Socio-Economic Impact of 2023 Turkey Earthquake Price Hikes: Insightful Analysis Using Transformer Models and XAI Models

by

Muhammed Yaseen Morshed Adib
22366020

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
September 2024

Declaration

The Authors of this research declares that

1. The work is conducted with sufficient groundwork. All the data are properly gathered. This is an original work for the completion M.Sc. in CSE thesis at Brac University.
2. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
3. All the references and aiding persons are addressed properly. .

Full Name of the Student & Signature:

Muhammed Yaseen Morshed Adib

22366020

Approval

The thesis/project titled "Socio-Economic Impact of 2023 Turkey Earthquake Price Hikes: Insightful Analysis Using Transformer Models and XAI Models" submitted by

1. Muhammed Yaseen Morshed Adib (22366020)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on September , 2024.

Examining Committee:

Examiner:
(External)

Dr. Md. Tauhid Bin Iqbal
Assistant Professor
Department of Computer Science and Engineering
East West University

Examiner:
(Internal)

Dr. Swakkhar Shatabda
Professor
Department of Computer Science and Engineering
Brac University

Supervisor:
(Member)

Dr. Farig Yousuf Sadeque
Associate Professor
Department of Computer Science and Engineering
Brac University

M.Sc. Coordinator:
(Member)

Dr. Md Sadek Ferdous
Professor
Department of Computer Science and Engineering
Brac University

Chairperson:
(Chair)

Dr. Sadia Hamid kazi
Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Natural disasters like the 2023 earthquake in Turkey have significant social and economic effects, making it important to use analytical methods for creating strong, disaster-ready communities. In our work, we analyze public sentiment on social media after the earthquake, focusing on the rise in prices that followed. We classify public reactions into three categories: negative, positive, and neutral. To do this, we use several machine learning models, deep learning models, and two transformer based models. By analyzing the connection between people's feelings and socio-economic factors like consumer spending, inflation, and price hikes, we aim to understand how public sentiment relates to policy decisions made in response to the crisis. Among all models tested, modified DistilBERT stood out, delivering the best performance with an accuracy of 82.20% and an F1-score of 84.30%. This shows that transformer-based models, particularly DistilBERT, are highly effective for sentiment analysis in this context. DistilBERT's strong precision, recall, and F1-score suggest that it could be a valuable tool for informing policy changes to reduce the socio-economic impacts of natural disasters. Additionally, we used Explainable AI to help explain the model's results, ensuring that policymakers can make informed decisions based on the data. Our research highlights the importance of advanced natural language processing (NLP) techniques for developing evidence-based policies in disaster management.

Keywords: NLP; Machine Learning; Deep Learning; Transformer; XAI; Sentiment Analysis, Earthquake

Acknowledgement

Several people are involved in this study project. The people who gathered the information from various sources are appreciated. It was a challenging task to annotate. We gratefully acknowledge those who assisted us with the annotation. The deserving supervisor Farig Yousuf Sadeque deserves a great deal of praise for his guidance through every step of the task accomplishment. Many well-known academics have offered their insights at every stage of the process to improve the caliber of the work. We also appreciate their help, and thank you.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Communities around the world face immense challenges due to natural disasters, which have serious socioeconomic impacts that require effective policy solutions. The 2023 earthquake in Turkey is a reminder of how environmental events can affect socio-economic stability. Following the earthquake, rising prices worsened conditions for those affected, emphasizing the need for evidence-based policies to build disaster resilience.

Disasters like earthquakes can cause significant socioeconomic disruptions, such as damaged infrastructure, loss of jobs, and higher costs of goods and services. These events often break supply chains and reduce the availability of essential goods, leading to price increases that further strain affected communities. The 2023 earthquake in Turkey had a profound impact, with many people facing higher prices for basic necessities, making recovery even harder. This highlights the importance of understanding and addressing public sentiment to effectively support those in need. Natural Language Processing (NLP) is an essential tool for identifying disaster-related posts on social media and understanding their semantic, spatial, and temporal context, enabling better preparedness and response in disaster-prone areas [1]. This helps improve preparedness and response in disaster-prone areas. By analyzing social media sentiment, NLP provides insights into the concerns and feelings of affected individuals.

Public sentiment is crucial in these situations because it reflects the immediate reactions, needs, and priorities of those affected by disasters. Machine learning models, deep learning models, and transformer-based models like DistilBERT and XLNet are essential for enhancing NLP techniques [2] [3]. This can analyze social media comments to inform policy making. These models help uncover public sentiments and guide more responsive and compassionate policy measures. Using NLP to analyze social media comments allows policymakers to understand public sentiment and adjust their strategies accordingly [4]. This ensures policies meet the real needs of people affected by disasters. It helps communities respond better and become stronger.

1.2 Motivation

Earthquakes and other natural disasters cause more than just physical damage. They also lead to financial struggles for those affected. [5]. After the 2023 earthquake in Turkey, prices went up, making life harder for people already dealing with the disaster. Studies show that after natural disasters, the cost of basic goods can increase by up to 30%, adding more pressure on families.

To help those affected, it's important to understand how they feel. Social media can provide insight into people's emotions and concerns. By using advanced computer programs to analyze what people post online, we can learn more about their thoughts and worries [6].

Our objective is to use these instruments to ascertain the opinions and effects of people following the earthquake in Turkey. For example, research has found that over 70% of people use social media to share their experiences and concerns after a disaster like the Turkey earthquake.

Our goal is to assist decision-makers in using this data to guide their decisions. We can assist leaders in developing strategies that truly benefit those who require assistance by integrating data on people's feelings with critical information. Ensuring the voices of individuals impacted by natural disasters be heard is the main goal of this study. In order for decision-makers to be truly impactful, we want to ensure that they have a thorough understanding of ordinary people's lives. The main goal of this research is to help people after disasters. We want to make their lives a bit easier during difficult times.

1.3 Objective

Our study aims to analyze social media data sentiment to examine the socio-economic effects of the 2023 earthquake in Turkey. In order to pinpoint important socioeconomic variables influencing the aftermath of the disaster, we hope to gather a sizable dataset of social media remarks from those impacted by the earthquake. In particular, our study focuses on using sentiment analysis methods to gather information from the gathered social media comments, especially concerning the opinions held over the earthquake-caused price increases and their wider socio-economic consequences. Furthermore, our goal is to investigate the relationship between major socioeconomic indicators and sentiment changes in order to provide information for evidence-based policy responses that are customized to the needs and perspectives of communities affected by disasters.

In addition, we would like to process and analyze the social media data in order to investigate the efficacy of different machine learning (ML) models, deep learning (DL) models and two advanced models in capturing complex sentiment expressions in communities affected by disasters. Furthermore, to make the models easier to understand and more transparent, especially with complex language data, we will use explainable AI techniques to interpret and extract useful insights from the data.

1.4 The 2023 Turkey Earthquake: Assessing the Socio-Economic Aftermath and Recovery Efforts

The earthquake that occurred in Turkey in 2023 had a significant and varied socioeconomic effect. The February 2023 earthquakes that rocked northern and western Syria, as well as southern and central Turkey, left large amounts of damage and casualties in their wake. The immediate effects on the economy included buildings and infrastructure being destroyed as well as output being disrupted. Early projections indicated that the reconstruction efforts would likely balance out the early negative effects, meaning that the net effect on Turkey's economic growth could be less than 1 percentage point for the year.

According to the World Bank, Turkey's 2021 GDP amounted to $34.2 billion in direct physical damages, with the potential for substantially greater recovery and reconstruction expenditures [7]. Together with leaving a huge number of people homeless, the earthquake had a profound social impact, particularly in areas with high rates of poverty and a high number of refugees. A thorough approach to catastrophe management and economic resilience planning is important, given the expected considerable overall cost to Turkey's economy.

1.5 Socio-Economic Impact of the 2023 Turkey Earthquake: Factors Contributing to Price Hikes

The Turkey earthquake of 2023 had a big impact on the economy of the nation, which included elements that led to price increases. Here are a few particulars:

1. **Costs of Infrastructure Damage and Reconstruction:** Buildings, roads, and utilities all sustained significant damage as a result of the earthquake. Costs rose as a result of the ensuing reconstruction operations, which demanded significant resources. Higher costs resulted from the need to rebuild or repair damaged infrastructure, which may have an indirect effect on the cost of goods and services [8].

2. **Supply Chain Interruptions:** Production and delivery of commodities were impacted by the disruption in supply chains caused by the earthquake. Businesses that depend on inputs from the impacted areas experienced difficulties locating supplies, which could have resulted in shortages and increased costs.

3. **Increased Demand for Construction Materials:** The need for construction materials including steel, cement, and wood increased as a result of the reconstruction activities. Prices for these materials rose due to the increasing demand, which had an impact on building expenses and, ultimately, inflation overall.

4. **Consumer Attitude and Purchase Patterns:** The aftermath of the earthquake might have had an impact on consumer confidence and purchasing patterns. Future uncertainty and safety worries may cause patterns of consumption to shift, which would impact demand and pricing [9].

5. **Impact of Multipliers on Supply Chains:** The consequences of the earthquake on supply chains were multiplicitous, meaning that problems in one industry could have a knock-on effect on others. For instance, damage to industries may have an impact on the amount of items produced both domestically and abroad, raising prices and costs.

6. **The impact of inflation on household expenditures:** The cost of the Minimum Expenditure Basket (MEB), which contains necessities for households, increased significantly. The cost of living increased following the earthquake due to the rise in food prices and rental costs, which put pressure on people to experience inflation.

In conclusion, there were a number of economic consequences of the 2023 Turkey earthquake, some of which led to price increases. The post-disaster inflationary pressures were impacted by a convergence of factors including reconstruction costs, labor market fluctuations, disruptions in supply, and policy actions.

1.6 Leveraging Public Sentiment from Social Media for Post-Disaster Policy Making

By utilizing the information found in social media feedback, decision-makers can shape policies that truly resonate with the public's concerns and feelings, resulting in improved strategies for disaster recovery.

1. **Real-Time Feedback:** Social media allows for prompt policy revisions by giving instant insights into the public's reactions to disaster response activities.

2. **Community Engagement:** Social media interaction with the community promotes trust and guarantees that impacted individuals' perspectives are heard and taken into account when creating recovery strategies.

3. **Identifying Urgent Needs:** By identifying the most pressing needs and concerns of disaster victims, sentiment analysis on the internet helps direct targeted governmental initiatives.

4. **Monitoring Public Mood:** Using social media to monitor public sentiment can help legislators assess the effectiveness of current initiatives and identify areas that require further focus.

5. **Enhancing Communication:** Social media facilitates a two-way conversation between government and public by providing a direct avenue for policy decisions to be communicated and comments to be received.

1.7 Contributions

The contributions of this research paper are discussed as follows:

1. This research investigates the socio-economic effects of the 2023 earthquake in Turkey, focusing on the interplay between public sentiment and economic indicators such as consumer spending, inflation, and price hikes. The analysis provides insights into the complex dynamics of disaster impact on society.

2. A comprehensive dataset has been compiled from 5000 social media comments post-earthquake, categorizing sentiments into negative, positive, and neutral. This dataset, annotated with the expertise of domain specialists, serves as a foundation for understanding public mood in response to the crisis.

3. Among the various machine learning models and state-of-the-art models evaluated, a modified DistilBERT emerged as the most effective after necessary hyperparameter tuning and changing the attention layer. The model is compared with state- of-the-art architecture

4. Utilized XAI model to interpret the performance of the modified model, aiding in evidence-based policy making.

1.7.1 Usability of the Research

Analytical methods are necessary to enable evidence-based policy-making for disaster-resilient communities, given the socio-economic effect of natural disasters. In this work, we perform sentiment analysis on social media data after the 2023 earthquake in Turkey, emphasizing the ensuing price increases. We examine public sentiment in three categories: negative, positive, and neutral, using machine learning models, deep learning models, DistilBERT, and XLNet. Through the correlation of mood movements with socio-economic variables including consumer spending, inflation, and price hikes, we shed light on the complex relationship between public perception and policy reactions to the crisis. DistilBERT's potential as a useful tool for understanding policy measures targeted at minimizing the socio-economic effects of natural disasters is highlighted by its high precision, recall, and F1 score. Our study advances our knowledge of the socioeconomic effects of disasters and emphasizes the value of using cutting-edge natural language processing (NLP) techniques to formulate evidence-based policies for disaster management.

1. Our custom-collected dataset, focusing on real-time social media comments after the 2023 Turkey earthquake, captures authentic public reactions, providing a valuable resource for policymakers.

2. This analysis offers a clear view of public emotions (negative, neutral, positive) toward price hikes after the earthquake, helping policymakers measures public concerns.

3. Utilizing various machine learning models, deep learning models and transformer based models sentiment analysis is conducted on social media data to discern public sentiment across three categories: negative, positive, and neutral.

4. Explainable AI techniques provide insightful information by making the model's predictions transparent, supporting data governance by offering clarity on the key factors.

5. These insights enhance data governance by ensuring effective data collection, processing, and analysis. Our research delivers impactful insights that guide informed decisions on disaster management, price regulation, and strategies for economic recovery after crises.

Chapter 2

Related Work

In recent years, sentiment analysis and language technologies have become increasingly important in understanding public responses to natural disasters and socio-economic issues. Twitter data, in particular, has proven to be a valuable source for analyzing societal reactions. After the earthquake in Izmir, researchers analyzed Twitter to examine tweet frequency, recurring themes, popular sentiments, and geographic distribution, providing critical insights into public emotions and attitudes in the aftermath of the disaster [10]. The analysis of such data helps paint a clearer picture of the collective consciousness of people affected by the disaster, offering both quantitative (tweet counts, location data) and qualitative (emotional tone, thematic elements) insights. This kind of data is essential for understanding how individuals react in times of crisis, how information spreads, and what issues or concerns are most prominent among the affected population. Similarly, another analysis of Twitter reactions following the earthquakes in Turkey and Syria revealed emotional responses such as empathy, concern, fear, and calls for action. This study explored how individuals and communities emotionally navigated the crisis, contributing to disaster management efforts [11]. By identifying these emotional trends, researchers can inform disaster relief organizations and government agencies about the public's emotional state, enabling more effective communication and intervention strategies during the aftermath of disasters.

