

# Relief Call Classification and NER Dataset on Twitter Data

1<sup>st</sup> Beyza Akdeniz

Computer Science Department. TOBB ETU

Ankara, Türkiye

beyzaakdeniz@etu.edu.tr

**Abstract**—This project aims to prepare labeled dataset for identify tweets containing calls for help shared during the Kahramanmaraş earthquake and then perform classification with SVM and two different BERTurk model. In addition prepare labeled Named Entity Recognition (NER) dataset to extract names and location information. Information shared on social media platforms during earthquakes and other disasters can be a crucial resource for disaster management and emergency response. Messages containing help requests can assist emergency response teams in addressing the needs and demands of people in the affected areas.

**Index Terms**—Tweets, NLP, Classification, NER, Earthquake, Humanitarian Aid, Event Detection

## I. INTRODUCTION

This project leverages Twitter data to enhance disaster response during the Kahramanmaraş earthquake by identifying and analyzing help calls. We aim to prepare two handy datasets that ready for used in advanced Natural Language Processing (NLP) techniques, classification and Named Entity Recognition (NER), to extract crucial information from tweets, such as names and locations of people in need. This data can be visualized on a map to assist emergency responders in prioritizing areas requiring urgent attention.

Our project workflow includes data collection, preprocessing, tweet classification and entity tagging. The goal is to create dataset can be used in map and analyze the help calls to provide real-time insights that aid in efficient resource allocation and decision-making during disaster relief operations.

## II. RELATED WORK

The number of Twitter users in Turkey is significant, with millions actively engaging on the platform. Twitter’s vast user base makes it an excellent source for data collection, ensuring a large volume of data. One of Twitter’s key advantages is its rapid data distribution, which is much faster than traditional online media or news outlets. Online media requires journalists to investigate and verify events, which can delay the dissemination of information. In contrast, many Turks use Twitter to discuss personal or public incidents, making it a valuable platform for real-time information. Therefore, text mining techniques are crucial for extracting meaningful insights from Twitter data.

Named Entity Recognition (NER) is a crucial NLP task aimed at identifying and categorizing entities such as names,

locations, and organizations within text. While extensively studied in English, NER research in Turkish remains limited, primarily due to the language’s morphological complexity and agglutinative nature, which complicates automated text analysis and information extraction. Conventional methods for Turkish NER, such as Conditional Random Fields (CRF), have been widely used. These methods rely on supervised learning and sequence tagging techniques but often fall short in handling the morphological richness of Turkish, leading to lower extraction accuracy. Recent shifts towards deep learning have seen models like BiLSTM and CRF integrated to enhance NER performance. However, these still face challenges with the language’s complexity. The introduction of models such as Google BERT, which leverage bidirectional context, has shown promise in improving NER tasks across various languages, including Turkish. BERT’s capability to capture nuanced context and manage diverse linguistic structures is particularly advantageous for Turkish NER. [1] Insights from this study on leveraging BERT for Turkish NER will inform our use of advanced NLP models like BerTURK and FastText in extracting help requests from tweets. By understanding how BERT handles Turkish’s linguistic intricacies, we can better adapt these models for the specific requirements of our project, ultimately enhancing the precision and utility of extracted information for emergency response teams.

Following the earthquakes in Southern and Central Turkey and Northern and Western Syria on February 6, 2023, Twitter emerged as a pivotal platform for real-time reporting and public sentiment analysis. Mandal et al. (2023) utilized machine learning algorithms including Logistic Regression, Random Forest, Decision Tree, and XGBoost to classify tweets from the disaster-affected regions between February 6 to February 21, 2023. Their study aimed to categorize tweets into specific themes to aid in the rapid allocation of relief efforts and resources. Previous research has demonstrated the effectiveness of leveraging Twitter data for various applications such as early warning systems, sentiment analysis, and real-time event monitoring across diverse domains. [2]

An existing dataset, HumAID, contains manually annotated tweets collected during 19 natural disaster events including earthquakes. One of the class labels is "requests or urgent needs" that represents messages "of urgent needs or supplies such as food, water, clothing, money, medical supplies or

blood.” In this study, we will extend this definition and target the detection of social media messages of calling help, not only supplies but also requesting rescue operation.[3]

