

**SMART CITY MODELING
USING BIG DATA TECHNIQUES
(BÜYÜK VERİ TEKNİKLERİ KULLANARAK
AKILLI ŞEHİR MODELLEME)**

by

Muhammet Sıddık EMEK

Thesis

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

in the

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

of

GALATASARAY UNIVERSITY

Supervisor: Assist. Prof. Dr. İsmail Burak PARLAK

June 2023

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude to my supervisor, Assistant Professor Dr. İsmail Burak Parlak, for his invaluable patience and constructive feedback. I could not have embarked on this journey without his generous provision of knowledge and expertise.

Moreover, I would like to extend my sincere thanks to my family, especially my wife, Hatice Azra Çelik Emek, for their unwavering belief in me, which kept my spirits and motivation high throughout this process.

Finally, it would be remiss of me not to acknowledge my friend Kadir Güzel, who inspired me to pursue my master's degree and supported me in its initial stages. I am grateful to him for his encouragement and guidance.

June 2023

Muhammet Sıddık EMEK

TABLE OF CONTENTS

LIST OF SYMBOLS	IV
LIST OF FIGURES	VII
LIST OF TABLES	IXII
ABSTRACT.....	XII
RÉSUMÉ	XII
ÖZET	XIV
1. INTRODUCTION	1
2. LITERATURE REVIEW	8
3. METHODOLOGY	11
3.1 Corpus	11
3.1.1 Data Gathering.....	11
3.1.2 Review of Corpus	16
3.1.3 Preprocessing	21
3.1.4 Automatic Class Prediction	24
3.2 Content-based Analysis	27
3.2.1 Word Clouds	27
3.2.2 Visualizing Co-occurrence	31
3.2.3 Visualizing Word Embeddings	34
3.2.4 Topic Modelling	38
3.3 Spatio-temporal Analysis.....	47
3.4 Sentiment Analysis	54
3.4.1 Models	55
3.4.2 Evaluation	62
4. CONCLUSION	73
REFERENCES.....	75

APPENDICES	79
A.1 Translation of Turkish words	79
A.2 Translation of Turkish phrases	84
BIOGRAPHICAL SKETCH	85



LIST OF SYMBOLS

AAC	: Automatic Annotated Corpus
API	: Application Programming Interface
AUC	: Area Under The Curve
BERT	: Bidirectional Encoder Representations from Transformers
Bi-LSTM	: Bidirectional Long Short-Term Memory
CBOW	: Continuous Bag-of-Words
CNN	: Convolutional Neural Network
CRF	: Conditional Random Field
DNN	: Deep Neural Networks
FN	: False Negative
FP	: False Positive
FPR	: False Positive Rate
HMM	: Hidden Markov Models
K-NN	: K-Nearest Neighbors
LBSN	: Location-Based Social Network
LSTM	: Long Short-Term Memory
MAC	: Manuel Annotated Corpus
ML	: Machine Learning
NB	: Naive Bayes
NLP	: Natural Language Processing
RNN	: Recurrent Neural Network
ROC	: Receiver Operating Characteristics
SGD	: Stochastic Gradient Descent
SMPs	: Social Media Platforms
SVM	: Support Vector Machines
TN	: True Negative
TP	: True Positive
TPR	: True Positive Rate

LIST OF FIGURES

Figure 3.1: Gathering the Data from Twitter and Foursquare	12
Figure 3.2: Foursquare comment published on Twitter	13
Figure 3.3: Annual heat map of tips by city	17
Figure 3.4: Annual heat map of tips by category group	18
Figure 3.5: Word Clouds of Final Dataset	19
Figure 3.6: Words with the highest frequency in Dataset	19
Figure 3.7: Preprocessing Steps	21
Figure 3.8: Steps of Automatic Class Prediction	24
Figure 3.9: List of CSV files created throughout the project	27
Figure 3.10: Word Clouds for Negative Tips	28
Figure 3.11: Words with the highest frequency in Negative Tips	28
Figure 3.12: Word Clouds for Neutral Tips	29
Figure 3.13: Words with the highest frequency in Neutral Tips	29
Figure 3.14: Word Clouds for Positive Tips	30
Figure 3.15: Words with the highest frequency in Positive Tips	30
Figure 3.16: Visualization of Words Co-occurrences for Positive Tips	31
Figure 3.17: Visualization of Words Co-occurrences for Neutral Tips	32
Figure 3.18: Visualization of Words Co-occurrences for Negative Tips	33
Figure 3.19: Visualization of Positive Tips Embeddings	35
Figure 3.20: Visualization of Neutral Tips Embeddings	36
Figure 3.21: Visualization of Negative Tips Embeddings	37
Figure 3.22: User Review Distribution in 2010	48
Figure 3.23: User Review Distribution in 2011	48
Figure 3.24: User Review Distribution in 2012	49
Figure 3.25: User Review Distribution in 2013	49
Figure 3.26: User Review Distribution in 2014	50

Figure 3.27: User Review Distribution in 2015	50
Figure 3.28: User Review Distribution in 2016	51
Figure 3.29: User Review Distribution in 2017	51
Figure 3.30: User Review Distribution in 2018	52
Figure 3.31: User Review Distribution in 2019	52
Figure 3.32: User Review Distribution in 2020	53
Figure 3.33: User Review Distribution in 2021	53
Figure 3.34: User Review Distribution in 2022	54
Figure 3.35: Optimal Hyperplane and margins for an SVM trained with two classes .	55
Figure 3.36: Multiclass classification using SVM	56
Figure 3.37: General structure of a Decision Tree	57
Figure 3.38: Random Forest working diagram	58
Figure 3.39: ROC curve for Negative Class with Machine Learning Models	66
Figure 3.40: ROC curve for Neutral Class with Machine Learning Models	67
Figure 3.41: ROC curve for Positive Class with Machine Learning Models	67
Figure 3.42: ROC curve for Negative Class with Deep Learning Models	68
Figure 3.43: ROC curve for Neutral Class with Deep Learning Models	69
Figure 3.44: ROC curve for Positive Class with Deep Learning Models	69
Figure 3.45: ROC curve for Negative Class with BERT Models	71
Figure 3.46: ROC curve for Neutral Class with BERT Models	71
Figure 3.47: ROC curve for Positive Class with BERT Models	72

LIST OF TABLES

Table 2.1: Related Works for Data and Sentiment Analysis	9
Table 2.2: Related Works for Smart Cities	10
Table 3.1: Sample tweet information fetched from Twitter by Twint API	13
Table 3.2: A sample dataset of venues information extracted from Foursquare	14
Table 3.3: A sample dataset of the tips of venue extracted from Foursquare.....	15
Table 3.4: Venue Category Groups	16
Table 3.5: Distribution of the venues according to location and category group	17
Table 3.6: Top 25 N-grams (bigram and trigram) in Dataset	20
Table 3.7: Generated Raw Dataset Statistics	20
Table 3.8: Samples of preprocessed data with and without stems	23
Table 3.9: Preprocessed Dataset Statistics	23
Table 3.10: Performance evaluation of pre-trained BERT models (Without Stems) ...	26
Table 3.11: Performance evaluation of pre-trained BERT models (With Stems)	26
Table 3.12: Wordlist for Top 6 Cities By Year	40
Table 3.13: Wordlist for Category Groups By Year	43
Table 3.14: Wordlist for Rating By Year	46
Table 3.15: Evaluation results for Machine Learning Models	66
Table 3.16: Evaluation results for Deep Learning Models	68
Table 3.17: Evaluation results for BERT Models	70

ABSTRACT

Smart city classification is a complex problem of semantic topics in city modeling. Urban area characteristics are analyzed through hierarchical tasks where location properties are gathered along with municipality records. The improvement of location-acquisition technologies has enabled individuals to incorporate a location aspect into their current online social networks in numerous ways, resulting in the development of location-based social networks (LBSN). These networks integrate the spatio-temporal attribute and real-time engagement of users. Thanks to the contribution of location dimension, LBSNs bridge the gap between the physical world and online social networking services and bring social networks closer to reality. Urban locations are graded in social networks with their popularity, quality, and customer-based interactions evaluated and shared as city insights in different social platforms. As a result, urban semantics can be combined with urban demographics and citizen sentiments in urban places. In this thesis, we focused on the smart city classification and modeling tasks in some important cities in Türkiye. First, we collected Turkish venue data from over ten categories from Foursquare and Twitter platforms. After preprocessing, we trained more than fifteen language models on our dataset and performed a comprehensive evaluation. Our findings revealed that the BERT model outperformed the other models, achieving a classification success rate of over 93%. In addition, by applying content-based and spatio-temporal analysis techniques to our dataset, we reached results that contributed to the advancement of our research. Our results show that location microblogging will be a promising benchmark in city semantics and modeling.

Keywords: Smart cities, Natural language processing, Artificial neural networks, BERT, Text classification, BERT, Location-based social networks, Content-based analysis, Spatio-temporal analysis

RÉSUMÉ

La classification des villes intelligentes est un problème complexe de sujets sémantiques dans la modélisation des villes. Les caractéristiques des zones urbaines sont analysées à l'aide de tâches hiérarchiques où les propriétés de localisation sont rassemblées avec les enregistrements de la municipalité. Le développement des technologies d'acquisition de localisation a permis aux individus d'intégrer un aspect de localisation dans leurs réseaux sociaux en ligne actuels de nombreuses façons, entraînant le développement des réseaux sociaux basés sur la localisation (RSSL). Ces réseaux intègrent l'attribut spatio-temporel et l'engagement en temps réel des utilisateurs. Grâce à la contribution de la dimension de localisation, les RSSL comblent le fossé entre le monde physique et les services de réseaux sociaux en ligne et rapprochent les réseaux sociaux de la réalité. Les emplacements urbains sont classés dans les réseaux sociaux avec leur popularité, leur qualité et leurs interactions basées sur les clients évaluées et partagées sous forme d'informations sur la ville dans différentes plateformes sociales. En conséquence, la sémantique urbaine peut être combinée avec la démographie urbaine et les sentiments des citoyens dans les lieux urbains. Dans cette thèse, nous nous sommes concentrés sur la tâche de classification des villes intelligentes dans certaines villes importantes de Turquie. Tout d'abord, nous avons collecté des données sur les sites turcs de plus de dix catégories à partir des plateformes Foursquare et Twitter. Après le prétraitement, nous avons formé plus de quinze modèles de langage sur notre ensemble de données et effectué une évaluation complète. Nos résultats ont révélé que le modèle BERT surpassait les autres modèles, atteignant un taux de réussite de la classification de plus de 93 %. De plus, en appliquant des techniques d'analyse basées sur le contenu et spatio-temporelles à notre ensemble de données, nous avons atteint des résultats qui ont contribué à l'avancement de nos recherches. Nos résultats montrent que le microblogging de localisation sera une référence prometteuse en sémantique et modélisation de la ville.

ÖZET

Akıllı şehir sınıflandırması, şehir modellemedeki anlamsal konuların karmaşık bir problemidir. Kentsel alan özellikleri, konum özelliklerinin belediye kayıtları ile birlikte toplandığı hiyerarşik görevler aracılığıyla analiz edilmektedir. Konum edinme teknolojilerinin gelişimi, bireylerin mevcut çevrimiçi sosyal ağlara çeşitli yollarla konum özelliğini dahil etmelerini sağlayarak, konum tabanlı sosyal ağların geliştirilmesine yol açmıştır. Bu ağlar, mekansal-zamansal özelliği ve kullanıcıların gerçek zamanlı katılımını bütünleştirir. Lokasyon boyutunun katkısı sayesinde, konum tabanlı sosyal ağlar fiziksel dünya ile çevrimiçi sosyal ağ hizmetleri arasındaki boşluğu doldurarak sosyal ağları gerçeğe yaklaştırır. Kentsel lokasyonlar, popülerlikleri, kaliteleri ve müşteri bazlı etkileşimleri ile sosyal ağlarda derecelendirilir, farklı sosyal platformlarda şehir içgörülerini olarak değerlendirilir ve paylaşılır. Sonuç olarak, kentsel semantik, kentsel demografik bilgiler ve kentsel yerlerdeki vatandaş duyguları ile birleştirilebilir. Bu tezde, Türkiye'deki bazı önemli şehirlerde akıllı şehir sınıflandırma ve modelleme çalışmalarına odaklandık. İlk olarak, Foursquare ve Twitter platformlarından ondan fazla kategoride Türkçe mekan verilerini topladık. Ön işlemeden sonra, veri setimiz üzerinde on beşten fazla dil modelini eğittik ve kapsamlı bir değerlendirme gerçekleştirdik. Bulgularımız, BERT modelinin diğer modellerden daha iyi performans gösterdiğini ve %93'ün üzerinde bir sınıflandırma başarı oranı elde ettiğini ortaya koydu. Ayrıca veri setimize içerik tabanlı ve mekansal-zamansal analiz teknikleri uygulayarak araştırmamızın ilerlemesine katkı sağlayan sonuçlara ulaştık. Sonuçlarımız, konum mikroblogunun şehir semantiği ve modellemesinde umut verici bir ölçüt olacağını gösteriyor.

1. INTRODUCTION

The swift advancements in Internet and mobile technologies have brought about a situation where the social activities occurring within a city are mirrored in virtual environments, and these traces reveal the relational ties of the city's dynamic structure. The data generated by mobile devices that are an integral part of users' daily lives offer significant insights into their behavioral trends in the city. This type of data referred to as "Big Data" comprises vast amounts of information with a finely-detailed resolution, enabling individuals to form logical conclusions about the formation of activity patterns within the city through a bottom-up approach.

Social media refers to the channels of communication between individuals through which they produce and exchange information and ideas in virtual communities and networks. They are increasingly being utilized to address the limitations of surveys and other sources of travel data, as they offer a more cohesive, less intrusive, and easily accessible approach to understand the intricacies of travel patterns. Furthermore, social media data incorporates the spatio-temporal attribute and real-time engagement of users. Social media serves as a powerful medium for individuals to exchange their sentiments in the form of opinions and perspectives on any given topic or location, resulting in a vast amount of unstructured information.

In recent times, there has been a significant increase in people's participation on Social Media Platforms (SMPs). Currently, over 4.5 billion individuals use the Internet, with social media users surpassing the 3.8 billion milestone. Nearly 60 percent of the world's population is connected to the Internet, and recent trends indicate that more than half of the world's total population will use social media by mid-year.¹

¹ <https://wearesocial.com/uk/blog/2020/01/digital-2020-3-8-billion-people-use-social-media/>

The enhanced accessibility of location acquisition technology, such as GPS and Wi-Fi, allows individuals to integrate a location aspect into their current online social networks in various ways. For instance, users can upload location-tagged photographs to a social networking platform like Flickr, provide commentary on an event using Twitter while being present at the event location, and share their current whereabouts on Foursquare to arrange a real-world group gathering.

Incorporating the location dimension in online social networking services connects the physical world with these virtual platforms, thus integrating social networks into reality. For instance, users can use their mobile phones to leave tips about a restaurant on a social networking site, allowing individuals within their social circle to reference it later when they visit the same restaurant. In this scenario, users generate their narratives regarding physical locations while browsing other individuals' experiences. Consequently, online social networking sites function as a platform that promotes the exchange of personal encounters.

Furthermore, individuals within an existing social network can broaden their social circle by forming new connections based on location-related interdependence. As location is a crucial aspect of user context, it provides detailed insights into an individual's interests and behaviors. For instance, people who share a liking for a particular restaurant can form a connection. In addition, individuals who frequently travel to the same mountain can communicate with each other to exchange their travel experiences.

A location-based social network (LBSN) not only involves adding a location dimension to an existing social network but also creates a new social structure based on the interdependence of individuals derived from their physical location in the world, as well as location-tagged media content such as photos, videos, and text. The physical location is determined by the individual's instantaneous location at a particular time and location history over a certain period. Interdependence refers to the coexistence of two individuals in the same physical location or sharing similar location histories and information such as common interests, behaviors, and activities derived from the individual's location and location-tagged data. LBSN data with geolocation information

is a significant component of smart city analysis. However, traditional social network analysis methods often fail to incorporate location information when generating recommendation services.

Using LBSN data can provide insights into the spatial distribution of tourist attractions and the travel patterns of tourists. It can also quantify the relationships between tourist facilities such as public transport, accommodations, and restaurants, leading to a better understanding of the dynamics of tourist activities. For example, LBSN data can be used to identify the popular routes and destinations of tourists, which can help in the planning of tourism infrastructure and allocating resources. Additionally, LBSN data can be used to analyze the tourists' behaviors, preference development of personalized tourism services, and targeted marketing strategies.

Foursquare, a prominent Location-Based Social Network (LBSN), allows its users to locate and explore desirable places based on the opinions, check-ins, ratings, and reviews provided by others. The platform also provides diverse features like place search and recommendations based on other users' experiences. Founded by Dennis Crowley and Naveen Selvadurai in 2009 in New York City, Foursquare has experienced substantial growth in its user base, check-ins, and venue information. In 2019, Foursquare recorded over 14 billion check-ins across more than 105 million places in over 190 countries. The platform boasts over 60 million registered users and over 50 million monthly active users (Metev, 2023). Initially, Foursquare began as a check-in app that enabled users to share their whereabouts and activities. However, with the emergence of similar features in other social networks like Facebook, Foursquare had to expand its offerings and create unique functionalities.

Three fundamental topics are necessary to explore Foursquare social network. Firstly, a "check-in" happens when a user visits a physical location and records it on Foursquare. The term "venue" refers to a physical place, such as a business or residence, where users can check-in and share their location with others. "Tips" enable users to share recommendations, warnings, and other information to help their friends have a better experience at a venue. Users receive points and badges for checking-in at various locations, and the person with the most check-ins at a venue becomes the "Mayor".

With real-time location information, users can discover nearby friends and engage in social activities such as dining out or shopping. Furthermore, users can add "tips" to venues, which are recommendations for other users regarding things to do, see, or eat at the venue. In this service, a venue is a primary factor determining the interdependence that links users. Additionally, user-generated content such as tips and badges are associated with a location.

On the other hand, Foursquare uses machine learning techniques to understand the user's preferences and provide personalized recommendations for the best places to visit in any city. Additionally, the company has modified its business model by introducing "Foursquare Pinpoint" a service that collects user data for commercial use and helps businesses gain insights into consumer behavior. Foursquare has also released a companion app called "Swarm" which enables users to compete with their friends to organize larger social gatherings.

There are several alternatives to access data from Foursquare. First, the Foursquare Places API is a popular option that enables developers to create location-based applications and services by accessing the information on users, venues, check-ins, photos, and tips. However, there are fees associated with using this API, and the free version has limitations in terms of data retrieval. Web scraping offers a viable alternative for obtaining desired data from Foursquare as it has no technical limitations, unlike APIs. It involves automatically retrieving the content of user-facing web pages, users to manually extract data by clicking through web pages and copy-pasting interesting parts into a spreadsheet, web scraping leverages computer programs to perform this task much more quickly and accurately. In this thesis, a combination of BeautifulSoup and Selenium for dynamic web scraping has been successfully implemented. Selenium was used to automate web browser interactions, while BeautifulSoup was utilized to extract pertinent information.

Selenium is a versatile tool initially developed for testing website behavior, but it has since become a widely used web browser automation tool for various tasks, including web scraping. The tool is highly popular and can automate various browsers through a middleware control known as the Selenium web driver. This protocol service acts as an

intermediary between the client and the browser, translating client commands to web browser actions. It also provides extensions for emulating user interaction with browsers, a distribution server for scaling browser allocation, and infrastructure for implementing the W3C WebDriver specification. This specification enables developers to write interchangeable code for all major web browsers.

Beautiful Soup is a popular Python library used for web scraping and parsing HTML and XML pages. With Beautiful Soup, it is possible to extract relevant information from a webpage, navigate through its structure, and modify its contents. One of the key benefits of using Beautiful Soup is that it can handle messy or poorly formatted HTML, cleaning up the data and presenting it in an easily-traversable XML structure. This saves time and effort that would otherwise be spent manually cleaning up the data.

Through the utilization of various analytical techniques, including content-based analysis, spatio-temporal analysis, and sentiment analysis on venue and user tips data collected from Foursquare using web scraping, valuable outcomes for the Smart Cities analysis were yielded.

Content analysis is a qualitative research technique commonly used to analyze various content features. With content analysis, researchers can identify, quantify, and analyze the presence, meaning, and relationships of particular words, themes, or concepts. This approach allows for the quantification of qualitative information by sorting data and comparing different pieces of information to summarize it into useful insights. In this thesis, content analysis techniques to perform various tasks, including generating word clouds based on user tips, identifying co-occurring and embedded words, and extracting topics based on spatio-temporal data were utilized.

Spatio-temporal analysis is becoming increasingly important in various fields as it enables researchers to identify patterns and trends in data over time and space. With the increasing availability of big data and advanced technologies, spatio-temporal analysis has become more feasible and efficient. Machine learning algorithms and data mining techniques are being used to identify patterns and relationships in large spatio-temporal datasets, and data visualization tools are being developed to help researchers

communicate their findings in a more accessible way. Overall, spatio-temporal analysis plays a crucial role in understanding complex systems in a dynamic world.

Sentiment analysis is a field of ongoing research that employs natural language processing (NLP), text analysis, and computational techniques to extract, quantify, and examine the emotional states present within a given piece of information or text dataset. As users generate an increasing amount of data through social media platforms, identifying and categorizing user-generated texts has become an important area of research and has led to the creation of several sub-fields that serve numerous purposes. Many corporations use sentiment analysis to evaluate products and tips on social media. In this thesis, we have applied sentiment analysis to user tips to comprehend users' emotions and recommendations and determine whether the text expresses positivity, negativity, or neutrality. Presently, the sentiment analysis problem can be addressed using three main approaches: lexicon-based techniques, machine-learning techniques, and hybrid techniques that combine both lexicon-based and machine learning-based approaches.

Lexicon-based techniques are the oldest and most straightforward approaches to sentiment analysis. They use pre-defined sentiment lexicons, which contain lists of words and their corresponding sentiment scores. The lexicon assigns a sentiment score to each word in the text, and the final sentiment score of the text is calculated by aggregating the scores of all the words in the lexicon. These techniques were initially designed for English texts but have since been extended to other languages.

The second approach for sentiment analysis is machine learning, which consists of two areas: traditional and deep learning models. Traditional models involve classical machine learning techniques such as Logistic Regression, Naive Bayes. These algorithms take lexical features, sentiment lexicon-based features, parts of speech, adjectives, and adverbs as inputs. The accuracy of these systems depends on the choice of features. In addition to traditional models, various deep learning models such as LSTM, RNN, and BERT can also be used for sentiment analysis. Deep learning models have recently shown the best results in terms of accuracy. However, these models may be less interpretable compared to lexicon-based methods, despite their high accuracy.

There are many studies on sentiment analysis in English, especially in news, movie and product reviews, customer feedback, and complaint management, but studies in Turkish are limited. Moreover, spatio-temporal text based studies are required on LBSN data to enable better classification tools in Turkish, especially for Smart city classification in Türkiye.

Firstly, based on the need for the Turkish dataset, we focused on constituting a valuable medium-sized Turkish dataset that contains 7083 unique venue records with many significant attributes like name, rating, address, location, categories, and almost 500K user reviews regarding the venue from Foursquare. Thus, this dataset will be made available to enable future studies in this area.

After preprocessing the obtained data, we applied the fine-tuned BERT model, a new approach, to our Turkish dataset for the sentiment classification problem on citizen feelings in urban locations. On the other hand, several classical machine learning models (Naive Bayes, Support Vector Machines, Logistic Regression, etc.) and deep learning (LSTM, Bi-LSTM) models have been trained and evaluated too.

We collected results from all classifiers and evaluated them using accuracy, precision, recall, F1 score, and ROC-AUC metrics. Then, we presented and compared the scores obtained from the Turkish corpus. Finally, the evaluation was concluded by identifying the best scores through a comparative analysis and proposing future steps.

In addition, different statistical, temporal, and spatial analysis techniques have been applied to visualize the image of cities in Türkiye, which is a promising way to better understand city insights and dynamics.

2. LITERATURE REVIEW

Most research conducted in this particular field has primarily focused on widespread languages like English, and there has been a little emphasis placed on languages with complex morphologies. The direct application of rules and algorithms used in English to languages like Turkish is impractical due to formal dissimilarities. The main focus of this research is the Turkish language, and we have surveyed recent studies carried out in this language. Furthermore, the methodology employed in this study has been examined in studies conducted in languages other than Turkish to expand its applicability.

Although most sentiment analysis research has concentrated on binary sentiment analysis, ongoing research efforts are directed toward multi-dimensional sentiment analysis. With the advent of BERT, many researchers have incorporated BERT-based techniques in sentiment analysis tasks, and the results have shown that BERT has substantially improved the performance of sentiment analysis models. However, recent studies have mostly utilized data from location-based social networks such as Foursquare, Twitter, Weibo, and others to identify various patterns, including human mobility, activity, urban planning, and venue classification.

Table 2.1 and Table 2.2 summarize relevant studies conducted on various analyses on social media big data. Table 2.1 focuses on research related to content-based and sentiment analyses, while Table 2.2 focuses on spatio-temporal and geo-mapping analysis in the context of smart cities. The Scope and Aim column describes the motivation behind each study, while the Methods column outlines the proposed techniques used to address the challenges associated with achieving their goals. The dataset column provides information on where the researchers obtained their data, the size of the dataset, and the language used. Finally, the Results column provides details on the outcomes of each study.