In addition to natural disasters, sentiment analysis has been used to investigate socio-economic issues, such as rising energy prices. From January 2021 to June 2022, Twitter data was analyzed using transformer-based models to classify sentiments and topic modeling techniques, such as BERTopic and LDA, to identify themes related to energy pricing [12]. The use of advanced transformer-based methods such as BERT allows researchers to capture the nuanced emotions behind these public reactions, while topic modeling helps to group related conversations into clusters, providing insights into the key issues discussed by the public, such as inflation, government policies, and potential solutions. This type of analysis is critical because socio-economic issues, especially those that affect day-to-day living, often spark strong reactions, and understanding these can help policymakers gauge the level of public concern and respond accordingly. Furthermore, a comparative analysis of text-based emotion recognition models found that BERT, RoBERTa, DistilBERT, and XLNet vary in their ability to accurately identify emotions in textual data, with the study helping to determine the most effective model for such tasks [13]. Comparative studies like these are essential for ensuring that sentiment analysis

techniques are not only accurate but also efficient, particularly when applied to large-scale datasets such as social media posts. Choosing the right model can lead to better sentiment interpretation, which is critical for informing decision-making processes.

The application of sentiment analysis has also been extended to price hikes in different linguistic contexts. One study leveraged an LSTM-ANN approach to analyze Bangla social media comments related to price hikes, showing the effectiveness of combining sequence modeling with neural networks for sentiment classification [14]. Price hikes are a common point of public discourse, and social media provides an outlet for the public to express their frustration or support regarding such changes. The use of sequence models like LSTM allows researchers to take into account the order and context of the words in these comments, making the analysis more precise. Neural networks like ANN can then be used to classify the overall sentiment expressed, giving a clearer picture of public opinion. Another study compared the performance of various machine learning algorithms, including neural networks, decision trees, and support vector machines, in analyzing public sentiments around fuel price increases, offering valuable insights for stakeholders and policymakers [15]. This comparison of machine learning algorithms is important for determining which model is best suited for sentiment analysis in various contexts. By finding the most effective method, analysts can provide more accurate data to policymakers, helping them understand public opinion more clearly and respond to price hikes in ways that may mitigate negative public reactions. Table 2.1 refers to the challenges of related work.

Table 2.1: Challenges Addressed from the Recent Literature

Reference of the paper	**Challenges**
Ağralı et al., 2022	Accurately interpreting data is challenging due to the the limited contextual understanding of tweets.
Hossain et al., 2023	Focusing solely on initial Twitter responses post-earthquake, without tracking sentiment changes over time.
Kastrati et al., 2023	Examining the complexity of energy price discussions necessitates a comprehensive and robust analysis.
Adoma et al., 2020	Fine-tuning language models for emotion recognition demands meticulous experimentation and validation to optimize performance
Saputra et al., 2023	Choosing the right machine learning algorithms for sentiment analysis is tricky due to their diverse strengths and weaknesses, shaped by data and task specifics.

Sentiment analysis has also proven beneficial in shaping public policy by aiding policymakers in understanding public opinion, identifying emerging issues, and assessing the impact of policies. Researchers have highlighted how these methods can enhance evidence-based policymaking by providing insights from social media data

throughout the policy cycle [16]. By analyzing the sentiment of social media users, policymakers can gain a real-time understanding of how the public perceives specific policies or issues. Sentiment analysis offers an ongoing mechanism to evaluate public response and adjust strategies accordingly. As policies are implemented, these methods can be used to track the success of the measures taken and provide feedback that may lead to policy refinement. This integration of data-driven insights allows for a more transparent and responsive policymaking process, enabling governments to address public needs and concerns more effectively.

Together, these studies demonstrate the power of advanced computational methods for sentiment analysis and their potential to support policymaking by addressing public concerns. Our research builds on this foundation by employing a combination of machine learning (ML) and deep learning (DL) models, including state-of-the-art transformer-based approaches like DistilBERT and XLNet, to analyze sentiment in social media comments. By integrating exploratory data analysis (EDA) with explainable artificial intelligence (XAI) techniques, our work provides interpretable insights that can be used to guide more informed decision-making.

Chapter 3

Background Study

3.1 Experimental Models

To validate the dataset, a good number of machine learning (ML), deep learning (DL) and transformer based architectures were utilized. This section provides a brief summary of these models.

3.1.1 Machine Learning Models:

1. **k-Nearest Classifier:** KNN, denoted as k-Nearest Neighbors, is a way of making predictions on new data by looking at similar examples. KNN predicts without complex models. It stores all training data. It finds the k closest data points based on distance for new data. In classification, the majority class of those k neighbors becomes the prediction. For regression, the average value of those neighbors is predicted. KNN finds the k closest points on the map. The new point is assigned the most common label from those neighbors in classification. The new point's value is predicted in regression as the average of its neighbors' values. It excels at making predictions based on similar past experiences. It stores all the training data like a reference map, allowing it to quickly find the k closest data points, which means the neighbors to a new arrival. Still, when it avoids complex models, KNN can become heavy with massive datasets, as searching the entire map gets time-consuming. Natural Language Processing (NLP) is utilized to preprocess raw texts and K-Nearest Neighbors (KNN) classification algorithm to classify the processed data [17].

$$\hat{y}_i = \text{mode}\left\{y_j \mid j \in \text{kNN}(\mathbf{x}_i)\right\} \tag{3.1}$$

   In k-Nearest Neighbors (k-NN) for classification, the predicted class $\hat{y}_i$ for an input $\mathbf{x}_i$ is determined by finding the most frequent class (mode) among the k nearest neighbors in the training data, where $\hat{y}_i$ represents the class label of the j-th nearest neighbor to $\mathbf{x}_i$.

2. **Decision Tree Classifier:** A Decision Tree Classifier helps decide what category something belongs to based on its characteristics. It works by repeatedly splitting the data into smaller groups based on certain features, aiming to have each group as pure as possible. It decides these splits by calculating which moves will reduce uncertainty about the data's classification at each step. Benefits It's easy to understand and explain to others. But it complicates if it gets

too fixated on the details in the training data, making wrong guesses when it sees new data. To solve the problem, sometimes it's best to limit how many questions it can ask or to use a bunch of decision trees together to make a better guess.

$$\hat{y}(\mathbf{x}) = \sum_{m=1}^{M} I(\mathbf{x} \in R_m) c_m \tag{3.2}$$

This equation illustrates how a decision tree classifier predicts an output for an input vector x. It checks each leaf node R_m to see if x falls within that region using the indicator function $I(x \in R_m)$. If true, it assigns the predicted class c_m of that leaf node to x. The final prediction $\hat{y}(x)$ is the sum of the classes of all relevant leaf nodes, thereby determining the class based on the tree's structure.

Decision trees in Natural Language Processing (NLP) are employed to model relationships between textual features and outcomes [18].

3. **XGBoost:** XGBoost (Extreme Gradient Boosting) is a highly efficient and flexible machine learning algorithm used primarily for classification tasks. It builds an ensemble of decision trees in a sequential manner, where each tree attempts to correct the errors of its predecessor. XGBoost uses a gradient boosting framework, optimizing a given loss function by adding new trees that predict the residuals of previous trees. This iterative process continues until the model's performance no longer improves significantly. XGBoost incorporates regularization techniques to prevent overfitting, enhancing the model's generalization capabilities. It also supports parallel processing, which accelerates training and makes it suitable for large datasets. Additionally, XGBoost handles missing data well and can automatically learn which features are important during the training process. However, the complexity of the model requires careful tuning of hyperparameters, such as the learning rate, maximum tree depth, and the number of trees, to achieve optimal performance. Despite its complexity, XGBoost's robustness and high accuracy make it a popular choice in competitive machine learning tasks.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i) \tag{3.3}$$

In XGBoost, the predicted value $\hat{y}_i$ for an input $\mathbf{x}_i$ is the sum of the predictions from all K trees in the ensemble, where $f_k(\mathbf{x}_i)$ represents the prediction from the k-th tree. In Natural Language Processing (NLP), XGBoost (Extreme Gradient Boosting) is often preferred for tasks requiring high predictive accuracy, such as text classification and sentiment analysis, due to its ability to handle sparse data efficiently, optimize feature selection through gradient boosting [19].

4. **Support Vector Machine:** Support Vector Machines (SVM) are a fundamental machine learning tool for classification tasks. SVM aims to identify the optimal hyperplane that best separates different classes in the feature space.

This hyperplane is selected to maximize the margin, which is the distance between the hyperplane and the nearest data points from each class, called support vectors. By maximizing this margin, SVM ensures a robust separation between classes. SVMs are versatile, capable of handling both linear and non-linear classification through the use of kernel functions, such as linear, polynomial, and radial basis function (RBF) kernels. Despite their power, selecting the appropriate kernel and tuning parameters like the regularization term and kernel coefficients can be complex and requires a deep understanding of the data and the model to achieve optimal performance. Support Vector Machines (SVM) are employed in Natural Language Processing (NLP) for tasks like text classification, sentiment analysis, and spam detection by converting text data into high-dimensional feature spaces [20].

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{3.4}$$

The equation represents the decision function of a Support Vector Machine (SVM), where w is the weight vector, x is the input feature vector, and b is the bias term, collectively determining the classification boundary.

5. **Random Forest:** Random Forests are utilized in Natural Language Processing (NLP) for various tasks such as sentiment analysis [21]. Random Forest is a powerful ensemble learning method used for classification tasks. It constructs a multitude of decision trees during training and outputs the mode of the classes predicted by the individual trees. Each tree in a Random Forest is built from a random subset of the training data, and at each split in the tree, a random subset of features is considered, ensuring diversity among the trees. This randomness helps to reduce overfitting and improve generalization. Random Forests are robust and handle a large number of input features well, offering high accuracy and the ability to handle missing data effectively. However, they can be computationally intensive, especially with a large number of trees, and interpreting the model can be more complex compared to a single decision tree. Hyperparameter tuning, such as the number of trees and the depth of each tree, is crucial for optimal performance.

$$\hat{y} = \text{mode}\left\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\right\} \tag{3.5}$$

In a Random Forest model, the predicted class $\hat{y}$ for a given input is determined by the most commonly predicted class among all individual decision trees $\text{mode}\left\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\right\}$.

3.1.2 Deep Learning Architectures

In this study, our attention is directed towards utilizing RNN-based architectures due to their impressive ability to capture semantics [22]. The sections are dedicated to briefly introducing several cutting-edge deep learning architectures within the relevant field.

1. **Long Short-Term Memory** The Long Short-Term Memory (LSTM) network is a widely used RNN architecture designed to capture long-term dependencies in sequential data. It features three gates: the Input gate, the

Forget gate, and the Output gate, which together regulate the flow of information through the network. The Input gate controls the addition of new information to the cell state, the Forget gate manages the removal of outdated information, and the Output gate determines the output of the LSTM cell. This architecture allows LSTMs to effectively learn and retain information over long sequences, making them well-suited for tasks involving complex dependencies and sequential patterns.

2. **Gated Recurrent Unit:** The Gated Recurrent Unit (GRU) is a well-known RNN architecture that serves as an alternative to the LSTM network. It features two gates: the Reset gate and the Update gate, which together facilitate effective semantic capturing. The Reset gate typically manages previously accumulated information, while the Update gate incorporates new information. A significant advantage of the GRU is its computational efficiency.

3. **Bidirectional Long Short-Term Memory:** The Bidirectional Long Short-Term Memory (BiLSTM) network is a powerful architecture for sentiment analysis classification tasks. BiLSTM enhances the standard LSTM by processing input data in both forward and backward directions, capturing context from both past and future states. This bidirectional approach allows the model to better understand the sentiment of a text by considering the full context of each word.

4. **Bidirectional Gated Recurrent Unit:** The Bidirectional Gated Recurrent Unit (BiGRU) network is an advanced architecture used for sentiment analysis classification tasks. BiGRU extends the standard GRU by processing input data in both forward and backward directions, allowing the model to capture context from both past and future words in the text. This bidirectional processing enhances the model's ability to understand sentiment by considering the entire context surrounding each word.

3.1.3 Transformer Based Model: XLNet

XLNet, introduced by Zhilin Yang et al. in 2019, is a transformer-based model designed to improve upon previous models like BERT by combining autoregressive modeling with bidirectional context learning. XLNet uses a unique Permuted Language Modeling (PLM) technique, enabling it to capture dependencies from multiple directions without corrupting input sequences.

Additionally, XLNet integrates the memory mechanism from Transformer-XL, allowing it to effectively model long-term dependencies in sequences. This combination of features makes XLNet more efficient for tasks that require understanding long-range context, and it has demonstrated superior performance over BERT in various NLP benchmarks.

3.1.4 Transformer Based Model: DistilBERT

DistilBERT is a streamlined, faster, and more efficient variant of BERT tailored for natural language processing tasks. It maintains around 97% of BERT's language

understanding capabilities while being 60% faster and utilizing 40% fewer parameters. This efficiency is achieved through knowledge distillation, enabling DistilBERT to emulate the performance of the larger BERT model. Despite its compact size, DistilBERT excels in various NLP applications, including text classification, sentiment analysis, named entity recognition, and question answering. Its speed and efficiency make it ideal for scenarios that require rapid inference and have limited computational resources. The details of the modified DistilBERT are discussed in subsection 4.7.

3.1.5 Differences Between DistilBERT and XLNet Models

Although both DistilBERT and XLNet are transformer-based models, each brings unique strengths to our research, offering valuable insights when applied to sentiment analysis of social media data.

1. **Model Architecture:**

   - **DistilBERT** is a distilled version of the BERT (Bidirectional Encoder Representations from Transformers) model. It reduces the size and computational cost while retaining about 97% of BERT's performance. It processes text bidirectionally, meaning it takes context from both the left and right sides of a word, making it highly efficient for tasks requiring a deep understanding of sentence structure.
   - **XLNet**, on the other hand, is based on a different paradigm called permutation language modeling. It improves upon BERT by capturing dependencies in a more generalized way, leveraging all possible permutations of word sequences. This enables XLNet to consider the order of words more effectively and model longer-range dependencies.

2. **Training Objectives:**

   - **DistilBERT** uses a masked language modeling (MLM) approach, where some words in a sentence are masked and the model is trained to predict them based on the context. This works well for understanding the general meaning of a sentence.
   - **XLNet**, however, employs autoregressive modeling, which predicts words based on all permutations of the input sequence. This allows the model to better understand the relationships between words in more complex and variable structures.

3. **Speed and Efficiency:**

   - **DistilBERT** is designed to be a lighter and faster alternative to BERT, making it highly suitable for applications where computational efficiency is critical without sacrificing much performance. This was beneficial in our research for handling large-scale social media data quickly and efficiently.
   - **XLNet**, while more resource-intensive than DistilBERT, offers improved performance on a variety of natural language processing tasks due to its

more sophisticated modeling of word order. However, this comes at the cost of slower training and inference times.

4. **Contextual Understanding:**

   - **DistilBERT** is more focused on capturing sentence-level understanding, making it well-suited for general sentiment analysis tasks.
   - **XLNet**, due to its ability to consider permutations of word sequences, offers a more granular approach, potentially capturing subtle nuances and long-range dependencies in the text. This may improve performance when dealing with complex sentence structures or nuanced sentiments.

