FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Faculty of Chemistry and
Earth Sciences

BACHELOR THESIS

AI-Based Classification of Disaster-Related Images: A Comparative Study of Models

Jakob Åke Glesmer
born 31.03.2003
in Ribnitz-Damgarten, Germany

first supervisor:

Prof. Dr. Alexander Brenning

second supervisor:
Dr.-Ing. Jens Kersten

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA
Faculty of Chemistry and Earth Sciences
Institute of Geography

Bachelor of Science

AI–Based Classification of Disaster-Related Images: A Comparative Study of Models

by Jakob Åke Glesmer

Student Number: 205162

Abstract

Natural disasters have become an increasingly severe threat. This study evaluates the reliability of image classification across multiple deep learning models in the context of natural disaster detection, using a subset of manually validated images derived from the GDELT dataset. Due to the absence of ground-truth labels and challenges in data quality, a smaller, curated sample was employed to ensure analytical validity. The evaluated models include EfficientNet-B1, ResNet-101, OpenCLIP, and CoCa. Results indicate that EfficientNet-B1 and ResNet-101—particularly when utilizing MEDIC's pretrained weights—achieved consistent and reliable performance, especially in distinguishing between disaster and non-disaster imagery. In contrast, OpenCLIP and CoCa exhibited lower classification accuracy, with CoCa performing weakest, primarily due to difficulties in interpreting abstract disaster categories and additional uncertainty introduced through semantic textual similarity. Identified sources of error include inconsistencies in image labeling, sampling biases, and ambiguities within both statistical and semantic evaluation procedures. Despite these limitations, the study highlights critical differences in model behavior and reliability, emphasizing the need for specialized fine-tuning when applying general-purpose vision-language models to disaster recognition tasks.

