ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF ARTS AND SOCIAL SCIENCES

AN EVALUATION OF ISTANBUL REAL ESTATE POSTS BASED ON TEXT ANALYSIS AND TOPIC MODELLING

MBA THESIS

İmge KIZILTAN

Department of Management

Management MBA Programme

JUNE 2018

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF ARTS AND SOCIAL SCIENCES

AN EVALUATION OF ISTANBUL REAL ESTATE POSTS BASED ON TEXT ANALYSIS AND TOPIC MODELLING

MBA THESIS

İmge KIZILTAN
(403161012)

Department of Management

Management Programme

Thesis Advisor: Assoc. Prof. Dr. Tolga KAYA

JUNE 2018

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ SOSYAL BİLİMLER ENSTİTÜSÜ

İSTANBUL GAYRİMENKUL İLANLARININ METİN MADENCİLİĞİ VE KONU MODELLEMESİ İLE ANALİZİ

YÜKSEK LİSANS TEZİ

İmge KIZILTAN
(403161012)

İşletme Anabilim Dalı

İşletme Yüksek Lisans Programı

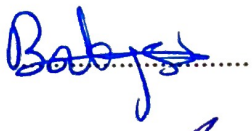
Tez Danışmanı: Doç. Dr. Tolga KAYA

HAZİRAN 2018

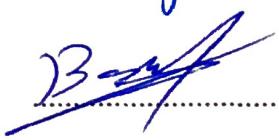
İmge KIZILTAN, a MBA student of ITU Graduate School of Arts and Social Sciences student ID 403161012, successfully defended the thesis entitled "AN EVALUATION OF ISTANBUL REAL ESTATE POSTS BASED ON TEXT ANALYSIS AND TOPIC MODELLING", which she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Assoc. Prof. Dr. Tolga KAYA**
Istanbul Technical University

Jury Members : **Assoc. Prof. Dr. Başar ÖZTAYŞİ**
Istanbul Technical University



Dr. Barış SOYBİLEN
Bilgi University



Date of Submission : 04 May 2018
Date of Defense : 05 June 2018

To my family,

FOREWORD

Firstly, I would like to thank my advisor, Assoc. Prof. Dr. Tolga KAYA for his understanding and valuable support. Despite his loaded schedule, he always had the willing to help and support me.

In addition, I would like to show my graditutes to my parents who, with their support have never left me alone throughout my studies.

May 2018 İmge KIZILTAN

TABLE OF CONTENTS

Page

ABBREVIATIONS

API : Application Programming Interface
CRISP-DM : Cross-industry standard process for data mining
FSBO : For Sale by Owner
HMM : Hidden Markov Model
HTML : Hyper Text Markup Library
HDI : Human Development Index
IR : Information Retrieval
LDA : Latent Dirichlet Allocation
NLP : Natural Language Processing
PLSI : Probabilistic Latent Semantic Indexing
SEMMA : Sample, Explore, Modify, Model, and Assess
TEM : Trans European Motorway
TM : Text Mining
TUIK : Turkish Statistical Institute
ZMOT : Zero Moment of Truth

SYMBOLS

α	: Dirichlet topic proportions hyperparameter
η	: Dirichlet topic hyperparameter
θ_d	: Topic proportion
z_{dn}	: Topic distribution
w_{dn}	: Observed words
β_k	: Topics
N_d	: Number of words in a document
K	: Number of topics
D	: Number of documents

LIST OF TABLES

LIST OF FIGURES

Page

EVALUATION OF ISTANBUL REAL ESTATE POSTS BASED ON TEXT ANALYSIS AND TOPIC MODELLING

SUMMARY

Real estate, which has a potential to trigger other sectors in the economy, is one of the most influential sectors in the global arena. There are many factors that distinguish this sector from the others. As real estate products are expensive investment tools, selling transactions are difficult to take place. Compared to other investments, people consider and elaborate more factors before making decisions. For this reason, real estate agencies and real estate agents play a key role in accelerating the slow-moving real estate sector.

With the rapid increase of the marketing activities of real estate stations on the internet, it has become much easier to compare the alternatives and find the desired real estate products. Today, potential property buyers check the internet before making decisions. Real estate websites can provide large scale data which is based on post records. Various analyses can be conducted using statistical and machine learning algorithms based on these data. In this study, some of the text mining techniques have been applied to the scraped data of a Turkish real estate website.

The purpose of this research is to investigate the topics embedded in the real estate post descriptions of Istanbul using the Latent Dirichlet Allocation (LDA) topic modelling method. The LDA can be considered as a technique which can detect the hidden topics in text by using the distributions of the observed words. By applying LDA to the announcements of real estate advertisements; the topics used to attract the attention of the consumers, the probabilities of the topics and the frequent words are determined. These figures are also evaluated by location, type of property (house/plot/office) and transaction type (sale/rent). Finally, human development levels of the districts are used to compare and understand the text patterns used in the posts of real estates at different regions of Istanbul.

To the author's knowledge, this study is the first attempt to apply LDA technique in the context of clustering real estate ads. Moreover, it is the first attempt to evaluate Istanbul real estate sector web posts by means of text mining and topic modelling tools. Fnally, investigating the relations between district based human development level and the real estate posts can be considered as a contribution to the gap in the literature.

İSTANBUL GAYRİMENKUL İLANLARININ METİN MADENCİLİĞİ VE KONU MODELLEMESİ İLE ANALİZİ

ÖZET

Günümüzde gayrimenkul, diğer bir çok sektörü tetikleyen, küresel öneme sahip sektörlerden biridir. Bu sektörü diğer sektörlerden ayıran bir çok etmen vardır. Öncelikte yatırım olarak pahalı bir alternatif olduğu için hem satışı, hem de süreçleri zorludur. İnsanlar ev alma kararı vermeden önce her faktörü düşünürler ve en ufak bir uyumsuzlukta olumsuz bir karar verme olasılıkları diğer yatırımlara kıyaslandığında daha yüksektir.. Bu nedenle aracılar olmadan, alıcılar ve satıcılar pazar eğilimlerini kaçırabilirler. Bir profesyonelin yardımı olmadan, satıcı potansiyel alıcının taleplerini tahmin edemeyebilir ve zarar edebilir. Aynı şekilde alıcı da tek başına isteklerine uyan gayrimenkulü bulmaya çalıştığında fazlasıyla zaman kaybedebilir. Bu nedenle gayrimenkul ajansları ve emlakçılar yavaş ilerleyen gayrimenkul sektörünü hızlanırdırmada kilit rol oynarlar. Bununla birlikte, devletin gayrimenkul alanında ve diğer alanlarda yaptığı yatırımlar ve sahip olduğu yetkinlikler dolayısıyla gayrimenkul sektörüne büyük etkisi vardır. Devletin yaptığı yatırımlarla birlikte emlak piyasası da şekillenmektedir.

Gayrimenkul satışlarının internet ortamında pazarlanmasının artması ile birlikte, alternatifleri karşılaştırmak ve istenileni en yakın şekilde bulmak fazlasıyla kolaylaşmıştır. Bugün, potansiyel emlak alıcıları, herhangi bir işlem yapmadan önce interneti kontrol ederler. Gayrimenkul web sitelerinin bir diğer olumlu yönü ise, kayıtlar tarafından oluşturulan, büyük veri olarak adlandırılabilecek bir veri tabanının varlığıdır. Bu veriler kullanılarak istatistiksel ve makine öğrenimi algoritmaları kullanılarak çeşitli analizlerin yapılabilir.

Veri madenciliği ve makine öğrenmesinde amaç, büyük veri setleri kullanarak saklı bilgiyi bulmaktır. Veri madenciliğinin bir çok alanı vardır ve bu araştırma, veri madenciliğinin metin madenciliği alanına odaklanmaktadır. Veri madenciliği gibi, metin madenciliği de birçok uygulamaya sahiptir. Bu araştırmada, metin madenciliği uygulamalarından bazıları, bir Türk gayrimenkul web sitesinden indirilen verilere uygulanmıştır. Veri görselleştirme, kelime çağrışımları, anlamsal analiz ve konu modellemesi ile İstanbul'daki emlak ilanlarının açıklamaları incelenmiştir.

Bu araştırmanın amacı, Gizli Dirichlet Tahsisi (GDT) konu modellemesi yöntemini kullanarak, gayrimenkul ilan açıklamalarının altında yatan konuları tespit etmektir. GDT'nin işleyişi, gözlemlenen kelimelerin metinler içindeki dağılımlarını kullanarak gizli konuları tanımlamak olarak açıklanabilir. Çalışmada, emlak ilanlarının açıklamalarına GDT uygulanarak; emlakçıların tüketicilerin ilgisini çekmek için kullandıkları konular, bu konularda en çok geçen kelimeler ve kullanılan konu olasılıklarının; emlağın bulunduğu konuma göre, emlağın tipine göre (ev, arsa, ofis) ve emlağın satılık veya kiralık olmasına göre değişiklik gösterip göstermediği araştırılmıştır.

Çalışmada üç ana konu üzerinde durulmuştur; gayrimenkul sektörüne genel bakış, metin madenciliği uygulamalarına genel bakış ve İstanbul'daki emlak ilanlarının metin madenciliği yöntemleri ile incelenmesi. Çalışmada R programlama dili kullanılarak; ısı haritaları, kelime ilişkileri, kelime bulutları gibi veri görselleştirmesi yöntemlerinden faydalanılmıştır. Bununla birkte doğal dil işleme modellerinden veri ilişkilendirmesi ve duygu analizi yöntemleri ile ilan detayları yorumlanmıştır. Ayrıca GDT'nin bir java uygulaması olan MALLET kullanılarak konu modellemesi yapılmş ve son olarak sonuçlar tartışılmıştır.

Tezin uygulama sürecinde, öncelikle veri, kelimeleri köküne ayırma, istenmeyen kelimeleri atma gibi işlemlerden geçirlerek analize uygun hale getirilmiştir. Ardından, veri MALLET'e aktarılmış ve algoritma işletilmiştir. Belgedeki sözcük sayıları ve verinin boyutu göz önünde bulundurularak 2 ila 20 arasında konu sayısı belirlenmiştir. Bu işlemlerin sonucunda, anlamlı sonuç veren altı konulu konu modellemesi seçilmiştir. Uygulama sonucunda ortaya çıkan konular sırasıyla, gayrimenkulün dış özellikleri, gayrimenkulün tipi, gayrimenkulün konumu ve ulaşım olanakları, inşaat şirketi ve emlak şirketinin tanımı, emlağın bulunduğu çevrenin özellikleri ve son olarak gayrimenkulün iç özellikleridir.

Bu araştırmaların sonucunda, bazı ilçelerdeki ilanların belirli konularda yoğunlaştığını ve bazı ilçelerdeki ilanlarda da kimi konuların üzerinde pek durulmadığı görülmektedir. Konu dağılımlarına bakıldığında, genellikle büyüyen ilçelerin (Çekmeköy, Ümraniye, Kâğıthane ve Eyüp) açıklamalarında evlerin dış özelliklerinin öne çıktığı saptanmıştır. Büyüyen ilçelerde daha fazla yeni gayrimenkul inşa ediliyor olması ve genellikle yeni inşa edilen konutların güvenlik ve kapalı otopark gibi ayırıcı dış özelliklere sahip olması bu durumun ardında yatan etmen olarak yorumlanabilir.
Arazi satışının yoğun olduğu ilçeler (Şile, Arnavutköy, Esenler) ikinci konuyu diğer ilçelerden daha fazla kullanmışlardır. Ayrıca, konu oranlarına, gayrimenkul tipi parametresi ile bakıldığında ofis türü gayrimenkul ilanlarında, ikinci konu olan gayrimenkul tipinin yoğun olarak kullanıldığı görülmüştür. Bu konuda, emlak türü hakkında detay vermek için arazi ve ofis gibi kelimeler kullanılmaktadır, bu yüzden bu sonuç da anlamlıdır.
Üçüncü konu (konum/ulaşım) incelendiğinde ilk göze çarpan bulgulardan biri konuyla ilgisi olmayan ilçelerin (Tuzla, Çatalca, Şile) genellikle merkezden uzak oluşlarıdır. Bu ilçelerin ulaşım imkânlarının diğer ilçeler kadar iyi olmadığı göz önünde bulundurulduğunda bu sonuç da anlamlıdır.
Bununla birlikte, dördüncü başlığın (inşaat firmaları ve emlak ajansları) en çok Kartal ilçesi tarafından kullanıldığı gözlemlenmiştir. Kartal, markalı inşaat firmaları tarafından birçok konutun inşa edildiği büyüyen bir bölgedir.
Beşinci konu (mahalle) incelendiğinde, Sancaktepe'nin en çok bu konuyu kullanan ilçe olduğu görülmektedir. Günümüzde, Sancaktepe, göç alan yeni sayılabilecek bir ilçedir ve tüketiciler tarafından diğer ilçelere göre daha az biliniyor olması muhtemeldir. Satılacak gayrimenkulün çevresinin daha detaylı anlatılmasının ardında böyle bir sebebin yatıyor olabileceği düşünülmektedir.
Tüketicilerin konut tercihini etkileyen faktörler çeşitlilik göstermektedir. Öte yandan, iki kişi aynı insani gelişme düzeyini paylaşıyorsa bu çeşitliliğin bir ölçüde azalıp azalmadığı araştırmaya değer bir konudur. Bu araştırmada, İstanbul'un farklı bölgelerindeki tercih kalıplarını anlamak amacıyla, ilçelerin insani gelişme düzeyleri kullanılarak İstanbul, dört bölgeye ayrılmıştır. İnsani gelişme indeksi kullanılarak oluşturulan coğrafi bölgelerin konu yüzdelerine bakıldığında üç parametre ayrıştırılmıştır.

İlk bölgedeki ilanlarda konum/ulaşım konusunun diğer üç bölgeye kıyasla daha fazla kullanıldığı görülmektedir. Bunun ardında, ilk bölgedeki insanlar için taşınmazın yerinin ve ulaşım olanaklarının önemli olması, bu insanların genellikle aktif bir çalışma hayatına sahip olmaları yatıyor olabilir. İkinci bölgedeki ilanlarda, inşaat firmaları ve emlak ajansları diğer bölgelere oranla daha çok vurgulanmıştır. Bu bölge içindeki ilçelerde yeni gayrimenkul projeleri başlamıştır. Dördüncü bölgede ise "mahalle" vurgusunun, mahalle özelliklerinin tanıtımının yoğun olduğu görülmüştür. Bu bölgedeki ilçeler diğer bölgeler kadar bilinirliği olmayan, İstanbul'un gelişmekte olan bölgeleridir. Bu bölgedeki gayrimenkuller için verilen ilanlarında daha fazla mahalle açıklaması bulunmasının ardında böyle bir neden yatıyor olabileceği düşünülebilir.

Bu çalışma bilindiği kadarıyla GDT yönteminin gayrimenkul sektöründe uygulamayı amaçlayan ilk çalışmadır. Yine, bilindiği kadarıyla, metin madenciliği ve konu modellemesi yöntemleri ile İstanbul Gayrimenkul piyasasında verilen ilanları inceleyen ilk çalışmadır. Son olarak çalışmanın, insani gelişme ile emlak piyasasındaki bölgesel farklılıklar arasındaki ilişkiye ışık tutmaya çalışması bakımından da yazına özgün bir katkı sağladığı düşünülmektedir.

1. INTRODUCTION

With the rapid growth of web usage, unlimited amount of content in different fields are created every day. Real estate is one of the fields that started with a glance and became one of the most used areas in the internet. Decades ago, real estate sales channels only consisted of newspaper advertisements and real estate offices. At that time, consumers made great efforts to find out what they were looking for. Nowadays, using real estate web pages, instant recommendations can be received according to the desired parameters.

Real estate is an expensive, complex market that needs consultation for both the seller side and the buyer side. Without the agents, the buyers and sellers may miss market trends. Real estate is an expensive investment. Therefore, buying and selling a real estate product takes time and often requires expertise. Trying to sell a high priced product has its own principles. Nowadays, professional intermediaries use both their offices and the internet to reach the customers.

Due to the rapid increase of real estate marketing activities in the internet medium, it became easier to compare the alternatives and find the desired products. Today, potential real estate buyers firstly check out the internet before taking any action. Another positive aspect of real estate in the world wide web is that, a database which can be referred to as big data is formed by the postings and various analyses can be conducted using this data with statistical and machine learning algorithms.

In this research, some of the text mining applications are applied on the scraped data of a Turkish real estate agent website. The descriptions of the real estate postings in Istanbul is investigated by means of data visualization, word association, semantic analysis and topic modelling.

The aim of this study is to identify the underlying topics of real estate ads by using the LDA topic modelling method. LDA identifies latent topics using observed words by their distribution [1]. By applying LDA to the announcements of real estate advertisements; the topics that are used to gain the interests of the consumers, the

frequent words in the topics and the probabilities of the topics are determined. These figures are also evaluated by location, type of property (office/house/land etc.) and transaction type (sale/rent). Finally, human development levels of the districts are used to compare and gain insight on the text patterns used in the real estate posts at different areas of Istanbul.

Demand of customers vary in many ways. A parameter which is compulsory to a person may be unnecessary for the other. Generally, this variation decreases if the two persons share the same human development level. In this research, human development levels of the districts are used to understand the choice patterns at different regions of Istanbul.

To the author's knowledge, this study is the first attempt to apply LDA technique in the context of clustering real estate ads. Moreover, it is the first study to evaluate Istanbul real estate sector web posts by means of text mining and topic modelling tools. Finally, investigating the relations between district based human development level and the real estate posts can be considered as a contribution to the gap in the literature.

This study consists of three main sections; an overview of real estate sector in the world and in Turkey, an overview of the text mining models and applications, and the application of text mining models using the database of real estate postings in Istanbul. Using R programming language; data visualizations such as heatmaps, correlation plots and word-clouds are produced. Similarly, natural language processing applications like data association and sentiment analysis are conducted. Finally, using MALLET, a java implementation of LDA, a topic modeling application is conducted.

2. REAL ESTATE SECTOR

2.1 Real Estate

Real estate is considered as land, joined to the land or fitted to the land that is immovable according to law. It has specific characteristics that differs from other markets. To give some examples, firstly real estate transactions occur much rarer than other sectors because, only a few people buy more than five houses in their life. Secondly, government has high impact on the price level of the real estate parcels in order to control the zoning, environmental and health code of the place. Thirdly, for an average family, buying a house is highly costly and decision-making process is harder than investing other assets. Also, there are few participants during the real estate transactions and lastly, real estate markets have more local reach, since the property cannot be moved [2]. According to a survey, nowadays buyers purchase houses for different reasons such as aspiration of having a house of their own, the need for a larger home, job relocation and other factors [3]. Although it is hard to buy a house, there are many different initiatives to buy a house.

Overall, real estate is a market that is expensive, highly complex and requires both legal and technical consult. Without the intermediaries, the buyers and the sellers are unaware of the market trends and this can lead to settling for unrealistic prices. Thus, the intermediaries are needed to overcome the complexity and the imperfections of the real estate market. The intermediaries which can be called realtors, help both the sellers and buyers through all the transaction phases and represent them [2,4]. Realtors give information about the neighborhood of the real estate such as schools, hospitals and other institutions, which may change the selling price and that can be an important data for buyers to make decision. Also, the real estate agents can give advices to fasten the selling process of the house such as, setting up open houses and creating advertisements. Setting up the house before the invitation of the potential buyers is important because, it can increase the purchase possibility [5,12].

The Real Estate Transaction

In his article, Crowston divided real estate transaction process to five steps which are listing, searching, evaluation, negotiation and execution and suggests that these processes can be applied worldwide despite the little details [5]. Firstly, the listing stage includes processes that occurs before and during the seller put his/her house on market. Determining the fields in the house that can be emphasized or needs repairing, setting the price and doing some paperwork can be examples of the processes. Later, the house is put on the market by giving an advertisement. If the seller cooperates with a real estate agent, the commission rate the agent will get is determined before listing and it is indifferent about how the buyer is willing to purchase the house. After these processes, the house is entered in a database by the agent.

Secondly, at the searching stage, the possible buyers search for the houses and try to find the one that suits their criteria [5]. They look for the houses in many platforms such as the internet, newspapers, local realtors or a walkthrough in the area. It is known that, buyers have different priorities and choices while they are searching for a house and likewise, the houses on sale have different properties. To this extent, the process of matching potential buyers to the desired home is not always easy. To make this easier, the buyer takes one step a time and shortlists the possible houses. For example, firstly he/she defines the price range, then the neighborhood and the process continues. The aim of the buyer is to find a home which is closest to what he/she desires and that meets all the criteria that he/she has determined [6]. On the other hand, the buyer cannot get access the whole real estate parameters by himself/herself and trying to find the best option will become costly after a while. In this case, the buyer needs to get a professional help from the agents.

Thirdly, at the evaluation stage, the top houses determined by the buyers' criteria with the help of the agents are evaluated. This is the stage where the real estate sector differentiates the other goods that can be bought from the internet. The goods that are sold in e-commerce sites are easily purchasable without thinking about it a second time, but hardly anyone buys a house with only the information on the website [5]. So, the agents show and give more information about the promising houses and help buyers to make a decision. Also without the agents, buyers can search and look for the for sale by owner houses (FSBO). On the other hand, FSBO houses are generally not listed in e-commerce sites and newspapers so it is harder to find [10].

The fourth stage is the negotiation stage where the potential buyer makes an offer to the house and the seller negotiates with the highest offer. At this stage, sometimes a counter-offer occurs and the agent will advise the buyer about the worth of the house. Before or after determining the price that will be offered, inspecting the house for defects can be done to have a deeper understanding of the worthiness of the place.

Lastly, at the execution stage, when the contractions and the paperwork are done, the sale can be closed. This process is generally taken care of a third party which can be a lawyer [5]. After all these steps, the buyer will be the owner of the house and the agent will get his/her commission from the buyer/seller.

