

Research Article


Large Language Models vs. Human Interpretation: Which is More Accurate in Text Classification?

Ahmet Hamdi Ozkurt, Emrah Aydemir, Yasin Sonmez


Abstract— Ekşi Sözlük is a widely used social network where numerous unusual events are discussed. In this context, it serves as a real-time news source for emergency response teams and digital news platforms. In this study, a dataset was compiled from comments shared on the Ekşi Sözlük platform regarding the Kahramanmaraş earthquake on February 6, 2023. These comments were classified into four categories: Source-Based Information, Emotional Reaction, Social Inference, and Personal Experience using the Gemma2 9B (9-billion-parameter) model, developed by Google with advanced natural language processing capabilities. A dataset of 500 comments in Excel format was analyzed, comparing the model outputs with human evaluations to assess classification accuracy. For this purpose, four evaluation columns were created for each comment based on category classification. The consistency between model-assigned categories and manually determined categories was examined using these columns. In cases where inconsistencies were detected, the model-generated explanations were subjected to qualitative evaluation. Model outputs that provide satisfactory explanations are considered acceptable, the manually classified category was assigned as the final evaluation. This process systematically resolved inconsistencies between model and human assessments, ensuring the final and validated category assignments for each comment. The highest accuracy values were observed for Social Inference (0.99), Source-Based Information (0.98), Personal Experience (0.88), and Emotional Reaction (0.83), respectively. In conclusion, this study presents a methodology for improving model performance through human supervision, contributing to the development of strategies for disaster management and crisis communication.

Index Terms— Natural Language Processing, Text Classification, Ekşi Sözlük, Gemma2 9B


Ahmet Hamdi Özkurt, is with Department of Management Information System University of Sakarya, Sakarya, Türkiye.(e-mail: hamdi.ozkurt@ogr.sakarya.edu.tr).

 <https://orcid.org/0009-0008-3220-4143>

Emrah Aydemir, is with Department of Management Information System University of Sakarya, Sakarya, Türkiye.(e-mail: emrahaydemir@sakarya.edu.tr).

 <https://orcid.org/0000-0002-8380-7891>

Yasin Sönmez, is with Department of Computer Technologies University, Batman, Türkiye (e-mail: yasinsonmez@batman.edu.tr).

 <https://orcid.org/0000-0001-9303-1735>

Manuscript received Mar. 05, 2025; accepted Apr. 14, 2025.

DOI: [10.17694/bajece.1652268](https://doi.org/10.17694/bajece.1652268)

I. INTRODUCTION

NATURAL DISASTERS are critical events that profoundly impact social order, where simultaneous, rapid, and accurate information is of vital importance. In addition to traditional news sources, social media platforms have emerged as significant information channels during disasters. Users share their emotions and thoughts regarding such events while also contributing to the flow of disaster-related information through social media [12]. This phenomenon holds substantial value in understanding public reactions to disasters and informing disaster management strategies. One of the frequently used platforms for sharing extraordinary events is Ekşi Sözlük, where users disseminate news, photographs, and observations related to disasters and emergencies. In this context, emergency response teams utilize data streams generated on social media platforms such as Ekşi Sözlük to identify affected regions and assess environmental impacts [15].

Social media data possess significant potential for enhancing crisis management during natural disasters and supporting post-disaster recovery efforts [17]. However, to effectively interpret, analyze, and utilize this data, machine learning techniques must be employed [7].

In this context, the primary aim of the study is to demonstrate how social media data can be classified using large language models during natural disasters. The research problem focuses on examining the potential contributions of rapidly and meaningfully distinguishing social media content in times of crisis to effective disaster management. Accordingly, comments related to the February 6, 2023 Kahramanmaraş Earthquake, collected from Ekşi Sözlük, were classified into four categories using the Gemma2 model.

The motivation of the study stems from the lack of models specifically designed to classify Turkish social media data in the context of disasters. As a contribution to the literature, this study offers an example of applying a large language model to Turkish-language data and proposes a method for the automatic and rapid classification of disaster-related content.

Finally, the classification performance of the model was evaluated using various metrics, and the outputs were compared with human annotations to analyze potential inconsistencies. In

doing so, the study demonstrates the potential of large language models in analyzing social media data during disaster events.

II. LITERATURE REVIEW

Natural disasters are crises in which the need for rapid, accurate, and reliable information reaches its peak. Among these, earthquakes are one of the most frequently encountered and devastating natural disasters on a global scale. Due to its geological location, Turkey is situated in an earthquake-prone region, posing a constant threat to the country. The February 6, 2023 Kahramanmaraş earthquake tragically reaffirmed this reality. The earthquakes, with magnitudes of 7.6 and 7.7, struck the Pazarcık and Elbistan districts of Kahramanmaraş, resulting in the loss of thousands of lives and leaving many individuals in urgent need of assistance [1].