One notable study proposed a comprehensive model specifically for NER in Turkish texts from platforms like Twitter, Facebook, and the Donanimhaber Forum. This model leverages Conditional Random Fields (CRF) and Bidirectional Long Short Term Memory (BiLSTM) to handle the complexities of social media language. To enhance performance, the model integrates a variety of features including word embeddings, character representations, morphological features, domain information, pattern-matching and Part-of-Speech (POS) Tags. This study will also utilize these methods; however, the focus will be on tweets related to disasters. [4]

The paper introduces a novel dataset tailored for Named Entity Recognition (NER) in Turkish social media, specifically Twitter. This work addresses the scarcity of publicly available datasets for Turkish NER, particularly in informal contexts like social media. It emphasizes the importance of recognizing diverse entity types beyond traditional categories such as person, location, and organization, including time, money, product, and TV-shows. Methodologically, the paper explores the efficacy of transformer-based models like BertTurk for NER tasks on the newly introduced dataset. This approach underscores the application of state-of-the-art NLP techniques to enhance entity recognition accuracy in Turkish tweets, thus broadening the scope of NER applications beyond traditional formal texts. [5] This study will use these techniques for disaster response.

### III. METHODOLOGY

#### A. Data Collection

We used the Kaggle dataset[6] to collect tweets related to the Kahramanmaraş earthquake using hashtags such as ”earthquake” and ”deprem” during and after the event for the first 24 hours (i.e., from 04:00 AM on February 6th to 04:00 AM on February 7th, 2023). Approximately 13,000 tweets were sampled, ensuring a balanced representation of both help and non-help requests. Hashtags were employed to filter relevant tweets from the Kahramanmaraş region and surrounding areas. Source dataset use these hashtags for filtering: Turkey and earthquake. We add additional hashtags. They are ahbap and afad that are human aid organisations the most used in relief tweets.

#### B. Data Preprocessing

Social media platforms are known for their informal language, frequent misspellings, and unconventional writing styles, which pose significant challenges for Natural Language Processing (NLP) tasks such as Named Entity Recognition (NER). In this study, we applied a series of preprocessing steps to enhance the quality and interpretability of the Twitter data collected during the Kahramanmaraş earthquake. Initially, tweets were filtered to select only those in Turkish, ensuring relevance to the target region and language. Dates were standardized to a common format and tweets were

filtered within the specified timeframe (from 04:00 AM on February 6th to 04:00 AM on February 7th, 2023). Turkish text normalization was performed using the Zemberek-Python library, which included deasciification to convert Turkish-specific accented letters and remove repetitive characters (e.g., converting ”sukran” to ”şükran”). This step was crucial for handling noisy text with inconsistent character encoding. The preprocessing pipeline involved several cleaning steps to prepare the text for NER analysis: URL Removal: Any URLs present in the tweets were removed using regular expressions. Punctuation Marks: Non-word characters such as punctuation marks were stripped from the text to focus on meaningful content. Hashtags and Mentions: Tags and mentions (e.g., @username) were removed as they do not contribute to entity recognition. Whitespace: Additional whitespace was reduced and standardized to improve the uniformity of text representation. All tweets were converted to lowercase characters to ensure consistency in text processing and analysis. This step prevents the model from treating words with different cases as distinct entities. Stopwords, common words that do not contribute significant meaning to the text (e.g., ”and”, ”the”, ”is”), were removed to focus on informative content relevant to NER tasks. By implementing these preprocessing steps, we aimed to enhance the quality of the dataset and improve the performance of subsequent NLP tasks, particularly in accurately extracting names, locations, and other entities from social media messages related to disaster response.

This comprehensive preprocessing approach not only cleansed the data but also prepared it for advanced NLP techniques such as Named Entity Recognition (NER) and classification, crucial for deriving actionable insights during disaster response efforts.