Table 2.1: Related Works for Data and Sentiment Analysis

Researchers	Dataset			Scope And Aim	Methods	Results
	Source	Language	Size			
Chen et al. (2016)	Foursquare	English	6.52 million of users and their all published tips with 2-classes	Measurement and analysis of Foursquare users and their published tips	Data Analysis, Sentiment Analysis	They found the concept of "happiness index" by analysis on groups of users who post positive tips. Also they presented the first comprehensive and unbiased picture of tips in Foursquare
Altunayyan et al. (2018)	Foursquare	Arabic	Approximately 1,000 restaurants info and 12,000 tips with 2-classes	Analyzing user behaviors and patterns by check-ins, comments, and ratings in Foursquare	Sentiment Analysis, Community Detection, LDA	Distinct communities were identified by categorizing venues on Foursquare according to their food category. Venues were assessed with respect to their service, price, rating, and category.
Av et al. (2018)	beyazperde.com	Turkish	156,816 movie reviews with 5-classes from beyazperde.com	Evaluating deep learning models for binary sentiment classification	Binary Sentiment Classification, Word2Vec, CNN, LSTM, BiLSTM	It was proposed that CNN-LSTM models exhibited superior performance when compared to other models.
Yüksel and Tan (2018)	Foursquare	Turkish English	A total of 7086 Turkish and 6943 English Foursquare tips which are manually labelled	Developing a real-time intelligent multilingual decision support system for social networks	Text mining, Sentiment Analysis, NLP, Naive Bayes, RT-SKDA	An accuracy rate of 84.49% was attained for the Turkish language, while a rate of 95% was achieved for the English language.
Yu et al. (2019)	Seven public widely-studied datasets used in many related works	English	A total of approximately 1 million different content across seven general datasets	Improving BERT-Based text classification with BERT4ITC a post-training approach for auxiliary sentence and domain knowledge	Sentiment Analysis, Question classification, Topic classification, BERT, Fine-Tuning	The BERT4ITC post-training strategy was found to produce superior results when compared to the original BERT model.
Azhar and Khodra (2020)	Airrooms	Indonesian	In total, 292,931 hotel reviews with 2-classes in two different datasets	Fine-tuning pretrained multilingual BERT model for Indonesian Aspect-based sentiment analysis (ABSA) and improving the performance	Aspect-based Sentiment Analysis, BERT, Fine-Tuning, Sentence-pair Classification	They were able to achieve a significant improvement of 8% in the F1-score when compared to the results obtained in their prior study.
Acikalin et al. (2020)	beyazperde.com, otelpuan.com	Turkish	53,400 movie reviews with 2-classes from beyazperde.com 11,600 hotel reviews with 2-classes from otelpuan.com	Turkish Sentiment Analysis Using BERT	Sentiment Analysis, BERT, fastText	An accuracy of 93.32% for 2-class sentiment analysis was obtained
Tang et al. (2020)	NLPCC 2018 [11]	Multilingual (Chinese, English, Japanese...)	6,738 multilingual code-switching texts with 5-classes (five basic emotions)	Fine-tuning BERT model for multi-label sentiment analysis in unbalanced code-switching text	Text mining, Sentiment Analysis, BERT, Fine-Tuning	The BERT-Base Chinese model was found to be better suited for code-switching texts present in Chinese microblogs.
Aksu and Karaman (2021)	Tripadvisor, Google Maps and Foursquare	Turkish	Totally 37,464 manual labeled reviews with 3-classes	Sentiment analysis regarding touristic sites using machine learning approaches Determining the effects of labeling, stemming, and negation on the success of sentiment analysis	Text mining, Sentiment Analysis, Machine Learning, TF-IDF	The F1-score achieved for 2-class sentiment analysis was 0.95, whereas for 3-class sentiment analysis, it was 0.80.
Başarlan and Kayaalp (2021)	Yelp, IMDB, Twitter	English	598,000 restaurants reviews with 2-classes from Yelp 50,000 movie reviews with 2-classes from IMDB 4,500 health-related tweets with 3-classes from Twitter	Determining the performance effects of different word embedding methods on Sentiment Analysis	TF-IDF, Word2Vec, GloVe, BERT, CNN, RNN, LSTM	Consequently, the BERT word embedding technique demonstrated the most outstanding performance when compared to the other word embedding methods.
Karayigit et al. (2022)	Instagram	Turkish	Totally 9,708 manual labeled reviews with 3-classes	Sentiment Analysis on Instagram comments regarding COVID-19 by using BERT-based Transfer Learning	Sentiment Analysis, Machine Learning, BERT, CNN, LSTM, GCR-NN, LDA	A macro-averaged F1-score of 0.7864 was attained

Table 2.2: Related Works for Smart Cities

Researchers	Dataset			Scope And Aim	Methods	Results
	Source	Language	Size			
Silva et al. (2012)	Foursquare	English	4,672,841 check-ins, and 1,929,237 venues	Visualizing the invisible image of Cities which is promising way to better understand the city dynamics. Human behaviors and Smart Cities	Data Analysis, Spatio-temporal Analysis, Network Coverage, Seasonality	Their study utilized transition graphs to map the movements of individuals within a Personal Social Network (PSN) across eight different cities. Based on these graphs, they were able to summarize the dynamics of each city.
Assem et al. (2016)	Foursquare	English	227,428 check-ins in New York City for about 10 months	Identifying functional regions in Cities taking into account temporal variations of geographic user activity	Spatio-Temporal Analysis, Clustering	Their study involved the analysis and visualization of the temporal functional regions that were generated for a 4-hour time interval in New York City.
Gallegos et al. (2016)	Twitter, Foursquare	English	6 million Tweets and 5,863 Foursquare check-ins	Investigating the characteristics of areas in a city that people happier when they visit. Human mobility and Smart Cities	Data Analysis, Sentiment Analysis, Spatio-temporal Analysis, Geo-mapping, SentiStrength	Their findings revealed that venues with check-ins in Los Angeles are inherently distinct from other areas. They observed that locations featuring a diverse range of activities and establishments, such as restaurants, gyms, and beaches, tend to attract a larger number of visitors.
AlQadi et al. (2020)	Google Maps	Arabic	26,000 restaurant reviews with 3-classes	Demonstrating Snapchat influencers how affect restaurant selection of the customers.	Data Analysis, Sentiment Analysis, Statistical Analysis	Their study indicated that over 50% of survey respondents visit restaurants based on recommendations from Snapchat celebrities.
Khan et al. (2020)	Weibo	Chinese	222,525 check-ins at 722 different venues for the time period of six months	Spatio-temporal analysis of tourists and residents in Shanghai Based on Location-Based Social Network's data. Smart Cities, Tourism	Spatio-temporal Analysis, Data Analysis, Geo-mapping, Classification	The study's findings indicate that tourists engage in more frequent spatial activities, particularly in the city center, when compared to the residents. The study also revealed that residents tend to visit suburban areas. Moreover, the temporal activities of tourists exhibit more significant variations, while those of residents reflect a relatively consistent behavior.
Khan et al. (2020)	Weibo	Chinese	441,471 check-ins by 144,582 users from 20,171 venues	The effect of various types of venues on the density distribution of residents for the city of Shanghai, China. Human mobility, Smart Cities, Recommendation systems	Spatio-temporal Analysis, Density Estimation, Geo-mapping, Classification	The density estimation results revealed that the highest concentration of check-ins occurs in the city center, as anticipated. This outcome can be attributed to the more abundant availability of resources in this area.
De et al. (2022)	Foursquare	English	138,140 check-ins from 20,000 users across 10,853 London venues	Deriving lifestyle patterns from Foursquare to extract hidden spatio-temporal behavior patterns. Human behaviour and interactions	Spatio-Temporal Analysis, Matrix Decomposition, Tensor Decomposition, Geo-mapping	Their study presented various aspects related to spatio-temporal and composite lifestyles.

3. METHODOLOGY

In this section, the smart city techniques will be discussed through classical machine learning and deep learning procedures, applied onto the Turkish dataset. Data collection, automatic labeling, data preprocessing, selected classifiers for sentiment analysis, training, classification, and evaluation processes are explained in detail.

3.1 Corpus

This chapter expounds on the procedural aspects of corpus development and preprocessing techniques employed in this thesis. The initial step comprises the collection of raw data from Twitter and Foursquare. To facilitate this study's analysis, modeling, and classification endeavors, we necessitated an annotated dataset, which we constructed manually and automatically. The preprocessing stage holds paramount significance for topic modeling and classical machine learning approaches, and its thorough execution provides valuable findings by the study's conclusion.

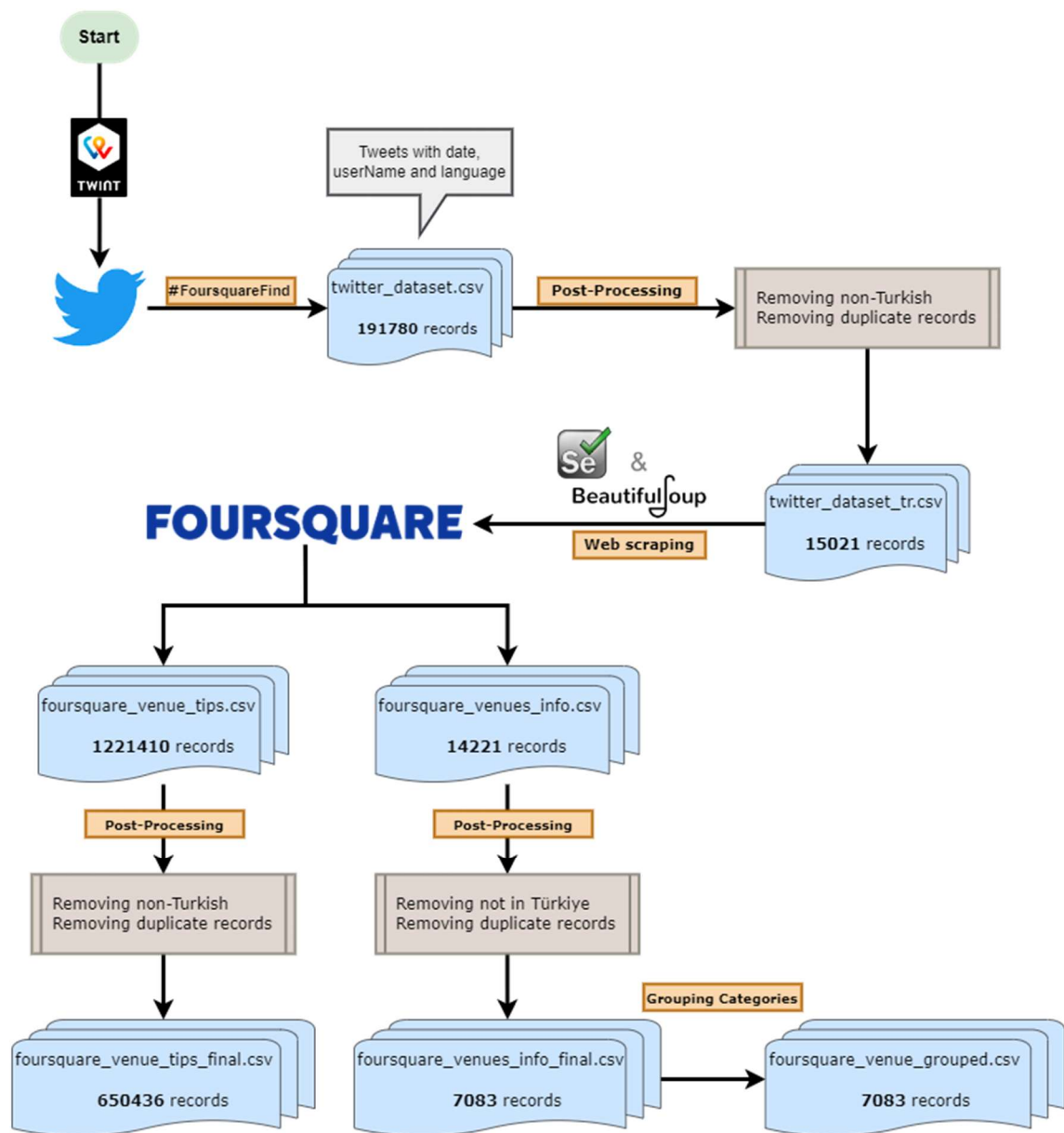
3.1.1 Data Gathering

Figure 3.1 shows the flowchart of data collection, post-processing, and the resulting final dataset files. Each step is explained in detail below.

First, Foursquare tips were pulled out on Twitter to obtain venue homepage URL information on foursquare. Twint API was preferred instead of Twitter API to accomplish this task.

Twint is a Python-based, sophisticated tool for scraping Twitter that enables the retrieval of Tweets from Twitter without using its API. Twint harnesses Twitter's search operators to extract Tweets associated with specific subjects, hashtags, trends, and users without requiring any authentication, API, or browser emulation. In addition, the Twitter API is confined to providing only the most recent 3200 tweets, while Twint is unrestricted by any such tweet download limitation.

Figure 3.1: Gathering the Data from Twitter and Foursquare



It is stated that Twint can only filter Tweets in specific languages with the Lang parameter. However, it was seen that this parameter did not work correctly and brought tweets in different languages. Therefore, after fetching tweets containing the #FoursquareFind hashtag, post-processing has been applied to obtain Turkish tweets. Figure 3.2 shows a sample Tweet containing the #FoursquareFind hashtag.

Figure 3.2: Foursquare comment published on Twitter



We retrieved 191780 tweet content from 2016 to the present, containing the hashtag #FoursquareFind. Among the obtained data, there are essential features of Tweets such as id, date, language, and user information. Afterward, we deleted tweets in languages other than Turkish, so we obtained 15021 original recordings for this thesis. Table 3.1 shows sample data obtained from Twitter containing the #FoursquareFind hashtag.

Table 3.1: Sample tweet information fetched from Twitter by Twint API

Date	Lang.	Username	Tweet
2022-01-21	tr	zlem**	hersey cok güzel 🍷🍷🍷"" #FoursquareFind @ Kukis https://t.co/JOABEfMvKC
2022-01-19	tr	zlem**	şefin yemekleri harika 😊 #FoursquareFind @ Beta Lake House https://t.co/EWDcPObsYr
2021-06-29	tr	blpn*	Muhteşem Koyları #FoursquareFind @ Demlik Apart Camping https://t.co/dEjGpQJC2V
2016-02-13	tr	rmnr**	İdeal 🍷🍷 #FoursquareFind @ Kahve Sokağı https://t.co/wOhyRPINjW

There are different alternatives to access Foursquare data, as such in Twitter. In the realm of location-based services, the Foursquare Places API offers a wealth of information pertaining to venues, users, photos, check-ins, and tips, thereby enabling enhanced location-based experiences. All this information can be obtained through different services with special usage fees. Although there is a free version of this API, it has limitations in obtaining data. The web scraping method, offers an efficient alternative for extracting desired data. While in Foursquare API, there are limitations, web scraping does not have any, at least in the technical sense.

In this thesis, the combined usage of Beautiful Soup and Selenium facilitated dynamic web scraping to a considerable extent. Utilizing the URL data of tips procured from Twitter in the preceding section, the web scraping technique was employed to obtain venue-related information from the Foursquare website. Initially, Selenium initiated the Chrome browser with the pertinent Foursquare venue page URL. Next, the HTML content of the page was retrieved through the library. Finally, the extracted HTML content was parsed with the aid of the Beautiful Soup library, allowing for the extraction of the requisite information.

We extracted 14221 unique venue information from Foursquare with significant attributes, including venueId, venue name, venue actual URL, location, category list, tip count, rating, geolocation, and address. On the other hand, we assigned a unique identifier to each venue to associate user tips. Table 3.2 shows a sample dataset containing venue information that was extracted from Foursquare.

Table 3.2: A sample dataset of venues information extracted from Foursquare

Venue Name	Location	Category	Tips Count	Rating	Coordinates
Cumba Künefe	Gaziantep	Tatlıcı	542	8.7	37.07412,37.34942
Mikla	Beyoğlu	Restoran	280	8.9	41.03046,28.97420
Otel Mavi Beyaz	Datça	Otel	291	7.1	36.67656,27.51145
Coffee Time	Erzurum	Kafe	79	7.8	39.96641,41.25562
Beylik Koyu	Bozcaada	Plaj	67	8.5	39.81592,25.99021

Then we deduplicated the venue information according to the venue homepage URL, which is unique on Foursquare. Thus, multiple tips from the same venue were prevented. In addition, we removed some venues because they are not in Türkiye. As a result, we obtained 7083 venue information that is unique and located in Türkiye.

The same method used for gathering venue information was also employed for fetching users' tips. The critical point in this process is that the tips on the web page are displayed with pagination logic. Therefore, each comment page must be individually called so all tips can be pulled out. On the other hand, the actual URL of the venue on Foursquare was needed to access the comment pages, and thankfully this information was obtained with the previous operation. The tips file was filled with the publish date and content of the tip extracted from Foursquare, also the identifier id of the venue located in the first file (Venue Information), which is used to make the relation between the venue and its tips.

As a result, over 1 million user tips about the venues have been obtained. However, all of these tips were not only in Turkish but also in different languages and found redundant. Therefore, post-processing is vital, and tips other than Turkish must be removed before Turkish sentiment analysis. Table 3.3 shows a sample dataset of the tips of venue extracted from Foursquare.

Table 3.3: A sample dataset of the tips of venue extracted from Foursquare

Venue ID	Publish Date	Comment
bf7e012f23d3436d9636	2022-04-18	Muhteşem bir müze
42f0-9e4a8cfa5a7bdc5a	2015-04-09	Doğayla iç içe görülmeye değer
60a14532e9bef98fe95d	2015-01-03	Ortam sıcak, ilgi desen sonsuz 😊👍
b60aa4dse9bef98fe95d	2014-05-11	Servisler rezalet ötesi
959b24225603fee2b1a1	2018-02-17	Very tasty premium burger.
6dd86475abc3fa4v23sda	2017-04-14	Herşey süper yok yok 😊👍
435197dba6e8cf2a9ef31	2016-07-07	Izgara sucuk 😊 Hamburger

On the other hand, when the category information in the venue information file was examined, more than 1000 different categories were encountered. Therefore, we need to group these categories to make an accurate analysis. According to some determined rules, this grouping was carried out with the application, and as a result, 15 different category groups were obtained. Table 3.4 shows sample category groups and related categories.

Table 3.4: Venue Category Groups

Category List	Category Group
Kafeterya, Kafe, Kahvaltı Noktası, Pastane, Nargile Kafe	Kafe
Restoran, Türk Restoranı, Kebapçı, Köfteci, Steakhouse	Restoran
Otel, Pansiyon, Kamp Alanı, Misafirhane, Hostel, Tatil Köyü	Konaklama Yeri
Festival, Tiyatro, Sinema, Fuar, Müze, Sanat Galerisi, Klüp	Kültür Sanat
Havalimanı, Liman, Gemi, Tren, Lounge, İskele, Vapur, Terminal	Ulaşım
Plaj, Koy, Park, Kamp, Şelale, Mağara, Göl, Bahçe, Orman, Nehir	Doğa

3.1.2 Review of Corpus

The final dataset prepared for this thesis consists of 2 CSV files : the Venue Information (foursquare_venues_info.csv) and the Tips of the users about the relevant venue (foursquare_venue_tips.csv). All this information was gathered from the Foursquare web page by using the method clarified before.

The venue information file has 7083 unique records with many significant attributes like name, rating, address, location, and categories. All of the venues are located in Türkiye, most of them in Istanbul city (2043/7083). On the other hand, more than 1000 different categories give information about the type of place, which are divided into 15 groups. Table 3.5 shows distribution of the venues according to the city and category group.

Table 3.5: Distribution of the venues according to location and category group

City Name	Frequency	Category Group Name	Frequency
İstanbul	2043	Restoran	2399
İzmir	859	Kafe	1650
Ankara	671	İşyeri	924
Muğla	313	Konaklama Yeri	350
Antalya	297	Doğa	274
Bursa	219	Açık Alan	266
Balıkesir	118	Alkollü Mekan	262
Kocaeli	116	Kültür Sanat Eğlence	200
Adana	107	Ulaşım	165
Mersin	96	Diğer	162
Gaziantep	89	Konut	146
Aydın	89	Yerleşim Yeri	104
Çanakkale	87	Spor Tesis	98
Manisa	69	Eğitim Kurumu	61
Konya	69	İbadethane	22

Figure 3.3: Annual heat map of tips by city

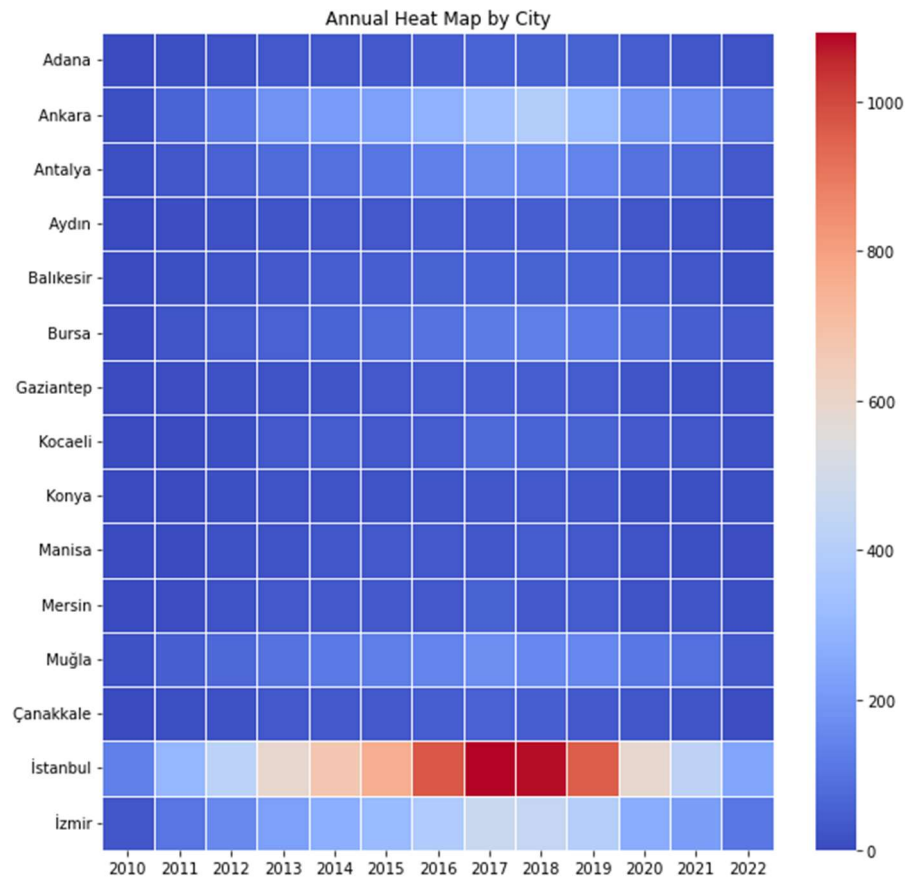
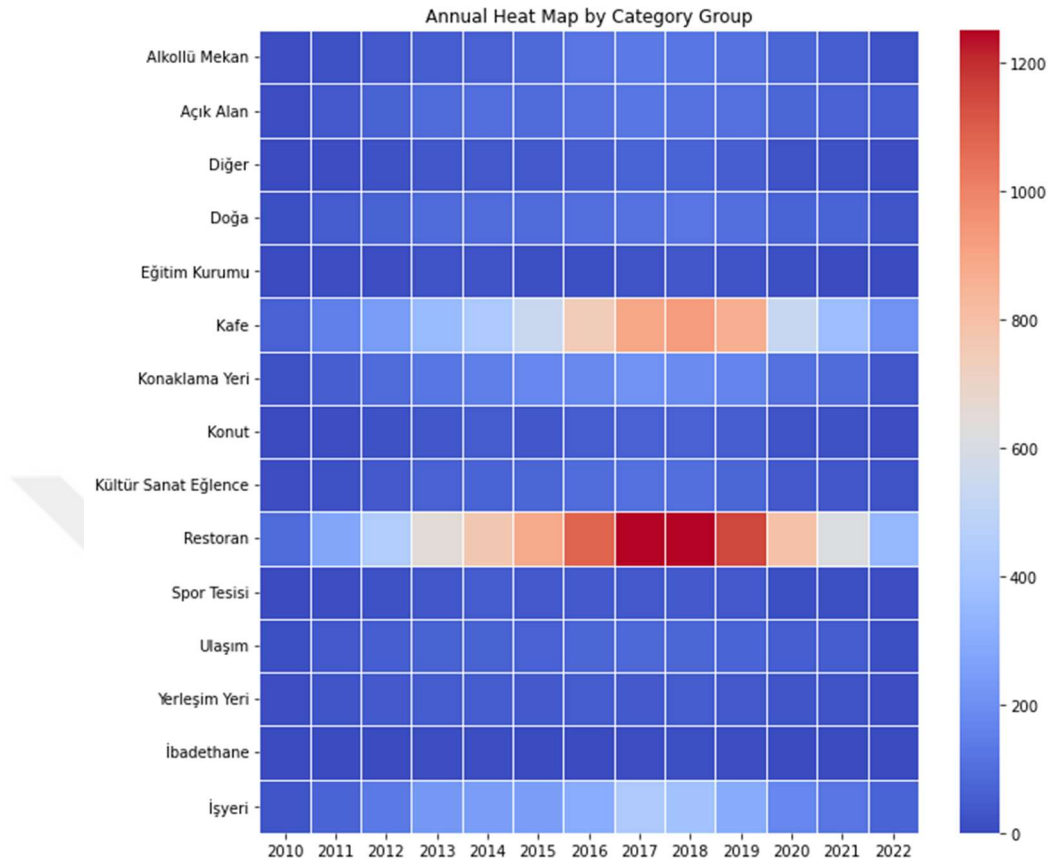


Figure 3.4: Annual heat map of tips by category group



The tips file includes the users's tips about the venue and the date of publication of the tip. Using the publish date information, the obtained annual heat map of the tips based on city and category group are illustrated in Figure 3.3 and Figure 3.4. According to the data, 2016 - 2019 stands out as the most intensive use years.

Word clouds render the text more legible and comprehensible and present a robust means of visualizing data by displaying graphical depictions of word frequency, with more prominence given to words that appear more frequently in a source text. Figure 3.5 shows the word clouds of the dataset and Figure 3.6 shows the frequency of the words in the dataset. They will help design the preprocessing operation steps and determine the stop-word list. For example, the words [Çok, çok, cok] are the same but can be treated as different, so preprocessing will make them unique. Also, the words [ve, bir, en, de, da, bi, bu] have high frequency and are meaningless, so that they will be removed as stop-words from the corpus.

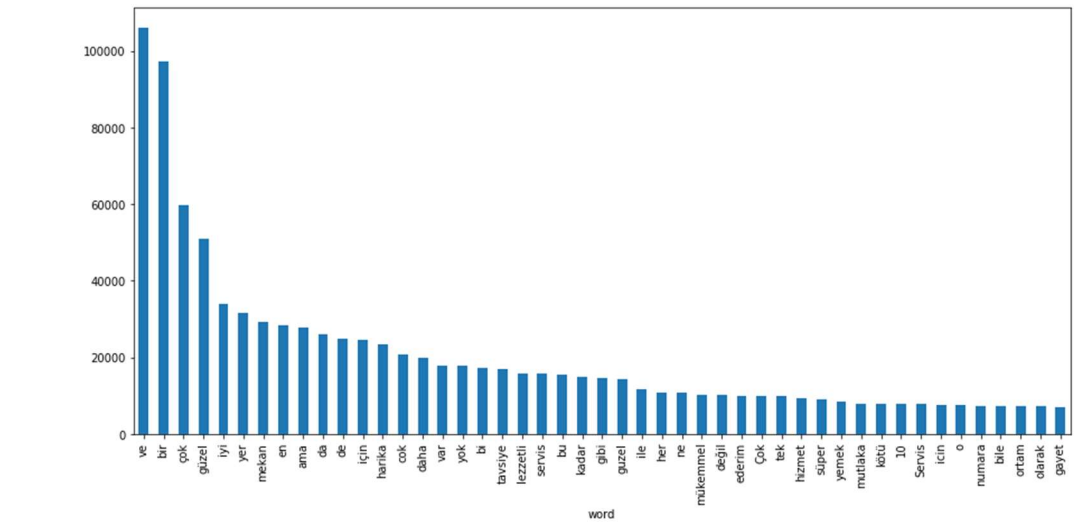


Table 3.6: Top 25 N-grams (bigram and trigram) in Dataset

Bigram	Frequency	Trigram	Frequency
bir yer	9645	güzel bir yer	801
bir mekan	6901	güzel bir mekan	753
çok güzel	6410	Çok güzel bir	749
en iyi	5832	çok güzel bir	729
güzel bir	5491	10 numara 5	649
tavsiye ederim	4165	daha ne olsun	638
en güzel	3956	Daha fazlasını oku	561
çok iyi	3906	Harika bir yer	509
bir yer	3351	gereken bir yer	476
bu kadar	3171	harika bir yer	467
bir mekan	2909	Tek kelime ile	452
10 numara	2434	bir o kadar	445
tavsiye ederim	2328	ve güler yüzlü	415
güler yüzlü	2322	için güzel bir	406
daha iyi	2175	güler yüzlü ve	404
bi yer	2164	çok güzel ama	373
harika bir	2156	bir yer :)	364
güzel ama	2121	ve güzel bir	348
o kadar	2087	numara 5 yıldız	346
Harika bir	2028	çok güzel ve	334
on numara	2024	herkese tavsiye ederim	316
Çok güzel	1892	tek kelime ile	313
cok guzel	1717	ve nezih bir	306
bir ortam	1638	için ideal bir	306
gereken bir	1622	yediğim en iyi	305

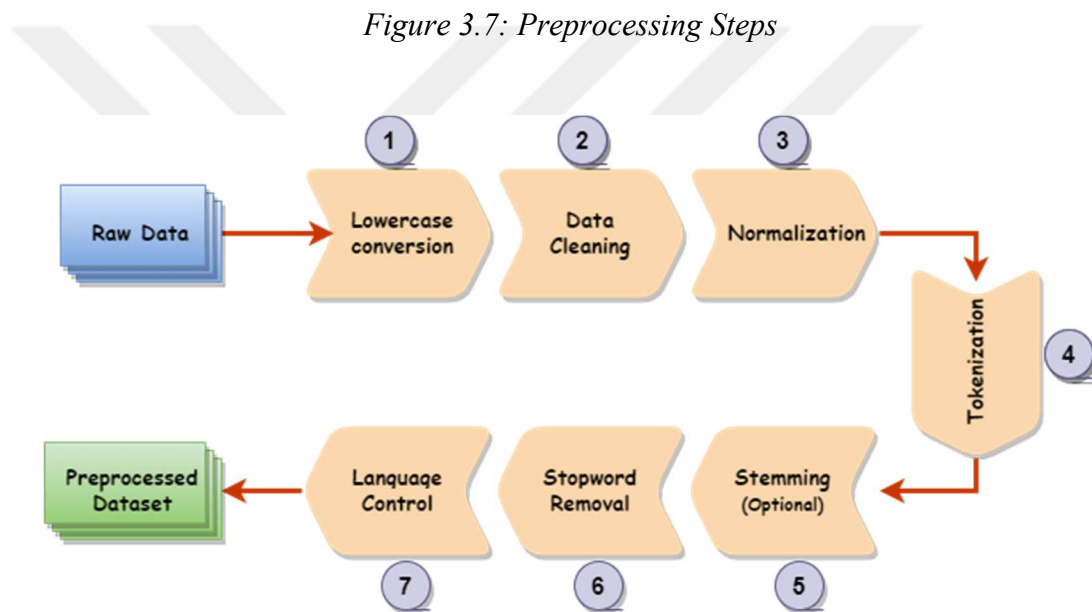
Finally, Table 3.7 show the statistics of generated raw dataset.