During extraordinary events such as earthquakes, social media serves as an effective channel for individuals to share their experiences, call for help, and exchange information. Platforms like Twitter and Ekşi Sözlük are widely utilized to analyze public reactions to disasters and assess crisis communication strategies [14]. Research conducted on these platforms provides critical insights for developing crisis management strategies, enhancing societal resilience, and accelerating post-disaster recovery processes.

In recent years, Natural Language Processing (NLP) techniques have been increasingly used to analyze and interpret social media data during disasters [4]. One of the key NLP methods, sentiment analysis, helps determine the public mood by classifying social media posts as positive or negative. Machine learning models are commonly employed to extract meaningful patterns from large datasets and address complex problems. In particular, recent advancements in NLP models have significantly improved the ability to analyze and classify public responses during disaster events [6] [7]. Models like Gemma2 are used for the classification and analysis of social media data, with their performance evaluated using metrics such as accuracy, precision, recall, and F1-score. These metrics provide essential indicators for assessing the classification success, reliability, and overall effectiveness of the model.

Studies in the literature suggest that machine learning and deep learning techniques demonstrate high performance in the classification and analysis of social media data. For instance, Vaswani et al. [16] compared traditional machine learning approaches with Transformer-based deep learning models in text classification tasks and found that Transformer-based models achieved significantly higher accuracy. Similarly, Vieweg et al. (2010) conducted a systematic review of studies examining the role of social media in crises following natural disasters. These studies integrated both quantitative and qualitative research methods to provide a comprehensive analysis. Additionally, Lindsay [9] highlighted the functional efficiency of social media tools during disaster and crisis events, emphasizing their role in accelerating the dissemination of information.

Recent studies highlight the effectiveness of large language models (LLMs) in social media analysis during crises. For

example, Pereira et al. [13] showed that LLMs improved the comprehensiveness of summaries in crisis communication and emergency response compared to traditional methods.

Similarly, Yang et al. [18] noted that LLMs provided accurate predictions and interpretable explanations in analyzing mental health data from social media. Lastly, McDaniel et al. [10] demonstrated that integrating additional event information improved success rates in classifying social media posts in the humanitarian aid context using a zero-shot classification approach.

In conclusion, existing studies support the effective use of NLP and machine learning techniques in the analysis of post-disaster social media data. Text classification methods form the foundation of social media data analysis, while techniques such as sentiment analysis have the potential to optimize response processes by ensuring rapid and accurate access to critical information in crisis situations. Therefore, further research in this field could provide significant contributions to crisis management and humanitarian aid efforts.

III. METHODOLOGY

A. DATA COLLECTION

Within the scope of this study, 7,250 comments related to the February 6, 2023, Kahramanmaraş earthquake were collected from the Ekşi Sözlük platform. These comments were obtained using the Selenium library in Python, ensuring that both timestamp and comment information were preserved. The extracted data were initially stored in JSON format and subsequently converted to CSV format for further processing.

B. DATA PREPROCESSING

Data preprocessing is the phase in which raw data is transformed into a structured and interpretable format. In other words, it involves converting initial raw data into final processed data that serves as the foundation for subsequent analyses [2].

In this study, the comments extracted in JSON format were cleaned by removing punctuation marks. Comments lacking semantic coherence or consisting solely of visual content were eliminated from the dataset. Additionally, numerical expressions within the comments were removed. However, double quotation marks were retained during this preprocessing phase.

C. CLASSIFICATION

After completing the data preprocessing stage, 6,895 comments remained from the original dataset of 7,250 and were subsequently classified into four categories (see Table 1).

In the classification process, Gemma2, a high-performance, fast, and efficient language model developed by Google, was utilized. This model is commonly employed in Natural Language Processing (NLP) tasks such as text generation and classification. Additionally, Gemma2 has 2B (2 billion parameters) and 27B (27 billion parameters) variants. The primary distinction among these models lies in their parameter count, which directly influences the model's information

processing capacity. As the number of parameters increases, the model's ability to process and learn from data improves proportionally.