2.2 Real Estate in the Internet Platform

It is known that, with the increase of internet usage, many sectors added internet as another channel to reach their customers and real estate sector is one of these. Buxmann made a research about the reasons of the internet usage in the real estate market and found out that with the internet, not only information about house sales are available online, but also websites offer extra services and facts to the visitors. To give examples, sales information can be compared, mortgage calculations can be done, the building can be seen on the map and "customized business directory" that gives facts about the schools, hospitals in the area, can be seen [2]. Also, Buxmann made a survey with real estate agents asking the initiative to start being in the internet. Looking at the results of the survey, attracting buyers is the highest initiative and producing new listings, giving information and advice, and advertising the company's brand are secondary reasons for being in the internet [2].

The listings at the internet are generally submitted by the local realtors or national companies that work nationwide [7]. At the listings, the agents provide information about the market prices, architectural styles and much more. Nowadays, thanks to Google Maps and its derivatives, potential buyers can see the neighborhood of a house and see the location of the schools, hospitals and other commodities near the house. Also, without going to the neighborhood, the buyer can see the visuals in street mode of the map. According to the Google's Zero Moment of Truth (ZMOT) handbook for marketers, the sales tunnel which focuses on attracting the customer and gaining steady customer is not realistic anymore because the buyers look for many parameters and

alternatives before deciding to buy an item [7,8]. In real estate case, it became much harder. So, attracting the customer in the right way became much more important in this era.

According to Google, the real estate related searches increased by 253% from 2008 to 2012 and it is found out that in the year 2012, 90% of the potential buyers searched the internet before buying a house. In the same research, it is found out that when the buyers are online, they firstly look for the explanation about the house in the posting. Comparing the prices, getting directions to the house, comparing the features and searching for the listings in the company's inventory are the secondary things that they look for [7,9].

3. LITERATURE REVIEW

3.1 Real Estate

Real estate is one of the sectors that involves social principles. Many social studies have been conducted in this field. One of the studies tries to find the determinants of repurchase intentions of real estate agents by developing a model focusing on potential buyers looking for Scandinavian agents. It is found out that, by including ethics as a core competence in a real estate job, repurchase intentions increase. By giving attention to the virtue ethics such as being fair caring, being kind, showing integrity etc., will increase the perceived ethicality of the company. Non-verbal behaviors are also important to earn positive relationships and to keep the customer [11].

Mulliner et al. proposed a methodology to evaluate the state and the condition of the real estate market by doing rating engineering to find out the investment rating. The model uses forms of social life, other sectors that the markets are in, safety of the investments, ability of the market to self-regulate and the needs and the desires of the potential real estate buyers as factors. Using the results, 14 real estate markets are segmented and classified [12].

In a different research, the housing price inequality made by the real estate agents in New York, is investigated by using regression methods and qualitative interviews. It is found out that, in specific neighborhoods, the house prices differ and the agents can change prices according to the race between potential customers and their concentration. Also with the power of their knowledge, the agents can change the potential buyer's perceptions of the house so that they perceive the house to be much more expensive [13].

Crowston et al. compared agent owned and agent represented houses and it is found out that when the contracts that are done with the agents involves commission, it gives the agent moral hazard and it has effects on both the selling price and the liquidity [14].

There are many factors which can impact buyers. A research study is prepared to check whether real estate agents have an impact on potential buyers' internet listings. An

online tour is prepared to track which agent the possible buyers choose and why. It is understood that the buyers do not decide on the houses according to the appearance of the agents [15].

In another research, online retailers and classic real estate realtors are compared by the means of e-tailing and internet related cost savings. It is understood that the real estate agencies do not figure the saving amount with e-retailing because of, the age of the real estate sector [22].

3.2 Text Mining

Text mining is a data science that is applied in many different fields. In the research of Fankild et. al, text mining to the biomedical abstracts is applied to extract disease and gene associations. A scoring diagram is built by using the co-occurrences of the human genes and the diseases inside the documents. This is done by using dictionary based tagger system. As a result, the dictionary based tagger system recognized the diseases and genes in the abstracts and using this co-occurrence of these parameters are calculated [69].

Similarly, at another research, the similarity between the headings of the medical subjects are investigated using text mining. This is applied by, figuring out the occurrences of two medical subject heading at the same article and by using the semantic relatedness of the headings. Also, the occurrence of the words inside the articles by the same author is investigated and by doing this, mostly used terms at medical subjects are found. As a result, relationship between medical subjects are extracted using text mining [70].

In another research that is applied by Thomas et. al, systematic review of the published studies is done. This research is conducted by using systematic and exhaustive searching methods after data is extracted. As a result, progressive data is found that, the offered system reduces workload of screening for systematic reviews by using text mining [71].

Also, text mining can be applied at tourism sector. In a research, text mining principles are applied on the reviews of the three to five-star hotels in China. In the research, category extraction and sentiment analysis are done to analyze the scraped data. As a result, themes of the hotels, the rising trends and hot topics at hoteling industry are

found. Also using the outcomes, the hotel owners can see the "key business metrics of growth, earnings and performance" indicators of their hotel [72].

In one of the researches, text mining is applied to predict markets. Using social media, high rate of information about the financial state of the markets could be gathered and in this research different pre-processing techniques, evaluation mechanisms and machine learning algorithms are investigated to review the data. As an outcome, it is seen that, from machine learning algorithms, support vector machine and Naive Bayes techniques are highly used because of their comprehensibility. Also, it is concluded that evaluation methods are highly subjective and generally the outcomes are compared with the actualization probabilities [73].

3.3 Latent Dirichlet Allocation

LDA is a kind of topic modelling algorithm. In a research by Tian and others, adaptation of LDA is used to automatically categorize software systems written in different programming languages. The proposed technique is used to index and analyze the source code by firstly; each system is recorded as collection of words which produces a document after, GibbsLDA is used by Tian et. al., to index the documents then, retrieval of the categories is done by clustering topics whom cosine similarity is greater than 0.8 and lastly, the naming of the topics that are formed is done [16].

Heafield and others searched if LDA can be used to extract business domain topics from the source code. To apply LDA, the corpus is treated as software system which consists of source code files and a source code file can be named as a document and inside a document there are data structures, files and functions served as words. To find the optimum number of topics, the maximum likelihood method which is proposed by Griffiths and Steyvers is done [17].

Tirunillai and Tellis did strategic brand analysis using big data. Firstly, LDA is applied to extract the valences with dimensions of the user generated contents. After, labelling of the dimensions, the variety of the dimensions is assessed [18].

Miller used LDA technique at the auditing sector. Using LDA, companies' announcements are investigated and from the topics that came out, the trends are understood [19].

In a research, quantitative ratings of the user reviews on the tourism sites is done. By using LDA, key dimensions of customer service that a hotel should propose to the client is observed. Also using the nationalities of the users, demographic segmentation to the topics are done [20].

Moro and Cortes applied LDA to the finance sector. At the research, text mining techniques are used to analyze the 219 articles about business intelligence in banking sector and LDA modelling is used to group the articles based on their topics. After topic modelling, it is found that, credit is the most used field in the banking industry [21].

From the literature review that is done, it is found out that there is no research found that involves both LDA and real estate sector. So, with this study, we aim to fill this gap.

4. METHODOLOGY

4.1 Data Mining and Machine Learning

With the progress of the web, more information becomes available online and it becomes more difficult to be able to quickly access it. This can be also applied to the real estate sector. The web which is a multimodal space, includes all types of information and analyzing this information with algorithmic tools brings decision makers leverage of knowing the current and future status of the sector they are in.

To analyze the big data that is generated by the internet, data mining and machine learning techniques are used.

To define data mining, it can be said that, it is the process of finding out the unknown, usable knowledge for example, patterns, anomalies etc. by processing large datasets. Using the web and other electronic platforms, these large datasets are formed without extra effort instead of storing it. In this scope, data mining is one of the most used and popular computer sciences that is used in many fields such as market analysis, business management, social studies and decision support [22,23].

Data Mining Steps

In data mining, there are standardized steps that are applied by data scientists. Sample-explore-modify-model-asses (SEMMA) and Cross-industry standard process for data mining (CRISP-DM) processes can be good examples. In SEMMA, as its name implies, the process includes sampling, exploring, modifying, modelling and assessing stages. In Figure 4.1, the summary of the SEMMA process can be seen.

Sample: In this stage, a portion of a large dataset is taken instead of trying to manipulate huge dataset. Doing this, the cost of processing time decreases and the computational performance increases. In this stage, selecting the data amount is important. The data should not be very big which may slow down the manipulation and it should not be very small which may result in loss of information.

Sampling can preserve the benefits of actual data if it is done successfully because generally data has general patterns and a successful sampling captures the big picture

of the data. On the other hand, if the actual data is not bigger than the sample size, there is no extra sampling needed.

At sampling, training to fit the model, validating to asses and overcome overfitting and lastly testing to test the assessment processes are done. Testing is done to see if it generalizes well.

Explore: At this stage, "the anomalies and unanticipated trends are searched" to have a basic knowledge about the data. This is done by exploring the data both visually and numerically and by looking for trends and clusters. When there is no outcome after the visual review, statistical analysis such as factor analysis, clustering and similarity analysis can be done to have a better knowledge of the data. Clustering before modifying can have a positive impact on the next stage because after clustering, the clustered data will be modified individually and the obtained result will be clearer and much more accurate [26].

Modify: This is the stage where the deep analyses of the variables are done. At this stage, the data is transformed before creating models. The transformation occurs with the light of the outcomes of the exploring stage. Also, narrowing the variables which are not useful can be done and the most important outcomes are selected. As it is known, data mining is a continuous process in which that the data is continuously generated and flows to the database. So, the new data that are formed should also be modified accordingly [24].

Model: At this stage, the combination of the variables and their dependencies are searched. This is done by using decision trees, artificial neural networks, support vector machines, logistic models, times series analysis another statistical model. Each of these models generate different kind of data and show different perspectives if the right variables form the data [24,25].

Assess: At the last stage, the models that are formed at the previous stage from the database, are evaluated by the terms of usage and reliability. Also, the models should include a different point of view that cannot be perceived directly before the analysis. Also at this stage, trial of the model is done with the dataset which is excluded from the sample dataset. By doing this, the usefulness of the model to the entire data set is tested. In addition, if there is a source of example from the real word, the model can also be tested if it is applicable to the real world. Lastly at this stage, the models are compared and decision making is done.

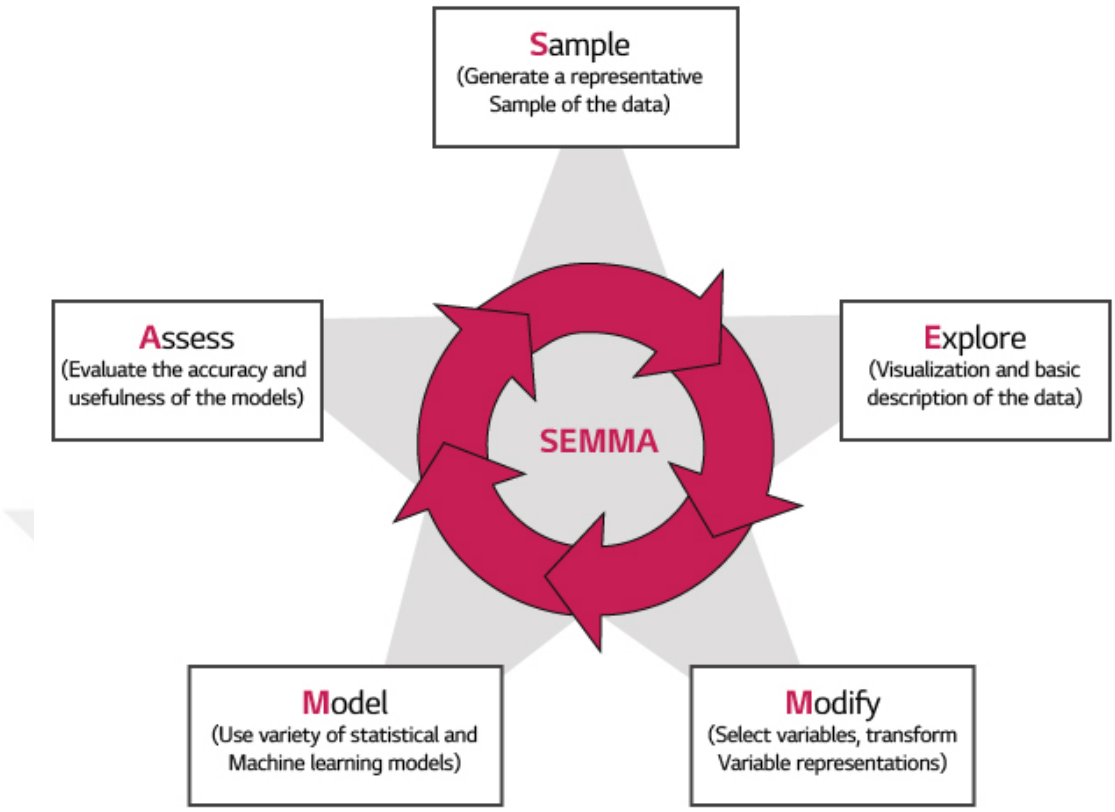


Figure 4.1 : Visualization of SEMMA Process [27].

Different from the SEMMA, at CRISP-DM, the phases consist of; business understanding where the objective of the mining that will be done is determined with a business perspective, data understanding where the initial data extraction and examining the data at a glance like the explore stage of SEMMA, data preparation where the data is manipulated before modelling like the modify stage at the other data mining process, modelling stage where the intensified modelling practices are applied just like the other modelling stage, evaluation stage where the models that are formed are evaluated and lastly deployment stage where the knowledge that is gained is used in business decision making process [25].

Machine Learning

According to Alpaydin, to develop a performance measure, machine learning (ML) uses experience or the sample data to program the computers [28]. Basically, in machine learning, firstly the task that will be done is formulated, then the examples

are obtained and split into two as training and test group. Lastly, computer tests the knowledge it gained from the training group and makes evaluation.

ML is done using different learning techniques which are supervised and unsupervised learning. At supervised learning, examples are given to the algorithm to learn from it. At this process, two sets are given to the algorithm which one of it is training data and the other is the testing data. Giving the training set, increases the accuracy of the outcome and the system uses training set to classify the data by analyzing it [29].

At unsupervised learning, the statistical composition of the input patterns is reflected with the learning of the systems. To compare it with supervised learning, there are no environmental evaluations and definite targets that correlates with every input at unsupervised learning [30]. Unlike the supervised learning, at this learning model, the correct results are not provided during training and it uses untaught primitives like neural competition and cooperation [31]. Generally clustering and labelling are done with this model.

4.2 Topic Modelling

Topic modelling is a concept that is highly used in computer sciences like information retrieval, natural language processing and text mining. As a brief history of topic modelling, firstly the text corpora is proposed [32], which gives every word a number that you can find out the overall word count of the document and by using tf-idf matrix the number of occurrences of each word in a document is counted and weighted, after to raise the level of statistical analysis of the documents latent semantic indexing is proposed [33] which makes linear combinations of the tf-idf matrix and " capture some aspects of basic linguistic notions such as synonymy and polysemy" [34]. Then the probabilistic latent semantic indexing (pLSI) is introduced [35] which is based on mixture decomposition derived from latent class model and analyses two-mode and co-occurrence data and it shows the documents as the list of mixing proportions and find a fixed set of topics by probability distribution. On the other hand, as Blei suggests, this does not provide probabilistic model at the level of documents, it just shows them as a list of numbers and no generative probabilistic model for these number is created. This is a problem because, showing documents as numbers will lead to a linear increase in parameters in the model and results in overfitting issues. In addition, the assignment of a probability to a document outside of the training set is

not vivid. To solve these issues LDA is proposed and with this model exchangeability of the words and documents can be captured [34].

4.3 Computer Sciences

4.3.1 Information Retrieval

To get information from more than thousand texts, there should be a method to find out the desired data using a computer interface. For example, the researcher should be able to search a keyword and get the related document as an outcome. Also, when new documents, related to the keyword that the user searches come out, a notification can be sent. To apply these processes there must be a solution to the problems like;

- matching the meaning of the searched word with the documents,
- finding similar documents,
- processing huge data when there is little time to give result.

In information retrieval, this type of problems is searched to be solved [36].

There are some information retrieval models to figure the goals of information retrieval. Firstly, when the IR models are classified at mathematical basis; set-theoretic models store and shows the documents as set of words, algebraic models show documents as vectors and matrices and lastly probabilistic models calculates the probabilities of documents with respect to the searched item, when the similarity increase the probability also increase [37].

Also, IR models can be classified by modelling term interdependencies. At this classification there are models like, models without term-interdependencies, models with immanent term independencies and models with transcendent term independencies. At models without term-interdependencies, the different terms and word are treated independently. At these models, generally vector space models are used. At the models with immanent term independencies, the interdependencies between the words are represented by showing degree. This degree is produced from the cooccurrences of the data. Lastly at models with transcendent term independencies there is an interdependence between the terms but it is not shown how the interdependencies are formed [36,38].

4.3.2 Text Mining

As a substitute of data mining, in text mining, getting unknown information about the documents or texts is the main objective. To reach this objective, some statistical analyses including topic modelling are done. In addition to topic modelling, some of the tasks below can also be applied;

- **Keyword in Context:** In this task, automatic labeling of the keyword is done to the documents using the keyword based knowledge discovery from text files [38].
- **Correlation Analysis:** At this process, correlations between all of words at the documents are found using the term-document matrix and highly correlated words are listed. The correlation is found by finding out the similarity measures and interpreting similar objects [38].
- **Information Visualization**: As its also named as visual text mining, this process uses the large text sources to produce a visual hierarchy and this has advantage of visual browsing capability to help the user to understand the data visually. At this application, zooming, scaling and creating sub-map actions can be done on the formed visual image. This application has an importance of narrowing down high number of documents with different perspectives and showing in a singular platform. This helps to explore the topics that forms the documents. To give an example of the usage of information visualization, a map can be created by searching for the events that may have relationship which finds out the connection between the crimes [39]. As it is stated at Figure 4.2, the relations of the words are showed as a bibliometric map. Correlation map can be good example of information visualization.

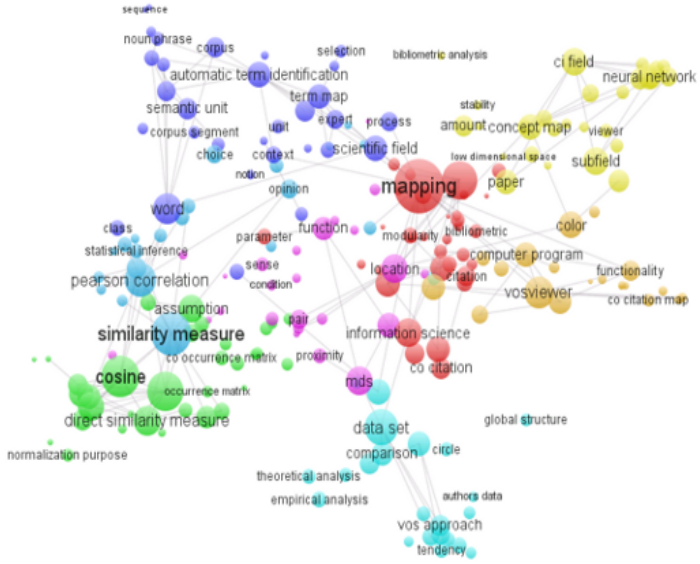


Figure 4.2 : Information Visualization Example [40].

- **Summarization:** This process is very useful for end users to find out the usability and the worthiness of a very long document without reading all of it. This task summarizes the document and the user will firstly read the summary before further reading. To make this process successful, maintaining the core points and meaning of the document while reducing the length of document should be done. This process is difficult because teaching a software to analyze and figure out the semantics and put the meaning of the document into words is hard [38].
- **Patent Analysis:** At this task, supervised and unsupervised techniques are used to analyze the patent documents. There are too many patent requests, which are longer than other text documents, to be supervised by patent approvers. Additionally, it is a sensitive area which, if something is missed while investigating, a problem may appear for the patent holder or requester. At this algorithm, clustering and an automatic procedure to create generic cluster titles are done in collaboration with co-word analysis and keyword extraction [41].
- **Trend detection/prediction:** At trend detection, assessment of the trends and emerging fields are investigated by looking at the changeover of the time using the structure and characterization of clusters. Also, the regression analysis as a statistical method can be applied to detect trends. In Figure 4.3, the words from different genres

of books are extracted and clustered and the trend of words throughout the history can be seen from the graph [36].

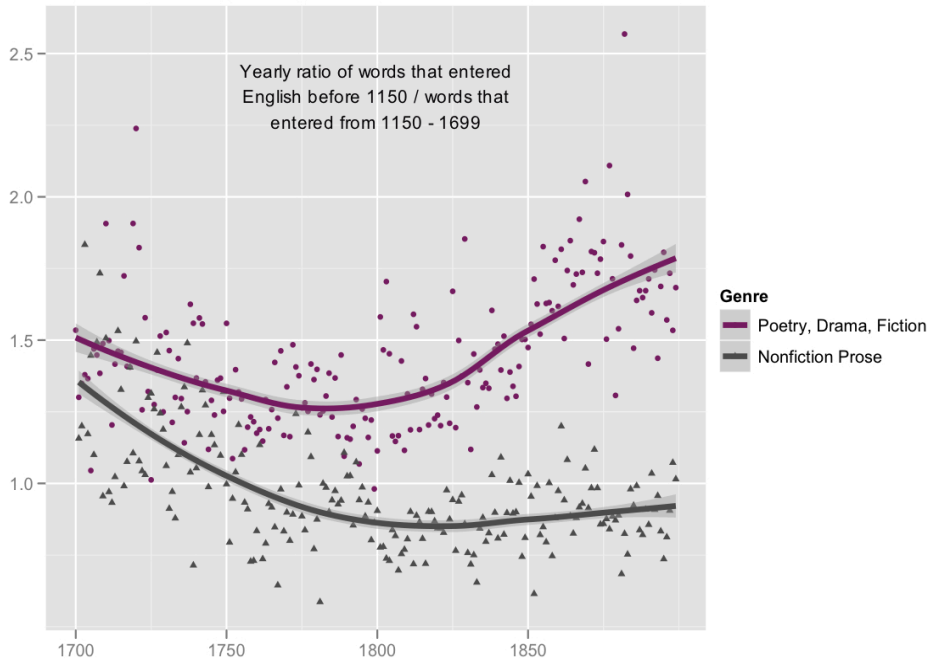


Figure 4.3 : Trend Detection Example [42].