#### C. Tweet Classification and

We took 2,000 of all of the tweets in the dataset and labeled them based on whether they contained a call for help or not. 553 of them contained relief calls. We used 1000 of them in the classification to train with a balanced dataset. A labeled dataset suitable for use in other projects was obtained.

Then we tokenized and classified tweets if they are emergency call or not with three different models. We trained SVM as a base case and also train two different BERTurk models. First one is Turkish uncased BERT model[7] and the second one is the pretrained model from A ”Twitter Corpus for Named Entity Recognition in Turkish”. We trained model with social media corpus because our dataset is also made by tweets. We fine-tuned the model and performed classification. Evaluation metrics are precision, recall, f1-score and accuracy.

A support vector machine (SVM) is a supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space.

SVMs were developed in the 1990s by Vladimir N. Vapnik and his colleagues, and they published this work in a paper titled ”Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing”<sup>1</sup> in 1995.

```

# Function to preprocess a single tweet
def preprocess_tweet(tweet):
    # Remove consecutive repetitive letters (more than two)
    tweet = re.sub(r'(\w)\1{2,}', r'\1\1', tweet)

    # Remove URLs
    tweet = re.sub(r'http\S+|www\S+|https\S+', '', tweet, flags=re.MULTILINE)

    # Remove punctuation marks
    tweet = re.sub(r'[^\w\s]', '', tweet)

    # Remove hashtags and mentions
    tweet = re.sub(r'[@#]\w+', '', tweet)

    # Remove additional white spaces
    tweet = re.sub(r'\s+', ' ', tweet)

    # Normalize sentence to avoid noisy text
    tweet = normalizer.normalize(tweet)

    # THIS PART NOT USED BECAUSE NORMALIZER ALREADY DOES THESE OPERATIONS

    # Convert to lowercase
    # tweet = tweet.lower()

    # Deascify
    # deascifier = Deascifier(tweet)
    # tweet = deascifier.convert_to_turkish()

    # Remove stop words
    # tokens = word_tokenize(tweet)
    # tweet = ' '.join(word for word in tokens if word not in turkish_stopwords)

    return tweet

```

Fig. 1. Preprocessing Steps

SVMs are commonly used within classification problems. They distinguish between two classes by finding the optimal hyperplane that maximizes the margin between the closest data points of opposite classes. The number of features in the input data determine if the hyperplane is a line in a 2-D space or a plane in a n-dimensional space. Since multiple hyperplanes can be found to differentiate classes, maximizing the margin between points enables the algorithm to find the best decision boundary between classes. This, in turn, enables it to generalize well to new data and make accurate classification predictions. The lines that are adjacent to the optimal hyperplane are known as support vectors as these vectors run through the data points that determine the maximal margin. [9]

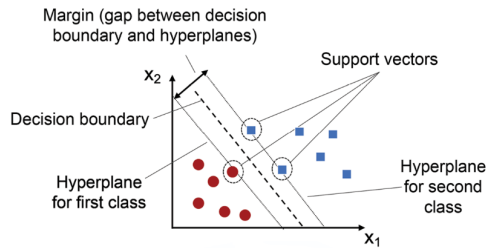


Fig. 2. SVM

BERT (Bidirectional Encoder Representations from Transformers) is a recent paper published by researchers at Google AI Language. It has caused a stir in the Machine Learning community by presenting state-of-the-art results in a wide

variety of NLP tasks, including Question Answering (SQuAD v1.1), Natural Language Inference (MNLI), and others.

BERT's key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling. This is in contrast to previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training. The paper's results show that a language model which is bidirectionally trained can have a deeper sense of language context and flow than single-direction language models. In the paper, the researchers detail a novel technique named Masked LM (MLM) which allows bidirectional training in models in which it was previously impossible. [10]