Table I
Definition and Examples of Comment Categories Used for Classification

CATEGORY	DEFINITION	EXAMPLE
Source-Based Information	Statements that are supported by verifiable data, academic sources, or credible authorities, aiming to inform or explain objectively.	According to a 2023 report by the World Health Organization, air pollution causes approximately 7 million premature deaths annually.
Emotional Response	Expressions that reflect the speaker's emotions, such as anger, fear, happiness, or sadness, often aiming to convey a personal or subjective response.	It breaks my heart to see how little is being done to protect our environment.
Social Inference	Comments that draw broader conclusions about society, culture, or collective behavior based on individual observations or specific events.	The increasing use of social media has fundamentally changed how people form and maintain relationships in modern society.
Personal Experience	Narratives or statements based on the individual's own life, experiences, or observations, usually shared in a subjective manner.	Last year, I volunteered at a refugee camp, and it completely changed my perspective on global crises.

Table II
Created Sample Excel

DATE	COMMENT	Source- Based Information	Emotional Reaction	Social Inference	Personal Experience
06.02.2023	Oh my God, Gaziantep is shaking very badly again, we are living in hell here, it is shaking like a cradle, please make it stop.	0	1	0	1
07.02.2023	The environment minister said on live broadcast that we did not leave any of our citizens hungry and exposed, right, you did not leave them hungry and exposed, you left them under the rubble.	0	1	1	0
07.02.2023	Malatya is in a very difficult situation, the food and shelter shortage of earthquake survivors is growing and no one is helping, please help please	0	1	0	1

IV. SAMPLING AND MANUAL PROCESSING

From the 6,895 classified comments, a balanced subset of 500 unique comments was randomly selected, consisting of 250 instances labeled as '0' and 250 instances labeled as '1' for each category. This random selection process was carried out to ensure data balance and was implemented using the following Python libraries and functions:

- Pandas: Utilized for data analysis and processing.
- NumPy: Used for numerical computations.
- Path: Employed for creating and managing file paths.

The Gemma2 model offers several advantages when compared to other existing large language models (LLMs). While higher-parameter models such as GPT-3.5 (175B) and GPT-4 (1.7T) demonstrate more advanced comprehension and text generation capabilities, their computational demands significantly limit their usability, particularly on local hardware. Although open-source alternatives like LLaMA 2 (7B–70B) offer greater modifiability, they fall short of achieving the same efficiency-performance balance that Gemma2 provides for specific tasks.

The 9B parameter model was selected for this study due to its balance between efficiency and computational resource consumption (e.g., processor, RAM), offering optimal performance for classification tasks. Additionally, the open-access availability of the Gemma2 model and its comparatively stronger support for the Turkish language have been key factors in its selection over alternative models [11].

During the classification process, each comment was assigned '1' if it belonged to a specific category and '0' otherwise. In the initial phase, the model was tasked solely with determining the appropriate category for each comment. Upon completion of the classification process, the annotated data was exported from a CSV file to an Excel format for further analysis.

- Linprog and Pulp: These functions were used for solving linear programming problems.

Subsequently, the randomly selected comments were also manually classified following the same methodology. Comments deemed meaningless, consisting only of visual content, or containing fewer than five words were excluded from the dataset. After this filtering process, the remaining 500 comments were reintroduced to the model for further evaluation.

The selection of only 500 comments from the dataset of 7,250 was driven by the dual aim of developing a balanced

classification model and enhancing the efficiency of the manual labeling process. Ensuring an equal number of examples from each category (250 labeled as '0' and 250 as '1') was essential for enabling the model to perform consistently across both

classes and for ensuring more reliable human intervention during the evaluation phase.

