IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CLASSIFICATION OF THE EFFECTS OF NATURAL DISASTERS ON STRUCTURES THROUGH SOCIAL MEDIA POSTS WITH MACHINE LEARNING METHODS

[1]Utku Süsoy, [2]Özlem Aktaş

[1]Dokuz Eylül University, The Graduate School of Natural and Applied Sciences, Department of Computer Engineering, İzmir, Turkey

[2]Dokuz Eylül University, Department of Computer Engineering, İzmir, Turkey

Abstract: Earthquakes cause massive damage to people and structures. The capacity to quickly assess damage over a large area is critical for successful disaster response. In recent years, social networks have proven to be a great capability to increase situational awareness and identify affected areas. In this context, this paper presents a method for assessing damage levels in earthquake zones utilizing social media data and the Naive Bayes, Support Vector Machine (SVM), Random Forest, and BERT classification algorithms. In order to compare different machine learning models in this study, we utilized our post-earthquake damage classification dataset consisting of Turkish Tweets, which we prepared and labeled for the problem. Identifying damaged structures from Turkish tweets after the earthquake is the first step of the solution which we propose. For spatial analysis, we need to extract the address information of the damaged structures from the tweets. Therefore, in this paper, we use Named Entity Recognition for address extraction and fine-tune the pre-trained BERT model with our own compiled detailed address detection NER dataset. Finally, in order to make disaster response more successful and effective, we obtained the latitude and longitude of damaged structures in the earthquake zone by obtaining the geographical coordinates of the addresses using the geocoder API, the address information obtained with the help of the NER model. Thus, rescue teams can intervene more effectively and increase their success rates.

Keywords **– Social Media, Twitter, Machine Learning, BERT, Named Entity Recognition (NER), Data Mining, Geocoding**

I. INTRODUCTION

Earthquakes are one of the most dangerous natural disasters. With their untimely occurrence, the damage they will leave in the impact area cannot be fully predicted. Earthquakes not only affect buildings but also have serious effects on people. However, after the earthquake, damage assessment works in the structures and rescue operations for people trapped under the rubble are vital. In these studies, the ability to quickly assess the damages in the affected area is critical for disaster management. Since time is of vital importance in natural disasters, authorities need accurate information on the spatial distribution of the damage in order to send aid to the affected areas quickly and quickly. In a devastated area, there is a race against time to rescue those trapped under the rubble. Rescue workers need to quickly identify the wrecked structures and start the necessary work in an organized manner. The sooner the spatial information of the demolished structures reaches the authorities, the more effective it will be for the teams to organize, plan the rescue process and take action.

As important as it is to reach people under rubble quickly in earthquakes, the damage caused by the earthquake on buildings should not be ignored. Earthquakes do not destroy every building. But many buildings are damaged on various scales. These damages pose a great danger for aftershocks and other possible earthquakes. For this reason, the authorities being aware of the damaged buildings immediately after the earthquake will prevent future losses of life and property.

With the widespread use of wireless internet and mobile phones, communication, and information transfer is much easier and more possible than before. In recent years, social media platforms (Twitter, Facebook, etc.) are one of the most common areas of communication today. It allows information to spread rapidly all over the world and increases social awareness. Immediately after natural disasters, the number of interactions about the region by the local people increases considerably on social media [1]. These interactions are mostly content describing the damage in the region, such as collapsed buildings, damaged structures, or calls for help. This kind of content is vital for authorities and rescue teams immediately after an earthquake. Because the sooner the authorities identify the damaged structures and affected areas, the sooner the rescue teams can organize and take action on the disaster area. As a result, the loss of life and property in the region can be minimized. However, manually identifying and analyzing damaged content on social media such as Twitter is a time-consuming, tiring, and costly process.

In this paper, we analyzed Turkish tweets using machine learning models on social media platforms such as Twitter immediately after the earthquake. Using various classification models, we classified the damaged structures in the region according to their damage status such as "weak", "violent" and performed the spatial analysis with NER to obtain the location information of the damaged content. In addition to all these, we compared the damage classification dataset we created with machine learning models such as Naive Bayes, SVM, Random Forest, and the currently popular language model such as BERT.

