

**TOPIC MODEL BASED RECOMMENDATION SYSTEM TO
IDENTIFY OPERATIONS THAT ARE MISSING IN THE
TREATMENT**

**TEDAVİDE EKSİK OLAN İŞLEMLERİ BELİRLEMEK İÇİN
KONU MODELİNE DAYALI ÖNERİ SİSTEMİ**

KAMURAN NUR KİRAZ

Asst. Prof. Dr. GÖNENÇ ERCAN
Supervisor

Submitted to
Graduate School of Science and Engineering of Hacettepe University
as a Partial Fulfillment to the Requirements
for the Award of the Degree of Master of Science
in Computer Engineering

2019

This work titled “**Topic Model Based Recommendation System to Identify Operations that are Missing in the Treatment**” by **KAMURAN NUR KIRAZ** has been approved as a thesis for the Degree of **Master of Science in Computer Engineering** by the Examining Committee Members mentioned below.

Prof. Dr. Fazlı CAN

Head

.....


Asst. Prof. Dr. Gönenç ERCAN

Supervisor

.....

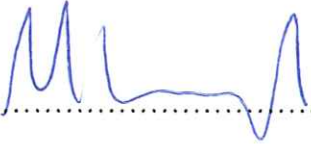

Asst. Prof. Dr. Aybar ACAR

Member

.....


Asst. Prof. Dr. Mehmet KÖSEOĞLU

Member

.....


Asst. Prof. Dr. Fuat AKAL

Member

.....


This thesis has been approved as a thesis for the Degree of **Master of Science in Computer Engineering** by Board of Directors of the Institute for Graduate School of Science and Engineering on/...../.....

Prof. Dr. Menemşe GÜMÜŞDERELİOĞLU

Director of the Institute of

Graduate School of Science and Engineering

ETHICS

In this thesis study, prepared in accordance with the spelling rules of Institute of Graduate School of Science and Engineering of Hacettepe University,

I declare that,

- all the information and documents have been obtained in the base of the academic rules
- all audio-visual and written information and results have been presented according to the rules of scientific ethics
- in case of using others works, related studies have been cited in accordance with the scientific standards
- all cited studies have been fully referenced
- I did not do any distortion in the data set
- And any part of this thesis has not been presented as another thesis study at this or any other university.

19 / 09 / 2019


Signature

NAME-SURNAME

Kamuran Nur KIRAZ

YAYINLANMA FİKRİ MÜLKİYET HAKKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan “Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge” kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H. Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir.

Enstitü / Fakülte yönetim kurulu gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ay ertelenmiştir.

Tezim ile ilgili gizlilik kararı verilmiştir.

19 / 09 / 2019


(İmza)

ÖĞRENCİNİN ADI-SOYADI
Komuran Nur KIRAZ

ABSTRACT

TOPIC MODEL BASED RECOMMENDATION SYSTEM TO IDENTIFY OPERATIONS THAT ARE MISSING IN THE TREATMENT

Kamuran Nur KİRAZ

Master of Science, Computer Engineering

Supervisor: Asst. Prof. Dr. Gönenç Ercan

September 2019, 79 pages

In the medical field, although it is extremely important and a legal obligation to record the procedures applied to patients by the health personnel, generally the operation lists are incomplete. Omissions in the operation lists can cause unexpected results for patients. In addition, inadmissible operation lists on billing operations applied to patients cause financial problems for both health institutions and patients because operation lists are used for invoicing process. Therefore, the main objective of this study is to develop an expert recommender system which can predict the omissions in the operation lists with a high success rate, which both threaten human health and cause economic problems for patients and medical centers. In this thesis study, we propose a new model different from the previous attempted solutions which tried to predict omissions in the operation lists using the Latent Dirichlet Allocation method, the proposed method uses the ICD-10 code as a new observed variable. The first experiments are carried out with Logistic Regression and Latent Dirichlet Allocation methods which had previously achieved success in this field. Precision, recall, F1 measure and MRR values are used as evaluation metrics, and the

results of the proposed model with the Logistic regression method and the classical Latent Dirichlet Allocation method are compared based on the evaluation metrics. According to the results of the experiments conducted on three different datasets, it is observed that the Proposed Method is 5% more successful than the LDA method and 13% more successful than the Logistic Regression method.

Keywords: Clinical order prediction, Topic models, Latent Dirichlet Allocation

ÖZET

TEDAVİDE EKSİK OLAN İŞLEMLERİ BELİRLEMEK İÇİN KONU MODELİNE DAYALI ÖNERİ SİSTEMİ

Kamuran Nur KİRAZ

Yüksek Lisans, Bilgisayar Mühendisliği

Tez Danışmanı: Dr. Öğretim Üyesi Gönenç ERCAN

Eylül 2019, 79 sayfa

Medikal alanda, sağlık personeli tarafından hastalara uygulanan operasyonları kayıt altında tutmak son derece önemli ve yasal bir zorunluluk olmasına rağmen, genellikle operasyon listeleri eksiktir. Operasyon listelerindeki ihmaller hastalar için beklenmeyen sonuçlara neden olabilmektedir. Ek olarak, operasyon listeleri faturalandırma sürecinde kullanıldığından, kabul edilemez operasyon listeleri hem sağlık kurumları hem de hastalar için finansal sorunlara neden olmaktadır. Bu nedenle, bu tez çalışmasının temel amacı, hem insan sağlığını tehdit eden hem de hastalar ve sağlık merkezleri için ekonomik sorunlara neden olan operasyon listelerindeki ihmalleri tahmin edebilen bir uzman tavsiye sistemi geliştirmektir. Bu tez çalışmasında, Latent Dirichlet Allocation yöntemini kullanarak operasyon listelerindeki ihmalleri tahmin etmeye çalışan önceki denenmiş çözümlerden farklı yeni bir model öneriyoruz, önerilen yöntem ICD-10 kodunu yeni gözlemlenen bir değişken olarak kullanıyor. İlk deneyler daha önce bu alanda başarı elde etmiş olan Lojistik Regresyon ve Latent Dirichlet Allocation yöntemleri ile

gerçekleştirilmiştir. Kesinlik, hatırlama, F1 ölçümü ve Ortalama Karşılıklı Sıra değerleri, değerlendirme ölçütleri olarak kullanılır ve önerilen modelin, Lojistik Regresyon ve klasik Latent Dirichlet Allocation yöntemiyle sonuçları değerlendirme ölçütlerine göre karşılaştırılır. Üç farklı veri setinde yapılan deney sonuçlarına göre, önerilen yöntemin LDA yönteminden %5, Lojistik Regresyon yönteminden %13 daha başarılı olduğu tespit edilmiştir.

Anahtar Kelimeler: Klinik düzen tahmini, Konu modelleri, Gizli Dirichlet Tahsisi

ACKNOWLEDGEMENTS

I would like to thank my first supervisor Asst. Prof. Dr. Gönenc Ercan (Hacettepe University) for his help, advice, and patience during this thesis. He has always been very supportive, informative and I have learned a lot from him. Without his supervision the achievements in this thesis would not be possible.

Also, I would like to thank my family and husband for their continuous and unparalleled love, help, support and patience.

Kamuran Nur KİRAZ

September 2019, Ankara

TABLE OF CONTENTS

ABSTRACT	i
ÖZET	iii
ACKNOWLEDGEMENTS	v
ABBREVIATIONS	ix
Table of Contents	vii
1. INTRODUCTION	1
1.1 Overview and Motivation	1
1.2 Contribution of the Thesis	3
1.3 Thesis Structure	3
2. RECOMMENDER SYSTEMS	5
2.1 Content Based Filtering	6
2.2 Collaborative Filtering	6
2.3.1 Memory Based Collaborative Filtering	7
2.3.2 Model Based Collaborative Filtering	9
2.3 Regression Models	11
2.3.1 Linear Regresion	11
2.3.2 Logistic Regression	12
3. PROBABILISTIC TOPIC MODELS	14
3.1 Latent Semantic Analysis	16
3.2 Probabilistic Latent Semantic Analysis	17
3.3 Latent Dirichlet Allocation	18
3.3.1 Inference and Parameter Estimation	24
3.4 Related Works	25
4. MEDICAL DATA	28
4.1 Data Understanding	30
4.3.1 K-Fold Cross Validation	34
5. PROPOSED MODEL	35
6. EXPERIMENTS	40
6.1 Data Preperation	40
6.2 Logistic Regression	42
6.3 Latent Dirichlet Allocation	43

6.4	Proposed Model.....	49
7.	RESULTS AND EVALUATION.....	52
7.1	Evaluation Metrics.....	52
7.2	Wilcoxon Signed-Ranks Test.....	53
7.3	Experimental Results and Evaluation.....	54
8.	CONCLUSION AND FUTURE WORK.....	64
9.	APPENDICES.....	65
	REFERENCES.....	80

ABBREVIATIONS

RS	Recommender System
CF	Collaborative Filtering
LR	Logistic Regression
LSA	Latent Semantic Analysis
PLSA	Probabilistic Latent Semantic Analysis
LDA	Latent Dirichlet Allocation
EMRs	Electronic Medical Records
ICD	International Classification of Diseases
WHO	World Health Organization

1. INTRODUCTION

1.1 Overview and Motivation

In today's society, intelligent systems that have become an indispensable part of our lives, serve humanity in many areas and industries. These intelligent systems are computer programs that are developed to automatically implement the activities of human experts in design, planning, diagnostics, summarizing, classification, controlling and recommendation. These systems can store specialist knowledge within a limited field and offer solutions by following the logical results. One of the first developed intelligent systems is the MYCIN [28] program which can diagnose some diseases. This system is developed to advise physicians to select appropriate treatment for patients with bacterial infections and uses clinical decision criteria [28]. In order to see the overall scope of intelligent systems, we can summarize the today's usage areas of these systems as follows:

- Control systems such as heavy industry, household appliances, automobiles, spacecraft
- Intelligent sensors such as high reliability for aircraft engines, barcode control
- Load systems such as economic load distribution, optimization and loss reduction, fault detection and forecasting, load estimation
- Military systems such as discovery, target information collection, automatic target recognition, fire control, navigation, manipulation
- Medicine such as classification, application, diagnosis, measurement analysis, medical image processing, organ analysis, artificial organ design, prediction of protein structures, DNA, gene, chromosome and cell analysis
- Finance such as risk profile analysis, assessment of appropriateness to credit, insurance fraud detection, stock market analysis, stock estimate
- Robotic

As it is seen, the application area of the intelligent systems is very wide but in this thesis, a study covering the fields of medicine and medical finance has been carried out. We have studied on a single special subject and the subject of our study is based on medical operations, procedures or medications applied to patients and diagnostic codes for patients.

In the medical field, it is a legal obligation to record the operations performed by the health personnel in the relevant patient file. There are main reasons why recording of the procedures applied to patients is extremely important in the medical field and we can list some of them as follows:

- Records are required to ensure and improve the safety and quality of the patient care and treatment process.
- There is such a requirement to protect patients and employees from potential risks and damages they may encounter during the health service delivery process.
- In order for the patients to be diagnosed correctly, the health personnel need all the information about the patient.
- In order to ensure safe surgical and drug applications, the recorded data should be provided to the personnel who performs the application.
- Patient files which includes procedures applied to patients are very important for improving communication security among healthcare providers. Because, it is important to transfer all the information about the patients to the health personnel who have seizures when changing the seizures.

Although, it is of utmost importance to record the procedures applied to patients, generally the operation lists are incomplete. Such omissions may cause serious or even vital consequences for health. For example, if an operation that should not be applied more than once is not included in the operation list, it can be reapplied. Conditions that may occur in this way may cause serious or even vital consequences for health.

In addition, inadmissible operation lists on billing operations applied to patients cause financial problems for both health institutions and patients since they are used for billing processes. There are several studies in the medical field related to erroneous billing results. Addressing medical coding and billing study [37] emphasizes the effect of erroneous billing on the rising costs of health care in the United States. According to the study, for inadequate medical records, some of the high risk areas are as follows [37]:

- Billing of services or products that are not actually provided
- Double billing for the same service or products

- Billing of services not provided
- Misuse of provider identification numbers

Another study in this area, “A Reliable Billing Method for Internal Medicine Resident Clinics: Financial Implications for an Academic Medical Center” [36], emphasizes the importance of the financial success of medical centers to appropriate coding and billing and consequently, the adequate recording of the procedures applied to patients. From the point of the health institutions, to be on the safe side in terms of accounting, they need to make sure that all expenses must be billed in detail. At this point, tremendous workload is required to be carried out manually by the accounting department of medical centers. Because, people working in this department have a responsibility to guarantee the accuracy of invoices provided to patients and all parties involved. In order to be sure of the accuracy of a bill, it is necessary to review all the electronically recorded data and to compare these data with those that are not electronically recorded, such as doctor's visitation notes.

Consequently, for a solution to all these problems, the main motivation of this research is to develop an intelligent and expert recommender system that can predict with a high success rate the omissions in the operation lists, which both threaten human health and cause economic problems for patients and medical centers.

1.2 Contribution of the Thesis

This thesis study's contributions can be summarized in two caption:

- We propose an extended LDA topic model with a new observable variable to predict the omissions in the operation lists with a high success rate.
- We compare the proposed method and the different methods previously achieved success in this field.

1.3 Thesis Structure

Section 1 gives an overview of the thesis study. It defines the research problem and its importance so presents the main motivation of the study. In this section, after defining the problem and motivation, the contributions of the thesis are listed in the different heading. This section also includes the structure of the whole thesis.

Section 2 includes general information about the Recommender Systems and methods. These methods are Content Based Filtering, Collaborative Filtering and Regression Models.

Section 3 gives an overview of the Probabilistic Topic Models, includes detailed information about the probabilistic topic model algorithms and underline the importance of them. In addition, in the last part of this section, the relates studies are summarized.

Section 4 define medical data and explain the features and structure of the medical data. The most important content of this section is to define the ICD codes used in the proposed model. In addition, the structure of the medical data used in the experiments is explained in detail in this section.

Section 5 contains a detailed description of the Proposed Model.

Section 6 includes the data preparation part for experiments and describes the experiments carried out in three different methods. These methods are respectively Logistic Regression, Latent Dirichlet Allocation and the Proposed Model.

Section 7 explains the results of experiments performed in section 7. In this section, the results of three methods are compared according to the evaluation metrics and Wilcoxon Signed-Ranks test.

Section 8 concludes the thesis and underlines the importance and findings of the study. Moreover, in this section includes some information about the possible future works to improve the study.

2. RECOMMENDER SYSTEMS

Since the existence of humanity, people have had to decide every day by being exposed to various options. However, decision-making is one of the most difficult things in life. In addition to being emotions and non-objective, it is an act that requires knowledge of historical data, events, processes or patterns that humanity has faced since its existence. For this reasons, people have always needed suggestions from others or advices from experts to make a decision. But, the source of suggestion or assistance also comes from other human beings and naturally it is limited ad insufficient. Computer-based suggestion systems allow people to expand their recommendations from people. They can store huge amounts of knowledge and apply data analysis techniques so they enable data mining from historical data and discover patterns, potentially providing finely-tuned personalization [18].

Therefore, recommender systems have been the focus of researchers for many years and the root of the recommender systems are based on the information retrieval studies.

Recommender systems can be defined as programs that aim to present the most appropriate items for specific users by estimating the elements that users can be interested in according to the interactions between each other and the elements in the system [19]. They can be developed mainly using collaborative filtering, content-based filtering and hybrid methods.

Collaborative filtering are described in the Section 2.1 and content-based filtering are described in the Section 2.2. Hybrid methods are used to combine these methods to eliminate the disadvantages of them.

Although this thesis does not focus directly on a recommendation system, recommendatin has a place in the study because the thesis' main purpose is to determine the missing and inaccurate items and suggest the findings for further examination by medical professionals.

2.1 Content Based Filtering

Content-based filtering basically uses similarities between the features of items to make a recommendation. In such systems, new items with features common to the user's past preferred items are suggested to the users. This method generates a profile of users by analyzing documents or items evaluated by the users. That is, it uses the past preferences of users.

The resulting profile is considered an example of user interests. Figure 2.1 shows an example of the matrix in which the similarities between the items are maintained.

	Item1	Item2	Item3	Item4	Item5	...
Item1	10	3	1	8	9	
Item2	3	10	8	1	2	
Item3	1	8	10	2	3	
Item4	8	1	2	10	9	
Item5	9	2	3	9	10	
...						

Figure 2.1 Similarity matrix between items

Content based filtering methods provide recommendations on an individual basis, ignoring other users' evaluations. Therefore, content based filtering does not require a large user community or database of evaluation points [20].

2.2 Collaborative Filtering

Collaborative filtering (CF) is an essential and commonly used recommendation algorithm. This basic algorithm is used in a wide variety of areas in our daily life. It is widely used in popular social networks such as Instagram, Twitter, Facebook and LinkedIn.

Marketing and advertising are the other fields where CF is used most widely and Amazon is one of the largest online marketing companies making suggestions using this algorithm. In addition, Netflix, the streaming service that has achieved successful original productions, has a strong share in its sector because of its strong infrastructure and advanced content suggestion system [39]. On the basis of the recommendation systems they use CF [38].

Collaborative filtering is the name given to all of the various techniques used for processing information or people's behaviors in order to make predictions about new information or people's behavior. CF is basically based on the similarity and the similarity can be between users or items and it is calculated according to past evaluations.

Memory and Model Based Collaborative Filtering are two approaches.

2.3.1 Memory Based Collaborative Filtering

Memory Based Collaborative Filtering methods compute the similarities between users or items named as neighbours and store these similarities. There are two types of Memory Based Collaborative Filtering algorithms. When the similarity bases on the user, it is named user based collaborative filtering, when the similarity bases on the items, it is referred to as item based collaborative filtering.

User Based Collaborative Filtering

User based collaborative filtering aims to find similar users and reach similar content among the content that users follow or like. This method is based on presenting the items that other users with similar qualities to the active user have liked in the past. Predictions are made using the user's similar preferences with other people. User-based CF uses the assumption that similar users love similar items [23]. Figure 2.2 shows an example of the user based rating matrix.

	Item1	Item2	Item3	Item4	Item5	...
User1	5	3	0	2	1	
User2	8	1	10	3	2	
User3	0	2	0	9	8	
User4	4	2	1	7	6	
...						

Figure 2.2 User – Based Rating Matrix

Item Based Collaborative Filtering

When using item based collaborative filtering methods making suggestions to a user, first of all the content that the user has voted the most is searched and suggestions or predictions are made to the user by finding the contents that are closest to them. According to the item based algorithms similar users like generally similar items [23]. Figure 2.3 shows an example of transposition of user-based matrix in Figure 2.2 to item-based matrix.

	User1	User2	User3	User4	...
User1	5	8	0	4	
User2	3	1	2	2	
User3	0	10	0	1	
User4	2	3	9	7	
User5	1	2	8	6	
...					