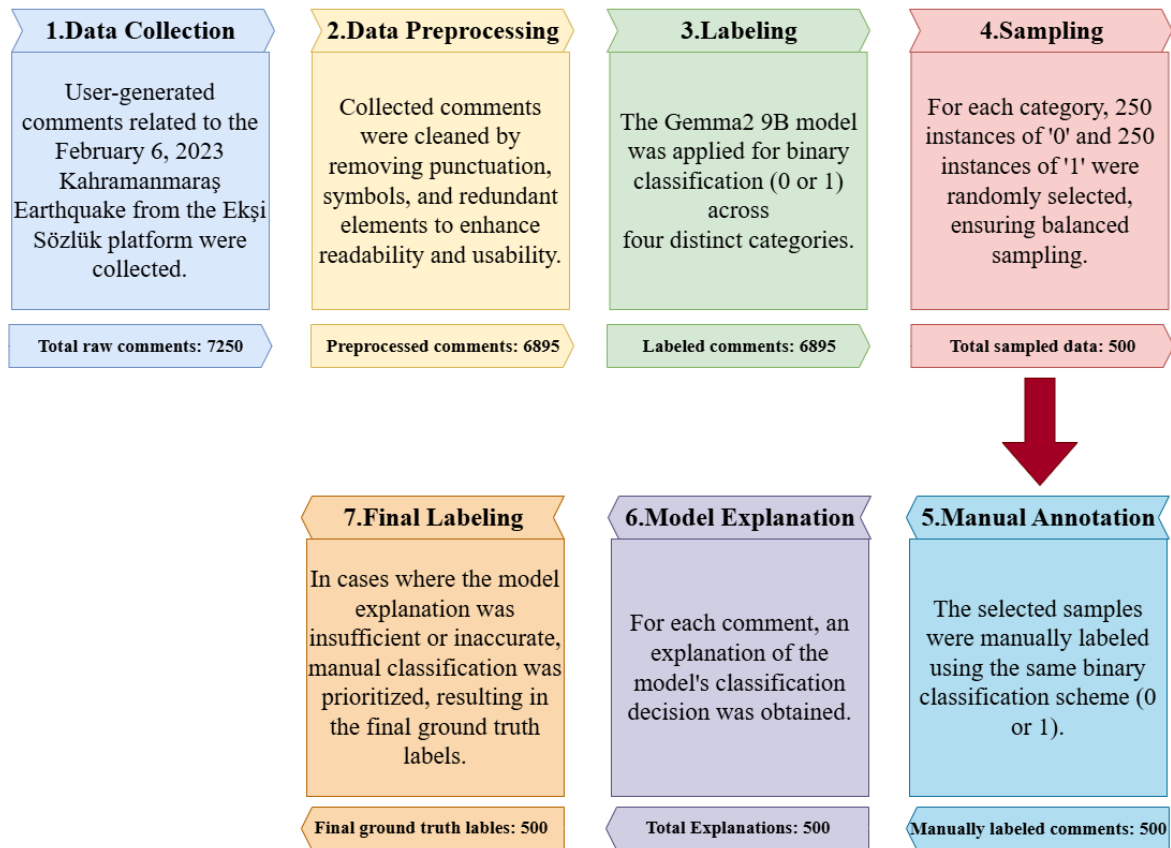


Fig. 1. Flow Diagram of the Study

V. MODEL EXPLANATION

During the classification process, discrepancies and inconsistencies were identified between the values assigned by the model and those assigned manually. To mitigate this issue,

an alternative approach was introduced to enhance the model's attentiveness and response generation. Specifically, modifications were made to the prompt content to ensure that the model provided more deliberate and well-considered classifications.

Table III
Category Based Descriptions of the Model

Explanation Based on Information Source	Explanation Based on Emotional Response	Explanation Based on Social Inference	Explanation Based on Personal Experience
NO because the comment does not refer to a specific event or source.	YES because the comment reflects feelings of sadness and concern, with phrases such as "they were in great shock" and "I hope our loss is not too great".	YES because the comment contains a call for social support and solidarity for the earthquake victims by saying "get well soon to the people in that region".	YES because the author's statement in the comment that "I talked to my loved ones in Antep and Nizip" reflects his personal experience.

VI. DETERMINATION OF FINAL GROUND TRUTH VALUES

In this study, four category-specific evaluation columns were created for the 500 comments stored in Excel format. These evaluation columns were utilized to assess the accuracy of the model outputs. The following steps were followed in the evaluation process:

- **Category-Based Consistency Check:** If there was

consistency between the values assigned by the model and those assigned manually, the respective evaluation column was left empty.

- **Category-Based Accuracy Assessment of Model Output:** In cases where discrepancies were identified between the model output and manual classification, the model's explanations were reviewed. If the explanation was deemed satisfactory, the model output was accepted in the evaluation column.

- **Category-Based Determination of Final Ground Truth Values:** If the model's explanations were found to be insufficient or unconvincing, the manually assigned classification was used as the final label in the evaluation column.

As a result of these procedures, discrepancies between the model outputs and manual classifications were resolved through the evaluation columns, and the final ground truth values were established.

VII. FINDINGS

To evaluate the predictive performance of the model across four distinct categories (*Source-Based Information*, *Emotional Response*, *Social Inference*, and *Personal Experience*), comparisons between the model-generated outputs and manually assigned labels were analyzed. These comparisons illustrate the statistical distribution of the model's predictions for each category and the accuracy rates of these predictions. Below, two tables containing these comparisons are presented,

along with detailed explanations regarding their implications.

Table IV
Comparison of Model Predictions and Manual Labels

Category	Model Prediction (0)	Model Prediction (1)	Manual Label (0)	Manual Label (1)
Source-Based Information	250	250	257	243
Emotional Reaction	250	250	303	197
Social Inference	250	250	258	242
Personal Experience	250	250	309	191

Table 4 shows the statistical distribution of the predictions made by the model for the four categories (coded as 0 and 1), along with the corresponding manual labels (coded as 0 and 1).