II. Related Works

The use of Online Social Networks (OSNs) for monitoring natural disasters has become a popular topic in recent research, and one approach involves geoparsing specific locations using Named Entity Recognition (NER) techniques to extract vital information from tweets, which is then scored using a Kernel Density Estimation function to visualize areas where tweets related to natural disasters are concentrated [3]. Another study builds upon previous research on the use of social media for disaster response and proposes an approach that applies Naive Bayes, support vector machine (SVM), and deep learning classification algorithms to assess earthquake damage at different scales, with validation using confusion matrix metrics, Spearman's rho, Pearson correlation, and Kendall's tau [4]. Also, Faxi et al. (2019) analyze the social media contents via a machine learning approach for identifying the damage-related social media data and classify social media data in future disaster events. Bernd et al. (2017) analyzed the social media contents with machine learning models and created a heat map by classifying the damage in the region. Also performed spatial and temporal analysis of the disaster area by combining machine learning models with (Latent Dirichlet Allocation) semantic extraction techniques. Priyanka et al. (2020), Performed sentiment analysis on Twitter by categorizing tweets as positive or negative. He also tested various machine learning models such as Logistic Regression (LR), Multinominal Naive Bayes (MNB), and Support Vector Machine (SVM) and compared the results. Also, another spatial-temporal pattern analysis was performed by Zou et al. (2021). Their aim is to categorize and analyze social media posts to better understand and analyze disaster responses and behaviors during real-time disasters.

To automate the extraction of operational information from social media posts to efficiently direct help resources, Hussein et al. (2018) combines multiple pre-trained unimodal convolutional neural networks that extract features from raw text and images independently, before a final classifier labels the posts based on both modalities. Another research is that classified social media content as damaged or undamaged by binary classification of user-generated data in social media with various machine learning models. Then, they aimed to help rescue teams by showing damaged tweets in the form of a heat map [4]. Also, Sajjad et al. (2021) improve their previous research that utilized the text messages from social media networks to create a map that depicted the scope of the devastation. To detect mentions of damage within these messages, they performed a machine-learning approach which called Support Vector Machine (SVM). A thorough damage map was created by focusing on tweets about damage. The results showed that the SVM classifier efficiently differentiated messages containing the damage, and its performance was measured using metrics such as F-score, precision, recall, and accuracy. The study also looked at the time patterns of damage and non-harm tweets, evaluating them daily and hourly.

Kumar et al. (2018) focused on location extraction on Twitter media. They aimed to find location-reporting words in tweets using Convolutional Neural Networks (CNN). Their findings can use in many problems such as emergency situations, location advertising, event localization, etc. Topcu et al. (2021) have worked on the Named Entity Recognition (NER) model with BERT to improve the performance of grammatically incorrect and long-written Turkish search queries in Search Engines. Yeniterzi et al. (2018) demonstrates a thorough analysis of the work on named-entity identification in Turkish as well as the data sources that different research initiatives have generated. Malmasi et al. (2015) studied on auto-detection of location expressions in social media contents. They presents a novel approach that finds addresses via Noun Phrase extraction and they compare their results with traditional entity findings methods such as Named Entity Recognition (NER) and Conditional Random Fields (CRF).

III. Dataset Collections

Two different datasets were created in order to obtain successful domain-specific results from the Classification and NER models separately. Post-Earthquake Damage Classification dataset was prepared for Classification and Location Oriented NER dataset was prepared for detailed address extraction.

3.1 Post-Earthquake Damage Classification Dataset

To create the Post-Earthquake Damage Classification dataset, the recent major earthquakes in Turkey were retrieved from Twitter using scraping tools. The dataset was created by fetching and labeling thousands of tweets from recent earthquakes in Turkey, such as the January 24, 2020 Elazığ Earthquake (6.8 Mw), 30 Ekim 2020 Izmir Earthquake(6.8 Mw) and 23 Kasım 2022 Düzce Earthquake (6.0 Mw). The tweets were taken in two different time intervals: long-term and short term.

- **Short-Term Scraping**: Time of the Earthquake – until 3rd day of the Earthquake
- **Long-Term Scraping:** Time of the Earthquake – until a month after the Earthquake

In the short-term scraping, it was tried to reach the tweets containing devastating damage of the earthquake. Contents, containing information about collapsed buildings, heavily damaged structures, and people who trapped under rubble were included in the scope. In long-term scraping aims to focuses on contents which containing information on weak and medium damaged structures after an earthquake. In order to facilitate the retrieval of relevant tweets, keyword-based and hashtag-based Twitter search methods were preferred in both time periods. Words such as "earthquake"(deprem), "debris"(hasar), "damage"(hasar), "column"(kolon), "crack"(çatlak), "afad" were searched to retrieve tweets suitable for the problem.

With this methodology, thousands of tweets were retrieved as described above and 6063 tweets were manually labeled as "None", "Weak" and "Violent" according to the damage levels of the tweets.

Table 1. Distribution of labels in the Post-earthquake Damage Classification Dataset

Labels	Number of Samples
Violent	1483
Weak	1565
None	2015

- **Violent:** Tweets containing information about heavily damaged and destroyed buildings and people who trapped under rubble.
- **Weak:** Tweets with information about lightly damaged and moderately damaged buildings.
- **None:** Tweets that do not contain any damage indications about buildings or the damage caused by the earthquake.