Figure 2.3 Item – Based Rating Matrix

As it is seen, the essential part of both methods is to compute the similarities. There are some measures of similarity for instance Cosine, Adjusted Cosine, Pearson Correlation, Mean Squared and Euclidean Distance.

Cosine distance, which is one of the most basic similarity distance, is used to find the similarity between vectors and the formulation (2.1) is as follows:

$$sim(a, b) = \cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|^2 * \|\vec{b}\|^2} \quad (2.1)$$

There are some potential challenges for memory based collaborative filtering, these are sparsity, scalability and some performance problems. Recommender systems which are memory based use large user-item sets, but generally the items that is matched with users are limited. Therefore, data sparsity may effect the accuracy. Huge user-item sets cause serious scalability problems.

Moreover, serious performance problems may occur as the number of users and elements increases in memory and the diversity on the basis of the user-item causes serious problems in finding the nearest neighbor.

2.3.2 Model Based Collaborative Filtering

Model Based Collaborative Filtering algorithms firstly learn a model offline using different machine learning algorithms and then this model provides item prediction online according to user ratings. These algorithms adopt a probabilistic approach and foresee the collaborative filtering process as calculating the expected value of a user estimate, taking into account the ratings of users on other elements [22]. When machine learning algorithms are involved, it is necessary to mention about two groups where machine learning is fundamentally separates; supervised and unsupervised learning. It is important to understand these learning algorithms and their distinctions to choose which method is more accurate for the problem.

- **Supervised Learning:** In this learning technique, the data which is used for learning is determined which output is produced for which input. Therefore, all training data has labels and a matching function is generated between the inputs and the labels of them. This function can be created by classification and regression algorithms [40].
- **Unsupervised Learning:** In this learning technique, a function is used to estimate an unknown structure over unlabelled data. It is not known which class the input data belongs to, or even what constitutes a class. This technique, aims to capture and model the patterns or hidden relationships in the dataset.

After mentioning this fundamental distinction clustering, classification and latent models are three common approaches for model based collaborative filtering.

- **Clustering :** The basis assumption of clustering is that similar users have same interests so they rate items similarly. The clustering model works by grouping similar users in the similar clusters and predicting the likelihood of a specific user in a specific class and calculating the conditional probability of rating from there [22].

Distance measures are used to find similar items such as Minkowski distance, Euclidian distance, Manhattan distance [25] and the formulas (2.2, 2.3, 2.4) of these distance measurements are as follows:

$$distance_{Minkowski}(u_1, u_2) = \sqrt[q]{\sum_j (r_{1j} - r_{2j})^q} \quad (2.2)$$

$$distance_{Euclidian}(u_1, u_2) = \sqrt{\sum_j (r_{1j} - r_{2j})^2} \quad (2.3)$$

$$distance_{Manhattan}(u_1, u_2) = \sum_j |r_{\{1j\}} - r_{\{2j\}}| \quad (2.4)$$

There are some popular clustering algorithms which are used commonly such as k-means and k-centroid [41].

- **Classification:** Classification learn a model from the dataset in which the class assignments are known. The goal of this approach is to predict the target class for new items using the learned model. The most commonly used technique is Bayesian Method. For the collaborative filtering problem, the Bayesian model describes that how to do learning and prediction about hypotheses from data with a probabilistic model. Mathematically Bayes' Theorem (2.5):

$$P(Rare|Pattern) = \frac{P(Pattern|Rare)P(Rare)}{P(Pattern)} \quad (2.5)$$

- **Latent Models:** This approach bases on the observation of coexistence of the user item pairs. The learning for the data set is performed indirectly through a hidden variable which is not included in the data set and it is mainly based on the probability.

Singular Value Decomposition, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation and Markov Chain decision process are some latent models. These models are discussed in detail in the Section 3.

2.3 Regression Models

When conducting scientific studies, it is necessary to evaluate the methods and approaches that have previously been successful in the particular field of interest and Logistic Regression method [10] is one of them for our study.

In data science, regression analysis is important approach for analyzing and modelling data. This technique is used to develop predictive models. Regression analysis investigates the relationship between variables, one of them is dependent variable which is named as target or criterion or class variable and the all others are independent which are named as predictor variables. In this way, the model contains multiple independent input variables and called Multi Variable Linear Regression. The other model which contains a single independent variable that offers a solution to fewer problems, is called the Single Variable Linear Regression.

The independent variables, the relationship between them and also their changes are used to estimate the dependent variable. For estimation, regression models need a mathematical equation which describes the target variable as a function of the predictor variables. Therefore, regression models are supervised algorithms and mainly, there are two main regression algorithms, namely linear regression and logistic regression.

2.3.1 Linear Regresion

Linear Regression model structure requires is that dependent variable is continuous and independent variable(s) can be continuous or discrete. The mathematical model which describes the relationship between a dependent and independent variable(s) is linear and its graphical representation is known as a straight line or regression line. The mathematical formulation (2.6) of straight line with one independent variabele is as follows:

$$y = a + b * x + e \tag{2.6}$$

where a is intercept, b is weight or coefficient which is related to independent variable x , y is target variable and e is error term. If there are more than one independent variable, formulation (2.7) changes as follows:

$$y = a + b * x_1 + c * x_2 + e \quad (2.7)$$

2.3.2 Logistic Regression

Logistic Regression is a powerful supervised classification algorithm in the machine learning field that is used for classification problems. This algorithm predicts dependent response variable using independent variable(s). The dependent variable mentioned here is the categorical target variable that is tried to predict. If the target variable has just two different values which are 0 and 1, this refers to the Binomial Regression model.

When the target variable has more than two different values, this refers to the Multinomial Regression model. The independent variables which are features or attributes and their relationships are used for prediction of dependent response variables.

Logistic function, also called the sigmoid function is the basis of logistic regression method, it refers infinite values as absolute values within a finite limit which is generally range of 0 to 1. Logistic function gives “S” shaped curve which is also named as sigmoid curve, mathematical function (2.8) and curve is showed as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2.8)$$

, where e is the Euler’s number and x is the value which is tranformed into the range 0 and 1.

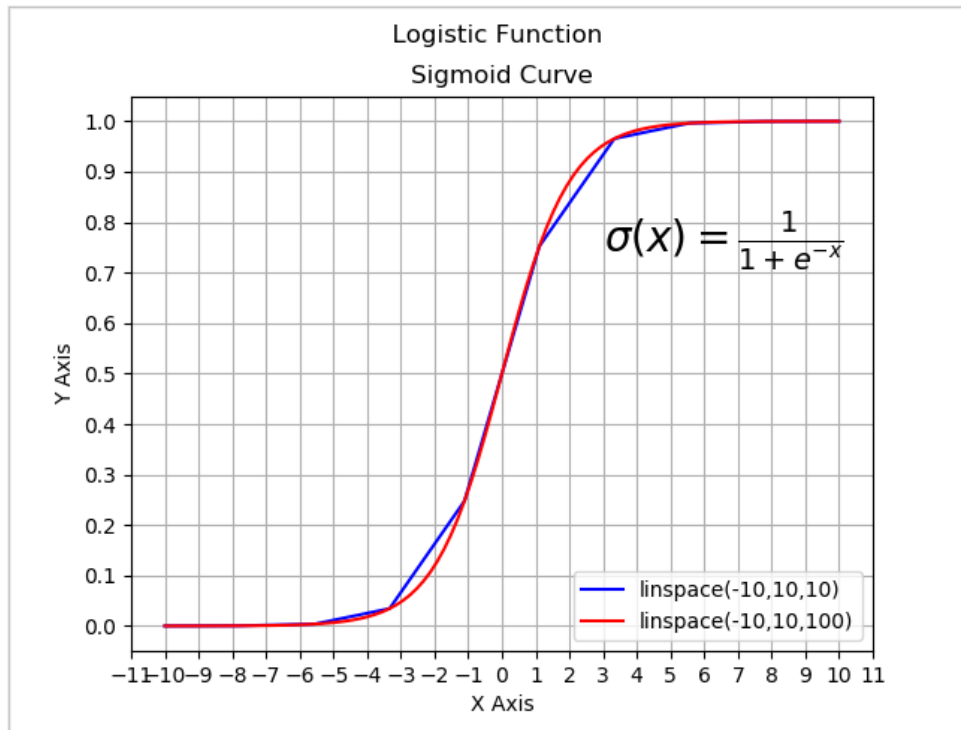


Figure 2.4 Sigmoid curve is the representation of Logistic Function

As can be seen from the graph, the minimum value that can be the output of this function is 0 and the maximum value is 1. When the value of x is 0, the output of the sigmoid function is 0.5, so for classification problems, if the output is greater than 0.5, the result can be evaluated as 1 and if it is small, the result can be evaluated as 0. Consequently, when using the logistic regression method to make predictions, we can calculate the probability of realization (or probability of being 1) and the probability of not being real (probability of being 0) for a case.

3. PROBABILISTIC TOPIC MODELS

With today's fast developing technology, we are faced with abundant data in many areas. In our daily lives in society, everyone is producing data with incredible speed especially in the web, social networks, e-mails, personal online conversations also government issues and digital archives.

Therefore, there are so many large scale scientific experiments, medical and climate data, electronic trading and advertising. So, there is a need for organizing, analyzing, modelling and understanding the huge mass of information we have today.

It is becoming increasingly important to transform this information into processable data and to obtain and organize the information we aim for. In this context, probabilistic modeling is one of the important recurring subjects we use to make sense of the data and it is a wide field which is covering topic models to be explained in detail after this section.

Mathematical and statistical models define and express the exact relationships between variables, such models are called deterministic models. However, in some cases relationship between variables can not be expressed exactly and a probabilistic component can be built in the model to express uncertainty, such models are called probabilistic models. They are based on the probability theory and because of that randomness is essential for predicting new events. Deterministic models produce one possible result for an event but probabilistic models produce a probability distribution over variables as a result.

Probabilistic models can be used in many fields for example machine learning, data mining, pattern recognition and etc.

Probabilistic topic models are used to analyze the content of documents and they can be adapted to many kinds of data such as text, images, genetic data and social networks [7]. In text mining field which is a sub-branch of machine learning, everything which is recorded in writing can be handled with a topic model as collections of documents such as newspaper articles, novels, medical data, twitter posts or blog posts and more.

Topic modeling is a type of statistical model for extracting the hidden topics which generate a document. The basis idea of topic modeling is that semantics in a document is governed by a variable that we cannot observe and called as a latent variable. Since this point is important for the topic model, it would be appropriate to mention about latent and observed variables in detail.

Latent variables are supposed to be within the structure of the model but not directly observed by means of mathematical model or by its side effects this means that latent variables are not directly measurable. Basically, the main idea on which latent variables are based, they are unobserved variables which are assumed to explain observed events. And observed variables basically can be considered as recorded and measured variables and they actually exist in data files. Therefore, a statistical model can reveal a direct relationship between observed variables. In this case, we need a common cause to explain this correlation. Latent variables explain this direct causal correlation between them. For example, suppose that our dataset includes ice cream consumption and air conditioner usage as observed variables. At this stage, temperature can be thought as an unobserved variable because it can explain the causal relationship. Generally, for topic models, one latent variable is used and is called “topic”. Because, one of the useful method to obtain beneficial information from a document is analyzing its topics. So, topic modeling can also be defined as a process of learning, identifying and extracting topics of a document.

As a result, the main purpose of topic models is to extract these hidden variables which shape the meaning of documents and all topic models bases on the following two assumptions:

1. First assumption is that each document is described as a statistical mixture of topics, each topic is a distribution over words and each word is underlined from one of those topics. The differences of topic models arise from the differences in the generative process of the models.
2. The second assumption is ‘Bag-of-words’. In view of to this assumption, the count of the words that generate a document is important for modeling however the order of words are ignored and not utilised to create a topic model. This assumption is based on the language of the probability theory [2] which suppose that words are exchangeable in

a document and documents in the corpus are also exchangeable. Therefore, words are not represented according to their sequences within the sentence or in the document so models can be considered as context-free.

Probabilistic topic models can be used by adapting to different domains. Medical field is one such domain.

3.1 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a statistical and unsupervised method that is used to determine the semantic relationship between words and sentences using the information of the usage of the related terms in the context [31]. Singular Value Decomposition (SVD) is used to find out the semantic relationship between words and sentences in the LSA method. SVD is one of the types of factorization of a matrix and this method is used to model semantic relationships. Latent Semantic Analysis algorithms usually contain three main steps. These steps are as follows:

1. Input Matrix Creation : An input document is showed by a matrix which generally rows are word vectors and columns are sentence vectors. In the first step, each cell of input matrix is filled with term frequency values of words for each sentences. Cell values show the importance of the words in sentences. Input matrix is sparse because not every word is used in every sentence. Representation of the word importances can be integrated using different weighting functions instead of raw term frequencies such as TF-IDF (term frequency–inverse document frequency), log entropy, root type or modified TF-IDF. Every approach which is used to create input matrix has an impact on the end result of the LSA method.
2. Singular Value Decomposition : SVD is a mathematical model which can correlate sentences and words. SVD decompose the input term document matrix created in the first step into three new matrices. Let the input document term matrix of size $m \times n$ be A , then SVD of A is defined as:

$$A = U\Sigma V^T ,$$

where Σ is a $n \times n$ diagonal matrix, V is an $n \times n$ orthonormal matrix and U is an $m \times n$ column orthonormal matrix.

3. Sentence Selection: In this step, various different algorithms can be used to select sentences according to the problem.

3.2 Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) is a new approach for document indexing from the category of topic models and it is a statistical latent variable model for co-occurrence data under a probabilistic framework. PLSA is based on the aspect model and it is based on the maximum likelihood principle.

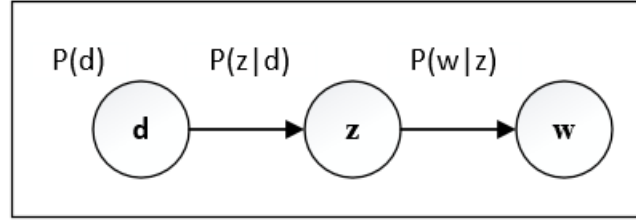


Figure 3.1 Graphical representation of PLSA

- Document is indicated as $d \in D = \{d_1, \dots, d_m\}$
- Word is indicated as $w \in W = \{w_1, \dots, w_n\}$
- Class variable is indicated as $z \in Z = \{z_1, \dots, z_k\}$

The likelihood function for each document-word can be showed as follows [32]:

$$(d, w) = P(d)P(w|d) = P(d) \sum_z^Z P(w|z)P(z|d) \quad (3.1)$$

According to the maximum likelihood principle $P(d)$, $P(z|d)$, $P(w|z)$ are calculated by the maximization of log likelihood function (3.2) [32]:

$$L = \sum_d^D \sum_w^W n(d, w) \log P(d, w) \quad (3.2)$$

Expectation Maximization (EM) algorithm prefers to use predictive criteria instead of using exact distance criteria to determine which cluster an object belongs to. EM based on the principle of maximum similarity and consists of two iterative steps. These steps are expectation (E-Step) and maximization (M-step). By estimating the parameters of the data observed in the E-Step, the best probabilities of the hidden variable are estimated. In step M, the predicted data is replaced and the maximum likelihood is calculated over the whole data to obtain new estimates of the parameters. Re-parameterized is done by using Baye's rule (3.3) [32]:

$$P(d, w) = \sum_z P(z) P(d|z) P(w|z) \quad (3.3)$$

In the expectation step, the following equation (3.4) is updated [32]:

$$P(z|d, w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')} \quad (3.4)$$

In the step M, the parameters are updated using the following equations to maximize the likelihood (3.5, 3.6, 3.7) [32]:

$$P(w|z) = \frac{\sum_d n(d, w)P(z|d, w)}{\sum_{d, w'} n(d, w')P(z|d, w')} \quad (3.5)$$

$$P(d|z) = \frac{\sum_w n(d, w)P(z|d, w)}{\sum_{d', w} n(d', w)P(z|d', w)} \quad (3.6)$$

$$P(d|z) = \frac{\sum_{d, w} n(d, w)P(z|d, w)}{\sum_{d, w} n(d, w)} \quad (3.7)$$

3.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [1] is a completely unsupervised probabilistic topic algorithm for modeling large unstructured text corpus. LDA models each input document as a mixture of hidden topics by the reason of that it allows words to be marked with different topics. Of course, this model is based on the bag-of-words assumption which describes in the Section 3.

This means that LDA describes how many percentage of information is in the document for each latent topic. In this way, it can be used in some tasks such as classification, summarization or finding similarity.

Generative model of LDA discovers latent topics for documents. According to the generative model, each document has a multinomial mixture over latent topics and each topic is defined by a multinomial distribution over words [3]. This model can be described by the following mathematical expressions.

1. Each distinct word (w) which is in the corpus has a probability for each latent topic (z) and it's mathematical expression (3.8) is as follows:

$$P(w|z) \tag{3.8}$$

The word-topic relationship refers to similarity between words this means that the words of the same topic can be assumed similar words.

Also, the discrete probability distribution over words for each topic can be used for document summarization because LDA assumes that each word is generated from one underlying topic so they can be grouped according to these top topics.

Due to the fact that each word has a probability for each topic, each topic can be represented as a multinomial distribution over words, represented by phi (ϕ) and it's definition (3.9) is as follows:

$$\sum_w^W p(w|z) = 1 \tag{3.9}$$

2. Depending on the same logic, each document (d) in the corpus has a probability for each hidden topic (z) and it's mathematical expression (3.10) is as follows:

$$P(d|z) \tag{3.10}$$

The discrete probability distribution over documents for each topic can be used to determine the similar documents. In addition, finding the underlying topic of documents is very important for summarization and other tasks of text mining.

Due to the fact that there are words tagged for different hidden topics in a document, each document has a multinomial mixture over these topics, represented by theta (θ) and the formulation of this expression (3.11) is as follows:

$$\sum_z p(d|z) = 1 \quad (3.11)$$

The multinomial distributions θ and ϕ have a symmetric Dirichlet prior with hyperparameters alpha (α) and beta (β) respectively [4]. Alpha (α) is related to the document-topic density, with higher alpha value documents are generated with more topics. As the value of alpha becomes smaller, the more uniform topics begin to disperse for each document therefore with lower alpha, LDA can discover fewer latent topics that generate a document.

The effect of the α :

- $\alpha < 1,0$ means that documents consist of few topics.
- $\alpha > 1,0$ means that documents consist of many topics.

Beta (β) is related to the word-topic density, high beta value means that each topic is generated with most of the words in the corpora and the same way lower beta value means that topics indicate fewer words of the corpora.

It is important to know some notes before the generative process steps. Firstly, assume that there are K latent topics that generate the document and M different documents in the corpus and documents consist of multiple topics. Also, words are generated independently from other words, this is based on the bag-of-word assumption.