Table 3.7: Generated Raw Dataset Statistics

Unique Record Count in Corpus	459,549
Total Word Count in Corpus	4,671,017
Unique Word Count in Corpus	310,785
Top Word Frequency in Corpus	106,145
Max Text Length	51
Average Text Length	10.16
Min Text Length	1

3.1.3 Preprocessing

Data preprocessing is an essential prerequisite for applying machine or deep learning models for classification purposes. The efficacy of the learning outcomes derived from problem-solving tasks relies heavily on providing pertinent, coherent, and intelligible data that can address a particular issue. Data preprocessing involves an iterative process to transform raw data into meaningful and well-suited formats. Various crucial steps are involved in data preprocessing, including data cleaning, data normalization, and tokenization, which are elucidated in Figure 3.7.



1. All of the tips were converted to lowercase.
2. A few data cleaning steps briefly explained below were applied.
 - The numbers 5 and 10 were converted into text for usage as bigram in sentiment analysis because of enriches the meaning by using them as prefixes of some words. For example, the statements such as “10 numara”, “5 yıldız” express strong positive emotions and can be distinctive for classification. Therefore, number 10 was converted to “on” and 5 to “beş”, so the above expressions were turned into “on numara” and “beş yıldız”.

- All non-Turkish characters were excluded to ensure no unwanted characters remain. Thus, texts consisting of letters other than the Latin alphabet, such as Arabic, Cyrillic, and Chinese, were quickly eliminated. On the other hand, numbers and punctuation were also deleted.
 - Some words may have consecutive duplicate characters because of misspellings, intentional, or for the purpose of emphasizing. For example, the word "fakaaat" has three consecutive letters, "a"; these letters have been singularized, so it was transformed into the correct form, "fakat".
 - Accented and old characters like â, î, û converted to the new ones like a, i, u.
3. Words in the corpus were normalized in terms of spelling by using the Zemberek NLP Java library. So misspelled words or informal speech in texts have been fixed.
 4. Tokenization has been applied for each tip, and sentences were transformed into token sets by using the Zemberek NLP Java library. Abbreviations, dates, emails, emojis, hashtags, mentions, metatags, newlines, spaces, tabs, and URLs have been deleted with tokenization process.
 5. Stemming, a constituent of linguistic morphological studies, is the technique employed to reduce derived words to their fundamental or root form. In the process of stemming, there is a trade-off between identifying the simplest stem and the most meaningful one. Besides using the Zemberek NLP Java library as Stemmer, some extra logical controls have been added. This step is optional, both datasets with stems and without stems have been used in classification.
 6. Stop-words, a set of frequently used and comparatively meaningless words, have been removed.
 7. Non-Turkish tips have been tried to detect with the Zemberek NLP Java library. Since it was not very successful, a second eye control was applied manually.

Table 3.8 shows samples of preprocessed data with and without stems.

Table 3.8: Samples of preprocessed data with and without stems

Raw Data	Preprocessed Data	Preprocessed with Stems
İç pilavv😄 tandır😄 sütlaaaaç♥️	iç pilav tandır sütlaç	iç pilav tandır sütlaç
Menu su çok ba sarılı, yıllardır beğenmediğim bir yemek hiç cikmadi	menüsü çok başarılı yıllardır beğenmediğim yemek hiç çıkmadı	menü çok başarılı yıl beğenmemek yemek hiç çıkmmak
Manzarasi konumu guler yuzlu calisanlari ve cayi harika	manzarası konumu güler yüzlü çalışanları çayı harika	manzara konum güler yüzlü çalışan çay harika
CardGusto ile bi ana yemek alana ikincisi bedavaymis. Super!!	cardgusto ana yemek alana ikincisi bedavaymış süper	ana yemek alan ikinci bedava süper
Bandırma'ya yolunuz düşüğünde gitmeniz şiddetle tavsiye olunur 10 numara 5 Yildiz :)	bandırma yolunuz düşüğünde gitmeniz şiddetle tavsiye olunur on numara beş yıldız	bandırma yol düşmek gitmek şiddet tavsiye olur on numara beş yıldız
Koftesi fuzel sulu geliyor biraz lastigimsi tadi olsa da tatlıları da guzel giderim bu arada foyatlaei bilmiyorum	köftesi güzel sulu geliyor biraz lastiksi tadı olsa tatlıları güzel giderim arada foyatlaei bilmiyorum	köfte güzel sulu gelmek biraz lastik tat olsa tatlı güzel giter ara bilmek
Tatlıları başarılı,Özellikle limonlu pastası oldukça güzel...👩	tatlıları başarılı özellikle limonlu pastası oldukça güzel	tatlı başarılı özellikle limonlu pasta oldukça güzel
One of the main course dishes! Must try!	-	-

After preprocessing, the dataset statistics changed as in Table 3.9.

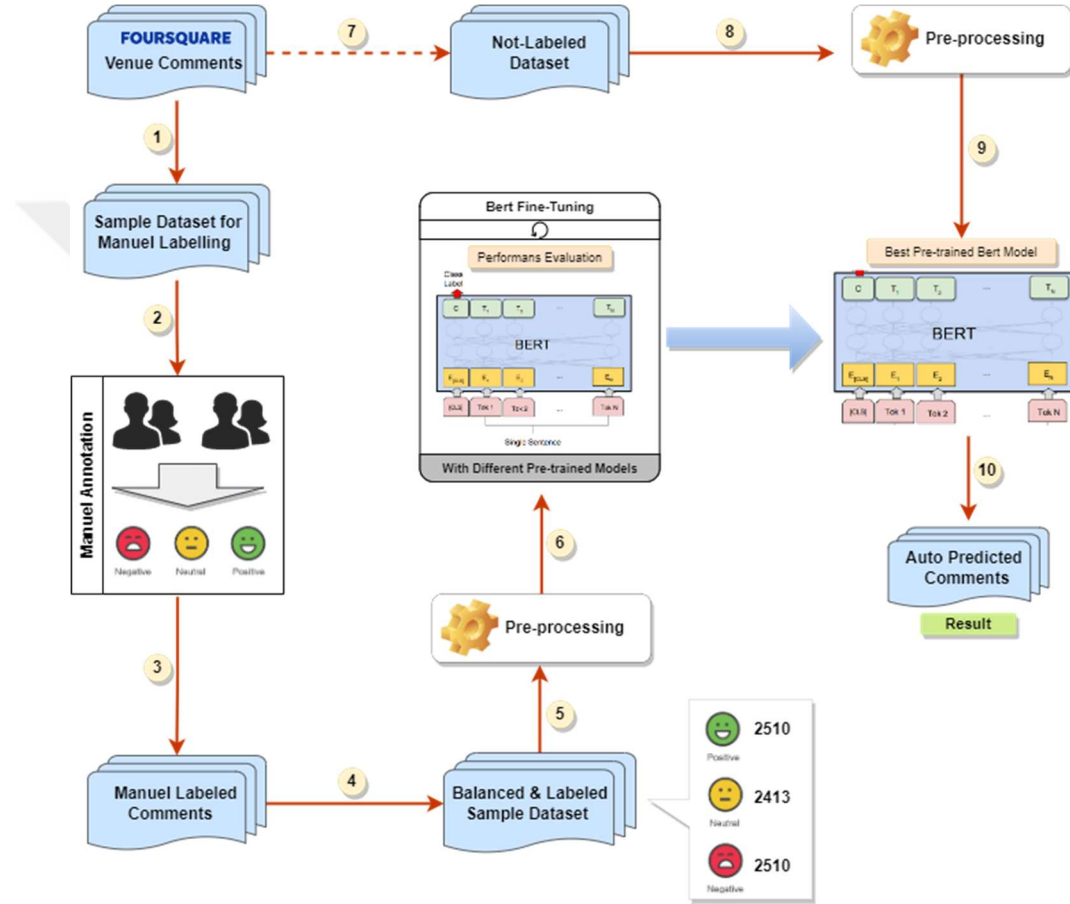
Table 3.9: Preprocessed Dataset Statistics

Unique Record Count in Corpus	435070
Total Word Count in Corpus	3811098
Unique Word Count in Corpus	221661
Top Word Frequency in Corpus	94103
Max Text Length	36
Average Text Length	8.75
Min Text Length	2

3.1.4 Automatic Class Prediction

Some significant steps for automatic class estimation to Foursquare user tips are described in Figure 3.8.

Figure 3.8: Steps of Automatic Class Prediction



First, a sample dataset was created with 12600 Foursquare tips and labeled by experts manually. As a result of this process, it was observed that an unbalanced dataset was formed. The distribution of labels was positive (50%), negative (30%), and neutral (20%). A balanced dataset should be used to make a successful classification. Therefore, a balanced Manual Annotated Corpus (MAC) was created with 2510 positive, 2510 negative, and 2413 neutral tips. After generating a balanced and annotated sample dataset, preprocessing steps described in section 3.1.3 have been applied.

Training a BERT model from scratch with a small dataset can result in overfitting since BERT is a sophisticated neural network architecture incorporating millions of parameters. Therefore, the following pre-trained BERT models, trained on extensive datasets, were used as a fine-tuning starting point.

- ✓ BERT Base Cased ²
- ✓ BERT Base Multilingual Cased (Devlin et al., 2018)
- ✓ BERT Base Turkish 128k Cased ³
- ✓ BERT Base Turkish Cased ⁴
- ✓ BERT Base Turkish Uncased
- ✓ DistilBERT Base Turkish Cased (Sanh et al., 2019)
- ✓ Convbert base turkish cased (Jiang et al., 2020)
- ✓ ConvBERT Base Turkish mc4 Uncased ⁵
- ✓ Electra base turkish cased discriminator (Clark et al., 2020)
- ✓ Electra base turkish mc4 cased discriminator ⁶

It was previously mentioned that Stemming is an optional step when preprocessing the dataset. In order to see the effect of the Stemming step on the fine-tuning process in terms of performance, two different datasets were created by applying both cases separately. Pre-trained BERT models were fine-tuned with these datasets, and the evaluation results are shown below.

According to the evaluation results in Table 3.10 and Table 3.11, it was found that the pre-trained BERT models fine-tuned with the preprocessed dataset without applying the Stemming step were better than the preprocessed dataset with the Stemming step. So, **BERT Base Turkish 128k Cased** pre-trained BERT has been chosen as the best model.

² <https://huggingface.co/bert-base-cased>

³ <https://huggingface.co/dbmdz/bert-base-turkish-128k-cased>

⁴ <https://huggingface.co/dbmdz/bert-base-turkish-cased>

⁵ <https://huggingface.co/dbmdz/convbert-base-turkish-mc4-uncased>

⁶ <https://huggingface.co/dbmdz/electra-base-turkish-mc4-cased-discriminator>

Table 3.10: Performance evaluation of pre-trained BERT models (Without Stems)

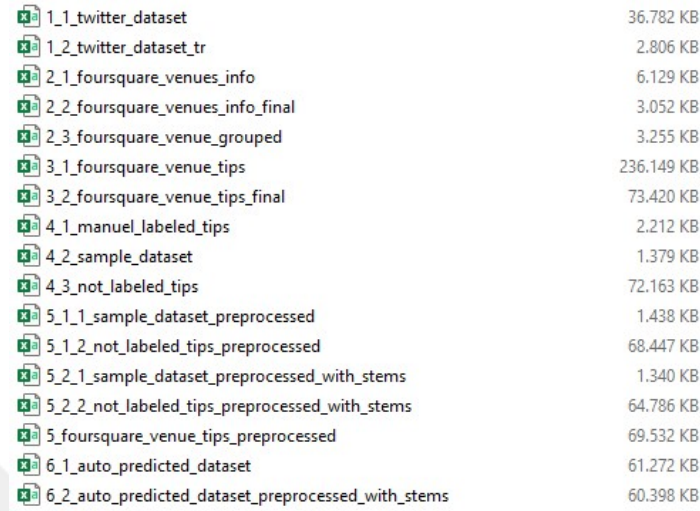
Pre-trained Model	Negative				Neutral				Positive				Accuracy
	Precision	Recall	F1	Support	Precision	Recall	F1	Support	Precision	Recall	F1	Support	
BERT Base Cased	79	73	75	534	59	67	63	466	83	78	81	487	73
BERT Base Multilingual Cased	81	75	78	534	60	67	63	466	81	79	80	487	74
BERT Base Turkish 128k Cased	86	85	85	534	70	77	74	466	89	84	87	487	82
BERT Base Turkish Cased	87	77	82	534	67	79	72	466	88	83	86	487	80
BERT Base Turkish Uncased	84	75	79	534	64	72	68	466	81	80	81	487	76
DistilBERT Base Turkish Cased	87	76	81	534	64	78	70	466	87	80	83	487	78
ConvBERT Base Turkish mc4	87	80	83	534	68	78	73	466	89	84	87	487	81

Table 3.11: Performance evaluation of pre-trained BERT models (With Stems)

Pre-trained Model	Negative				Neutral				Positive				Accuracy
	Precision	Recall	F1	Support	Precision	Recall	F1	Support	Precision	Recall	F1	Support	
BERT Base Cased	78	70	73	526	55	63	59	474	76	75	76	486	69
BERT Base Multilingual Cased	79	74	76	526	58	68	63	474	82	74	78	486	72
BERT Base Turkish 128k Cased	83	83	83	526	69	76	73	474	89	80	84	486	80
BERT Base Turkish Cased	84	81	82	526	69	80	74	474	92	81	86	486	81
BERT Base Turkish Uncased	81	77	79	526	63	68	66	474	81	79	80	486	75
DistilBERT Base Turkish Cased	82	78	80	526	64	73	68	474	86	79	82	486	77
ConvBERT Base Turkish mc4	81	84	83	526	68	74	71	474	90	80	84	486	79

Finally, after preprocessing step, all of the not labeled tips have been auto-predicted by the best pre-trained and fine-tuned BERT model, thus creating an Automatic Annotated Corpus (AAC). Figure 3.9 displays a comprehensive list of the CSV files that were created and utilized in the end-to-end process of generating the final AAC. These files were created and organized in a specific order to ensure that the final AAC was accurate.

Figure 3.9: List of CSV files created throughout the project



1_1_twitter_dataset	36.782 KB
1_2_twitter_dataset_tr	2.806 KB
2_1_foursquare_venues_info	6.129 KB
2_2_foursquare_venues_info_final	3.052 KB
2_3_foursquare_venue_grouped	3.255 KB
3_1_foursquare_venue_tips	236.149 KB
3_2_foursquare_venue_tips_final	73.420 KB
4_1_manuel_labeled_tips	2.212 KB
4_2_sample_dataset	1.379 KB
4_3_not_labeled_tips	72.163 KB
5_1_1_sample_dataset_preprocessed	1.438 KB
5_1_2_not_labeled_tips_preprocessed	68.447 KB
5_2_1_sample_dataset_preprocessed_with_stems	1.340 KB
5_2_2_not_labeled_tips_preprocessed_with_stems	64.786 KB
5_foursquare_venue_tips_preprocessed	69.532 KB
6_1_auto_predicted_dataset	61.272 KB
6_2_auto_predicted_dataset_preprocessed_with_stems	60.398 KB

3.2 Content-based Analysis

Content-based analysis is a research method commonly used in social sciences and humanities for analyzing textual or visual data. It involves a systematic and objective approach to identifying patterns, themes, and insights from the content of the data. Content-based analysis typically involves breaking down data into smaller, manageable units for analysis, coding those units based on predetermined criteria, and analyzing the results to draw conclusions about the content. This method can be applied to various data sources, including written documents, social media posts, and audio or visual recordings.

3.2.1 Word Clouds

Word clouds provide a potent approach for representing data graphically by depicting the frequency of words in a given text. The visual display accords greater significance to words that occur more frequently, enhancing the text's readability and comprehension. Word clouds and frequency diagrams for negative, neutral, and positive users' tips are shown below. Figures 3.10, 3.12, and 3.14 depict word clouds for each class, while figures 3.11, 3.13, and 3.15 display words with the highest frequency.