3.2 Location-Oriented NER Dataset

Categorizing Turkish tweets according to the damage status of buildings is an important process. But getting detailed addresses of these damaged buildings is much harder and more important to help people. Existing NER datasets are insufficient to train NER models capable of capturing detailed and comprehensive addresses. For this reason, we changed the labels of the MilliyetNER Tür et al. (2003) dataset slightly to make the existing dataset capable of capturing locations in more detail. In addition, we scraped and manually labeled 500 current news articles which are containing detailed addresses and enriched the MilliyetNER] Tür et al. (2003) dataset.

Table 2. Frequency distribution of entities in the Location Oriented NER Dataset

Labels	Entity Count
Location	13066
Organization	9588
Person	16924
Total Words	550775

IV. METHODOLOGY

In this paper, machine learning solutions to extract location information from damage-related social media contents are proposed. Figure 1 shows the proposed approach which includes both damage assessments in social media posts such as tweets and then extraction of geocodes from contents that are classified as damaged. According to the Figure 1, relevant earthquake tweets scrape from Twitter media immediately after the earthquake. Then, machine learning methods are used to find out which tweets contain damaged information. In this way, the tweets are classified with the most appropriate tag according to the level of damage. After the identification of damage classification on tweets, some location clues try to find inside of the damaged related tweets via Named Entity Recognition. Finally, using the locations which found in the tweets, the coordinates of the place where the damage occurred or the wreckage area are accessed via the Geocode API.

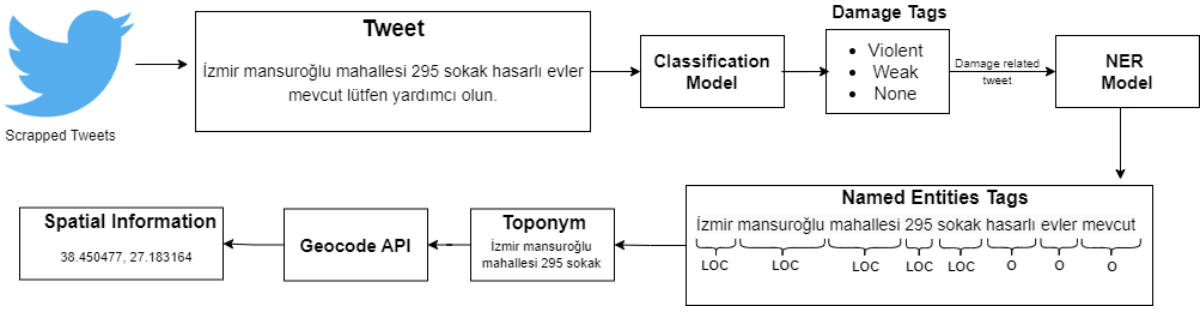

Figure 1. General machine learning flow for location extraction from damaged tweets

4.1 Data Preprocessing

To lessen any statistical "noise," the dataset needs to be preprocessed before performing the real semantic and spatial analysis on the Tweets. The full pipeline for semantic and geographical analysis is shown in Figure 2. The preprocessing and vectorizing processes are outlined in the paragraphs that follow. It should be emphasized that the order in which the preprocessing and vectorizing stages are completed is plainly crucial to ensuring the highest level of results reliability.