In our research, we applied both models to explore their different strengths. While DistilBERT's efficiency made it ideal for quick and accurate sentiment classification, XLNet's advanced contextual understanding allowed us to capture deeper relationships in the data, providing a comprehensive sentiment analysis of social media comments following the 2023 Turkey earthquake.

3.2 Explainable AI Techniques

In our research, we utilize Explainable AI (XAI) techniques to enhance transparency and interpretability in the decision-making processes of our machine learning models. These techniques allow us to comprehend how models arrive at specific predictions, providing insights into the influence of input features—such as sentiment, price hike relatability, and region—on the output. By making AI systems more interpretable, XAI fosters trust in the models' outcomes, which is crucial for deriving reliable insights from complex datasets. This transparency enhances the credibility of our findings and supports evidence-based analyses of socio-economic impacts.

3.2.1 Local Interpretable Model-agnostic Explanations (LIME)

Local Interpretable Model-agnostic Explanations (LIME) is a methodology designed to clarify the decision-making processes of advanced predictive models by offering explanations for individual predictions. It does this by generating a simplified, interpretable model that approximates the reasoning of the complex model for a specific instance, thereby illuminating the factors influencing that particular outcome. LIME achieves this by introducing slight variations to the data point and observing the resulting changes in predictions. By iteratively modifying the data points and applying an interpretable model, LIME captures the decision-making process near the instance of interest. Its strength lies in producing models that are understandable and reveal the rationale behind the predictions.

$$\theta_{\text{next step}} = \theta_{\text{current}} - \eta \cdot \nabla_\theta J(\theta; \mathbf{x}^{(i)}, y^{(i)}) \tag{3.6}$$

- $\xi(x)$ represents the explanation model for instance x.
- f denotes the complex model being explained.
- g symbolizes the simple, interpretable model chosen to approximate f locally, where g belongs to a family of models G.

- L is the loss function quantifying the discrepancy between f and g in the vicinity of x, which defines this locality.
- $\Omega(g)$ measures the model g's complexity, promoting simplicity in the explanation.

Despite its effectiveness in enhancing the transparency of model predictions, LIME faces challenges. Its explanations are localized, focusing solely on specific instances rather than the overall logic of the model. Additionally, the choice of the local interpretive model and its neighborhood can significantly impact the accuracy and relevance of these explanations. Nevertheless, LIME serves as a valuable tool for illuminating the predictive mechanisms of complex models, fostering greater understanding and trust in their outputs by translating them into more accessible and interpretable forms.

3.2.2 SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) is a method grounded in cooperative game theory that aims to provide insights into the contributions of individual features toward a model's prediction. SHAP computes the Shapley values, which quantify how much each feature contributes to the difference between the actual prediction and the average prediction of a model. By assigning these contributions in a fair and consistent manner, SHAP offers a robust framework for interpreting the behavior of complex machine learning models.

The core strength of SHAP lies in its theoretical foundation, ensuring consistency and accuracy in attributing the prediction to features. Unlike local methods that focus only on specific instances, SHAP provides both local and global interpretability by aggregating the feature contributions across multiple predictions. SHAP's methodology involves calculating the marginal contribution of each feature by considering all possible combinations of features in the prediction process, making it a powerful tool for feature attribution. This comprehensive approach allows for deeper insights into model behavior, fostering transparency and trust in predictive models.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right] \tag{3.7}$$

In this equation, ϕ_i denotes the Shapley value for feature i, representing its contribution to the prediction, $f(S)$ is the prediction for a subset of features S excluding feature i, N refers to the set of all features, and S is any possible subset of N excluding i. The term $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ assigns a weight to each subset S, ensuring fair attribution of contributions across all features.

Explainable AI techniques such as LIME and SHAP are highly effective for ensuring data quality and deriving insightful analyses. By utilizing machine learning models like DistilBERT alongside these XAI techniques, we achieve accurate analysis of large datasets, providing reliable insights into complex issues. These techniques explain how models arrive at their decisions, enhancing transparency and trust in the outputs. This framework not only improves interpretability and accountability but also facilitates the continuous enhancement of analyses.

Chapter 4

Research Methodology

Figure 4.1 represents the overall workflow of this research. Initially, data were collected from diverse sources, including social media platforms like Facebook, YouTube, Twitter, and online news portals. Annotated by multiple individuals, the collected data underwent scrutiny to ensure accurate sentiment assignments. Subsequently, exploratory data analysis (EDA) was conducted to unveil meaningful patterns within the dataset. Preprocessing techniques were then applied. After processing the data were utilized across various machine learning models such as Support Vector Machines (SVM) and Random Forest, as well as deep learning architectures like Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM). Additionally, state-of-the-art models such as distilBERT and XLNet were employed in the analysis. The models' performance was evaluated using various metrics such as Precision, Recall, and F1 score. Finally, the modified model was interpreted through the lens of explainable artificial intelligence (XAI), Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP).

4.1 Data Collection Procedure

The data collection process began with the selection of topics, focusing on events such as the "Turkey Earthquake 2023" and its economic repercussions, specifically the "Price Hike due to Turkey Earthquake 2023". Data were sourced from various social media platforms, including Facebook, YouTube, Twitter, and online news portals, chosen for their extensive user base and diverse discussions. Relevant content was identified through tailored search queries, utilizing filters to refine results. Posts and discussions were pinpointed based on keywords, hashtags, and mentions, prioritizing those with high engagement and diverse viewpoints. A systematic sampling strategy was implemented to capture diverse geographic locations. Collected data were securely stored and organized for analysis, with proper documentation to track the source and context of each data point.

4.2 Dataset Description

Subsequently, the dataset was created and several columns were combined based on the results to comprehend the social and economic effects of the Turkey Earthquake 2023, especially with reference to the subsequent price increases. An additional

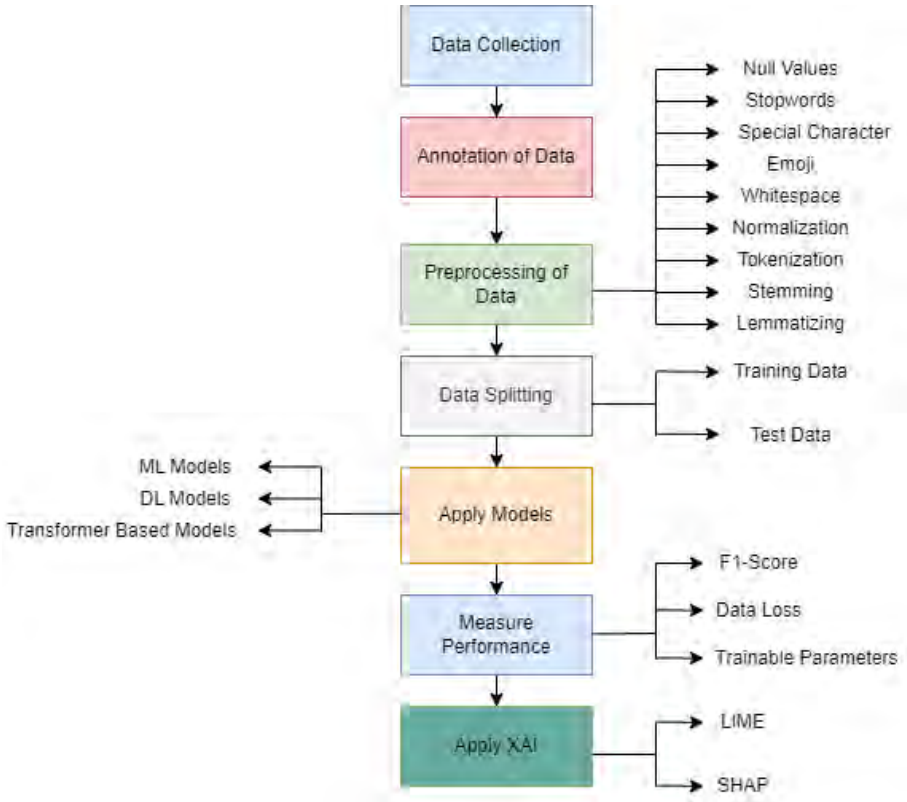


Figure 4.1: Methodology of Research Work

column was set aside to compile remarks in English that users posted on different social networking sites. After that, three participants helped assign the necessary polarity. Table 4.1 provides a detailed description of each attribute in the dataset. Here, the term "Comment" refers to user-provided language that expresses their thoughts about the Turkey Earthquake of 2023 and its effects on the economy. These comments have one of three sentiments: 0, 1, or 2. A score of 0 indicates negativity, a score of 1 indicates positivity, and a score of 2 indicates neutrality. Whether the statement was made before or after the earthquake is indicated by the second property, Time. In this case, 0 indicates remarks made prior to the earthquake, 1 indicates remarks made during the earthquake, and 2 indicates remarks made following the earthquake. The Region feature helps to explain regional differences in public opinion by indicating the comment's geographic origin. In this case, 1 represents comments from Turkey and 0 represents comments from other countries. Price_Hike_Relatability, the last characteristic, establishes whether or not the comment is related to the price increases brought on by the earthquake. Here, a 1 indicates that the criticism is about the price increase, while a 0 indicates it is not.

4.3 Exploratory Data Analysis

The main goal when we start Exploratory Data Analysis (EDA) is to get a clear picture of how the data is structured. At this early stage, we dive deep to fully understand what's in our dataset. A thorough EDA helps us spot any odd data points, known as outliers. In our research, the EDA gave us a detailed look at things like whether the data is good quality and if everything's there that should

Table 4.1: Attribute Description from the Dataset

Attribute Name	**Attribute Description**
Comments	Contains textual comments from social media users and news portal readers in English.
Sentiment	Represents the polarity of the comments such as Positive, Negative and Neutral.
Time (Before, After)	Indicates the period when the comment was made in relation to the earthquake.
Region	Specifies the geographic origin of the comment
Price_Hike_Relatability	Indicates whether the comment is related to the price hike following the earthquake.

be. We have also used a bunch of different ways to show the data, which helps us see patterns more clearly. On top of that, we have carefully picked out the steps to get the data ready for analysis. Preprocessing steps are also carefully identified with the aid of available programming libraries. We have summarized all the key points we found from our data exploration in the next few sections.

4.3.1 Statistical Analysis of the Dataset

The dataset's statistical overview starts by outlining key attributes such as mean, standard deviation, minimum, and maximum values. Table 4.2 provides a detailed summary using these statistical measures. During this process, the "Comment" attribute was excluded, as it holds no relevance to the statistical analysis. Understanding the dataset through these descriptive statistics offers several advantages. Primarily, it provides a clear and comprehensive insight into the data.

Table 4.2: Dataset Description

Statistic	**Sentiment**	**Time (Before, During, After)**	**Region**	**Price Hike Relatability**
Count	5,052	5,052	5,052	5,052
Mean	1.00	1.55	0.76	0.65
Std	0.81	0.65	0.43	0.48
Min	0.00	0.00	0.00	0.00
25%	0.00	1.00	0.00	0.00
50%	1.00	2.00	1.00	1.00
75%	1.00	2.00	1.00	1.00
Max	2.00	2.00	1.00	1.00

The correlation matrix presented in Table 4.3 provides insights into the relationships between key attributes in the dataset. Notably, The primary correlation observed is between sentiment and time, suggesting that the earthquake had a significant

impact on public sentiment. Additionally, The moderate negative correlation between sentiment and price hike relatability indicates that price hikes were a concern and were more likely to obtain negative sentiments. The strong correlation between region and price hike relatability suggests that Turkey might be more affected by price increases or have more public discussion about them.

Table 4.3: Correlation Among Data Points

Attributes	**Sentiment**	**Time (Before, After)**	**Region**	**Price Hike Relatability**
Sentiment	1.000	-0.500	-0.400	-0.600
Time (Before, After)	-0.500	1.000	0.300	0.600
Region	-0.400	0.300	1.000	0.700
Price Hike Relatability	-0.600	0.600	0.700	1.000

4.3.2 Data Quality Checking using Kappa Score

In our study, we have applied the Fleiss' Kappa score as a way to quantify the agreement among multiple raters. This score is highly effective for monitoring and assessing how well raters are in sync. It serves as an effective method to ensure the consistency and reliability of the data we are analyzing. Moreover, it aids in decision-making by illustrating the degree of agreement present. In fields like socio-economics, Fleiss' Kappa is particularly useful for identifying areas of inconsistency. In our analysis, we used Fleiss' Kappa to measure the level of agreement among different reviewers, calculated using the following mathematical formula.

Table 4.4: Interpretation of Fleiss Kappa Score

Kappa Score Range	**Interpretation**
0.01-0.09	Poor Agreement
0.10 - 0.20	Slight Agreement
0.21 - 0.40	Fair Agreement
0.41 - 0.60	Moderate Agreement
0.61 - 0.80	Substantial Agreement
0.81 - 1.00	Almost Perfect Agreement

The formula for Kappa score for three persons is:

$$\kappa = \frac{\overline{P} - P_e}{1 - P_e}$$

Where:

- $\overline{P}$ is the average observed agreement across all raters.

- P_e is the expected agreement, calculated as the probability of random agreement.

Table 4.4 represents the values associated with the data quality. In our case, the Fleiss' Kappa score is 83%, which resides in the category of almost perfect agreement. This reflects that the inter-annotator agreement is strong enough to process the dataset. Consequently, the comments in the dataset can be reliably fed into various models.

4.3.3 Data Annotation and Assistance

For the purpose of our research, particularly in annotating the data, we received valuable assistance from Sumaiya Hoque, who completed her Master's degree in English Literature from International Islamic University of Chittagong (IIUC). Her expertise in understanding language proved instrumental in the annotation process.

Sumaiya Hoque helped us formulate the guidelines for annotation, ensuring that sentences were accurately categorized based on sentiment—negative, positive, or neutral. These guidelines, which formed the foundation for consistent and reliable data labeling, were used by the team to assess social media comments related to the Turkey Earthquake 2023 and its economic consequences. Her contribution greatly enhanced the quality of our dataset.

Rules for Sentiment Annotation

The following rules were applied during the annotation process to classify sentences into different sentiment categories:

Negative Sentiment (0):

- **Criticism of price hikes:** Sentences that express frustration or complaints about rising prices due to the earthquake.
- **Expression of loss or damage:** Sentences describing personal or collective loss (economic or physical) caused by the earthquake.
- **Negative emotions:** Sentences containing words or phrases indicating sadness, anger, or frustration related to the earthquake or its consequences.
- **Blame or dissatisfaction:** Sentences blaming authorities, governments, or businesses for inadequate response or failure to manage post-earthquake economic challenges.