Contents

1 Introduction **1**

1.1 Motivation: The Need for Rapid Identification of Disasters 1

1.2 Background: Image Processing with Convolutional Neural Networks 2

1.3 Research Questions 3

2 Related Work **4**

2.1 Social Media–Based Disaster Information Processing 4

2.2 Disaster Information Processing Based on Other Sources 5

3 Data **6**

3.1 Source of Data 6

3.2 Data Acquisition and Preparation 6

3.3 Ground Truth Sampling 7

4 Models and Methods **10**

4.1 Resnet101 and Efficentnet-B1 10

4.2 OpenAIs Contrastive Language-Image Pre-training 12

4.3 Coca: Contrastive Captioners are Image-Text Foundation Models 12

5 Experiments and Results **13**

5.1 Experiments 13

5.1.1 Mean Reciprocal Rank 13

5.1.2 Accuracy, Precission, Recall and F1 Score 14

5.1.3 Konfusionmatrix 15

5.1.4 Mapping Coca captions with Semantic Textual Similarity 15

5.2 Results 16

5.2.1 MEDIC models Efficentnet-b1 and Resnet101 16

5.2.2 OpenClip 20

5.2.3 CoCa 24

6 Discussion **29**

6.1 Interpretaition of Results 30

6.1.1 Similarity Between Efficient Net B1 and ResNet-101 30

6.1.2 Difference in Klassifikation from CoCa and OpenCLIP 30

6.1.3 Comparison with Existing Research 32

6.2 Possible Sources of Error 34

6.2.1 Possible Errors Made in the Sampling Prozess 34

6.2.2 Possible errors Made in the Usage of the Models 35

6.2.3 Possible errors Made in the Statistical valuation of the Results . . . 35

6.2.4 Inclination trought difference in Training Data 35

6.2.5 Difficulty in compairing CoCa Results 36

7 Conclusion and Future Work **40**

7.1 Conclusion 40

7.2 Future Work 40

References **41**

List of Figures

1 Network architecture of the classic CNN LeNet-5 (Liu, 2018) 2
2 Flow Chart of All Operations 7
3 Image from the GDELT dataset excluded in samples with category `Earthquake` depicting a seismograph in a symbolic manner 9
4 Image from the GDELT dataset excluded in samples with category `Fire` depicting multiple fires in one image 9
5 A confusion matrix showing the ground truth classes versus the classes predicted by Efficient-Net B1 18
6 A confusion matrix showing the ground truth classes versus the classes predicted by ResNet 101 20
7 OpenCLIP heatmap of rank frequencies per category. Darker colors indicate higher frequencies. 23
8 OpenCLIP heatmap of rank frequencies per category. Darker colors indicate higher frequencies. 23
9 A confusion matrix showing the ground truth classes versus the classes predicted by OpenCLIP 24
10 Heatmap of rank frequencies per category. Darker colors indicate higher frequencies. 28
11 Heatmap of rank frequencies per category. Red colors indicate higher frequencies. 28
12 A confusion matrix showing the ground truth classes versus the classes predicted by CoCa 29
13 Image form the GDELT Data included in Samples depicting a rather green drought faced landscape 34
14 Image form the GDELT Data included in Samples with Category `No Disaster` who was Classified as `Fire` 36
15 One of 5 Images from the GDELT Data included in Samples with Category `No Disaster` who was Classified as `No Disaster` 37
16 Image form the GDELT Data included in Samples with Category Blizzard who was Classified as Landslide 37
17 Image form the GDELT Data included in Samples with Category `Hurricane` who was Classified as `Landslides` 38
18 Only Image form the GDELT Data included in Samples with Category `Hurricane` who was Classified as `Hurricane` 39

List of Abbreviations

ANN	**A**rtificial **N**eural **N**etwork
CNN	**C**onvolutional **N**eural **N**etwork
GDELT	**G**lobal **D**atabase of **E**vents, **L**anguage, and **T**one
CoCa	**Co**ntrastive **Ca**ptioner
OpenCLIP	**Open**source **C**ontrastive **L**anguage-**I**mage **P**re-training
EFF	**Eff**icent-Net B1
RES	**Res**Net 101
STS	**S**emantic **T**extual **S**imitlarity

Fonts

`This font indicates the mention of a class.`

1 Introduction

This introduction outlines the necessity of disaster detection and classification, as well as the use of social media and news data for these purposes. It establishes the motivation to ensure the reliability and security of models applied in this context. Subsequently, the background of the study concerning image recognition is presented, concluding with the formulation of the research question.

1.1 Motivation: The Need for Rapid Identification of Disasters

On July 14, 2021, a severe flood disaster transformed the Ahr Valley in Rhineland-Palatinate and North Rhine-Westphalia into a landscape of destruction within just a few hours. Entire settlements were devastated, hundreds of people lost their lives, and tens of thousands were displaced. Similarly, recurrent large-scale wildfires in California, particularly in the Los Angeles area, demonstrate how rapidly natural hazards can escalate, threatening human lives, destroying infrastructure, and causing long-term environmental damage. This trend is further exacerbated by climate change.

Natural disasters such as floods, earthquakes, storms, and wildfires pose a constant and severe threat to human life, infrastructure, and the environment. In today's interconnected digital landscape, images of such events spread within minutes across social media, news websites, and other online platforms. News organizations, in particular, provide highly visual coverage of disasters due to the significant public attention these events attract.

This abundance of publicly shared imagery represents a valuable data source for the automated detection and assessment of natural disasters. In the critical hours following an event, rapid and reliable information about the scale, nature, and geographic location of the disaster is essential. Modern computer vision models, trained specifically on images from social media and online news sources, offer the potential to automate and accelerate this process substantially.

The use of images as an information source can provide an essential complement to established methods that primarily rely on text and other web-based data. Visual data capture contextual and situational details that textual or metadata-based approaches may overlook, thereby enriching the overall analytical framework and contributing to a more comprehensive understanding of events and phenomena.

These methods can facilitate the rapid identification of severely affected areas, enhance situational awareness, and support the targeted coordination of emergency response operations. In doing so, they hold the potential to reduce both the material impact of disasters and the extent of human suffering.