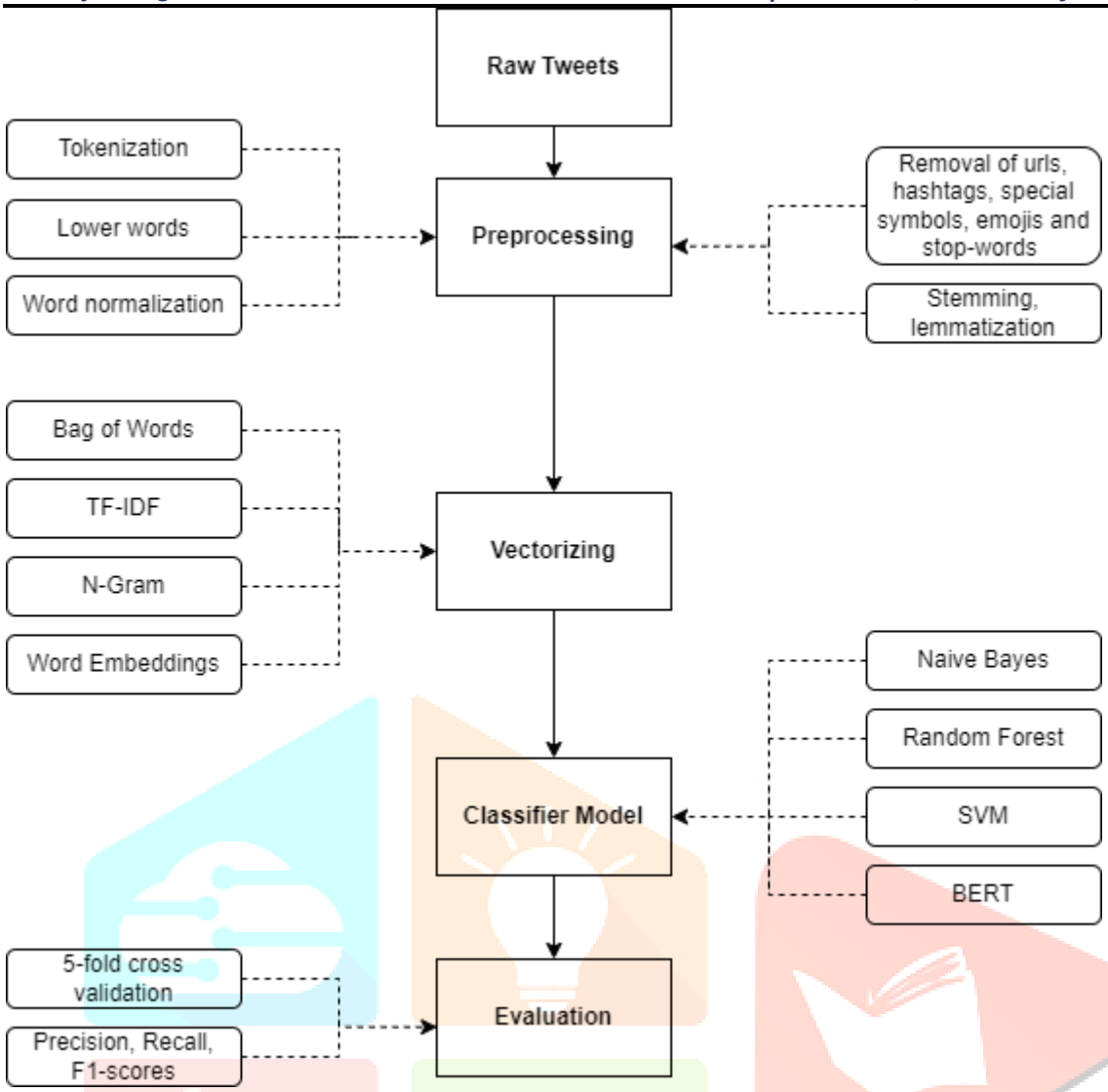

Figure 2. Process flow diagram for identifying damage-related tweet

Tokenization

In this step, the words in each Tweet sentence are separated from each other by the white space character. This allows each token to be handled and processed individually and the remaining processing steps to be performed relatively easily.

Tokenization is an important step in text preprocessing because many natural language processing (NLP) algorithms and models rely on tokens as their basic unit of input. By breaking a text document down into tokens, we can perform various operations on the text, such as counting the frequency of individual words or phrases, analyzing the syntactic structure of sentences, or training machine learning models to recognize patterns in the text.

Removal of Potential Noise

Tweets usually contain content such as URLs (https://…), hashtags (#earthquake), person tagging (@someone), and emojis that can create noise when they are not processed. Since the above-mentioned contents contain non-specific and interpretable semantic information on their own, it should be preferred to try to extract contextual meanings that will contribute to the subject by processing them separately. However, in this article, we preferred deleting structures that could potentially cause noise.

In addition, words which are known as stopwords are considered useless, meaningless, and commonly used in sentences were deleted from tweets. Also, punctuation and other special symbols have been removed from tweets and words. Finally, duplicate tweets and empty tweets (that have no characters left after the cleaning) were removed from the data sets.

Tokens to Lowercase

This step aims to reduce the feature space by converting all uppercase and lowercase forms of the same word into the same form. Thus, all capitalized words are minimized, and as a result, the feature space is reduced, the contextual meaning is better understood and the training process becomes more faster and efficient.

Word Normalization

Tweets contain a lot of typos, such as incomplete and misspelled words. Since there is no specific pattern to the mistakes made in misspelled words, they increase the feature space considerably. This makes it difficult to grasp the contextual meaning. As a solution to this, it is aimed to identify the incorrect words in the text data, find the words that are similar to the correctly written word, and replace them with the correct version of the incorrect word. Thus, it is aimed to narrow the feature space of the words and correct the words that want to give the same meaning. In this way, it is thought to facilitate the comprehension of contextual meaning.

