As a result, GSDMM method is more successful than LDA. The main reason for this is that the complaint texts are composed of short texts,
- The most popular topic modelling method LDA's failure in short texts is seen in this study,
- More consistent and successful results can be achieved with DMM based models in short texts,
- Gensim and Mallet applications, which are 2 different LDA Libraries, are compared and as a result, GenSim is more successful than Mallet.

In order for this study to give more successful results, improve suggestions can be listed as follows:

- The dataset can be expanded,
- It can be worked with dataset containing longer texts,
- Pre-processing steps can be more frequent. Thus, more successful results can be obtained by working with minimum stop-words and maximum keywords,

- Multiple topic modelling methods can be used together,
- Other DMM-based methods such as GPU-DMM can be used,
- Different topic modelling methods such as ATM, BTM, TTM, ASTM, TATM, SATM can be tried and compared together. The success of the methods can vary depending on your dataset and the results you want.

More consistent results can be obtained by changing the parameters of the methods used, such as alpha, beta, and the number of topics according to your dataset and the desired outputs.

ACKNOWLEDGMENT

Thanks to TÜBİTAK for their support to the project numbered 1919B011902805 within the scope of TÜBİTAK-2209-A University Students Research Projects Support Program 2019/2.

REFERENCES

- [1] S. Prasad, "Use of Natural Language Processing to Improve Complaint Classification in Customer Complaint Management System", *Journal of Critical Reviews*, Vol.7, No.14, 2020, pp.2642-2652.
- [2] F. Kalyoncu, E. Zeydan, İ. O. Yiğit, A. Yıldırım, "A Customer Complaint Analysis Tool for Mobile Network Operators." 2018 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining, Barcelona, Spain, 2018.
- [3] R. Liang, W. Guo, D. Yang, "Mining product problems from online feedback of Chinese users", *Kybernetes*, Vol.46, No.3, 2017, pp.572-586.
- [4] K. Bastani, N. Hamed, S. Jeffrey, "Latent Dirichlet allocation (LDA) for topic modelling of the CFPB consumer complaints", *Expert System with Applications*, Vol.127, 2019, pp.256-271.
- [5] W. Mai, M. Wei, J. Zhang, F. Yuan, "Research on Chinese text and application based on the Latent Dirichlet Allocation." 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering. Shenzhen, China, 2020.
- [6] B. Atıcı, S. İlhan Omurca, E. Ekinici, "Product aspect detection in customer complaints by using latent dirichlet allocation." 2017 International Conference on Computer Science and Engineering. Antalya, Turkey, 2017.
- [7] X. He, H. Xu, X. Sun, J. Deng, X. Bai, J. Li, "Optimize collapsed Gibbs sampling for biterm topic model by alias method." 2017 International Joint Conference on Neural Networks. Anchorage, Alaska, 2017.
- [8] R. Albalawi, T. H. Yeap, M. Benyoucef, "Using Topic Modelling Methods for Short-Text Data: A Comparative Analysis," *Frontiers of Artificial Intelligence*, Vol.3, No.42, 2020, pp.1-14.
- [9] E. Ekinici, S. İlhan Omurca, "An Aspect-Sentiment Pair Extraction Approach Based on Latent Dirichlet Allocation for Turkish," *International Journal of Intelligent Systems and Applications in Engineering*, Vol.6, No.3, 2018, pp.209-213.
- [10] D. M. Blei, "Probabilistic topic models," *Communications of ACM*, Vol.55, No.4, 2012, pp.77-84.
- [11] J. Yin, J. Wang, "A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering." 20th ACM SIGKDD International Joint Conference on Knowledge discovery and data mining. New York, USA, 2014.
- [12] C. C. Aggarwal, C. Zhai, *Mining text data*, Springer, 2012, pp.77-128.
- [13] G. Salton, C. S. Yang, C. T. Yu, "A theory of term importance in automatic text analysis", *Journal of the American Society for Information Science*, Vol.26, No.1, 1975, pp.33-44.
- [14] K. Nigam, A. K. McCallum, S. Thrun, T. M. Mitchell, "Text classification from labeled and unlabeled documents using EM", *Machine Learning*, Vol.39, No.2/3, 2000, pp.103-134.
- [15] I. Akef, X. Xu, J. S. Munoz Arango "Mallet vs GenSim: Topic modeling for 20 news groups report", 2016.

- [16] P. Stiff, *Analysis of Remarks Using Clustering and Keyword Extraction*, Master Thesis, 2018, p. 7.

BIOGRAPHIES



SEVINÇ İLHAN OMURCA is an Associate Professor at the Kocaeli University Computer Engineering Department in Turkey. She has Ph.D. at the Kocaeli University Electronics and Communication Engineering. Her main research interest includes text mining, sentiment analysis, natural language processing, machine learning and data mining earned.



EKİN EKİNCİ is an Assistant Professor of Computer Engineering Department at Sakarya University of Applied Sciences in Turkey. She has received her BS. in Computer Engineering from Çanakkale Onsekiz Mart University in 2009 and MS. in Computer Engineering from Gebze Technical University in 2013 and Ph.D. degree in Computer Engineering from Kocaeli University in 2019. Her main research interest includes text mining, sentiment analysis, natural language processing and machine learning.



ENES YAKUPOĞLU received the B.S. degree in computer engineering from the Kocaeli University in 2020. From 2018 to September 2020, he worked as an Intern Engineer at Argelabs Information Technologies company. He has been working as a Computer Engineer at the same company since September 2020. His research interests include artificial intelligence, natural language processing, machine learning, deep learning, digital marketing, mobile and web application developing.



EMİRHAN ARSLAN is a student at the Kocaeli University Computer Engineering Department. From July 2018 to August 2018, he was an intern at Kocaeli University, Image Processing Laboratory. His research interests are natural language processing, text mining, big data, data visualizations and internet of things.



BERKAY ÇAPAR is a student at the Kocaeli University Computer Engineering Department. From July 2018 to August 2018, he was an intern at Kocaeli University, Image Processing Laboratory. His research interests are natural language processing, artificial intelligence, big data, web and mobile programming applications.