6th Intercontinental Geoinformation Days

igd.mersin.edu.tr

# Getting disaster information through the web

**Lutfiye Kuşak**[*1]

[1]*Mersin University, Engineering Faculty, Department of Geomatics Engineering, Mersin, Turkiye*

**Keywords**
Disaster
News
Internet
Text Mining
Sentiment Analysis

**Abstract**

In the world and in Türkiye, there are numerous geological, climatic, biological, social, and technical disasters. While some of these disasters occur for natural reasons, others are caused on by humans. Numerous factors, such as population growth, varying meteorological conditions, and an unplanned and unregulated rise in impermeable surface area, negatively impact the impact on areas of disasters. Today, disaster management, documentation, and monitoring are the primary concerns in various countries. For example, there is not any doubt about the amount of work, time, and expense involved in gathering disaster data. Alternative techniques should be considered as supporting evidence in addition to terrestrial observations. Data collection via the internet is a topic of interest nowadays, especially given how widely it is used. This study has therefore looked into how to collect disaster data from online sources.

## 1. Introduction

Numerous financial and environmental costs result from disasters. This condition causes people to die, get injured, lose their residences, and suffer economic and psychological damage. Disasters are typically divided into two categories: natural and man-made. The categories of geological, climatic, biological, social, and technological disasters are addressed all over the world. Geological disasters include earthquake, landslide, rockfall, volcanic explosion, mudflow, and tsunami. Climate disasters include heat waves, cold waves, droughts, hailstorms, tornadoes, lightning, hurricanes, typhoons, floods, cyclones, tornadoes, blizzards, avalanches, and extreme snowfall. Biological disasters include erosion, forest fires, epidemics, and insect invasion. Among the social disasters are migration, wars, terrorist acts, and fires. Technological disasters include mining accidents, incidents involving biological, nuclear, and chemical weapons, accidents involving industrial equipment, and accidents involving transportation. Nuclear, biological, and industrial disasters are the most common created by humans'. Disasters like avalanches, earthquakes, floods, and storms develop quickly but disasters like famine and drought take time to develop. One of the most frequent types of natural disasters is the climactic one.

Records should be maintained on a regular basis in order to control disasters and take precautions. In terms of time, effort, and cost, collecting and storing disaster data is a hard task. For this reason, remote sensing data collected using a variety of platforms, including UAV, satellite, and airborne, is employed in research together with ground measurements and observations. In order to carry out disaster analysis and management in a healthy way, the attribute data, which includes information such as time, disaster location, and impact area of these data, must be very carefully recorded. This is why online resource-based feature extractions have been used in studies recently.

As of April 2023, approximately 5 billion users are actively using the internet, while there are 4.8 billion active social media users (Statista, 2023). This number of users means a large data set for all commercial and non-commercial environments. For this reason, many studies are carried out to extract information by making the data meaningful.

Web usage mining, web content mining, and web structure mining are the three primary categories that web mining is analyzed under. Web mining is a sub-branch of data mining that attempts to analyze and process information on the web. The main focus of his research is web content mining, which involves activities like identifying the languages used on the pages of

* **Corresponding Author**

*(lutfiyekusak@mersin.edu.trl) ORCID ID 0000-0002-7265-245X

**Cite this study**

Kusak, L. (2023). Using online news to find disasters. Intercontinental Geoinformation Days (IGD), 6, 335-340, Baku, Azerbaijan

websites, disclosing the word density, locating the keywords, and categorizing the websites. Image processing or natural language processing techniques are employed for this.

Web mining generally uses text mining methods. However, unlike text mining, the preprocessing process is a little different because web pages contain html tags. Today, there are many studies in which social media content is used as a data set.

### 1.1. Related Works

Web content mining has been the subject of many studies. These studies may be intended to find social media content, track national and international news, and highlight people's emotional responses to some significant events.

In the study conducted by Duan et al., up-to-date data set was obtained for social media contents and management information system. Tweets were analyzed using text mining methods and Supervised machine learning algorithms (Duan et al., 2023).

In the study conducted in 2020, one-week agenda topics for Kuwait were determined using Twitter data. It has been determined that the most topics are opened on religion, emotions, laws and education in the selected date range, and different topics are trending in different locations every day of the week. In the study, as in all previous studies, data crawling was done first, then data filtering was performed and word frequencies were determined. Subject headings were created according to the results obtained (Almatar et al., 2020).