In addition, abbreviations in addresses indicating place, space, and location have been corrected in a common structure for easier spatial analysis (abbreviations indicating place, space in addresses such as sk, sok, skk, cadd, cd, mah, mh have been corrected as sokak (street), cadde (avenue), mahalle (neighborhood) respectively).

Root Operations

Stemming and lemmatization are important techniques in natural language processing cause of reducing vocabulary size and improving text classification results. Both techniques help reduce the number of unique words in the text data. By transforming different forms of the same word into a single word, the vocabulary size can be reduced, Thus reducing the number of unique features and helps improve the accuracy of the classification model.

In this study, Snowball Stemming and Zemberek's lemmatization algorithm were used as root operations and word normalizations. In this way, it is aimed to provide the above-mentioned benefits and to provide better results of machine learning and a better understanding of the contextual meaning.

Vectorizing Techniques

One of the most difficult difficulties in machine learning is selecting the most relevant variables to include in a model. This can be a time-consuming and difficult procedure, but advances in machine learning have made it simpler and more effective. Vectorizing is one such advancement, which includes transforming text data into numerical vectors that can be utilized as input to machine learning models. Vectorizing can assist in identifying the most informative words or phrases in a text collection and removing noise or extraneous information.

There are several ways to vectorize text data, including term frequency-inverse document frequency (TF-IDF), bag-of-words, and n-grams. TF-IDF assigns weights to each term in a document based on how often it appears in that document and across the entire dataset, while bag-of-words simply counts the occurrences of each term in a document. Both approaches can be effective in different situations, and it often takes experimentation to determine which is the most appropriate for a given dataset. In addition to TF-IDF and bag-of-words, n-gram vectorizing is an other vectorizing technique for text data. N-grams are contiguous sequences of n words in a text, and n-gram vectorizing is the process of counting the occurrences of all n-grams in a text and using the counts as features for machine learning models. N-gram vectorizing, which takes into account the sequence in which words appear in a sentence or document, can capture more information about the relationships between words in a sentence or document than bag-of-words. However, when n grows larger, the number of distinct n-grams in a text grows exponentially, potentially resulting in high-dimensional feature spaces and the risk of overfitting.

Another way to represent text data as numerical vectors is through word embeddings, which use neural networks to create dense, low-dimensional representations of words or phrases. Word embeddings have become a popular approach in natural language processing (NLP), as they can capture the semantic relationships between words and phrases in a way that is difficult to achieve with other methods.

Using vectorization as a feature selection technique and incorporating word embeddings can lead to significant improvements in machine learning models' performance. By identifying and representing the most informative features of a dataset, vectorization can reduce noise and increase the signal-to-noise ratio, making it easier for a model to identify patterns and make accurate predictions.

In this study, we try all the techniques mentioned above and observe their effects on model performances. By processing both preprocessing techniques and vectorization methods in the Turkish damage classification dataset, we measure their performance among themselves. As a result, we try to find out which technique or method gives more effective results with our dataset and problem specific.

4.2 Damage Estimation

In this study, we evaluated the damage caused by the earthquake on buildings in 3 stages: None, Weak and Violent. Tweets taken in certain periods after the earthquake are considered as damaged tweets if they contain damage information about the structures. If the building is slightly damaged due to the damage, that is, if the building is safe for people to live in, the building is classified as "weak". Or, if the building has collapsed or the bearing parts of the building have been severely damaged to endanger the safety of the building, it is classified as “violent”. In this way, the damage caused by the earthquake on the buildings in the region can be determined immediately after the earthquake and at certain intervals, and the building stocks of the region can be analyzed.

4.3 Spatial Analysis

Filtering tweets by detecting damaged structures in tweets is an important step, but it is not enough on its own. Because the location information of the damaged structures is needed in order for the aid to reach the region. In this study, our main goal to obtain the addresses of damaged buildings using the Named Entity Recognition (NER) model. Considering the address information as an entity, we marked the words describing the region where the mentioned structure is located as Location Entity (B-LOC, I-LOC). Then, we accessed the coordinates of the damaged structures through the Geocode API with the obtained address information from NER model. In this way, we aimed to facilitate the access of rescue teams or authorities to damaged areas or structures more quickly and effectively.

4.4 Used Models

We provide a general description of the models used in multi-class classification and Named Entity Recognition (NER) problems and examine their advantages. Multi-class classification models are used to understand the extent of damage caused by earthquakes on buildings. After the detection of damaged structures, entity recognition was performed using the BERT model to access the location information of the structures. As mentioned in the rest of the chapter, Naive Bayes, Random Forest, Support Vector Machine (SVM), and BERT models are discussed respectively.