Positive Sentiment (1):

- **Supportive or hopeful statements:** Sentences expressing hope for recovery or acknowledging positive actions taken to mitigate the earthquake's effects.
- **Praise for resilience or aid efforts:** Sentences that praise individuals, communities, or organizations for their efforts in helping others or managing the crisis.

- **Gratitude:** Sentences showing appreciation for aid, international help, or any form of assistance during or after the earthquake.

Neutral Sentiment (2):

- **Informational or factual:** Sentences that simply provide information or state facts about the earthquake without expressing any emotional tone or opinion.
- **Balanced viewpoint:** Sentences that mention both negative and positive aspects of the earthquake's impact without leaning toward a strong emotional response.
- **Discussion without judgment:** Sentences that discuss the situation objectively, focusing on the event itself or consequences without displaying clear sentiment.

We would also like to thank a few more individuals who helped in the annotation process voluntarily. They are: Tanvir Rahman, Lecturer, Stamford University Bangladesh; Nahid Hasan, Lecturer, Southeast University; Nazrul Islam, Lecturer, Southeast University; Mashiwat Tabassum Waishy, Lecturer, Stamford University Bangladesh; Sovon Chakraborty, Lecturer, ULAB; and Shanta Maria Shithil, Lecturer (On Leave), Stamford University Bangladesh. Their contributions were essential in ensuring the quality and consistency of the data annotations used in this research.

4.3.4 Data Exploration and Analysis

First, we examined the timing of the social media comments related to the 2023 Turkey earthquake. Three periods were considered: before, during, and after the event. The primary focus was on the volume of comments made during each period. As shown in the data, the majority of comments were made after the earthquake, with a total of 3174 comments. During the earthquake, there were 1453 comments, and before the event, only 425 comments were recorded. This distribution highlights a significant increase in social media activity following the earthquake, reflecting a common trend where public engagement spikes in response to major events. The data suggests that the earthquake had a considerable impact on people, leading to a notable increase in comments in the aftermath of the disaster.
Figure 4.2 illustrates the number of comments made before, during, and after the earthquake, emphasizing the increased activity in the post-event period. Additionally, this analysis provides insight into how public attention and social media discussions are heavily influenced by significant events such as natural disasters, with the aftermath period showing the highest levels of engagement.

Then we considered the relatability of social media comments to the price hike, focusing on two categories: comments related to the price hike and comments made in general. Figure 4.3 illustrates an interesting trend observed in the dataset. It is evident from the data that a significant portion of the comments, 3277 to be precise, are related to the price hike. This indicates that the price hike has been a prominent topic of discussion among the commenters. In comparison, there are 1775

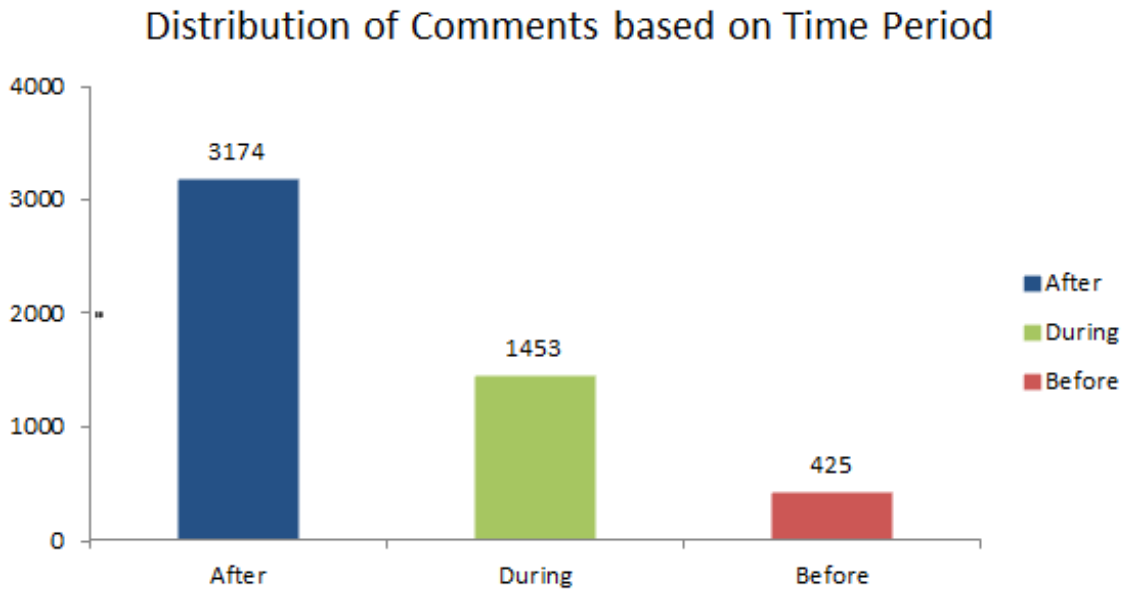


Figure 4.2: Distribution of Social Media Comments by Time Period

comments made in general, not specifically addressing the price hike. This disparity highlights the considerable impact and concern that the price hike has generated among the public. The large number of comments related to the price hike suggests that economic issues are a major concern for the population, prompting more people to express their opinions and experiences. This could also indicate sensitivity to economic changes and a greater awareness of financial matters among the general public. Understanding this distribution can help stakeholders measure the public sentiment and address the underlying issues more effectively.

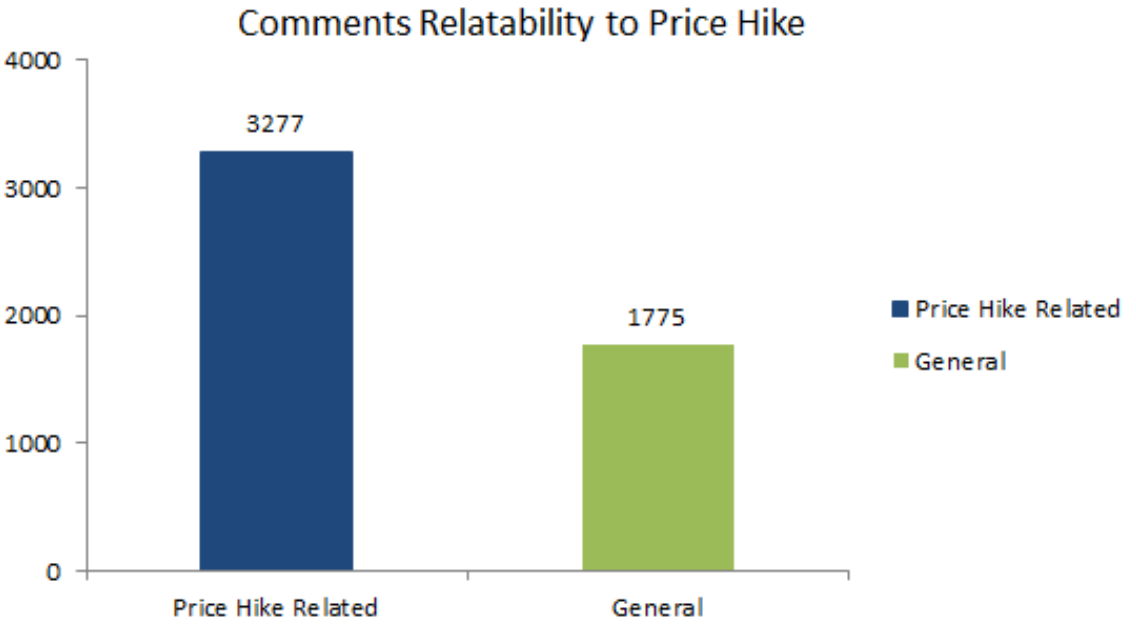


Figure 4.3: Comments Repeatability to Price Hike

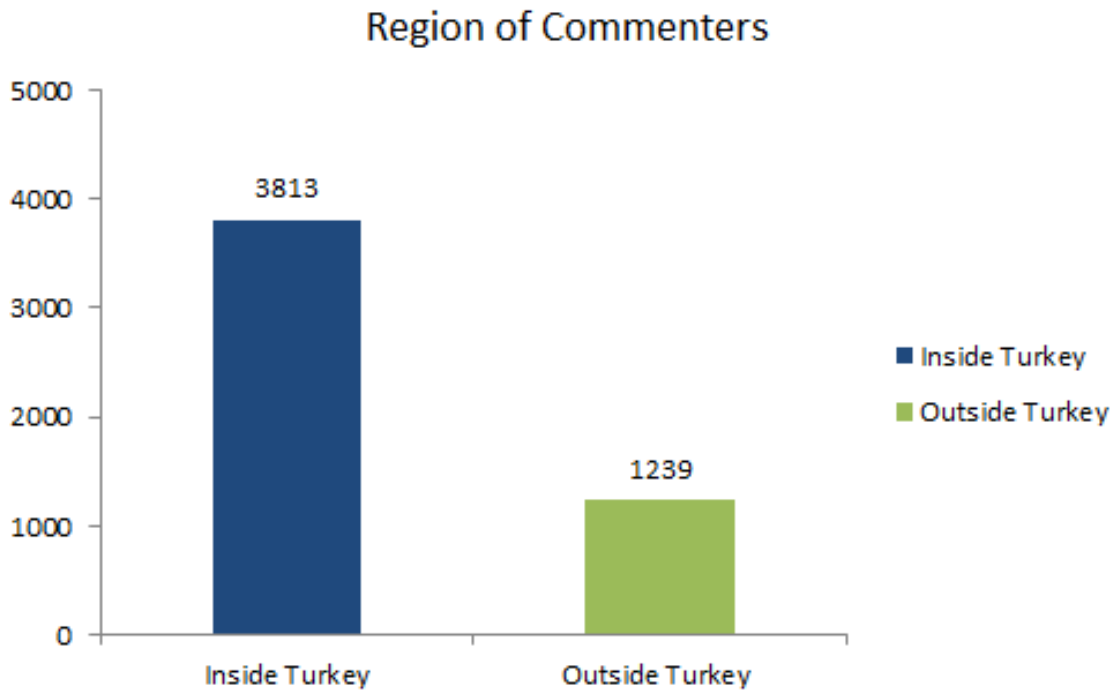


Figure 4.4: Region of Commenters

We are now examining the geographic distribution of social media comments, focusing on two categories: comments from Turkish people and comments from individuals outside Turkey. Figure 4.4 presents an insightful analysis based on the data. From our dataset, it is clear that a majority of the comments, specifically 3813, are from Turkish people. This suggests that the events and issues being discussed have a strong local impact, prompting a significant response from those directly affected. In comparison, there are 1239 comments from individuals outside Turkey, indicating considerable international interest and concern as well. The higher number of comments from Turkish people highlights the direct connection and relevance of the issues to the local population. This can be attributed to the immediate impact and personal experiences that prompt more extensive commentary from those within the country. On the other hand, the significant number of comments from outside Turkey represents the global awareness and solidarity regarding the situation.

Data are collected from different social media platfroms as We mentioned earlier that. 4.5 shows the details break down number of sources in the dataset.

Lastly, in our analysis, we initially had a dataset of approximately 5500 comments, predominantly labeled with a higher number of negative sentiment values. To address the issue of class imbalance, we chose to remove some of the negative comments, ensuring an equal representation of negative, positive, and neutral sentiments. As a result, the final counts are 1685 negative, 1684 positive, and 1683 neutral comments. This decision was made to avoid the use of resampling techniques, as we aimed to retain all unique data points. Maintaining unique data is beneficial because it preserves the diversity of opinions and experiences, which can lead to more generalizable insights. By keeping a balanced dataset without losing the richness of the original comments, we enhance the model's ability to understand and interpret

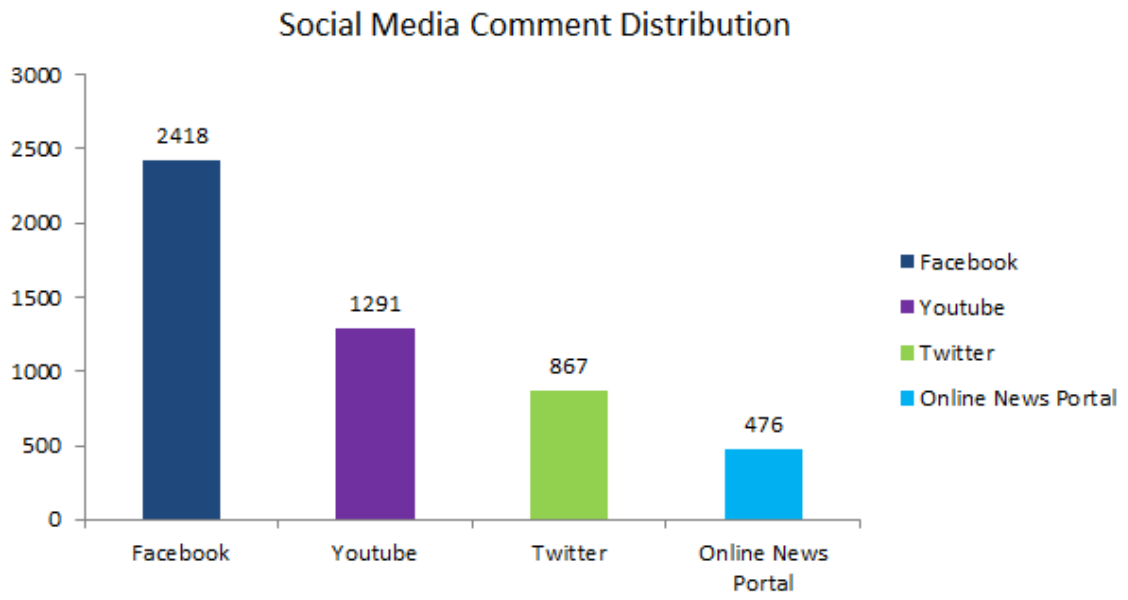


Figure 4.5: Ratio of number of sources in the dataset

sentiment accurately across different contexts.

Understanding this distribution of comments helps to measure both local and international sentiments, providing valuable insights into how the events are perceived by different groups. This information can be crucial for policymakers, organizations, and other stakeholders in addressing the concerns and responses effectively.

Thus, the above explorations conclude some important facts that are

1. The majority of social media comments regarding the 2023 Turkey earthquake were made after the event, indicating a significant surge in public engagement following the disaster.

2. Comments related to the price hike outnumbered general comments, highlighting substantial public concern and discussion surrounding economic issues in the dataset.

3. Turkish people contributed a higher number of comments compared to individuals outside Turkey, indicating strong local impact and international interest in the events discussed.

In the next phase, researchers have focused on preprocessing the comments.