Text Mining Process

Data can be found in different forms, some can be easily analyzed and some must be transformed before analyzing. Generally, data mining is done when the data is in a record or variable form, but when the data is not in these forms, text mining techniques are used to analyze the text data. Texts are not formed by the standardized rules. Each text differentiates in terms of its language and its meaning. Thus, the computer cannot understand the data scraped form the texts. In text mining, information extraction from natural language texts is done by firstly obtaining key contents of the text and then categorizing these key contents [43].

At the beginning of the text mining process, documents are transformed to a machine-readable format before they are analyzed. If this step is passed, the outcome of the analysis can be wrong and misleading. The processes as step by step can be seen below:

- **Noise removal:** Most declarations contain general data on the discharging organization, for example, the address of central command or contact data. This kind of words does not give any significant data regarding the main objective, so these

parts are deleted and it is concentrated to the fundamental body of the document. Also, some basic words and expressions, presented by the distributing organization, for example, brand names of the organizations are generally expelled from the corpus.

- **Stop word removal:** Purported stop words, for example, the, is, from. These words don't contribute to the analysis that will be done and that is why, these words are deleted from the corpus.
- **Stemming:** In this process, the arched words are reduced to their stems. This process is generally done using the Porter stemming algorithm.
- **Document term matrix:** In the last advance of the pre-processing stage, Document term matrix is formed to numerically speak to which words or expressions are available in an article. After, tf-idf weighting (term frequency-inverse document frequency) is applied to distinguish the most used words in the records [44].

4.3.3 Natural Language Processing

Natural Language Processing (NLP) is one of the principles that is part of topic modelling and can be processed with machine learning. NLP deals with computer and human interaction in both written and spoken natural language and the data of this context are texts, speeches and linguistic information. On the other hand, machine learning deals with teaching computers to learn from data presented as examples. So, for computer to understand the spoken natural language the concept of machine learning is used to make the computers learn from the examples that are given in language processing [45]. To apply machine learning or other programming languages on natural language processing problems, the text that will be analyzed should be transformed to a structured form and it is done at text mining. Also, producing algorithms involving NLP is hard because the computer algorithms are formed generally by programming languages and programming languages are precise, explicit and systematic. On the other hand, our daily speech involves implicit parameters and has complex variables such as slang, regional dialect, idioms and proverbs [46]. Generally, Markov Model is used in NLP tasks.

Some of the tasks that involves both text mining and NLP are;

- **Sentiment analysis:** This analysis is used to figure the sentimental idea behind the text. There are some studies that trains computer using human interaction and they try

to increase the effectiveness of the sentimental detection [38]. Momentarily, Naïve Bayes approach is the most successful classifier that is used to analyze sentiment. At Naïve Bayes classifier, prior probabilities of the elements are measured which are the probabilities the system experienced previously. The algorithm elements, are regarded as classes. In addition, when new data enters to the algorithm, local area of the new data is clustered and the probabilities of the elements in this cluster are measured which is called likelihood. Using the prior probabilities and the likelihood, new data is classified. [49] In Figure 4.4, sentiment analysis of the Twitter posts about the 2016 USA Election candidates are done and graphed using the posting date.

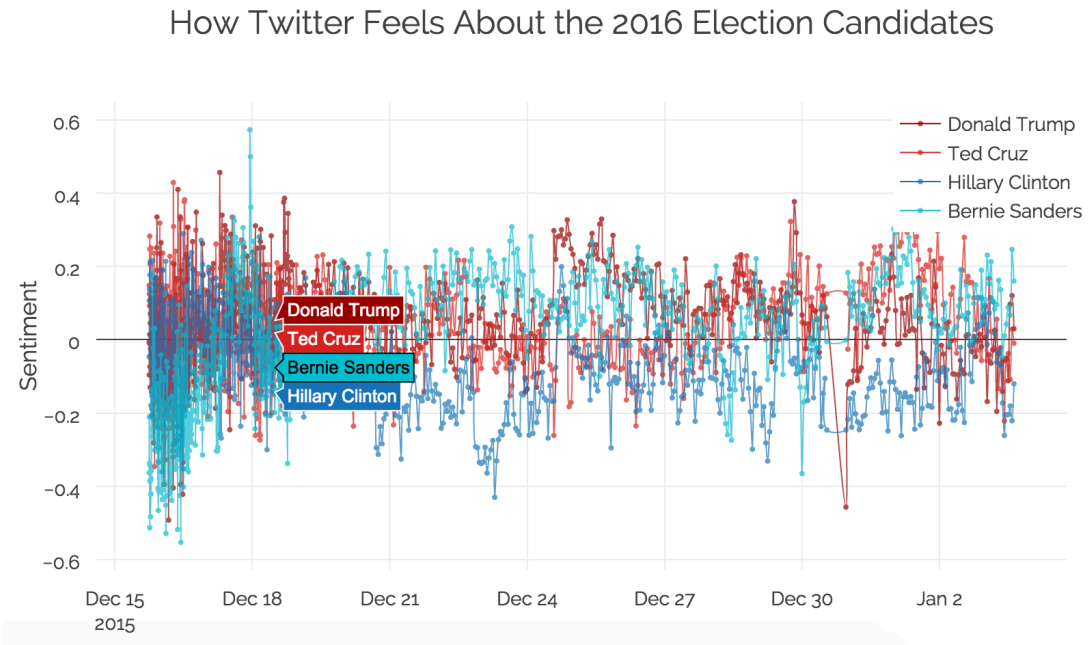


Figure 4.4 : Sentiment Analysis Example [50].

- **Question Answering:** It is one of the examples of NLP and it focuses on, giving the best solution to the questions that are asked. This application is highly used at online websites to help the consumers. It uses many text mining principles. For example, it uses information retrieval principles to get the data like people, place, time, event or categorize the questions like where, when, who how etc. to get the answers. Also, aside from the websites, this model is also used at the O&A spaces of the public places like hospital, school and mall. Basically, it produces a supply where people ask questions to a computer and gets the answer [52].

- **Paraphrase Detection:** At plagiarism detection, it is hard to locate the paraphrased sentences. An algorithm is produced to detect the paraphrased paragraphs by using the common methods that are used to paraphrase such as; ordering word/phrase

differently, inserting or deleting the original source, placing alternative words. These methods are combined to form a paraphrase detection model. Word embedding models are used for placing alternative words, vectoring the words in a sentence is done to figure the re-order of the words and lastly word level edit distance is calculated to figure the insertion and deletions from the phrased sentence [51].

Markov Model

Markov Model is used on randomly changing systems to model them. Modeling is done by predicting that the future state of an action comes from its present state. The model only uses the current state, it does not use the previous states. At the algorithm of the Markov Model, Markov Chains are used. Markov chain is a set of states where there is a subsequent action between the states. At Figure 4.5, an example of a Markov chain can be seen [47,48].

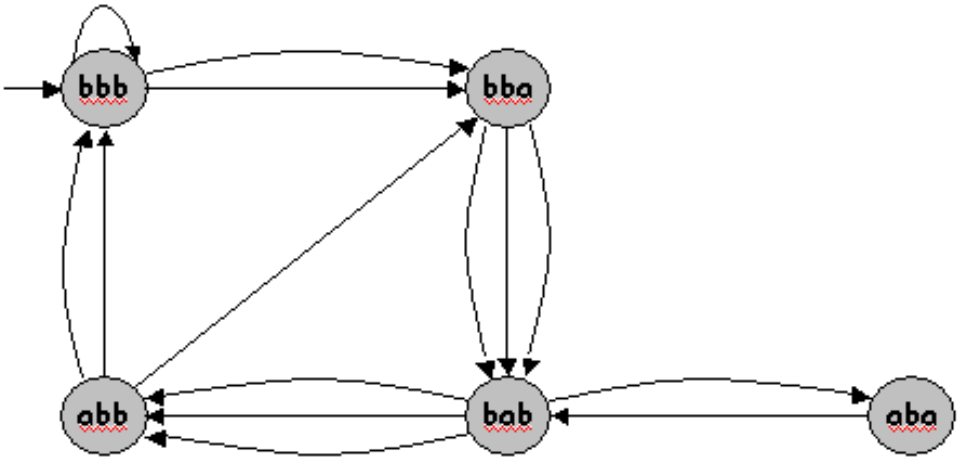


Figure 4.5 : Markov Chain Example [47].

As it can be seen from the figure, action starts with the bbb state and after it has 3 possibilities to move on. 1/3 is to remain at that state and 2/3 is to go to the next state which is bba. Using this chain, the possibilities of moving through states are calculated [47].

Hidden version of Markov Model is generally used at NLP tasks. At the Hidden Markov Model (HMM) the states are not visible, but the output comes from the states are visible. So, the posterior process of Markov Model is done on HMM. Knowing the probabilities, the sequence of the states is determined [48].

4.4 Latent Dirichlet Allocation

LDA is one of the unsupervised learning tools that is used to find out the unknown topics of the texts. The texts are generally form a large corpus that it is impossible/takes very long time to classify. Using LDA, the keywords used to form documents generates the latent which is unobserved topics. At the algorithm, the documents are seen as mixture of topics that the words are split with certain probabilities and a topic is a probability distribution over set of words *p(t/d)*. At the same time, a document is seen as a probability distribution over topics *p(w/t)*. To make sense, LDA makes the topic generation by doing collective probability distribution using both observed and hidden data [53,54].

To give more detail about the process of LDA;

At the algorithm, firstly a random distribution over topics is determined for the first document. Later, for each word at the first document, a topic is chosen from the distribution over topics randomly and a word is chosen randomly from the previously assigned topic. This process is not finished even if all the documents are distributed over topics. LDA uses distribution over distribution so; this process repeats over and over. Also, the position of words in the documents are not important. The system uses bag of words model while selecting the words [54]. At Figure 4.6, the visualization of the LDA can be seen.

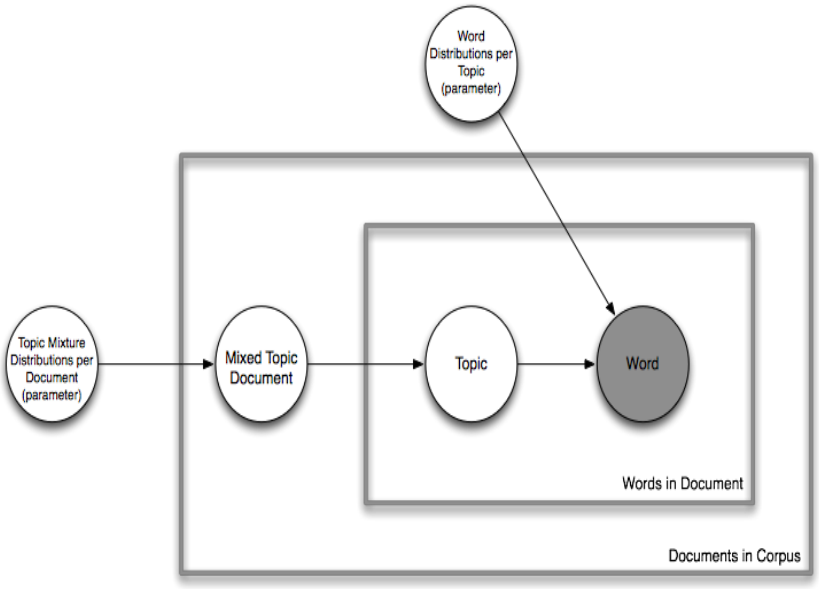


Figure 4.6 : LDA Model [56].

In addition to the PLSI, which can assign multiple topics to individual documents by sampling a distribution of topics each time a word is drawn, instead of each time a document is created, LDA introduces latent variables to the scheme and now there is two corpus-level parameters. That is why, Bayesian sampling method is used to do the random sampling. Using the Dirichlet distribution, randomly assigning of the words to the topics are done.

Dirichlet Distribution

Dirichlet is a type of exponential distribution that is the conjugate prior of the likelihood of multinomial distribution. In Dirichlet, prior probability distribution is used which is doing probability distribution before the fact is unknown, deciding using the unknown parameter. At Dirichlet, firstly an item is assigned to a probability and after, the second items position is determined whether it will be at the same position with the first item or a new category. For a new category, simply the formula below is used;

$\alpha/\alpha+n-1$

For the existing category;

$nx/\alpha+n-1$

formula is used [57].

Simple Gibbs Sampling

As a training method, generally Gibbs sampling is used. Gibbs sampling is a strong method that uses Monte Carlo Markov-chain algorithm. It produces "a sample from a joint distribution when only conditional distributions of each variable can be efficiently computed [58]."

4.4.1 LDA model

LDA, is a generative probabilistic algorithm used to model a corpus. LDA accepts the fact that, each document can be seen as probabilistic distribution over unknown topics, and the topic distributions in all documents have the common Dirichlet prior. Also, each unknown topic in the model can be seen as a probabilistic distribution over words

and lastly one of the other things that shares common Dirichlet prior is word distributions over topics.

In a corpus named *D* which consist of *D* documents and the document *d* which has N number of d words ($d \in \{1,..., M\}$), modelling of LDA is done on D, regarding the steps below;

1. Multinomial distribution β_k is chosen for topic k ($t \in \{1,..., K\}$) from a Dirichlet distribution with parameter η.
2. A multinomial distribution θ_d is selected for document d ($d \in \{1..., D\}$) from a Dirichlet distribution with parameter α.
3. For a word w_n ($n \in \{1..., N_{d\}}$) in document *d*,
    - From $\theta_{d,}$ z_n topic is selected which is a multinomial probability.
    - From β_{zn} , w_n word is selected, which is also a multinomial probability [54].

When LDA process occurs, words in documents are the only observed variables while others are the latent variables which are β and θ; and hyper parameters which are α and η. To draw a conclusion to the latent variables and hyper parameters, the probability of observed data *D* is computed and maximized using the formula below:

$$p(\beta, \vartheta, z, w) = [\prod_{i=1}^{K} p(\beta_i|\eta)] [\prod_{d=1}^{D} p(\vartheta_d|a) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n})]$$

If the formula is divided into parts,

$\prod_{i=1}^{K} p(\beta_i|\eta)$

comes from

$(\beta_d|\eta) \sim Dir(\beta)$

And like β,

$\prod_{d=1}^{D} p(\vartheta_d|a)$

comes from

$(\theta_d | a) \sim Dir(a)$ and the Dirichlet distribution can be formalized as [60];

$$p(\vartheta | a) = \frac{\Gamma(\sum_{i=1}^{k} a_i)}{\prod_{i=1}^{k} \Gamma(ai)} \vartheta_1^{a_1 - 1} .. \vartheta_k^{a_k - 1}$$

Also,

$$[\prod_{d=1}^{D} p(\vartheta_d | a) \prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n})] = p(\theta, z, w | a, \beta)$$

and $p(\theta, z, w | a, \beta)$ means "given parameters a and β, joint distribution of a topic mixture θ a set of N topics z, and a set of N words w"

The graph version of the formula is stated in Figure 4.7.

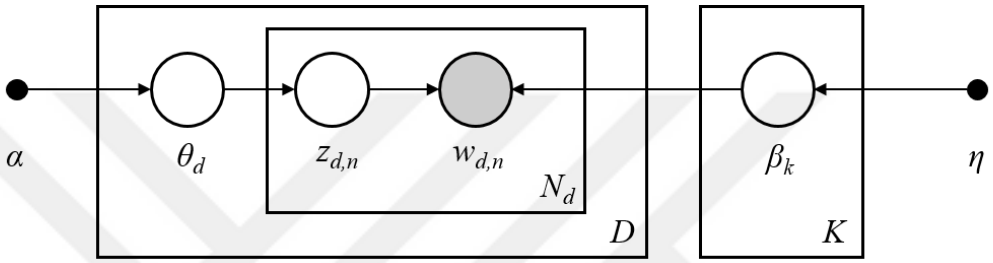


Figure 4.7 : Detailed Version of LDA Model [34].

Where;

α = Dirichlet topic proportions hyperparameter, controls division of documents into topics

η = Dirichlet topic hyperparameter, controls division of topics into words

θ_d = Topic proportions for each document

z_{dn} = Topic distribution to each word

w_{dn} = Observed words

β_k = Topics

N_d = number of words in a document

K= number of topics

D= number of documents

As it can be seen from the graph that the only observed variable is w_{dn} and the rest of them are latent.

The uniqueness of LDA from other clustering applications is that "the Dirichlet on the topic proportions can encourage sparsity" which helps latent variables to be zero. To explain, the model of LDA forms double sparsity where there is a compromise between document topic distribution and topic word distribution. So, when the system allows documents to have fewer topics, there will be higher words per topic. This balance varies from corpus to corpus and LDA adapts to all by this process. To summarize, LDA does not look for only co-occurrence, it also shows flexibility by using sparsity [61].

In addition, LDA uses it's algorithm to learn the distributions that will describe the probabilities that forms latent topics.

LDA implementations

There are many LDA implementation options in the internet from different programming languages. Table 4.1 shows some of the alternatives and their descriptions.

Table 4. 1: LDA Implementations [62].

Name	Author	Language	Description
lda-c	David M. Blei	C	Variational inference for LDA
lda	J. Chang	R	Uses Gibbs sampling and it is fast
Stanford Topic Modeling Toolbox	D.Ramage and E.Rosen	java	Trains topic models using LDA, Labeled LDA, or PLDA new
MALLET	UMASS	java, R	Uses Gibbs sampling, fast, scalable
Mr. LDA	U. Maryland	java	Uses variational Bayes
Gensim	R. Rehurek	python	Topic modelling alternatives

5. APPLICATION: REAL ESTATE SECTOR IN ISTANBUL

5.1 Real Estate in Turkey

In Turkey, construction is one of the most important sectors that investors and government invests and being its complement, real estate sector also have high market value. The key players of real estate production in Turkey are, government projects 10%, branded real estate producers 3% and other small to middle construction firms [68].

The factors that affect real estate prices in Turkey are, the place of the house (city, the district of the city), the neighborhood (hospital, school, park, the view, ease of public transportation), age of the building, properties of the building etc. For example, when a subway infrastructure is planned in an area, automatically sale and rental prices go up and this leads others to think of a real estate bubble. On the other hand, at a research which is done at 2015, it is evaluated that there is no real estate bubble, but there can be some price differences between some fields according to the new projects [67]. The most important projects that are done and ongoing are; Yavuz Sultan Selim Bridge, Osmangazi Bridge, North Marmara Railway, Third Airport in Arnavutköy/Istanbul, Avrasya Tunnel, Kanal Istanbul, Istanbul Finance Center in Ataşehir/Istanbul, Bakü-Tiflis-Kars Railway, 1915 Çanakkale Bridge, 3 Floored Sub-Sea Tunnel, City Hospitals and Zigana Gateway.

5.1.1 Real estate in Istanbul

Istanbul is the most important market in the Turkish real estate sector. It is one of the Word cities that gets attraction by both local and international investors. To understand the importance of Istanbul better the facts below can explain the current situation:

- When Istanbul's house sales rate with respect to Turkey is investigated, it is seen that the ratio varies between 19.4% to 17.4%.
- Istanbul is a city which continues to get migration even though it is one of the most crowded cities in the World.

- According to the TUIK data, even though the rate of real estate purchase decreased in 2017 and 2018, the house sales goes on actively and the foreign investor rate increases over time [63].

In many districts of Istanbul, branded real estate projects are ongoing and as a result, those districts develop and their real estate prices increase. For example, as a district of Istanbul, Şile is a touristic place which consists of forest and seaside and with the increase of real estate projects, the district is developing. The zone of the real estate projects is generally determined by the transportation and other projects of government. At Table 5.1, ongoing projects in Istanbul can be seen.

Table 5. 2: Government Projects in Istanbul[63].

Project Name	Budget	Current Situation
3rd Bridge	4.5 Billion TL	Open for Business
3rd Airport	10.3 Billion €	Under Construction
Istanbul Finance Center-Ataşehir	4.5 Billion TL	Under Construction
Avrasya Tunnel	1.3 Billion $	Open for Business
Kanal Istanbul	5.5 Billion $	Studies about the root is carried out
Big Istanbul Tunnel	3.5 Billion $	Proposal Evaluation

As it can be seen at Table 5.1, in Arnavutköy, 3rd airport will be built so the worth of the district increased perceptibly. Also in Tuzla and Pendik, metro and Havaray projects are ongoing and these districts can be good examples of price increase. In addition, with the connection roads and metro stations, Beykoz and Üsküdar gained investors interest. Lastly at Kartal, urban transformation is in progress and with the new real estate projects, Kartal is one of the developing districts [64].

5.2 Web Scraping Using API

With the permission of the authorized company to investigate Istanbul real estate posting data, a real estate company's website is scraped. The scraping process is done with the help of "ParseHub" API. The scraping process begins at the in dept searching page of the website. 24 postings are listed in the first page and to see all of the postings in a page, scrolling is needed. To view the other entries, the user should pass to the

next page. Using the API, each posting page is investigated using set of instructions. The instructions that are taught to the API can be seen at Figure 5.1 and 5.2. Firstly, scrolling of the page is done to see all the data in a page. After, the appointment of the parts, where the clicking action will be done to get to the actual real estate posting page, are marked. Lastly using the arrow key, the API passes to the next page and continue to apply the same process.