Table V
Evaluation Results of Model Estimates

Category	Model Acceptance 0	Model Acceptance 1	Model Rejection 0	Model Rejection 1	Total Prediction	Correct Prediction	Percentage
Source-Based Information	250	240	0	10	500	490	%98
Emotional Response	235	180	15	70	500	415	%83
Social Inference	248	246	2	4	500	494	%99
Personal Experience	247	194	3	56	500	441	%88

Table 5 presents the number of predictions labeled as 0 and 1 for both accepted and rejected instances across each category, along with the total and correctly classified predictions. For instance, in the "Social Inference" category, the model correctly predicted 248 instances as 0 and 246 as 1, with only 6 misclassifications. These results provide a basis for evaluating the overall classification performance of the model and understanding the distribution of success across categories. Moreover, they serve as input for the calculation of classification metrics such as accuracy, precision, recall, and F1-score, which are discussed in detail in the following section.

A. 1. CONFUSION MATRIX ANALYSIS AND CLASSIFICATION METRICS

The classification performance of the model is quantitatively assessed using the confusion matrix. In the confusion matrix, columns represent the final true values, while rows represent the predicted values. In this context, the evaluation columns indicate the actual values in the classification, and the model outputs represent the predicted values. The following Python libraries were used to construct the confusion matrix:

- Pandas: Used for data processing and reading, particularly for reading columns of data in Excel format.

- NumPy: Used for numerical computations, including percentage calculations in the confusion matrix.
- Scikit-learn: Used for generating the confusion matrix and calculating classification metrics (Accuracy, Precision, Recall, F1-Score).
- Seaborn: Used for the visualization of the confusion matrix.
- Matplotlib: Used for the customization of the confusion matrix visualization.

The classification metrics derived from the confusion matrix are utilized to measure the classification process's performance in greater detail. These metrics include:

- Accuracy
- Precision
- Recall
- F1-Score

Table VI
Confusion Matrix Structure

	REAL POSITIVE (1)	REAL NEGATIVE (0)
PREDICTION POSITIVE (1)	TP	FP
PREDICTION NEGATIVE (0)	FN	TN

Table VII
Classification Metrics and Explanations

Classification metrics	Computation formulas	Definition
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$	It is the overall accuracy rate of the model's classification process.
Precision	$\frac{TP}{TP + FP}$	It represents the proportion of classes predicted as positive by the model that are actually positive.
Recall	$\frac{TP}{TP + FN}$	It measures how accurately the model predicts within the true positives.
F-1 Score	$\frac{2 * Precision * Recall}{Precision + Recall}$	The harmonic mean of Precision and Recall is used for their combined evaluation.

To evaluate the overall performance of the model, it is necessary to examine the category-specific confusion matrices and classification metrics as summarized in Table 7.

- True Positive (TP): The number of positive instances correctly classified by the model.
- True Negative (TN): The number of negative instances correctly classified by the model.
- False Positive (FP): The number of negative instances incorrectly classified as positive by the model.
- False Negative (FN): The number of positive instances incorrectly classified as negative by the model [3].

The classification metrics used for performance evaluation, along with their formulas and definitions, are provided below.