1.2 Background: Image Processing with Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a specialized form of Artificial Neural Networks (ANNs) developed for the recognition and processing of image data (Liu, 2018). Inspired by the structure of the visual pathway in animals, artificial neurons in a CNN do not respond to the entire input image but only to a small local region known as the receptive field (Liu, 2018). This localized connectivity gives rise to convolutional layers, where filters (or convolutional kernels) are applied across the image to detect simple features like edges or textures, generating feature maps that indicate where these features occur (Liu, 2018).

Following convolutional layers, pooling layers reduce the dimensionality of the feature maps, enhancing robustness to small translations, scalings, or rotations (Liu, 2018). Techniques like max-pooling and average-pooling are commonly used. Activation functions, such as the Rectified Linear Unit (ReLU), introduce non-linearity, allowing the network to model complex relationships (Liu, 2018). At the end of a CNN, fully connected layers integrate the extracted features for final classification, producing an output vector representing the probabilities of belonging to different classes (Liu, 2018).

Compared to traditional ANNs, which use fully connected layers, CNNs significantly reduce the number of parameters through local connections and weight sharing (Ankith and Akshaya, 2021; Liu, 2018). This allows CNNs to automatically extract relevant image features without requiring manual feature engineering. A prominent example is LeNet-5, a classic CNN developed for handwritten digit recognition on the MNIST dataset (Liu, 2018). LeNet-5 consists of convolutional layers, subsampling (pooling) layers, and fully connected layers, taking as input 224×224 grayscale images and ultimately classifying digits using a Radial Basis Function (RBF) (Liu, 2018).

In summary, CNNs—through their specific structure of convolution, pooling, activation,

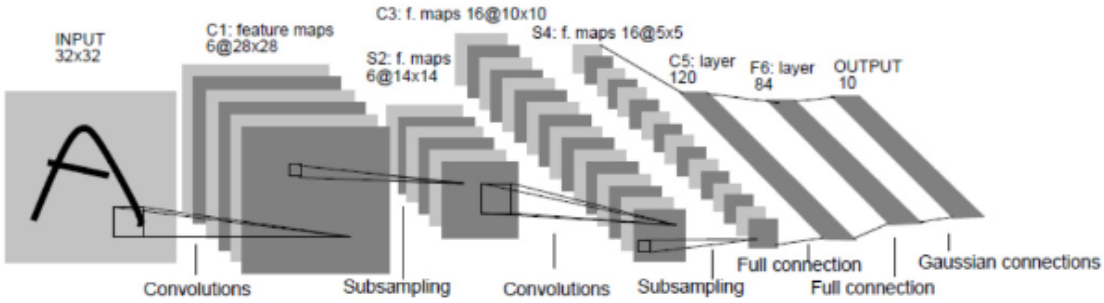


Figure 1 Network architecture of the classic CNN LeNet-5 (Liu, 2018)

and classification—offer an efficient solution for image classification tasks. Compared to their ANN predecessors, CNNs substantially reduce the number of parameters and enable automatic, hierarchical feature extraction (Ankith and Akshaya, 2021; Liu, 2018; Hamadain et al., 2023; Goodfellow et al., 2016).

The convolution operation itself enables hierarchical feature representation, where lower layers detect simple structures such as edges, while higher layers capture more complex shapes and objects (Liu, 2018; Hamadain et al., 2023).

1.3 Research Questions

The objective of this thesis is to explore the potential of such models for automated disaster detection based on social media and news imagery. The focus lies on three specific models—MEDIC, CoCa, and OpenCLIP—which classify or describe images in terms of different disaster types (Alam et al., 2023). The central research question is: To what extent can the outputs of these models be considered reliable and applicable in operational disaster response scenarios?

2 Related Work

The following section presents related work, encompassing studies that utilize both social media as a data source and traditional sources such as sensors, satellite imagery, and aerial photography. These works address both the filtering and classification of such data, aiming to enable real-time disaster detection through machine learning and deep learning techniques.