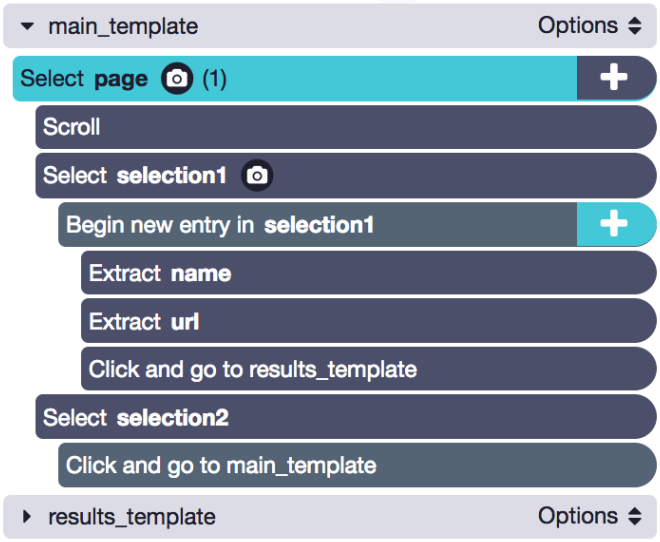


Figure 5.1: API Instructions Part One.

The second part of the instructions occur at the real estate posting pages. From this page, the parameters specified are extracted and at each posting extraction, new row is created at an excel file. At Figure 5.2, the requests of data extraction can be seen.

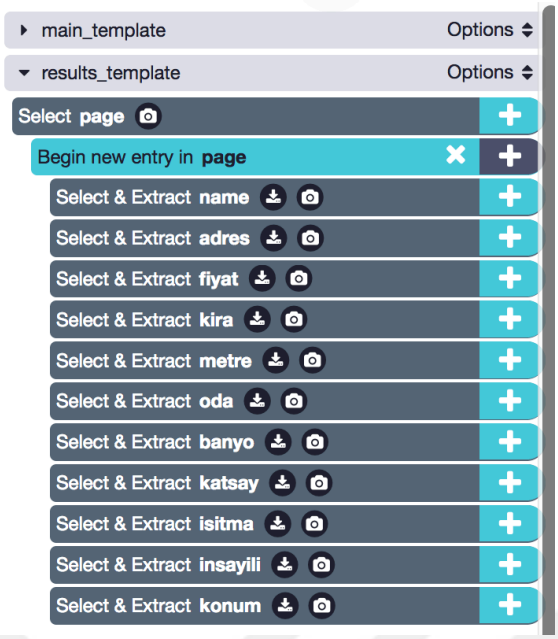


Figure 5.2: API Instructions Part Two.

The outcome of web scraping is an csv file that consists of; topic, URL,detail of the post, location, price, rental type, meter-square, construction date and the coordinate parameters of the house that is posted, as it can be seen from Figure 5.3 Some of the entry details could not be extracted automatically, so the missing are taken manually.

selection1_name	selection1_url	detail	selection1_page_adres	selection1_	selection1_page_kira	sele	sele	selection1_page_konum_url
ŞİŞLİ MERKEZDE MET	http://www.turyap.com.tr/ilan-detayi/sa	' ŞİŞLİ MERKEZDE,	İstanbul / Şişli	650,000 TL	Satılık Apartman Dairesi Kc	110 m²	2018	https://www.google.com/maps/@41.06303377693878,28.98895857
Sarıyer Acarlar Uyum	http://www.turyap.com.tr/ilan-detayi/sa	Turyap Sarıyer Maden	İstanbul / Sarıyer / Köyler / Zekeriy	590,000 TL	Satılık Apartman Dairesi Kc	100 m²	1993	https://www.google.com/maps/@41.18864705787779,29.04789522
Sarıyer Turyap tan Ter	http://www.turyap.com.tr/ilan-detayi/sa	3.Köprü ve 3.Köprü	İstanbul / Sarıyer / Bahçeköy / Mad	1,060,000 TL	Satılık Apartman Dairesi Kc	147 m²	2012	https://www.google.com/maps/@41.19208723617951,29.03674442
ESENTEPE EMEKLİ Sİ	http://www.turyap.com.tr/ilan-detayi/sa	ESENTEPE'DE KENTSEL	İstanbul / Şişli / Esentepe / Esentep	1,200,000 TL	Satılık Apartman Dairesi Kc	100 m²	1978	https://www.google.com/maps/@41.06426219999999,29.01091180
MERTER TURYAPTAN	http://www.turyap.com.tr/ilan-detayi/sa	TOPKAPI E-5	İstanbul / Zeytinburnu / Maltepe / M	1,200,000 TL	Satılık Residans Konut	136 m²	2016	https://www.google.com/maps/@41.0236571469164 1,28.91698532
Şişli - Bomontide Fırsa	http://www.turyap.com.tr/ilan-detayi/sa	Şişli - Bomonti'de Fırsat	İstanbul / Şişli / Ergenekon Mah.	690,000 TL	Satılık Apartman Dairesi Kc	145 m²	2018	https://www.google.com/maps/@41.060040882992 29,28.98820302
Şişli - Paşa Mahallesin	http://www.turyap.com.tr/ilan-detayi/sa	Şişli - Paşa Mahallesinde	İstanbul / Şişli / Duatepe / Paşa Ma	270,000 TL	Satılık Apartman Dairesi Kc	60 m²	2018	https://www.google.com/maps/@41.05305900945117,28.97135556
Şişli - Paşa Mahallesin	http://www.turyap.com.tr/ilan-detayi/sa	Şişli - Paşa Mahallesinde	İstanbul / Şişli / Duatepe / Paşa Ma	280,000 TL	Satılık Apartman Dairesi Kc	65 m²	2018	https://www.google.com/maps/@41.05305900945117,28.97135556
Şişli - Paşa Mahallesin	http://www.turyap.com.tr/ilan-detayi/sa	Şişli - Paşa Mahallesinde	İstanbul / Şişli / Duatepe / Paşa Ma	370,000 TL	Satılık Apartman Dairesi Kc	65 m²	2018	https://www.google.com/maps/@41.05305900945117,28.97135556
Zekeriyaköy de güvenl	http://www.turyap.com.tr/ilan-detayi/sa	İnanlar inşaat ın ödüllü	İstanbul / Sarıyer / Köyler / Uskumr	1,999,000 TL	Satılık Villa Konut	214 m²	2015	https://www.google.com/maps/@41.21309562488036,29.03330297
Zekeriyaköy Ormanad	http://www.turyap.com.tr/ilan-detayi/sa	Zekeriyaköy	İstanbul / Sarıyer / Köyler / Zekeriy	5,500,000 TL	Satılık Villa Konut	410 m²	2014	https://www.google.com/maps/@41.19940619234525,29.02857775
ŞİŞLİ MERKEZDE PRO	http://www.turyap.com.tr/ilan-detayi/sa	MERKEZİN TA KENDİSİ	İstanbul / Şişli	455,000 TL	Satılık Apartman Dairesi Kc	88 m²	2017	https://www.google.com/maps/@41.06318951311723,28.98635041
3+2 - 305m2 BELGRAD	http://www.turyap.com.tr/ilan-detayi/sa	3 KATLI (TRİPLEX) 6 +1	İstanbul / Sarıyer / Bahçeköy / Bah	1,200,000 $	Satılık Apartman Dairesi Kc	230 m²	2014	https://www.google.com/maps/@41.17768725095163,28.99136805
FIRSAT! MERTER TUR	http://www.turyap.com.tr/ilan-detayi/sa	NEF 13 TE İÇERİSİNDE	İstanbul / Zeytinburnu / Maltepe / M	345,000 TL	Satılık Apartman Dairesi Kc	49 m²	2016	https://www.google.com/maps/@41.0167684,28.907491499999992

Figure 5.3: Scraped Data Format.

The crawling action occurred from 08.04.2018 to 12.04.2018. The postings that are published between these dates are scraped by looking at the district of the posting. No more than 300 posting from the same district is taken. After the web scraping process, dataset that consists of 5000 real estate postings, which includes all districts in Istanbul except the Adalar district, is created.

5.3 R/R Studio

R programming language is used to do data visualization, association and sentiment analysis at the progress of text mining. The packages below are included to R:

SnowballC: It is used for stemming Turkish words**.**

Ggmap: It is used for generating maps.

Ggplot2: It is used to plot graphs and visuals on the map.

RSentiment: NLP application to do sentiment analysis. This package can also use Turkish words as an input.

Wordcloud: It is used to generate wordcloud.

TM- The text mining package can be used in multiple applications such as**;** top words, frequency analysis, correlation plots, association analysis, data processing, LDA application

5.4 Data Processing

Both using Microsoft Excel and R programming language, the actions below are done:

- The corpus is prepared before it is imported to the language programs. For example MALLET uses txt format, so the csv data is transformed to the txt.
- To avoid differentiation between the same words, all letters are converted to lowercase.
- Turkish letters such as "ü,ı,ş,ç,ö,ğ" are transformed to "u,i,s,c,o,g" for avoiding unprocessed letters by the programming languages like R and MALLET.
- Numbers are eliminated from the dataset.
- Using Excel spelling tool, wrong words are fixed.
- Since there is no automatic stop word removal for Turkish, selected stop words are deleted manually.
- Elimination of the extra spaces is done.
- Using the SnowballC package which adapt with Turkish, stemming of the words are done.
- After all data cleansing is done, the words are transformed to vectors where a number is set to a word and the frequency is listed.
- From the location link that is scraped using the map in the entry, latitude and longitude is extracted for further use.

- The prices of the houses are transformed to the same unit using the exchange rate at 29.04.2018.

5.5 Data Visualization

The dataset that is scraped consists of 38 districts. At Figure 5.4, the real estate posting rates from each district can be seen.

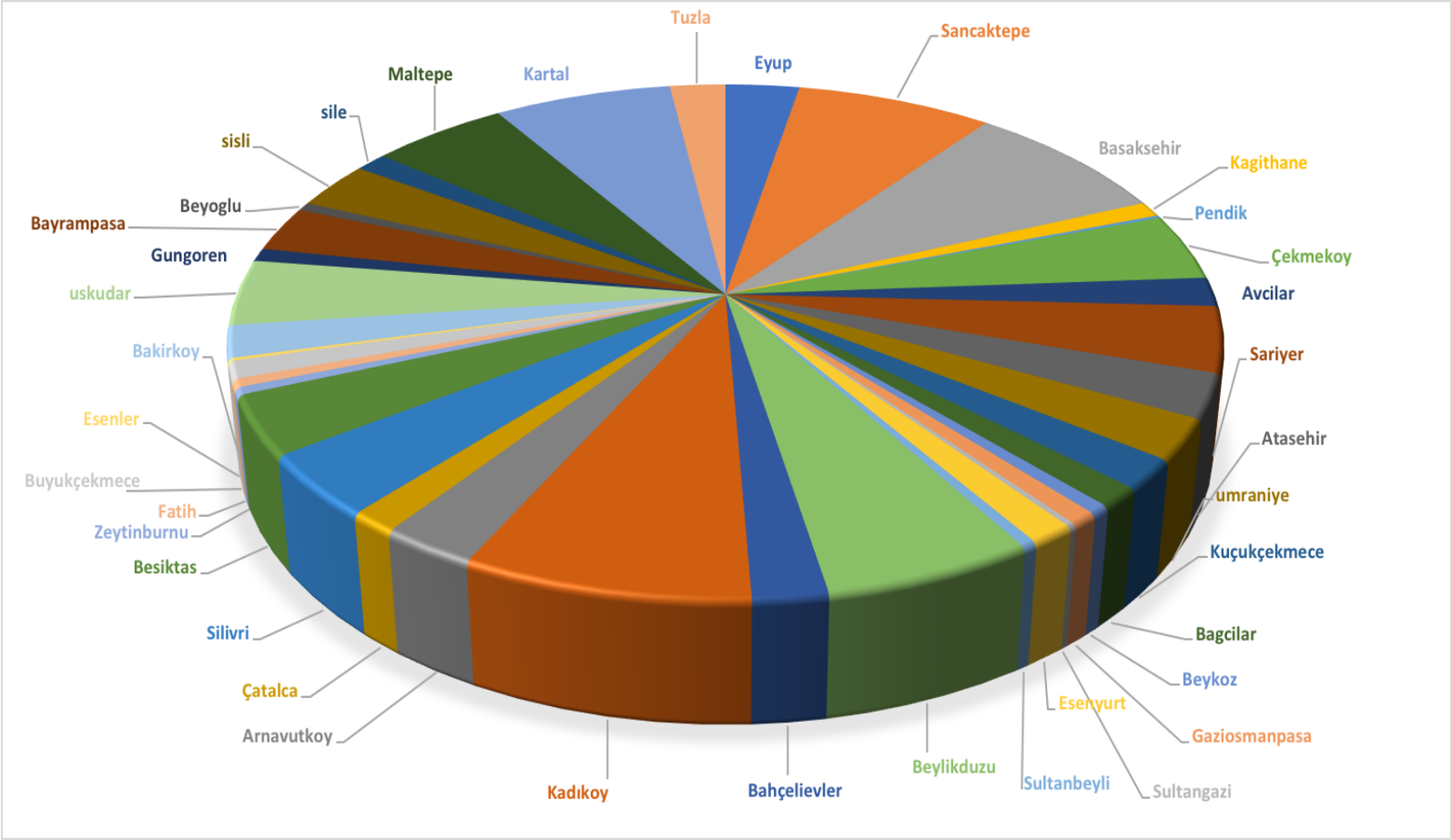


Figure 5.4: Real Estate Posting Rates from Each District.

5.5.1 Visualization using Maps

Using the R package (ggmap) and Google API, maps from Google Maps can be scraped using the parameters of latitude and longitude. After the map and the data is prepared, different kinds of maps can be generated. Using the location data of the real estate postings dataset, different kinds of maps can be created. For example, at Figure 5.5, heatmap of the real estates for sale can be seen.

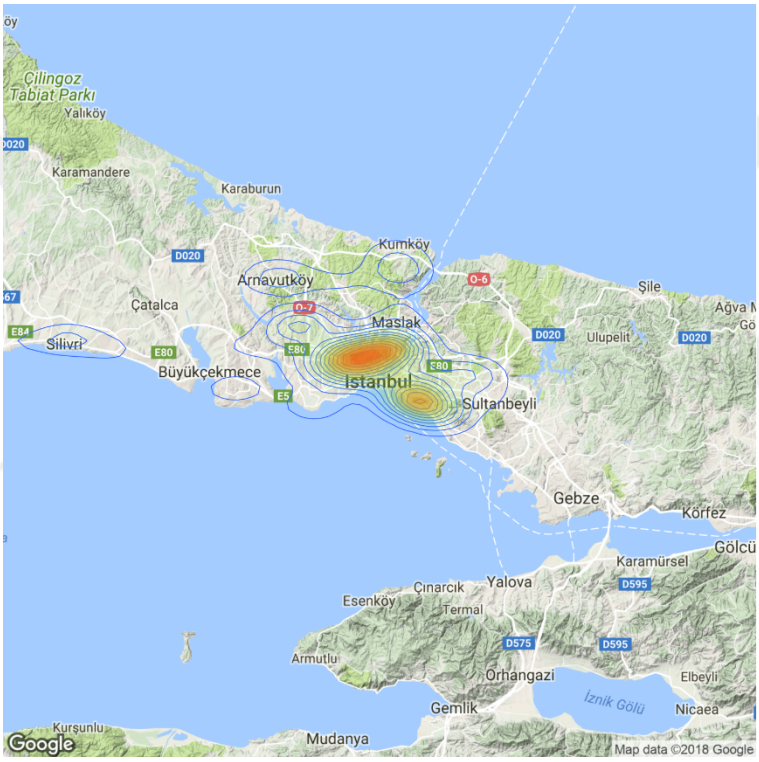


Figure 5.5: Posting Rates for Sale.

From the map, it can be seen that, the majority of the data comes near the Bosporus. Also at Figure 5.6 the price range of the houses in different zones can be seen. At the map, if a house is expensive than others, it is plotted in dark purple and if it is less expensive, it is plotted with light purple dots. Knowing this, it can be stated that sea cost houses are more expensive than other. On the other hand, just like the heat map, the plots which have the same coordinates overlap and create a darker color. So, looking at the previous map, the house prices at Beşiktaş and Kadıköy may not be as expensive as it seems.

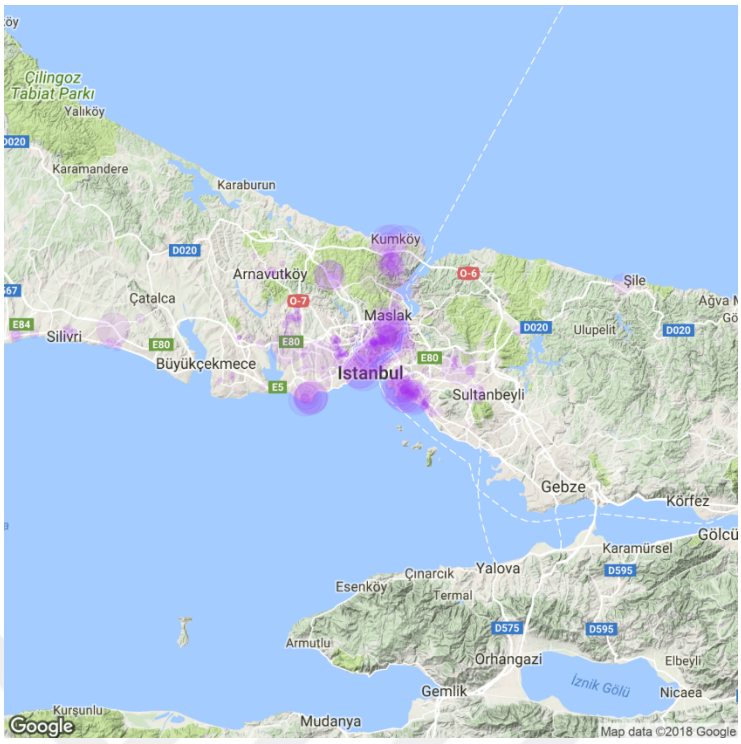


Figure 5.6: Price Distribution of Real Estate for Sale.

The same process is applied to the real estate on rent and the outcomes can be seen from Figure 5.7.

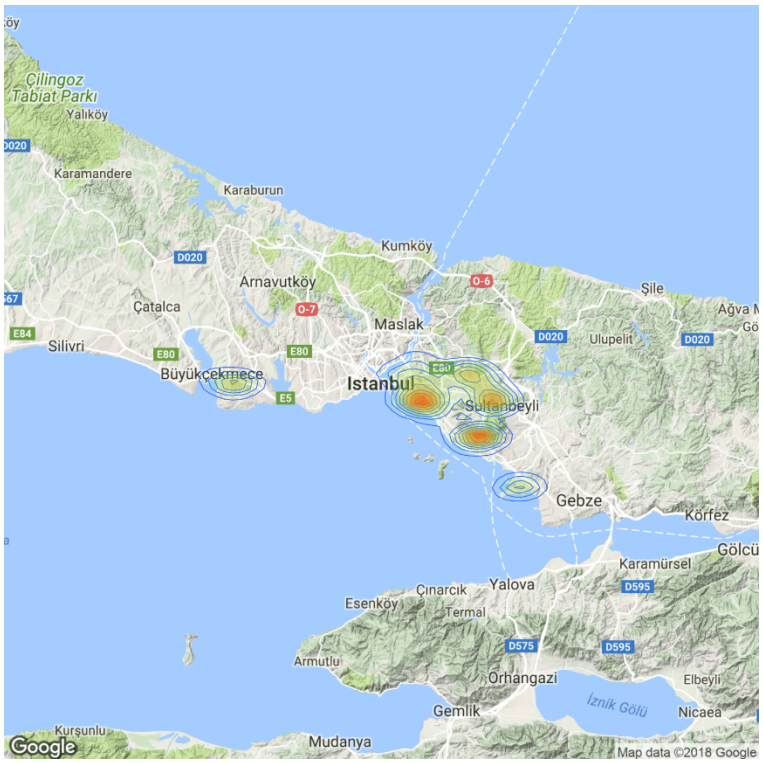


Figure 5.7: Rental Posting Rates.

It can be seen from the graph that, unlike the real estate for sale, rental postings are generally located in Asian side. On the other hand, from some of the rental data, the locations could not be retrieved correctly, so because of the missing parameters, some data could not be plotted. At Figure 5.8, the rent amount of the real estates can be seen.

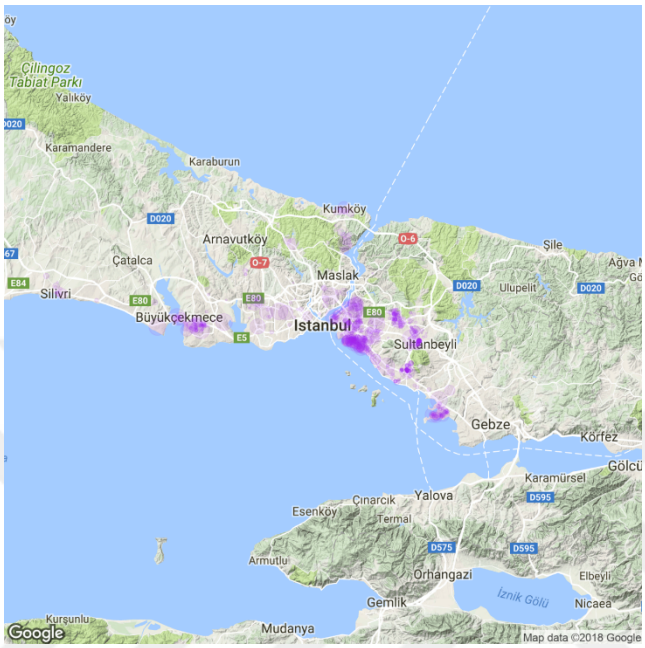


Figure 5.8: Price Distribution of Rentals.