4.4 Preprocessing of the dataset

Preprocessing data is essential for supplying data to several architectures so they can accurately interpret semantics. It has a sizable amount of operations that are possible on the dataset's pattern and language. Making sure the abnormality in the dataset does not lead the model to become overfit is another crucial step. There are several different preprocessing methods available for English. Because of this, we have only included the essential processes needed to offer the applied models.

The adopted preprocessing techniques are stated below:

1. **Dropping Null Values:** The primary task is to remove any null values from the dataset. We have used the Python NLTK library for various operations. This step is essential for ensuring data integrity before feeding it into machine learning (ML), deep learning (DL) models and transformer based models.

2. **Lowercasing:** To ensure uniformity and avoid case sensitivity issues, all text is converted to lowercase. This preprocessing step is applied explicitly in both ML and DL models, helping to standardize the input data and making it easier for the models to recognize and analyze words without distinguishing between different cases (e.g., "Apple" and "apple" are treated as the same word). This simplification contributes to more accurate sentiment analysis and reduces the dimensionality of the feature set.

3. **Removing Stopwords and Tags:** In the English language, stopwords are common words that do not add significant meaning to the text and can create ambiguity in analysis. Therefore, all stopwords have been removed to enhance the clarity of the data. This step is explicitly applied in both ML and DL models. Additionally, HTML tags have been eliminated from the text for all the models, as they do not provide meaningful information for the analysis and can interfere with understanding the content.

4. **Special Character and Punctuation Removal:** All special characters from the comments have been removed. Moreover, punctuation marks have been eradicated from the sentences since they also do not contribute meaningful information in this context. This preprocessing step is explicitly applied in both ML and DL models, ensuring cleaner input data for analysis.

5. **Emoji Conversion:** In our sentiment analysis, while emojis visually express emotions, they do not significantly enhance the understanding of the text's sentiment. By converting emojis into their corresponding text descriptions, we simplify the analysis while retaining their expressive value. This approach is applied explicitly in every ML and DL and transformer based models, allowing them to consider both words and visual emotions, potentially improving the sentiment analysis's accuracy and efficiency.

6. **Creating Dictionary:** A dictionary has been created to identify unique words. The word definitions are also created using this dictionary. There are 29,172 words available in the dataset, with 3,854 unique words. After closely understanding the words, it can be observed that most of them are related to earthquakes and their aftereffects.

7. **Lemmatizing:** In our data preprocessing workflow, lemmatization is an essential step for normalizing the text data. By converting words to their base or dictionary forms, lemmatization provides consistency and lowers the dataset's dimensionality. We apply lemmatization explicitly in both ML and DL models to treat various inflected forms of a word as a single entity, thereby improving the coherence of our feature set. This preprocessing step facilitates subsequent text analysis tasks and enhances model performance by minimizing noise and improving data representativeness.

8. **Word Embedding:** The cardinal purpose of using word embedding is to represent data in a dense vector in a vector space. Learning distributed representations of words based on their context in a sizable corpus of text data is the fundamental target behind word embedding. In this research, we have utilized word embedding techniques such as Word2Vec for both ML and DL models. Additionally, we explored GloVe and TF-IDF embeddings in these models; however, the results indicated that Word2Vec outperformed both GloVe and TF-IDF in terms of accuracy and relevance. For the DistilBERT and XLNet models, no additional embedding technique has been adopted as these models incorporate their own contextual embeddings into their architecture.

4.5 Hyperparameter Details of Deep Learning Models

In our research, The LSTM model utilizes 64 and 32 LSTM units, an embedding vector length of 64, and three dense layers. Dropout is set to 0.4, with recurrent dropout at 0.25, and a batch size of 16. The model trains over 50 epochs with a validation split of 0.2. Lastly, Early stopping is utilized to avoid overfitting by terminating the training process once the validation loss ceases to show improvement. The parametric details of the LSTM architecture that we impelemented are presented in Table 4.5.

Hyperparameter Name	**Value**
Number of epoch	50
Activation function	Softmax
LSTM Units	64, 32
Embedding vector length	64
Dropout	0.4
Recurrent Dropout	0.25
Dense Layers	3
Batch Size	16
Validation Split	0.2

Table 4.5: Hyperparameter details of LSTM

The Gated Recurrent Unit (GRU) model shares a similar configuration, using 64 and 32 units in its GRU and three dense layers. The embedding vector length is also 64, with a dropout rate of 0.4 and a recurrent dropout rate of 0.15. The GRU model employs a smaller batch size of 8 while training for 50 epochs with a 0.2 validation split.
Table 4.6 provides the GRU's parametric details that we have implemented.
For both the BiLSTM and BiGRU models, we employ Word2Vec for word embedding, which converts text data into continuous vector representations that capture semantic meanings. The models are trained using Stratified K-Fold cross-validation to ensure robust performance and generalizability. Early stopping is also implemented to prevent overfitting by halting training when the validation loss stops improving.

Hyperparameter Name	Value
Number of epoch	50
Activation function	Softmax
GRU Units	64,32
Embedding vector length	64
Dropout	0.4
Recurrent Dropout	0.15
Dense Layers	3
Batch Size	8
Validation Split	0.2

Table 4.6: Hyperparameter details of GRU

Hyperparameter Name	Value
Number of epoch	50
Activation function	Softmax
LSTM Units	128, 64, 32
Dense Layers	3
Dropout	0.4
Batch Size	16
Validation Split	0.2
Number of Folds	5

Table 4.7: Parametric details of BiGRU and BiLSTM

Table 4.7 reflects the parametric details of the BiGRU and BiLSTM architecture. i want whole of this in a more precise and standard structure. keep the table same. just upgrade the write up

4.6 Modified XLNet Model

We opted for the XLNet architecture to tackle sentiment analysis classification tasks, harnessing the model's robustness and adaptability to sequence-based tasks. XLNet, an extension of the transformer architecture, introduces permutation-based training, enhancing its ability to capture bidirectional context and dependencies within the text data. This model was trained and evaluated on a dataset of textual comments, with preprocessing steps including tokenization using the XLNet tokenizer. The data was then divided into training and testing sets to prepare for model training.

The XLNet model architecture consists of a pre-trained XLNet model for sequence classification, augmented by additional layers for classification. After tokenization, the data is fed into the XLNet model, and the resulting representations are passed through a classifier layer to predict sentiment labels. During training, we utilized an AdamW optimizer with a fixed learning rate, along with a cross-entropy loss function to optimize the model's parameters.

To handle the data efficiently, we employed DataLoader instances to manage batch processing during both training and evaluation phases. The model was trained over

130 epochs, with the training loss monitored and recorded for each epoch to assess convergence. After training, the model's performance was evaluated on the test set, and classification metrics such as precision, recall, and F1-score were computed using a classification report.
Table 4.8 shows the necessary changes made to the original XLNet architecture.

Parameter Name	**Value**
Maximum sequence length	Tokenizer-specific, adjusted dynamically
Number of epochs	30
Batch Size	16
Learning rate	2e-5
Number of classes	3 (negative, positive, neutral)
Dropout rate	0.2

Table 4.8: Parametric details of the modified XLNet model

By adopting XLNet for sentiment analysis and implementing these changes, we aimed to leverage the model's sophisticated architecture and training mechanisms, enhancing its capability to capture and classify sentiments effectively from textual data.

4.7 Modified DistilBERT Model

We employed a DistilBERT-based architecture for sentiment analysis classification tasks, taking advantage of the efficiency and performance of transformer-based models. DistilBERT, a compact and faster variant of BERT, retains effective language understanding capabilities while being computationally efficient. The model was trained and evaluated on a dataset of text data, which was pre-processed by splitting into training and testing sets, followed by tokenization using the DistilBERT tokenizer.

The text data was tokenized with the DistilBERT tokenizer, converting the text into input IDs and attention masks suitable for the model. The tokenized text was then used to create datasets for input to the model, with a maximum sequence length of 128. A custom dataset class, TurkeyEarthquakeDataset, was defined to handle the encodings and labels.

The model architecture includes an initial DistilBERT layer followed by a pre-classifier and classifier layer, with dropout applied for regularization. The DistilBERT model was fine-tuned for sentiment analysis, with three classes: negative, positive, and neutral. The model was trained using the AdamW optimizer with a learning rate scheduler to adjust the learning rate dynamically throughout the training process.

Table 4.9 represents the hyperparameter details of modified DistilBERT model.
The working procedure of the modified DistilBERT model is explained in 4.6.

Training was conducted over 30 epochs with a batch size of 16. The training process included calculating the loss using the CrossEntropyLoss function, backpropagation,

Parameter Name	Value
Maximum sequence length	128
Number of epochs	30
Batch Size	16
Learning rate	5e-5
Number of classes	3 (negative, positive, neutral)
Dropout rate	0.2 (sequence classification)

Table 4.9: Parametric details of the modified DistilBERT model for sentiment analysis

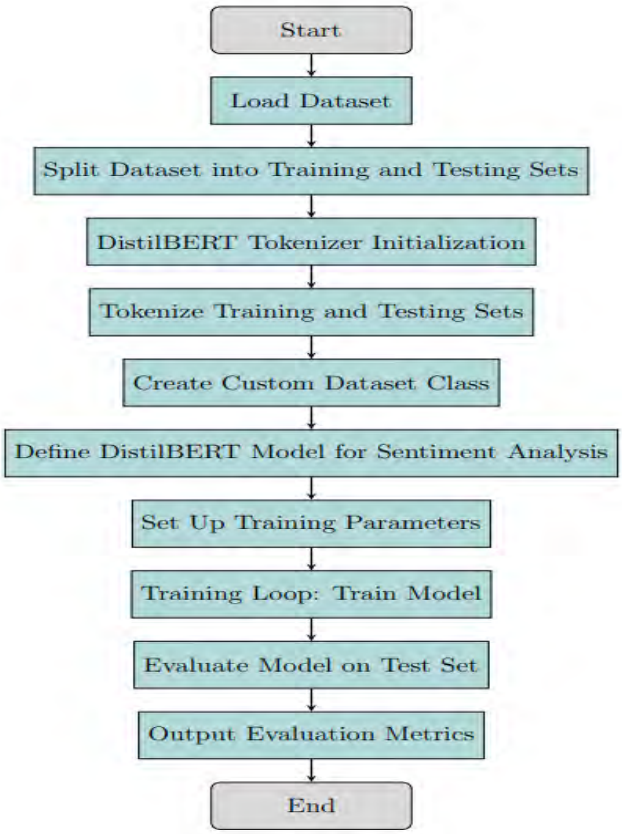


Figure 4.6: Workflow of Modified DistilBERT Model

and updating model parameters using the optimizer. A linear learning rate scheduler was employed to manage the learning rate.

Figure 4.7 shows the methodology of the modified model.
Evaluation of the model was performed on the test set, with predictions compared to true labels to measure classification performance. The results were evaluated using accuracy and classification reports. The training and evaluation process demonstrated the model's ability to effectively capture and classify sentiments from text data, leveraging the power of transformer-based models while optimizing for computational efficiency.

The implementation and training of the DistilBERT model for sentiment analysis showcased its capability to efficiently capture and classify sentiments from text

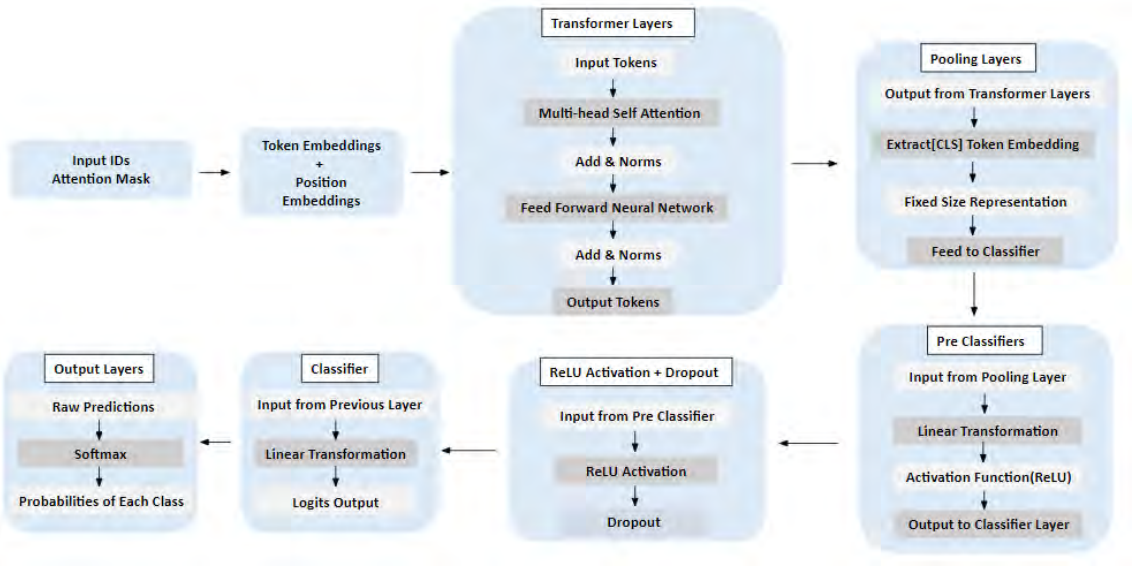


Figure 4.7: Methodology of Modified DistilBERT Model

data, demonstrating the utility of transformer-based models in natural language processing tasks.
Table 4.10 refers to the modification that we made in the DistilBERT model.

4.8 Performance Metrics

Although various performance metrics have been used in the field of sentiment analysis, our primary focus has been on metrics that directly reflect the model's classification abilities—namely, precision, recall, accuracy, F1-score, data loss, and the number of trainable parameters. In the case of deep learning and transformer-based architectures, particularly for multiclass classification tasks like ours, relying solely on precision, recall, and accuracy can be insufficient for a comprehensive evaluation. Therefore, we have also employed the macro F1-score to provide a more holistic view of our model's performance across different sentiment classes.