The generative process of Latent Dirichlet Allocation for each N words $W = \{W_1 \dots W_n\}$ from a document is as follows:

1. Randomly sample a multinomial distribution θ_i over topics from a Dirichlet (α) distribution (where $i = 1, \dots, M$)
 - $\theta_i \geq 0, \sum_i \theta_i = 1$ (where $i \in \{1 \dots M\}$)
 - $\theta_{i,k} =$ probability of document $i \in \{1 \dots M\}$ has topic $k \in \{1 \dots K\}$

2. For each words W_n in the document

2.1 Sample a random topic Z_n , $n \in \{1..k\}$ from a multinomial distribution over topics

$$P(Z_n = i | \theta) = \theta_i$$

2.2 Sample a random word W_n from the multinomial distribution $P(W_n | Z_n, \beta) = \phi$, a multinomial conditioned on the corresponding topic Z_n .

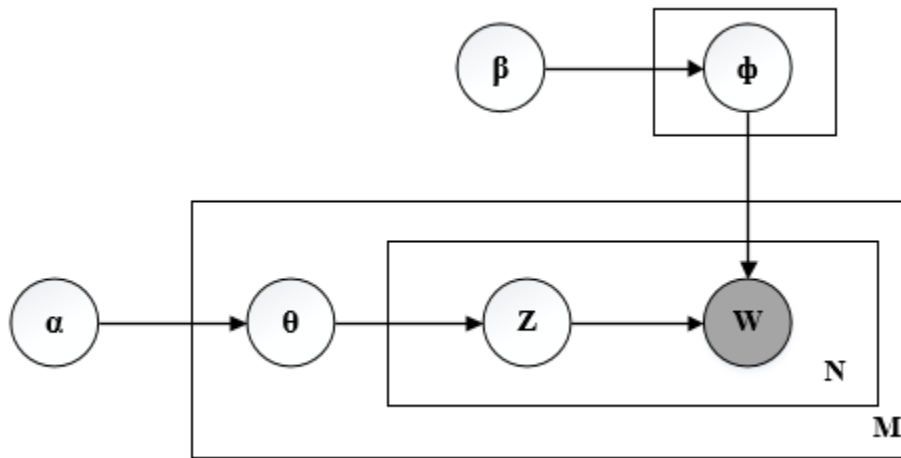


Figure 3.2 Graphical Representation of LDA

Figure 3.2 shows the plate diagram of the LDA graphical model. The boxes represent repeats and stand for plates. The big plate indicate documents in the corpus and its inner plate represents words in a document together with the topic assignments so this plate means document topic assignments.

The small plate on the down side of Figure 3.2 represents the topic assignments for each distinct word in the corpus.

The dark node in Figure 3.2 means that the variable expressed by that node is the observed variable. The concept of observed and latent variable was explained in detail under Section 3. The observed variable in the LDA model is the word and there is only one latent variable which is the topic in the basic model of the LDA.

N	Word count
D	Document count
K	Topic count
α	Distribution of topics in a document.
θ	A dirichlet prior that represents topic document distribution as K dimensional vector.
Z	N dimensional vector represents topic assignments for document.
W	Vocabulary for all documents in a corpus.
ϕ	Probability distribution of words over topics.
β	Distribution over the vocabulary.

Figure 3.3 Definition of symbols used in LDA's graphical model representation.

According to the LDA's graphical model representation, there are three levels for modeling documents:

1. First level is corpus level and hyperparameters α and β are corpus level parameters. The reason for being first level is that they are specified before start the generative process of LDA.
2. Secondly, θ and ϕ are document level parameters, they are sampled once for every documents in the corpus.
3. Third level is word-level and its parameters are Z and W. W already expresses words and Z is generated for each word of all documents in the corpus.

One of things to note about the LDA generative process is that each hidden topic has a different probability of generating each word and the word probabilities are kept in a $k \times N$ matrix.

Secondly, k latent topics indicates k dimensional Dirichlet distribution so dimensionality of Dirichlet is specified firstly and it is fixed. In this context, θ is the k dimensional Dirichlet random variable and it can take values between 1 and k . This means that if, $\theta_i \geq 0, \sum_{i=1}^k \theta_i = 1$ probability density is formulated (3.12) as follows [1] :

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1} \quad (3.12)$$

This formulation is called as Dirichlet Distribution with parameters α and $\Gamma()$ is the Gamma distribution. Generally, LDA has a Symmetric Dirichlet hyperparameter where all the α are equal.

Given the hyperparameters α and β , the joint distribution of a topic mixture θ , a set of N topics \mathbf{z} , and a set of N words \mathbf{w} is given by [1]:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^N p(Z_n|\theta)p(w_n|z_n, \beta) \quad (3.13)$$

, where $p(Z_n | \theta)$ is simply θ_i for the unique i and integrating over θ and summing over z , the marginal distribution of a document is formulated (3.14) as follows [1]:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^N \sum_{Z_n} p(Z_n|\theta)p(W_n|Z_n, \beta) \right) d\theta \quad (3.14)$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus [1]:

$$p(D|\alpha, \beta) = \prod_{d=1}^M \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{Z_{dn}} p(Z_{dn}|\theta_d)p(W_{dn}|Z_{dn}, \beta) \right) d\theta_d \quad (3.15)$$

In Section 3, exchangeability theory and bag-of-words assumption is mentioned and LDA is based on them. In LDA, there is another basis assumption, words are generated by topics and all words are exchangeable in a document.

3.3.1 Inference and Parameter Estimation

The generative process of LDA defines a joint probability distribution over the latent and observed variables. Joint probability is used to perform data analysis to calculate the conditional distribution of hidden variable, considering the observed variables, and this conditional distribution is also called the posterior distribution [7]. However calculating the posterior distribution of the latent variables is a key inferential problem and formulated (3.17) as below [1]:

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \quad (3.16)$$

This formulation can be marginalized over the hidden variables to obtain a normalized distribution [1] :

$$p(w|\alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^k \theta_i^{\alpha_i-1} \right) \left(\prod_{n=1}^N \sum_{i=1}^k \prod_{j=1}^V (\theta_i \beta_{ij})^{w_n^j} \right) d\theta \quad (3.17)$$

Unfortunately, using this estimation method is untractable and messy work so other approximate parameter estimation algorithms such as Gibbs Sampling or Variational Inference are needed.

Gibbs Sampling, is an example of the Markov chain monte carlo algorithm, is presented as a parameter estimation method for LDA by Griffiths and Steyvers in 2014. Markov chain Monte Carlo algorithm is used to sample the posterior distribution over the parameters [4]. Instead of estimating the model parameters directly, we only evaluate the posterior distribution on just document and topic and then use the results to infer θ and ϕ [4].

3.4 Related Works

Since the health field contains many differences on an individual basis, it is very troublesome to generalize and make inferences based on acquired medical data in this context. In addition, recording and storing electronic data is very difficult due to the fact that health personnel are directly focused on human health. However, there are studies conducted with the limited medical data that we have, and this area is attracting interest from researchers. In this section, we briefly overview some medical studies related to the problem that this thesis is trying to provide a solution for and some of the studies also related the techniques which are used in this thesis to solve the problem.

“Towards a Collaborative Filtering Approach to Medication Reconciliation” [11] is a study conducted by Hasan, Duncan and Padman which aims to develop some techniques for automatic detection of omissions in medication lists, identifying drugs that are forgotten and incomplete while prescribing to the patient. In this paper, they focus on the verification step in the medication reconciliation process. The authors use five computational and statistical methods of collaborative filtering for the problem of medication reconciliation and these are Drug Popularity, Co-occurrence counting, K-Nearest Neighbors (KNN), Logistic Regression, Drugs with Regularization and Random. At the end of the study, they present the results of these methods comparatively, they have achieved the best result with the logistic regression method.

Doctor AI is published by Choi, Bhadori, Schuetz, Stewart and Sun as an intelligent clinical decision support system to predict clinical events via recurrent neural networks (RNN) [14]. They use large historical data in electronic health record (EHR [9]) and there are two main contributions. Firstly, their study demonstrates that RNNs usage for representing patient status and prediction for new diagnosis, medication and visit time. Secondly, they improve the performance of the RNN in both accuracy and speed using Skip-Gram embedding.

Gong & Liu [33] summarization algorithm uses LSA method to identify semantically important sentences for extraction. The algorithm applies the first two steps of LSA and after performing SVD, V^T matrix is used for sentence selection. Rows of V^T are topic vectors and columns are sentence vector. The first row is the most important topic so firstly a sentence is selected which is the most related to the first topic, then the same process continues with the other topics in order until the summary reaches the expected size.

“Generic text summarization using probabilistic latent semantic indexing” [34] study presented a strategy to generate extractive summary of document using PLSA with two different approaches to select sentences. The first of them is Document Topic Only approach, it finds the main topic of input document and selects summary sentences which are related the main topic. In this approach, term frequency matrix can be used or it can be implemented as a graph based approach with sentence similarity matrix instead of term frequency matrix. The other approach is Multiple Topic and it presents different way for sentence selection. This approach takes the advantage of the fact that PLSA divides the input document into different topics and it selects the most important sentences belonging to different topics as summary sentences.

“Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets” is one of the important studies which is conducted by Chen, Goldstein, Asch, Mackey and Altman [8]. The objective of this investigation is to develop a LDA model for hospital admissions, learn clinical order patterns and compare the prediction ability of this model according to the preconstructed order sets. The authors develop this model on the first 24 hours of structured electronic health records (EHRs [9]) which contains several types of data and they use Gibbs Sampling to predict the preconstructed order sets. As a result, their probabilistic model provides a clinical decision support for hospital admissions with the precision 16%, and recall 35% of success.

An extremely increasing amount of data on the web we are facing today is a problem that needs to be overcome in the medical field. “Incorporating Statistical Topic Models in the Retrieval of Healthcare Documents” is developed as a solution to this problem by Caballero and Akella [12]. The basis aim of this study is to develop a method that combines Statistical Topics Models, Language Models and Natural Language Processing to retrieve healthcare related documents.

Authors improve a language model covering as a statistical topic model, noun phrase extraction, the query expansion using discharge summaries to determine related documents for retrieval task. Generalized Latent Dirichlet Allocation (GD-LDA) [13] is used as a topic model and Gibbs Sampling is used for inference.

“Probabilistic Author-Topic Models for Information Discovery” is another paper which is presented as a new text modeling technique for analyzing information from text corpus by Stevyers, Smyth and Griffiths [4]. The authors propose a new model, unlike the classical LDA's generative process. This difference arises from the assumption that each author is represented by a probability distribution over topics instead of each document is represented by a probability distribution over topics. The other important issue is handled with this paper is that documents can have more than one author. This paper contribution is that, in addition to the words that are observed variables for extracting document's topics process with LDA, the coauthors of documents are added as a new observed variables. The introduced author-topic model of LDA can automatically extract information about authors, topics, and documents and model is used Markov chain Monte Carlo algorithm for inference.

4. MEDICAL DATA

In today's fast-growing information age, the amount of the data produced is growing incredibly and a large portion of data which humanity has is stored in electronic form. Actually, the explosion in the amount of data provides great advantages in all fields of life but this thesis focuses on the medical field that is a very special field. Because, medical data creates historical record and support basic research to anticipate future health problems also it creates guidelines for good clinical practice.

The basic definition of medical data is any observation of a patient that is health related information associated with regular patient care or as part of a clinical trial program. Medical data represents tremendous amount of complex, heterogeneous, structured or unstructured data. In addition, it is more difficult to analyze human medical data than to analyze other living things because humans provide observations that are not included in any other data such as pain, hallucinations, visual and auditory sensations and many others. The major challenges of medical data related to collection, distribution, analysis and usage may be organized under four general titles [15] as follows;

- **Heterogeneity of medical data:** Medical data is gathered from very different sources such as anatomic images, research papers, laboratory data, the physician's observations and interpretations [15]. All these sources make the medical data heterogeneous.
- **Legal and ethical issues:** Medical data are obtained from human beings. When people are the subject, legal and ethical subjects are very important. In this context, there are two important points: preventing the abuse of patients and their data. These issues directly relate to intimacy and security.
- **Statistical philosophy:** Data mining methods, especially statistics and its basic assumptions must differ for medical data.

Special status of medicine: Medicine has the most special status among all sciences because it is directly related to life.

Medical data structure is very important because where and how the data is captured and stored determine how the data is analyzed and managed. Medical data can be structured, unstructured or semi-structured. Medical data can be stored as structured using a specific pattern. For example, laboratory results can be stored with specific codes for specific tests with the result values in structured way. Not only laboratory results, but also most diagnosis, procedure orders, medications and many others can be stored as structured data using certain patterns. Structured data can be examined in two different categories: structured and coded data elements and structured but uncoded data elements.

Structured and coded data elements can be analyzed and managed easily and effectively. The coded data can be randomly assigned numbers or internationally valid codes can be used such as ICD which stands for International Classification of Diseases. Since ICD is included in the data set used for this thesis study, we elaborate on this coding standard. ICD is defines as the diagnostic classification standard for all clinical and research purposes by Word Health Organization (WHO) [27].

ICD codes continue to be developed since the first version, each new version is offered with new features. The ICD included in the data set used in our study includes the 10th version codes namely as ICD-10 codes and this version is widely used today. The basic difference ICD-10 brings is the alphanumeric code structure. There are 4 levels in the classification structure of ICD-10. Each level is an detailed version of a higher one. An example for the ICD-10 class A is as follows:

- A00 Cholera
 - A00.0 Cholera, *Vibrio cholorea* 01, biovar cholera-dependent
 - A00.1 Cholera, *Vibrio cholerae* 01, biovar eltor
 - A00.9 Cholera, unspecified
- A01 Typhoid and paratifo
 - A01.0 Typhoid
 - A01.1 Parathy A
 - A01.2 Paratypes B
 - A01.3 Paratypers C

The use of ICD codes allows for systematic and meaningful regulation of diseases and this allows for easy examination and evaluation of medical data.

On the other hand, unstructured data can not be analyzed and managed easily because it is captured and stored as free text and provides the most comprehensive information. For example clinical notes and findings, radiology reports, test results and many others are generally stored with flexible documentation so they are unstructured. Semi-structured data is a combination of structured and unstructured data. As a result, the structure of medical data is especially important and has a big role to determine how and by which methods the data is handled.

4.1 Data Understanding

In section 3, we discussed medical data, features and challenges in detail. In this section, we examine the data set used for this thesis according to the features detailed in Section 4 and explain how it is processed to be used in the experiments.

In medical centers, various types of things are considered hospital expenses for instance surgery fees, blood tests, consultations, examinations, radiology results, hospitalization fees and all used medicine and materials such as serum, glove, injector, oxygen mask, plaster, stopper and etc. Our medical data set includes all of these hospital expences, therefore the data set is very heterogeneous. The heterogeneity we mentioned for our data set also includes a variety of medicines. The fact that the data is so heterogeneous is an important issue that should be considered in data preparation, statistical method selection and evaluation of results. Another important issue is that since the source of medical data is people, legal and ethical issues are essential. Actually, we are working on anonymized information, so is not possible to identify individual patients from this data. Data obtained from the billing module of a health informatics system. Details of lab results, doctor notes or patient details are not included in the dataset. Only billing information is provided with anonymized dates for the operations.

There are three different data sets used in this research. The first data set includes patient reception data from various medical departments. The second data set contains the patient reception data for the ward of internal diseases. Third set includes the patient reception data from department of general surgery.

The structure and content of the data set is an important part of understanding the problem. First of all, our medical data set is structured and includes coded data elements. Each row in the data set gives information about a procedure applied to the related patient and each column in the data set represents a feature for the procedure applied to the related patient. The data set consists of 11 columns and each column header in the data set is shown in the Table 4.1 with the feature of the information it contains.

Column Header	Column Features
Patient ID	The unique id for each patient
Patient Reception	Patient reception number
Operation Type	Type of procedure applied to patient
Operation ID	Coded information of the procedure applied to the patient
Operation Name	Description of the procedure applied to the patient
Table	Material knowledge used for the procedure applied to the patient
Diagnosis Number	How many diagnosis codes are used for the patient
Diagnosis Codes	ICD-10-CM Codes
Diagnosis Names	Explanation of the diagnostic codes for the patient
Operation Date	Date information of the procedure applied to the patient
Operation Time	Time information of the procedure applied to the patient

Table 4.1 Explanation of thesis data set columns

In addition, examples of data that each column can contain are summarized in the Table 4.2.

Column Header	Column Example Data
Patient ID	1148158839 / 727499184 etc.
Patient Reception	20771069-1 / 20732453-2 etc.
Operation Type	Consumables / Other / Bed / Drug / Surgery / Assay / Dental / Blood / Consultation / Examination / Radiology
Operation ID	10036 / 787931893 etc.
Operation Name	Nebülizatör ile ilaç uygulaması / Refakat / İntravenöz enjeksiyon / STOPER (AJUTAJ) / ENJEKTÖR 10 CC etc.
Table	Operations Performed / Used Material / Used Medicine
Diagnosis Number	4
Diagnosis Codes	A49.9,K21,S52.50,W19,
Diagnosis Names	Bakteriyel enfeksiyon, tanımlanmamış#Düşme, tanımlanmamış#Gastro-özofajial
Operation Date	01.01.2017
Operation Time	03:58:00

Table 4.2 Data examples of thesis data set columns

Since the data set we are working on is structured and includes coded data elements, standard text preprocessing steps which are tokenization, stopwords removal, lemmatization or stemming are unnecessary for the data set. However, specific preparation steps are applied to the dataset before the implementing the data mining methods.

The procedures applied for each patient reception in the data set are shown in different rows together with the details specified in the Figure 3.1. However, we are interested in only Patient Id, Operation Id and Diagnosis Code columns from the data set and we consider each Patient Id information as a document and the Operation Ids in the rows recorded with that Patient Id as the words belonging to that document. For each patient reception, a single list of diagnosis codes is used in the dataset, and we make a separate input data set of diagnosis codes list that matches each patient reception.