5.5.2 Word Analysis

Word Frequencies

Using the R package, the top words used in the dataset can be listed. At Figure 5.9 below, frequencies of words higher than 500 are listed.

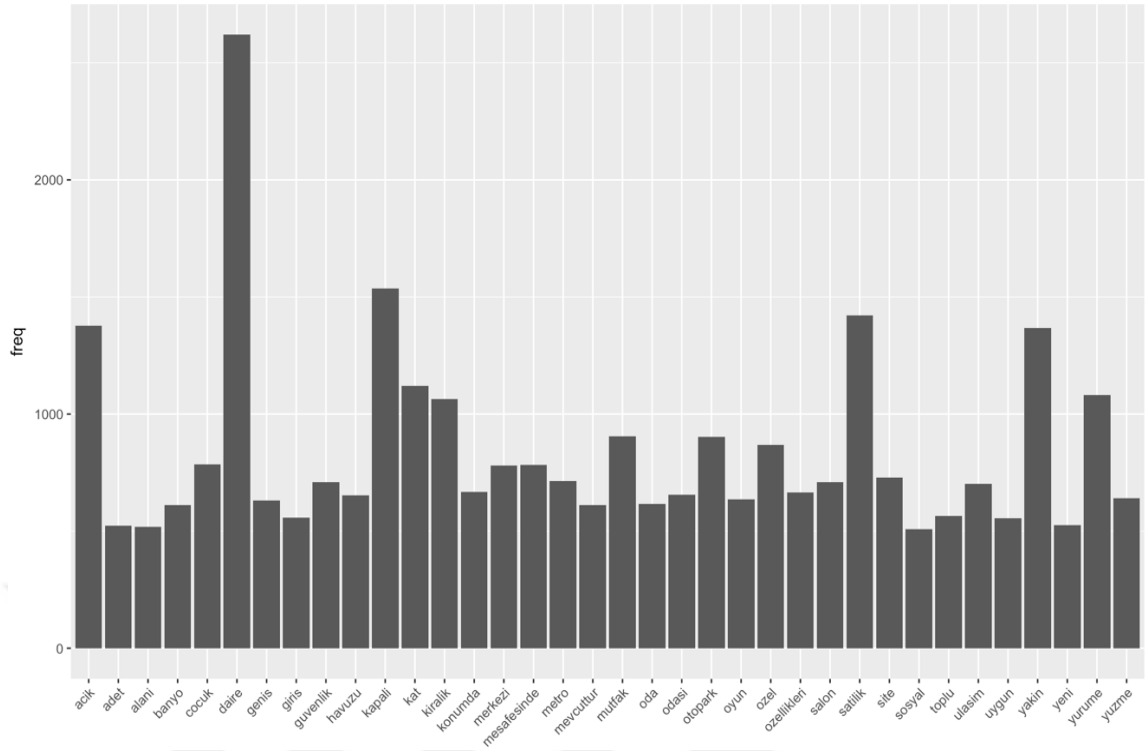


Figure 5.9: Frequencies of Words higher than 500.

Looking at the graph it can be seen that the top 5 words used are “Daire, kapalı, satılık, açık, yakın” which respectively means flat, closed, for sale, closed and near.

Word Cloud

Using the R package overall word-cloud is formed to see the most used words visually. At Figure 5.10 most used words from the word cloud can be seen.

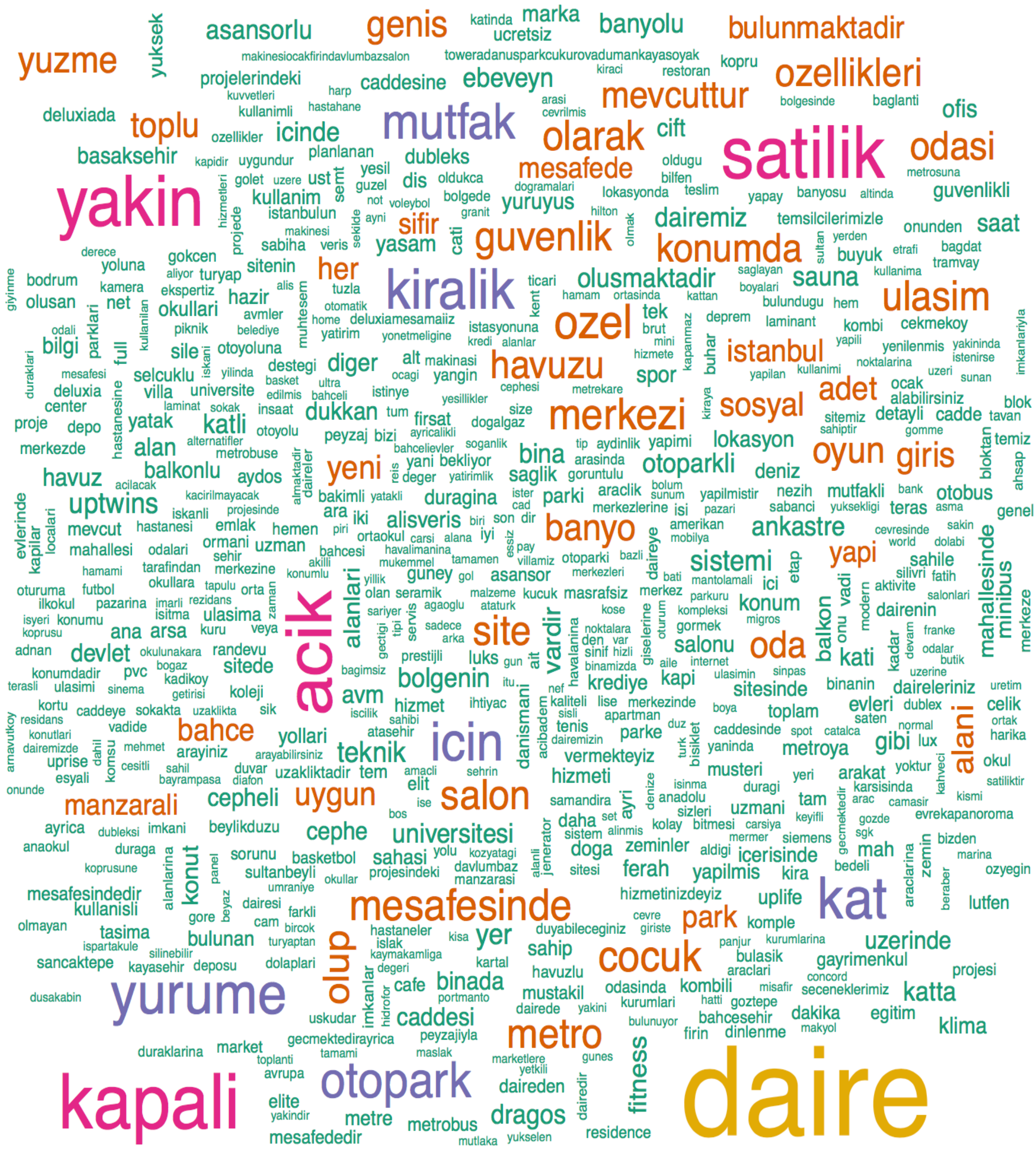


Figure 5.10: Overall Word Cloud.

Correlation Plot

Using TM package of R correlation between low frequency words can be plotted. By using this technique, correlation between more than two items can be observed. At Figure 5.11, the correlation plot of the words which have lower frequency than 200 can be seen.

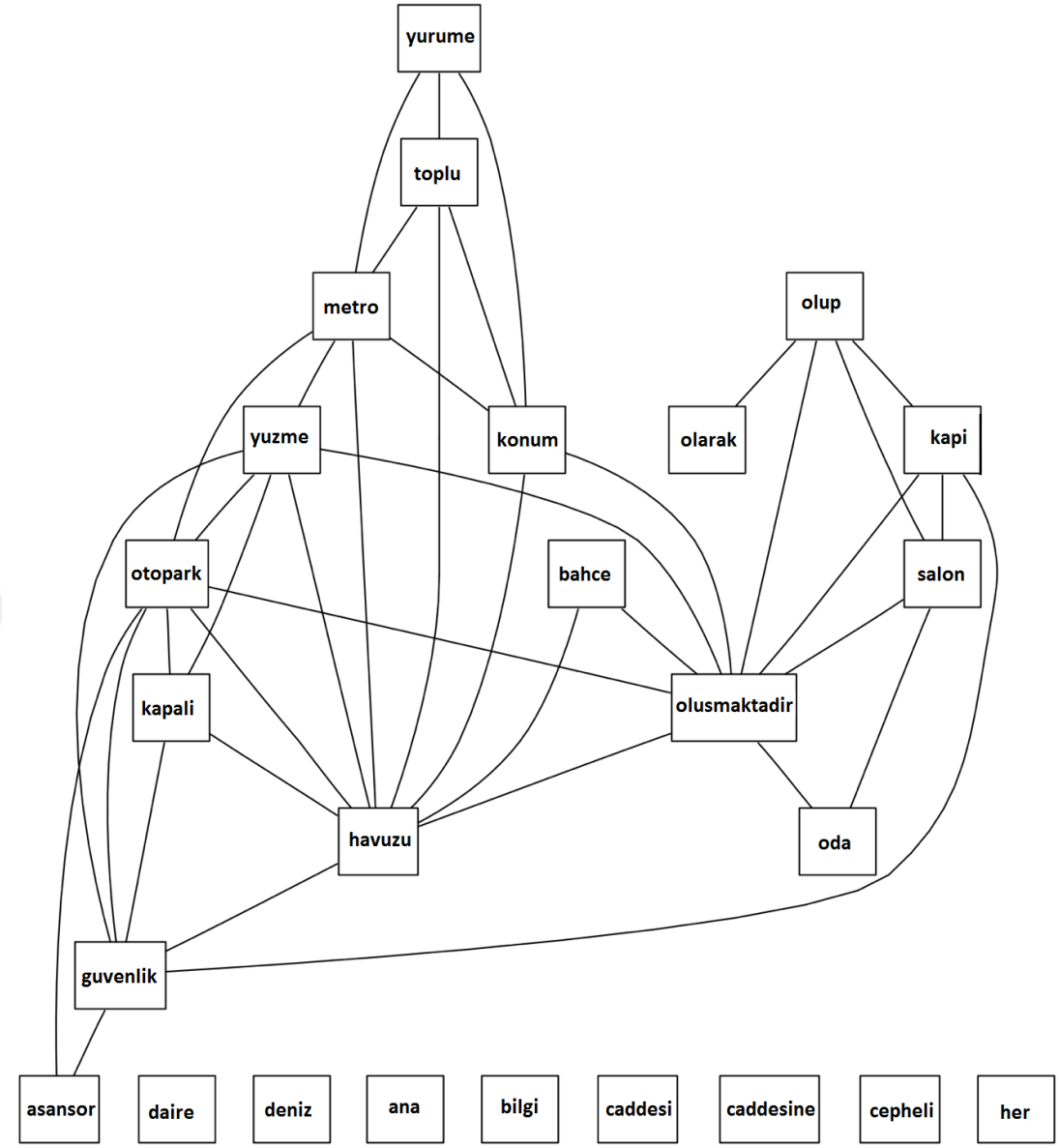


Figure 5.11: Correlation plot using the low frequency words.

By looking at Figure 5.11, to give an example; metro, walking, mass and position words correlate with one another and correlate individually with other words like pool.

5.6 MALLET

MALLET is a Java based topic model package that consists of LDA applications that uses Gibbs sampling which gives opportunity to use hyperparameter optimization and turning unsupervised LDA to supervised LDA by using training set.
At the application, data is loaded by txt. format. Each document should be saved as txt. and form a corpus inside a file. After the dataset is imported (using the import-dir --input /data/topic-input command), the commands below are used to do the topic modelling[65]:
--keep-sequence:
--remove-stopwords:
--num-topics
--num-iterations
--optimize-interval
--optimize-burn-in
Output commands:
--output-state [FILENAME]
--output-doc-topics [FILENAME]
--output-topic-keys [FILENAME]

Using the codes above, MALLET application of LDA is done trying the parameters of 2-20 topics and 200-2000 iterations. Likewise, using optimize-interval command with 10-50 intervals hyperparameter optimization is tested but as a result no critical change occurred. Since the hyperparameter optimization is not used, the Dirichlet parameter of the outcome stayed the same which is $\alpha = 5/$ Topic and it is equal to 0,8333. Lastly, the most meaningful topics are found at 6 topics at 1000 iterations. The found topics and top 15 words at each topic are listed at the Table 5.2.

Table 5. 2: Topic Name Generation.

Topic Number	Most Used Words	Topic Name	Simplified Topic Name
1	kapali acik otopark guvenlik site yuzme sistemi cocuk havuz fitness alani saat oyun ozellikleri odasi kat ofis dukkan giris bina alan	Outer properties of real estate	Outer Property
2	uygun uzerinde arsa toplam kati alani buyuk yer bulunmaktadir daire yakin yurume cok	Type of real estate	Type
3	mesafesinde otoparkli merkezi genis konumda toplu sifir yeni ferah ulasım metro	Transportation possibilities at the location of real estate	Location/Transportation
4	satilik kiralik uptwins park yapi teknik konakları uplife bilgi konut gayrimenkul hizmet elit elite daireleriniz	Real Estate Construction Firms and Real Estate Agents	Construction firms and agents
5	ozel cocuk universitesi sosyal acik devlet oyun havuzu mevcuttur okullari yollari metro tem aydos egitim	Neighborhood Advantages of the Real Estate	Neighborhood
6	salon oda mutfak banyo yakin giris ozellikleri dairemiz olusmaktadir içinde bahce katli zeminler balkon ankastre	Inner Properties of Real Estate	Inner Property

After the topics are formed, the keywords are translated to English. At Table 5.3 below the last version of the topics and keywords can be seen.

Table 5. 3: Last Version of Topics and Keywords.

Topic Number	Most Used Words	Simplified Topic Name
1	closed open parking-lot security site swimming system child pool fitness space hour/clock gaming properties room floor office shop entrence building area	Outer Property
2	affordable above land total floor-of space big place found	Type
3	flat near walking very distance with-parking-lot central wide location mass zero new roomy transportation metro	Location/Transportation
4	for-sale rental Uptwins Park Yapı Teknik Konakları Uplife information house real-estate service elit elite flats	Construction firms and agents
5	private child university social open public state gaming/play pool exists schools roads metro Tem Aydos education	Neighborhood
6	living-room room kitchen bathroom near entrance properties flat consist-of inside yard floor grounds balcony encastered	Inner Property

As it can be seen from Table 5.3, there are some proper nouns that needs explanation. Firstly, at topic 4, “Uptwins, Park, Yapı, Teknik, Konakları, Uplife” words are parts of the naming of construction firms. Also “Tem” word form topic 5 comes from the abbreviation of Trans European Motorway and “Aydos” is the name of one of the forests in Istanbul. Also while translating the words from Turkish to English it is seen that, some of the words in Turkish has two meanings, so in English format both meanings are written. As it is said previously, LDA enables positioning of the parameters (both documents and words) more than once. At the Table 5.3 , it can be seen that, the word floor is placed in both 2nd and 6th topic.

Later, using the word- topic proportions taken form MALLET outcomes, wordclouds for each topic is created At Figure 5.12, word-clouds of the topics can be seen.

After the topics are formed, the distribution of topics over districts and other parameters are analyzed.

Figure 5.12: Wordclouds of Topics Generate

6. RESULTS

6.1 Natural Language Processing Models

6.1.1 Word Association

By using NLP applications of TM package on R, the association of a selected word, with the other words can be found out. This model is firstly tested on the word "Asansör" which can be translated as elevator. The outcomes are shown at Figure 6.1.

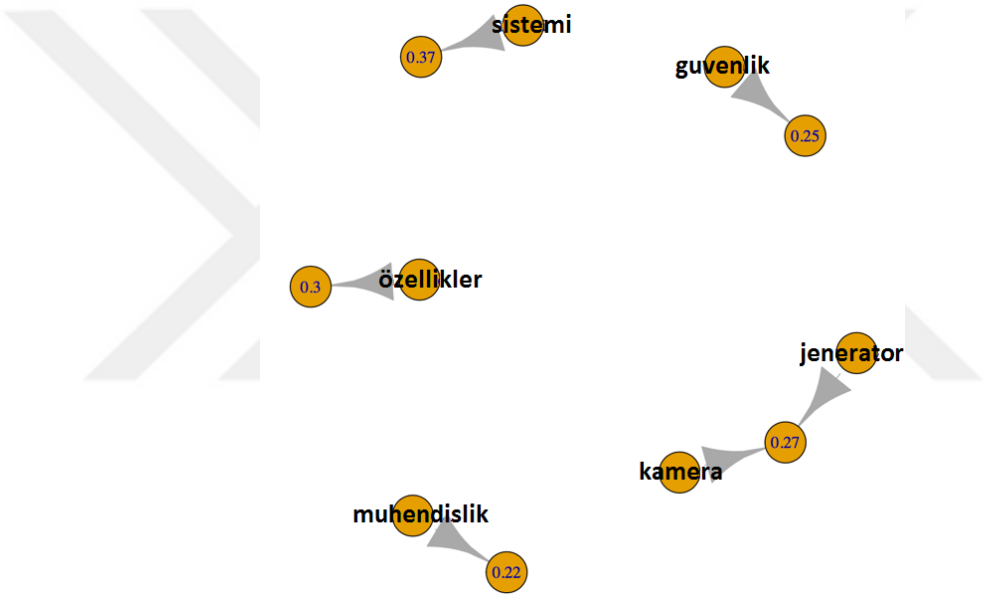


Figure 6.1: Word Association done on word "Asansör".

The word association is then searched within each district in Istanbul. The outcomes generally show the advantages and the properties of that district. The results are listed below:

- Arnavutköy: Associated with airport (0.32) because the third airport will be close to this District, Waterway (0.26) because KanalIstanbul project takes place near and Rising (0.38) because the district will rise with third airport and Kanal Istanbul projects.
- Bahçelievler: Associated with airport (0.36), Metrobus (0.45) which both are near the district.

- Beykoz: Associated with bridge (0.28), Stream (0.50) which both is near the district.
- Beyoğlu: Associated with cinema (0.44), Entertainment (0.76), Theater (0.76), Pasaj (Marketplace) (0.44), Consulate (0.22) because all of them are the major properties of Beyoğlu.
- Çatalca: Associated with farm (0.22), which there are plenty at the district.
- Eyüp: Associated with Mosque (0.34), Pierre Lotti (0.34), which are the major touristic locations of Eyüp.
- Fatih: Associated with Havaray (0.30), which is planned to be built near the district.
- Kartal: Associated with Residence (0.77), which there are plenty at the district.
- Sultangazi: Associated with Dönüm (area scale of land) (0.37), which there are plenty empty lands at the district.
- Zeytinburnu: Associated with Atelier (0.20), which there are plenty at the district.
- Esenler: Associated with metro (1.00), which is near the district.
- Esenyurt: Associated with Vialand (0.23), which is one of the biggest entertainment places in Istanbul and it is at Esenyurt.
- Şile: Associated with landscape(0.23), fireplace(0.82) which many house in the district owns and tourism (0.52), because Şile is one of the touristic sites at Istanbul.
- Ataşehir: Associated with highway (0.53), which is near the district and finance (0.35), because Ataşehir will be the finance capital of Istanbul.
- Sultanbeyli: Associated with airport (0,84), Özyeğin University (0,88), Bilfen College (0.90) and highway (0.83) which are near the district.
- Tuzla: Associated with Bazzar (0.58), Seaside (0.58) and Villa (0.57) which all can be find in the district.
- Başakşehir: Associated with Havaray (0.47), which is planned to be built near the district.
- Çekmeköy: Associated with bridge (0.22) which is near the district and supermarkets Carefour(0.41), Migros(0.31) which are important for the citizens that are living in districts because the district is not at the center location. There aren't too many shops.
- Kadıköy: Associated with street (0.59), bridge (0.23) which both are near the district.
- Beşiktaş: Associated with Shared taxi (0.22), which is an important transportation unit for that district
- Şişli: Associated with Metrobus(0.23) and Courthouse (0.28) which both are near the district.

- Sarıyer: Associated with Koç University (0.30) which is inside the district and dorm (0.48) because of the University.

6.1.2 Sentiment Analysis

Using the R package, the overall sentiment analysis of the real estate posting descriptions are done. The result can be seen at Figure 6.2.

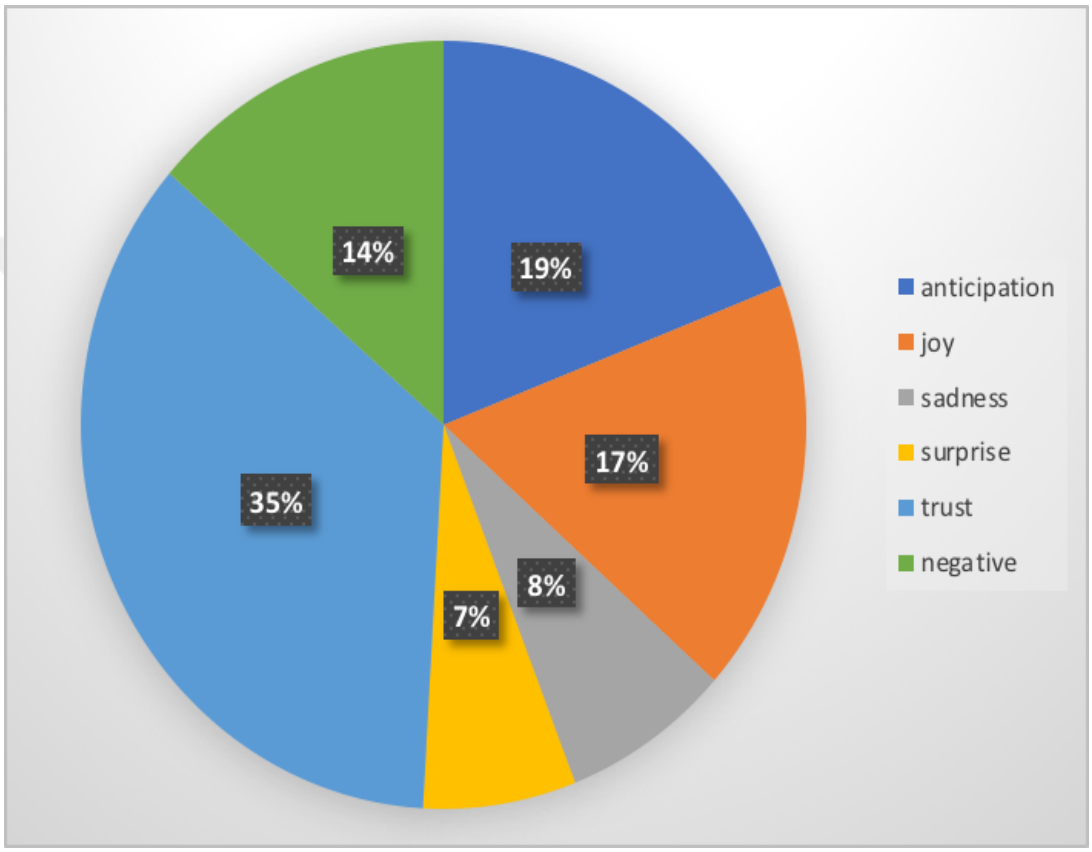


Figure 6.2: Sentiment Analysis Results of Total Data.