Naive Bayes

Naive Bayes is a probabilistic model based on the Bayes theorem that assumes the input features are independent of one another. It is extensively used in natural language processing (NLP) for text classification tasks such as spam detection and sentiment analysis. Given the input features, Naive Bayes calculates the likelihood of each class and chooses the class with the highest probability as the forecast. Naive Bayes has the advantage of being computationally efficient and capable of handling huge datasets with many features. Also works well even with a small dataset. It also works well when the data is high dimensional and the independence assumption is met. Finally, the Naive Bayes has the advantage of not requiring hyper-parameter tuning in compared to the other methods.

Random Forest

Random forest is an ensemble learning method for improving classification accuracy by combining numerous decision trees. Random forest constructs numerous decision trees from distinct subsets of the training data and uses majority voting to integrate their predictions. Random forest has the advantage of being able to handle high-dimensional data (i.e., many features) and being resistant to noise and overfitting.

Support Vector Machines (SVM)

The main principle behind SVM is to discover the hyperplane that best separates the data into multiple classes. The hyperplane is a decision boundary that categorizes data points depending on their characteristics. In this particular case of text classification, the features might be the words in the text or the frequency of certain words in the text. SVM works by mapping the input features into a higher-dimensional space where the data points are more likely to be linearly separable. This is accomplished by the use of a kernel function, which maps the input features into a higher-dimensional space without actually computing the coordinates of the data points in that space. SVM determines the hyperplane that best splits the data into various classes after mapping the data points into higher-dimensional space. This hyperplane is known as the greatest margin hyperplane, and it is the hyperplane that maximizes the distance between the nearest data points from each class. Finally, SVM is a robust and effective method for handling non-linearly separable data and imbalanced datasets and is commonly used in NLP for tasks.

BERT

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model that has produced cutting-edge outcomes on a variety of NLP tasks. It is built on the transformer architecture and is trained using a masked language modeling task on massive volumes of text data. Question answering, sentiment analysis, and language translation are all common uses for BERT. On the other hand, since the transformers trained with a limited token limit such as 256 or 512, the sentences are truncated during the training phase, there is a loss of meaning in the interrupted sentences. As a result, as the number of tokens increases, their performance and tendency to comprehend the whole content decreases. For this reason, character-limited content such as Twitter tweets is relatively easier to interpret. BERT has the advantage of being able to handle sophisticated verbal occurrences such as sarcasm and negation. It also excels at a wide range of NLP tasks without the need for task-specific feature engineering.

V. RESULTS AND DISCUSSIONS

In this section, we analyze our experimental results for both damage classification and Named Entity Recognition problems using our Twitter dataset and discuss our different preprocessing methodologies that we use in model training.

5.1 Damage Classification

Table 3. Performance evaluation of the selected classifier models based on different preprocessing techniques.
Abbreviation: P = Precision, R = Recall, Lemma = Lemmatization, Stem = Stemming

Models	Bag of Words			TF-IDF			N - gram			Preprocessing Technique			
	P	R	F1	P	R	F1	P	R	F1	Base Cleaning	Lemma	Stem	Word Normalize
SVM	0.72	0.68	0.70	0.72	0.71	0.71	0.75	0.69	0.71	+	-	-	-
	0.71	0.67	0.68	0.71	0.67	0.68	0.73	0.71	**0.72**	+	+	-	-
	0.72	0.70	0.71	0.74	**0.71**	**0.72**	0.74	**0.71**	**0.72**	+	-	+	-
	0.74	0.68	0.70	0.74	0.68	0.70	**0.76**	**0.71**	**0.72**	+	-	-	+
Random Forest	0.70	0.68	0.69	0.70	0.68	0.69	0.71	0.68	0.69	+	-	-	-
	0.70	0.69	0.69	0.70	0.70	0.70	0.71	0.70	0.70	+	+	-	-
	0.70	0.67	0.68	0.70	0.66	0.67	0.71	0.68	0.69	+	-	+	-
	0.72	0.70	0.71	0.72	0.69	0.70	0.72	0.70	0.71	+	-	-	+
Naive Bayes	0.68	0.66	0.67	0.70	0.61	0.63	0.72	0.68	0.70	+	-	-	-
	0.69	0.64	0.66	0.69	0.64	0.66	0.72	0.65	0.67	+	+	-	-
	0.68	0.65	0.66	0.68	0.65	0.66	0.72	0.66	0.68	+	-	+	-
	0.68	0.66	0.67	0.68	0.66	0.67	0.73	0.68	0.70	+	-	-	+

We experiment many machine learning models with tweets containing damage to detect the damage caused by the earthquake on buildings. As seen in Table 3, each model was trained separately by applying different preprocessing methods and different vectorization techniques. The dataset is divided into two as 70% training and 30% test dataset. The hyper-parameters of the 70% training set were optimized by applying by 5-fold cross-validation for each model. We also used grid-search techniques at each fold to find the best hyperparameters for each model. After the training, the performances of each model were evaluated according to the precision, recall, and f1 scores. As can be seen in Table 3, almost all models produced similar results. The biggest reason for this can be thought of as too much noise in the data set and limited data sources. However, the SVM model based on N-gram vectorization processed with "Word Normalization" stands out as the model that achieves the best scores compared to other models. Because tweets contain a lot of typos, much better results can be obtained with a proper word correction algorithm.