We can make a sampling of this assessment as follows:

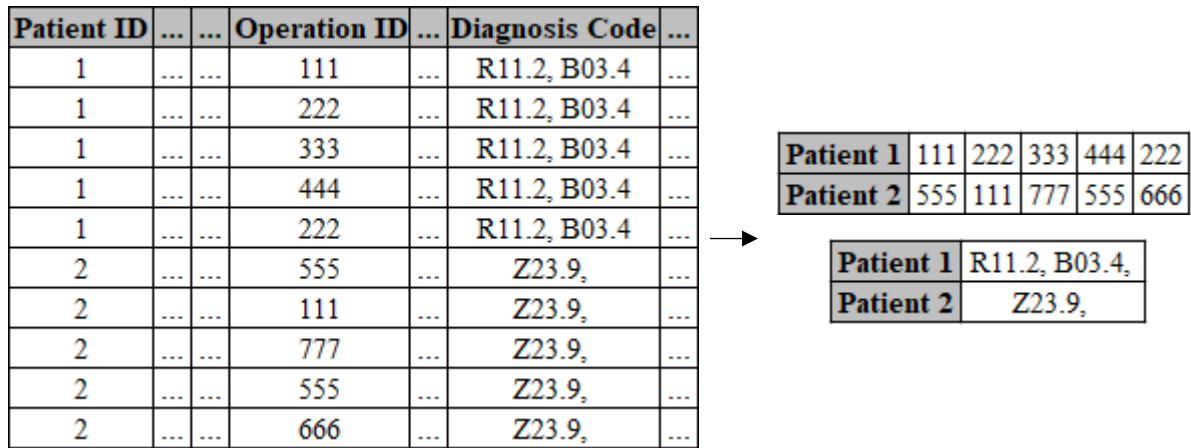


Figure 4.1 Sample of data selection and transformation from medical data set

The handling of three different data sets is the same. Two different input data are created for each different data set. In the first input data, each line contains operation ids belonging to a patient reception, and in the second input data, each line contains diagnosis codes. There is an important point we have noticed here, different patient receptions can be recorded at different times with the same Patient Id, such data are available in the data set, we take each patient reception to consider it as a different patient. Because a patient may have been admitted to the hospital for different reasons at different times, accepting these entries as a single patient causes misleading results. There is an important point to be mentioned here, there are two input data which are defined and exemplified are created from the beginning. The first input data consisting of operation ids and it is used in all data mining methods. The second input data consisting of diagnosis codes is not used in all data mining methods.

The section 6 includes detailed information on which data sets are used in which data mining method.

4.3.1 K-Fold Cross Validation

In data mining, it is important to generalize all the data in the training set. This generalization allows to make predictions for new data using the model. However, overfitting and underfitting are some of the main challenges we encounter when dealing with data.

- Overfitting occurs when a model learns the details and noises in the training data to such an extent that it negatively affects the model's predictive performance on new data. Namely, it is a kind of memorization of the training data.
- Underfitting occurs when a model cannot learn training data or cannot be generalized to new data. Such a model predictive performance is poor on training data so detection of underfitting is easy and the algorithm used can be changed as a solution.

In data mining studies, the data set is divided into training and test sets to test the success of the applied method. In addition, one of the main objectives is to understand how the model performs on the data set that it has not seen before.

However, selected training and test sets may cause some errors or overfitting due to distribution and using the same training and test data set does not allow these errors to be recognized. The important point we are interested in here is to see if there is any overfitting for our learning model while using the data set for this study. For this purpose, the most well known method is k-fold cross validation. With this method, preparing and using the data set and evaluating the results process is recommended as follows:

1. Divide the training data set into k random parts.
2. Use the k-1 part for training, 1 part for testing and repeat this step k times.
3. Collect the values obtained in each round and evaluate the average of the model performance.

This validation method is used in the study except the logistic regression method which is accepted as baseline.

5. PROPOSED MODEL

The proposed model uses the Latent Dirichlet Allocation algorithm on its basis and uses ICD-10 codes as an unused observable variable in the medical recommender systems so far developed and estimates missing operation ids for patient receptions labelled with specific ICD-10 codes. The experiment and its basic objectives of the model developed as a different version of LDA can be defined simply as follows:

1. Using the operation ids applied to the related patient reception and ICD-10 codes labeled to the related patient reception, attempting to estimate the operation ids which are not applied to the patient and may be missed.
2. Using the operation ids applied to the related patient and ICD-10 codes labeled to the related patient reception, attempting to estimate the operation ids which are applied to the patient and may be incorrect.

The proposed probabilistic topic model discovers latent/hidden topics that generate a patient reception in two statistical stages. Each ICD-10 Code is represented by a probability distribution over topics and each topic is represented as a probability distribution over operation ids for that topic. Although a new patient reception is not generated, for inference there are two main observations:

- A patient reception and operation ids which are applied to the related patient reception
- A patient reception and ICD-10 codes which are labeled to the related patient reception

In other words, the main aim is to find the missing operation ids more accurately by using both the operation ids applied to the patient receptions and the defined ICD-10 codes. The model not only discovers which topics are expressed in a patient reception, but also which ICD-10 codes are associated with each topic. In this model, in addition to patient receptions and operation ids, ICD-10 codes are used as observed variable. As in the LDA model, a single hidden variable is used, again called as topic. Figure 5.1 shows the plate diagram of the proposed model.

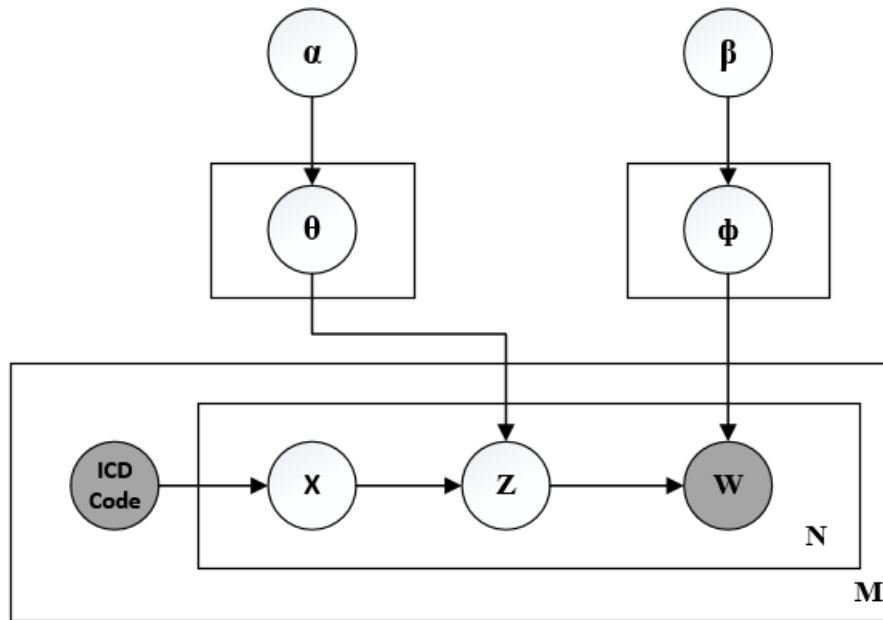


Figure 5.1 The proposed model graphical representation

Each patient reception has one or more ICD-10 codes, so the generative process of the proposed model is summarized as follows for the set of ICD-10 codes:

1. Choose an ICD-10 Code
2. Choose a topic for the given ICD-10 Code
3. Choose an operation id given the topic

The Figure 5.2 shows graph construction of the proposed model for medical data set. In the medical data set, there are one or more ICD-10 codes defined for each patient reception, ICD-10 codes labelled for each patient reception in the graph representation are expressed as a set.

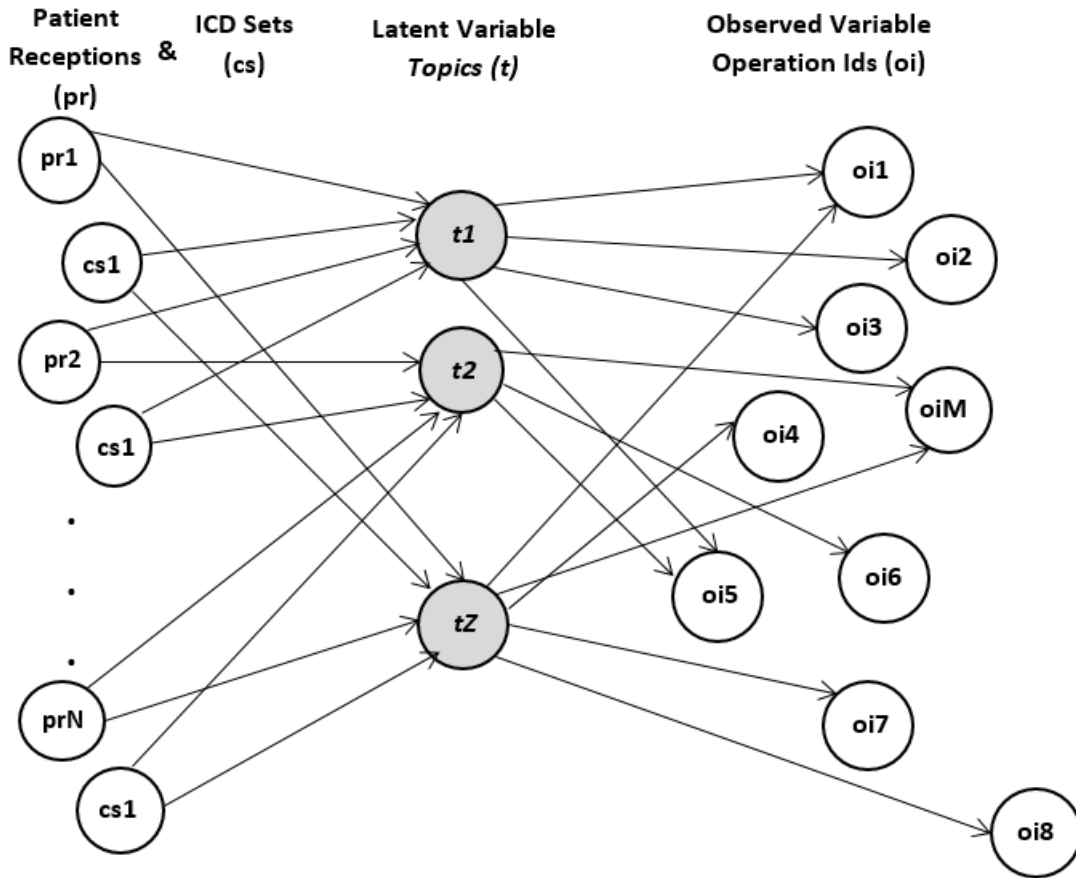


Figure 5.2 Graph construction of the Proposed Model

Basically, the process is to be able to assign ICD-10 codes and operation ids to latent variables by using operation ids applied for each patient reception and ICD-10 codes labelled to each patient reception.

The training and inference part of the proposed model is different from LDA. For training step, first of all, according to the data in the training set, random initialization step is performed for the following distributions and sum of distributions:

- Operation Id – Topic Distribution: Distribution of operation ids (oi) assigned to each hidden topic (z) is represented as follows:

$$P(oi, z) \tag{5.1}$$

Each operation id has a probability for each latent topic so each topic can be represented as a multinomial distribution over operation ids and the basic assumption can be formulated is as follows:

$$\sum_{k=1}^K P(o_i, z) = 1 \quad (5.2)$$

- Sum of Operation Id – Topic Distribution: Total number of operation ids assigned to each topic is formulated as follows:

$$\sum_z^Z P(z) \quad (5.3)$$

- ICD-10 Code – Topic Distribution : Each ICD-10 code (ic) in the corpus has a probability for each hidden topic (z) and mathematical expression of the distribution of ICD-10 codes assigned to each topic is as follows:

$$P(ic|z) \quad (5.4)$$

Each ICD-10 code has a multinomial mixture over latent topics, and the basic assumption can be formulated is as follows:

$$\sum_{t=1}^T P(ic, z_t) = 1 \quad (5.5)$$

- Sum of ICD-10 Code – Topic Distribution: Total number of operation ids assigned to each patient reception:

$$\sum_{ic}^{IC} P(ic) \quad (5.6)$$

For each operation id of each ICD-10 code, a random topic is selected from the initially selected topic number range, assigned to the topic distribution of the relevant operation id, and added to the topic distribution of the relevant ICD-10 codes.

As the second step, learning by using Gibbs Sampling algorithm is performed and distributions and sum of distributions are changed to a more accurate probability in each iteration.

When the number of initial iteration count is completed, the final theta and phi values are calculated as follows:

$$Theta = \sum_{ic}^{IC} (\sum_z^Z (P(ic, z) + alpha) / (\sum_{ic}^{IC} P(ic) + Z * alpha)) \quad (5.7)$$

$$Phi = \sum_z^Z (\sum_{oi}^{OI} (P(oi, z) + beta) / (\sum_z^Z P(z) + OI * beta)) \quad (5.8)$$

After the learning is completed, the learned model is saved and used in the inference phase. In the inference phase, instead of random initialization, the learned model is loaded first. That is, in the training phase distributions are initialized using the learned model instead of the randomly filled distributions at the initial stage using random initialization. Then as in the training phase, Gibbs Sampling algorithm is performed and distributions and sum of distributions are changed to a more accurate probability in each iteration. When the number of initial iteration count is completed, the final theta and phi values are calculated using the formulas (5.7), (5.8) in the same training phase for the test set. At the end of this phase, the probabilistic topic distributions of each patient reception's ICD-10 codes and each word in the test set are obtained.

The next step is to infer omissions according to obtained distributions. Each patient reception's ICD-10 codes are a weighted mixture of multiple topics and the patient reception contents are expected to be generated from a proportional mixture of the operation ids associated with each topic. The main aim is to calculate the conditional probability for every possible operation ids which can be suggested for the patient receptions in the test set. However, since each patient reception may have more than one ICD-10 code defined, the conditional probability has been calculated using the ICD-10 code with the maximum probability for each topic.

The mathematical formula of the conditional probability mentioned is as follows:

$$\begin{aligned} & P(\text{Operation } Id_i | \text{ICD} - 10 \text{ Code}_j) \\ &= \sum_z^Z P(\text{Operation } Id_i | \text{Topic}_z) \\ & * P(\text{Topic}_z | \text{ICD} - 10 \text{ Code}_j) \end{aligned} \tag{5.9}$$

6. EXPERIMENTS

In this section, the experiment is explained in detail by using the Logistic Regression method. Then, Latent Dirichlet Allocation and Topic-ICD Code experiments which are considered as the main experiments are given in detail.

6.1 Data Preperation

Data preperation and evaluation the results are processed according to the K-Fold Cross Validation method which is detailed in Section 4.1.1. The data set containing operation ids is divided into 10 parts to be able to apply the 10-fold cross validation. The ICD codes of the patient receptions included in each part are also divided into 10 parts according to the related patient receptions. In this way, the experiment is repeated 10 times with 10 different data sets. For each experiment, random omissions are created in the test sets. For each patient reception in the test set, randomly 10% of the operation ids are selected, removed from the test set and stored separately. If 10% of the operation ids applied to the relevant patient reception is less than 1, at least 1 operation id is selected as random omission. The code that makes this selections does not know which operation ids it is asked to predict, so it is forced to keep a distributional contextual representation of every operation id. However, no revisions are made to the input data set containing ICD-10 codes. A simulation of the data preparation step can be visualized as follows:

6.2 Logistic Regression

Generally, collaborative filtering methods are used by online retailers to propose the right product to the customers but the “Towards a Collaborative Filtering Approach to Medication Reconciliation” [11] study which is mentioned in Section 3.4, adapts this recommender mechanism to identify the omissions in the medication lists of patients. And, this study shows that collaborative filtering methods are also successful in the medical field.

Since this research has become one of the important stones for data mining in medical field, Logistic Regression which is one of the most successful methods as the main study of this thesis has been applied as the first solution. Logistic regression method is implemented with Python and some Python libraries were used. The first experiment and its basic objectives can be defined as using the operation ids applied to the related patient reception, attempting to estimate the operation ids that are not applied to the patient and may be missed.

As it can be understood from the simple description above, only data sets containing operation ids are used in this experiment. However, a specific data preparation process was carried out for this method.

Firstly, the data set containing operation ids is converted to Patient Reception-Operation Id two dimensional data using Dataframe library which is the two dimensional container of Panda in Python. This data structure stores the count of each distinct operation id is applied for each patient reception and it is used throughout the entire implementation. Second part of the data preparation for Logistic Regression detailed how dependent and independent variables are prepared in the train and test sets. In Section 2.3.2, the properties of these variables are mentioned. However, in this thesis, how to determine dependent and independent variables for medical data set is an important point because the data set which is used is not a labelled set and in fact there is only one kind of variable, that is, the operation ids. For this reason, the data set is processed with a loop and each cycle, a different operation id is assigned as a dependent or class variable, and each patient reception is labelled according to whether or not this operation id is applied. Then, the label column is separated from the train and test data sets.

The train set with no operation id for the label and its label column are given to Python's Logistic Regression classifier to perform learning. Once the learning is completed, only the test set is used without label column and the probability of applying the operation id which is used as a label to each patient is estimated. For estimation, the `_predict_proba_lr` method of the Logistic regression classifier is used. After the possibility of being applied to each patient reception for all distinct operation ids in the data set, the highest probability operation ids are recommended. Precision and recall values are calculated on how many of the operation ids that are proposed for the relevant patient reception are found to be correct.

6.3 Latent Dirichlet Allocation

In Section 3, the features of probabilistic topic modeling and some models are detailed. As a solution to the problems addressed by this thesis study, topic models are considered as a solution at the point where the operations which are applied for each patient reception are wondered why they are applied together. At this point, if hidden topics which are hold the operations together for each patient reception are found, the omissions can be estimated using these latent topics. In addition, it is necessary to identify the topics for which each word is relevant to estimate the operations that may have been improperly applied. Considering all these, LDA is one of the most popular generative probabilistic topic modelling method for text can be considered as a proper solution.

The LDA experiment and its basic objectives can be defined simply same as Logistic Regression method as using the operation ids applied to the related patient reception, attempting to estimate the operation ids which are not applied to the patient and may be missed.

The probabilistic topic distributions of each patient reception and each operation id in the test set are two essential probabilistic distributions which are intended to achieve with Latent Dirichlet Allocation. And this experiment is examined under 5 different headings; data preparation, graph construction, parameter estimation, training and inference.

As it can be understood from the simple description above in the introduction part, only data sets containing operation ids are used in this experiment same as Logistic Regression method.

Generally, LDA is used for the objective of document classification and its graph construction is shown through documents and words. In Section 3.3, detailed explanation of the LDA is made on this scope. However, the LDA can be easily adapted to many different areas. In this experiment, an adaptation of the model to the medical field is examined. The Figure 6.2 shows graph construction of the model for medical data set.