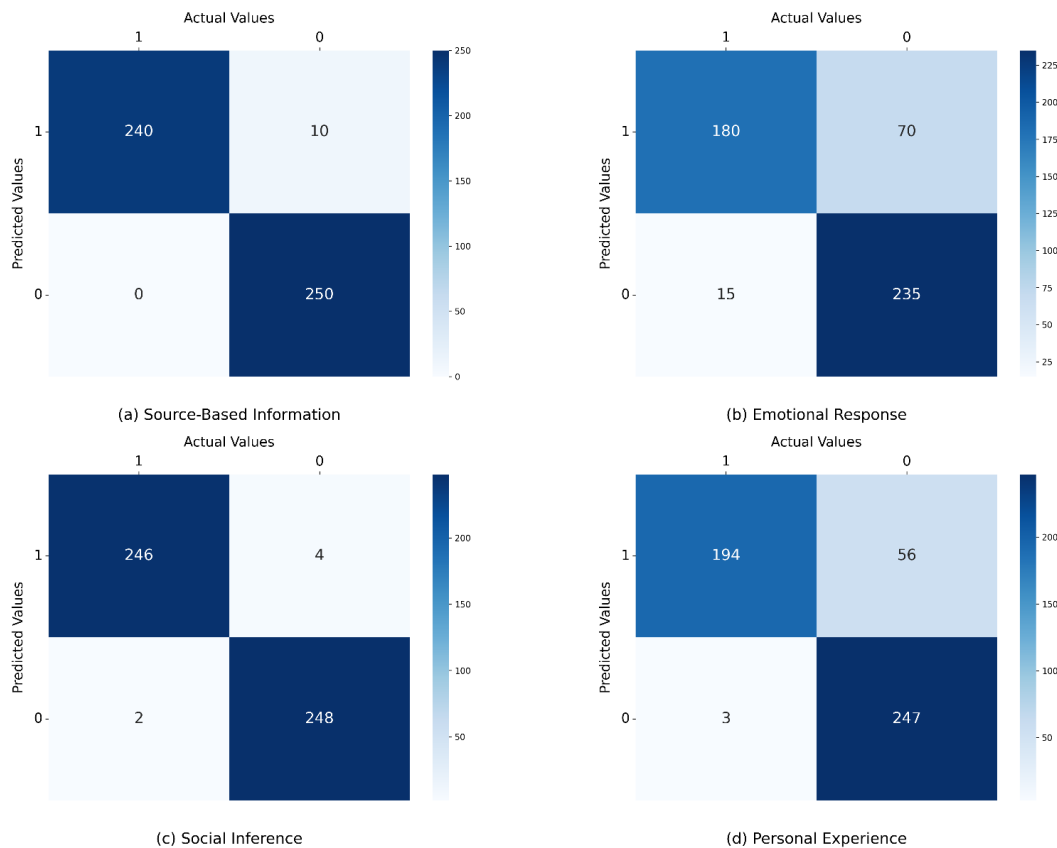


Fig. 2. Confusion Matrices for Categories

B. CATEGORY-BASED CONFUSION MATRIX

1) Source Information

Upon examining the confusion matrix for the Source Information category presented in Figure 2(a), it has been

observed that the model made 250 true negative (TN) and 240 true positive (TP) predictions. Additionally, it was found that there were 0 false positive (FP) and 10 false negative (FN) classifications. This indicates that the model has not tended to

classify non-source information as source information (FP=0), and has correctly identified the vast majority of comments containing source information (low FN=10).

2) Emotional Response

Upon evaluating the confusion matrix for the Emotional Response category presented in Figure 2(b), it has been observed that the model made 235 true negative (TN) and 180 true positive (TP) predictions. Additionally, the model produced 70 false positive (FP) and 15 false negative (FN) predictions. This finding suggests that the model incorrectly classified some non-emotional response comments as emotional responses (high FP=70) and was able to correctly identify a portion of the comments that contained emotional responses.

3) Social Inference

Upon examining the confusion matrix for the Social Inference category presented in Figure 2(c), it has been observed that the model made 248 true negative (TN) and 246 true positive (TP) predictions. Moreover, there were 4 false positive (FP) and 2 false negative (FN) classifications. In this context, the model's classification performance in this category is quite successful, as indicated by the low false positive rate (FP=4) and the high true positive rate (TP=246).

4) Personal Experience

Upon reviewing the confusion matrix for the Personal Experience category presented in Figure 2(d), it has been observed that the model made 247 true negative (TN) and 194 true positive (TP) predictions. Additionally, there were 56 false positive (FP) and 3 false negative (FN) classifications. This indicates that the model incorrectly classified some non-personal experience comments as personal experiences (FP=56), but successfully identified the majority of comments containing personal experience (low FN=3).

Table VIII
Performance of Category-Based Classification Metrics

Categories	Accuracy	Precision	Recall	F-1 Score
Source-Based Information	0.98	0.96	1.00	0.97
Emotional Response	0.83	0.72	0.92	0.80
Social Inference	0.99	0.98	0.99	0.98
Personal Experience	0.88	0.77	0.98	0.86

C. CATEGORY-BASED CLASSIFICATION METRICS

1) Source Information

Upon examining Table 8, it can be observed that the classification metrics for the Source Information category demonstrate generally successful performance (Accuracy: 0.98, Recall: 1.00, Precision: 0.96, F1-Score: 0.97). The high recall indicates that the model is able to successfully identify comments containing source information, while the high

precision indicates that the model can also largely classify comments that do not contain source information correctly.

2) Emotional Response

When reviewing the classification metrics for the Emotional Response category in Table 8, it is observed that the recall value is high (0.92), the accuracy value is 0.83, the precision value is 0.72, and the F1-Score is 0.80. These values suggest that the model is successful in detecting comments containing emotional responses (high recall), but also incorrectly classifies some comments that do not contain emotional responses as emotional (lower precision).