In addition to the statistical machine learning models, the Deep learning-based BERT model was trained with the same data set. Transformer architecture, which has been a state of art in text classification problems, has overcome other models in our problem of detecting damaged structures from tweets. As can be seen in Table 4, the models were trained separately with different preprocessing methods and the evaluation results are shown in Figure 3. The best scores were obtained in base cleaning and word normalization preprocessing techniques.

On the Twitter dataset, which contains a lot of noise, the transformer architecture seems to give better results than statistical machine learning models. The main reason for this is that BERT processes billions of words in a pre-trained dataset. This allows even previously unseen or rare examples to be understood and predicted. Also, because BERT is a pre-trained language model, it can process text in both directions (i.e. both forward and backward) as it reads it. This allows it to better understand the context of the text and make better predictions.

Table 4. Performance evaluation of the pre-trained BERT classifier models based on different preprocessing techniques.

Pretrained BERT Model	Precision (P)	Recall (R)	F1-Score
Base Cleaning	**0.79**	**0.80**	**0.79**
Lemmatization	0.75	0.76	0.75
Word Normalization	0.78	**0.80**	**0.79**
Stemming	0.75	0.77	0.76

5.2 NER

The pre-trained BERT model was used to determine the addresses of damaged buildings from tweets. In this way, detailed location information of the damaged buildings was extracted from the tweets. Before putting the tweets into the BERT tokenizer, all letters were reduced by preprocessing. The aim is for BERT to capture the context in sentences and locations while leaving punctuation untouched. The BERT model was trained for 20-25 epochs with the existing Location Oriented NER Dataset and the validation loss was reduced to 0.059 without overfit. As can be seen in Figure 4, the evaluation metrics of our model gave very good results after training. In fact, it is obvious that the results are even better than the classification process. The main reason for this is that the NER dataset is much larger and fine-tuning is done on a location-specific basis. Looking at the B-LOC and I-LOC values in Figure 4, we can see that the model is quite successful in capturing long addresses. It is also seen that the model does a good job in capturing names. This can help us to reach people trapped under the rubble of collapsed buildings.