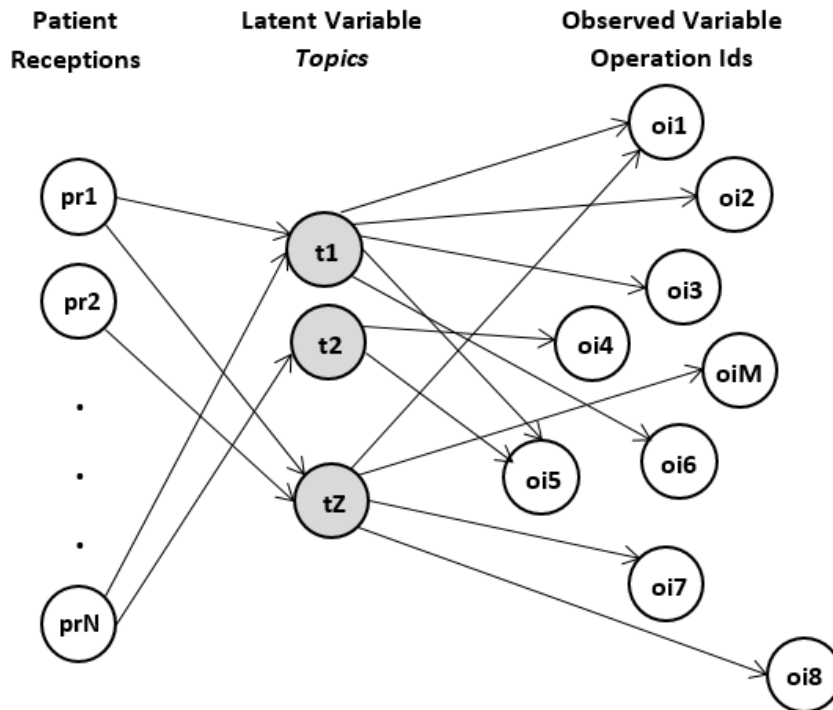


Figure 6.2 Graph construction of LDA

Basically, the process is to be able to assign patient receptions and operation ids to latent variables by using operation ids applied for each patient reception.

The Latent Dirichlet Allocation algorithm requires some initial parameters for training and inference. Some parameters required for LDA can be explained as follows:

- alpha (α) and beta (β) are hyperparameters which are described in detail in Section 3.3.
- ntopics is the number of topics to be used in topic distributions.
- niters is the number of iterations to use for sampling.

At this point, it is important that the alpha, beta and ntopic parameters are selected correctly according to the data set. Experiments with different input parameter values are performed on the first part of the 10 different experimental sets prepared for using in the k-fold cross validation method to find the correct parameters for the data set.

First of all, experiments are performed to find the optimum alpha value. The more uniform topics begin to disperse for each document with lower alpha value therefore experiments are made for alpha values smaller than 1. Precision and recall values calculated for different alpha values with 0.1 beta and 50 number of topics are shown in the graph below:

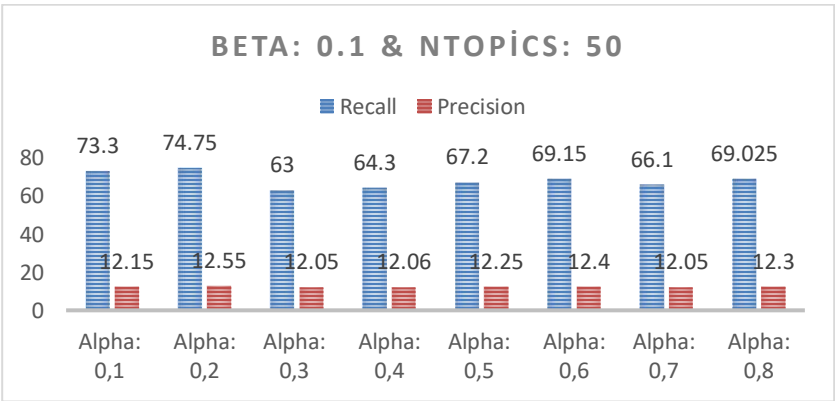


Figure 6.3 Optimum alpha value estimation graph

As can be seen from the chart above, the optimum alpha value is 0,2. Then, secondly, experiments are performed to find the optimum ntopics value.

The Figure 6.4 shows the calculated precision and recall values for different topic numbers:

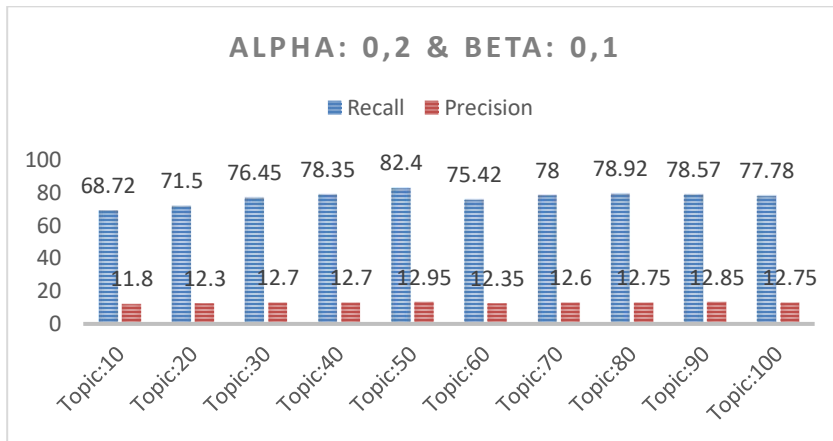


Figure 6.4 Optimum ntopics value estimation graph

According to the graphical representation of the experiments' results, the optimum number of topics is determined as 50. The last parameter estimation experimental set is done for the optimum beta value. The graph below shows the precision and recall results of experiments for different beta values with the optimum alpha and ntopic parameter values:

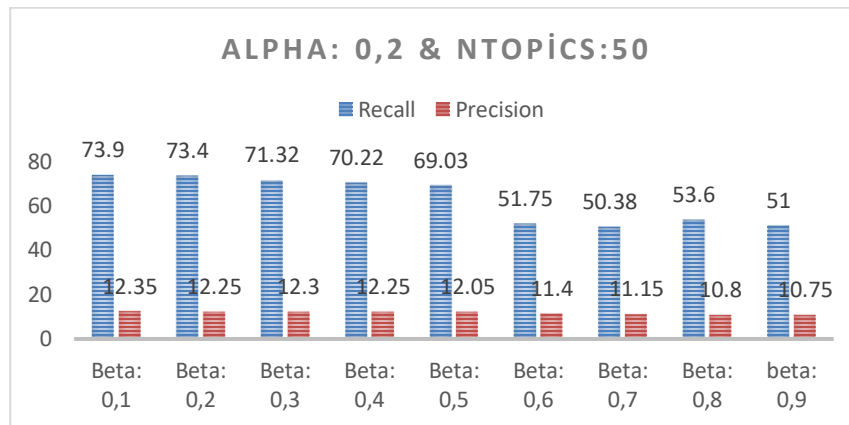


Figure 6.5 Optimum beta value estimation graph

According to the results of the test for optimum beta value, beta was chosen as 0,1. The optimum values determined for the input parameters are used in all experiments performed in this section.

In Section 3.3, as the detailed explanation of the LDA is done through the context of the word-document, in this section, it is explained in detail for the training and inference processes how the LDA is applied to the medical data set.

For LDA training process, according to the data in the training set, random initialization step is performed firstly for the following distributions and sum of distributions:

- Operation Id – Topic Distribution: Distribution of operation ids (oi) assigned to each hidden topic (z) is represented as follows:

$$P(oi, z) \tag{6.1}$$

Each operation id has a probability for each latent topic so each topic can be represented as a multinomial distribution over operation ids and the basic assumption can be formulated is as follows:

$$\sum_{k=1}^K P(oi_k, z) = 1 \tag{6.2}$$

- Sum of Operation Id – Topic Distribution: Total number of operation ids assigned to each topic is formulated as follows:

$$\sum_z^Z P(z) \tag{6.3}$$

- Patient Reception – Topic Distribution : Each patient reception (pr) in the corpus has a probability for each hidden topic (z) and mathematical expression of the distribution of patient receptions assigned to each topic is as follows:

$$P(pr|z) \tag{6.4}$$

Each patient reception has a multinomial mixture over latent topics, and the basic assumption can be formulated is as follows:

$$\sum_{t=1}^T P(pr, z_t) = 1 \tag{6.5}$$

- Sum of Patient Reception – Topic Distribution: Total number of operation ids assigned to each patient reception:

$$\sum_{pr}^{PR} P(pr) \quad (6.6)$$

For each operation id of each patient reception, a random topic is selected from the initially selected topic number range, assigned to the topic distribution of the relevant operation id, and added to the topic distribution of the relevant patient reception.

As the second step, learning by using Gibbs Sampling algorithm is performed and distributions and sum of distributions are changed to a more accurate probability in each iteration.

When the number of initial iteration count is completed, the final theta and phi values are calculated as follows:

$$Theta = \sum_{pr}^{PR} \left(\sum_z^Z (P(pr, z) + alpha) / \left(\sum_{pr}^{PR} P(pr) + Z * alpha \right) \right) \quad (6.7)$$

$$Phi = \sum_z^Z \left(\sum_{oi}^{OI} (P(oi, z) + beta) / \left(\sum_z^Z P(z) + OI * beta \right) \right) \quad (6.8)$$

After the learning step is completed, the learned model is saved and used in the inference phase. In the inference phase, instead of random initialization, the learned model is loaded first. One of the basic and important points here is about the data set. The used data set for inference differs from the data set used for learning. In other words, the code never saw the patient entries which are in the test data set, in the learning phase. That is, in the training phase distributions are initialized using the learned model instead of the randomly filled distributions at the initial stage using random initialization. Then as in the training phase, Gibbs Sampling algorithm is performed and distributions and sum of distributions are changed to a more accurate probability in each iteration.

When the number of initial iteration count is completed, the final theta and phi values are calculated using the formulas (6.7, 6.8) in the same training phase for the test set. At the end of this phase, the probabilistic topic distributions of each patient reception and each word in the test set are obtained.

The next step is to infer omissions according to obtained distributions. Each patient reception is a weighted mixture of multiple topics and the patient reception contents are expected to be generated from a proportional mixture of the operation ids associated with each topic. This study does not interested in to generate a new patient reception. Instead, the aim is to calculate the conditional probability for every possible operation ids that can be suggested for the patient receptions in the test set. The mathematical formula of the conditional probability mentioned is as follows:

$$\begin{aligned}
 & P(\text{Operation } Id_i | \text{Patient Reception}_j) \\
 &= \sum_z^Z P(\text{Operation } Id_i | \text{Topic}_z) \\
 & * P(\text{Topic}_z | \text{Patient Reception}_j)
 \end{aligned} \tag{6.9}$$

6.4 Proposed Model

In Section 5, the structure of the Proposed Model are explained in detail. As a solution to the problems addressed by this thesis study, as distinct from the LDA, we used ICD-10 codes as a new observed variable in the Proposed Model when grouping the operations under the hidden topics. This section discusses the important points of the Proposed model experiment.

As can be understood from the above simple description of the Proposed Model, two input data sets are used in this experiment, one of them contains operation ids for patient receptions and the other contains ICD-10 codes for related patient receptions. Section 6.1 describes how to prepare input data sets.

The proposed method is based on the Latent Dirichlet Allocation algorithm therefore it is very important that alpha and beta hyperparameters values and how many topics is used for modelling. The most appropriate values of parameters for medical datasets are quite parallel

with the LDA according to the experimental results of the selected representative datasets and the parameters to be used according to the experimental results are as follows:

- Alpha: 0.2
- Beta: 0.1
- Topic count: 50

In Section 5, training and inference steps are explained in detailed. The Proposed model as well as Logistic Regression and Latent Dirichlet Allocation, makes suggestions for the omissions according to the calculated maximum conditional probability values. The operations with the highest probability for some topics can be displayed in the Table 6.1:

TOPIC10	TOPIC20	TOPIC30	TOPIC40	TOPIC50
PARACEROL 10 MG/ML İV 100 ML 12 FLAKON	OKSAPAR 6.000 ANTI-XA İÜ/0,6 ML İV/SC 2 ENJEKTÖR	DELİX 5 MG 28 ÇENTİKLİ TABLET	UNACEFİN 1.000 MG İV 1 FLAKON	PAROL 500 MG 30 TABLET
SERUM İZOPLEN-M %5 DEKSTROZ 500 ML MX SETSİZ TÜRKİPSAN	Anti CMV IgM (Microparticle immunoassay-MEIA or similar)	Sedimentasyon	SERUM İZOPLEN-M %5 DEKSTROZ 500 ML MX SETSİZ TÜRKİPSAN	Sedimentasyon
Akciğer grafisi P.A. (Tek yön)	İntravenöz ilaç infüzyonu	Sağlık kurulu raporu	METRONİDAZOLE FRESENIUS %0,5 100 ML İV SOLÜSYON	Glukoz
CONTRAMAL 100 MG/2 ML İM/İV/SC 5 AMPUL	Demir (Serum)	Serbest T3	FUROMİD 20 MG/2 ML İM/İV 5 AMPUL	HBsAg (Kemoluminesans veya benzeri)
SEVORANE LİKİT %100 250 ML SOLÜSYON	PARACEROL 10 MG/ML İV 100 ML 12 FLAKON	Serbest T4	Glikolize hemoglobin (Hb A1C)	Anti HIV (Kemoluminesans veya benzeri)
ENJEKTÖR UCU NO:22 (SİYAH)	ZOLAMİD 15 MG/3 ML İV 3 ML 5 AMPUL	Amilaz	ABO+Rh tayini (Forward gruplama)	Tam Kan (Hemogram)
ANTİBAK. POLİGLAKTİN 2/0 Y.İ. 26MM 70CM 1/2 suture-76	STERİL CERRAHİ ÖNLÜK	TSH	ANTI-NAUSEA 10 MG/2 ML İM/İV 5 AMPUL	Laktik Dehidrogenaz (LDH)
ARİTMAL %2 100 MG/5 ML İV 5 AMPUL	STERİL DİSTİLE SU 500 ML	Lipaz	DİKLORON 75 MG/3 ML İM 10 AMPUL	Sodyum (Na) (Serum ve vücut sıvılarında, herbiri)

Table 6.1: Highest possible operations for topic 10, 20, 30, 40 and 50

When the operations in the Table 6.1 are examined, it can be seen that similar operations are gathered under the same subject. For example, if we examine for Topic 10, Paracerol and Contramal are analgesic medication. Sevorane is both an analgesic and anesthetic medication. In addition, Aritmal can be defined as a drug used in cardiac surgery. If we evaluate these findings together with the Serum İzoplen, chest X-ray, Injector Tip, Antibacterial Polyglactine, we can conclude that there may be procedures related to cardiac surgery. When we examine the operations included in Topic 30, Free T3, Free T4 and TSH are directly related to thyroid. Amylase and Lipase are enzymes secreted from the pancreas. Delix is a drug related to blood pressure. Sedimentation is about the analysis of blood. According to these findings, we can say that the associated hormones and enzymes that are required for a blood test are collected under the same subject. For Topic 50, HbsAg for hepatitis B and Anti-HIV are blood values for HIV. These tests also complete the sedimentation e whole blood test. It can be said that similar operations related to blood tests are collected under this subject.

7. RESULTS AND EVALUATION

7.1 Evaluation Metrics

There are two concepts that are frequently used as evaluation metrics; precision and recall. Also, the F1 measure is the harmonic mean of precision and recall [42] therefore, we use F1 measure as an other evaluation metric. The last metric used in the experiments is Mean Reciprocal Rank (MRR).

Precision or confidence represents the proportion of estimated as positive cases that are correctly real positive cases [35] and can be formulated as follows:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (7.1)$$

As it can be understood from the formula, precision is calculated as the ratio of the correct results in the information brought to the whole information. The precision calculation we use in our experiments is as follows:

$$Precision = \frac{True\ Predictions}{Recommended\ Top\ N} \quad (7.2)$$

Recall or Sensitivity is the proportion of real positive cases that are correctly estimated as positive cases [35] and can be formulated as follows:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (7.3)$$

As it can be understood from the formula, recall is calculated by the ratio of the correct results brought to the correct results that need to be brought. The recall calculation we use in our experiments is as follows:

$$Recall = \frac{True\ Predictions}{Omissions} \quad (7.4)$$

The F1 measure calculation formula that we use as the evaluation metric is as follows:

$$F1 \text{ Measure} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (7.5)$$

The Reciprocal Rank (RR) performs the calculation using the first order in which the searched data is located [43]. In our experiments, the calculation of RR is based on the rank of correct predicted operations. When RR is averaged the Mean Reciprocal Rank (MRR) [43]. In the experiments, MRR is calculated as follows:

$$MRR = \frac{1}{True \text{ Prediction}} \sum_{i=1}^{True \text{ Prediction}} \frac{1}{Rank_i} \quad (7.6)$$

7.2 Wilcoxon Signed-Ranks Test

The Wilcoxon Signed-Ranks test orders the differences of the two classifiers' performance for each data set [44]. The non-parametric tests can be used to analyze data that we do not have accurate information. For instance, data does not contain any distribution assumptions. Because of that, this test is applied to analyze the results of LDA and the Proposed Model. The F1 Measure results of LDA and the Proposed Model for the k data set obtained using the k-fold cross validation method are analyzed with this test.

Firstly, the difference between the F1 Measure results of LDA and the Proposed Model are obtained and let's call this value d_i . The differences are sorted according to their absolute values and then average orders are assigned to the relationships [44]. The positive orders are collected as R^+ and the negative ranks are collected as R^- . These collections can be formulated as follows.

$$R^+ = \sum_{d_i > 0} Rank(d_i) + \frac{1}{2} \sum_{d_i = 0} Rank(d_i) \quad (7.7)$$

$$R^- = \sum_{d_i < 0} Rank(d_i) + \frac{1}{2} \sum_{d_i = 0} Rank(d_i) \quad (7.8)$$

The test is finalized by comparing the small of the sums ($\min(R^-, R^+)$) with the critical value. The table of critical values for confidence levels 0.05 and 0.10 is shared below. The Table 7.1 is used to evaluate the test results.

#data sets	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$w_{0.05}$	5	6	7	7	8	9	9	10	10	11	12	12	13	13	14	15	15	16	17	18	18
$w_{0.10}$	5	6	6	7	7	8	9	9	10	10	11	12	12	13	13	14	14	15	16	16	17

Table 7.1: The two-tailed sign test critical values at $\alpha = 0.05$ and $\alpha = 0.10$.

To say that the difference between the LDA and the Proposed Model is significant, the smaller of the sums must be equal or less than the critical value.

7.3 Experimental Results and Evaluation

The size of data sets are important in order to evaluate the experimental results correctly. The Table 7.2 lists the significant data numbers of the three different data sets used in the experiments.

Data Sets	Source Department	Patient Reception	Distict Operation Id	Distinct ICD-10 Codes	Average Number of ICD-10 Codes Per Patient Reception
First Data Set	Various Departments	372	1620	182	1,4
Second Data Set	Internal Diseases	450	1013	163	2,02
Third Data Set	General Surgery	4109	1967	410	1,26

Table 7.2: The size of the data sets.

Precision, Recall, F1 Measure and Mean Reciprocal Rank (MRR) values are used as the evaluation metrics. According to the evaluation metrics, the experimental results of the Logistic

Regression, Latent Dirichlet Allocation and the Proposed Model for three different data sets are shown in the Table 7.3, 7.4 and 7.5. Since the F1 Measure value is the harmonic mean of the Recall and Precision values, three methods are compared for three different data sets using F1 Measure. Also, three methods are compared with precision-recall curves for three different data sets. According to the comparative tables, graphs and curves, results are evaluated and discussed. In addition, the Wilcoxon Signed-Ranks Test results are also included in this section.