In the Sewol ferry disaster that took place in South Korea in 2014 and 300 people died, it was tried to understand the trauma and modes of the public by using social media data. For this, twitter data belonging to 3 different periods, pre-event, during the event and after the event, were analyzed using natural language processing and text mining. It has been determined that anger and sadness data are high in the period of approximately 1 month after the event, and anger increases in every 5-day period at various times until May 30 (Woo et al., 2015).

Web mining techniques are often applied in healthcare environments. Preprocessing steps were carried out with NLP and data cleaning methods in the study in which the comments made by the patients on the web sites were acquired by web scraping method. The data was analyzed using Nvivo software, correlation analysis, and sentiment analysis methods, and it was presented using a word cloud, tabular sheet, and graphic representations (Hameed, 2023).

Geotagged information makes it easier to show the location of data and to make decisions. In the 2017 study, two separate investigations were conducted on Twitter, consisting of geotagged and non-geotagged flu and a movie. According to the findings of the study, geotagged data is significantly more consistent with events, noise data is much lower, and correlations are high (Issa et al., 2017).

Web mining applications may be used in the monitoring, tracking, and analysis of epidemics in the context of biological disasters. COVID-19, which affects the entire world, is one of the diseases for which social media is employed in the context of biological disasters. The correspondence linked to COVID-19 over a period of around one month was analyzed in a study conducted in China utilizing Weibo, China's most popular social platform. The most time periods of the comments submitted on COVID-19 were gathered using the Seasonal-trend decomposition approach based on Loess (STL), one of the time series analyses. Furthermore, LDA, a Topic extraction algorithm, and Random Forest, a classification model, were merged. Furthermore, the Kernel density analysis indicated the densities of Weibo messages associated with COVID-19 based on geographical locations (Han et al., 2020).

COVID-19 has had numerous results, not just in terms of health, but also in terms of economics. To study these effects, data on agriculture and COVID-19 were acquired from the WEChat and WEibo platforms using a web crawler and text mining. First, data denoising was conducted, followed by word segmentation. LDA was used to create high frequency word groupings and determine the topic. The impact of the COVID-19 pandemic on the agriculture industry was disclosed following these steps (Pan et al., 2020).

The use of social media in disaster-related studies was examined by Said et al. and Zhang et al. in 2019.The studies in the first review article examined were grouped under three main headings. The first title is disaster detection in texts from Disaster and related social media content, the second title is disaster-related visual content analysis in social media, and the third title is disaster detection in satellite images (Said et al., 2019).In the second review study, in which social media was examined for public information and disaster warnings, the social media platforms used in the first place were evaluated, and then the regional distribution of such studies on disasters was examined (Zhang et al., 2019).

Studies conducted in general terms without focusing on any specific disaster are also included in the literature. In a study conducted in 2019, a very complex approach was put forward in which multimedia and satellite images were evaluated together. The disasters that occurred were first questioned with the countries where they occurred, then content analysis was made using social media texts, pictures and videos with the help of crawlers. Coordinate information obtained through tweets and satellite images were matched and presented to the user (Ahmad et al., 2019).

There are studies on the use of social media in disaster management. By using social media platforms such as Facebook and Twitter, the process of revealing previously occurring disasters was carried out (Kankanamge, Yigitcanlar, & Goonetilleke, 2020).

It is observed that the use of social media, especially Twitter, increases remarkably when natural disasters occur. For this reason, NLP techniques have been used in the study on information extraction to be used in future disasters based on the sharings of previously experienced disasters. Sentiment analysis, topical modeling procedures were applied (Karimizarani et al., 2023).

In a study conducted to understand the dynamics of communication when natural disasters occur, calls made by people on the web and on the phone were evaluated according to gender, ethnicity, and educational status (Pourebrahim et al., 2019).

Mobile social media data is used for real-time disaster damage assessment model. Accordingly, among the data sets obtained from Weibo, the damage data is divided into two main headings as the effect of people and the effect of structures. While data such as loss and injury were used to affect people, words such as collapse were used for structures such as buildings, traffic and agriculture. Topical models were created with LDA. The DUTIR dictionary was used for sentiment analysis. This dictionary is mainly used in Chinese natural language processing. Each category is numbered according to emotional intensity (Shan et al., 2019).