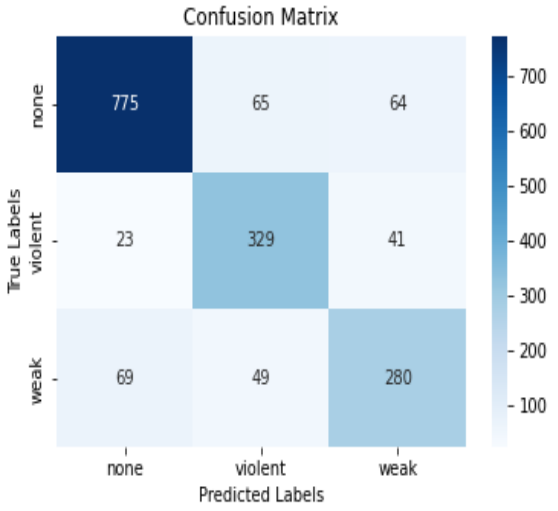

Figure 3. Pre-trained BERT model Confusion Matrix on Damage Classification

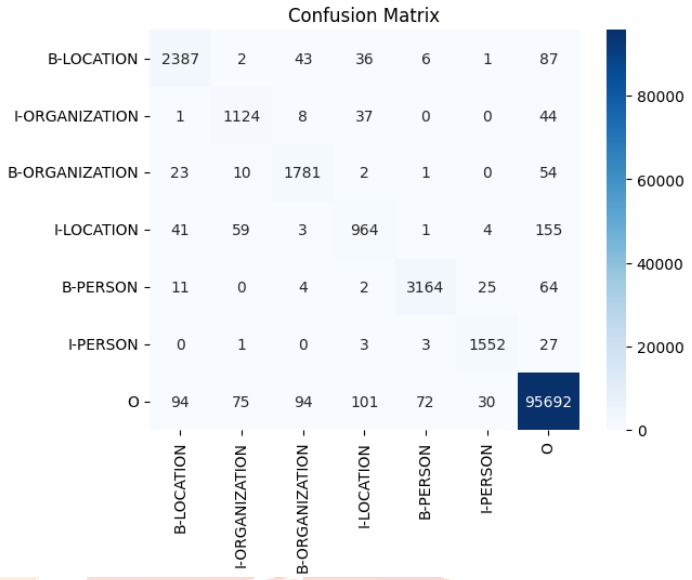

Figure 4. Pre-trained BERT model Confusion Matrix on NER task

VI. CONCLUSION

In this study, we performed the damage assessment analysis according to the levels of the damaged buildings in the earthquake zone after the earthquake and the location determination of the damaged building. We compared various machine learning and deep learning models such as Naive Bayes, Support Vector Machine (SVM), Random Forest, and BERT when detecting damaged tweets. We compared the models among themselves with the Post-Earthquake Damage Classification Dataset in the Turkish tweeters we created. In addition, we analyzed the effects of different preprocessing techniques and vectorization processes on the training process and their performance. As a result of these experiments, the pre-trained BERT model, which is state of art in language modeling today, has been superior to other models.

Another area of our work is to reach the addresses of the damaged buildings by performing a spatial analysis of the tweets which contain the damaged buildings. To this end, we used the Named Entity Recognition (NER) model to extract the location names mentioned in the tweets in detail. For this purpose, we extend existing pre-prepared NER datasets, added on top of them, and detailed the address contents. Finally, we used the obtained address information through the geocoder API to access the exact locations (latitude, longitude) of the damaged structures. Collaborating with the government and public groups to enhance the rapid detection of disaster regions, the identification of missing people, and the real-time location of damaged areas utilizing our suggested technique is part of our future work.

REFERENCES

[1] Bernd Resch, Florian Usländer & Clemens Havas (2017): Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment, Cartography and Geographic Information Science

[2] Gökhan T. Dilek H. Kemal O. 2003. A statistical information extraction system for Turkish. Cambridge University Press. Natural Language Engineering , Volume 9 , Issue 2 , June 2003 , pp. 181 - 210

[3] Aldo Hernandez-Suarez, Gabriel Sanchez-Perez, Karina Toscano-Medina , Hector Perez-Meana, Jose Portillo-Portillo , Victor Sanchez and Luis Javier García Villalba. Using Twitter Data to Monitor Natural Disaster Social Dynamics: A Recurrent Neural Network Approach with Word Embeddings and Kernel Density Estimation. 11 April 2019.

[4] Ahadzadeh, S.; Malek, M.R. Earthquake Damage Assessment in Three Spatial Scale Using Naive Bayes, SVM, and Deep Learning Algorithms. Appl. Sci. 2021, 11, 9737

[5] Faxi Yuan1 and Rui Liu, Ph.D. Identifying Damage-Related Social Media Data during Hurricane Matthew: A Machine Learning Approach. M. E. Rinker, Sr. School of Construction Management, Univ. of Florida. 2019

[6] Hussein Mouzannar Yara Rizk Mariette Awad. Damage Identification in Social Media Posts using Multimodal Deep Learning. Department of Electrical and Computer Engineering American University of Beirut. May 2018

[7] Ahadzadeh, S.; Malek, M.R. Earthquake Damage Assessment Based on User Generated Data in Social Networks. Sustainability 2021, 13, 4814.

[8] Priyanka Harjule, Astha Gurjar, Harshita Seth, Priya Thakur. Text Classification on Twitter Data. Department of Computer Science. IIIT Kota. 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE-2020), 07-08 February 2020, (IEEE Conference Record 48199)

[9] Lei Zou, Nina S. N. Lam, Heng Cai, and Yi Qiangy. Mining Twitter Data for Improved Understanding of Disaster Resilience. Department of Environmental Sciences, Louisiana State University y Department of Geography, University of Hawaii–Manoa**.**

[10] Shervin Malmasi, Mark Dras. Location Mention Detection in Tweets and Microblogs. Centre for Language Technology Macquarie University Sydney, NSW, Australia. 2015 Pacific Association for Computational Linguistics, 2021

[11] Reyyan Yeniterzi, Gökhan Tür, and Kemal Oflazer. Turkish Named-Entity Recognition. © Springer International Publishing AG, part of Springer Nature 2018 K. Oflazer, M. Saraçlar (eds.), Turkish Natural Language Processing, Theory and Applications of Natural Language Processing

[12] Abhinav Kumara, Jyoti Prakash Singh. Location reference identification from tweets during emergencies: A deep learning approach. Department of Computer Science and Engineering National Institute of Technology Patna, India

[13] Berkay Topcu, İlknur Durgar El-Kahlout, TR-SEQ: Named Entity Recognition Dataset for Turkish Search Engine Queries, 2021, Turkcell Technology, İstanbul, Turkey