First of all, the first data set Recall, Precision, F1 measure and MRR results according to the 10-fold cross validation are shown the Table 7.3. In addition, for three methods F1 Measure comparative results are shown in the figure 7.1. The more detailed experimental results are represented in appendixes A and B.

Suggested Operations	Method	Precision	Recall	F1 Measure	MRR
Top 5	Logistic Regression	0,095	0,263472	0,13964	0,21083
	Latent Dirichlet Allocation	0,299	0,27035	0,28395	0,28853
	Proposed Model	0,35132	0,31927	0,33453	0,32738
Top 10	Logistic Regression	0,087568	0,39997	0,14367	0,26238
	Latent Dirichlet Allocation	0,21775	0,3661	0,27307	0,35901
	Proposed Model	0,26282	0,41963	0,32321	0,39104
Top 30	Logistic Regression	0,070877	0,575585	0,12621	0,21253
	Latent Dirichlet Allocation	0,128875	0,568675	0,21012	0,47386
	Proposed Model	0,14902	0,61597	0,23999	0,49412
Top 50	Logistic Regression	0,065	0,6	0,11729	0,23570
	Latent Dirichlet Allocation	0,099505	0,66308	0,17304	0,50639
	Proposed Model	0,10551	0,69923	0,18335	0,52056
Top 70	Logistic Regression	0,053	0,66	0,09812	0,13861
	Latent Dirichlet Allocation	0,074515	0,72618	0,13516	0,52546
	Proposed Model	0,08185	0,75453	0,14769	0,53858
Top 90	Logistic Regression	0,046	0,71	0,08640	0,14100
	Latent Dirichlet Allocation	0,06185	0,76264	0,11442	0,53567
	Proposed Model	0,06672	0,80097	0,12319	0,55914

Table 7.3: First data set results according to the evaluation metrics

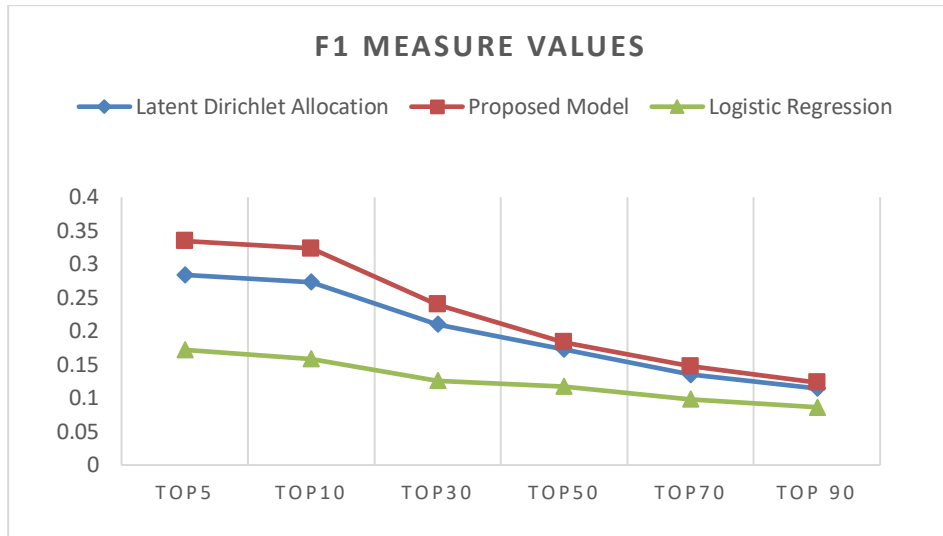


Figure 7.1: First data set F1 Measure results for three methods

According to the all evaluation metrics, the predictions of the proposed model for the first data set including patient reception data from various medical departments are more accurate than the Logistic regression and Latent Dirichlet Allocation. As seen in addition, even if the number of suggested operation id changes, the success sequence for the three methods does not change. However, it can be said that the increase in the suggested number of words prevents us to distinguish between the three methods.

The second data set contains the patient reception data for the department of internal diseases. For this data set Recall, Precision, F1 measure and MRR results according to the 10-fold cross validation are shown the Table 7.4. In addition, for three methods F1 Measure comparative results are shown in the figure 7.2. The more detailed experimental results are presented in appendixes C and D.

Suggested Operations	Method	Precision	Recall	F1 Measure	MRR
Top 5	Logistic Regression	0,10204	0,28718	0,15057	0,23639
	Latent Dirichlet Allocation	0,37315	0,33712	0,35186	0,35610
	Proposed Model	0,65877	0,46436	0,53731	0,40227
Top 10	Logistic Regression	0,08367	0,39396	0,13803	0,22900
	Latent Dirichlet Allocation	0,27764	0,41743	0,33162	0,40990
	Proposed Model	0,49755	0,57990	0,5298	0,44309
Top 30	Logistic Regression	0,05510	0,602295	0,11010	0,22398
	Latent Dirichlet Allocation	0,186587	0,646136	0,28887	0,48795
	Proposed Model	0,27156	0,746316	0,39507	0,50771
Top 50	Logistic Regression	0,046	0,73	0,08654	0,23094
	Latent Dirichlet Allocation	0,14815	0,77630	0,24836	0,52268
	Proposed Model	0,18391	0,81397	0,29835	0,54942
Top 70	Logistic Regression	0,033	0,78	0,06332	0,19820
	Latent Dirichlet Allocation	0,11379	0,84780	0,20034	0,53799
	Proposed Model	0,13886	0,85050	0,23765	0,55288
Top 90	Logistic Regression	0,033	0,85	0,06353	0,12401
	Latent Dirichlet Allocation	0,09374	0,88615	0,16936	0,55315
	Proposed Model	0,11206	0,87255	0,19785	0,57920

Table 7.4: Second sata set results according to the evaluation metrics

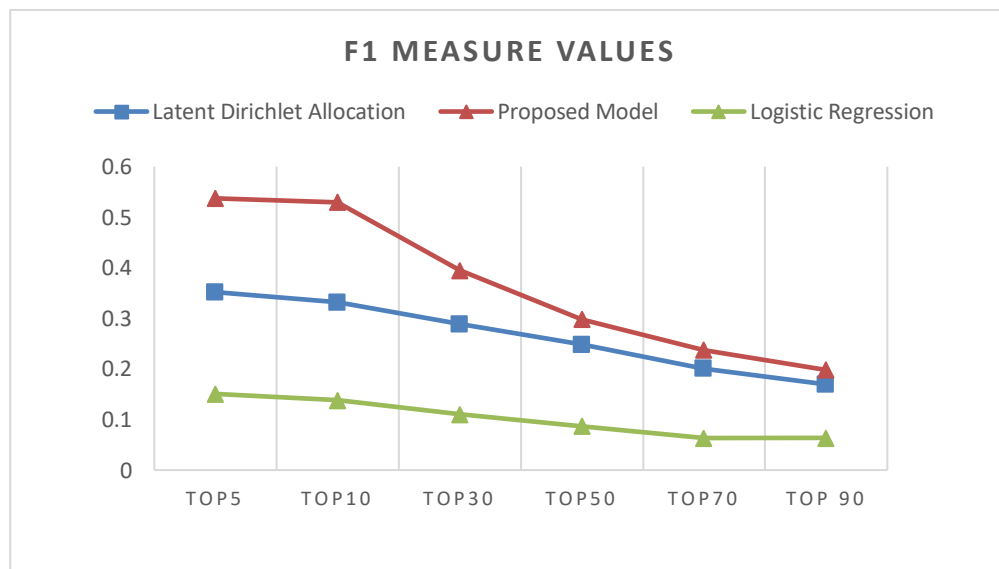


Figure 7.2: Second Data Set F1 Measure Results For Three Methods

According to the all evaluation metrics, the difference between the three methods is clearly seen. The predictions of the proposed model for the second data set including patient reception data from internal diseases departments are more accurate than the Logistic regression and LDA. It can be said that the experimental results from the second data set and the first data set are very similar. As for the results of the first experiment, even if the number of suggested operation id changes, the success sequence for the three methods does not change. However, it can be said that the increase in the suggested number of words prevents us to distinguish between the three methods.

The third data set contains the patient reception data for the department of general surgery. For this data set Recall, Precision, F1 measure and MRR results according to the 10-fold cross validation are shown the Table 7.5. In addition, for three methods F1 Measure comparative results are shown in the figure 7.3. The more detailed experimental results are presented in appendixes E and F.

Suggested Operations	Method	Precision	Recall	F1 Measure	MRR
Top 5	Logistic Regression	0,10617	0,30789	0,15790	0,20655
	Latent Dirichlet Allocation	0,37315	0,33712	0,35422	0,34610
	Proposed Model	0,44792	0,37512	0,40830	0,37341
Top 10	Logistic Regression	0,09359	0,48350	0,15682	0,24258
	Latent Dirichlet Allocation	0,27764	0,41743	0,33348	0,35990
	Proposed Model	0,35033	0,47142	0,40195	0,41629
Top 30	Logistic Regression	0,05781	0,64897	0,10617	0,21386
	Latent Dirichlet Allocation	0,18658	0,64582	0,28952	0,45523
	Proposed Model	0,21255	0,63407	0,31838	0,45983
Top 50	Logistic Regression	0,046	0,73	0,08654	0,19735
	Latent Dirichlet Allocation	0,14815	0,76730	0,24835	0,49668
	Proposed Model	0,14078	0,76602	0,23784	0,50870
Top 70	Logistic Regression	0,033	0,78	0,06332	0,19352
	Latent Dirichlet Allocation	0,11379	0,82180	0,19990	0,49799
	Proposed Model	0,12287	0,84090	0,21441	0,53422
Top 90	Logistic Regression	0,033	0,85	0,06353	0,17694
	Latent Dirichlet Allocation	0,09374	0,87015	0,16926	0,53315
	Proposed Model	0,11059	0,88348	0,19657	0,54713

Table 7.5: Third data set results according to the evaluation metrics

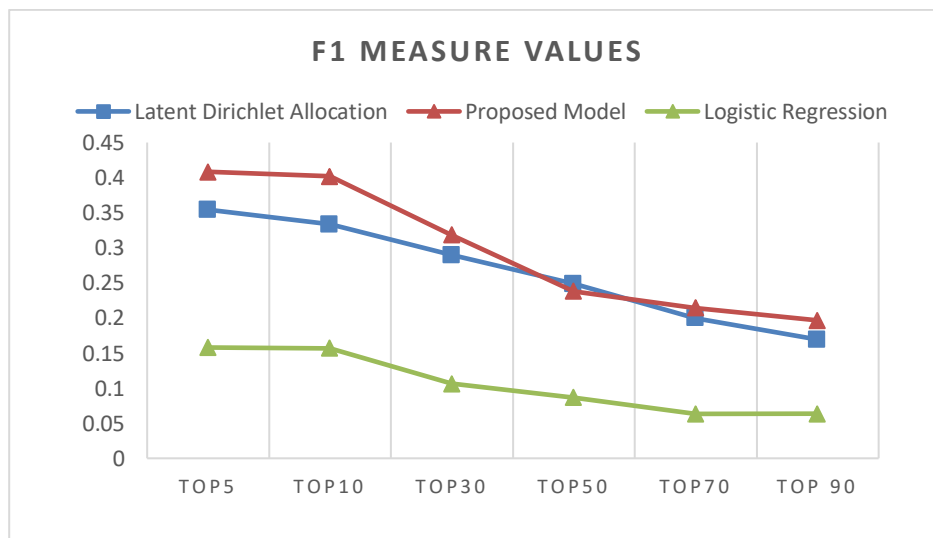


Figure 7.3: Third Data Set F1 Measure Results For Three Methods

For the third data set including patient reception data from general surgery department, evaluation metrics show that the predictions of the Proposed Model are more accurate than the Logistic regression and Latent Dirichlet Allocation. As seen in addition, even if the number of suggested operation id changes, the success sequence for the three methods does not change. However, at a point it is observed that the Proposed Model predicts worse than LDA, which may be due to the fact that the number of ICD-10 codes per patient reception shown in table 7.2 is less in the third dataset than in the other datasets. The small number of ICD-10 codes per patient is a factor preventing the separation of operation ids.

The Wilcoxon Signed-Ranks test is applied to prove that the success of the Proposed Method is significant. Wilcoxon Signed-Ranks test described in section 5.2 is used to compare the LDA and the Proposed Model with F1 measures. According to the table of critical values (Table 7.1) for the Wilcoxon’s test, for a confidence level of $\alpha = 0.05$ and $N = 10$ data sets of 10-fold cross validation, critical value is 9. For three data sets, the test results according to different suggestion counts are shown the Table 7.6, 7.7 and 7.8:

First Data Set				
Suggested Operations	R^-	R^+	$\min(R^-, R^+)$	Critical Value
Top 5	7	48	7	9
Top 10	2	53	2	9
Top 30	3	52	3	9
Top 50	13	42	13	9

Table 7.6: Wilcoxon Signed-Ranks test steps for first data set.

Second Data Set				
Suggested Operations	R^-	R^+	$\min(R^-, R^+)$	Critical Value
Top 5	0	55	0	9
Top 10	0	55	0	9
Top 30	0	55	0	9
Top 50	5	50	0	9

Table 7.7: Wilcoxon Signed-Ranks test steps for second data set.

Third Data Set				
Suggested Operations	R^-	R^+	$\min(R^-, R^+)$	Critical Value
Top 5	0	55	0	9
Top 10	0	55	0	9
Top 30	0	55	0	9
Top 50	6	49	6	9

Table 7.8: Wilcoxon Signed-Ranks test steps for third data set.

According to the Wilcoxon's test, we can say that the difference between the LDA and the Proposed Method is significant if the smaller of the sums is less than the critical value.

For the first data set, in general the Proposed Model is significantly successful than LDA. But, it is seen that the success of the Proposed Model is insignificant as the number of suggestions increases. This situation arises from the fact that the difference between the two methods is not fully differentiated as the number of suggestions increases. Because as the number of suggestions increases, the effect of how well the method predicts decreases.

For the second data set, the Wilcoxon's test results on F1 measure indicate that the Proposed Model is significantly successful than the LDA. When the test results are compared for three data sets, it is seen that the most significant success is obtained in this data set.

For the third data set, according to the results of the Wilcoxon Signed-Ranks test on F1 measure, the Proposed Method has achieved the significant success on this data set.

In addition to the F1 measure graphs and the Wilcoxon Signed-Ranks test results, the Precision-Recall curves are important for evaluating the results from a different angle and more clearly. The Precision-Recall curves for the three different data sets used in the experiments can be displayed in the Figure 7.4 and 7.5.

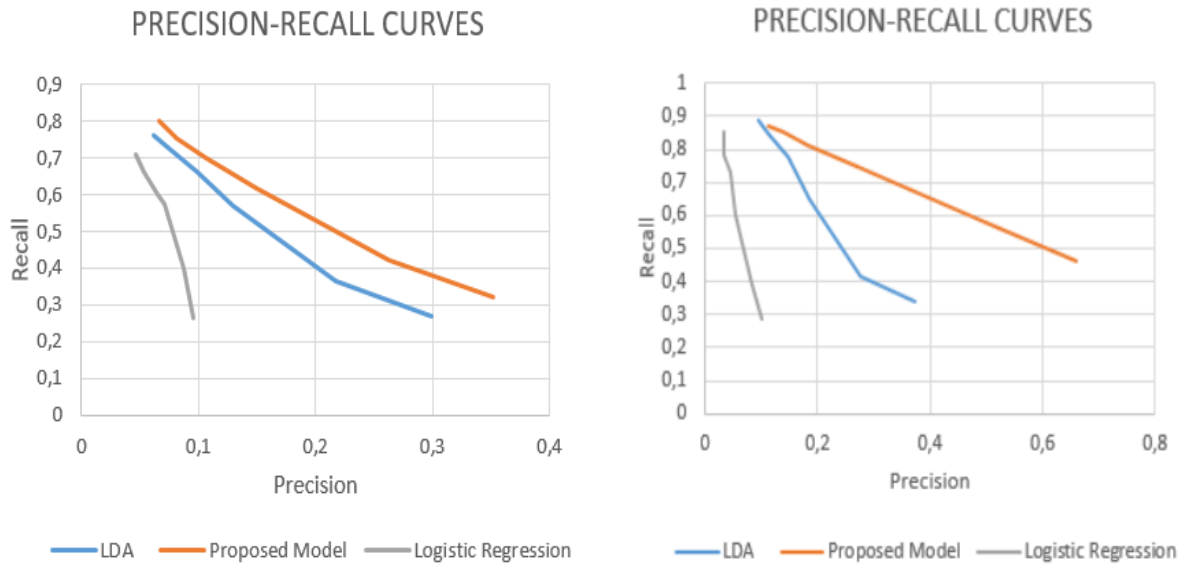


Figure 7.4: The curve on the left is the Precision-Recall curve for the first data set and the curve on the right for the second data set.

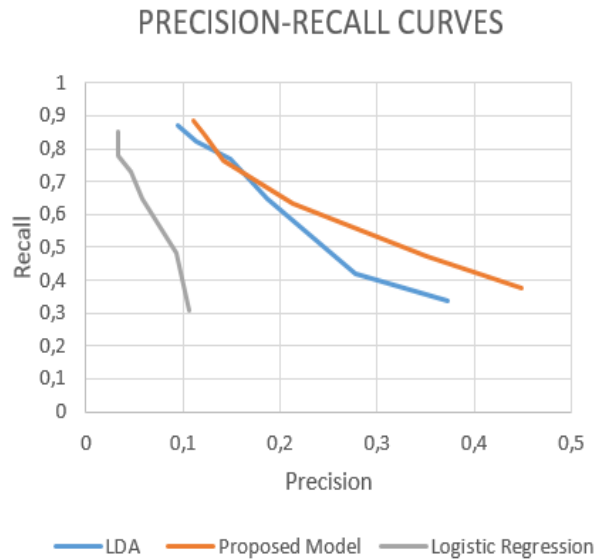


Figure 7.5: The Precision-Recall curve for the third data set.

In the precision-recall curve, the curve that closest to the upper right corner is most successful. According to the precision-recall curves showing the results of the experiment, for the first data set and second data set, it can be said that the Proposed Model is the most successful method. For the third data set, it can easily be said that the proposed method is generally better. Although the first data set is small, it contains patient reception data from different departments and its results are very important. Because the experiments performed in this data set show that the Proposed Method is able to distinguish the patient receptions to different departments quite well and the estimation rate is quite successful. The second and third data sets are larger than the first data set and the results are important because they show that the Proposed Model works well in large data sets.