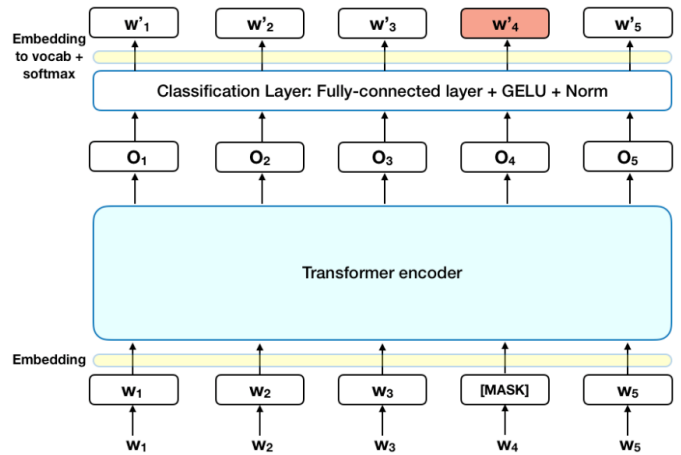


Fig. 3. BERT

```

# Define training arguments
training_args = TrainingArguments(
    output_dir="./results",
    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,
    num_train_epochs=1,
    logging_dir='./logs',
    logging_steps=10,
    evaluation_strategy="epoch",
    save_strategy="epoch",
    learning_rate=5e-5,
    gradient_accumulation_steps=2
)

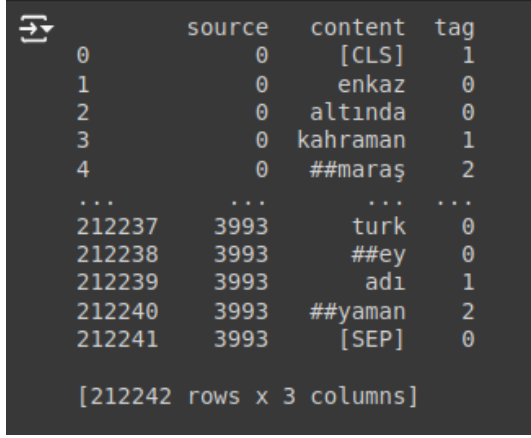
```

Fig. 4. Model Parameters

#### D. Named Entity Recognition (NER) Dataset

After classifying tweets we prepared another dataset for extracting names, locations, and city's (PERSON, LOC, CITY) information from the identified help call tweets. We performed IOB tagging on dataset. We used these tags for mentioned purpose: ['B-CITY' 'OTHER' 'I-CITY' 'B-ADDR' 'I-ADDR'

'B-PER' 'I-PER']. There are subtokens in our dataset. Tokens starting with are sub-word tokens, typically used in tokenization by models like BERT. We used that technique because Turkish is an agglutinative language. It is a much more effective method to receive attachments as subtokens in such language.



	source	content	tag
0	0	[CLS]	1
1	0	enkaz	0
2	0	altında	0
3	0	kahraman	1
4	0	##maraş	2
...	...	...	...
212237	3993	turk	0
212238	3993	##ey	0
212239	3993	adı	1
212240	3993	##yaman	2
212241	3993	[SEP]	0

[212242 rows x 3 columns]

Fig. 5. NER Dataset

By mapping and analyzing social media help calls, this project aims to provide actionable insights for emergency response teams, thereby enhancing the effectiveness of disaster relief operations in real-time.

#### IV. EVALUATION

To assess the performance of our classification and NER models, we will use Precision, Recall, F1-score and Accuracy: Evaluate how accurately and comprehensively the models identify and classify help requests. We compare the metrics of two BERTurk models and SVM

**Model Limitations:** This study implements in Türkiye region and focus on Kahramanmaraş Earthquake.

No visual analysis was performed on tweets containing images. Therefore, the locations mentioned there were overlooked.