The formulas for precision, recall, accuracy, F1-score, and macro F1-score for multiclass classification are provided below.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4.1}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4.2}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \tag{4.3}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

Aspect	Traditional DistilBERT	Modified DistilBERT
Dataset Splitting	Not specified	Explicitly splits the dataset into train and test sets using `train_test_split`
Tokenizer Initialization	Uses `DistilBertTokenizer`	Same as traditional
Tokenization	Tokenizes input data using `DistilBertTokenizer`	Same as traditional
Dataset Class	Uses standard data loading and tokenization methods	**Custom `TurkeyEarthquakeDataset` class to handle encodings and labels**
DataLoader Batch Size	Typically varies, often larger	Uses a batch size of 16
Model Definition	DistilBertForSequenceClassification from Hugging Face	**Custom `DistilBERTForSentimentAnalysis` class with additional pre-classifier layer and ReLU activation**
Pre-Classifier Layer	Not present	**Added pre-classifier layer with ReLU activation**
Dropout Rate	Defined within `DistilBertConfig`	**Custom dropout defined within the model class**
Optimizer	Uses `AdamW`	Same as traditional
Learning Rate Scheduler	Uses learning rate scheduler	Same as traditional
Training Loop	Standard training loop	**Custom training loop with explicit loss calculation and backward pass**
Loss Function	Cross-Entropy Loss (implicit within Hugging Face model)	**Cross-Entropy Loss (explicitly calculated using `nn.CrossEntropyLoss()`)**
Evaluation Method	Standard evaluation loop	Custom evaluation loop with accuracy and classification report
Training Epochs	Typically varies	Set to 30 epochs

Table 4.10: Differences between Traditional DistilBERT and Modified DistilBERT Code

$$\text{Macro F1 score} = \frac{1}{n}\sum_{i=1}^{n} \text{F1 score for class } i \tag{4.5}$$

Evaluation Metrics for Multiclass Classification

Given that our task involves three-class classification (negative, positive, and neutral sentiments), we have opted to use macro F1 scores for evaluation. The macro F1-score treats all classes equally, providing insight into how the model performs across all sentiment classes without considering their frequency. Since our dataset has been balanced (with 1685 negative, 1684 positive, and 1683 neutral sentences), utilizing this metric allows us to monitor overall performance while ensuring fair treatment across all classes.

4.9 Training Set Up

For the training of DistilBERT and XLNet models, we utilized cloud-based platforms to ensure efficient and uninterrupted execution of our experiments. Specifically, we used Google Colab and Kaggle, which provide access to high-performance GPUs and

other necessary resources without the need for local high-configuration hardware components.

Google Colab

Google Colab, a free Jupyter notebook environment provided by Google, was one of the primary platforms used. It offers:

1. **GPU and TPU:** Support Access to high-performance NVIDIA GPUs and TPUs, allowing for faster training and inference.
2. **Ease of Use:** A user-friendly interface that facilitates quick setup and execution of machine learning experiments.
3. **Integrated Libraries:** Pre-installed libraries and tools essential for training models, such as TensorFlow, PyTorch, and the Hugging Face Transformers library.

Kaggle

Kaggle, a platform known for its vast data science community and resources, was also employed in our research. Key features include:

1. **GPU Acceleration:** Availability of high-performance NVIDIA GPUs similar to Google Colab, ensuring efficient model training.
2. **Datasets and Kernels:** Access to numerous datasets and pre-built kernels, which significantly accelerated the experimentation process.
3. **Community Support:** A supportive community where we could share insights and troubleshoot issues with fellow researchers.

Rationale for Using Cloud Platforms:

Initially, we attempted to train our models on a local setup using Jupyter Notebook without GPU support. However, this approach led to frequent kernel disconnections and significantly slower performance, making it impractical for our needs.
By leveraging these cloud platforms, we ensured that our models were trained effectively, monitored performance metrics, and optimized the architecture without the limitations posed by local hardware.

Chapter 5

Experimental Result Analysis

5.1 Result Analysis

In this section, we will present and analyze the results obtained from various machine learning (ML) models, deep learning (DL) models, and Transformer-based models. We applied several traditional machine learning algorithms, including K-Nearest Neighbors (KNN), Decision Tree Classifier, XGBoost, Support Vector Machines (SVM) and Random Forest to our dataset. Each model's performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. A comparative analysis of these models will be discussed to identify the most effective model for our classification tasks.

In addition to traditional machine learning models, we implemented deep learning architectures to leverage their ability to capture complex patterns in the data. This section will cover the performance evaluation of various deep learning models, highlighting their strengths and weaknesses in handling our specific dataset.

Transformer-based models, such as DistilBERT and XLNet, were employed due to their state-of-the-art performance in natural language processing tasks. These models were fine-tuned on our dataset, and their results were compared to those of the machine learning and deep learning models. We will discuss the effectiveness of these models in capturing the insights of our data and their overall performance metrics.

5.1.1 Result Analysis of Machine Learning Models

In our study, we applied several machine learning models to a given dataset to observe and compare their performance. The models tested included Support Vector Machine (SVM), Random Forest, Decision Tree Classifier, XGBoost, and K-Nearest Neighbour (KNN). Below is a detailed summary of our findings based on the average precision, recall, F1-score, and accuracy of each model.

As shown in Table 5.1, we first trained and tested the K-Nearest Neighbour (KNN) model. The KNN model achieved an average accuracy of 60.24%, with an average precision of 65.60%, recall of 64.60%, and an F1-score of 65.10%. KNN performed well in terms of recall for neutral sentiments but was slightly behind the Random Forest and SVM models in overall accuracy. Next, we evaluated the Decision Tree

Classifier, which showed an average accuracy of 67.40%. It achieved an average precision of 70.90%, recall of 70.3%, and an F1-score of 70.60%. The Decision Tree model exhibited the highest recall for positive sentiments among the models, though its overall accuracy was lower than both the SVM and Random Forest models. We then tested the XGBoost model, which resulted in an average accuracy of 67.30%. The model's average precision was 70.52%, recall was 70.20%, and F1-score was 70.36%. XGBoost demonstrated balanced performance across all categories but did not surpass the Random Forest in terms of overall accuracy. After that, we trained and tested the SVM model. The SVM demonstrated a reasonable performance with an average accuracy of 71.32%, and it achieved an average precision, recall, and F1-score of 72.60%, 72.9%, and 72.20%, respectively. While the SVM exhibited high precision for positive sentiments, its recall for the same category was relatively lower. Lastly, we evaluated the Random Forest model, which delivered an average accuracy of 71.60%. It showed balanced performance across all sentiment categories, with average precision and recall both at 74.12%, and an average F1-score of 73.76%. The Random Forest model had higher recall for positive sentiments compared to the SVM.

Our comprehensive analysis revealed that the Random Forest model exhibited the highest and balanced performance across all sentiment categories. The SVM also demonstrated strong performance, particularly in terms of precision and recall for positive and neutral sentiments. The Decision Tree and XGBoost models, while effective, did not outperform the Random Forest model in overall accuracy. These findings suggest that Random Forest is a good choice for sentiment classification tasks in this context.

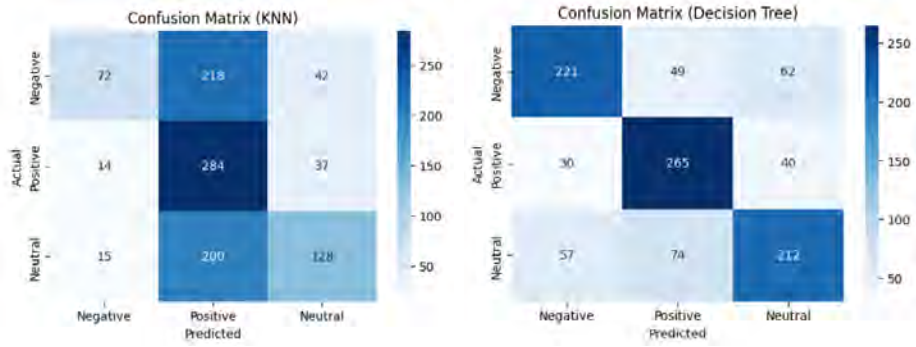
Table 5.1: Performance Metrics for Different Machine Learning Models

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
KNN	60.24%	65.60%	64.60%	65.10%
Decision Tree	67.40%	70.90%	70.3%	70.60%
XGBoost	67.30%	70.52%	70.20%	70.36%
SVM	71.32%	72.60%	72.9%	72.20%
Random Forest	71.60%	74.12%	73.4%	73.76%

Now we present the confusion matrices for the various machine learning models applied to our sentiment analysis task in Figure 5.1. These confusion matrices provide a visual representation of the model's performance by showing the distribution of predicted and actual sentiment classes. The models included in this analysis are K-Nearest Neighbors (KNN), Decision Tree, XGBoost, Support Vector Machine (SVM), and Random Forest. By comparing these matrices, we can assess the strengths and weaknesses of each model in terms of accurately classifying the sentiment categories: Negative, Positive, and Neutral.

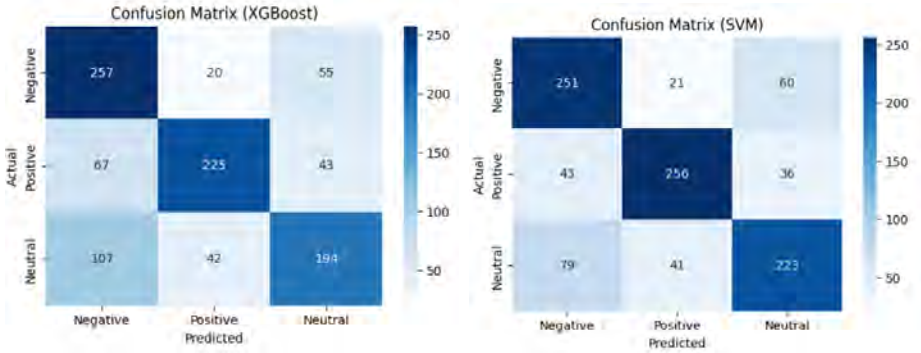
5.1.2 Result Analysis of the Deep Learning Models

Several deep learning models were implemented on a given dataset to analyze and compare their performance. These models included Long Short-Term Memory



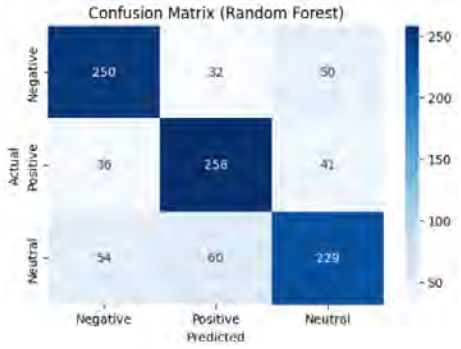
(a) Confusion Matrix of K-Nearest Neighbour

(b) Confusion Matrix of Decision Tree



(c) Confusion Matrix of XGBoost

(d) Confusion Matrix of SVM



(e) Confusion Matrix of Random Forest

Figure 5.1: Comparison of Confusion Matrices of Applied ML Models

(LSTM), Gated Recurrent Units (GRU), Bidirectional Long Short-Term Memory (BiLSTM), and Bidirectional Gated Recurrent Units (BiGRU). The detailed findings based on the average precision, recall, F1-score, and accuracy for each model are summarized below.

As shown in Table 5.2, the LSTM model was first trained and tested. The results indicated that LSTM performed well, achieving an average accuracy of 72.60%. The model had an average precision of 73.93%, recall of 74.67%, and F1-score of 74.30%. Although the LSTM showed high precision for positive sentiments, its recall for the same category was comparatively lower. Next, the GRU model was assessed, yielding an average accuracy of 72.29%. It demonstrated balanced performance across all sentiment categories, with an average precision of 73.50%, recall of 72.76%,

and an F1-score of 73.13%. The GRU model exhibited higher recall for positive sentiments than the LSTM. The BiLSTM model was also tested and achieved an average accuracy of 74.26%. It recorded an average precision of 77.36%, recall of 76.02%, and F1-score of 76.69%. The BiLSTM model had the highest recall for positive sentiments among the models and the highest overall accuracy. The BiGRU model was then evaluated, attaining an average accuracy of 73.30%. The average precision was 74.98%, recall was 74.60%, and F1-score was 74.79%. The BiGRU model demonstrated balanced performance across all categories and closely followed the BiLSTM in overall accuracy.

The analysis showed that the BiLSTM model had the highest overall accuracy and balanced performance across all sentiment categories. The BiGRU and GRU models also exhibited strong performance, especially in terms of precision and recall for positive and neutral sentiments. While the LSTM model was effective, it did not outperform the BiLSTM and BiGRU models in overall accuracy. These results suggest that BiLSTM and BiGRU are robust choices for sentiment classification tasks in this context.

Table 5.2: Performance Metrics of Deep Learning Models

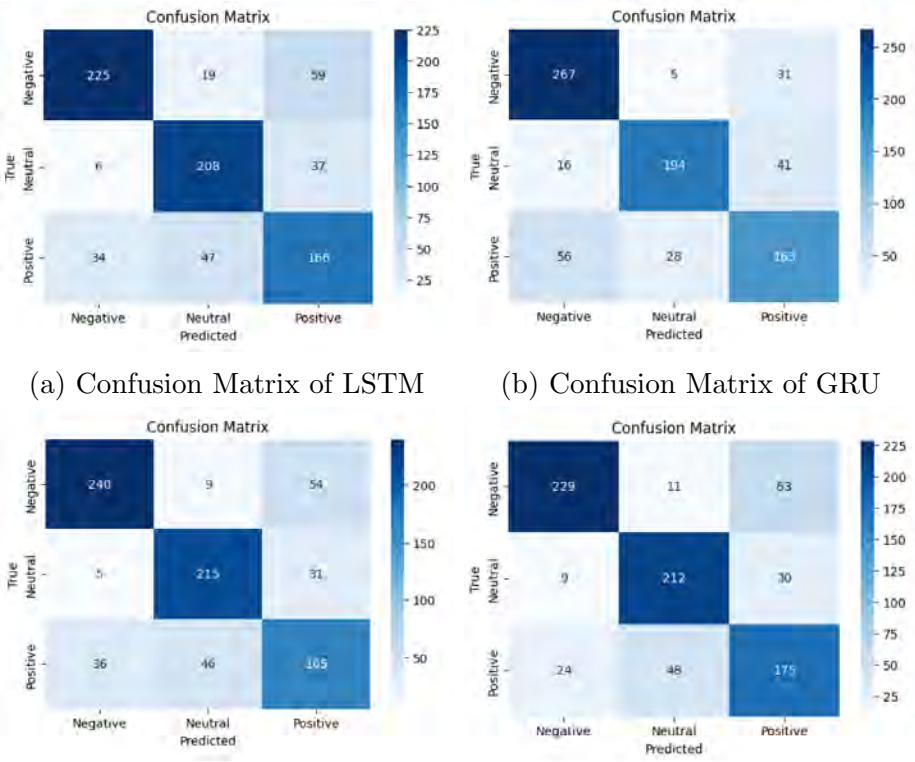
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
LSTM	72.60%	73.93%	74.67%	74.30%
GRU	72.29%	73.50%	72.76%	73.13%
BiLSTM	74.26%	77.36%	76.02%	76.69%
BiGRU	73.30%	74.98%	74.60%	74.79%

In Figure 5.2, we present the confusion matrices for the deep learning models applied to our sentiment analysis task. These matrices showcase the classification performance of Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), Bidirectional LSTM (BiLSTM), and Bidirectional GRU (BiGRU) models. By analyzing the confusion matrices, we can observe how each model predicts sentiment classes such as Negative, Positive, and Neutral. These matrices provide insights into the distribution of true and false predictions, helping us assess each model's strengths and weaknesses. Ultimately, this comparison helps identify the most suitable deep learning architecture for sentiment classification in this context.