During the disaster, classification and location estimation studies were carried out with the help of tweets. With the help of the Twitter API as a data set, the data was obtained for the floods that occurred in the southern and eastern regions of India. In the study, a total of 32400 tweets were examined, the Tweets were collected in English, Hindi and other regional languages. In the study, the event was evaluated under two main headings: low importance and high importance. According to the information of whether there is an address in the tweets of high importance, information is extracted for the rescue teams, and if not, it is determined whether the user has location information with the help of previous tweet records. For this, the Markov model was used and it was planned to be sent to the rescue teams when the location was determined (Singh et al., 2019).

It is also noteworthy that social media data is used in studies for a specific disaster such as flood, earthquake, storm.With the effective use of social media, many information, including disaster information, have begun to be shared by many people who are experts in the subject or not. Posts on Weibo on the flood disaster that occurred after heavy rains in Wuhan in 2016 were examined by Fang and his friends. Hotspot analyzes of the flood-affected areas in Wuhan, the largest metropolis of China, were revealed with the help of Weibo topics. Frequency analysis method was preferred as the method in the study (Fang et al., 2019).

Social media data can be used to decide the severity of the disaster. The tweets shared during the flood disaster in South East Queensland were analyzed. First of all, it was tried to understand the trend topics of the people with descriptive analysis. With content analysis, tweets have been made meaningful. With the help of spatial analyses, regions where the disaster is intense were determined with the help of geotagged tweets (Kankanamge, Yigitcanlar, Goonetilleke, et al., 2020).

In another remarkable study, sentiment analysis was conducted with the help of twitter data in the flood disaster in Jakarta (Saddam et al., 2023).

When the literature is examined, it is seen that the focus of attention of the studies is the studies on storms and other climatic disasters. The reason why this issue is specifically addressed is that climatic disasters are much more common, as stated in the introduction.

In another study in which Twitter was preferred, text and image content were evaluated together. For the study in which the Harvey, Maria and Irma hurricanes were evaluated, sentiment classification and clustering methods were preferred in the evaluation of the text content, and the most shared entity was revealed with topic modeling and Named Entity recognition. Especially in hurricanes, the preferred organizations for aid and the people who were mentioned the most were revealed. Finally, sentiment analysis, tweets and the levels at which people felt disasters were tagged on the map (Alam et al., 2020).

The power of social media can be used to evaluate the impact of the disaster on people after a disaster has occurred. In the study conducted in 2019, the shares of people affected and unaffected by the disaster after Hurricane Sandy were evaluated. Twitter data were analyzed on the topics of faith-based, community, assets and financial. In addition, the correlation between demographic, socioeconomic, spatial attributes and topics in the region was revealed with the Dirichlet regression model (Jamali et al., 2019).

In another study to extract useful and immediate information from social media content, Machine learning and Rule-based classification were used. In the study using Weibo, the heavy storm that occurred in Zhengzhou on July 20 was discussed. As a result of the natural disaster, the city was seriously flooded, floods occurred in the streams and landslides took place. In the disaster, where 380 people were reported dead or missing, the economic loss was quite high (Tchounwou, 2004).

Investigation of Hurricanes Sandy, Harvey and Irma with CNN model using real-time social media data. In the study, in which approximately 5000-15,000 tweets were examined for each hurricane, 5 main topics were examined: warning and advice, damage, information resources, aid and people (Yu et al., 2019).

Another study under the title of special disasters is about earthquakes. The texts, pictures, videos and links shared by Weibo users after the earthquake in Sichuan constitute the main subject of the study. First of all, time-dependent analyzes of the data were made and the temporal intensities of the shares related to the earthquake were revealed. In addition, maps showing the location distribution of the shared information were created. The frequencies of the words from the shared texts were revealed. In the texts shared for the earthquakes that occurred at 3 different times, the information about the earthquake and its results were revealed with sentiment analysis (He et al., 2023).

## 2. Method

The aim of the study is to evaluate the results of the web mining study conducted in two different languages, Turkish and English.

Data crawling, data pre-processing, topic modelling and sentiment Analysis were done in the study.