Figure 3.10: Word Clouds for Negative Tips



Figure 3.11: Words with the highest frequency in Negative Tips

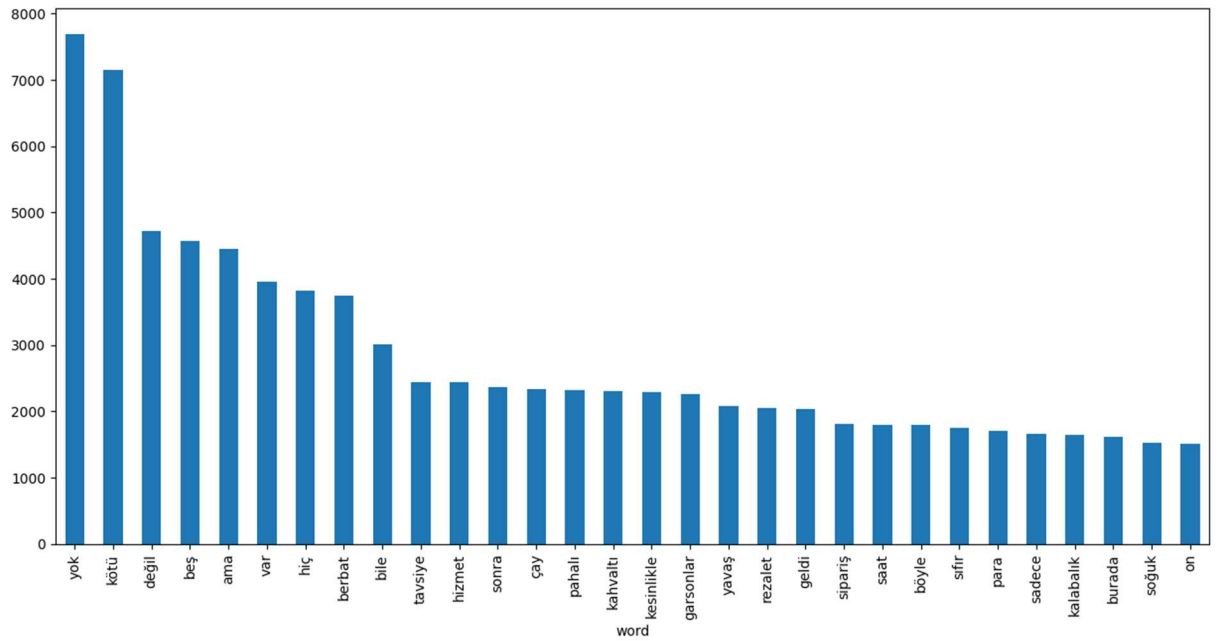


Figure 3.12: Word Clouds for Neutral Tips



Figure 3.13: Words with the highest frequency in Neutral Tips

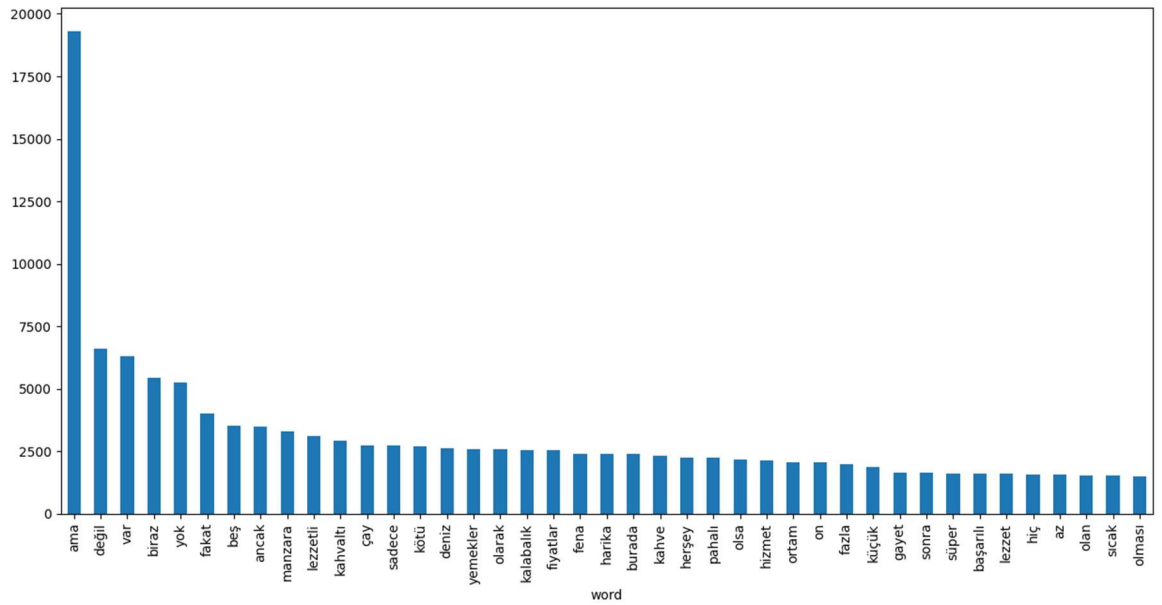


Figure 3.14: Word Clouds for Positive Tips



Figure 3.15: Words with the highest frequency in Positive Tips

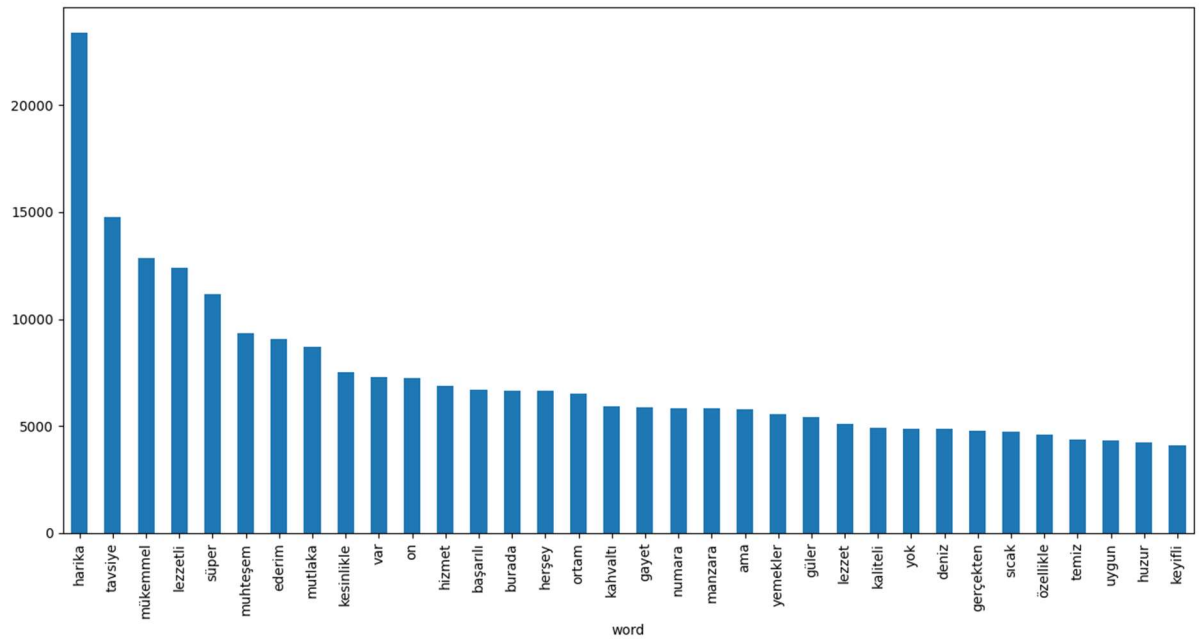


Figure 3.16: Visualization of Words Co-occurrences for Positive Tips

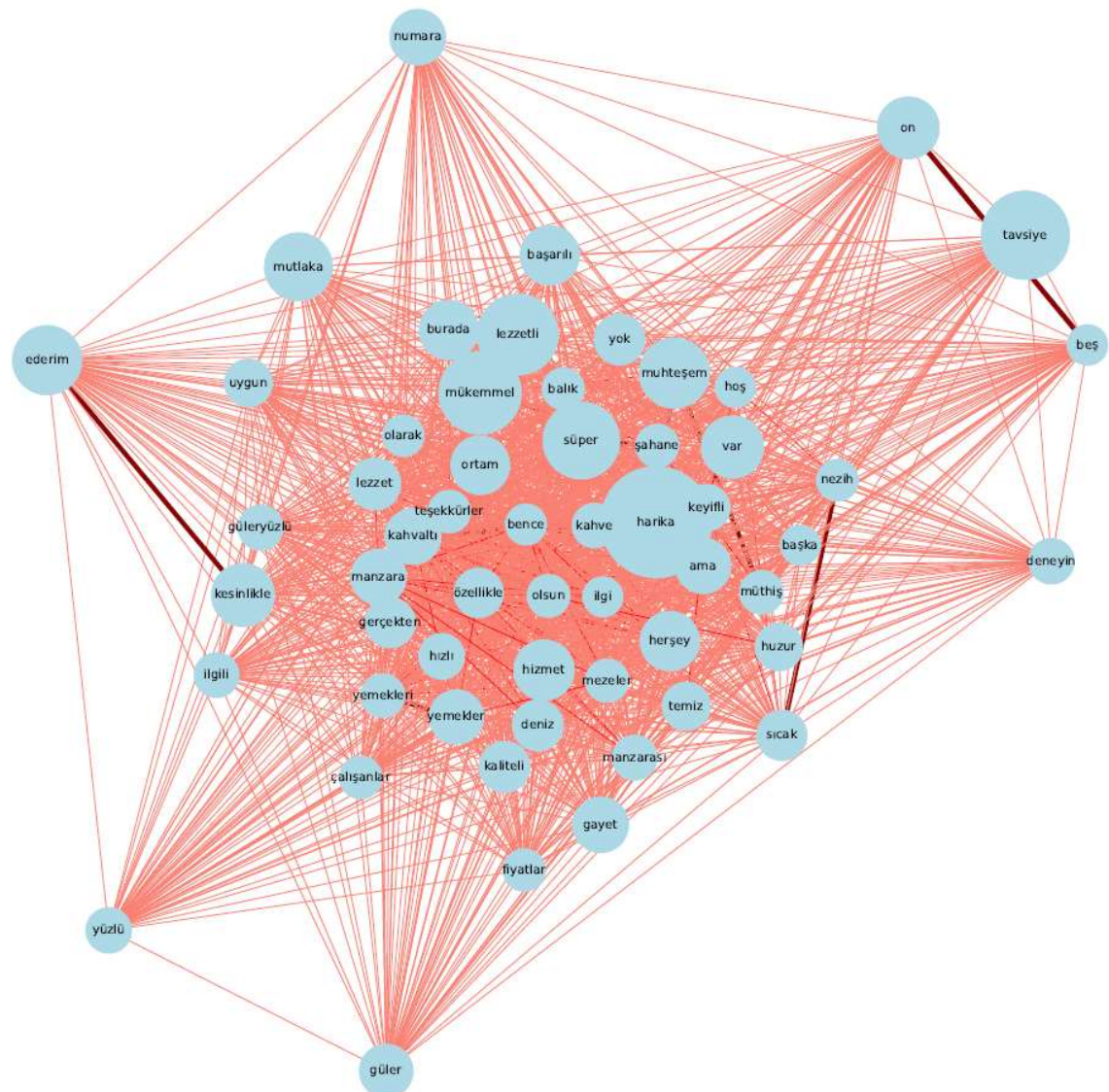


Figure 3.17: Visualization of Words Co-occurrences for Neutral Tips

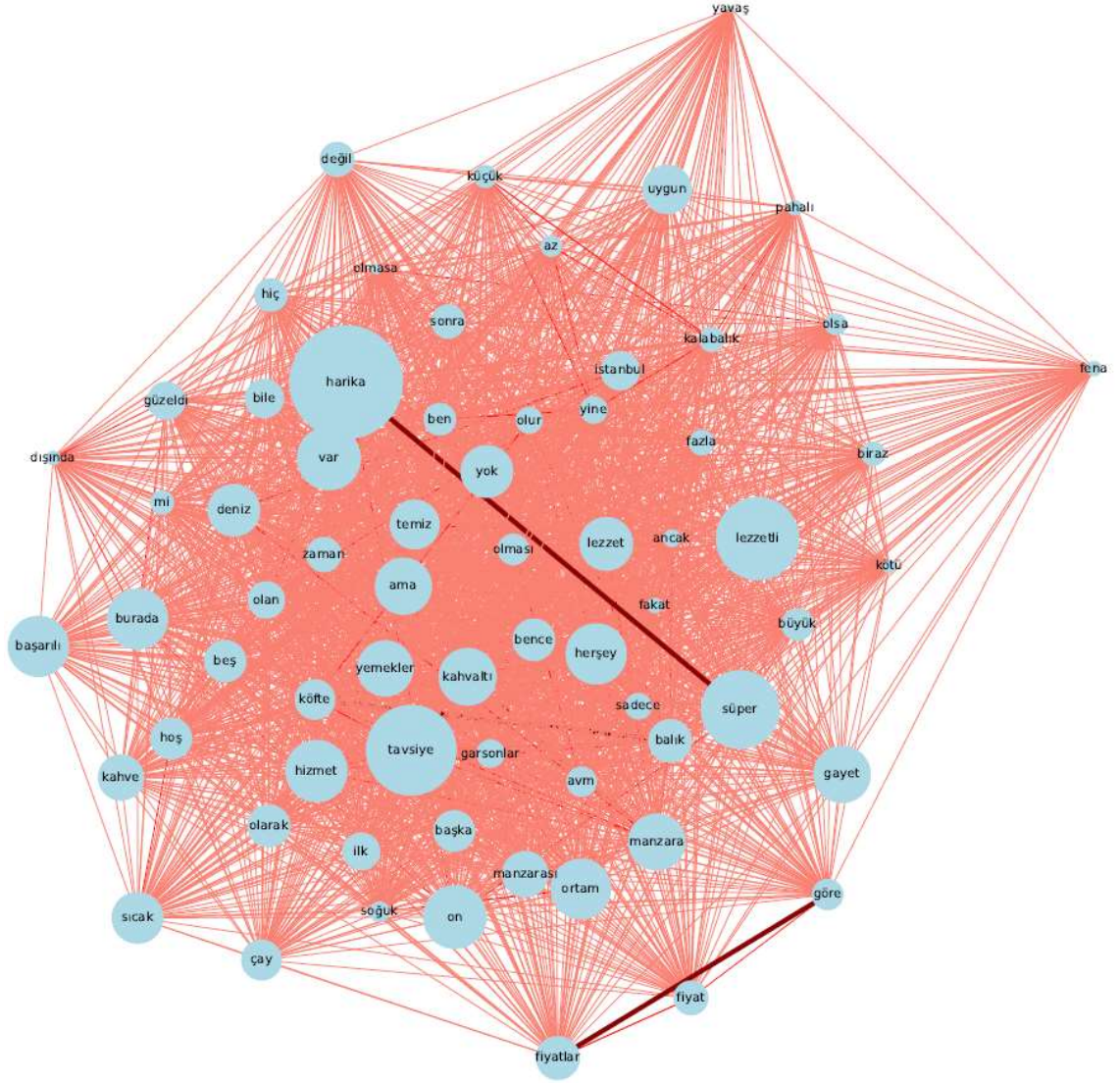
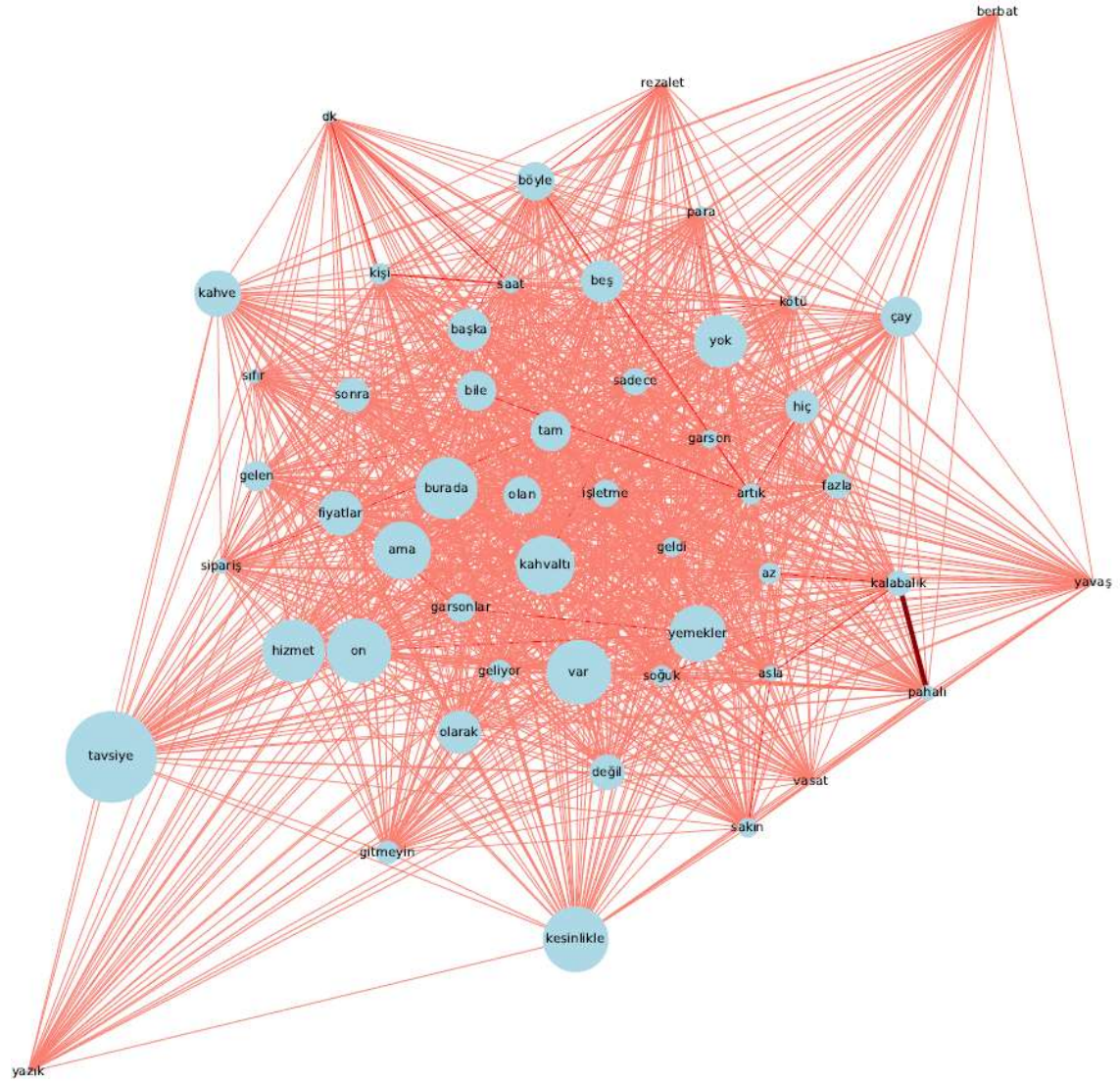


Figure 3.18: Visualization of Words Co-occurrences for Negative Tips



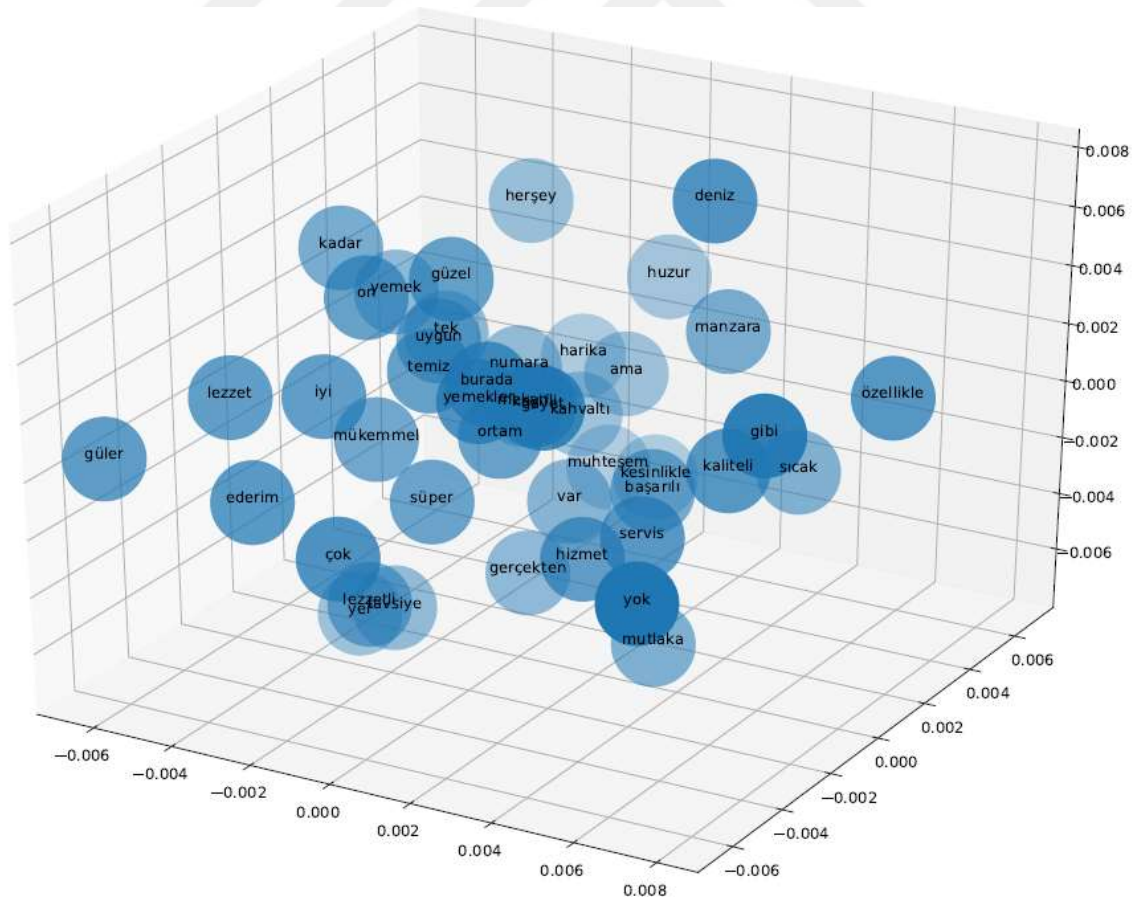
3.2.3 Visualizing Word Embeddings

Word embedding is a technique used in language modeling and feature learning that involves the mapping of each word to a vector of actual values. This process enables words with similar meanings to have comparable representations. The primary application of word embedding is identifying semantic and syntactic similarities among words within a document or their associations with other words within the same context. It is a method used to model natural language for implementing feature engineering in NLP studies, wherein natural language words, sentences, or documents are transformed into vector numbers. This technique reduces the complexity of words and their high-dimensional space into a vector space with smaller dimensions.

It is worth noting that Word2vec (Mikolov et al., 2013) is a widely utilized word embedding system comprising models such as skip-gram and continuous bag-of-words (CBOW). These models are grounded on the likelihood of words occurring in close proximity to each other. Specifically, the skip-gram model aims to predict the words surrounding a given word, whereas CBOW predicts a target word based on the context words.

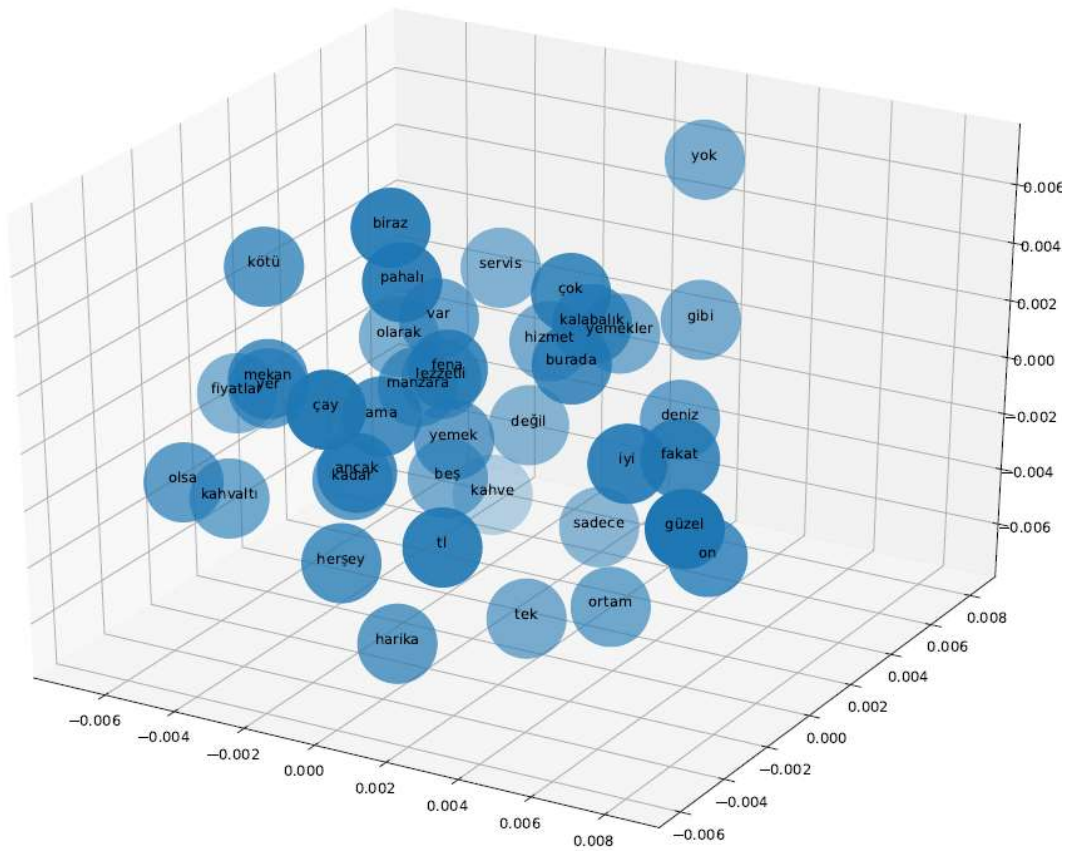
Principal Component Analysis (PCA) is a statistical method used to identify patterns in high-dimensional data by transforming the data into a lower-dimensional space while retaining as much of the original variation as possible. This reduction in dimensions makes it easier to analyze and visualize the data. The primary goal of this method is to find a new set of dimensions that can better capture the variability in the data. By reducing the data to two or three dimensions, visualization becomes much more straightforward.

In this thesis, one of the content-based analysis methods is visualizing the word embeddings in the Automatic Annotated Corpus (AAC). To accomplish this, Word2Vec is implemented and trained using positive, neutral, and negative tips to obtain word embeddings. The resulting embedding vectors are then subject to dimensionality reduction using PCA, which reduces the number of dimensions to three. This technique



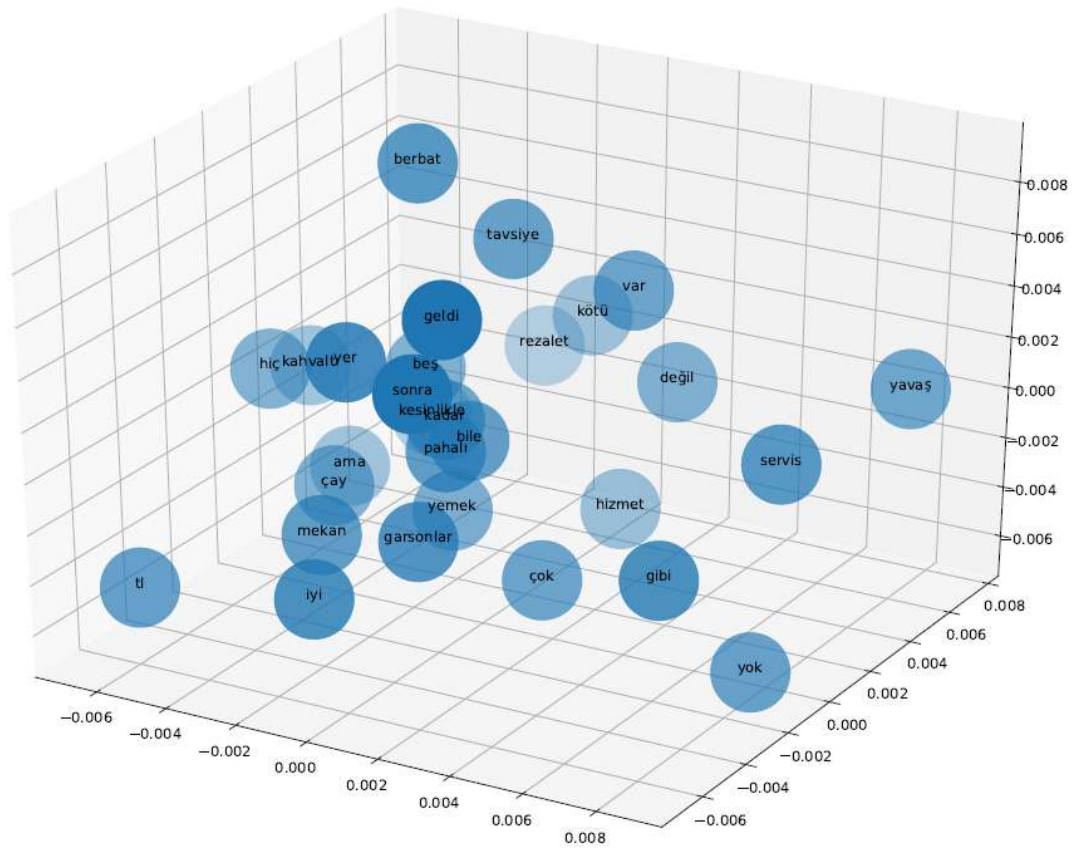
Upon analyzing the embeddings of Neutral tips depicted in Figure 3.20, one may observe the proximity of words such as "iyi" (good), "güzel" (nice), and "fakat" (but). The occurrence of the word "but" after two positive words suggest the possibility of neutral tips containing both positive and negative sentiments.

Figure 3.20: Visualization of Neutral Tips Embeddings



Finally, considering the embeddings of the Negative tips in Figure 3.21, it becomes apparent that words such as "rezalet" (disgraceful) and "kötü" (bad), "kesinlikle" (definitely) and "pahalı" (expensive) are closely clustered together.

Figure 3.21: Visualization of Negative Tips Embeddings



3.2.4 Topic Modelling

Topic Modeling is a statistical technique employed to uncover latent topics or themes present in a corpus of documents. It is widely used in natural language processing (NLP) and has applications in various domains such as text mining, information retrieval, social network analysis, and computational social science.

There are several approaches to topic modeling, but the most widely used ones are Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF). LDA assumes that each document is a mixture of a few topics, and each topic is a distribution over words. NMF, on the other hand, decomposes the term-document matrix into two matrices: a matrix of topics and a matrix of coefficients that assign weights to the topics in each document.

Latent Dirichlet Allocation (LDA) is a statistical model used for topic modeling to identify hidden or latent topics in a collection of documents. LDA assumes that each document is a mixture of a few topics, and each topic is a distribution over words. In other words, LDA assumes that a document is a probabilistic combination of a set of topics, where each topic is a probability distribution over words. The model works by assigning each word in a document to a topic, and each document is represented as a probability distribution over topics. The critical assumption of LDA is that each document comprises multiple topics, and each topic has a probability distribution of words associated with it. It allows for the discovery of hidden or latent topics present in the collection of documents. The model has been widely used in NLP, information retrieval, and other related fields. (Blei, 2003)

This section presents the topic analysis outcomes of the AAC dataset using the Latent Dirichlet Allocation method available in the Gensim library. The objective is to model negative, neutral, and positive topics in the created corpora based on the publication year of the tips and extracted the most appropriate word sets. We applied three different analyses on user tips based on the city, category, and evaluation score of venue.

Initially, we grouped the venues by their city, category group, and rating score. Subsequently, we identified the top 6 based on their popularity in the city and category group as the topics. We further categorized these venues into three ranges based on their rating score: low (below 6), medium (between 6 and 8), and high (above 8). Then, for each topic, Latent Dirichlet Allocation extracted the most appropriate words annually.

Tables 3.12, 3.13, and 3.14 displays the LDA-generated output of 20 words associated with each topic. The words in the tables are weighted according to their relation to the subject and are presented in descending order from top to bottom.

In Table 3.12, the list of topics is based on the venue's location and includes the most popular cities: Istanbul, Ankara, Izmir, Mugla, Antalya, and Bursa. We can identify patterns or trends across the different cities and their associated venues and also determine the most frequently occurring words related to each city's venues. For example, while the word “Kahvaltı” (breakfast) is trendy in Istanbul, Izmir, and Bursa, it is not so popular in Ankara and Muğla. In cities such as Muğla and Antalya, located on the coast and are touristic, the words “Manzara” (landscape) and “Deniz” (sea) are also popular. On the other hand, considering the people's economic situation, the word “Fiyat” (price) started to become popular among cities since 2015.

In Table 3.13, the topics are organized according to the venue's category groups, such as “Yemek Mekanı” (Food), “Kültür, Spor, Eğlence” (Culture, Sports, Entertainment), and “Ulaşım, Konaklama” (Transportation, Accommodation). By applying LDA to each category group separately, we can gain insights into the key factors people care about in each type of venue. For example, “Manzara” (landscape) and “Lezzet” (taste) in the Food category are the elements customers look for most in a place.

In Table 3.14 arranges its topics based on the rating score ranges of the venues, such as low, medium, and high. By applying LDA to each rating score range separately, we can gain insights into the key factors that customers associate with different levels of satisfaction. This can help venue owners and managers to better understand what drives customer satisfaction and dissatisfaction and make more informed decisions about how to improve their offerings.