5.1.3 Result Analysis of the XLNet Architecture

We evaluated various models on a given dataset to compare their performance. These models included traditional machine learning (ML) models, deep learning (DL) models, and a transformer-based model, XLNet. Below is a detailed summary of our findings based on the average precision, recall, F1-score, and accuracy of XLNet, along with a comparison to the ML and DL models.

The results indicated that XLNet outperformed all other models, achieving an average accuracy of 81.20%. The model also demonstrated an average precision of 82.36%, recall of 82.24%, and F1-score of 82.30%. XLNet showed high precision and recall for all sentiment categories, significantly surpassing the performance of both the ML and DL models.



(a) Confusion Matrix of LSTM (b) Confusion Matrix of GRU

(c) Confusion Matrix of BiLSTM (d) Confusion Matrix of BiGRU

Figure 5.2: Comparison of Confusion Matrices of Applied DL Models

Table 5.3 shows the result analysis of the XLNet model.

Table 5.3: Performance Metrics of XLNet Model

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
XLNet	81.20%	82.36%	82.24%	82.30%

vspace1em In this section, we present the confusion matrices for the deep learning models applied to our sentiment analysis task. These matrices showcase the classification performance of Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), Bidirectional LSTM (BiLSTM), and Bidirectional GRU (BiGRU) models. By analyzing the confusion matrices, we can observe how effectively each model predicts sentiment classes such as Negative, Positive, and Neutral. These matrices provide insights into the distribution of true and false predictions, helping us assess each model's strengths and weaknesses. Ultimately, this comparison helps identify the most suitable deep learning architecture for sentiment classification in this context.

5.1.4 Result Analysis of the Modified DistilBERT Architecture

We evaluated multiple models on a given dataset to assess their performance. These models encompassed traditional machine learning models, deep learning models, and a transformer-based model, DistilBERT. Here's a comprehensive summary of

our findings based on the average precision, recall, F1-score, and accuracy of DistilBERT, along with a comparison to the machine learning (ML) and deep learning (DL) models.

Our analysis revealed DistilBERT as the top performer among all models, demonstrating an average accuracy of 82.20%. With an average precision of 84.81%, recall of 83.79%, and F1-score of 84.30%, DistilBERT exhibited remarkable precision and recall across all sentiment categories, surpassing both the machine learning and deep learning models. The performance metrics of DistilBERT are outlined in Table 5.4.

Table 5.4: Performance Metrics of DistilBERT Model

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
DistilBERT	82.20%	84.81%	83.79%	84.30%

Table 5.5: Comparison of Performance Metrics Across ML, DL, DistilBERT, and XLNet Models

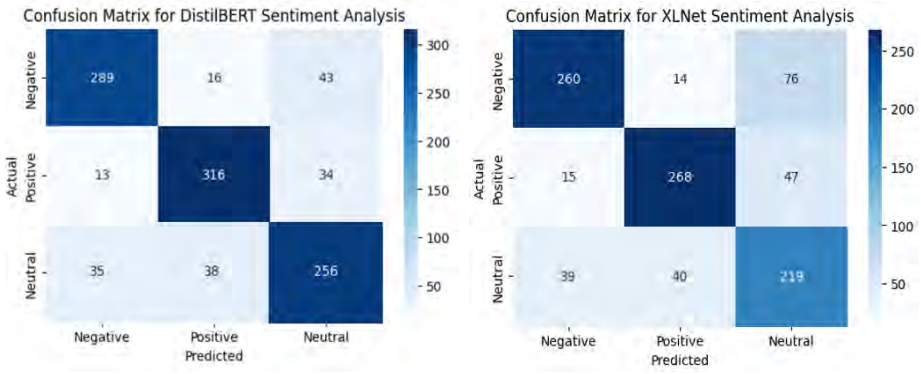
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
KNN	60.24%	65.60%	64.60%	65.10%
Decision Tree	67.40%	70.90%	70.3%	70.60%
XGBoost	67.30%	70.52%	70.20%	70.36%
SVM	71.32%	72.60%	72.9%	72.20%
Random Forest	71.60%	74.12%	73.4%	73.76%
LSTM	72.60%	73.93%	74.67%	74.30%
GRU	72.29%	73.50%	72.76%	73.13%
BiLSTM	74.26%	77.36%	76.02%	76.69%
BiGRU	73.30%	74.98%	74.60%	74.79%
XLNet	81.20%	82.36%	82.24%	82.30%
DistilBERT	**82.20%**	**84.81%**	**83.79%**	**84.30%**

Table 5.5 presents a comparison of the results of all applied models. The comparison highlights the superior performance of transformer-based models, particularly DistilBERT and XLNet, over both traditional machine learning (ML) and deep learning (DL) models across all evaluated metrics. DistilBERT achieved the highest accuracy at 82.20%, followed closely by XLNet with an accuracy of 81.20%. DistilBERT excelled with balanced and superior precision 84.81%, recall 83.79%, and F1-score 84.30%.

Here, we present the confusion matrices for the transformer-based models utilized in 5.3. These matrices provide a clear depiction of each model's performance by displaying how they classify the sentiment categories of Negative, Positive, and Neutral. Through the confusion matrices, we can evaluate how well these transformer architectures manage the complexity of our dataset.

Among the traditional ML models, Random Forest performed the best with an accuracy of 71.60% and F1-Score 73.76%.

The performance of DL models, including LSTM, GRU, BiLSTM, and BiGRU, was also examined. BiLSTM and BiGRU outperformed other DL models with accuracies



(a) Confusion Matrix of DistilBERT (b) Confusion Matrix of XLNet

Figure 5.3: Comparison of Confusion Matrices of Applied Transformer Based Models

of 74.26% and 73.30%, respectively. Despite their strong precision and recall for positive and neutral sentiments, these models did not surpass the transformer models in overall accuracy.

In conclusion, DistilBERT notably outperformed both ML and DL models, achieving an accuracy of 82.20% and also DistilBERT excelled with balanced and superior precision 84.81%, recall 83.79%, and F1-score 84.30% While XLNet also showed strong performance with an accuracy of 81.20%. Therefore, transformer-based models, especially DistilBERT, emerged as the most robust choices for sentiment classification tasks in this context.Figure 5.4 show the details of the F1-Score Comparison for all the applied Models.

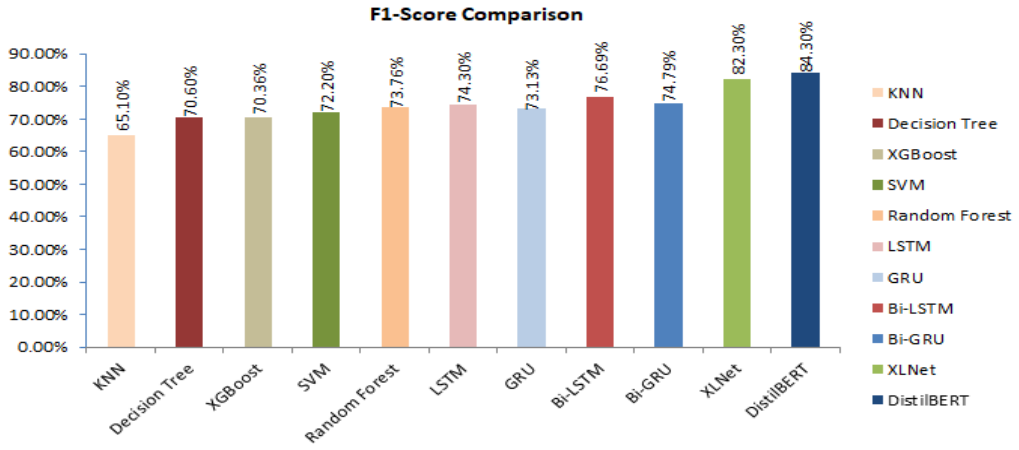


Figure 5.4: F1-Score Comparison of all the models

5.2 Performance Comparison: DistilBERT vs. XLNet Using McNemar's Test

In our sentiment analysis task, we compared the performance of two transformer-based models: DistilBERT and XLNet. Both models have demonstrated strong

capabilities in various natural language processing tasks, particularly in sentiment classification. However, our evaluation using McNemar's test reveals that DistilBERT significantly outperforms XLNet on our dataset.

5.2.1 McNemar's Test Results

McNemar's test is a statistical test used to compare the performance of two models based on the differences in their predictions. Specifically, it evaluates the disagreement between the two models on the same data points, making it a suitable choice for comparing classifier performance.
The results of the test were as follows:

- **Test Statistic**: 85.8
- **p-value**: 0.03918

Since the p-value is less than the commonly accepted threshold of 0.05, we conclude that the difference in performance between DistilBERT and XLNet is statistically significant. This indicates that the performance improvement observed with DistilBERT is unlikely to be due to random chance, confirming its superiority over XLNet in this task.

5.2.2 Interpretation of Results

The statistically significant p-value of 0.03918 suggests that DistilBERT's predictions are more accurate than those of XLNet on the same dataset. In the context of this study—sentiment analysis of social media comments following the 2023 Turkey earthquake—this result is particularly noteworthy.

- DistilBERT correctly classified a larger proportion of data points than XLNet, as indicated by the higher number of cases where XLNet made incorrect predictions while DistilBERT made correct ones.
- The McNemar's test statistic (85.8) reflects this imbalance in performance, where DistilBERT consistently outperforms XLNet in terms of predictive accuracy.

In summary, DistilBERT significantly outperforms XLNet in this sentiment analysis task, as confirmed by McNemar's test. This finding highlights the importance of model selection and suggests that DistilBERT is better suited for tasks involving informal, high-volume textual data.

5.3 Model Interpretation Using Explainable AI

In modern machine learning applications, models like DistilBERT provide powerful predictions, but their complexity can make it difficult to understand how decisions are being made. Explainable AI (XAI) techniques are designed to bridge this gap by offering insights into the model's inner workings. This is crucial for gaining trust in the predictions, ensuring fairness, and enabling informed decision-making, especially in sensitive areas such as policy making.

XAI allows us to interpret the reasoning behind individual predictions, which is particularly useful for identifying patterns, biases, and key drivers behind the results in large datasets. By implementing different XAI methods, we can uncover deeper insights into the data, helping us refine models and gain more nuanced understanding of the underlying phenomena.

In this report, we implemented two XAI techniques—LIME and SHAP—to achieve a more granular interpretation of our DistilBERT model. These methods help us analyze model behavior, providing transparency in how the model arrives at its predictions, and offering a clearer view of the factors driving sentiment classifications.

5.3.1 Local Interpretable Model-agnostic Explanations (LIME)

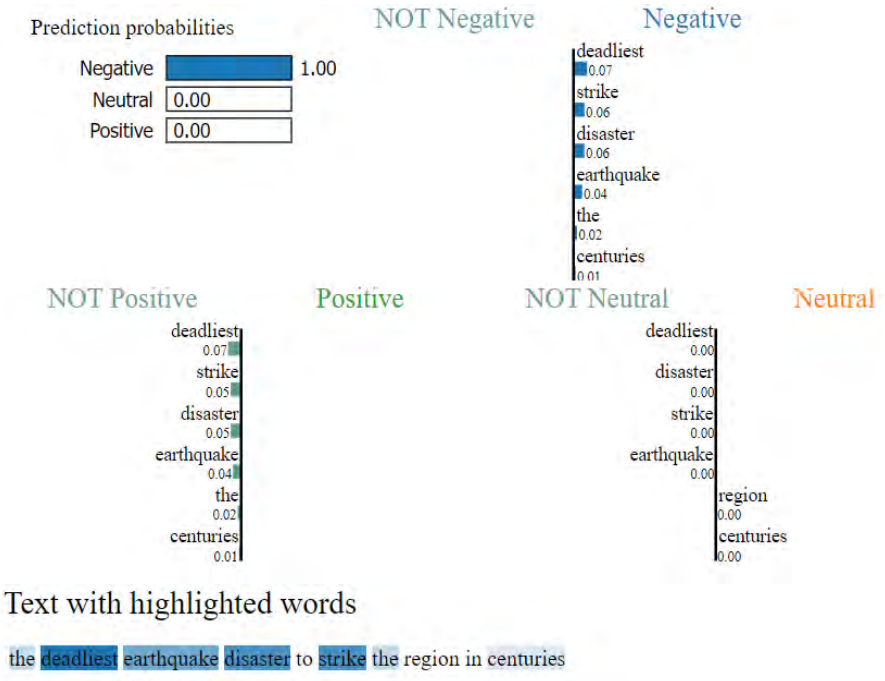


Figure 5.5: Model Interpretation Using Explainable AI - LIME

LIME (Local Interpretable Model-agnostic Explanations) is designed to explain the predictions of any model by approximating its behavior locally around a specific instance. It achieves this by perturbing the input data and observing how the model's predictions change, thereby creating a simpler, interpretable model for that specific instance.

Figure 5.5 illustrates the application of LIME to one of our test sentences. The modified DistilBERT model successfully captured the semantic meaning of the sentence and classified it as negative. LIME highlights the contributions of individual words, helping us understand why the model made this particular prediction. For example, the words *"heartbreaking"* and *"devastating"* heavily influenced the model to lean towards a negative sentiment. LIME's ability to make these localized explanations brings transparency to the decision-making process.