### 2.1. Study Area

In the study carried out on the scale of the world and Türkiye, web pages were crawled according to both scales.

### 2.2. Data Set

The investigation made use of Wikipedia data. Some of the word titles were derived from the most common natural disasters in the world and in Türkiye. As a result, two distinct content analyses were carried out, one in Turkish and one in English (Table 1).

**Table 1.** Query words

| Query (TR) | Query (EN) |
|---|---|
| Deprem | Earthquake |
| Afet | Disaster |
| Fay Hattı | Fault line |
| Yağmur | Rain |
| Fırtına | Storm |
| Dolu | Hail |
| Heyelan | Landslide |
| Toprak | Soil |
| Toprak Kayması | Mudslide |
| Kaya Düşmesi | Rock fall |
| Pandemi | Pandemic |
| Corona | Corona |
| Veba | Plague |
| Sel | Flood |

### 2.3. Method

#### 2.3.1. Data Crawling

Data crawling is a process of extracting data from various web sources. Crawling data is quite similar to what major search engines do. Data crawling is a technique for identifying web links and extracting information from them.

#### 2.3.2. Data Preprocessing

Transformation, Tokenization, filtering can be used in this stage.

Transformation: In this section, lower keys, parse HTML and remove URL operations are done.

Tokenization: At this stage, word punctuation, regular expression and extraction of data showing only word characteristics were performed.

Filtering: This stage consists of detecting and creating stop words in order to extract Turkish words, and filtering punctuation marks and remove numerical information if you don't need.

#### 2.3.3. Topic Modeling

Topic modeling is a statistical modeling technique that employs unsupervised Machine Learning to discover clusters or groupings of related terms within a body of text. This text mining method understands unstructured data by using semantic structures in text rather than predefined tags or training data. Latent Dirichlet Allocation (LDA) is a topic model that is used to assign text in a document to a certain topic.

## 3. Results

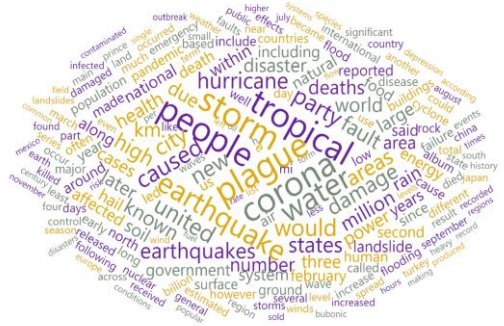
### 3.1. Preprocessing

Wikipedia was preferred in the study. 5 articles per query in Turkish Wikipedia and 3 articles in English Wikipedia were examined in order to make the process faster for this mini study.

In the study, both Turkish and English Wikipedia web pages were obtained as a result of the disaster words queries.

According to the search results created by using the query words in table, a total of 79 articles were found in Turkish searches and a total of 59 articles in English web sites. As a result of the pre-processing, word clouds were created. Frequency of word clouds in Turkish is 294, and 2419 in English (Figure 1).