2.1 Social Media–Based Disaster Information Processing

Comprehensive surveys in crisis informatics, such as those by (Imran et al., 2018) and (Imran et al., 2015), provide a broad overview of computational strategies for leveraging social media in disaster management. These studies examine core methodologies including filtering, clustering, summarization, semantic enrichment, and credibility analysis, while emphasizing persistent challenges such as real-time adaptation, misinformation detection, and multimodal data integration. Their findings underscore that effective disaster-response systems depend on robust pipelines that can accurately filter, classify, and verify social media content before transforming it into actionable intelligence for emergency decision-makers.

Building upon these survey insights, the following works exemplify focused implementations that operationalize the principles outlined above, demonstrating practical applications of social media data processing for disaster management.

A considerable body of research explores the use of social media for disaster management. (Sangameswar et al., 2017) demonstrates that Twitter streams processed in R-Studio can serve as effective sources for detecting and locating disasters. Inputs were live Twitter messages, analyzed via geoparsing and sentiment analysis, with outputs providing real-time maps of disaster locations and public sentiment. This work underscores the value of social media for rapid situational awareness.

(Nguyen, Alam, et al., 2017) focuses on managing visual social media content, introducing a system that filters irrelevant and duplicate images using deep learning–based relevancy detection combined with perceptual hashing (pHash). The system processes raw Twitter images and outputs a reduced, high-quality dataset, improving annotation efficiency and downstream analysis.

(Colombo, 2018), within the E2mC project, develops a machine learning–based filter to classify Twitter posts as relevant or irrelevant for Civil Protection agencies during earthquakes, floods, and fires. The system uses Gaussian Naive Bayes and Random Forest classifiers, taking text and metadata as input and producing labeled outputs. The significance and correctness of the classification were analyzed using accuracy, precision, recall, and F1 score, confirming the model's reliability for operational use.

Together, these studies illustrate how social media data can be processed, filtered, and classified to generate actionable insights for emergency responders, emphasizing the critical role of relevance detection, deduplication, and rigorous evaluation in creating effective disaster-response pipelines.

2.2 Disaster Information Processing Based on Other Sources

Social Media as data is powerful but the traditional sources such as sensor networks, satellite imagery, and aerial data, have continuously adapted to emerging developments and, therefore,remain relevant today. (LeoNila et al., 2024;Toan et al., 2019 and Pi et al., 2020) are leveraging various forms of non-social media data to enhance detection accuracy, reduce false alarms, and support timely emergency responses.

(LeoNila et al., 2024) present a comprehensive forest fire detection framework integrating smart sensor networks, computer vision, and deep learning models such as Artificial Neural Networks and CNN's. The system processes real-time environmental sensor data and visual inputs to identify fire events with high precision. Experimental evaluations demonstrated impressive results, achieving detection accuracies of 95% with CNNs and 89% with ANNs, along with a low false alarm rate. The research underscores the effectiveness of hybrid sensor–vision systems in diverse forestry conditions and suggests future improvements in energy efficiency, scalability, and adaptability through reinforcement learning techniques.

In a complementary direction, (Toan et al., 2019) propose a satellite-based wildfire detection framework utilizing deep learning to analyze high-resolution imagery for early-stage fire identification. Their model outperforms traditional approaches with a reported 94% F1-score and achieves detection speeds 1.5 times faster than existing systems. Moreover, the authors introduce a visualization dashboard that supports real-time alerting and monitoring, effectively addressing challenges such as cloud cover and nighttime observation. This satellite imagery approach significantly enhances situational awareness and early warning capabilities in large-scale wildfire management.