As it can be seen from the graph, positive feelings are much more than negative ones because these posting are commercial and to attract customers generally positive feelings are imposed using the right words.

6.2 LDA

After the topics are found using MALLET, the percentages are investigated. From the Figure 6.3, it can be seen that, the location and the transportation possibilities is the most used subject when the details of a house is given. To comment about this fact,

real estate agents may forecast that the home buyers and potential lease-holders look for the ease of transportation in Istanbul. Overpopulation of Istanbul and increment of traffic led people to think about the escape ways. From an article from real estate sector, it is stated that when metro station is built near a neighborhood, the prices of the houses in that neighborhood increases while the station is constructed[64]. The other topics are respectively the type of real estate, outer property of the houses, the information about construction firms and real estate agents, inner property of the houses and lastly the neighborhood of the real estate.

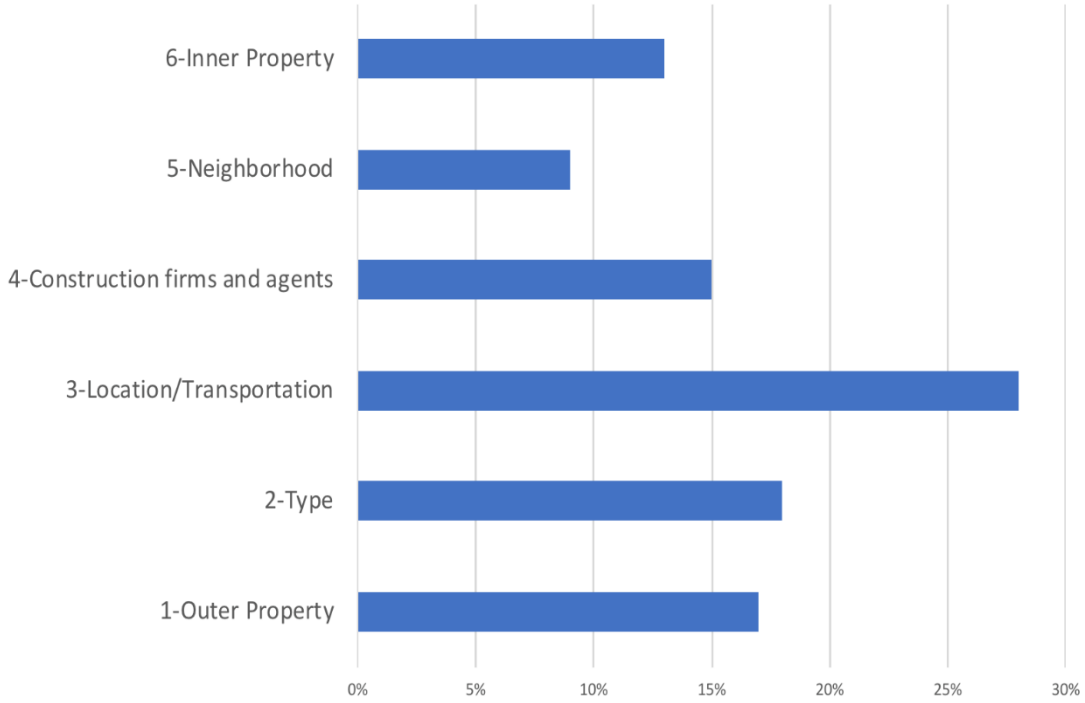


Figure 6.3: Overall Topic Percentages.

6.2.1 Topic Distribution of Districts

After the percentage of the topics are found, the percentage variation of the topics with respect to the districts is studied. In Figure 6.4, the topic percentages for each District can be seen.

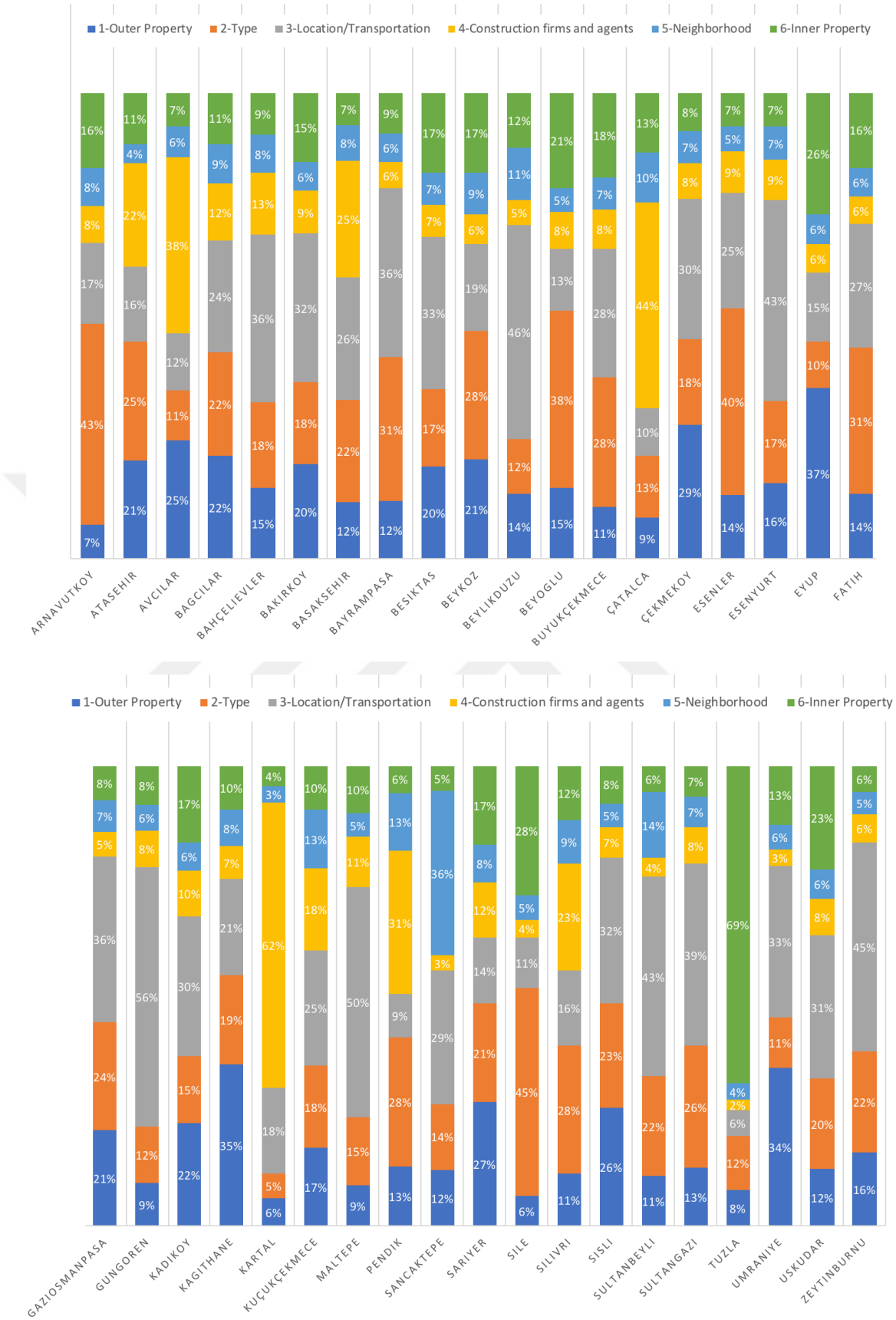


Figure 6.4: Topic Percentages of Districts.

After topic distributions of the districts are observed, topics of the districts outside the average percentages will be observed. Starting from the first topic, the average percentage of Outer Property is 16.69% and the Figure 6.5 shows the percentages of the Districts on this topic.

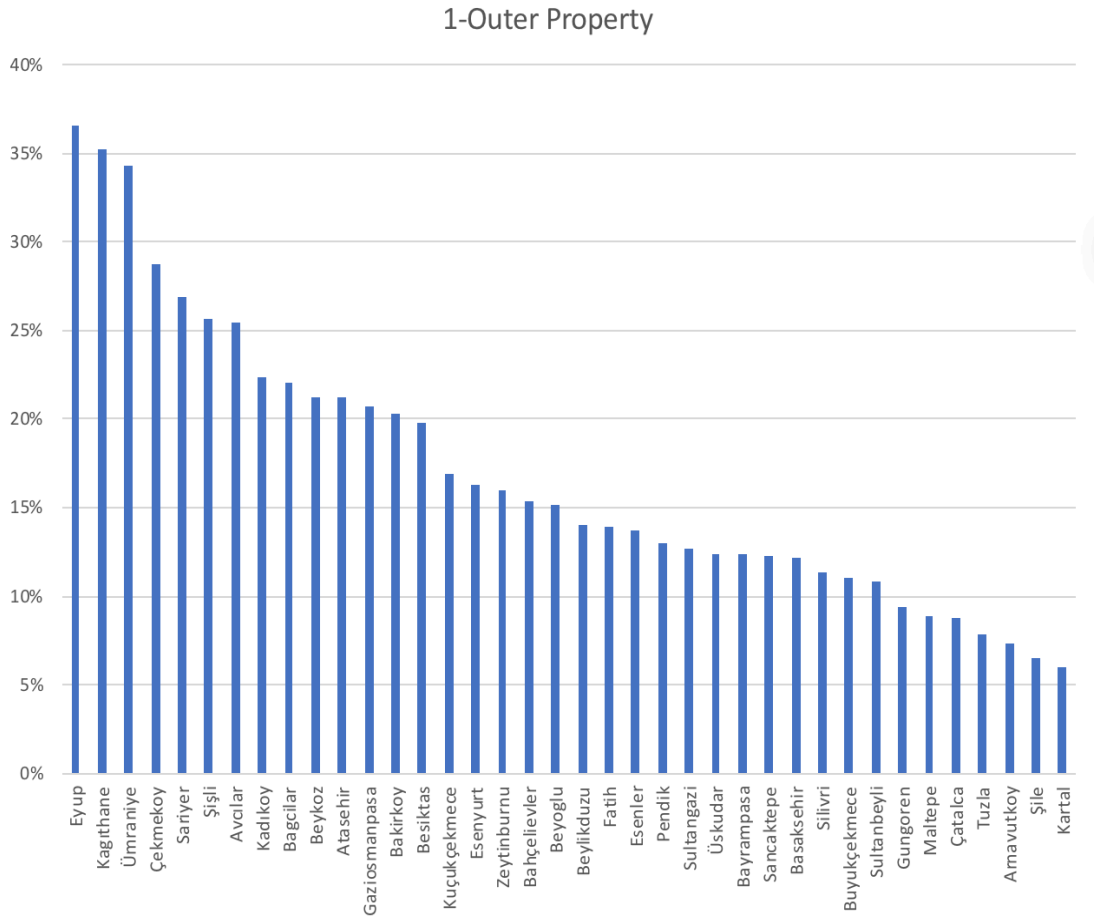


Figure 6.5: Topic 1 Percentages.

As it can be seen from the graph; real estate posts which shows the houses in Eyüp, Kağıthane, Ümraniye and Çekmeköy has high percentage of Outer Property topic which includes the properties of the houses outside the lot such as parking lot, fitness room, swimming pool etc. Postings from Tuzla, Arnavutöy, Şile, Kartal districts have lower percentage of this topic.

Next, to investigate the second topic, the average percentage of Type is 18.44% and the Figure 6.6 shows the percentages of the Districts on this topic.

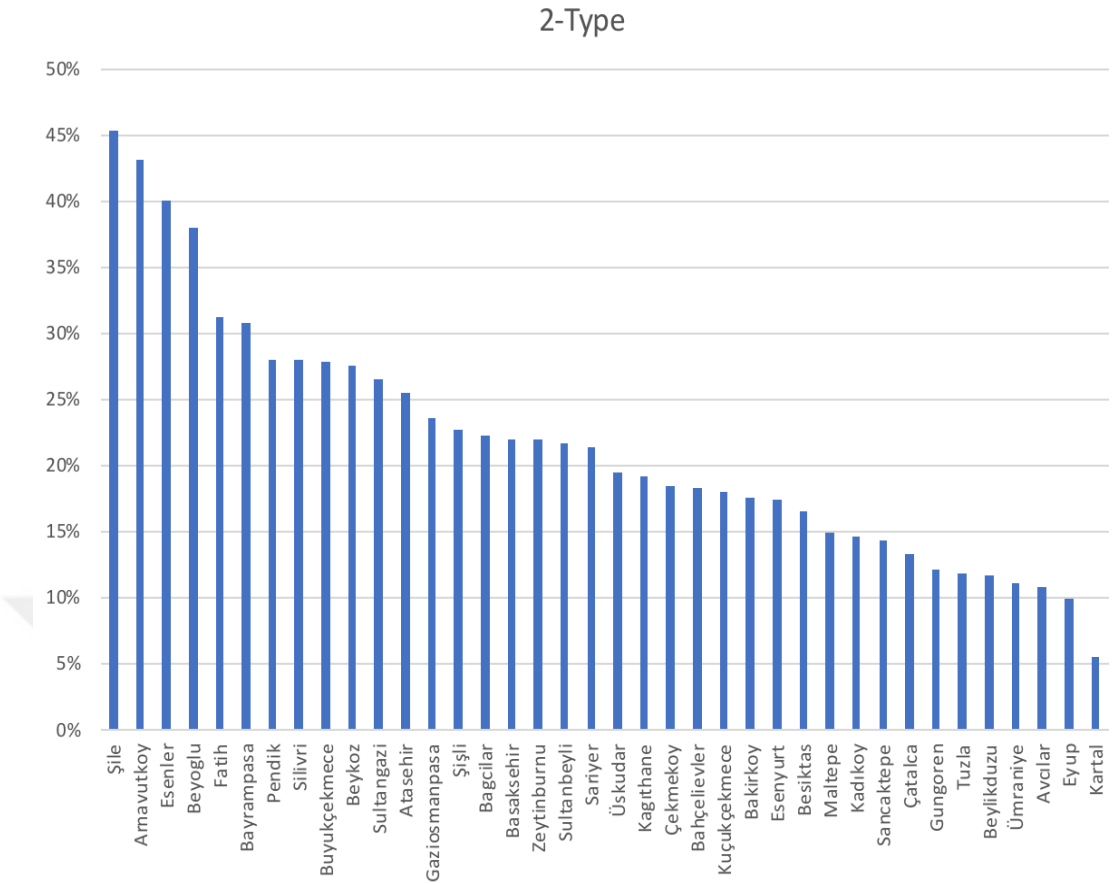


Figure 6.6: Topic 2 Percentages.

The graph shows that; real estate posts from the real estates in Şile, Arnavutköy, Esenler and Beyoğlu includes high percentage of Type topic which includes real estate types like flat office shop land etc. It can be assumed that, at Şile and Arnavutköy there are plenty of empty lands. Postings from Umraniye, Avcılar, Eyüp Districts have lower percentage of this topic.

After, third topic is searched. Knowing that the average percentile of Location/Transportation is 27.22% and using the average value; the Figure below, which shows the percentages of the Districts on Location/Transportation, can be interpreted.

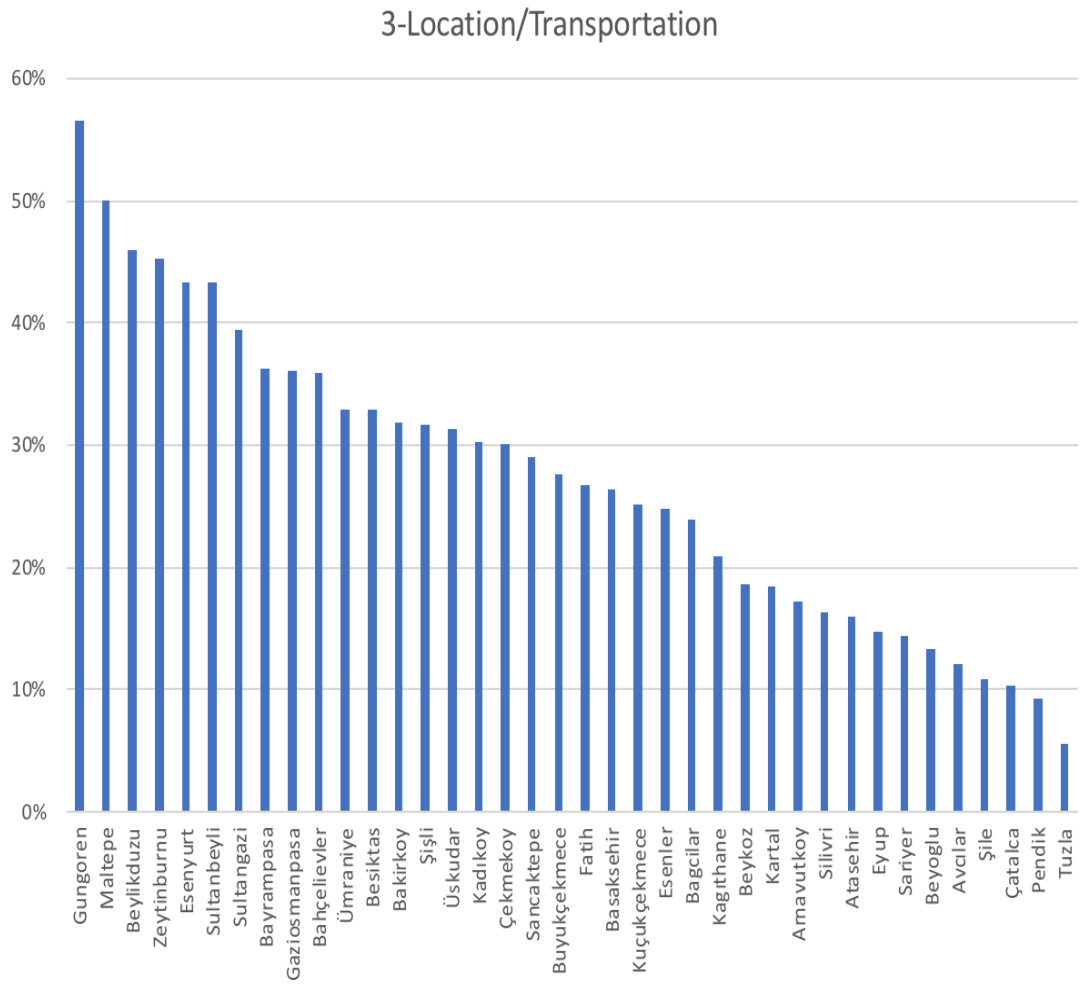


Figure 6.7: Topic 3 Percentages.

As it can be seen from the graph; real estate posts from the real estates in Güngören, Maltepe , Beylikdüzü and Zeytinburnu includes high percentage of Location/Transportation topic. Postings from Avcilar, Şile, Çatalca, Pendik, Tuzla districts have lower percentage of this topic.

At the fourth topic, the average percentage of construction firms and agents is 14.9% and the Figure 6.8 shows the percentages of the districts on this topic.