a. Turkish Wikipedia Disaster Results

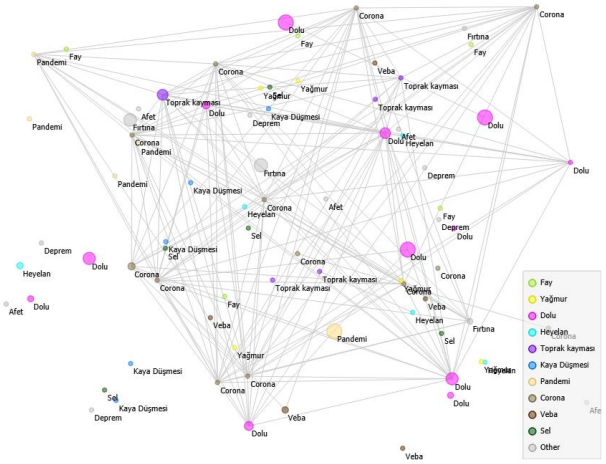



b. English Wikipedia Disaster Results

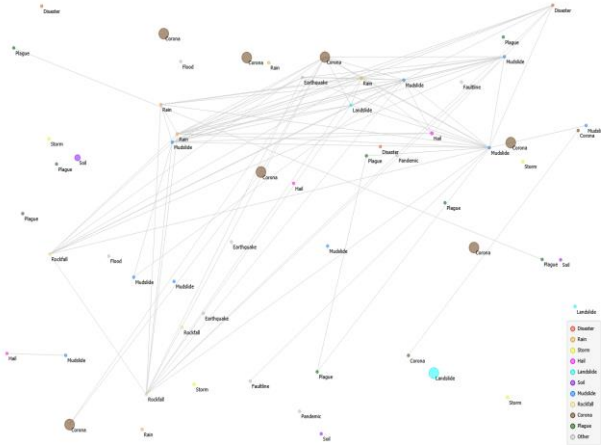
**Figure 1.** Word Cloud Results After preprocessing

### 3.2. Topic modeling Results

In addition, in the study, topic determination was made with the help of word frequencies according to the LDA (Latent Dirichlet allocation). LDA is a particularly popular method for fitting a topic model. 10 different topics have been determined (Figure 2).




a. Topic 2 and Query Results of Turkish Wikipedia



b. Topic 3 and Query Results of English Wikipedia

**Figure 2.** Topic modeling results

## 4. Conclusion

Since the study is only for Wikipedia data, it contains more academic information. Therefore, the results obtained should be improved, even if they are not worthless. In such studies, other platforms such as twitter and news, which show much more variability, may be preferred. However, supporting local data with this type of data can provide many advantages for both the government and local governments in the management of disasters.

## Acknowledgement

The open source software Orange Data Mining and free information resource Wikipedia were used by the author. Thank you to both supporters.

## References

Ahmad, K., Pogorelov, K., Riegler, M., Conci, N., & Halvorsen, P. (2019). Social media and satellites: Disaster event detection, linking and summarization. Multimedia Tools and Applications, 78(3), 2837–2875. https://doi.org/10.1007/s11042-018-5982-9

Alam, F., Ofli, F., & Imran, M. (2020). Descriptive and visual summaries of disaster events using artificial intelligence techniques: Case studies of Hurricanes Harvey, Irma, and Maria. Behaviour & Information Technology, 39(3), 288–318. https://doi.org/10.1080/0144929X.2019.1610908

Duan, H. K., Vasarhelyi, M. A., Codesso, M., & Alzamil, Z. (2023). Enhancing the government accounting information systems using social media information: An application of text mining and machine learning. International Journal of Accounting Information Systems, 48, 100600. https://doi.org/10.1016/j.accinf.2022.100600

Fang, J., Hu, J., Shi, X., & Zhao, L. (2019). Assessing disaster impacts and response using social media data in China: A case study of 2016 Wuhan rainstorm. International Journal of Disaster Risk Reduction, 34, 275–282. https://doi.org/10.1016/j.ijdrr.2018.11.027

G. Almatar, M., Alazmi, H. S., Li, L., & Fox, E. A. (2020). Applying GIS and Text Mining Methods to Twitter Data to Explore the Spatiotemporal Patterns of Topics of Interest in Kuwait. ISPRS International Journal of Geo-Information, 9(12), 702. https://doi.org/10.3390/ijgi9120702

Hameed, A. Z. (2023). A hybrid Fifth Generation based approaches on extracting and analyzing customer requirement through online mode in healthcare industry. Computers and Electrical Engineering, 106, 108550. https://doi.org/10.1016/j.compeleceng.2022.108550

Han, X., Wang, J., Zhang, M., & Wang, X. (2020). Using Social Media to Mine and Analyze Public Opinion Related to COVID-19 in China. International Journal of Environmental Research and Public Health, 17(8), 2788. https://doi.org/10.3390/ijerph17082788

He, W., Yuan, Q., & Li, N. (2023). Research on the Characteristics of Internet Public Opinion and Public Sentiment after the Sichuan Earthquake Based on the Perspective of Weibo. Applied Sciences, 13(3), 1335. https://doi.org/10.3390/app13031335

Issa, E., Tsou, M.-H., Nara, A., & Spitzberg, B. (2017). Understanding the spatio-temporal characteristics of Twitter data with geotagged and non-geotagged content: Two case studies with the topic of flu and Ted (movie). Annals of GIS, 23(3), 219–235. https://doi.org/10.1080/19475683.2017.1343257