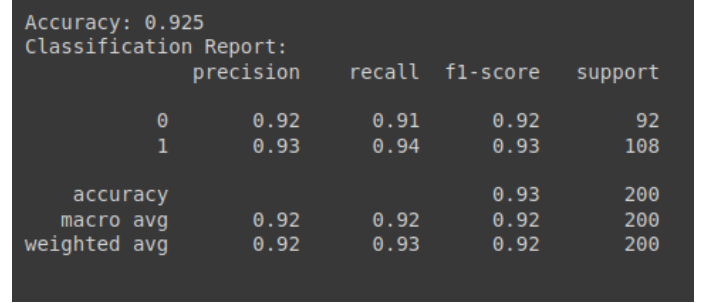
#### V. TOOLS AND LIBRARIES

**Programming Language:** Python is chosen for its versatility and extensive libraries supporting data analysis and natural language processing (NLP). **Kaggle:** Used to gather real-time tweets relevant to the earthquake, enabling the collection of large datasets efficiently. **spaCy:** Employed for its powerful named entity recognition (NER) capabilities, essential for extracting names and locations from tweets. **NLTK:** Assists with text preprocessing tasks such as tokenization and stop-word removal. **Zemberek-Python:** Tailored for Turkish text processing, providing tools for morphological analysis and lemmatization. **Scikit-learn:** Utilized for implementing and evaluating machine learning models, including various classification algorithms. **PyTorch:** Used for building and training deep learning models, specifically Bi-LSTM networks for

enhanced NER performance. **Matplotlib/Seaborn:** Generate visual plots to analyze data distributions and model performance.

#### VI. RESULTS

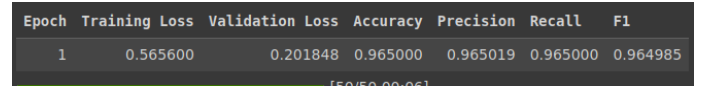
The SVM model showed reasonable performance in tweet classification, but main BERTurk model significantly outperformed it. But the berturk-sunlp-ner-turkish underperformed. The main BERTurk model, loodos/bert-base-turkish-uncased variant achieved the highest accuracy and F1-score, demonstrating the importance of the general purpose models.



Accuracy: 0.925  
Classification Report:

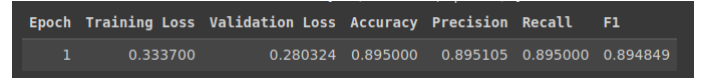
	precision	recall	f1-score	support
0	0.92	0.91	0.92	92
1	0.93	0.94	0.93	108
accuracy			0.93	200
macro avg	0.92	0.92	0.92	200
weighted avg	0.92	0.93	0.92	200

Fig. 6. SVM Results



Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.565600	0.201848	0.965000	0.965019	0.965000	0.964985

Fig. 7. BERTurk Results



Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.333700	0.280324	0.895000	0.895105	0.895000	0.894849

Fig. 8. Sunlp Results

#### VII. CONCLUSION

In conclusion, this project aims to utilize Twitter data to enhance disaster response efforts during the Kahramanmaraş earthquake by identifying and analyzing help requests through advanced NLP techniques. This study demonstrates the effectiveness of advanced machine learning models, particularly BERTurk, in the classification and analysis of Turkish tweets. The results highlight the superior performance of BERT-based models over traditional classifiers like SVM. Future work could involve integrating these models into real-time social media monitoring tools, expanding the approach to other languages and exploring the use of additional contextual features to further enhance performance. The two datasets can be used in another applications.

#### REFERENCES

- [1] "Named Entity Recognition on Morphologically Rich Language: Exploring the Performance of BERT with varying Training Levels" Yuksel Pelin Kilic, Duygu Dinc, Pinar Karagoz (2020).
- [2] "Predicting Themes of Tweets on Earthquakes in Turkey Syria for Real-Time Classification" Subhanan Mandal, Mamatha Alugubelly, Manoj Jayabalan (2023).

- [3] "HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning" Firoj Alam, Umair Qazi, Muhammad Imran, Ferda Ofli (2021).
- [4] "Real-time event detection in social media streams through semantic analysis of noisy terms" Taiwo Kolajo, Olawande Daramola and Ayodele A. Adebisi (2022).
- [5] "A Twitter Corpus for Named Entity Recognition in Turkish" Buse Carik, Reyhan Yeniterzi (2022)
- [6] <https://www.kaggle.com/datasets/swaptr/turkey-earthquake-tweets/data>
- [7] [loodos/bert-base-turkish-uncased](#)
- [8] [busecarik/berturk-sunlp-ner-turkish](#)
- [9] <https://www.ibm.com/topics/support-vector-machine>
- [10] <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>