3) Social Inference

Upon evaluating the classification metrics for the Social Inference category presented in Table 8, it is evident that the values for accuracy, recall, precision, and F1-Score are very high (Accuracy: 0.99, Recall: 0.99, Precision: 0.98, F1-Score: 0.98). These values indicate that the model has very low false positive (FP) and false negative (FN) rates.

4) Personal Experience

According to the data presented in Table 8, it is observed that the model demonstrates high performance in the Personal Experience category (Accuracy: 0.88, Precision: 0.77, Recall: 0.98, F1-Score: 0.86). The high recall indicates that the model correctly identifies the majority of comments containing personal experience, while the lower precision value suggests that the model also incorrectly classifies some comments that do not contain personal experience as personal experience.

VIII. DISCUSSION AND CONCLUSION

This study aimed to automatically classify the comments on Ekşi Sözlük related to the 6th February 2023 Kahramanmaraş earthquake and categorize them into four distinct categories (Source Information, Emotional Response, Social Inference, Personal Experience) using the Gemma2 model. This work highlights the importance of analyzing social media data in crisis situations to obtain quick and accurate information. The results obtained indicate that the classification performance of the model varies significantly across categories.

The analyses show that the model demonstrated its highest classification performance in the Social Inference category, while its lowest performance was observed in the Emotional Response category. In the Source Information and Personal Experience categories, the model demonstrated satisfactory classification performance with accuracy rates of 98% and 88%, respectively. These findings suggest that the model's classification performance varies by category, and the confidence in the results may fluctuate based on the category. While the accuracy rate may be very high in one category, the opposite can be observed in another category. In this context, it is suggested that instead of fully relying on machine learning models, category-specific approaches may be necessary for different classification categories. The higher classification performance in the Social Inference category may be due to the language being more objective and tending to express social inferences directly. Conversely, in the Emotional Response category, the intensity of ambiguous, complex expressions

might make it difficult for the model to comprehend.

Table 8 clearly demonstrates this situation in the classification metrics. The model exhibited the highest classification performance in the Social Inference category (Accuracy: 99%), indicating that it can accurately classify texts containing social inferences, evaluations, or general comments. On the other hand, the performance in the Emotional Response category is the lowest (Accuracy: 83%). In the Source Information (Accuracy: 98%) and Personal Experience (Accuracy: 88%) categories, the model performed quite well. Another reason for the low classification performance in the Emotional Response category may be the differences and subjectivity of emotional expressions in the dataset. The model may struggle with categorizing different emotional tones.

In conclusion, this study revealed that the classification performance of the Gemma2 model in categorizing the comments on Ekşi Sözlük regarding the 6th February Kahramanmaraş Earthquake varies based on the category. The model exhibited low performance in the Emotional Response category, while achieving high success in the Social Inference category.

These findings highlight important points to consider when developing natural language processing-based sentiment analysis and automatic text classification models. To improve the model's classification performance in the Emotional Response category, various sentiment analysis techniques, such as sentiment lexicons, can be utilized, or a specialized model for this category could be trained. For example, pre-trained sentiment analysis models can be used to help the model understand complex emotional expressions. Additionally, adding data from broader and diverse sources (social media, news websites, etc.) could further enhance the model's classification performance.

In particular, for categories with lower performance, fine-tuning techniques can be applied to improve the model's performance. Fine-tuning the model on category-specific datasets may enhance classification performance for those particular categories. Additionally, approaches such as ensemble learning can leverage the strengths of different models. For instance, combining the predictions of a sentiment analysis-focused model with the Gemma2 model for the Emotional Response category could improve classification performance. By developing an optimized ensemble model for each category in this manner, the overall system performance can be enhanced.

Some limitations of the study should be considered. Firstly, the data being collected from only a single social media platform (Ekşi Sözlük) restricts the generalizability of the findings. Additionally, the relatively small size of the dataset may have limited the model's classification capacity and could have affected the reliability of the results. Furthermore, the potential presence of biases within the dataset should not be overlooked. In particular, the demographic profile of Ekşi Sözlük users may prevent the analyzed content from fully representing the broader population. For these reasons, future studies should consider using larger and more diverse datasets

from various social media platforms to enhance the model's performance and strengthen the generalizability of the findings. These suggestions will contribute to the development of strategies for disaster management and crisis communication.