As a result, it can be concluded that according to evaluation metrics and test results the Proposed Model is more successful than Logistic Regression [11] and Latent Dirichlet Allocation [8] methods which have been applied as solutions for similar problems and have achieved certain successes.

8. CONCLUSION AND FUTURE WORK

In the information age we are in, incredibly fast growing data has become one of the biggest challenges to be addressed. Building statistical models is a popular research field because of the large amount of data we are facing and want to use in the best way. The investigation for a new models, which has gained importance in every aspect of life, has also become extremely important in the medical field. In this research, we present a study to eliminate the negative effects of the incomplete operation lists. We suggest a new probabilistic topic model for predicting the potential omissions in the operation lists. So far, there are solutions offered to similar problems with different methods in medical field. Logistic regression (LR) and Latent Dirichlet Allocation (LDA) are the most successful ones. The proposed model is based on the LDA and Gibbs sampling algorithms for more efficient learning and inference. Distinctively, we have developed a model using the previously unused ICD-10-CM codes as a new observed variable. The precision, recall, F1-Measure and MRR values used as evaluation metrics are compared for LR, LDA and the Proposed Method. According to the experimental results for the three different data sets, it is observed that the Proposed Method is 5% more successful than the LDA method and 13% more successful than the Logistic Regression method. Therefore, the most effective estimation results according to evaluation metrics are obtained with the Proposed Method. As a result, this study contributes to the field of data mining with medical data by developing an improved version of an existing probabilistic topic model.

In the future, we plan to continue to investigate the medical dataset and evaluate new possible observable variables that may contribute to the model we have developed. Because, there is some information such as the operation type and branch name that we think can improve our estimation in the data set. In addition, as data mining is a highly popular field in the world, researches continue and these researches result in great developments. For this reason, we plan to try to adapt the new methods as a solution for our problem. One of the methods we think can be adapted to our problem is BERT[31], which is one of the most promising new methods.

9. APPENDICES

Appendix A: 10 Fold Cross Validation Latent Dirichlet Allocation Experimental Results For The First Data Set

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,36	0,249	0,294384236	0,26034344
1	0,2	0,2715	0,230328738	0,29496965
2	0,2	0,271	0,23014862	0,265190627
3	0,17	0,23925	0,198766035	0,263710553
4	0,2	0,33975	0,251783233	0,378706051
5	0,39	0,2775	0,324269663	0,322498773
6	0,38	0,2945	0,331830986	0,27415663
7	0,265	0,23725	0,250358387	0,257407488
8	0,3	0,26775	0,282959049	0,283454444
9	0,525	0,256	0,344174136	0,284909733
Average	0,299	0,27035	0,283954158	0,288534739

Table 9.1: Latent Dirichlet Allocation Experimental Result with Top 5 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,25	0,31225	0,277678968	0,286778536
1	0,1425	0,34975	0,202496191	0,372341525
2	0,1475	0,35775	0,208879268	0,356409066
3	0,1375	0,348	0,197116375	0,358543854
4	0,1675	0,42125	0,239692144	0,462947095
5	0,2825	0,3215	0,300740894	0,326913133
6	0,2625	0,48425	0,340450285	0,436139059
7	0,1775	0,32775	0,230284513	0,302757422
8	0,1925	0,3385	0,245428437	0,329450737
9	0,4175	0,4	0,408562691	0,357835142
Average	0,21775	0,3661	0,273077931	0,359011557

Table 9.2: Latent Dirichlet Allocation Experimental Result with Top 10 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,15925	0,579	0,249795462	0,449421916
1	0,07525	0,4315	0,128151455	0,378449437
2	0,08225	0,50625	0,141509133	0,418231812
3	0,086	0,54575	0,148585675	0,472380367
4	0,10525	0,66	0,181548514	0,539312234
5	0,16325	0,56825	0,253634484	0,49367449
6	0,15075	0,68325	0,247002248	0,55636443
7	0,0935	0,5895	0,161400439	0,509179552
8	0,10675	0,566	0,179622445	0,524686821
9	0,2665	0,55725	0,360563581	0,39693248
Average	0,128875	0,568675	0,210129713	0,473863354

Table 9.3: Latent Dirichlet Allocation Experimental Result with Top 30 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,1245	0,764252525	0,214119086	0,52116223
1	0,0645	0,574124785	0,115971223	0,426423823
2	0,0655	0,581145782	0,117730757	0,440369369
3	0,066	0,591122559	0,118742199	0,490754393
4	0,07755	0,722222222	0,140060712	0,548988816
5	0,12255	0,660251415	0,206728832	0,583520632
6	0,11845	0,774526751	0,205476108	0,588695834
7	0,073	0,643752626	0,131130156	0,522629509
8	0,079	0,652756426	0,140942411	0,543150252
9	0,204	0,666666667	0,312404288	0,398248484
Average	0,099505	0,663082176	0,17304249	0,506394334

Table 9.4: Latent Dirichlet Allocation Experimental Result with Top 50 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,0965	0,817751425	0,172628689	0,525502283
1	0,0415	0,62955759	0,077867058	0,505996277
2	0,0475	0,649751566	0,088528161	0,481034145
3	0,0495	0,677777778	0,092261859	0,508327724
4	0,05455	0,734251526	0,10155513	0,551643225
5	0,09355	0,827251975	0,168091347	0,621762567
6	0,09225	0,840752146	0,166257679	0,581215895
7	0,0535	0,684752427	0,099245877	0,53021314
8	0,05675	0,685752575	0,104825114	0,558019961
9	0,15955	0,714252525	0,260834656	0,390904893
Average	0,074515	0,726185153	0,135160925	0,525462011

Table 9.5: Latent Dirichlet Allocation Experimental Result with Top 70 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,07975	0,845254755	0,14574858	0,518686856
1	0,03775	0,679654785	0,07152717	0,553371552
2	0,03925	0,667714314	0,07414175	0,479167232
3	0,0431	0,754252546	0,08154056	0,53511957
4	0,0444	0,765251312	0,08393035	0,552776045
5	0,07751	0,881532457	0,14249125	0,619844655
6	0,07575	0,867752369	0,13933668	0,579308242
7	0,044	0,702753636	0,08281489	0,534407184
8	0,04725	0,730752575	0,08876078	0,566469989
9	0,12975	0,731515285	0,22040621	0,417530924
Average	0,06185	0,762643403	0,11442055	0,535668225

Table 9.6: Latent Dirichlet Allocation Experimental Result with Top 90 Operation Id

Appendix B: 10 Fold Cross Validation Proposed Model Experimental Results for the First Data Set

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,33	0,23991894	0,27784039	0,244446598
1	0,242105263	0,332505339	0,280194247	0,338010767
2	0,2	0,179305047	0,189087963	0,17913336
3	0,248648649	0,302534417	0,272957493	0,339138071
4	0,261538462	0,304848784	0,281537703	0,312623332
5	0,461538462	0,335760302	0,388728292	0,370568358
6	0,375	0,375918998	0,375458937	0,355656911
7	0,394444444	0,541361786	0,456370436	0,533638804
8	0,310526316	0,240198563	0,270872001	0,29624454
9	0,689473684	0,340390544	0,455769442	0,304400506
Average	0,351327528	0,319274272	0,334534863	0,327386125

Table 9.7: The Proposed Model Experimental Result with Top 5 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,27	0,33550907	0,299210875	0,309141514
1	0,178947368	0,49729652	0,263188785	0,478600013
2	0,169230769	0,27973166	0,210882698	0,247482692
3	0,172972973	0,35851094	0,23335684	0,372945141
4	0,192307692	0,36339579	0,251514731	0,355685005
5	0,358974359	0,39666984	0,376881875	0,404854103
6	0,2525	0,5104258	0,337863826	0,464138559
7	0,233333333	0,59061171	0,334511142	0,521577235
8	0,231578947	0,44099301	0,303684078	0,426910423
9	0,568421053	0,42315768	0,485149037	0,329130926
Average	0,26282665	0,4196302	0,323214574	0,391046561

Table 9.8: The Proposed Model Experimental Result with Top 10 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,170833333	0,555315004	0,261286319	0,394331607
1	0,093859649	0,654456788	0,164174088	0,599712556
2	0,098290598	0,554271616	0,166971631	0,452036478
3	0,1	0,548353836	0,169152646	0,437908046
4	0,116239316	0,562346592	0,192655882	0,459749689
5	0,178632479	0,636038582	0,278927665	0,58155576
6	0,141666667	0,627226327	0,231129855	0,53049891
7	0,124074074	0,75504434	0,213125845	0,621451337
8	0,123684211	0,61371346	0,205877148	0,479736155
9	0,342982456	0,652999055	0,449741722	0,384259105
Average	0,149026278	0,61597656	0,239990467	0,494123964

Table 9.9: The Proposed Model Experimental Result with Top 30 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,129	0,66735	0,21621	0,429498443
1	0,06105	0,69079	0,11219	0,608801439
2	0,06974	0,61069	0,12519	0,452691618
3	0,07189	0,69052	0,13023	0,516843221
4	0,07846	0,63785	0,13973	0,467765615
5	0,12974	0,71041	0,21941	0,580938155
6	0,1065	0,68375	0,18429	0,556197801
7	0,08556	0,81485	0,15485	0,615418057
8	0,08737	0,71542	0,15572	0,530099755
9	0,23579	0,77072	0,3611	0,447429815
Average	0,10551	0,69923	0,18335	0,520568392

Table 9.10: The Proposed Model Experimental Result with Top 50 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,10143	0,71619	0,17769	0,430322032
1	0,04624	0,72965	0,08697	0,633294766
2	0,05385	0,64111	0,09935	0,453446619
3	0,05907	0,81757	0,11019	0,602292234
4	0,06044	0,66882	0,11086	0,475111496
5	0,10037	0,77558	0,17773	0,588429024
6	0,08429	0,73917	0,15132	0,55738574
7	0,06587	0,87991	0,12257	0,634685674
8	0,06729	0,76891	0,12376	0,565510062
9	0,1797	0,80844	0,29404	0,445328398
Average	0,08185	0,75453	0,14769	0,538580604

Table 9.11: The Proposed Model Experimental Result with Top 70 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,08306	0,77143	0,14997	0,43958202
1	0,03947	0,75334	0,07502	0,62837439
2	0,04416	0,70895	0,08314	0,501647662
3	0,04685	0,83086	0,08869	0,60013574
4	0,04929	0,73029	0,09234	0,53738074
5	0,08091	0,83134	0,14747	0,603700632
6	0,07278	0,80569	0,1335	0,564598641
7	0,05278	0,89646	0,09969	0,644322475
8	0,05468	0,80421	0,10239	0,575448831
9	0,14327	0,87714	0,24632	0,496218244
Average	0,06672	0,80097	0,12319	0,559140938

Table 9.12: The Proposed Model Experimental Result with Top 90 Operation Id

Appendix C: 10 Fold Cross Validation Latent Dirichlet Allocation Experimental Results for The Second Data Set

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,32244898	0,335056689	0,32863	0,365469357
1	0,317460318	0,355102041	0,33523	0,371190416
2	0,321088435	0,389818594	0,35213	0,426658848
3	0,377324263	0,362834467	0,36994	0,366802044
4	0,385941043	0,346712018	0,36528	0,362049251
5	0,407709751	0,308707483	0,35137	0,317484205
6	0,409070295	0,313265306	0,35481	0,326550371
7	0,376417234	0,338684807	0,35656	0,360826168
8	0,421315193	0,319977324	0,36372	0,32031564
9	0,392743764	0,301133787	0,34089	0,343669755
Average	0,373151927	0,337129252	0,35186	0,356101606

Table 9.13: Latent Dirichlet Allocation Experimental Result with Top 5 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,236734694	0,399410431	0,29727	0,40904967
1	0,236961451	0,434535147	0,30668	0,435211623
2	0,233786848	0,45675737	0,30927	0,469729808
3	0,267573696	0,443015873	0,33364	0,434974314
4	0,279365079	0,434421769	0,34005	0,425508613
5	0,306575964	0,407709751	0,34998	0,387383525
6	0,305895692	0,38324263	0,34023	0,362615654
7	0,284353742	0,421315193	0,33954	0,417860515
8	0,313605442	0,402675737	0,3526	0,366616983
9	0,311564626	0,391247166	0,34689	0,390072257
Average	0,277641723	0,417433107	0,33162	0,409902296

Table 9.14: Latent Dirichlet Allocation Experimental Result with Top 10 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,174172336	0,627709751	0,27268	0,514465391
1	0,169251701	0,651746032	0,26872	0,437667314
2	0,155827664	0,674943311	0,2532	0,455276324
3	0,173015873	0,668185941	0,27486	0,53475526
4	0,182199547	0,653605442	0,28496	0,525513631
5	0,199251701	0,650453515	0,30506	0,49832149
6	0,205442177	0,643424036	0,31144	0,469311831
7	0,186870748	0,649818594	0,29027	0,505069462
8	0,211451247	0,622993197	0,31574	0,464383167
9	0,208390023	0,618480726	0,31174	0,474797375
Average	0,186587302	0,646136054	0,28887	0,487956124

Table 9.15: Latent Dirichlet Allocation Experimental Result with Top 30 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,141678005	0,757619048	0,23872	0,535809724
1	0,132018141	0,78553288	0,22605	0,546897338
2	0,122494331	0,798208617	0,21239	0,511268581
3	0,136417234	0,796575964	0,23294	0,529516615
4	0,142993197	0,776054422	0,24149	0,547382945
5	0,157732426	0,780748299	0,26244	0,531861349
6	0,163809524	0,777755102	0,27062	0,493275786
7	0,147800454	0,786258503	0,24883	0,534862312
8	0,167755102	0,742675737	0,27369	0,489291267
9	0,168843537	0,761655329	0,27641	0,506692689
Average	0,148154195	0,77630839	0,24836	0,522685861

Table 9.16: Latent Dirichlet Allocation Experimental Result with Top 50 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,108390023	0,834784581	0,19187	0,545139011
1	0,100294785	0,86877551	0,17983	0,589838646
2	0,092539683	0,857346939	0,16705	0,595085401
3	0,10414966	0,863219955	0,18587	0,474828724
4	0,108979592	0,840680272	0,19295	0,562107917
5	0,122517007	0,850566893	0,21418	0,544984596
6	0,127664399	0,849705215	0,22198	0,504346105
7	0,113877551	0,864512472	0,20125	0,546046229
8	0,128390023	0,804353742	0,22143	0,494776531
9	0,131133787	0,844104308	0,227	0,522754651
Average	0,113793651	0,847804989	0,20034	0,537990781

Table 9.17: Latent Dirichlet Allocation Experimental Result with Top 70 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,089410431	0,869478458	0,16215	0,546191207
1	0,082131519	0,898095238	0,1505	0,593980946
2	0,077029479	0,895963719	0,14186	0,60308996
3	0,085873016	0,898548753	0,15676	0,575547945
4	0,090453515	0,883038549	0,1641	0,565929527
5	0,100839002	0,887845805	0,18111	0,552384618
6	0,105124717	0,889614513	0,18803	0,512471159
7	0,093650794	0,910861678	0,16984	0,557105188
8	0,105396825	0,84755102	0,18748	0,496957083
9	0,107573696	0,880589569	0,19173	0,527846109
Average	0,093748299	0,88615873	0,16936	0,553150374

Table 9.18: Latent Dirichlet Allocation Experimental Result with Top 90 Operation Id

Appendix D: 10 Fold Cross Validation Proposed Model Experimental Results for The Second Data Set

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,514285714	0,465655477	0,48876	0,441982657
1	0,6	0,43444208	0,50397	0,390170516
2	0,706122449	0,437698154	0,54041	0,37639296
3	0,620408163	0,497391045	0,55213	0,455245852
4	0,689795918	0,4593837	0,55149	0,397766422
5	0,604081633	0,419544154	0,49518	0,337394159
6	0,710204082	0,506413916	0,59124	0,421006547
7	0,873469388	0,374481421	0,52422	0,272588444
8	0,657142857	0,689313248	0,67284	0,612805787
9	0,612244898	0,359312544	0,45285	0,31738134
Average	0,65877551	0,464363574	0,53731	0,402273468

Table 9.19: The Proposed Model Experimental Result with Top 5 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,414285714	0,568035474	0,47913	0,466052172
1	0,473469388	0,546388518	0,50732	0,423824653
2	0,528571429	0,529539132	0,52905	0,405921376
3	0,434693878	0,609353123	0,50741	0,523764278
4	0,53877551	0,548607324	0,54365	0,41272571
5	0,465306122	0,567328407	0,51128	0,366488379
6	0,491836735	0,617393362	0,54751	0,476502316
7	0,671428571	0,524858416	0,58916	0,3297113
8	0,473469388	0,78579782	0,5909	0,647058979
9	0,483673469	0,501775909	0,49256	0,378936522
Average	0,49755102	0,579907748	0,5298	0,443098568

Table 9.20: The Proposed Model Experimental Result with Top 10 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,240136054	0,776606736	0,36684	0,530802886
1	0,268707483	0,681309536	0,38541	0,422702464
2	0,332653061	0,697607884	0,45049	0,473507592
3	0,237414966	0,740767545	0,35958	0,558201289
4	0,297278912	0,702499874	0,41777	0,443173674
5	0,229931973	0,776243035	0,35478	0,451138815
6	0,230612245	0,753693617	0,35316	0,508543485
7	0,353741497	0,753046488	0,48136	0,498241678
8	0,239455782	0,8947359	0,3778	0,652955678
9	0,285714286	0,686656794	0,40352	0,537843232
Average	0,271564626	0,746316741	0,39507	0,507711079

Table 9.21: The Proposed Model Experimental Result with Top 30 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,164081633	0,872232877	0,2762	0,583636299
1	0,180816327	0,730684559	0,28989	0,526957844
2	0,225714286	0,763584331	0,34843	0,487792582
3	0,168571429	0,809645654	0,27904	0,589184323
4	0,20244898	0,765843214	0,32024	0,565658335
5	0,15755102	0,856020396	0,26612	0,481059048
6	0,149387755	0,792279766	0,25138	0,511778317
7	0,233061225	0,816721948	0,36264	0,518257737
8	0,15755102	0,932630663	0,26956	0,658607605
9	0,2	0,80011878	0,32001	0,571310511
Average	0,183918367	0,813976219	0,29835	0,54942426

Table 9.22: The Proposed Model Experimental Result with Top 50 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,124489796	0,899805052	0,21872	0,587721459
1	0,139067055	0,779543762	0,23603	0,549984468
2	0,171428571	0,806196265	0,28274	0,500535823
3	0,134402332	0,887164314	0,23344	0,635363397
4	0,152478134	0,787788312	0,2555	0,564506987
5	0,118075802	0,884125791	0,20833	0,48579261
6	0,110204082	0,852994167	0,19519	0,55484775
7	0,173760933	0,841319142	0,28803	0,520695549
8	0,11574344	0,942223263	0,20616	0,661610026
9	0,148979592	0,823878661	0,25233	0,467791426
Average	0,138862974	0,850503873	0,23765	0,55288495