Table 3.12: Wordlist for Top 6 Cities By Year

	< 2013	2013	2014	2015	2016	2017	2018	> 2018
İstanbul	manzara çay kahvaltı kahve garson mutlaka tatlı lezzet lezzetli başarılı fiyat yavaş sıcak zaman kalabalık insan balık boğaz tat tavuk	manzara kahvaltı insan kahve çay lezzetli lezzet garson zaman mutlaka çalışan herkes avm tat başarılı deniz balık fiyat tatlı huzur	manzara kahvaltı kahve çay lezzet lezzetli garson çalışan avm insan fiyat tatlı tat huzur mutlaka balık deniz kalite zaman sıcak	manzara kahve kahvaltı lezzetli fiyat lezzet tatlı çalışan garson çay başarılı avm tat balık insan et kalabalık mutlaka zaman pizza	kahve manzara kahvaltı lezzetli fiyat lezzet çalışan çay tatlı başarılı garson et menü köfte meze tat sıcak pizza insan zaman	kahvaltı manzara lezzetli kahve fiyat lezzet tatlı çalışan çay meze garson başarılı et köfte menü kalite tat sıcak mutlaka personel	kahvaltı lezzetli fiyat kahve manzara lezzet çalışan tatlı çay meze garson nargile başarılı köfte güler börek tat sıcak kalite ekmek	sıcak lezzetli fiyat kahvaltı kahve manzara dakika bebek lezzet tatlı çalışan meze teşekkür çay başarılı garson insan kalite ilgili et
Ankara	pasta tuz burger garson kahvaltı çorba çay mutlaka limonata lezzetli tatlı masa sıcak berbat zaman limonlu lezzet kahve başarılı bira	garson insan çorba zaman çalışan lezzetli çay mutlaka sıcak tuz lezzet müzik tavuk berbat masa hoş kahve başarılı müşteri fiyat	pide garson lezzet lezzetli çorba çay insan çalışan sıcak pasta başarılı kahvaltı tatlı mutlaka pizza tat zaman kahve fiyat ilgi	pide lezzetli pizza fiyat çorba lezzet garson et başarılı kahve çalışan et kahvaltı çay çorba tat kalite pasta çay kalabalık mutlaka avm	pide lezzetli fiyat lezzet tatlı pasta kahve çalışan başarılı kahvaltı et çay çorba ciğer dondurma garson tat limonlu mutlaka köfte	kahve lezzetli fiyat lezzet döner çalışan tatlı et kahvaltı salon başarılı müzik düğün ilgi ürün kalite tat uygun sıcak	kahve kahvaltı lezzetli waffle fiyat tatlı çalışan börek ürün lezzet insan dondurma doğum taze parti efsane özel sıcak kişi malzeme	kahve kahvaltı lezzetli waffle fiyat tatlı çalışan börek ürün lezzet insan dondurma doğum taze parti efsane özel sıcak kişi malzeme

	< 2013	2013	2014	2015	2016	2017	2018	> 2018
İzmir	kahvaltı dondurma sakızlı tavuk mutlaka başarılı çay garson kahve damla müzik tat balık insan kalite lezzetli tatlı olan fiyat menü	kahvaltı deniz kahve lezzet dondurma garson lezzetli çalışan insan balık sıcak tat huzur mutlaka waffle çay manzara tavuk ilgi zaman	kahvaltı deniz kahve dondurma garson lezzet lezzetli çalışan manzara balık kumru fiyat çay insan personel personel müzik nargile sıcak hızlı mutlaka	kahvaltı fiyat lezzetli meze lezzet çalışan kahve deniz garson dondurma mutlaka başarılı personel sıcak çay tat manzara temiz çeşit	kahvaltı kahve fiyat lezzetli dondurma çalışan lezzet meze deniz başarılı garson mutlaka çay personel manzara balık pizza menü insan uygun	kahvaltı lezzetli nargile fiyat pide lezzet pizza ilgi çalışan kahve başarılı meze dondurma köfte deniz temiz personel mutlaka garson ilgili	kahvaltı fiyat lezzetli lezzet kahve dondurma meze pizza çalışan burger başarılı deniz köfte temiz işletmek garson çay personel tatlı manzara	fiyat lezzetli kahvaltı lezzet kahve meze pizza çalışan dondurma balık temiz personel başarılı güler ilgili deniz tat işletmek manzara menü
Muğla	deniz balık bodrum mutlaka meze manzara marina fiyat lezzetli marmaris plaj kaliteli insan başarılı otel kahvaltı kalamar rakı olan tatlı	manzara bodrum deniz balık mutlaka marmaris kahvaltı marina otel lezzet kahve meze lezzetli tat huzur insan yol cennet personel tatlı	deniz manzara balık bodrum otel kahvaltı lezzetli müzik fiyat meze lezzet insan kaliteli kalite tatil pizza mutlaka kahve huzur marmaris	manzara deniz balık lezzetli bodrum fiyat meze kahvaltı pizza tatlı lezzet personel mutlaka çalışan marina müzik güler kalamar uygun başarılı	manzara meze fiyat deniz lezzetli balık bodrum kahvaltı lezzet mutlaka uygun personel otel tatlı güler işletmek samimi müzik sıcak	deniz fiyat meze balık kalamar lezzetli lezzet manzara pide bodrum çalışan kahvaltı pizza oda ilgili ızgara temiz uygun mutlaka sıcak	lezzetli deniz fiyat meze pizza balık lezzet kahvaltı bodrum manzara çalışan mutlaka işletmek güler kalamar plaj uygun müzik kalite sıcak	kahvaltı fiyat deniz meze lezzetli çalışan kahve balık lezzet pizza temiz işletmek ahtapot ilgili pide sahip otel manzara güler mutlaka

	< 2013	2013	2014	2015	2016	2017	2018	> 2018
Antalya	manzara balık otel deniz para mutlaka tatlı fiyat oda kahvaltı kabak köfte piyaz lezzet dondurma temiz döner akşam lezzetli su	manzara otel deniz tatlı lezzetli köfte oda kahvaltı çorba piyaz mutlaka personel lezzet çıkmak kabak sıcak fiyat çalışan kalite tam	manzara deniz balık otel lezzetli balık lezzetli kokoreç tatlı insan çalışan tat mutlaka personel köfte müzik sıcak fiyat gece çıkmak kahve	deniz manzara otel lezzetli balık fiyat kokoreç temiz personel işletmek çalışan köfte kahve meze çorba et oda mutlaka tatlı	deniz fiyat ipek nargile lezzetli kahvaltı samimi manzara uygun güler temiz otel personel yüz alanya müzik çalışan su canlı teşekkür	teşekkür nargile deniz güler personel çalışan fiyat otel huzur çay lezzetli olan temiz manzara belediye güleryüzlü lezzet taze çeşit güler	deniz lezzetli nargile fiyat ürün taze temiz kahvaltı uygun plaj kahve lezzet kalite kaliteli personel otel su ilgi müzik çalışan	fiyat deniz çalışan temiz kahvaltı uygun personel lezzetli lezzet güler manzara künefe ekmek katmer plaj işletmek tatlı kaliteli nargile teşekkür
Bursa	kahvaltı lezzetli kahve başarılı lezzet zaman hızlı kafa mutlaka dere köfte et tam çocuk manzara otel sakin şehir büfe herkes	kahvaltı kahve et çay lezzet mutlaka lezzetli personel huzur tavuk tat güler nezih fiyat çalışan soslu sıcak müzik bonfile kalite	kahvaltı et lokum lezzet kahve çay lezzetli huzur garson manzara personel mutlaka sıcak su tat fiyat dere tatlı köfte bonfile	kahvaltı et lokum lezzet fiyat lezzetli kahve müzik köfte manzara garson meze kalite sıcak pide çay menü balık huzur başarılı	kahvaltı fiyat lezzetli lokum lezzet et kahve tatlı çay çalışan manzara personel başarılı garson pide kalite pasta pizza meze menü	kahvaltı et fiyat lezzetli lokum lezzet çay pide çalışan kahve meze personel kalite başarılı sıcak garson tatlı köfte manzara ilgi	kahvaltı fiyat lokum lezzetli lezzet et çay kalite meze kahve başarılı ciğer personel köfte kebab masa tat tatlı garson	kahvaltı fiyat lezzetli et lezzet kahve meze lokum personel köfte kalite kaliteli kebab çalışan balık güler tat çay başarılı ürün

Table 3.13: Wordlist for Category Groups By Year

	< 2013	2013	2014	2015	2016	2017	2018	> 2018
Yemek Mekanı	kahvaltı manzara kahve çay garson lezzetli lezzet mutlaka tatlı fiyat başarılı çorba tavuk tat köfte tost balık yavaş istanbul sıcak	kahvaltı manzara lezzet lezzetli garson kahve çay tat mutlaka köfte tatlı başarılı çalışan balık fiyat başarılı sıcak tavuk kebab müzik çorba	kahvaltı lezzet manzara kahve lezzetli garson çay köfte tatlı fiyat tat çalışan balık mutlaka sıcak başarılı et pide kebab pizza	kahve kahvaltı lezzetli manzara lezzet fiyat köfte çalışan tatlı et çay garson pide tat başarılı meze balık kebab mutlaka pizza	kahvaltı lezzetli kahve fiyat lezzet manzara köfte çalışan tatlı çay et meze başarılı pide garson menü mutlaka tat pizza sıcak	kahvaltı mantı lezzetli lezzet fiyat kahve manzara çalışan tatlı manzara el başarılı zaman meze et mutlaka güleryüzlü tat çay üzerine	lezzetli fiyat kahvaltı kahve lezzet çalışan meze köfte tatlı manzara çay et kebab başarılı tat uygun pide pizza personel nargile	lezzetli fiyat kahvaltı kahve lezzet çalışan meze manzara tatlı köfte kebab pizza güler başarılı çay et pide personel ilgili garson
Kültür, Spor, Eğlence	salon fuar film sinema kitap koltuk spor zaman izlemek reklam bilet lastik mutlaka istanbul aralık otopark beklemek gereken aynı sahip	salon mutlaka gereken spor hoca gezmek film büyükşehir fuar el eskişehir ders yılmaz sinema herkes insan insan izlemek heykel sağlık sıra	salon sinema film spor fuar müze oyun gereken hoca reklam izlemek mutlaka herkes gezmek insan zaman ımax koltuk emek girmek	salon sinema film fuar koltuk spor izlemek müze rahat reklam hoca bilet alan zaman fiyat girmek ses pahalı	salon sinema film fuar koltuk spor izlemek reklam müze ses insan alan cinemamaximum fiyat ımax temiz küçük izmir bilet	salon sinema koltuk film müze spor fuar insan izlemek gezmek mutlaka çocuk reklam rahat alan keyifli herkes heykel el	salon müze film spor sinema fuar koltuk girmek çocuk fiyat gezmek mutlaka izlemek reklam gereken heykel bilet alan insan yol	müze salon sinema film çocuk gezmek girmek fuar para mutlaka alan spor insan gereken zaman ankara rahat heykel keyifli hoca

	< 2013	2013	2014	2015	2016	2017	2018	> 2018
Ulaşım, Konaklama	otel	otel	otel	otel	otel	otel	otel	otel
	oda	deniz	deniz	oda	oda	oda	oda	kahvaltı
	deniz	insan	oda	deniz	temiz	temiz	temiz	temiz
	yol	oda	temiz	temiz	deniz	kahvaltı	kahvaltı	oda
	istanbul	yol	insan	kahvaltı	kahvaltı	personel	deniz	sıcak
	temiz	hava	hava	manzara	personel	çalışan	personel	dakika
	insan	huzur	personel	personel	manzara	manzara	çalışan	bebek
	çay	istanbul	manzara	çalışan	havuz	havuz	insan	işletmek
	kalmak	temiz	huzur	havuz	kalmak	güler	manzara	çalışan
	şehir	manzara	kahvaltı	huzur	çalışan	kalmak	fiyat	manzara
	trafik	şehir	çalışan	insan	fiyat	deniz	kalmak	açık
	manzara	kalmak	kalmak	fiyat	insan	yüzlü	şehir	ekip
	vapur	zaman	tatil	kalite	huzur	fiyat	huzur	huzur
	yeni	personel	havuz	kalmak	güler	aile	havuz	koala
	hava	çay	yol	gülyüzlü	tatil	ilgili	sakin	deniz
	kahvaltı	çalışan	şehir	huzurlu	şehir	insan	tatil	teşekkür
	eski	sıcak	sakin	yol	gülyüzlü	teşekkür	güler	tatil
	huzur	beklemek	istanbul	hava	küçük	küçük	kalite	zaman
	moda	sen	olan	tatil	işletmek	kalite	zaman	rahat
	havuz	kahvaltı	çay	sahip	sıcak	tatil	hava	kaliteli
Açık Alan, Doğa	manzara	deniz	deniz	deniz	deniz	deniz	deniz	deniz
	deniz	manzara	manzara	manzara	manzara	manzara	sahil	sıcak
	tuz	huzur	huzur	huzur	plaj	sahil	manzara	manzara
	girmek	hava	sahil	sahil	sahil	huzur	kalabalık	dakika
	inmek	sahil	hava	yürümek	huzur	su	insan	bebek
	mutlaka	insan	insan	plaj	temiz	temiz	plaj	insan
	kale	su	plaj	insan	yürümek	plaj	su	sahil
	çıkma	plaj	yürümek	temiz	doğa	yürümek	huzur	temiz
	zaman	istanbul	istanbul	kalabalık	su	girmek	hava	hava
	karadeniz	yürümek	su	mutlaka	insan	mutlaka	temiz	huzur
	sahil	mutlaka	temiz	hava	kalabalık	doğa	doğa	plaj
	su	sakin	çay	doğa	mutlaka	insan	yol	kalabalık
	insan	kalabalık	doğa	su	girmek	kalabalık	yürümek	doğa
	izmir	gereken	sakin	istanbul	hava	hava	sakin	yol
	huzur	kafa	kafa	yol	istanbul	sakin	girmek	su
	mavi	temiz	kalabalık	girmek	yol	yol	gezme	gezme
	yürümek	çay	gereken	kafa	sakin	gereken	işletmek	yürümek
	olan	doğa	mutlaka	zaman	kum	park	mutlaka	fiyat
	sakin	zaman	kum	sakin	gereken	istanbul	olan	sakin
	müthiş	soğuk	gezme	çay	işletmek	şezlong	kafa	girmek

	< 2013	2013	2014	2015	2016	2017	2018	> 2018
İşyeri , İbadethane	avm alışveriş merkezi mağaza kalabalık zaman sinema ankara insan otopark çıkma açmak girmek iş kat istanbul hastane sadece olan yol	avm alışveriş insan hastane iş çalışan zaman mağaza beklemek herkes merkezi sen personel kalabalık girmek otopark kaliteli bulmak ilgi çıkma	avm alışveriş insan hastane mağaza otopark iş doktor merkezi bulmak sinema olan kalabalık gezmek teşekkür ilgi para girmek zaman zorlu	avm alışveriş mağaza merkezi kalabalık insan otopark zaman gezmek istanbul marka hava kaliteli ferah açık marina girmek izmir fiyat bulmak	avm alışveriş mağaza bira otopark merkezi kalabalık kaliteli iş fiyat hava çıkma insan istanbul ferah marka çalışan bulmak açık alan	avm bira fiyat burger hamburger kaliteli teşekkür alışveriş ürün çalışan mağaza insan başarılı iş hastane ilgi uygun bulmak kalite istanbul	burger hamburger fiyat avm alışveriş iş lezzetli ürün kaliteli ekmek hastane uygun insan özel patates bulmak teşekkür kalite personel saç	hamburger avm fiyat burger ürün kaliteli alışveriş patates lezzetli temiz insan mağaza ilgili satmak hayat olan özel kalite uygun ekmek
Diğer	balık para tünel adliye km akvaryum ankara açmak kar mevcut ceza sakın kapı hemen hız dr fazla sahip yıkamak girmek	tünel balık iş mutlaka dünya insan bol ses akvaryum araba dağ adliye manzara girmek hava herkes geçmek yaptırmak temiz ok	tünel manzara yol adliye fazla insan balık herkes adalet okumak hava çay akvaryum istanbul huzur düğün bol girmek zor gereken	balık tünel girmek akvaryum fiyat çiçek mutlaka tat iş zaman manzara araç eğlenceli ankara güler şarap bol zevkli fazla	şarap mutlaka fiyat iş teşekkür balık kaliteli çöp manzara lezzetli ilgi hayat sadece tünel şiş huzur çay içeri insan dış	şarap fiyat tünel geçmek pahalı şiş çöp akvaryum teşekkür mayıs iş zaman huzur beklemek araba manzara et açmak murat kalite	şarap fiyat üretim et olan araba kaliteli koy akvaryum kırmızı merinos iş yıkamak taksi lezzetli çöp girmek damızlık pahalı para	şarap fiyat kalite para lezzetli iş çöp şiş gazipaşa taksi et ürün hızlı randevu bitki çalışan lazım tören kaliteli krep

Table 3.14: Wordlist for Rating By Year

	< 2013	2013	2014	2015	2016	2017	2018	> 2018
Low	çay	çay	kahve	kahvaltı	kahvaltı	fiyat	fiyat	sıcak
	kahve	kahve	çay	fiyat	fiyat	kahvaltı	kahvaltı	dakika
	garson	garson	garson	çay	lezzet	lezzetli	lezzetli	bebek
	kahvaltı	çalışan	kahvaltı	kahve	lezzetli	lezzet	lezzet	kahvaltı
	fiyat	kahvaltı	fiyat	lezzetli	çay	çay	çalışan	fiyat
	berbat	lezzet	çalışan	personel	garson	çalışan	börek	lezzetli
	çalışan	fiyat	insan	çalışan	kahve	et	çay	lezzet
	ekmek	berbat	berbat	kebab	çalışan	döner	temiz	çalışan
	avm	yavaş	lezzetli	manzara	manzara	ilgi	köfte	kahve
	lezzetli	lezzetli	lezzet	temiz	menü	kahve	midye	temiz
	sipariş	insan	teşekkür	lezzet	kebab	kebab	personel	manzara
	yavaş	mutlaka	manzara	köfte	köfte	nargile	kebab	personel
	başarılı	nargile	personel	garson	personel	personel	pahalı	çay
	nargile	manzara	sipariş	berbat	et	garson	kalite	para
	lezzet	sıcak	sıcak	döner	para	köfte	berbat	pahalı
	sıcak	zaman	avm	insan	berbat	pide	uygun	insan
	müzik	personel	güler	hızlı	çorba	başarılı	para	döner
	kalabalık	otel	köfte	tat	başarılı	manzara	menü	kalite
	çorba	başarılı	otel	meze	meze	işletmek	güler	köfte
	insan	tat	müşteri	pizza	temiz	temiz	işletmek	güler
Medium	kahvaltı	manzara	kahvaltı	kahve	kahvaltı	kahvaltı	kahvaltı	kahvaltı
	manzara	kahvaltı	manzara	kahvaltı	fiyat	lezzetli	fiyat	fiyat
	kahve	kahve	lezzet	manzara	lezzetli	fiyat	lezzetli	lezzetli
	mutlaka	lezzetli	lezzetli	lezzetli	kahve	kahve	kahve	kahve
	lezzetli	lezzet	köfte	fiyat	manzara	lezzet	lezzet	lezzet
	çay	garson	kahve	lezzet	lezzet	manzara	çalışan	çalışan
	garson	deniz	çay	çalışan	çalışan	çalışan	manzara	manzara
	fiyat	çay	garson	garson	köfte	meze	meze	meze
	lezzet	mutlaka	deniz	köfte	çay	köfte	tatlı	köfte
	başarılı	köfte	çalışan	çay	meze	çay	köfte	deniz
	tost	insan	pide	tatlı	et	tatlı	çay	personel
	deniz	tat	fiyat	pide	başarılı	başarılı	garson	başarılı
	tatlı	çalışan	tat	deniz	deniz	garson	uygun	güler
	tavuk	avm	insan	et	garson	et	deniz	çay
	köfte	balık	mutlaka	başarılı	personel	uygun	başarılı	tatlı
	çorba	fiyat	balık	meze	tatlı	deniz	personel	temiz
	istanbul	tost	tatlı	tat	mutlaka	personel	nargile	pide
	zaman	sıcak	sıcak	kebab	balık	pide	tat	ilgili
	kalabalık	zaman	başarılı	mutlaka	pide	mutlaka	temiz	kebab
	yavaş	başarılı	personel	kalite	kebab	sıcak	kalite	garson

	< 2013	2013	2014	2015	2016	2017	2018	> 2018
High	manzara	manzara	manzara	manzara	lezzetli	lezzetli	lezzetli	kahve
	deniz	deniz	deniz	deniz	manzara	fiyat	fiyat	lezzetli
	mutlaka	mutlaka	huzur	lezzetli	kahve	manzara	kahve	fiyat
	istanbul	insan	lezzet	kahve	fiyat	lezzet	manzara	kahvaltı
	kahvaltı	huzur	kahve	lezzet	kahvaltı	pizza	lezzet	lezzet
	çay	istanbul	insan	tatlı	lezzet	kahve	kahvaltı	deniz
	tatlı	lezzet	mutlaka	fiyat	deniz	kahvaltı	deniz	çalışan
	insan	balık	lezzetli	kahvaltı	tatlı	deniz	çalışan	manzara
	balık	hava	kahvaltı	pizza	çalışan	çalışan	meze	temiz
	lezzet	çay	tatlı	mutlaka	pizza	mutlaka	et	tatlı
	lezzetli	zaman	pizza	et	mutlaka	et	tatlı	pizza
	su	kahvaltı	tat	köfte	köfte	tatlı	mutlaka	meze
	yol	gereken	çay	insan	insan	meze	pizza	sıcak
	sıcak	tat	istanbul	balık	başarılı	başarılı	balık	mutlaka
	başarılı	tatlı	sıcak	başarılı	et	köfte	köfte	kebab
	hava	sıcak	balık	tat	sıcak	balık	temiz	güler
	fiyat	lezzetli	hava	çalışan	çay	sıcak	başarılı	teşekkür
	girmek	su	çalışan	istanbul	tat	temiz	sıcak	ilgili
	zaman	yol	temiz	huzur	teşekkür	personel	tat	et
	tat	temiz	köfte	sıcak	menü	tat	insan	başarılı

3.3 Spatio-temporal Analysis

Spatio-temporal Analysis is an important field of research that plays a critical role in understanding complex systems that vary over space and time. It offers a range of tools and techniques that can be used to analyze and visualize spatio-temporal data, providing valuable insights for a wide range of applications.

Since most of the venues used in this thesis have geographic location information, the temporal and spatial distribution of the interpretations of the space can be easily made. In addition, the time information in the dataset allows for examining the data on a temporal basis using maps. Figures 3.22 to 3.34 illustrate the annual distribution of tips in Türkiye using heat maps.

The heat map displays heat points in various colors, such as red, yellow, green, and blue, representing regions with the highest to lowest number of reviews posted, respectively.