Jamali, M., Nejat, A., Ghosh, S., Jin, F., & Cao, G. (2019). Social media data and post-disaster recovery. International Journal of Information Management, 44, 25–37. https://doi.org/10.1016/j.ijinfomgt.2018.09.005

Kankanamge, N., Yigitcanlar, T., & Goonetilleke, A. (2020). How engaging are disaster management related social media channels? The case of Australian state emergency organisations. International Journal of Disaster Risk Reduction, 48, 101571. https://doi.org/10.1016/j.ijdrr.2020.101571

Kankanamge, N., Yigitcanlar, T., Goonetilleke, A., & Kamruzzaman, Md. (2020). Determining disaster severity through social media analysis: Testing the

methodology with South East Queensland Flood tweets. International Journal of Disaster Risk Reduction, 42, 101360. https://doi.org/10.1016/j.ijdrr.2019.101360

Karimiziarani, M., Shao, W., Mirzaei, M., & Moradkhani, H. (2023). Toward reduction of detrimental effects of hurricanes using a social media data analytic Approach: How climate change is perceived? Climate Risk Management, 39, 100480. https://doi.org/10.1016/j.crm.2023.100480

Pan, D., Yang, J., Zhou, G., & Kong, F. (2020). The influence of COVID-19 on agricultural economy and emergency mitigation measures in China: A text mining analysis. PLOS ONE, 15(10), e0241167. https://doi.org/10.1371/journal.pone.0241167

Pourebrahim, N., Sultana, S., Edwards, J., Gochanour, A., & Mohanty, S. (2019). Understanding communication dynamics on Twitter during natural disasters: A case study of Hurricane Sandy. International Journal of Disaster Risk Reduction, 37, 101176. https://doi.org/10.1016/j.ijdrr.2019.101176

Saddam, M. A., Dewantara, E. K., & Solichin, A. (2023). Sentiment Analysis of Flood Disaster Management in Jakarta on Twitter Using Support Vector Machines. Sinkron, 8(1), 470–479. https://doi.org/10.33395/sinkron.v8i1.12063

Said, N., Ahmad, K., Riegler, M., Pogorelov, K., Hassan, L., Ahmad, N., & Conci, N. (2019). Natural disasters detection in social media and satellite imagery: A survey. Multimedia Tools and Applications, 78(22), 31267–31302. https://doi.org/10.1007/s11042-019-07942-1

Shan, S., Zhao, F., Wei, Y., & Liu, M. (2019). Disaster management 2.0: A real-time disaster damage assessment model based on mobile social media data—A case study of Weibo (Chinese Twitter). Safety Science, 115, 393–413. https://doi.org/10.1016/j.ssci.2019.02.029

Singh, J. P., Dwivedi, Y. K., Rana, N. P., Kumar, A., & Kapoor, K. K. (2019). Event classification and location prediction from tweets during disasters. Annals of Operations Research, 283(1–2), 737–757. https://doi.org/10.1007/s10479-017-2522-3

Statista (2023), (https://www.statista.com/statistics, Access Date: 10 June, 2023

Tchounwou, P. (2004). Environmental Research and Public Health. International Journal of Environmental Research and Public Health, 1(1), 1–2. https://doi.org/10.3390/ijerph2004010001

Woo, H., Cho, Y., Shim, E., Lee, K., & Song, G. (2015). Public Trauma after the Sewol Ferry Disaster: The Role of Social Media in Understanding the Public Mood. International Journal of Environmental Research and Public Health, 12(9), 10974–10983. https://doi.org/10.3390/ijerph120910974

Yu, M., Huang, Q., Qin, H., Scheele, C., & Yang, C. (2019). Deep learning for real-time social media text classification for situation awareness – using Hurricanes Sandy, Harvey, and Irma as case studies. International Journal of Digital Earth, 12(11), 1230–1247. https://doi.org/10.1080/17538947.2019.1574316

Zhang, C., Fan, C., Yao, W., Hu, X., & Mostafavi, A. (2019). Social media for intelligent public information and warning in disasters: An interdisciplinary review. International Journal of Information Management, 49, 190–207. https://doi.org/10.1016/j.ijinfomgt.2019.04.004