Table 9.23: The Proposed Model Experimental Result with Top 70 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,099546485	0,914057468	0,17954	0,587919418
1	0,112244898	0,788222239	0,19651	0,550162385
2	0,139002268	0,836240514	0,23838	0,511014985
3	0,110430839	0,925509259	0,19732	0,657324653
4	0,123129252	0,80200178	0,21348	0,563885701
5	0,094331066	0,909681641	0,17094	0,59425928
6	0,087981859	0,867936926	0,15977	0,55066438
7	0,141043084	0,881687591	0,24318	0,542575543
8	0,092743764	0,962361785	0,16918	0,665297247
9	0,120181406	0,837863499	0,21021	0,568939952
Average	0,112063492	0,87255627	0,19785	0,579204354

Table 9.24: The Proposed Model Experimental Result with Top 90 Operation Id

Appendix E: 10 Fold Cross Validation Latent Dirichlet Allocation Experimental Results for The Third Data Set

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,32244898	0,335056689	0,328631958	0,365469357
1	0,317460318	0,355102041	0,33522782	0,371190416
2	0,321088435	0,389818594	0,352131115	0,326658848
3	0,377324263	0,362834467	0,369937535	0,366802044
4	0,385941043	0,346712018	0,365276295	0,362049251
5	0,407709751	0,308707483	0,351368016	0,317484205
6	0,409070295	0,313265306	0,354814385	0,326550371
7	0,376417234	0,338684807	0,356555543	0,360826168
8	0,421315193	0,319977324	0,363719598	0,32031564
9	0,392743764	0,301133787	0,340891319	0,343669755
Average	0,373151927	0,337129252	0,354227125	0,346101606

Table 9.25: Latent Dirichlet Allocation Experimental Result with Top 5 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,236734694	0,399410431	0,297272752	0,30904967
1	0,236961451	0,434535147	0,306682355	0,335211623
2	0,233786848	0,45675737	0,30927452	0,369729808
3	0,267573696	0,443015873	0,333636743	0,334974314
4	0,279365079	0,434421769	0,340051858	0,325508613
5	0,306575964	0,407709751	0,349983227	0,387383525
6	0,305895692	0,38324263	0,34022856	0,362615654
7	0,284353742	0,421315193	0,339543221	0,417860515
8	0,313605442	0,402675737	0,352602599	0,366616983
9	0,311564626	0,391247166	0,346888821	0,390072257
Average	0,277641723	0,417433107	0,333480201	0,359902296

Table 9.26: Latent Dirichlet Allocation Experimental Result with Top 10 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,174172336	0,627709751	0,272682668	0,414465391
1	0,169251701	0,651746032	0,268719681	0,337667314
2	0,155827664	0,674943311	0,253198157	0,455276324
3	0,173015873	0,668185941	0,274860971	0,43475526
4	0,182199547	0,650453515	0,284661982	0,49832149
5	0,199251701	0,650453515	0,305056311	0,49832149
6	0,205442177	0,643424036	0,311442328	0,469311831
7	0,186870748	0,649818594	0,290268038	0,505069462
8	0,211451247	0,622993197	0,315737469	0,464383167
9	0,208390023	0,618480726	0,311742102	0,474797375
Average	0,186587302	0,645820862	0,289526166	0,45523691

Table 9.27: Latent Dirichlet Allocation Experimental Result with Top 30 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,141678005	0,757619048	0,238715238	0,435809724
1	0,132018141	0,77553288	0,225627885	0,466897338
2	0,122494331	0,788208617	0,2120364	0,491268581
3	0,136417234	0,786575964	0,232509876	0,469516615
4	0,142993197	0,766054422	0,241000733	0,547382945
5	0,157732426	0,770748299	0,261872963	0,531861349
6	0,163809524	0,767755102	0,270009389	0,493275786
7	0,147800454	0,776258503	0,24832043	0,534862312
8	0,167755102	0,732675737	0,273002851	0,489291267
9	0,168843537	0,751655329	0,275746444	0,506692689
Average	0,148154195	0,76730839	0,248355222	0,496685861

Table 9.28: Latent Dirichlet Allocation Experimental Result with Top 50 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,108390023	0,824784581	0,191600627	0,445139011
1	0,100294785	0,80877551	0,178459171	0,489838646
2	0,092539683	0,837346939	0,166660791	0,495085401
3	0,10414966	0,843219955	0,185399806	0,474828724
4	0,108979592	0,810680272	0,192131045	0,462107917
5	0,122517007	0,820566893	0,213201391	0,544984596
6	0,127664399	0,819705215	0,220921533	0,504346105
7	0,113877551	0,844512472	0,200692849	0,546046229
8	0,128390023	0,804353742	0,221434866	0,494776531
9	0,131133787	0,804104308	0,2254939	0,522754651
Average	0,113793651	0,821804989	0,199906639	0,497990781

Table 9.28: Latent Dirichlet Allocation Experimental Result with Top 70 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,859478458	0,089410431	0,161971207	0,546191207
1	0,858095238	0,082131519	0,149914188	0,493980946
2	0,895963719	0,077029479	0,141862488	0,50308996
3	0,878548753	0,085873016	0,156453605	0,575547945
4	0,863038549	0,090453515	0,163745191	0,565929527
5	0,877845805	0,100839002	0,180898067	0,552384618
6	0,869614513	0,105124717	0,187574228	0,512471159
7	0,890861678	0,093650794	0,169484706	0,557105188
8	0,84755102	0,105396825	0,187479697	0,496957083
9	0,860589569	0,107573696	0,191242126	0,527846109
Average	0,093748299	0,87015873	0,169260932	0,533150374

Table 9.28: Latent Dirichlet Allocation Experimental Result with Top 90 Operation Id

Appendix F: 10 Fold Cross Validation Proposed Model Experimental Results for The Third Data Set

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,411337869	0,391587556	0,401219804	0,408225291
1	0,436281179	0,395774104	0,415041636	0,402370171
2	0,4	0,405292651	0,402628933	0,416941651
3	0,415873016	0,373983049	0,393817217	0,377234813
4	0,438548753	0,384278549	0,409623935	0,383469263
5	0,418594104	0,361324653	0,38785673	0,363751491
6	0,560909091	0,376121798	0,450294944	0,344736754
7	0,395464853	0,331469901	0,360650512	0,326525815
8	0,521088435	0,389719482	0,445930061	0,369423026
9	0,481179138	0,341718078	0,399630979	0,341507828
Average	0,447927644	0,375126982	0,408307638	0,37341861

Table 9.29: The Proposed Model Experimental Result with Top 5 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,326303855	0,484345933	0,389919167	0,451010886
1	0,340136054	0,510539801	0,408270653	0,455775307
2	0,3025	0,49198804	0,374647256	0,458665843
3	0,315873016	0,444306827	0,369240355	0,4068714
4	0,343764172	0,506968366	0,409711755	0,449679692
5	0,337414966	0,448802277	0,385218224	0,407232771
6	0,454090909	0,52179026	0,485592347	0,39624816
7	0,304988662	0,408902635	0,349382793	0,372835044
8	0,407256236	0,484131804	0,442379048	0,406105742
9	0,370975057	0,412469471	0,390623407	0,358555563
Average	0,350330293	0,471424541	0,401955159	0,416298041

Table 9.30: The Proposed Model Experimental Result with Top 10 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,203930461	0,662418075	0,31185422	0,50472997
1	0,202418745	0,685582412	0,312555295	0,497049652
2	0,177954546	0,647633555	0,279193304	0,498830609
3	0,186470144	0,593469786	0,283776717	0,446041293
4	0,1994709	0,631220376	0,303145344	0,481477421
5	0,211942555	0,590689295	0,311954225	0,455352198
6	0,266666667	0,721749036	0,389444257	0,441468935
7	0,187755102	0,540681742	0,278722188	0,40655003
8	0,251473923	0,669373047	0,365597915	0,452550694
9	0,237490552	0,597892822	0,33994906	0,414283062
Average	0,212557359	0,634071014	0,318383991	0,459833386

Table 9.31: The Proposed Model Experimental Result with Top 30 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,147755102	0,784541928	0,248676267	0,556048939
1	0,143537415	0,794497726	0,243146861	0,529728942
2	0,126772727	0,776375047	0,217955876	0,566681535
3	0,146281179	0,745632232	0,244579711	0,513298264
4	0,135714286	0,761661202	0,230379161	0,507194186
5	0,161004566	0,731685889	0,263931956	0,481017523
6	0,123681818	0,807320579	0,214501869	0,465885598
7	0,136621315	0,759436204	0,231581502	0,472214398
8	0,143877551	0,765732358	0,242239438	0,504738021
9	0,142562358	0,73334254	0,238717793	0,49022656
Average	0,140780832	0,766022571	0,237849338	0,508703397

Table 9.32: The Proposed Model Experimental Result with Top 50 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,12457726	0,837814767	0,216902603	0,562725782
1	0,128940719	0,855580238	0,224107228	0,558871571
2	0,118863636	0,8366266	0,208153838	0,584800987
3	0,117677357	0,844704296	0,206576192	0,562032437
4	0,123376623	0,862500224	0,215873546	0,523279504
5	0,142999676	0,815167526	0,243315972	0,543257288
6	0,11948052	0,862765905	0,209893803	0,477738766
7	0,12457726	0,820375038	0,216307372	0,49758029
8	0,113851636	0,832764511	0,200316891	0,523700226
9	0,114402332	0,840735591	0,201399421	0,508305334
Average	0,122874702	0,84090347	0,214418143	0,534229218

Table 9.33: The Proposed Model Experimental Result with Top 70 Operation Id

Iteration Number	Precision	Recall	F1-Measure	MRR
0	0,113121693	0,86788315	0,200154794	0,56569266
1	0,117427564	0,880292433	0,207213639	0,564305528
2	0,110959596	0,871898133	0,196865654	0,592762798
3	0,117906274	0,866606282	0,207571385	0,573885021
4	0,112290249	0,881262095	0,199198645	0,577781744
5	0,120629882	0,87340604	0,211982012	0,547480471
6	0,101666667	0,898243247	0,182659249	0,48681886
7	0,093953137	0,919512069	0,170486451	0,517294787
8	0,108012094	0,866271314	0,192075073	0,526678824
9	0,109952129	0,909512617	0,196186968	0,518666613
Average	0,110591928	0,883488738	0,196577052	0,54713673

Table 9.34: The Proposed Model Experimental Result with Top 90 Operation Id

REFERENCES

- [1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.
- [2] Aldous, David J. "Exchangeability and related topics." *École d'Été de Probabilités de Saint-Flour XIII—1983*. Springer, Berlin, Heidelberg, 1985. 1-198.
- [3] Paul, Michael J., and Mark Dredze. "A model for mining public health topics from Twitter." *Health* 11 (2012): 16-6.
- [4] Steyvers, Mark, et al. "Probabilistic author-topic models for information discovery." *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004.
- [5] Neal, Radford M. "Probabilistic inference using Markov chain Monte Carlo methods." (1993).
- [6] Heath, David, and William Sudderth. "De Finetti's theorem on exchangeable variables." *The American Statistician* 30.4 (1976): 188-189.
- [7] Blei, David M. "Probabilistic topic models." *Communications of the ACM* 55.4 (2012): 77-84.
- [8] Chen, J. H., Goldstein, M. K., Asch, S. M., Mackey, L., & Altman, R. B. (2017). Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. *Journal of the American Medical Informatics Association*, 24(3), 472–480.
- [9] ONC. Health information technology: standards, implementation specifications, and certification criteria for electronic health record technology, 2014 edition; revisions to the permanent certification program for health information technology. Final rule. *Fed Regist.* 2012;77:54163–292.
- [10] Hasan, Sharique, et al. "Automatic detection of omissions in medication lists." *Journal of the American Medical Informatics Association* 18.4 (2011): 449-458.
- [11] Hasan, Sharique, et al. "Towards a collaborative filtering approach to medication reconciliation." *AMIA Annual Symposium Proceedings*. Vol. 2008. American Medical Informatics Association, 2008.
- [12] Barajas, Karla L. Caballero, and Ram Akella. "Incorporating Statistical Topic Models in the Retrieval of Healthcare Documents." *CLEF (Working Notes)*. 2013.

- [13] Caballero, Karla L., Joel Barajas, and Ram Akella. "The generalized dirichlet distribution in enhanced topic detection." Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012.
- [14] Choi, Edward, et al. "Doctor ai: Predicting clinical events via recurrent neural networks." Machine Learning for Healthcare Conference. 2016.
- [15] Cios, Krzysztof J., and G. William Moore. "Uniqueness of medical data mining." Artificial intelligence in medicine 26.1-2 (2002): 1-24.
- [16] Mate, Sebastian, et al. "Ontology-based data integration between clinical and research systems." PloS one 10.1 (2015): e0116656.
- [17] Trevor, Hastie, Tibshirani Robert, and Friedman JH. "The elements of statistical learning: data mining, inference, and prediction." (2009).
- [18] Ekstrand, Michael D., John T. Riedl, and Joseph A. Konstan. "Collaborative filtering recommender systems." Foundations and Trends® in Human-Computer Interaction 4.2 (2011): 81-173.
- [19] Lu, J., Wu, D., Mao, M., Wang, W., and Zhang, G. 2015. Recommender system application developments: a survey. Decision Support Systems, 74, pp. 12-32.
- [20] Wu, M. L., Chang, C. H., and Liu, R. Z. 2014. Integrating content-based filtering with collaborative filtering using co-clustering with augmented matrices. Expert Systems with Applications, 41(6), pp. 2754-2761.
- [21] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web, pages 285–295. ACM.
- [22] Sarwar, Badrul Munir, et al. "Item-based collaborative filtering recommendation algorithms." Www 1 (2001): 285-295.
- [23] Levinas, Claudio Adrian. An analysis of memory based collaborative filtering recommender systems with improvement proposals. MS thesis. Universitat Politècnica de Catalunya, 2014.
- [24] Hao, Fang, and Rachael Hageman Blair. "A comparative study: classification vs. user-based collaborative filtering for clinical prediction." BMC medical research methodology 16.1 (2016): 172.

- [25] Thi Do, Minh-Phung & Van Nguyen, Dung & of Loc Nguyen, Academic Network. (2010). Model-based approach for Collaborative Filtering.
- [26] Logistic Regression for Machine Learning, <https://machinelearningmastery.com/logistic-regression-for-machine-learning/> (March, 2019)
- [27] World Health Organization <https://www.who.int/classifications/icd/en/> (April, 2019)
- [28] Shortliffe, Edward H., et al. "Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system." *Computers and biomedical research* 8.4 (1975): 303-320.
- [29] Deerwester, Scott, et al. "Indexing by latent semantic analysis." *Journal of the American society for information science* 41.6 (1990): 391-407.
- [30] Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." *Discourse processes* 25.2-3 (1998): 259-284.
- [31] Ozsoy, Makbule Gulcin, Ilyas Cicekli, and Ferda Nur Alpaslan. "Text summarization of turkish texts using latent semantic analysis." *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 2010.
- [32] T. Hofmann. Probabilistic latent semantic indexing. In *Proc.of SIGIR '99*, pages 50–57, 1999.
- [33] Y. Gong, X. Liu: Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, United States 2001, pp. 19-25
- [34] Bhandari, Harendra & Ito, Takahiko & Shimbo, Masashi & Matsumoto, Yuji. (2008). Generic text summarization using probabilistic latent semantic indexing.
- [35] Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).
- [36] Kapa, Suraj, et al. "A reliable billing method for internal medicine resident clinics: financial implications for an academic medical center." *Journal of graduate medical education* 2.2 (2010): 181-187.

- [37] Adams, Diane L., Helen Norman, and Valentine J. Burroughs. "Addressing medical coding and billing part II: a strategy for achieving compliance. A risk management approach for reducing coding and billing errors." *Journal of the National Medical Association* 94.6 (2002): 430.
- [38] Su, Xiaoyuan, and Taghi M. Khoshgoftaar. "A survey of collaborative filtering techniques." *Advances in artificial intelligence 2009* (2009).
- [39] Koren, Yehuda. "Factorization meets the neighborhood: a multifaceted collaborative filtering model." *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008.
- [40] Russell, Stuart J., and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited., 2016.
- [41] Xu, Rui, and Donald C. Wunsch. "Survey of clustering algorithms." (2005).
- [42] Lipton, Zachary C., Charles Elkan, and Balakrishnan Naryanaswamy. "Optimal thresholding of classifiers to maximize F1 measure." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Berlin, Heidelberg, 2014.
- [42] Craswell, Nick. "Mean reciprocal rank." *Encyclopedia of Database Systems* (2009): 1703-1703.
- [43] Demšar, Janez. "Statistical comparisons of classifiers over multiple data sets." *Journal of Machine learning research* 7.Jan (2006): 1-30.



HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
THESIS ORIGINALITY REPORT

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
TO THE DEPARTMENT OF COMPUTER ENGINEERING

Date: 19/09/2019

Thesis Title / Topic: TOPIC MODEL BASED RECOMMENDATION SYSTEM TO IDENTIFY OPERATIONS THAT ARE MISSING IN THE TREATMENT

According to the originality report obtained by my thesis advisor by using the *Turnitin* plagiarism detection software and by applying the filtering options stated below on 19/09/2019 for the total of 65 pages including the a) Title Page, b) Introduction, c) Main Chapters, d) Conclusion sections of my thesis entitled as above, the similarity index of my thesis is 10 %.

Filtering options applied:

1. Bibliography/Works Cited excluded
2. Quotes excluded
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Science and Engineering Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

Date and Signature

Name Surname: KAMURAN NUR KIRAZ

Student No: N16126568

Department: COMPUTER ENGINEERING

Program: COMPUTER ENGINEERING

Status: Masters Ph.D. Integrated Ph.D.

19.09.2019

ADVISOR APPROVAL

APPROVED.

Asst. Prof. Dr. GÖNENÇ ERCAN

ÖZGEÇMİŞ

Adı Soyadı : Kamuran Nur KİRAZ
Doğum yeri : Ankara
Doğum tarihi : 07/08/1989
Medeni hali : Evli
Yazışma adresi : Bağlıca Mahallesi, 1154. Sokak, Bahçelievler Sitesi 5. Kısım 1/25
Etimesgut/ANKARA
Telefon : 506 895 31 41
Elektronik posta adresi : nur.seker.k@gmail.com
Yabancı dili : İngilizce

EĞİTİM DURUMU

Lisans : Çankaya Üniversitesi – Bilgisayar Mühendisliği (Burslu)
Çankaya Üniversitesi – Elektronik ve Haberleşme Mühendisliği(Burslu)
Yüksek Lisans : Hacettepe Üniversitesi – Bilgisayar Mühendisliği

İş Tecrübesi

Innova Bilişim çözümleri 2012