Figure 3.22: User Review Distribution in 2010

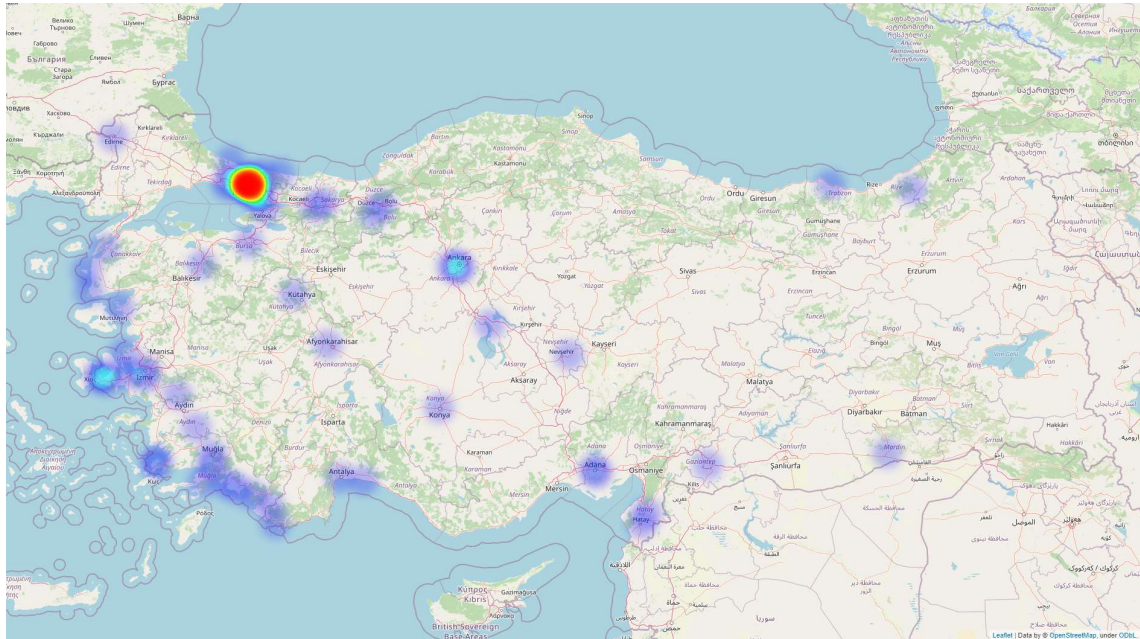


Figure 3.23: User Review Distribution in 2011

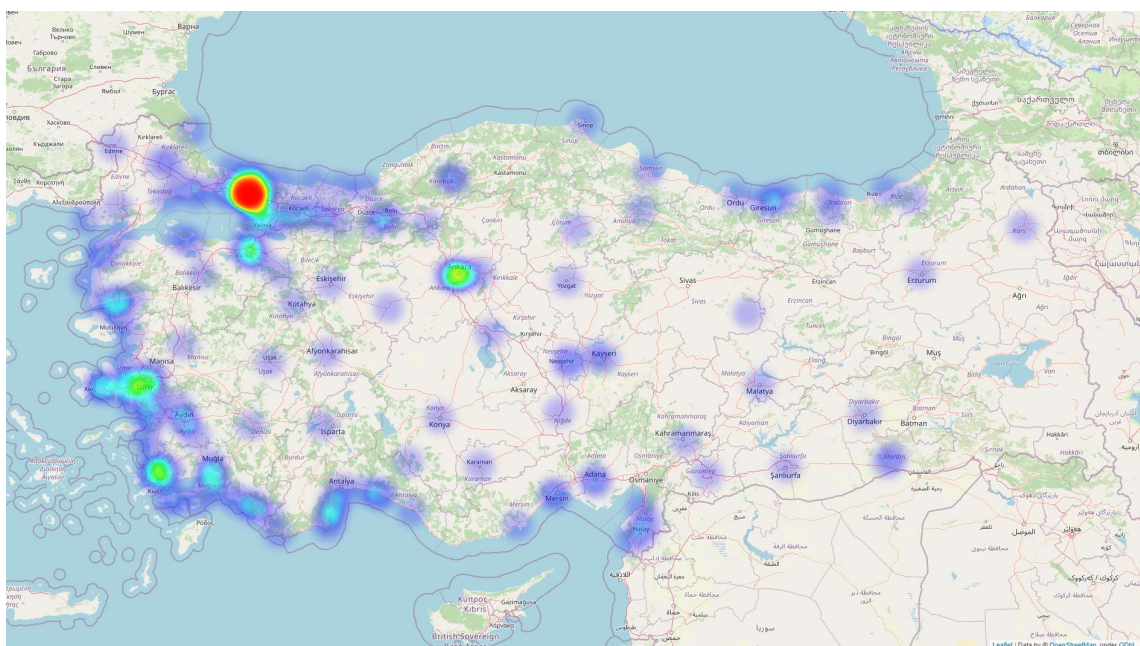


Figure 3.24: User Review Distribution in 2012

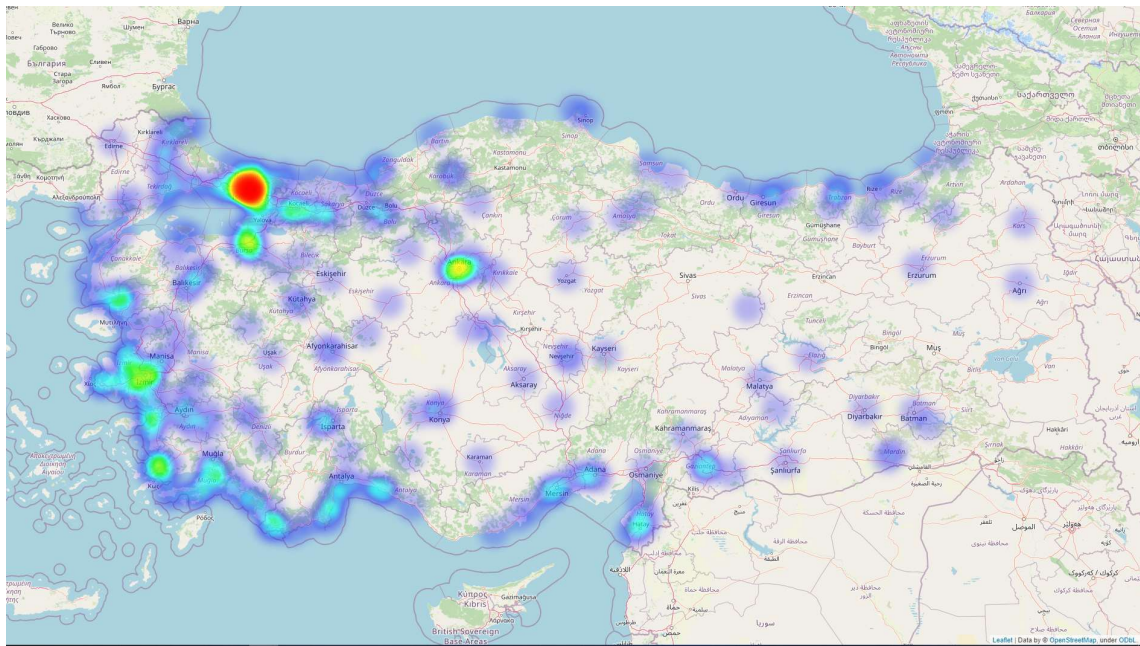


Figure 3.25: User Review Distribution in 2013

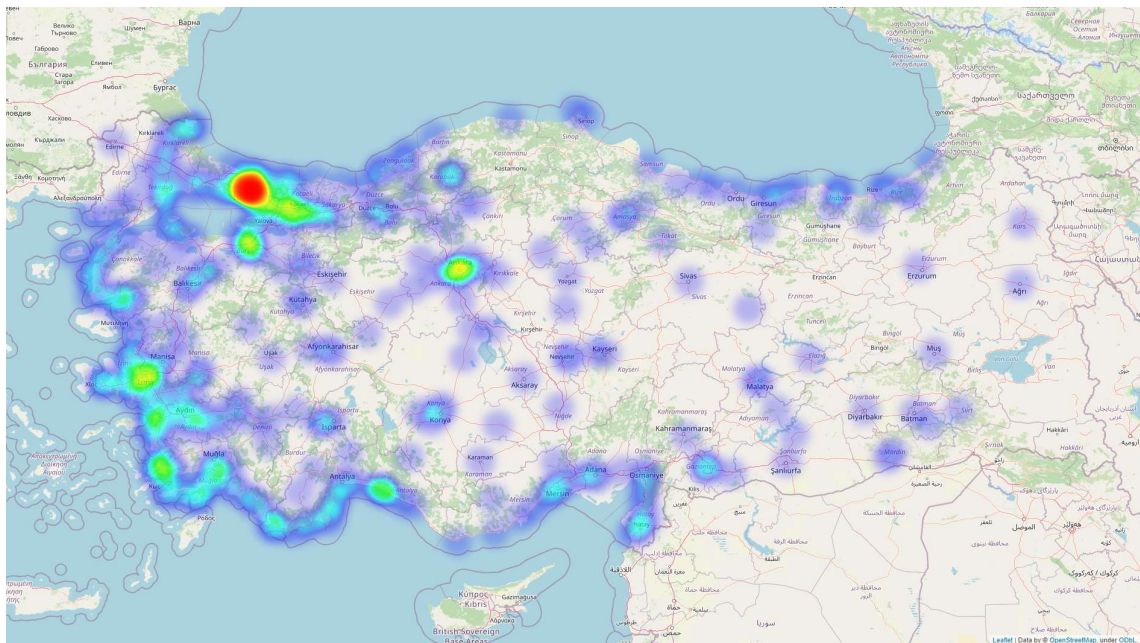


Figure 3.26: User Review Distribution in 2014

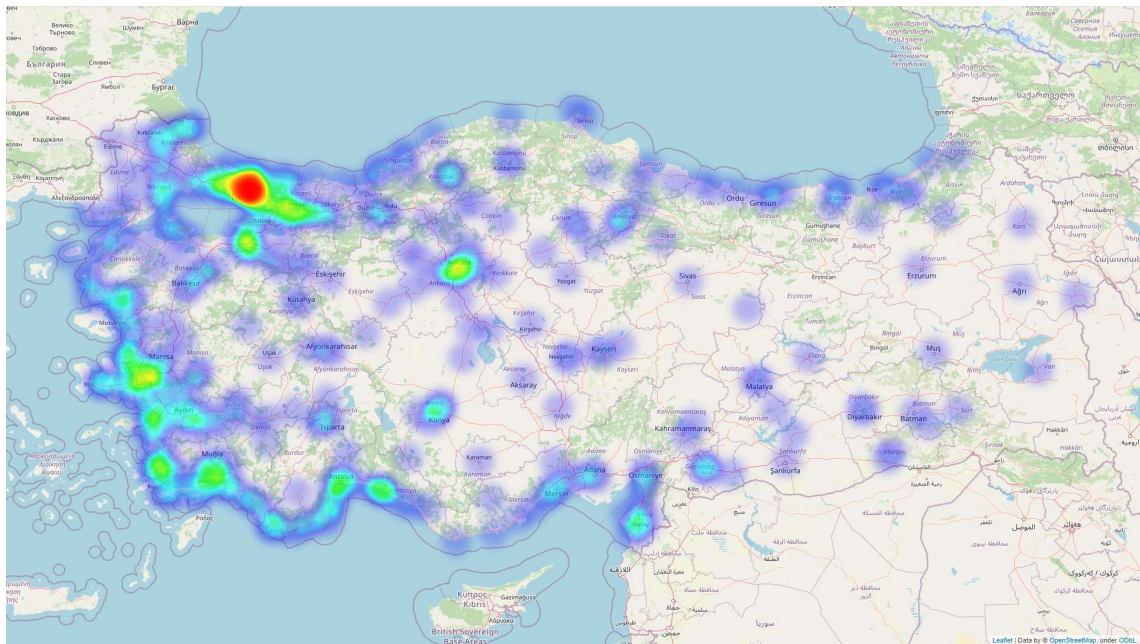


Figure 3.27: User Review Distribution in 2015

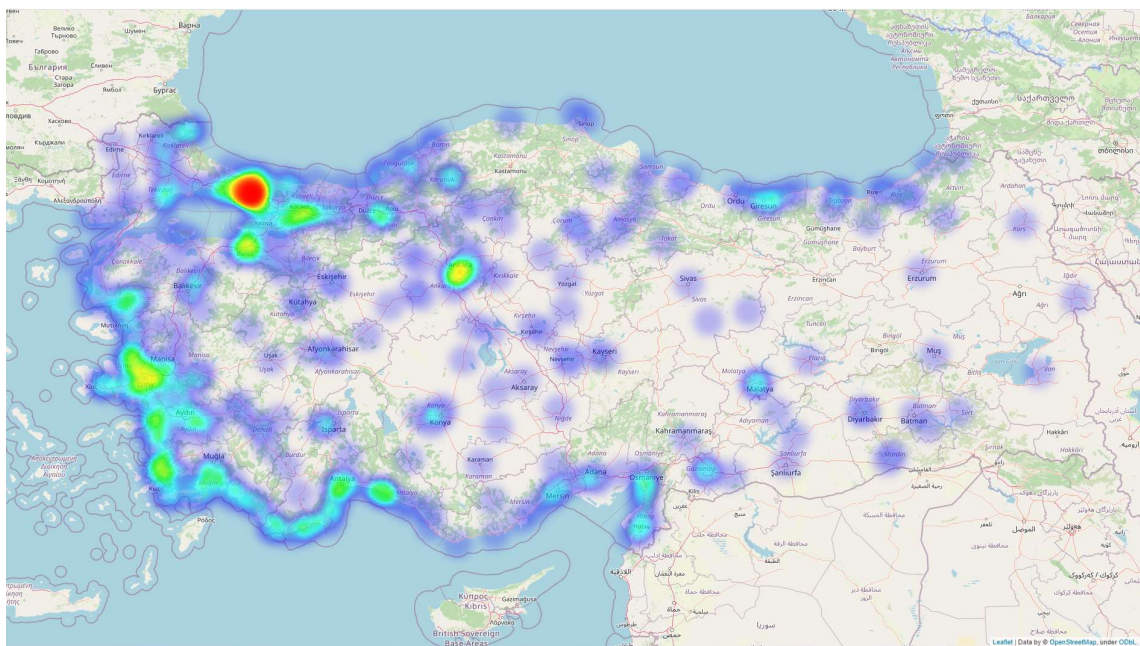


Figure 3.28: User Review Distribution in 2016

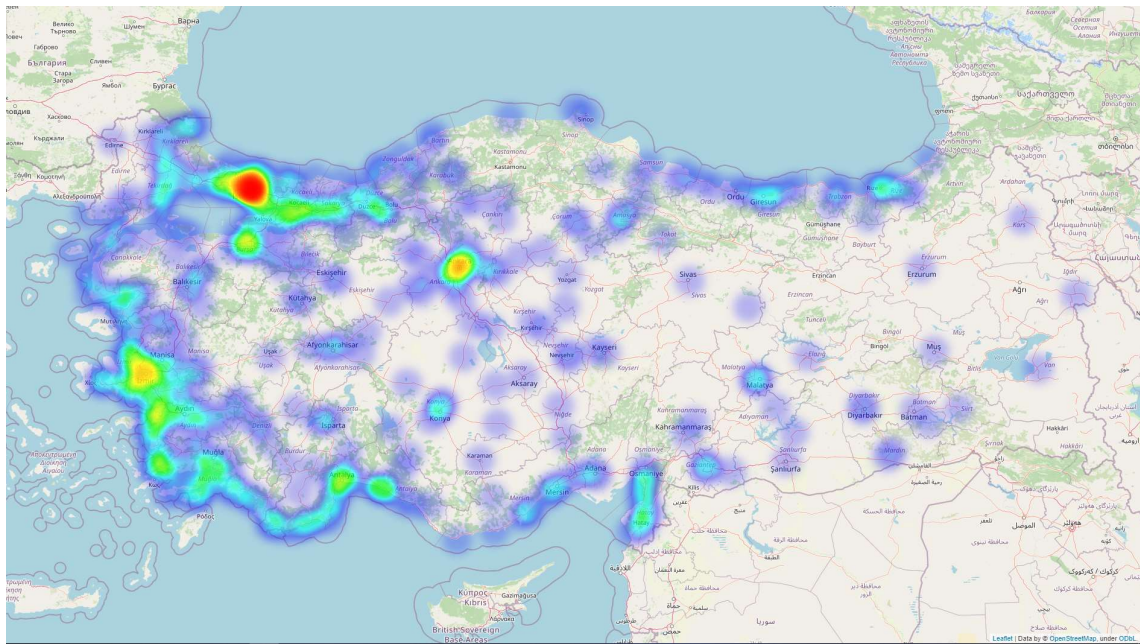


Figure 3.29: User Review Distribution in 2017

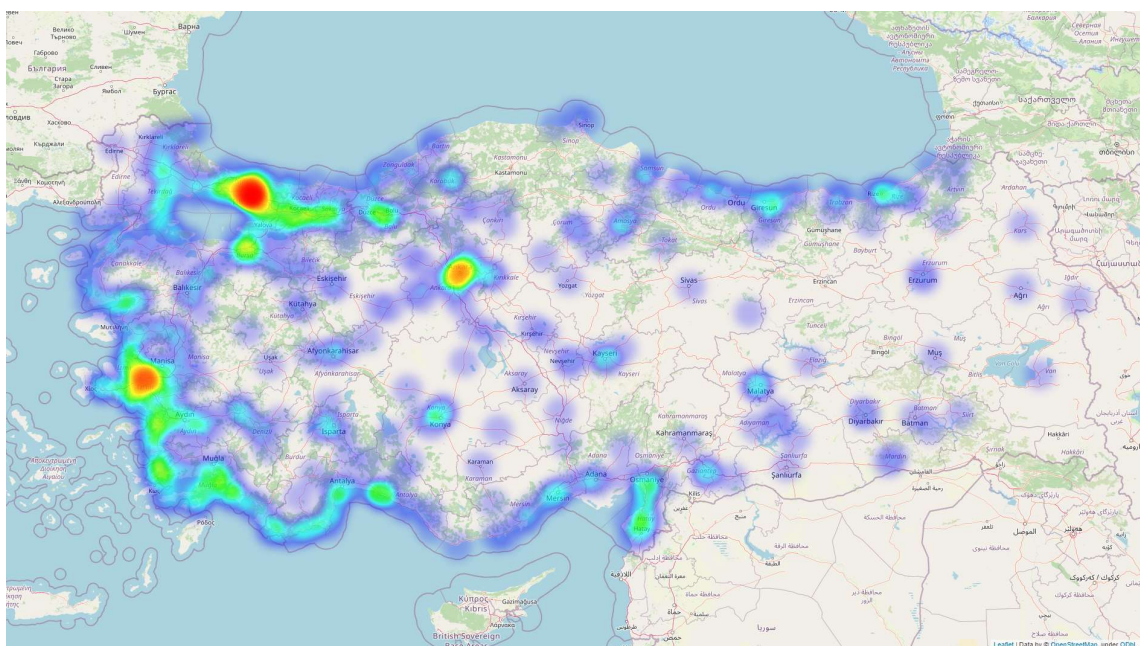


Figure 3.30: User Review Distribution in 2018

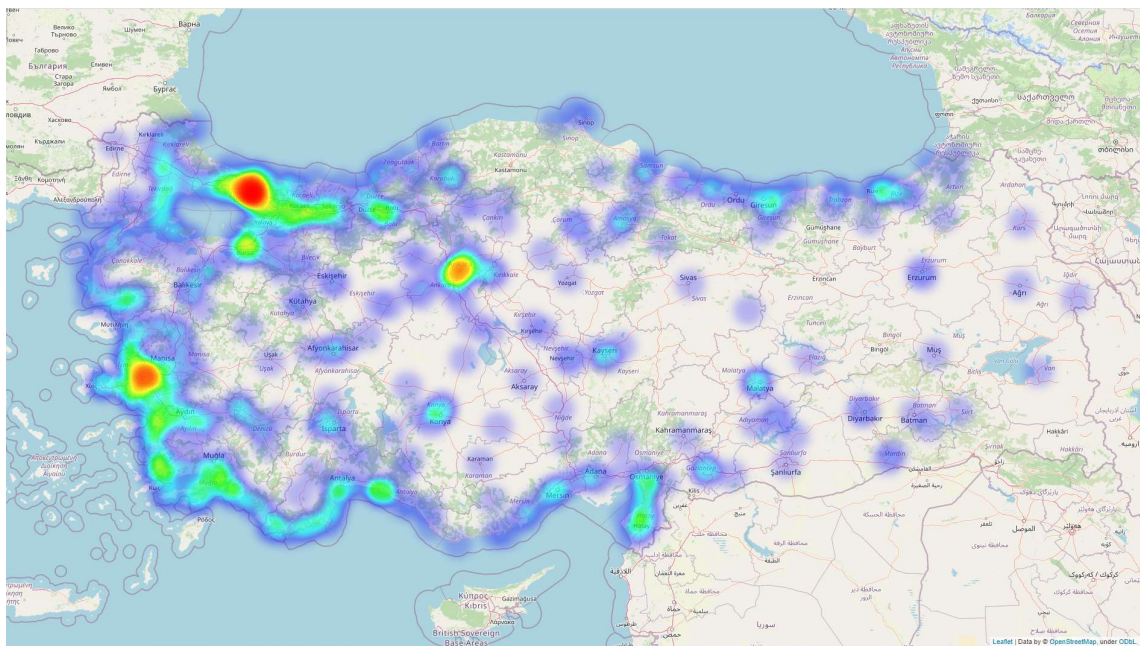


Figure 3.31: User Review Distribution in 2019

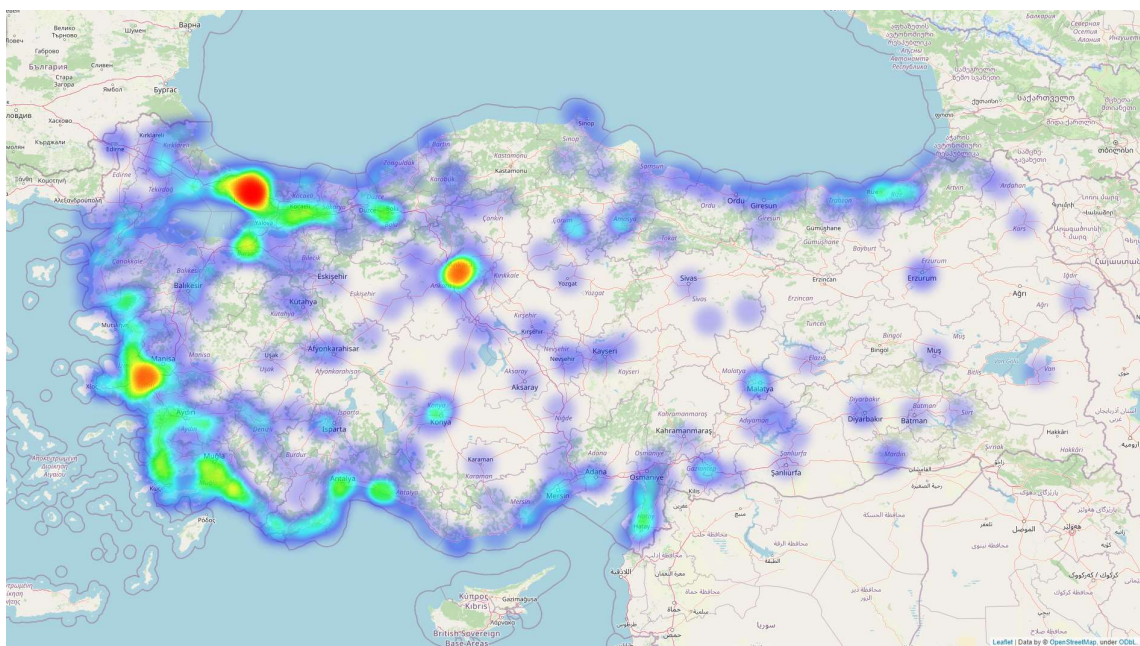


Figure 3.32: User Review Distribution in 2020

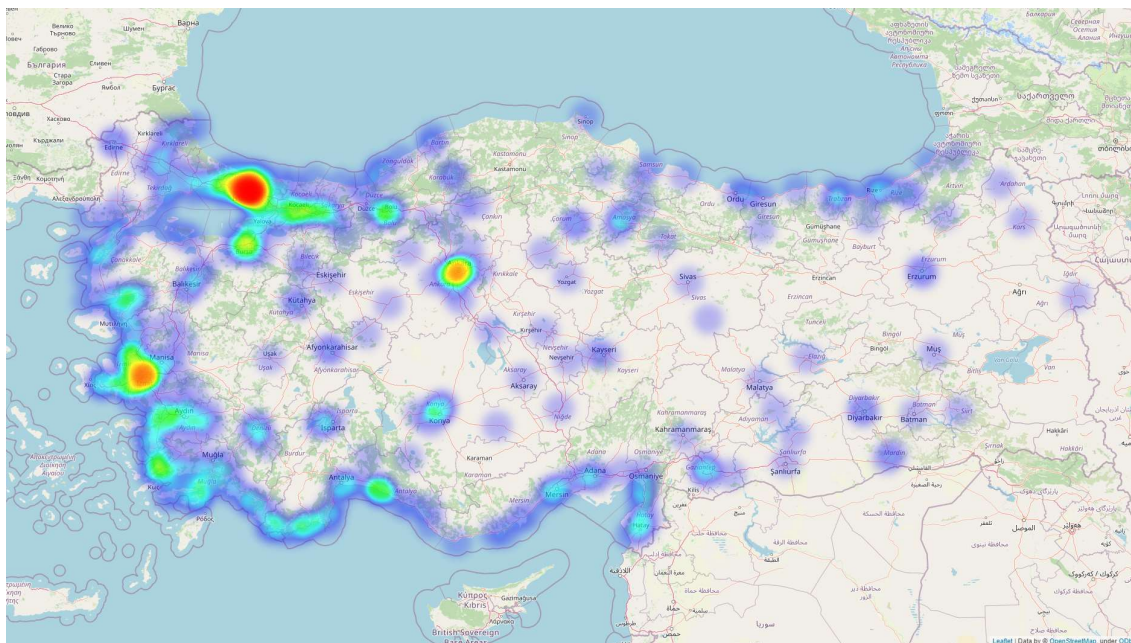


Figure 3.33: User Review Distribution in 2021

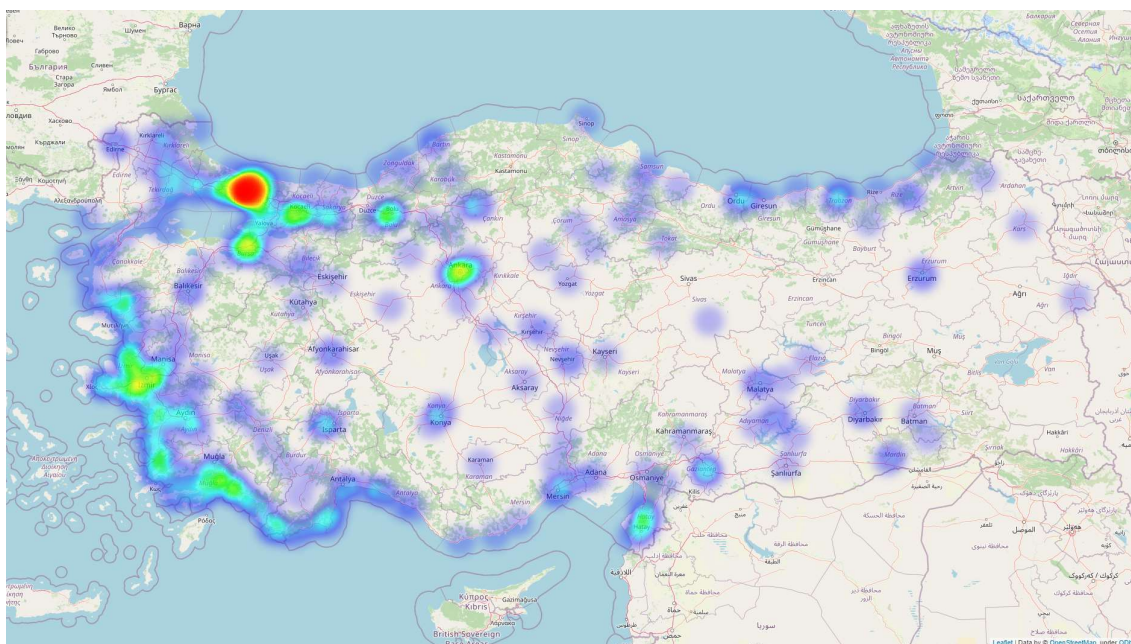
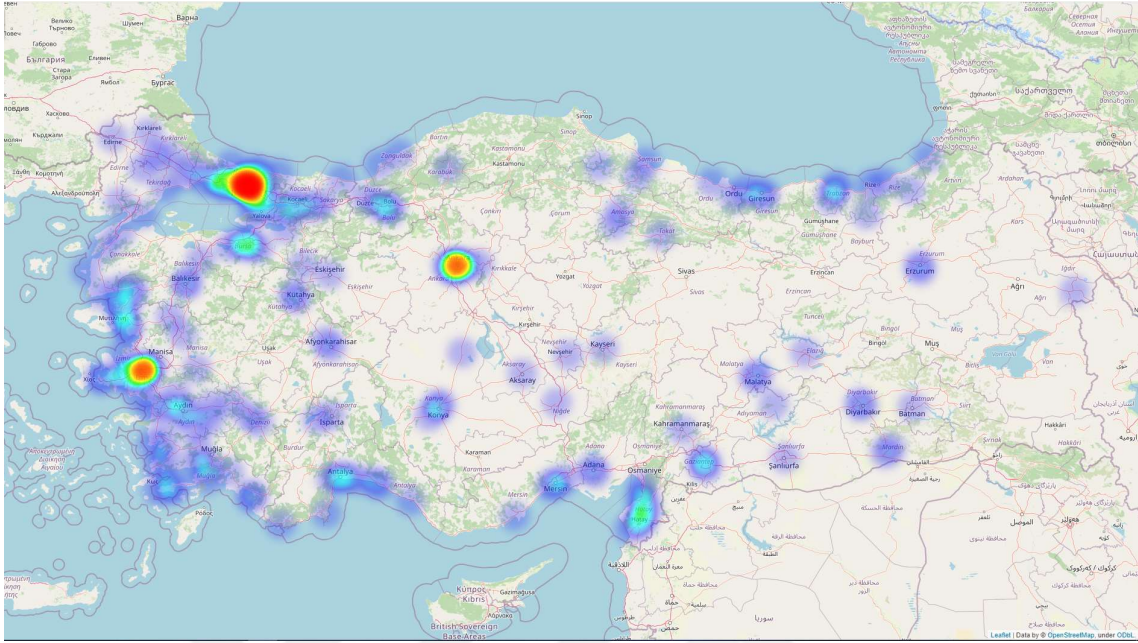


Figure 3.34: User Review Distribution in 2022



3.4 Sentiment Analysis

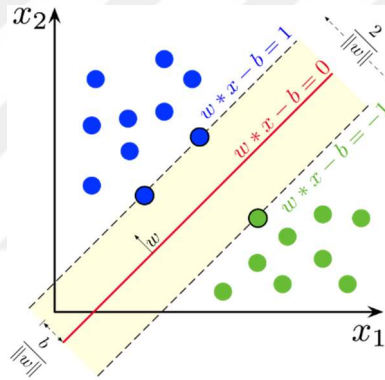
In this section, sentiment analysis results will be discussed through classical machine learning and deep learning models applied onto the Turkish dataset. Sentiment analysis, also known as opinion mining, is a subfield of natural language processing (NLP) that focuses on text analysis and computational techniques to extract, quantify, and examine the emotional states present within a given piece of information or text dataset.

For the purpose of comparative analysis with other methods, we have opted to utilize classical machine learning algorithms such as Logistic Regression, Linear SVC, Decision Tree, Random Forest, and Multinomial NB, as well as deep learning models such as LSTM and Bi-LSTM, as baselines. Furthermore, in order to achieve the highest level of accuracy, the state-of-the-art transformer-based language models have been implemented. These models were trained on 80% of the dataset and evaluated with the remaining 20%.

3.4.1 Models

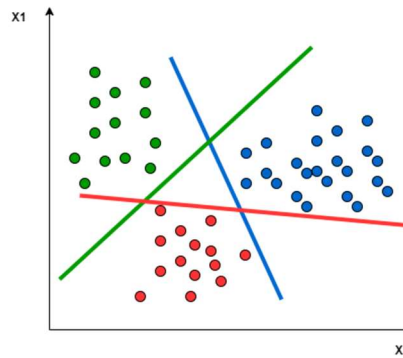
Support Vector Machine (SVM) is a powerful supervised machine learning algorithm utilized for classification and regression analysis. The primary goal of the SVM algorithm is to discover the optimal hyperplane in an N-dimensional space, which separates the data points into different classes. In essence, SVM performs complex data transformations based on the selected kernel function, aiming to maximize the separation boundaries between data points. The data points located at the minimum distance to the hyperplane are called Support Vectors, which determine the hyperplane's position and orientation (Figure 3.35).

Figure 3.35: Optimal Hyperplane and margins for an SVM trained with two classes



SVM is primarily designed for binary classification problems and does not directly support multiclass classification. However, several strategies exist to use SVM for multiclass classification (Figure 3.36) by transforming the problem into multiple binary classification tasks. One common approach is the one-vs-all (OVA) method, where a separate binary SVM classifier is trained for each class against all the other classes. In contrast, the one-vs-one (OVO) method trains binary classifiers for each pair of classes and then combines their predictions to determine the final class. Another method is the error-correcting output codes (ECOC), which encode the class labels into a set of binary codes and then train multiple binary classifiers to recognize each code. These binary classifiers are then combined to make multiclass predictions. In this thesis, we used the Stochastic Gradient Descent (SGD) Classifier, which is an optimization algorithm designed to determine the parameter values of a function that minimizes a cost function.

Figure 3.36: Multiclass classification using SVM



Naive Bayes (NB) is a popular and effective supervised learning algorithm based on Bayes' Theorem, stated mathematically as the following equation and commonly used for text classification problems where the number of features (e.g., words in a document) is a high-dimensional training dataset. On the other hand, it is a probabilistic classifier that predicts the class label of an object based on the probability of its features. The Naive Bayes classifier is known for its simplicity and efficiency in building classification models. It is highly scalable and can handle large datasets with many features. It is a fast algorithm that can quickly make predictions on new data.

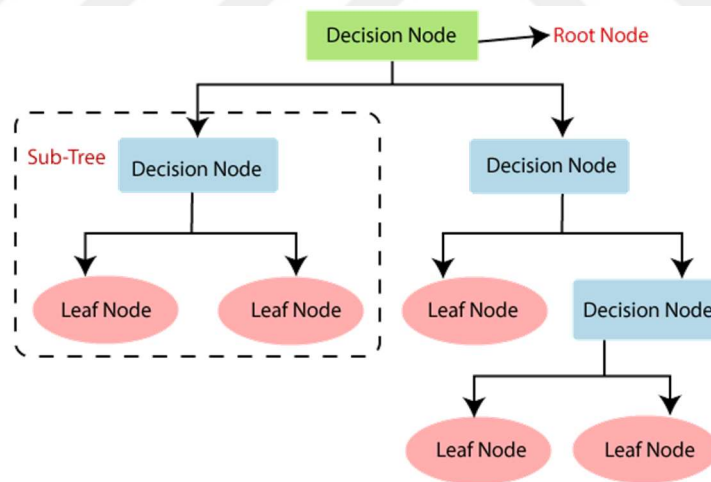
The method employed in this thesis, Multinomial Naive Bayes, models the distribution of words in a document as a polynomial. The document is treated as a sequence of words, and it assumes that the occurrence of each word is independent of the others. Consequently, it computes the probability of each label for a given sample and outputs the label with the highest probability.