Extending the application of deep learning to post-disaster analysis, (Pi et al., 2020) investigate the use of CNNs for detecting ground objects of interest (GOIs) in aerial videos collected by drones and helicopters after natural disasters. Their findings emphasize the impact of data acquisition parameters on model performance, particularly showing that CNNs achieve superior results when trained and tested on data captured from similar altitudes. The study further demonstrates that balanced datasets and pre-training with VOC dataset weights substantially improve detection accuracy. These insights highlight the importance of dataset design, pre-training strategy, and viewpoint consistency in aerial-based disaster assessment.

3 Data

This section describes the datasets used in this study. It begins by outlining their sources, structure, and content, followed by explanations of how the data were pre-filtered, selected, obtained, prepared, and, in some cases, transformed. The necessity of data sampling for this study is also discussed, including a detailed description of how the sampling was performed and the criteria applied.

3.1 Source of Data

The source of the images used in this study is the Global Database of Events, Language, and Tone (GDELT) (Leetaru, 2013–2022b). GDELT is a news aggregator database that collects and classifies online news articles, including the URLs of the original sources. The DLR submitted a query to the multilingual GDELT database for the period from January 1 to January 8, 2025. As a result, information on online news articles containing the GDELT GKG theme `NATURAL_DISASTER` was extracted.

Using this approach, metadata from approximately 85,000 news articles were retrieved and stored by the DLR in a MongoDB database. The HTML code of the extracted websites was then archived in Web ARChive (WARC) format and stored in an object store designed to handle large volumes of unstructured data. Subsequently, the images referenced in the source code were requested. To reduce the total image volume, a filter was applied to ensure that only images with a resolution of 150×150 pixels or higher were retained.

3.2 Data Acquisition and Preparation

The data were provided by the DLR and stored in a MongoDB database. This data included the metadata received from GDELT. Important keys included `objectstore`, containing the path to the images in the object store, and `V2ENHANCEDTHEMES`, which contained a list of GKG themes related to keywords extracted from the text and organized into themes, effectively serving as categories describing the text (Leetaru, 2013–2022a).

In the MinIO object store, the images were stored in JPG, PNG, and JPEG formats. The dataset initially contained approximately 470,000 images, which needed to be prefiltered. Filtering criteria were based on the metadata in MongoDB. A second collection was created to contain only the filtered metadata. From the GDELT collection, the structure was simplified to retain fewer keys, and the values were made more easily accessible. Entries without image links or without `V2ENHANCEDTHEMES` values indicating disasters were removed.

Once filtering was complete, images could be downloaded from the object store by matching their entries in MongoDB. Only images from articles, websites, or reports that contained references to natural disasters in context were downloaded. The images were organized into seven folders, each containing up to 50,000 images, resulting in a total of 329,177 images for further use. Although a significant portion of irrelevant images remained, the

dataset was intentionally reduced for the intended use.

The images are mostly in JPG format and depict a wide variety of content, with only a few actually representing disaster events. As a result, the dataset contains a substantial proportion of irrelevant data, which makes it highly realistic. This forces the models to cope with heterogeneity and a large proportion of non-relevant images, thereby testing their robustness under realistic conditions.

As described in the chapters on the individual models, image transformations were applied prior to model processing. These included resizing images to 224×224 pixels and normalizing them based on ImageNet standards to ensure compatibility with the models. The implementations of these transformations were provided for each model except CoCa (Kersten, 2022; Andersson, 2025). The CoCa transformation process includes resizing images to 288×288 pixels and applying normalization. For this type of transformation, CoCa includes both an image encoder component and a text encoder component (Yu et al., 2022).