REFERENCES

- [1] AFAD, "06 Şubat 2023 Pazarcık-Elbistan Kahramanmaraş (Mw 7.7; Mw 7.6) depremleri raporu," Deprem ve Risk Azaltma Genel Müdürlüğü, 2023.
- [2] Y. Argüden and B. Erşahin, Veri madenciliği: Veriden bilgiye, masraftan değere, ARGE Danışmanlık Yayınları, 2008.
- [3] Z. Bakan and F. Kanbay, "Makine öğrenmesi yöntemleri ile eğitim başarısına etki eden faktörlerin modellenmesi," İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, vol. 23, no. 45, pp. 27–41, 2024. [Online]. Available: <https://doi.org/10.55071/ticaretfd.1442084>
- [4] G. Burel and H. Alani, "Crisis event extraction service (CREES)—Automatic detection and classification of crisis-related content on social media," in Proc. 15th Int. Conf. Inf. Syst. Crisis Response and Manage., 2018.
- [5] C. Coşkun and A. Baykal, "Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması," in Akademik Bilişim Konferansı (AB'11) Bildirileri, 2011, pp. 51–58.
- [6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2019.
- [7] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "AIDR: Artificial intelligence for disaster response," in Proc. 22nd Int. Conf. World Wide Web, 2015, pp. 159–162.
- [8] O. H. Kwon et al., "Sentiment analysis of the United States public support of nuclear power on social media using large language models," Renewable and Sustainable Energy Reviews, vol. 200, 114570, 2024. [Online]. Available: <https://doi.org/10.1016/j.rser.2024.114570>
- [9] B. R. Lindsay, "Social media and disasters: Recent United States experiences," J. Contingencies Crisis Manage., vol. 19, no. 1, pp. 1–7, 2011. [Online]. Available: <https://doi.org/10.1111/j.1468-5973.2011.00639.x>
- [10] E. L. McDaniel, S. Scheele, and J. Liu, "Zero-shot classification of crisis tweets using instruction-finetuned large language models," in 2024 IEEE Int. Humanitarian Technol. Conf. (IHTC), Nov. 2024, pp. 1–7.
- [11] M. Özkan and G. Kar, "Türkçe dilinde yazılan bilimsel metinlerin derin öğrenme tekniği uygulanarak çoklu sınıflandırılması," Mühendislik Bilimleri ve Tasarım Dergisi, vol. 10, no. 2, pp. 504–519, 2022. [Online]. Available: <https://doi.org/10.21923/jesd.973181>
- [12] L. Palen and S. B. Liu, "Citizen communications in crisis: Anticipating a future of ICT-supported public participation," in Proc. SIGCHI Conf. Human Factors Comput. Syst., 2007, pp. 727–736.
- [13] J. Pereira, R. Lotufo, and R. Nogueira, "Large language models in summarizing social media for emergency management," arXiv preprint arXiv:2401.03158, 2024.
- [14] C. Reuter and M. A. Kaufhold, "Fifteen years of social media in emergencies: A retrospective review and future directions for crisis informatics," J. Contingencies Crisis Manage., vol. 26, no. 1, pp. 41–57, 2018. [Online]. Available: <https://doi.org/10.1111/1468-5973.12196>
- [15] O. Sevil and N. Kemaloglu, "Olağandışı olaylar hakkındaki tweet'lerin gerçek ve gerçek dışı olarak Google BERT modeli ile sınıflandırılması," Veri Bilimi, vol. 4, no. 1, pp. 31–37, 2021.
- [16] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst., vol. 30, 2017.
- [17] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: What twitter may contribute to situational awareness," in Proc. SIGCHI Conf. Human Factors Comput. Syst., 2010, pp. 1079–1088.
- [18] K. Yang et al., "MentalLaMA: Interpretable mental health analysis on social media with large language models," in Proc. ACM Web Conf. 2024, May 2024, pp. 4489–4500.

BIOGRAPHIES



Ahmet Hamdi Özkurt is going on his Bachelor's degree in Management Information System in Sakarya University. He is a third year student. He continues to develop himself in the field of artificial intelligence, large language models, database and mobile programming.



Emrah Aydemir was received the M.S. degrees in computer teaching from the University of Elazig Firat, in 2012 and the Ph.D. degree in informatics from Istanbul University, Turkey, TR, in 2017. From 2012 to 2015, he was an Expert with the Istanbul Commerce University. Since 2021, he has been an Associate Professor with the Management Information System, Sakarya University. He is the author of three books, more than 60 articles, and more than 40 conference presentation. His research interests include artificial intelligence, microcontroller, database and software



Yasin Sönmez was born in Diyarbakır, Turkey in 1986. He received the B.S. degree from the Firat University, Technical Education Faculty, Department of Electronics and Computer Education in 2010, M.S. degree in computer science from the Firat University in 2012 and Ph.D. degree department of software engineering at Firat University in 2018. His research interests include, artificial intelligence, and information security.