Logistic Regression is a widely used supervised machine learning algorithm for binary classification tasks. It models the probability of a binary response variable as a function of one or more independent variables. The algorithm finds the coefficients that minimize the difference between predicted probabilities and actual binary outcomes. It is called "logistic" because it uses a logistic function to transform the output of a linear regression into a probability value between 0 and 1. This probability is then used to make a binary classification decision.

In this thesis, we have employed the Multinomial Logistic Regression method to estimate categorical assignment or the likelihood of category membership for a dependent variable, taking into account numerous independent variables. Specifically, we have utilized this approach to conduct multiclass classification.

Decision Tree algorithm is a supervised machine learning technique that can be utilized for both classification and regression tasks. It is characterized by a tree-like structure in which internal nodes represent the features of a dataset, branches indicate the decision rules, and each leaf node corresponds to the outcome. The two main types of nodes in a Decision Tree are the Decision Node and Leaf Node. Decision nodes enable decision-making by branching into multiple branches, whereas leaf nodes represent the output of those decisions and do not have any other branches. The diagram below illustrates the general structure of a decision tree. Figure 3.37 provides general structure of a Decision Tree algorithm.

Figure 3.37: General structure of a Decision Tree

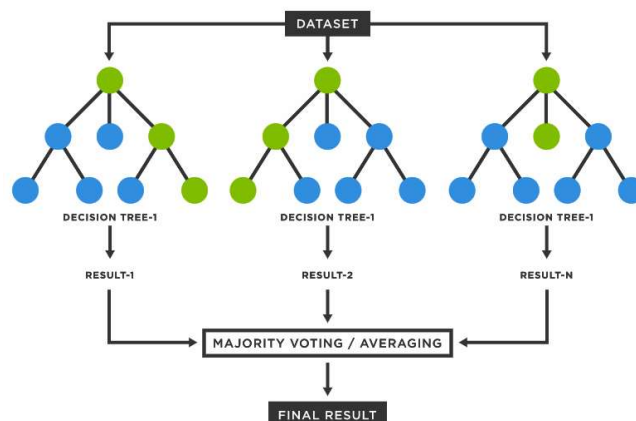


To predict the class of a given dataset, the algorithm starts from the tree's root node and compares the values of the root attribute with the corresponding record attribute. Based on this comparison, it follows a branch and jumps to the next node. The algorithm repeats this process by comparing the attribute value with the sub-nodes for the next node until it reaches the leaf node of the tree. This decision-making process is similar to the way humans make decisions in real life, making the algorithm easy to understand and providing a robust and fast way to analyze data structure. However, a complex tree structure with many branches can lead to longer processing times.

Random Forest algorithm is a powerful ensemble learning method commonly used in classification and regression problems. It combines multiple decision trees to create a single model, resulting in a more accurate and robust model than a single decision tree. The name "Random Forest" comes from the fact that it grows a forest of decision trees. During training, each decision tree in the forest is trained on a randomly selected subset of the training data and a randomly selected subset of the features. This helps to reduce overfitting and improve the model's generalization performance. When predicting a new data point, the Random Forest algorithm aggregates the predictions of all the decision trees in the ensemble. For classification tasks, it assigns the class label that receives the most votes from the decision trees. For regression tasks, it takes the average of the numerical values predicted by the decision trees. This approach provides a more robust and accurate prediction than a single decision tree and can handle complex datasets with high-dimensional input features.

Random Forest algorithm possesses several desirable features such as its capacity to effectively manage high-dimensional data, its ability to resist overfitting, and its robustness when processing noisy data. Furthermore, this algorithm is versatile in its ability to accommodate various data types, including both categorical and numerical data. However, a drawback of the Random Forest algorithm is its interpretability, which may present challenges due to the model's composition of multiple decision trees. As a result, comprehending the fundamental relationships between the target variable and its features can be challenging. Figure 3.38 provides a visual representation of the Random Forest algorithm's workflow.

Figure 3.38: Random Forest working diagram



Adaptive Boosting (AdaBoost) is an ensemble learning technique that has gained popularity for classification and regression tasks. The approach employed by AdaBoost entails amalgamating several weak learners to produce a powerful classifier or regressor.

The fundamental concept behind AdaBoost involves iteratively training a sequence of weak learners, each focusing on the examples that previous weak learners misclassified. AdaBoost assigns a higher weight to the misclassified examples in each iteration to enhance their selection by the following weak learner.

At the end of the training process, AdaBoost integrates the sequence of weak learners into a robust classifier or regressor by allocating a weight to each weak learner based on its performance on the training data. Then, the assigned weights are utilized to combine the predictions of the weak learners into a final prediction.

K-Nearest Neighbor (KNN) is a non-parametric machine learning algorithm prevalent in classification and regression tasks. KNN is classified as non-parametric since it makes no assumptions about the underlying data distribution. The operation of KNN entails identifying the nearest neighbors to a specific data point using a similarity metric, usually Euclidean distance. The algorithm then classifies or predicts the value of the data point based on the majority class or the average value of the K nearest neighbors. The selection of K is often determined using cross-validation or other hyperparameter tuning techniques.

KNN is an uncomplicated and intuitive algorithm that is simple to comprehend and apply. It performs well for small to medium-sized datasets with low dimensionality and is robust to noise and outliers in the data. However, KNN may be computationally intensive and memory-intensive for large datasets, and it may need help in high-dimensional spaces.

Long Short-Term Memory (LSTM) is a type of deep learning model used for sequence modeling and prediction tasks. It is a variation of the recurrent neural network (RNN) architecture designed to address the problem of vanishing gradients, which can occur when training deep neural networks. (Hochreiter and Schmidhuber, 1997)

LSTM networks are composed of recurrent cells with a more complex structure than those in a traditional RNN. These cells have three gates - an input gate, an output gate, and a forget gate - that allow them to selectively read, write, and erase information from their internal state over time. This situation makes them well-suited for tasks with essential long-term dependencies between input sequences. LSTM networks have been successfully applied to a wide range of tasks, including natural language processing, speech recognition, image captioning, and video analysis. They are also commonly used in applications that require time-series prediction, such as stock market forecasting, weather prediction, and energy consumption forecasting.

Bidirectional Long Short-Term Memory (Bi-LSTM) is a type of recurrent neural network (RNN) that employs two layers of LSTM cells to process input sequences in both forward and backward directions. This approach enables the network to capture not only the temporal dependencies in the forward direction but also those in the backward direction, making it particularly suitable for tasks that require context awareness.

In a conventional LSTM network, the input sequence is processed sequentially from left to right, and the output is generated at the end of the sequence. By contrast, in a Bi-LSTM network, the input sequence is processed in both directions, and the output is obtained by combining the outputs from the two layers of LSTM cells. At each time step, the forward and backward layers of the Bi-LSTM network receive the same input but process it in opposite directions. The outputs from both layers are merged to produce the final output for that time step. This enables the network to capture both short-term and long-term dependencies in the input sequence, regardless of their direction. Besides, Bi-LSTM is proficient in augmenting the volume of data accessible to the network and enhancing the contextual information accessible to the algorithm by including information about the adjacent words that precede and follow a specific word within a sentence.

It has been used in various applications, including speech recognition, natural language processing, and video analysis. They are particularly useful for tasks such as named entity recognition and sentiment analysis, where the surrounding context heavily influences the meaning of a word or phrase.

BERT (Bidirectional Encoder Representations from Transformers), a pre-trained bidirectional deep natural language processing model developed by Google researchers in 2018, leverages both left and right context in all layers to encode the meaning of the unlabeled text. (Devlin et al., 2019). BERT is considered a powerful model, available in two versions: BERT Large with 24 transformer blocks, 1024 hidden layers, and 340M parameters, and BERT Base with 12 transformer blocks, 768 hidden layers, and 110M parameters. These models are pre-trained on large amounts of text data and can be fine-tuned for various natural language processing tasks. During the development and research stages, BERT achieved remarkable performance in 11 natural language understanding tasks, including sentiment analysis, semantic role labeling, sentence classification, and disambiguation of polysemous words. Some brief information regarding BERT models given below.

BERT Base Cased is a case-sensitive pre-trained natural language processing model that forms the base architecture of BERT and designed to process English text data

BERT Base Multilingual Cased is a case-sensitive pre-trained model within the BERT architecture designed to work with multiple languages and has been pre-trained on a large corpus of text data from 104 different languages.

BERT Base Turkish Cased is a case-sensitive pre-trained BERT model specifically designed for the Turkish language. This model is trained on a filtered and sentence-segmented version of the Turkish OSCAR corpus, a recent Wikipedia dump, and a special corpus provided by Kemal Oflazer.

BERT Base Turkish Uncased has been trained on the same dataset as BERT Base Turkish Cased, however, it is not sensitive to letter case.

BERT Base Turkish 128k Cased has been trained on a larger dataset of Turkish language compared to the previous model, and it is sensitive to letter case.

DistilBERT Base Turkish Cased is a variant of the DistilBERT model, which is a more compact and efficient version of the BERT architecture. It is specifically trained on 7GB of the original training data that was used for training BERTurk, a BERT model specifically designed for the Turkish language.

ConvBERT Base Turkish Cased is a model that utilizes a span level dynamic convolution technique to enhance BERT performance. It has been trained on the same dataset used to train Bert Base Turkish Cased.

ConvBERT Base Turkish mc4 Uncased is a model that is based on the ConvBERT architecture and has been trained on the recently published Turkish portion of the multilingual C4 corpus, which was compiled by the AI2 team.

3.4.2 Evaluation

This thesis employs various performance measures to assess the categorization of Foursquare tips as negative, neutral, or positive. Some of the metrics utilized in this research are delineated as follows.

True Positive (TP) refers to the situation where a model correctly predicts the presence of a particular label in a data point. For example, if the actual user tip class value is Neutral and the model prediction is also neutral, it is a true positive (TP) scenario for the Neutral class. This means the model correctly identified the Neutral class for the given user tip.

False Positive (FP) refers to the situation where a model predicts the presence of a label in a data point, but in reality, the data point does not have that label. For example, suppose the model predicts the presence of a Neutral class for a particular text, but in reality, the text does not contain the Neutral class. Then, the model made a false positive (FP) prediction for the Neutral class.

True Negative (TN) refers to the situation where a model correctly predicts the absence of a particular class in a data point. For example, suppose the model predicts the absence of a Neutral class for a particular text; in reality, the text does not contain a Neutral class. Then, the model made a true negative (TN) prediction for the Neutral class.

False Negative (FN) refers to the situation where a model predicts the absence of a class in a data point, but the data point does contain that class. For example, Suppose the model predicts the absence of a Neutral class for a particular text, but in reality, the text does contain a Neutral class; it is a false negative scenario for the Neutral class.

Accuracy is a commonly used performance metric that assesses the quality and overall correctness of a classification model. Formula 3.1 shows the calculation of the classification accuracy metric, which is obtained by dividing the number of correctly classified examples by the total number of classified examples. However, it may be misleading in certain situations, particularly in imbalanced datasets where one class has significantly fewer instances than the others. When dealing with certain scenarios, alternative performance metrics like precision, recall, and F-score may offer a more precise and accurate evaluation of the model's performance.

$$\text{Accuracy} = \frac{\text{Number of correctly classified tips}}{\text{Total number of classified tips}} = \frac{(TN + TP)}{(TP + FN + FP + TN)} \quad (3.1)$$

Recall, also known as Sensitivity or Probability of Detection, is a performance measure utilized in binary and multi-class classification tasks to assess a model's capacity to identify all pertinent cases among all relevant examples. It is computed by determining the proportion of examples from a specific class that the model has accurately classified as belonging to that class. For instance, recall quantifies the number of user reviews correctly predicted as positive from the total number of positive user reviews. This metric can be calculated using the formula 3.2, and in this thesis, it was computed separately for three distinct classes.

$$\text{Recall} = \frac{(\text{TP})}{(\text{TP} + \text{FN})} \quad (3.2)$$

Precision, also known as Positive Predictive Value, is a performance measure used in binary and multi-class classification tasks to assess a model's capability to identify all relevant cases among the retrieved examples. It is calculated by determining the proportion of accurate predictions among all predictions made for a specific class. For instance, precision calculates the total number of user reviews that were correctly predicted as positive from among the total number of user reviews predicted as positive. This metric can be calculated using the formula 3.3, and in this thesis, it was computed separately for three distinct classes.

$$\text{Precision} = \frac{(\text{TP})}{(\text{TP} + \text{FP})} \quad (3.3)$$

F-score or F-measure is a performance metric utilized in binary and multi-class classification tasks to evaluate the overall accuracy of a model. Formula 3.4 shows the calculation of the classification F-score metric, which is computed as the harmonic mean of the precision and recall, and it accounts for both false positives and false negatives. In this thesis, the F-score of the classifiers was compared to determine which classifier performed better. This metric can be calculated using the formula below, and in this thesis, it was computed separately for three distinct classes.

$$\text{F-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (3.4)$$

The Receiver Operating Characteristics (ROC) curve and the Area Under the Curve (AUC) score are commonly used performance metrics for classification problems across various threshold settings. The ROC curve is a graphical representation of the binary classifier system's diagnostic ability, plotted by depicting the true positive rate (TPR) against the false positive rate (FPR) along a probability curve as the discrimination threshold varies. The AUC score is a quantification of a classifier's discriminative power between classes. In this thesis, ROC curves were generated using the "one class versus the rest" methodology for multiclass classification, resulting in individual curves for each class. Consequently, the AUC score can also be computed individually for each class. Therefore, a higher AUC indicates a better predictive performance of the model for negative, neutral, and positive classes.

Evaluation Results

In order to evaluate the performance of the models, a range of performance metrics were employed, which included Accuracy, Precision, Recall, and F-scores for the negative, neutral, and positive classes. Receiver Operating Characteristic (ROC) curves were also generated, and the corresponding Area Under Curve (AUC) scores were calculated to compare and contrast the models based on their true positive and false positive rates. The results obtained from the models trained on both the balanced dataset and the Automatic Annotated Corpus, wherein an equivalent number of samples were selected from each class, are presented below.

Table 3.15 displays the performance metrics of all machine learning models trained on the balanced AAC dataset, ranked according to their accuracy values. The Logistic Regression model recorded the highest accuracy value, while the following four models obtained comparatively lower accuracies. In contrast, the Decision Tree and K-Nearest Neighbor models performed the worst. ROC curves and AUC scores were employed to evaluate the performance of each model for individual classes, and the results from Figures 3.39, 3.40, and 3.41 indicate the need for improvements in some of the models.

Table 3.15: Evaluation results for Machine Learning Models

Machine Learning Model	Negative				Neutral				Positive				Accuracy
	Precision	Recall	F-Score	Support	Precision	Recall	F-Score	Support	Precision	Recall	F-Score	Support	
Logistic Regression	86	85	85	13196	78	79	78	13402	88	88	88	13402	84
Adaptive Boosting	84	76	80	13196	68	78	73	13402	86	82	84	13402	79
Multinomial Naive Bayes	79	85	82	13196	73	59	65	13402	78	87	82	13402	77
Linear SGD	78	78	78	13196	67	71	69	13402	83	77	80	13402	75
Random Forest	84	55	66	13196	57	77	65	13402	79	78	78	13402	70
Decision Tree	82	22	35	13196	40	89	55	13402	82	42	55	13402	51
K-Nearest Neighbor	65	15	24	13196	37	86	51	13402	74	29	42	13402	44

Figure 3.39: ROC curve for Negative Class with Machine Learning Models

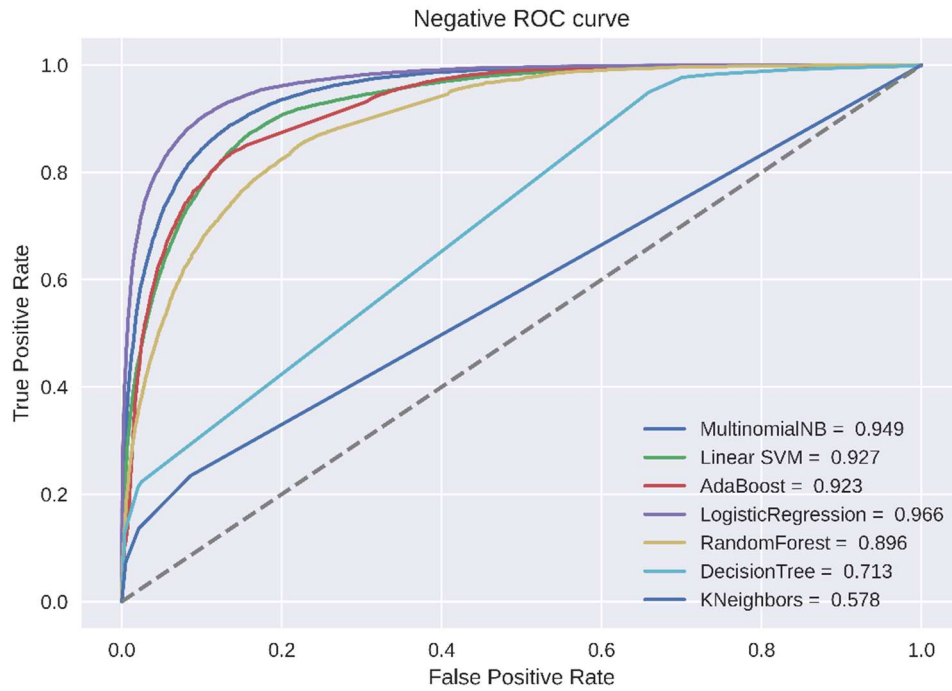


Figure 3.40: ROC curve for Neutral Class with Machine Learning Models

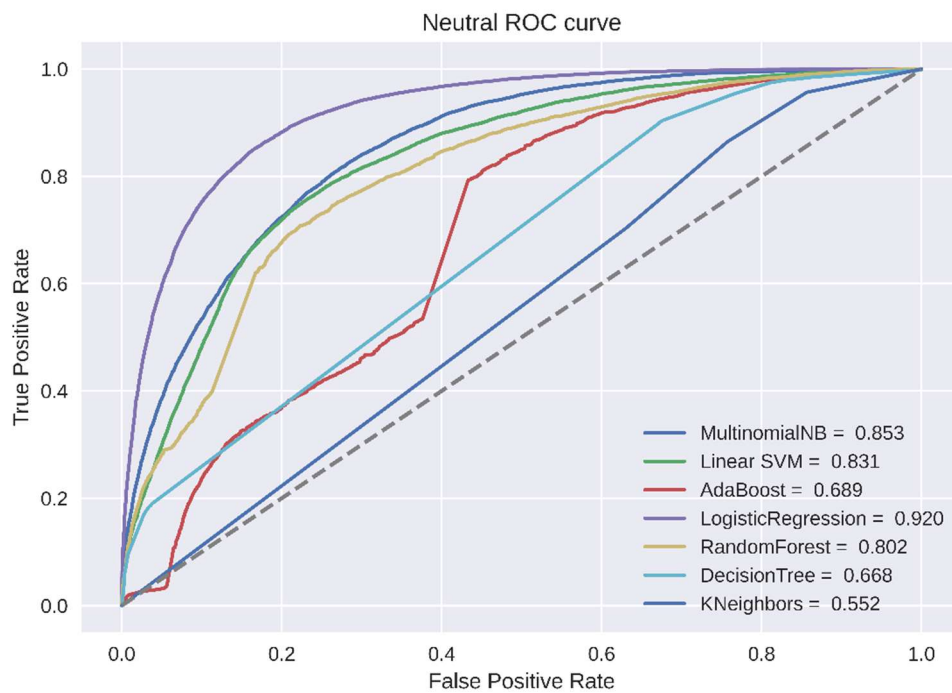


Figure 3.41: ROC curve for Positive Class with Machine Learning Models

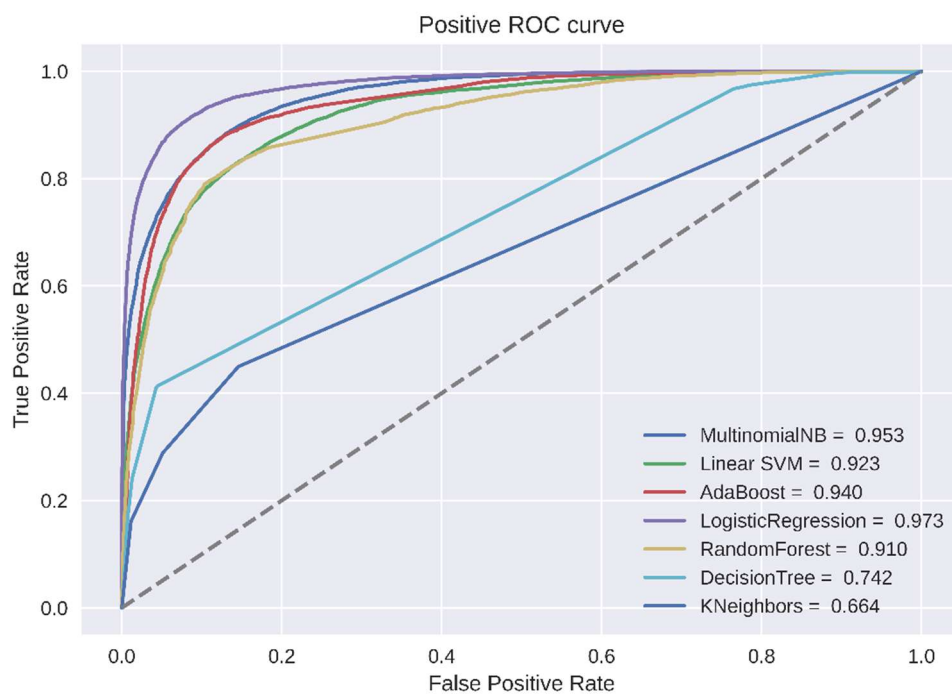


Table 3.16 illustrates the performance metrics of the Dense, LSTM, and Bi-LSTM deep learning models trained on the balanced AAC dataset, with accuracy values as the basis for ranking. When the LSTM and Bi-LSTM models exhibited the highest accuracy values, which were very close to each other, the Dense model obtained comparatively lower accuracies. ROC curves and AUC scores were utilized to assess the performance of each model for individual classes, and Figures 3.42, 3.43, and 3.44 present the results.

Table 3.16: Evaluation results for Deep Learning Models

Deep Learning Model	Negative				Neutral				Positive				Accuracy
	Precision	Recall	F-Score	Support	Precision	Recall	F-Score	Support	Precision	Recall	F-Score	Support	
Bi-LSTM	84	81	83	13069	74	80	77	13466	88	83	85	13465	81
LSTM	84	82	83	13069	75	76	76	13466	85	86	85	13465	81
Dense Model	77	84	80	13069	72	74	73	13466	89	79	84	13465	79

Figure 3.42: ROC curve for Negative Class with Deep Learning Models

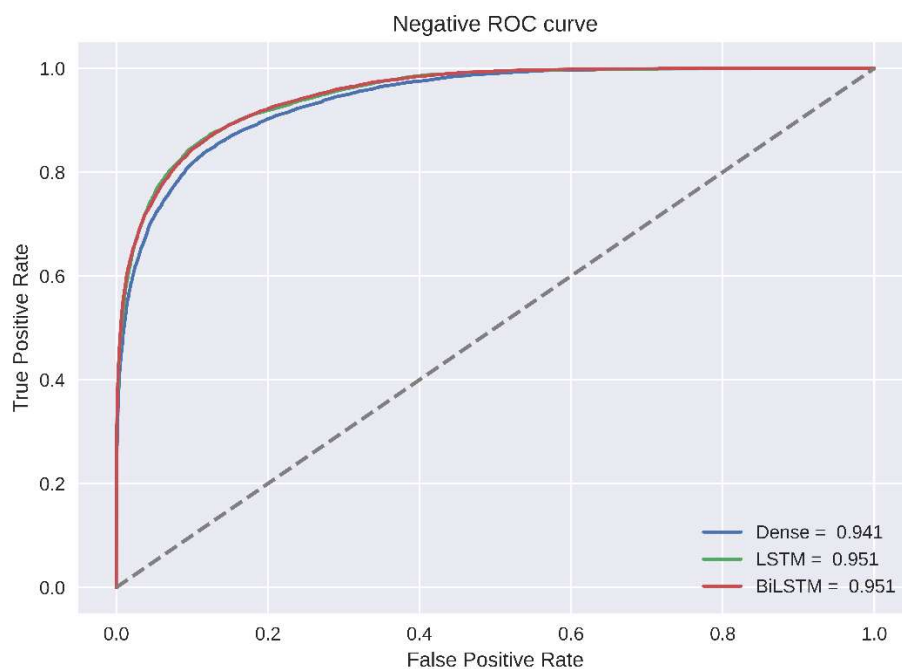


Figure 3.43: ROC curve for Neutral Class with Deep Learning Models

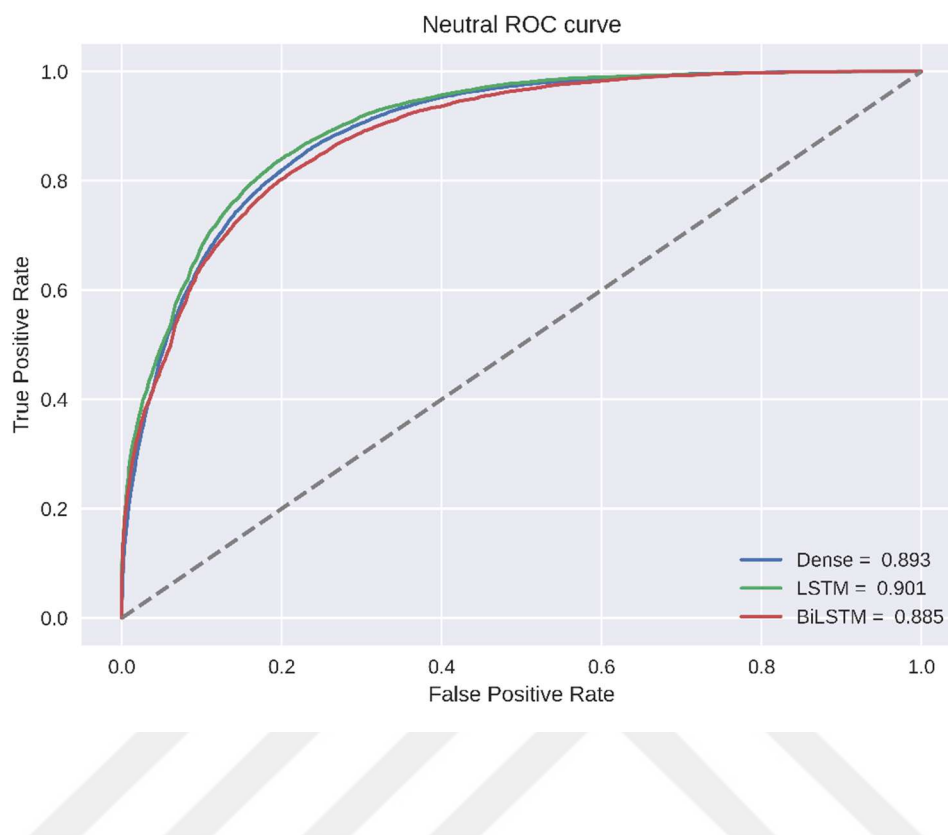


Figure 3.44: ROC curve for Positive Class with Deep Learning Models

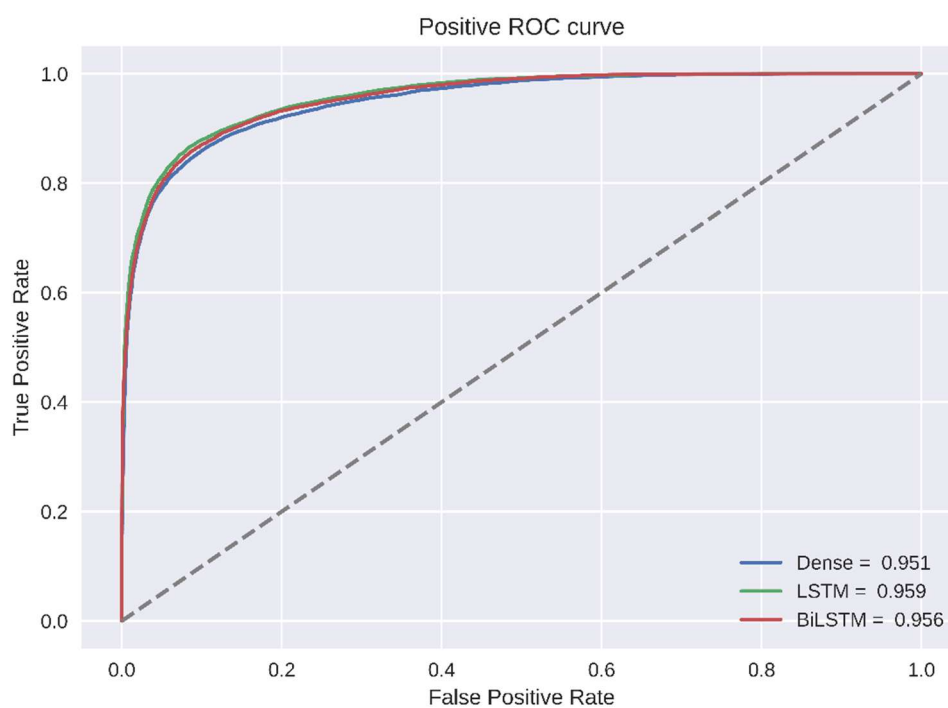


Table 3.17 presents the performance metrics of all pre-trained BERT models trained on the balanced AAC dataset, ranked according to their accuracy values. The overall accuracy performance of the models was high, with all models achieving an accuracy of over 90%. The model that achieved the highest accuracy was "BERT Base Turkish 128k Cased". ROC curves and AUC scores were utilized to assess the performance of each model for individual classes, and Figures 3.45, 3.46, and 3.47 present the results.