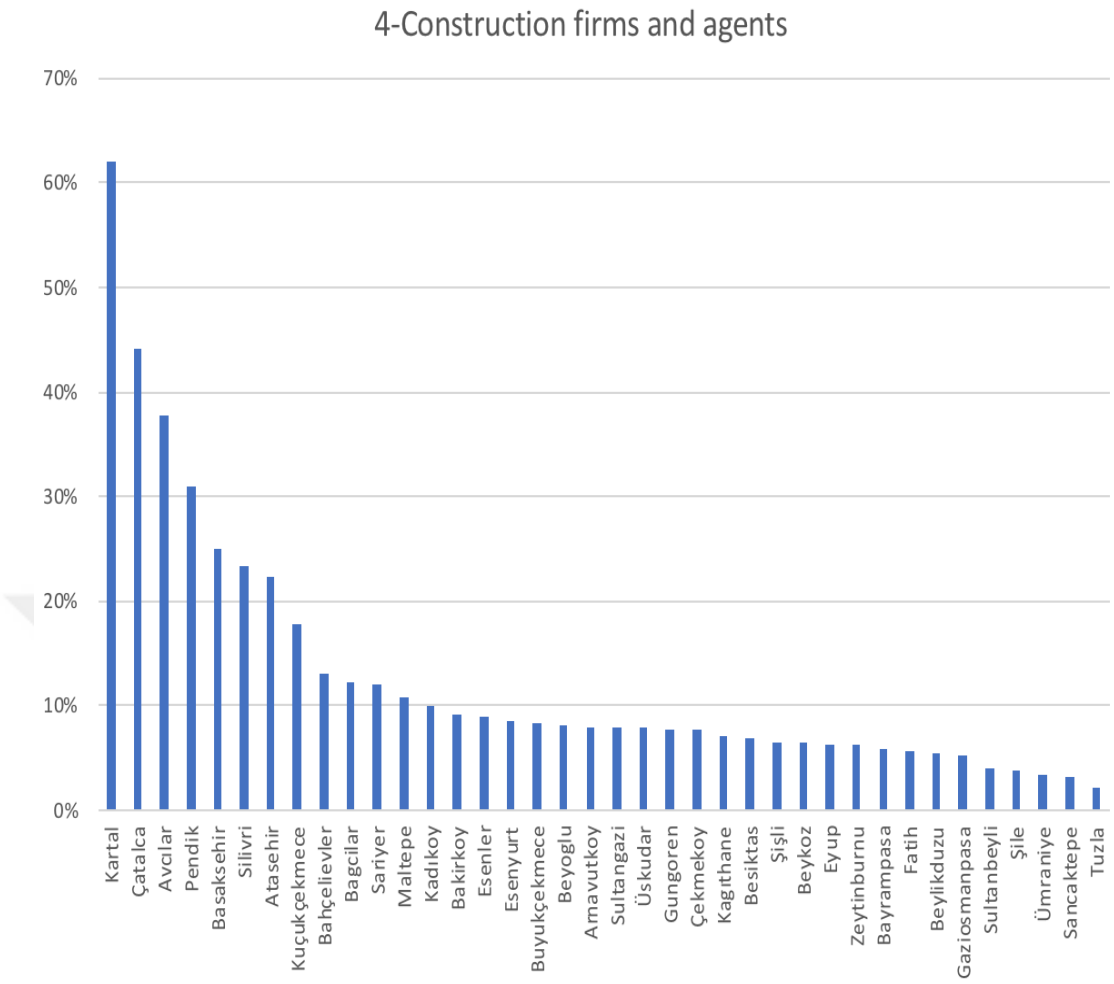


Figure 6.8: Topic 4 Percentages.

To explain the graph above; real estate posts from the real estates in Kartal, Çatalca, Avcılar, Pendik and Başakşehir includes high percentage of construction firms and agents topic. Postings from, Şile, Ümraniye, Sancaktepe and Tuzla Districts have lower percentage of this topic. It can be seen by itself, Kartal increases the average. At this topic generally the advertisement of the construction firm and real estate agents are done and mostly these construction firms creates big lots like residences. Also it can be seen that ,Tuzla has more than 3 topics that has percentage lower than 5%. This means that Tuzla will have peak at one subject.

After, when the second topic is investigated, the average percentage of Neighborhood is 9.07% and the Figure 6.9 shows the percentages of the Districts on this topic.

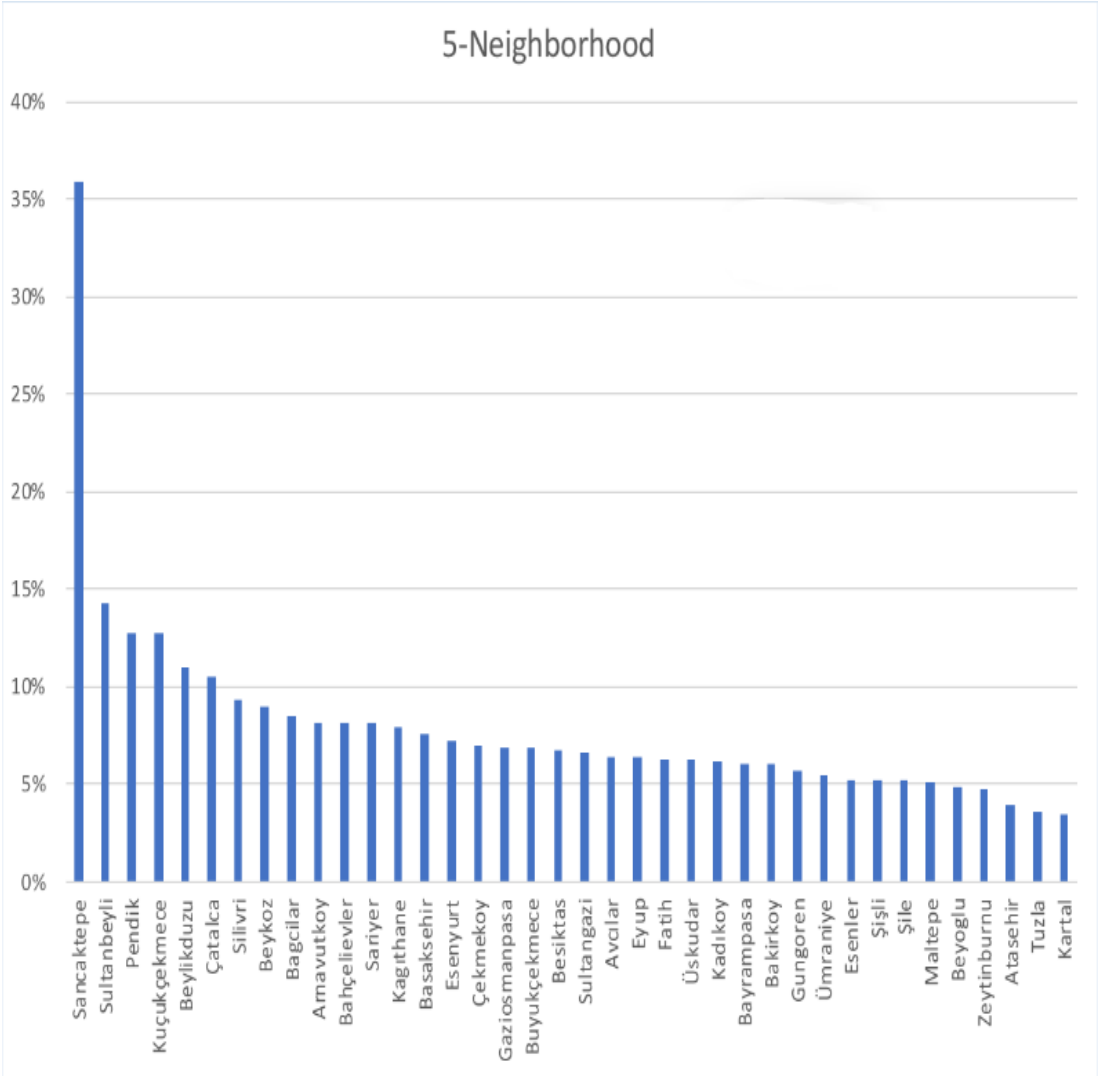


Figure 6.9: Topic 5 Percentages.

Using the graph it is seen that ; real estate posts from the real estates in Sancaktepe includes high percentage of Neighborhood topic. Generally the percentages are stable instead of Sancaktepe. Sancaktepe increases the average by itself. This can be the result of the location of Sancaktepe. It is a newly growing localization zone that the area should be explained in detail.

Lastly at the sixth topic, the average percentage of Inner Property is 13.07% and the Figure 6.10 shows the percentages of the Districts on this topic.

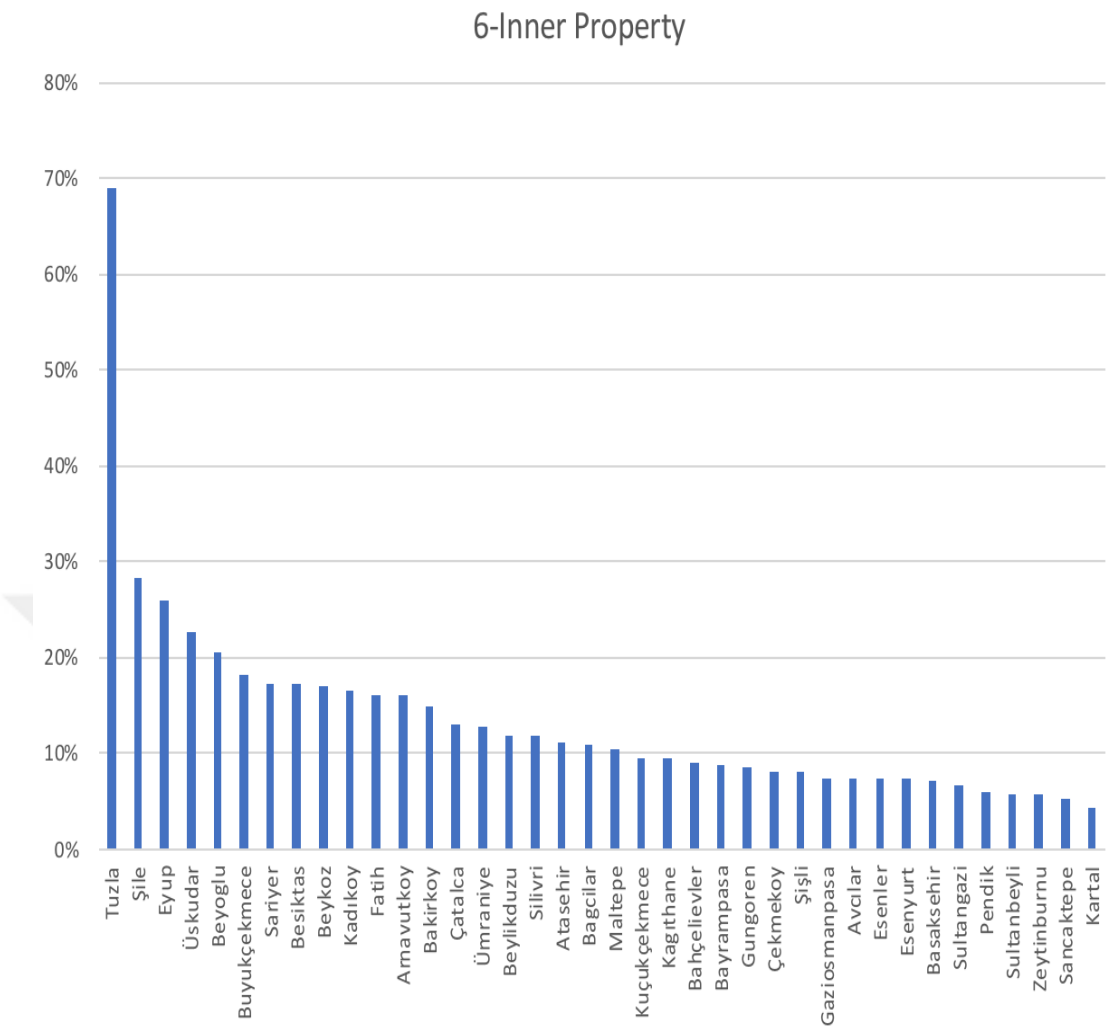


Figure 6.10: Topic 6 Percentages.

As it is stated before, Tuzla peaked the percentage at this topic and changed the Tuzla changed the average of the last topic. Generally other districts have similar percentages despite Tuzla.

To add before passing to another subject, districts like Bağcılar, Kağıthane and Başakşehir's topic percentages are very stable which means that their postings have high amount of diversity that each topic is seen at near rates. Also at Beyoğlu and Kadıköy, it is figured that they both have high percentages in 1st and 6th topics which are the outer and inner properties of the house. This may be occurred because they are both urban transformation zones.

6.2.2 Topic Distribution According to Rental/for Sale

After, the topic percentages of rentals and real estates for sale can be contrasted using the Figure 6.11.

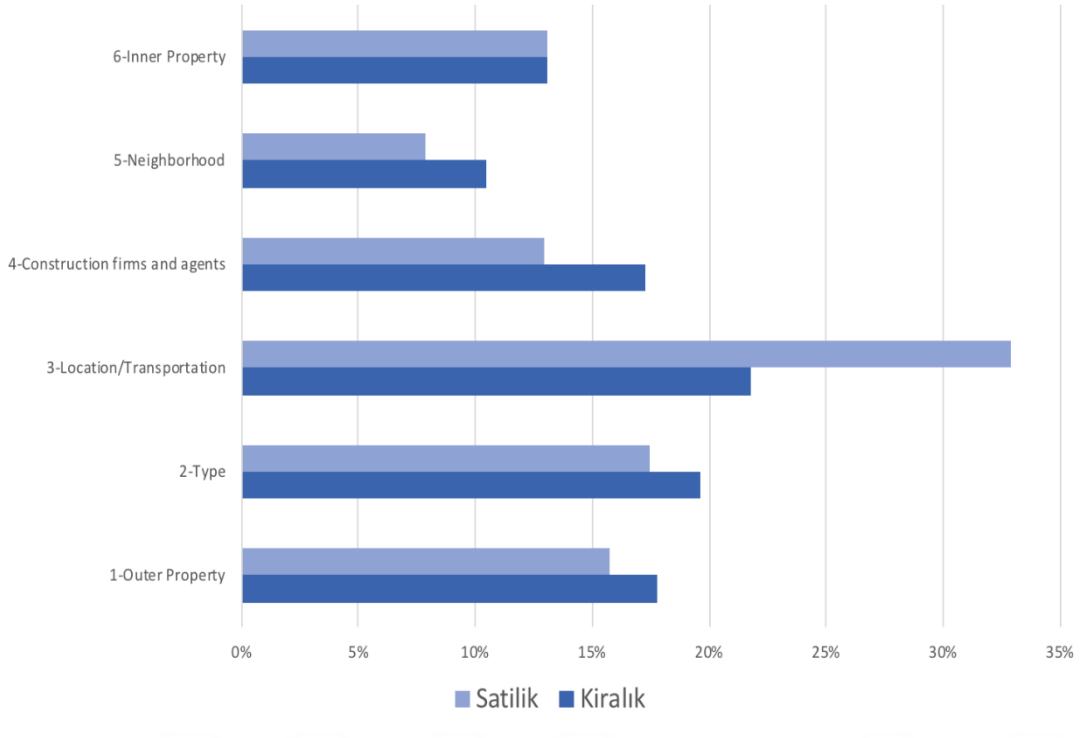


Figure 6.11: Rental/Sale Topic Percentages.

As it can be seen from the figure, there is only one obvious difference and it is located at 3rd topic. The location theme is used more on the real estates that are at sale.

6.2.3 Topic Distribution According to the Real Estate Type

The topic percentages of real estate types can be contrasted using the figure below. From the Figure 6.12, it can be seen that, topic 2 is used more at the office type of real estate and topic 4 is used more while defining residences.

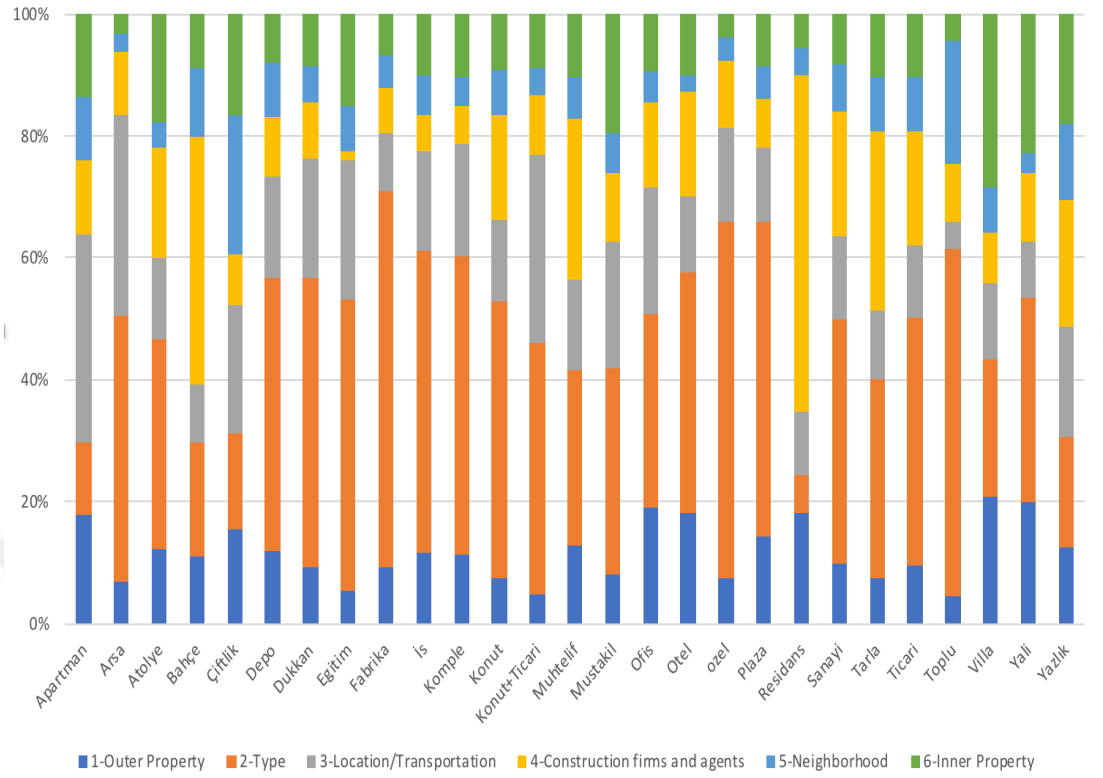


Figure 6.12: Real Estate Type Topic Percentages.

6.2.4 Topic Distribution According to Geographical Zones

From the HDI (human development index) 2017 report of INGEV the districts are rated with respect to the Governance, Social Coverage, Economic Status, Education, Health, Social Life, Municipal Environmental Performance and Transportation. Istanbul districts are rated in an order of; Beşiktaş, Kadıköy, Şişli, Bakırköy, Maltepe, Üsküdar, Sarıyer, Ataşehir, Ümraniye, Beyoğlu, Fatih, Avcılar, Beylikdüzü, Tuzla, Çekmeköy, Başakşehir, Pendik, Kartal, Küçükçekmece, Bayrampaşa, Eyüp, Silivri, Beykoz, Esenler, Kağıthane, Bahçelievler, Güngören, Gaziosmanpaşa, Büyükçekmece, Zeytinburnu, Sultanbeyli, Esenyurt, Arnavutköy, Çatalca, Bağcılar, Sancaktepe, Sultangazi, Adalar and lastly Şile [66]. Using this rating, Istanbul is divided into four zones to test the hypotheses of if the topics and words, that are used at the explanations of the posts, varies different zones of a city. The divided zones can be seen at Table 6.1.

Table 6. 3: Division of Geographical Zones.

District	Zone
Ataşehir	1
Bakırköy	1
Beşiktaş	1
Kadıköy	1
Maltepe	1
Sarıyer	1
Şişli	1
Üsküdar	1
Avcılar	2
Basakşehir	2
Beylikdüzü	2
Beyoğlu	2
Çekmeköy	2
Fatih	2
Kartal	2
Pendik	2
Tuzla	2
Ümraniye	2
Bahçelievler	3
Bayrampaşa	3
Beykoz	3
Büyükçekmece	3
Esenler	3
Eyüp	3
Gaziosmanpaşa	3
Gungören	3
Kağıthane	3
Küçükçekmece	3
Silivri	3
Zeytinburnu	3
Arnavutkoy	4
Bağcilar	4
Çatalca	4
Esenyurt	4
Sancaktepe	4
Şile	4
Sultanbeyli	4
Sultangazi	4

After the zones are set, the topic distributions of different zones are observed. At the Figure 6.13, the topic percentages of the geographical zones can be seen.

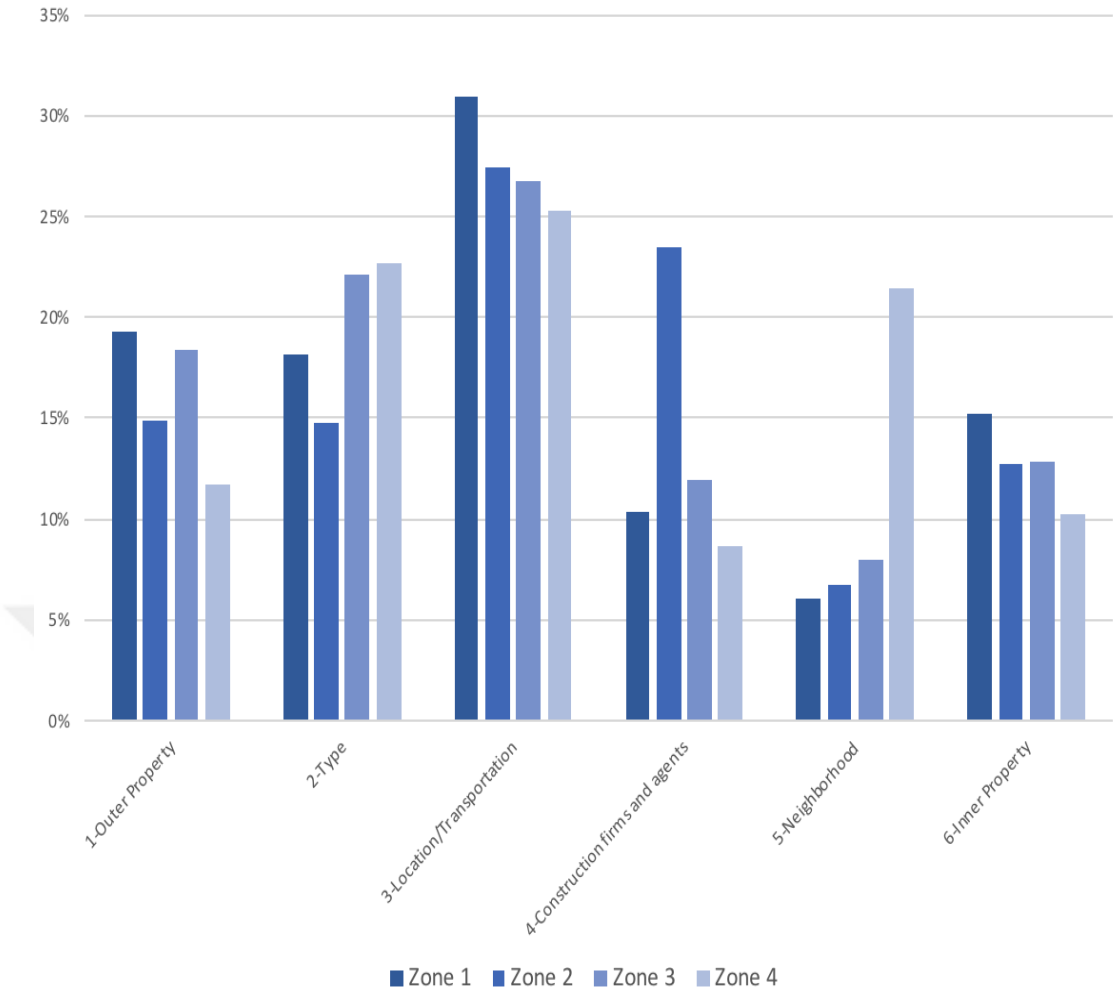


Figure 6.13: Geographical Zones Topic Percentages.