5.3.2 SHapley Additive Explanations (SHAP)

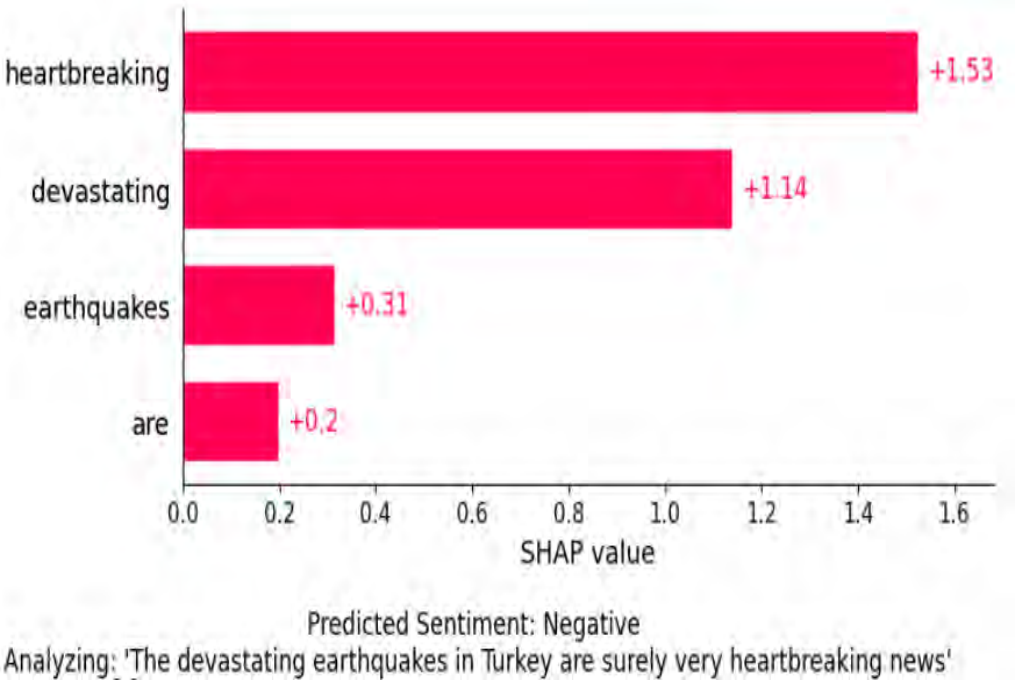


Figure 5.6: Model Interpretation Using Explainable AI - SHAP

SHAP (SHapley Additive exPlanations) provides a unified framework for interpreting model predictions by computing the contribution of each feature to the prediction. It is based on Shapley values from cooperative game theory and provides a global understanding of how features influence the output, while also explaining individual predictions.

In Figure 5.6, considering the sentence *"The devastating earthquakes in Turkey are surely very heartbreaking news"*. SHAP calculated the contribution of each word to the sentiment classification. The words *"heartbreaking"* and *"devastating"* have the highest SHAP values, indicating that they played the most significant role in driving the model's prediction toward a negative sentiment. Other words, such as *"earthquakes"* and *"are"*, had smaller but still notable contributions. These SHAP values help us quantify the importance of each feature in the final classification.

By applying both LIME and SHAP, we gained a more comprehensive understanding of the model's behavior. While LIME provides local interpretability, SHAP offers both local and global insights into feature contributions. Together, these techniques enhance transparency and help us derive more informed and actionable insights from the model's predictions.

Chapter 6

Explainable AI and Transparency in Policy Formulation

Today, public policymakers have the opportunity to make data-driven, evidence-based decisions by analyzing the vast amounts of policy-related data generated from various sources such as e-services, mobile apps, and social media. Machine learning and artificial intelligence technologies facilitate and automate the analysis of these large datasets, enabling a shift toward data-driven decision-making. However, the implementation and use of AI tools in public policy development come with significant technical, political, and operational challenges. For instance, AI-based policy solutions must be transparent and explainable to policymakers [23].

6.1 Explainable AI in Policy Making

The decision-making process has been completely transformed in recent years by the incorporation of machine learning and artificial intelligence into the formulation of public policy. These tools enable decision-makers to examine massive datasets produced by multiple sources, including social media, mobile applications, and e-services. In order to guarantee that the insights obtained from these technologies are comprehensible and practical, explainable artificial intelligence (XAI) must be employed. We used the DistilBERT model, a condensed form of the BERT model, to examine textual data from several sources in relation to the Turkey Earthquake 2023. DistilBERT, which is renowned for its effectiveness and performance in situations involving natural language processing, offered insightful information about the disaster's effects. We applied the Local Interpretable Model-agnostic Explanations (LIME) technique to improve the model's outputs' interpretability. LIME makes the AI process transparent by producing explanations for the model's predictions that are understandable to humans. Policymakers may make key evidence-based decisions for disaster-prone areas by utilizing DistilBERT and LIME.
The benefits of this approach include:

1. **Improved Decision-Making**: AI models can identify patterns and trends that may not be immediately apparent, enabling more informed and accurate policy decisions.

2. **Increased Public Trust**: The transparency provided by explainable AI fosters public trust, as policymakers can clearly communicate the rationale behind

their decisions.

Overall, the application of explainable AI in analyzing the Turkey Earthquake 2023 dataset demonstrates its potential in enhancing policy formulation for disaster management and response.

6.2 Why Transparency is Important in Policy Making?

There are various reasons why policy making must be transparent. Transparency guarantees that procedures and choices are transparent, comprehensible, and justified to all parties involved, including the general public, in the era of data-driven decision-making.

Key Reasons for Transparency in Policy Making:

1. **Accountability**: Transparent processes hold policymakers accountable for their decisions. When the decision-making process is open and clear, it is easier to identify and address any mistakes or biases.

2. **Public Trust**: Transparency fosters trust between the government and the public. When people understand how decisions are made and have access to the underlying data and rationale, they are more likely to support and comply with policies.

3. **Informed Participation**: Transparency allows for informed participation from various stakeholders, including citizens, experts, and advocacy groups. This collaborative approach can lead to more robust and well-rounded policies.

Incorporating transparency into AI-based policy making involves several strategies, including:

- **Use of Explainable AI Techniques**: Methods like LIME help in making AI models more interpretable.

- **Clear Communication**: Policymakers should clearly communicate the data, methods, and reasoning behind their decisions.

In summary, transparency is fundamental to effective and democratic policy making. It promotes accountability, builds public trust, and encourages informed participation, ultimately resulting in more successful policy outcomes.

6.3 Steps Towards Effective Data-Driven Policy Making

Data governance is essential for managing and regulating data assets to ensure accuracy, accessibility, and security, serving as the foundation of informed decision-making processes. It establishes the protocols for handling data effectively, ensuring

that the information policymakers rely on is both high-quality and trustworthy. This is especially important in the context of data-driven policy making, where decisions must be supported by transparent, reliable evidence. Proper data governance enables organizations to increase transparency and accountability in policymaking by ensuring that all decisions are based on solid, defensible data.

In our research, we ensured the collection and analysis of high-quality, reliable custom-collected data from social media responses after the 2023 Turkey earthquake. By focusing on sentiment analysis and socio-economic impacts, we provide valuable insights into public sentiment regarding price hikes, demonstrating how data governance improves policy relevance. The use of advanced analytics and explainable AI in our study further reinforces this by making model predictions transparent and accountable. This transparency is crucial in helping stakeholders understand the key factors driving the data-driven insights, a core goal of data governance.

While data-driven policy making is a large and complex task, requiring extensive analysis and comprehensive data integration, our research demonstrates a small yet significant contribution. By ensuring the proper management data, we provide a strong basis for future policy decisions that can be more accurate, equitable, and responsive to real-time public needs. Our study represents an important step in the broader effort to achieve effective data governance and impactful, evidence-based policy making.

Chapter 7

Conclusion and Future Work

We underscore the critical role of sentiment analysis in facilitating evidence-based policy-making for disaster-resilient communities, particularly in the aftermath of natural disasters. By analyzing social media data following the 2023 earthquake in Turkey, we categorized public sentiment into negative, positive, and neutral responses, leveraging advanced machine learning models such as Support Vector Machines (SVM), Random Forest, DistilBERT, and XLNet. Our findings reveal a significant correlation between public sentiment and socio-economic variables such as consumer spending, inflation, and unemployment rates. DistilBERT, in particular, demonstrated exceptional precision, recall, and F1 score, proving its efficacy as a powerful tool for understanding and guiding policy measures aimed at mitigating the socio-economic impacts of natural disasters. Explainable AI models have been used to interpret the model, and the we observed the attempted result. This research highlights the importance of incorporating state-of-the-art natural language processing (NLP) techniques in disaster management policies, ultimately contributing to more resilient and adaptive communities.

7.1 Limitations of this Research

Despite the significance of this research, several limitations remain unaddressed. Some of these limitations are outlined below:

1. **Data Quality and Representativeness:**
   (a) The sentiment analysis relies on social media data, which may not be representative of the entire affected population. Individuals without access to social media or those who choose not to express their views online are excluded, potentially leading to a bias in the sentiment captured.
   (b) The accuracy of sentiment analysis is dependent on the quality and authenticity of the social media posts. The presence of bots, fake accounts, or coordinated disinformation campaigns can distort the sentiment analysis results.
2. **Model Limitations:**
   (a) While models like DistilBERT and XLNet show high performance in sentiment classification, they may still struggle with understanding the con-

text of certain posts, especially those involving sarcasm, idioms, or local dialects.

(b) The models used were pre-trained on general corpora and might not be fully optimized for the specific context of disaster-related sentiment, which may require domain-specific training data.

3. **Temporal Dynamics:**

   (a) The timing of data collection relative to the earthquake and subsequent events may affect the sentiment captured. For instance, immediate reactions might differ significantly from sentiments expressed weeks or months later.

4. **Interpretation of Sentiment:**

   (a) Sentiment analysis tools can sometimes misinterpret the context, leading to incorrect classifications that might affect the overall findings.

By acknowledging these limitations, the research can provide a more balanced and nuanced understanding of the findings and their implications, while also highlighting areas for future improvement and investigation.

Future Work

Future research should address the limitations identified in this study to enhance the applicability of sentiment analysis in disaster management. Firstly, expanding data sources beyond social media to include traditional media, surveys, and official reports could provide a more comprehensive and representative understanding of public sentiment. Improving model accuracy by incorporating domain-specific training data and fine-tuning models to handle local dialects and idiomatic expressions is crucial. Additionally, employing advanced techniques to detect and filter out bots and fake accounts will enhance data quality. To better understand the temporal dynamics of sentiment, future studies should consider data collection and analysis to capture long-term sentiment shifts. Establishing causal relationships between sentiment and socio-economic variables using sophisticated econometric models will provide deeper insights into the impact of disasters on public perception and economic behavior. Finally, exploring more sentiment categories and additional dimensions such as fear, anger, and hope can offer a richer understanding of the emotional landscape of affected communities. These enhancements will contribute to more effective and evidence-based policy-making for disaster resilience.

Bibliography

[1] M. A. Sit, C. Koylu, and I. Demir, "Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: A case study of hurricane irma," pp. 8–32, 2020.

[2] I. Lauriola, A. Lavelli, and F. Aiolli, "An introduction to deep learning in natural language processing: Models, techniques, and tools," vol. 470, Elsevier, 2022, pp. 443–456.

[3] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[4] L. Hagen, Ö. Uzuner, C. Kotfila, T. M. Harrison, and D. Lamanna, "Understanding citizens' direct policy suggestions to the federal government: A natural language processing and topic modeling approach," in *2015 48th Hawaii International Conference on System Sciences*, IEEE, 2015, pp. 2134–2143.

[5] A. Evgenidis, M. Hamano, and W. N. Vermeulen, "Economic consequences of follow-up disasters: Lessons from the 2011 great east japan earthquake," *Energy Economics*, vol. 104, p. 105 559, 2021.

[6] W. Chung and D. Zeng, "Dissecting emotion and user influence in social media communities: An interaction modeling approach," *Information & Management*, vol. 57, no. 1, p. 103 108, 2020.

[7] S. Akar, "Natural disasters and syrian refugees: The case of kahramanmaraş earthquakes in türkiye," 2024.

[8] T. Wang, J. Chen, Y. Zhou, *et al.*, "Preliminary investigation of building damage in hatay under february 6, 2023 turkey earthquakes," *Earthquake Engineering and Engineering Vibration*, vol. 22, no. 4, pp. 853–866, 2023.

[9] M. Özhavzalı, "A qualitative study on changing consumer behaviors after the earthquake (clothing shopping)," *Dynamics in Social Sciences and Humanities*, vol. 5, no. 1, pp. 17–23,

[10] Ö. Ağrali, H. Sökün, and E. Karaarslan, "Twitter data analysis: Izmir earthquake case," *Journal of Emerging Computer Technologies*, vol. 2, no. 2, pp. 36–41, 2022.

[11] M. M. Hossain, M. S. Amin, F. Khairunnasa, and S. T. H. Rizvi, "Tweet trends and emotional reactions: A comprehensive sentiment analysis of twitter responses to the earthquake in turkey and syria," 2023.

[12] Z. Kastrati, A. S. Imran, S. M. Daudpota, M. A. Memon, and M. Kastrati, "Soaring energy prices: Understanding public engagement on twitter using sentiment analysis and topic modeling with transformers," *IEEE Access*, vol. 11, pp. 26 541–26 553, 2023.

[13] A. F. Adoma, N.-M. Henry, and W. Chen, "Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition," pp. 117–121, 2020.

[14] S. Chakraborty, M. B. U. Talukdar, M. Y. M. Adib, S. Mitra, and M. G. R. Alam, "Lstm-ann based price hike sentiment analysis from bangla social media comments," in *2022 25th International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2022, pp. 733–738.

[15] H. Saputra, R. Rahmaddeni, and F. Fazri, "Comparison of machine learning algorithms in analyzing public opinion sentiments against fuel price increases," *CESS (Journal of Computer Engineering, System and Science)*, vol. 8, p. 138, Jan. 2023. DOI: 10.24114/cess.v8i1.41911.

[16] A. Ceron and F. Negri, "Public policy and social media: How sentiment analysis can support policy-makers across the policy cycle," *Rivista Italiana di Politiche Pubbliche*, vol. 10, no. 3, pp. 309–338, 2015.

[17] F. Shamrat, S. Chakraborty, M. Imran, *et al.*, "Sentiment analysis on twitter tweets about covid-19 vaccines using nlp and supervised knn classification algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 1, pp. 463–470, 2021.

[18] S. S. Ismail, R. F. Mansour, R. M. Abd El-Aziz, and A. I. Taloba, "Efficient e-mail spam detection strategy using genetic decision tree processing with nlp features," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 7 710 005, 2022.

[19] S. Ghosal and A. Jain, "Depression and suicide risk detection on social media using fasttext embedding and xgboost classifier," *Procedia Computer Science*, vol. 218, pp. 1631–1639, 2023.

[20] A. Kumar, J. M. Chatterjee, V. G. Díaz, *et al.*, "A novel hybrid approach of svm combined with nlp and probabilistic neural network for email phishing," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 1, p. 486, 2020.

[21] J. Antony Vijay, H. Anwar Basha, and J. Arun Nehru, "A dynamic approach for detecting the fake news using random forest classifier and nlp," in *Computational Methods and Data Engineering: Proceedings of ICMDE 2020, Volume 2*, Springer, 2020, pp. 331–341.

[22] P. Lavanya and E. Sasikala, "Deep learning techniques on text classification using natural language processing (nlp) in social healthcare network: A comprehensive survey," in *2021 3rd international conference on signal processing and communication (ICPSC)*, IEEE, 2021, pp. 603–609.

[23] T. Papadakis, I. T. Christou, C. Ipektsidis, J. Soldatos, and A. Amicone, "Explainable and transparent artificial intelligence for public policymaking," *Data & Policy*, vol. 6, e10, 2024.