Table 3.17: Evaluation results for BERT Models

Pre-trained BERT Model	Negative				Neutral				Positive				Accuracy
	Precision	Recall	F1	Support	Precision	Recall	F1	Support	Precision	Recall	F1	Support	
BERT Base Turkish 128k Cased	93	94	94	13187	90	89	90	13260	95	95	95	13553	93
BERT Base Turkish Cased	93	93	93	13187	89	89	89	13260	94	95	94	13553	92
ConvBERT Base Turkish mc4 Uncase	93	94	94	13187	90	88	89	13260	94	95	94	13553	92
Electra Base Turkish Cased Disc.	93	94	94	13187	90	88	89	13260	93	95	94	13553	92
Electra Base Turkish mc4 Uncased D.	93	94	93	13187	90	88	89	13260	94	94	94	13553	92
DistilBERT Base Turkish Cased	93	92	88	13187	87	88	88	13260	94	93	93	13553	91
BERT Base Turkish 128k Uncased	91	93	92	13187	88	85	86	13260	93	93	93	13553	90
BERT Base Turkish Uncased	90	92	91	13187	87	85	86	13260	92	93	93	13553	90

Considering the classification results made with various models, the **BERT Base Turkish 128k Cased** exhibited the best results. Consequently, the BERT model demonstrates high accuracy in the Turkish dataset, as evidenced by the close and favorable outcomes obtained by all BERT models employed in this investigation. Following BERT's performance, the Logistic Regression machine learning model ranked second. Regrettably, deep learning models failed to reach the expected level of accuracy and lagged behind Logistic Regression.

Figure 3.45: ROC curve for Negative Class with BERT Models

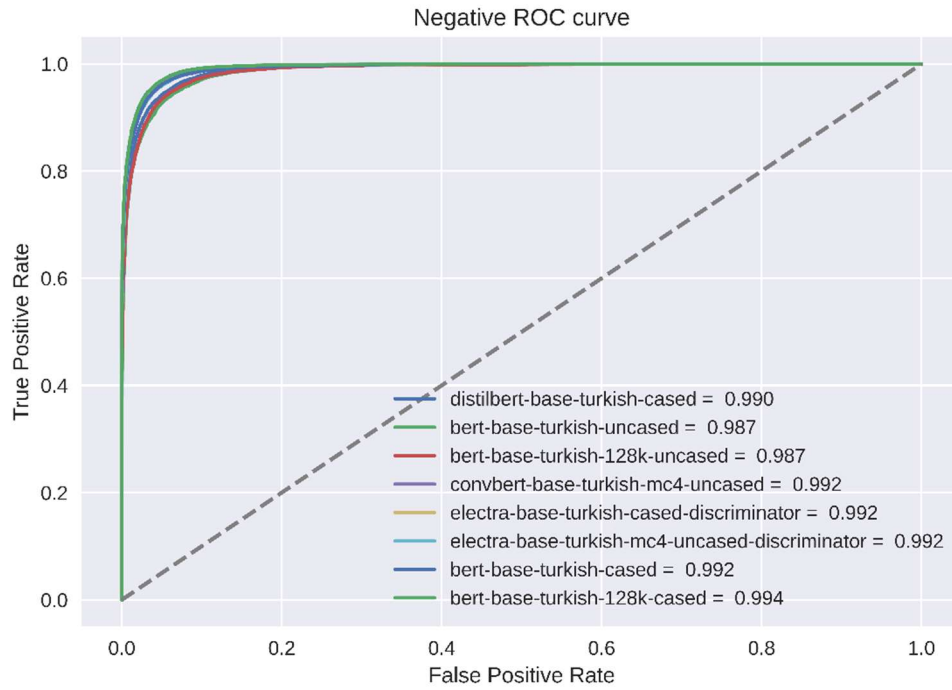


Figure 3.46: ROC curve for Neutral Class with BERT Models

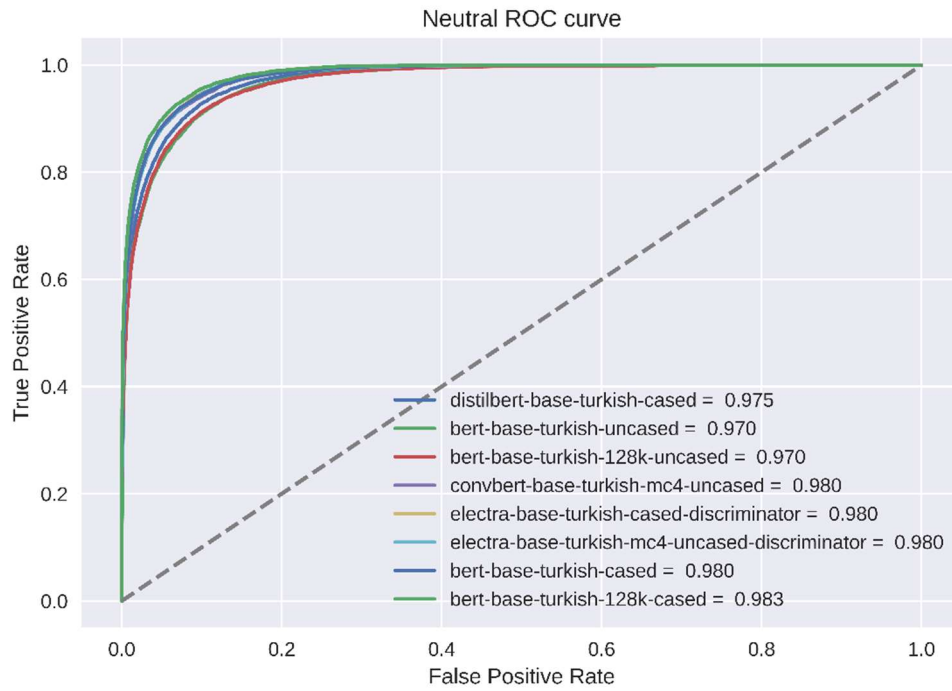
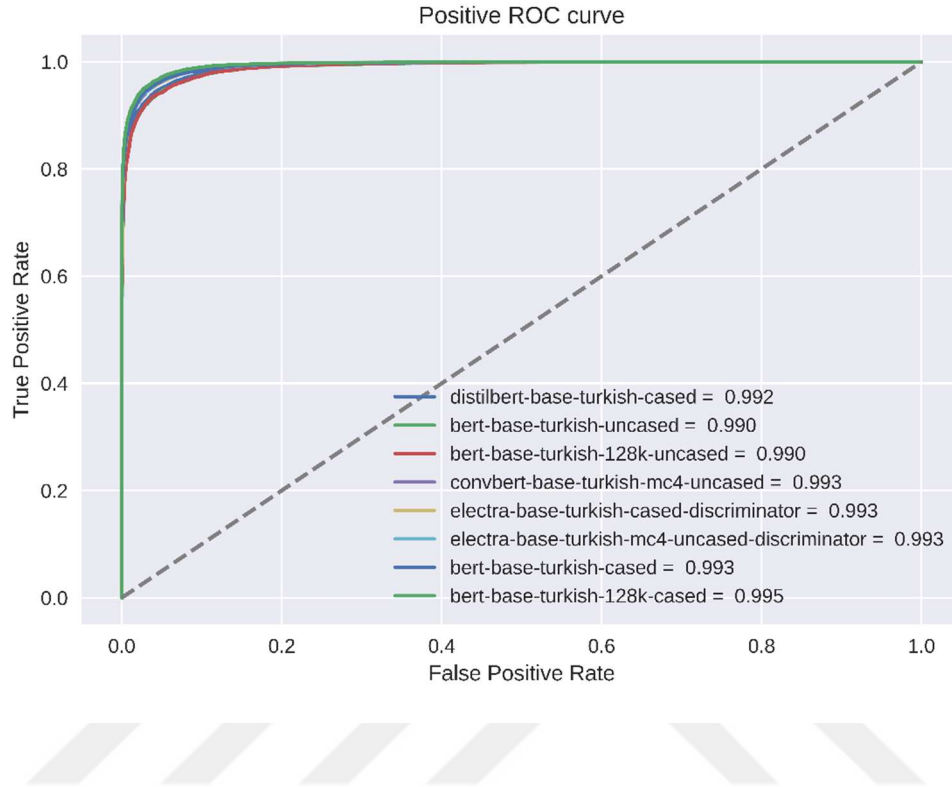


Figure 3.47: ROC curve for Positive Class with BERT Models



4. CONCLUSION

Recently, the rapid advancements in Internet and communication technologies have led to virtual environments, where activities in a city leave digital traces that reveal the city's dynamic structure and relationships. Additionally, critical components of a city's tourism industry, such as restaurants, public transport, and accommodations, serve as operational bases. Online social media platforms have assumed a crucial role in facilitating sharing of individuals' experiences related to these components.

Social media platforms have emerged as a powerful medium of communication for individuals to express their opinions and viewpoints concerning brands, topics, or places, consequently generating copious amounts of unstructured information. Sentiment analysis has emerged as the most widely employed text classification technique to evaluate this data, determining whether the underlying sentiment expressed is positive, negative, or neutral. This methodology facilitates a more profound comprehension of people's perceptions and sentiments toward a specific subject matter.

In this thesis, we have created a medium-sized Turkish dataset that can be used for multiple analyses by obtaining valuable data containing venue information and related user tips using various methods and techniques. During the process from the raw data collection to the generation of the automatically classified dataset, a balanced dataset (MAC) labeled by the experts was first created. Automatic prediction was made by fine-tuning to the BERT model after preprocessing and the AAC data set that we will use in our analysis has been created. Content-based and sentiment analyses were conducted on this dataset, allowing for a detailed analysis and classification of tips. As a result, inferences were drawn on the venues and cities, providing a unique perspective on smart city analysis.

The content-based analysis involves utilizing various methods and techniques, including spatio-temporal and word analysis. The word analysis approach encompasses techniques such as word clouds, co-occurrences of words, word embeddings, and topic modeling, to gain insights across all categories. Initially, we created annual heat maps of tips categorized by city and category groupings. This approach helped us to identify which city and category groups required further investigation. Next, we utilized word clouds to identify the most frequently used words in our MAC and AAC datasets, it allowed us to gain deeper insights into the dataset's content. We plotted the co-occurrences of words to understand how they co-exist in a tip. We also utilized word embeddings, which are vector representations of words created by the Word2Vec method. Finally, we applied LDA's topic modeling to extract the words associated with each topic based on their probability. Topics became intuitively distinguishable by examining words obtained with LDA's topic modeling approach.

Various techniques, ranging from traditional machine learning approaches to state-of-the-art language models, have been utilized for the purpose of conducting the classification process. The down-sampling technique was progressively tested to address the unbalanced data issue, and a balanced sample dataset of Auto Annotated Compilation was created. Additionally, not implementing stemming in the preprocessing phase has been shown to improve fine-tuning performance with BERT. A total of more than fifteen language models were trained in our dataset and comprehensively evaluated through Precision, Recall, and F1 metrics. At the same time, their ROC curves were plotted, and AUC scores were calculated. The BERT models exhibited high accuracy performance, with all models achieving an accuracy of over 90%. The "BERT Base Turkish 128k Cased" model achieved the highest accuracy, which was 93%. Among machine learning algorithms, the Logistic Regression model achieved the highest accuracy of 84%. The LSTM and Bi-LSTM models demonstrated accuracy values of 81%, which were very close to each other.

For future work, a more comprehensive and balanced Corpus will be constructed by gathering data from platforms other than Foursquare, with the aim of enhancing the accuracy performance. Furthermore, the estimation of venue rating scores can be achieved by analyzing the collected tips.

REFERENCES

- A. N. Azhar and M. L. Khodra (2020). Fine-tuning Pretrained Multilingual BERT Model for Indonesian Aspect-based Sentiment Analysis, 2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA), Tokoname, Japan, pp. 1-6, doi: 10.1109/ICAICTA49861.2020.9428882.
- Akın, A.A. & Akın, M.D., Zemberek, is an open source NLP library for Turkic languages. Structure, 2007.
URL : <https://github.com/ahmetaa/zemberek-nlp>
- AKSU, Muhammed & KARAMAN, Ersin. (2021). Analysis of Turkish Sentiment Expressions About Touristic Sites Using Machine Learning. Journal of Intelligent Systems: Theory and Applications. 4. 103-112. 10.38016/jista.854250.
- Alrumayyan, N., Bawazeer, S., AlJurayyad, R., Al-Razgan, M. (2018). Analyzing User Behaviors: A Study of Tips in Foursquare. In: Alenezi, M., Qureshi, B. (eds) 5th International Symposium on Data Mining Applications. Advances in Intelligent Systems and Computing, vol 753. Springer, Cham. https://doi.org/10.1007/978-3-319-78753-4_12.
- Ay, Betul & Talo, Muhammed & Hallac, Ibrahim & Aydin, Galip. (2018). Evaluating deep learning models for sentiment classification. Concurrency and Computation: Practice and Experience. 30. 10.1002/cpe.4783.
- Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, J. Mach. Learn. Res. 3 : 993–1022.

- Chen, Y., Yang, Y., Hu, J., Zhuang, C.: Measurement and analysis of tips in foursquare. In: 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), pp. 1–4 (2016).
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.
- Gallegos, Luciano & Lerman, Kristina & Huang, Arhur & Garcia, David. (2016). Geography of Emotion: Where in a City are People Happier?. 569-574. 10.1145/2872518.2890084.
- H. Assem, L. Xu, T. S. Buda and D. O'Sullivan (2016). Spatio-Temporal Clustering Approach for Detecting Functional Regions in Cities, 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), San Jose, CA, USA, 2016, pp. 370-377, doi: 10.1109/ICTAI.2016.0063.
- Jiang, Z., Yu, W., Zhou, D., Chen, Y., Feng, J. and Yan, S. (2020). Convbert : Improving bert with span-based dynamic convolution.
URL: arxiv.org/abs/2008.02496
- Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.
- Karayigit, Habibe & Akdagli, Ali & Aci, Cigdem. (2022). BERT-based Transfer Learning Model for COVID-19 Sentiment Analysis on Turkish Instagram Tips. Information Technology and Control. Vol. 51. 409-428. 10.5755/j01.itc.51.3.30276.
- Khan, Naimat & Wan, Songze & Luo, Yu-Feng. (2020). Spatiotemporal Analysis of Tourists and Residents in Shanghai Based on Location-Based Social Network's Data from Weibo. ISPRS International Journal of Geo-Information. 9. 70. 10.3390/ijgi9020070.

- Khan, Naimat & Wan, Wanggen & Yu, Shui & Muzahid, A.A.M & Khan, Sajid & Hou, Li. (2020). A Study of User Activity Patterns and the Effect of Venue Types on City Dynamics Using Location-Based Social Network Data. *International Journal of Geo-Information*. 9(12). 733. 10.3390/ijgi9120733.
- M. S. Başarslan ve F. Kayaalp (2021). Sentiment Analysis on Social Media Reviews Datasets with Deep Learning Approach, *Sakarya University Journal of Computer and Information Sciences*, c. 4, sayı. 1, ss. 35-49, doi:10.35377/saucis.04.01.833026
- Metev, D. (2023) 41+ Must Know Foursquare Statistics in 2023
URL : <https://review42.com/resources/foursquare-statistics/>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality, *Neural and Information Processing System (NIPS)*.
- R. AlQadi, H. Al-Nojaidi, L. Alabdulkareem, M. Alrazgan, N. Alghamdi and M. M. Kamruzzaman (2020). How Social Media Influencers Affect Consumers' Restaurant Selection: Statistical and Sentiment Analysis, *2nd International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, pp. 1-6, doi: 10.1109/ICCIS49240.2020.9257636.
- S. De, U. Jassat, A. Grace, W. Wang and K. Moessner. (2022) Mining Composite Spatio-Temporal Lifestyle Patterns from Geotagged Social Data, *2022 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, Espoo, Finland, pp. 444-451, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics55523.2022.00027.
- S. Yu, J. Su and D. Luo (2019) Improving BERT-Based Text Classification With Auxiliary Sentence and Domain Knowledge, in *IEEE Access*, vol. 7, pp. 176600-176612, 2019, doi: 10.1109/ACCESS.2019.2953990.

- Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019). Distilbert, a distilled version of bert : smaller, faster, cheaper and lighter.
URL: arxiv.org/abs/1910.01108
- T. H. Silva, P. O. S. V. d. Melo, J. M. Almeida, J. Salles and A. A. F. Loureiro (2012). Visualizing the Invisible Image of Cities, IEEE International Conference on Green Computing and Communications, Besancon, France, 2012, pp. 382-389, doi: 10.1109/GreenCom.2012.62.
- T. Tang, X. Tang and T. Yuan (2020). Fine-Tuning BERT for Multi-Label Sentiment Analysis in Unbalanced Code-Switching Text, in IEEE Access, vol. 8, pp. 193248-193256, doi: 10.1109/ACCESS.2020.3030468.
- U. U. Acikalin, B. Bardak and M. Kutlu (2020). Turkish Sentiment Analysis Using BERT, 28th Signal Processing and Communications Applications Conference (SIU), Gaziantep, Türkiye, pp. 1-4, doi: 10.1109/SIU49456.2020.9302492.
- Y. Wang, H. Sun, Y. Zhao, W. Zhou and S. Zhu (2020). A Heterogeneous Graph Embedding Framework for Location-Based Social Network Analysis in Smart Cities, in IEEE Transactions on Industrial Informatics, vol. 16, no. 4, pp. 2747-2755, doi: 10.1109/TII.2019.2953973.
- Yüksel, Asım & Tan, Gulsah. (2018). A real-time social network-based knowledge discovery system for decision making. *Automatika*. 59. 262-274. 10.1080/00051144.2018.1531214.
- Z. Wang, S. Li, F. Wu, Q. Sun, and G. Zhou (2018) Overview of NLPCC 2018 shared task 1: Emotion detection in code-switching text, in Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput., pp. 429–433.
- Zhenmin Dong and Chunfang Guo 2021 J. Phys.: Conf. Ser. 1792 012056

APPENDICES

Appendix A

A.1 Translation of Turkish words

Turkish	English
açık	open
açmak	to open
adalet	justice
adliye	courthouse
ahtapot	octopus
aile	family
akşam	evening
akvaryum	aquarium
alan	area
alışveriş	shopping
araba	car
araç	vehicle
aralık	december
avm	mall
aynı	same
bahçe	garden
balık	fish
başarılı	successful
bebek	baby
beklemek	to wait
çıkmaq	to leave

Turkish	English
köfte	meatball
köfteci	meatball maker
krep	crepe
kum	sand
kumru	dove
küçük	small
kültür sanat	art and culture
künefe	künefe
lastik	tyre
lazım	required
lezzet	flavor
lezzetli	tasty
liman	port
limonata	lemonade
limonlu	lemon
lokum	turkish delight
lounge	lounge
mağara	cave
mağaza	shopping centre
malzeme	material
fuat	fair

Turkish	English
çiçek	flower
çocuk	child
çorba	soup
çöp	rubbish
dağ	mountain
dakika	minute
damızlık	stud
damla	drop
deniz	sea
dere	stream
ders	lesson
dış	external
doğa	nature
doğum	birth
doktor	doctor
dondurma	ice cream
düğün	wedding
dünya	world
efsane	myth
eğlenceli	entertaining
ekip	team
ekmek	bread
el	hand
emek	work
eski	old
et	meat
fazla	more
ferah	spacious
festival	festival
film	film
fıyat	price
ipek	silk

Turkish	English
garson	waiter
gece	night
geçmek	to go past
gemi	boat
gereken	required
gezmek	to travel
girmek	to enter
göl	lake
gülyüzlü	smiling
hamburger	hamburger
hastane	hospital
hava	weather
havalimanı	airport
havuz	swimming pool
hayat	life
hemen	immediately
herkes	everyone
heykel	statue
hız	speed
hızlı	fast
hoca	teacher
hostel	hostel
hoş	pleasant
huzur	peace
huzurlu	peaceful
ızgara	grill
içeri	inside
ilgi	interest
ilgili	relating to
inmek	to get down
insan	person
konaklama yeri	accommodation place

Turkish	English
İskele	dock
iş	work
işletmek	operate
izlemek	to watch
kabak	courgette
kafa	head
kafe	cafe
kafeterya	cafeteria
kahvaltı	breakfast
kahve	coffee
kalabalık	busy
kalamar	squid
kale	castle
kalite	quality
kaliteli	quality
kalmak	to stay
kamp	camp
kamp alanı	camping area
kapı	door
kar	snow
karadeniz	black sea
kat	floor
kebab	kebab
kebabçı	kebab shop
keyifli	pleasant
kırmızı	red
kişi	person
kitap	book
klüp	club
koala	koala
kokoreç	kokorec
koltuk	armchair

Turkish	English
koy	bay
köfte	meatball
köfteci	meatball maker
krep	crepe
kum	sand
kumru	dove
küçük	small
kültür sanat	art and culture
künefe	künefe
lastik	tyre
lazım	required
lezzet	flavor
lezzetli	tasty
liman	port
limonata	lemonade
limonlu	lemon
lokum	turkish delight
lounge	lounge
mağara	cave
mağaza	shopping centre
malzeme	material
mantı	mantı
manzara	view
marina	marina
marka	brand
masa	table
mavi	blue
mayıs	may
menü	menu
merinos	merino
merkezi	central
mevcut	available

Turkish	English
meze	appetizer
midye	mussel
misafirhane	guesthouse
moda	fashion
mutlaka	definitely
müşteri	customer
müthiş	wonderful
müze	museum
müzik	music
nargile	hookah
nargile kafe	hookah cafe
nehir	river
nezih	decent
oda	room
ok	arrow
okumak	to read
olan	the one which.
orman	forest
otel	hotel
otopark	car park
oyun	game
özel	special
pahalı	expensive
pansiyon	hostel
para	money
park	park
parti	party
pasta	cake
pastane	patisserie
patates	potatoes
personel	employee
pide	pita bread

Turkish	English
pizza	pizza
plaj	beach
rahat	comfortable
rakı	raki
randevu	meeting
reklam	advert
restoran	restaurant
saç	hair
sadece	only
sağlık	health
sahil	beach
sahip	owner
sakın	beware
sakızlı	gummy
sakin	beware
salon	living room
samimi	sincere
sanat galerisi	art gallery
satmak	to sell
sen	you
ses	sound
sıcak	hot
sıra	order
sinema	cinema
sipariş	order
soğuk	cold
soslu	with sauce
spor	sport
steakhouse	steakhouse
su	water
şarap	wine
şehir	city

Turkish	English
piyaz	bean salad
şezlong	sunbed
şiş	skewer
taksi	taxi
tam	full
tat	taste
tatil	holiday
tatil köyü	holiday village
tatlı	sweet
tatlıcı	confectioner
tavuk	chicken
taze	fresh
temiz	clean
terminal	terminal
teşekkür	thanks
tiyatro	theatre
tost	toast
tören	ceremony
trafik	traffic
tren	train
tuz	salt
tünel	tunnel

Turkish	English
şelale	waterfall
türk restoranı	turkish restaurant
ulaşım	transport
uygun	suitable
üretim	production
ürün	product
üzerine	over
vapur	ferry
waffle	waffles
yaptırmak	have it made
yavaş	slow
yeni	new
yıkamak	wash
yılmaz	indomitable
yol	path
yürümek	to walk
yüz	face
yüzlü	faced
zaman	time
zevkli	enjoyable
zor	difficult
zorlu	tough

A.2 Translation of Turkish phrases

Turkish	English
Muhteşem bir müze.	An amazing museum.
Doğayla iç içe görülmeye değer.	It is worth seeing in touch with nature.
Ortam sıcak, ilgi desen sonsuz.	The environment is warm, the pattern of interest is endless.
Servisler rezalet ötesi.	The services are beyond disgrace.
Herşey süper yok yok.	Everything is super.
Izgara sucuk Hamburger.	Grilled Sausage Hamburger.
İç pilav tandır sütlaç.	Internal pilaf is tandoori rice pudding.
Menüsü çok başarılı yıllardır beğenmediğim yemek hiç çıkmadı.	The menu is very successful, there has never been a dish that I did not like for years.
Manzarası konumu güler yüzlü çalışanları çayı harika.	The view, the location, the friendly staff, the tea is great.
Cardgusto ana yemek alana ikincisi bedavaymış süper.	If you buy Cardgusto main course, the second one is free.
Bandırma yolunuz düştüğünde gitmeniz şiddetle tavsiye olunur on numara beş yıldız.	When you go to Bandırma, it is highly recommended that you go, number ten, five stars.
Köftesi güzel sulu geliyor biraz lastiksi tadı olsa tatlıları güzel giderim arada fiyatları bilmiyorum.	The meatballs are nice and juicy, if they taste a little rubbery, I can go with the desserts, by the way, I don't know the prices.
Tatlıları başarılı özellikle limonlu pastası oldukça güzel.	Desserts are successful, especially the lemon cake is quite nice.

BIOGRAPHICAL SKETCH

Muhammet Siddık EMEK, born in **** in *****, obtained his bachelor's degree in Computer Engineering from Ege University in 2006. He pursued his academic career by enrolling in the master's program at Galatasaray University in 2020, and he has been focused on artificial intelligence, deep learning, and natural language processing. With a professional background of 16 years, he has served as a software developer in numerous private companies.