At the figure, it can be seen that, there are observable differences at topic 3, 4 and 5. At topic 3, Zone 1 has higher percentage (0.3) than other zones and looking at topic 4, Zone 2 has higher percentage (0.23) than other zones while the other zones are approximately (0.1). Also, it is seen that 5th topic is mostly used by zone 4.

7. CONCLUSION

Real estate is one of the sectors that has global importance and at the same time, it is one of the fields that can be actively tracked anytime using internet. Nevertheless, selling and buying actions are not easy as tracking and reaching the information. That is why, although reaching to the potential buyers through the internet is easy, there is still a need for intermediaries in this sector. As it is figured in literature, customers look for many parameters and alternatives before deciding to buy an item and the potential buyers of real estate look for the descriptions first. Knowing this, the intermediaries should know the customer desires to advertise the real estate. As it is known, customer wants change person to person, thus the realtor should understand the potential buyer and propose the real estate accordingly. Claiming that the realtors understand the wishes of the customer, real estate postings of Istanbul realtors are investigated. The following questions are aimed to be answered:

- What are the topics that the real estate agents use to gain consumer interest?
- Which words from the topics of real estate posts are the most frequent ones?
- Do topics differ by type of real estate (office, house, apartment, villa) or rental (for rental or sale) or the place of the real estate (district)?
- Do these parameters differ by geographical zones based on human development levels?
- What kind of sentiments are dominant in the descriptions of the real estate posts?

To answer the first four questions above LDA is used. After, data preparation is done by applying stemming, eliminating stop words etc. Then, the data is imported and the algorithm is implemented. Because of the word counts at the documents and the size of the corpus, the algorithm is operated assigning 2 to 20 topics. The one that has most significant outcome is selected which is six topics. The topics that are constructed from the words of the latent topics are; outer property, type, location/transportation, construction firms and agents, neighborhood and lastly the inner property of the real

estate. Using topics and the topic proportions, analysis regarding the district, rental type and rental/for sale is done.

As an outcome, it is seen that some districts peaked at some topics and some districts lacked at some topics. When we look at topic 1 distributions, it is seen that the districts that are growing (Çekmeköy, Ümraniye, Kağıthane and Eyüp) used the outer properties of the houses. This finding can be considered plausible as in the growing districts, new houses are built and generally the houses that are newly built include outer properties such as security and closed parking lot.

When second topic is investigated, the districts (Şile, Arnavutköy, Esenler) that land is sold used this topic more than other districts. Also from the Real Estate type topic percentages, it is seen that while promoting Office type real estate topic 2 is used more than other types. At this topic, words like “land” and “office” are used to provide details about the type of the real estate.

The districts (Tuzla, Çatalca, Şile) that lack the third topic (location/transportation) are generally far from the center and the transportation possibilities are not as good as the other districts.

At the forth topic (construction firms and agents), Kartal peaked and increased the average. Kartal is a growing district where many residences are built by the branded construction firms, so this result is also acceptable.

Sancaktepe increases the average of the overall fifth topic (neighborhood) percentage. Sancaktepe is a newly grown district that allows migration and it is less known by the others. So, describing the neighborhood is good for the potential buyers.

Like Kartal and Sancaktepe, Tuzla draws attention at the topic 6 (inner property). The real estate agents may have only used the parameters such as room number, kitchen area in this district.

Lastly, looking at the topic percentages of geographical zones created using human development index, three parameters are differentiated. It is seen that the first zone uses topic 3 more than other zones. It is meaningful because, to the people who are at the first zone, the location of the real estate may be more important. Transportation possibilities may be more crucial for them since they generally have an active working life. The second zone is remarkable at the topic 4. This finding can also be considered plausible as the districts inside this zone are generally rising and new real estate projects are undertaken. Lastly, the last zone peaked at topic 5 (neighborhood topic). This outcome also makes sense because the districts in these zones are not known as

well as the other zones so explanation of the neighborhood is needed. As a result of this outcome, we can assume that the people living at zone 1 cares about the location and transportation possibilities, people who live in zone 2 look for newly built houses that are built by branded construction firms and lastly people living in the last zone care about the neighborhood of the house.

To answer the question regarding the feelings of the realtor which he/she is trying to point out while writing the descriptions of real estate posts, semantic analysis is done. As it is forecasted, realtors had positive feelings while writing descriptions.

In addition, at the word association analysis, the outcomes of the word associations of district generally show the advantages and the properties of that district. For further study, knowing the results from the word association it can be said that, this data type can be used as the training data. By using Markov Chains in NLP, the patterns can be learnt by training the data. After the patterns are learnt, word prediction and autocomplete for real estate processes can be done. On the other hand, this rate of (5000) data may not be enough for the training process. To get accurate outcomes larger dataset should be used.

The limitations of the study include having a small data set and using Turkish language for the analysis. If the dataset were bigger, more discrete topics would be extracted. The Turkish word stemming algorithm may have decreased the accuracy of the project as the performance of Turkish algorithm for stemming is not as good as those of the ones used for English language.

In the future, the corpus used in this study can be enlarged so as to include other web based real estate databases of Istanbul. Moreover, the methods used in this study can be applied to other cities of Turkey. The topic models and other word usage patterns of Real Estate sector can be compared with other e-commerce categories like second hand goods, books, music etc. In addition, the relationship between the topic modeling parameters and the real-estate features like price and size (m^2) could be investigated. Similarly, the association between text mining indicators and post based web analytics features like duration and hit counts could be evaluated. Finally, more advanced NLP techniques based on supervised and unsupervised Deep Learning methodologies could be used to evaluate the real estate patterns of both Istanbul and other cities/regions of Turkey.

REFERENCES

[1] Feuerriegel, S., Ratku, A., & Neumann, D. (2016). Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation. In *System Sciences (HICSS), 2016 49th Hawaii International Conference on* (pp. 1072-1081). IEEE.

[2] Buxmann, P., & Gebauer, J. (1998). Internet-based intermediaries-the case of the real estate market. In *ECIS* (pp. 60-74).

[3] The National Association of Realtors. (2012). *The Digital House Hunt:Consumer and Market Trends in Real Estate.* Retrieved From https://www.nar.realtor/sites/default/files/documents/Study-Digital-House-Hunt-2013-01_1.pdf

[4] Stapleton, C. O., & Williams, M. R. (2004). *California Real Estate Principles*. Dearborn Real Estate.

[5] Crowston, K., & Wigand, R. T. (1999). Real estate war in cyberspace: an emerging electronic market?.

[6] Ford, J. S., Rutherford, R. C., & Yavas, A. (2005). The effects of the internet on marketing residential real estate. *Journal of Housing Economics*, *14*(2), 92-108.

[7] Low, G. S., & Lamb Jr, C. W. (2000). The measurement and dimensionality of brand associations. *Journal of Product & Brand Management*, *9*(6), 350-370.

[8] Lecinski, J. (2011). ZMOT: Winning the zero moment of truth. Google.

[9] Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic analysis of the digital economy* (pp. 89-118). University of Chicago Press.

[10] Huang, B., & Rutherford, R. (2007). Who you going to call? Performance of realtors and non-realtors in a MLS setting. *The Journal of Real Estate Finance and Economics*, 35(1), 77-93.

[11]-Le, N. Q., & Supphellen, M. (2017). Determinants of repurchase intentions of real estate agent services: Direct and indirect effects of perceived ethicality. *Journal of Retailing and Consumer Services*, *35*, 84-90.

[12]-Mulliner, E., Smallbone, K., & Maliene, V. (2013). An assessment of sustainable housing affordability using a multiple criteria decision making method. *Omega*, *41*(2), 270-279.

[13]- Besbris, M., & Faber, J. W. (2017). Investigating the relationship between real estate agents, segregation, and house prices: Steering and upselling in New York State. In *Sociological Forum* (Vol. 32, No. 4, pp. 850-873).

[14]-Crowston, K., Sawyer, S., & Wigand, R. (2015). Social Networks and the Success of Market Intermediaries: Evidence From the US Residential Real Estate Industry. *The Information Society*, *31*(5), 361-378.

[15]- Schlauch, A., & Laposa, S. (2001). E-tailing and Internet-related real estate cost savings: A comparative analysis of E-tailers and retailers. *Journal of Real Estate Research*, *21*(1-2), 43-54.

[16] Tian, K., Revelle, M., & Poshyvanyk, D. (2009). Using latent dirichlet allocation for automatic categorization of software. In *Mining Software Repositories, 2009. MSR'09. 6th IEEE International Working Conference on* (pp. 163-166). IEEE.

[17] Maskeri, G., Sarkar, S., & Heafield, K. (2008). Mining business topics in source code using latent dirichlet allocation. In *Proceedings of the 1st India software engineering conference* (pp. 113-120). ACM.

[18] Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463-479.

[19] Miller, G. S. (2017). Discussion of "the evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation". *Journal of Accounting and Economics*, *64*(2-3), 246-252.

[20] Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467-483.

[21] Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*, *42*(3), 1314-1324.

[22] Hand, D. J. (2007). Principles of data mining. *Drug safety*, *30*(7), 621-622.

[23] Soman, K. P., Diwakar, S., & Ajay, V. (2006). *Data mining: theory and practice [with CD]*. (pp. 1-5). PHI Learning Pvt. Ltd..

[24] Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. (pp. 9-35). Springer Science & Business Media.

[25] Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

[26] Obenshain, M. K. (2004). Application of data mining techniques to healthcare data. *Infection Control & Hospital Epidemiology*, *25*(8), 690-695.

[27] LG CNS (n.d.). *SEMMA Star* [infographic] Retrieved from http://blog.lgcns.com/1268

[28] Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

[29] Learned-Miller, E. G. (2014). Introduction to Supervised Learning. *I: Department of Computer Science, University of Massachusetts*.

[30] Dayan, P., Sahani, M., & Deback, G. (1999). Unsupervised learning. *The MIT encyclopedia of the cognitive sciences*.

[31] Bharambe, A., Ravindran, R., Suchdev, R., & Tanna, Y. (2014, August). A robust anomaly detection system. In *Advances in Engineering and Technology Research (ICAETR), 2014 International Conference on* (pp. 1-7). IEEE.

[32] Salton, G., & Michael, J. (1983). McGill. 1983. *Introduction to modern information retrieval*.

[33] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391.

[34] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

[35] Hofmann, T. (1999, July). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 289-296). Morgan Kaufmann Publishers Inc.

[36] Ponweiser, M. (2012). Latent Dirichlet allocation in R.

[37] Hiemstra, D. (2009). Information retrieval models. Information Retrieval: searching in the 21st Century, 1-17

[38] Wikipedians. (n.d.). Information Retrieval. (pp. 9-10)

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

[40] Van Eck, N. J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. *arXiv preprint arXiv:1109.2058*.

[41] Hu, J., Li, S., Yao, Y., Yu, L., Yang, G., & Hu, J. (2018). Patent Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification. *Entropy*, *20*(2), 104.

[42] Underwood, T., & Sellers, J. (2012). The emergence of literary diction. *Journal of Digital Humanities*, *1*(2), 1-2.

[43] Dolgun, M. Ö., Özdemir, T. G., & Oğuz, D. (2009). Veri madenciliğinde yapısal olmayan verinin analizi: Metin ve web madenciliği, İstatistikçiler Dergisi, 2, 48-58.

[44] Feuerriegel, S., Ratku, A., & Neumann, D. (2016). Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation. In *System*

Sciences (HICSS), 2016 49th Hawaii International Conference on (pp. 1072-1081). IEEE.

[45] Hladká, B., Holub, M., & Kriz, V. (2013). Feature engineering in the NLI shared task 2013: Charles University submission report. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 232-241).

[46] Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

[47] Sedgewick B. (2004). COS 126:Markov Model of Natural Language. Retrieved From:
http://www.cs.princeton.edu/courses/archive/spr05/cos126/assignments/markov.html

[48] Maldron M. (2015) 10 More Common NLP Terms Explained for the Text Analysis Novice Retrieved from: http://blog.aylien.com/10-more-common-nlp-terms-explained-for-the-text/

[49] **Url-1<** Retrieved from http://www.statsoft.com/textbook/naive-bayes-classifier> date retrieved 01.04.2018.

[50] Url-2< Retrieved from: http://moderndata.plot.ly/elections-analysis-in-r-python-and-ggplot2-9-charts-from-4-countries/> date retrieved 29.03.2018.

[51] Thompson, V. (2017). Methods for Detecting Paraphrase Plagiarism. *arXiv preprint arXiv:1712.10309*.

[52] Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support?. *Journal of biomedical informatics*, *42*(5), 760-772.

[53] Url-3<http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/ date retrieved 01.04.2018.

[54] Jelodar, H., Wang, Y., Yuan, C., & Feng, X. (2017). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *arXiv preprint arXiv:1711.04305*.

[55] Wood, J., Tan, P., Wang, W., & Arnold, C. (2017). Source-LDA: Enhancing probabilistic topic models using prior knowledge sources. In *Data Engineering (ICDE), 2017 IEEE 33rd International Conference on* (pp. 411-422). IEEE.

[56] Url-4<http://mcburton.net/blog/joy-of-tm/> date retrieved 20.04.2018.

[57] Url-5<http://phyletica.org/dirichlet-process/>date retrieved 22.04.2018.

[58] Carlo, C. M. (2004). Markov chain monte carlo and gibbs sampling. *Lecture notes for EEB*, *581*.

[59] Url-6<http://www.ece.umd.edu/~smiran/LDA.pdf>date retrieved 26.04.2018.

[60] Chinaei, H., & Chaib-draa, B. (2016). A Few Words on Topic Modeling. In *Building Dialogue POMDPs from Expert Dialogues* (pp. 7-19). Springer, Cham

[61] Url-7<https://stats.stackexchange.com/questions/266059/is-sparsity-of-topics-a-necessary-condition-for-latent-dirichlet-allocation-lda>date retrieved 22.04.2018.

[62] Url-8<http://www.cs.columbia.edu/~blei/topicmodeling_software.html>date retrieved 01.05.2018.

[63] Emlak Konut. (2014). "Gayrimenkul ve Konut Sektörüne Bakış"Date Retrieved 10.04.2018, adress: http://emlakkonut.com.tr/_Assets/Upload/Images/file/EKGYO-Aralık-Konut-Raporu-Final.pdf

[64] Url-9<http://www.hurriyet.com.tr/ekonomi/kobi/istanbulda-gayrimenkul-sektorunun-canlandigi-ilceler-40487165> date retrieved 15.04.2018.

[65] Url-10< http://mallet.cs.umass.edu/topics.php> date retrieved 23.04.2018.

[66]-Şeker, M., Bakış, Ç.,& Dizeci, B. (2017). *İnsani Gelişme Endeksi – İlçeler (İGE-İ) 2017: Tüketiciden İnsana Geçiş.* Retrieved from http://ingev.org/raporlar/IGE_RAPOR_2018.pdf

[67] Url-11< https://www.milliyetemlak.com/dergi/konut-fiyatlarinda-balon-var-mi-yok-mu/> date retrieved 18.04.2018

[68] Url-12< http://www.sde.org.tr/userfiles/file/T%C3%BCrkiye%20Konut%20Sekt%C3%B6r%C3%BC%20Geli%C5%9Fmeler%20%E2%80%93%20Beklentiler%20Raporu.pdf> date retrieved 18.04.2018

[69] Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X., & Jensen, L. J. (2015). DISEASES: Text mining and data integration of disease–gene associations. *Methods*, *74*, 83-89.

[70] Smalheiser, N. R., & Bonifield, G. (2016). Two Similarity Metrics for Medical Subject Headings (MeSH):: An Aid to Biomedical Text Mining and Author Name Disambiguation. *Journal of biomedical discovery and collaboration*, *7*.

[71] O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches.*Systematic reviews*, *4*(1), 5.

[72] Tian, X., He, W., Tao, R., & Akula, V. (2016). Mining online hotel reviews: a case study from hotels in China.

[73] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*(16), 7653-7670.

APPENDICES

APPENDIX A: R Codes

APPENDIX A

R code for Istanbul Map

```
library(ggmap)
library(ggplot2)

# load the data
emlak <- read.csv("~/Desktop/emlak.csv", sep = ";")
emlak$lon = as.numeric(as.character(emlak$lon))
emlak$lat = as.numeric(as.character(emlak$lat))
emlak$fiyat = as.numeric(as.character(emlak$fiyat))
Istanbul_map <- get_map(location = "Istanbul", zoom = 9)

ggmap(Istanbul_map, extent = "device") + geom_density2d(data = emlak, aes(x =
lon, y = lat), size = 0.3) +
  stat_density2d(data = emlak,
                 aes(x = lon, y = lat, fill = ..level.., alpha = ..level..), size = 0.01,
                 bins = 16, geom = "polygon") + scale_fill_gradient(low = "green", high =
"red") +
  scale_alpha(range = c(0, 0.3), guide = FALSE)

#expensiveness
tx_circle_size <- 0.0007

emlak$build_price <- as.numeric(emlak$fiyat)
ggmap(tartu_map_g_str, extent = "device") + geom_point(aes(x=lon, y=lat),
                                        data=emlak, col="purple", alpha=0.1,
                                        size=emlak$build_price*tx_circle_size) +
  scale_size_continuous(range=range(emlak$fiyat)) +
  ggtitle("Expensiveness of Houses in Istanbul")
```

R code for Wordcloud

```
cname <- file.path("~","desktop","toplam")
library (tm)
docs <- Corpus(DirSource(cname))
docs <- tm_map(docs,tolower)
docs <- tm_map(docs,removeNumbers)
docs <- tm_map(docs,removePunctuation)
docs <- tm_map(docs,stripWhitespace)
docs <- tm_map(docs, removeWords, c("bir", "cok", "i<c3><a7>"))
dtm <- DocumentTermMatrix(docs)
library(wordcloud)
m <- as.matrix(dtm)
v <- sort(colSums(m),decreasing=TRUE)
head(v,14)
words <- names(v)
d <- data.frame(word=words, freq=v)
wordcloud(d$word,d$freq,min.freq=50, scale=c(5, .4), colors=brewer.pal(6, ""))
```

R code for Word Association and Most frequent words

```
cname <- file.path("~","desktop","toplam")
library (tm)
docs <- Corpus(DirSource(cname))
docs <- tm_map(docs,tolower)
docs <- tm_map(docs,removeNumbers)
docs <- tm_map(docs,removePunctuation)
docs <- tm_map(docs,stripWhitespace)
docs <- tm_map(docs, removeWords, c("bir", "cok"))
dtm <- DocumentTermMatrix(docs)
findFreqTerms(dtm, lowfreq = 4)
#for association
findAssocs(dtm, terms = "zeytinburnu", corlimit = 0.2)

head(d, 10)
barplot( d[1:20,]$freq, las = 2, names.arg = d[1:20,]$word,
       col ="lightblue", main ="Most frequent words",
       ylab = "Word frequencies")
```

R code for Correlation Plots

```
cname <- file.path("~","desktop","toplam")
library (tm)
library(SnowballC)
docs <- Corpus(DirSource(cname))
docs <- tm_map(docs,tolower)
#docs <- tm_map(docs, stemDocument, language = "turkish")
docs <- tm_map(docs,removeNumbers)
docs <- tm_map(docs,removePunctuation)
docs <- tm_map(docs,stripWhitespace)
# specify your stopwords as a character vector
docs <- tm_map(docs, removeWords, c("bir", "cok" , "olarak", "olup" , "icin"))
docs <- tm_map(docs, removeWords, c('i<c3><a7>'))
dtm <- DocumentTermMatrix(docs)

plot(dtm,
    terms=findFreqTerms(dtm, lowfreq=200)[1:30],
    corThreshold=0.2)
#Plotting Word Frequencies

freq <- sort(colSums(as.matrix(dtm)), decreasing=TRUE)
head(freq, 14)
wf <- data.frame(word=names(freq), freq=freq)
library(ggplot2)
af<- subset(wf, freq>500)
ggplot(af,aes(x=word,y=freq))  + geom_bar(stat="identity") +
theme(axis.text.x=element_text(angle=45, hjust=1))
```

R code for Sentiment Analysis

```
library(tm)
library(SnowballC)
library(RSentiment)
emlak <- fread(file = "~/Desktop/sentiment.csv",stringsAsFactors = T)
library(topicmodels)
library(tidytext)
library(magrittr)
library(dplyr)
library(stringr)
emlakdetay <- fread(file = "~/Desktop/sentiment.csv",select = c("V4"))
library(syuzhet)
emlakdetay1 <- as.character(emlakdetay)
my_text_values <- get_nrc_sentiment(emlakdetay1, cl=NULL, language="turkish")
my_text_values[1:10]
```

CURRICULUM VITAE

Name Surname: İmge KIZILTAN

Place and Date of Birth: İstanbul 13.12.1994

E-Mail: imgekzltn@gmail.com

EDUCATION:

- **B.Sc.:** 2016, Istanbul Technical University, Engineering, Textile Engineering

PROFESSIONAL EXPERIENCE:

- 2017-… Arçelik Pazarlama A.Ş, Sales Operations Associate