3.3 Ground Truth Sampling

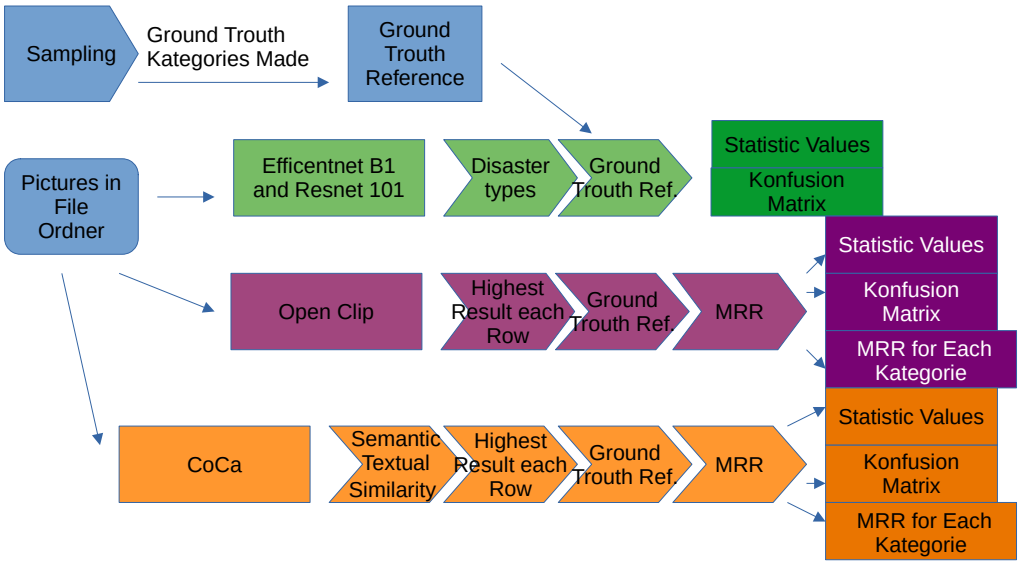


Figure 2 Flow Chart of All Operations

The image dataset, consisting of 329,177 images, had already been substantially reduced from approximately 470,000 images during preprocessing. The reduction primarily relied on the values of the key `V2ENHANCEDTHEMES`. These themes are keywords present in the texts of the websites or articles from which the images were scraped. Images from the

same article or website were assigned all the themes present in that article or website.

At first glance, this may suggest a strong correlation between images and themes; however, this can be misleading. For example, if a website covers multiple topics and includes images related to very different subjects, all images will nonetheless be associated with all themes from that website, regardless of their individual content. If one of the 72 themes relevant to this study (specifically those describing disasters) is present, all images associated with that theme are included, even if most of the images do not accurately reflect that theme.

Evidence of a substantial number of such false-positive images is apparent: after the prior selection process, the themes should retain strong descriptive power. However, a large proportion of the 329,177 images do not depict disasters, demonstrating that the themes are often inaccurate and consequently lack meaningful descriptive validity.

Since the primary objective of this study is to evaluate the reliability of the models, a high-confidence dataset is required. A reference set of hand-annotated results must exist against which the model classifications can be measured; in other words, ground truth data are necessary. To obtain these data, images must be sampled. Manual validation of these images involves substantial effort, which is limited for this study. Therefore, the number of sampled images must be sufficiently representative while remaining manageable in size.

The procedure for sampling images to create the ground truth dataset was as follows: initially, 160 images that, upon visual inspection, did not depict any disasters were collected into a folder representing the `No Disaster` category. Subsequently, 20 images were collected for the `Blizzard` category and stored in a separate folder. Blizzard images depicted snow, heavy snowfall, or dense snow cover. The same procedure was applied for the categories `Drought, Earthquake, Fire, Flood, Heavy Rains, Hurricane,` and `Landslides`.

For `Drought`, images showed drought conditions, such as dried vegetation, cracked soil, or desert-like landscapes. `Earthquake` images primarily showed destruction of stone buildings without fire damage or bullet holes, to avoid alternative causes of damage. `Fire` images depicted fire or destruction caused by fire, often including nighttime scenes with characteristic orange glows or showing firefighters in action. `Flood` images showed elevated water levels, submerged streets, or flood-related destruction. The `Heavy Rains` category included rainfall without flooding, often showing people under umbrellas, making these images more challenging for classification. `Hurricane` images depicted wind-related damage, such as high waves, tornadoes, fallen trees, or damaged buildings, ensuring damage involved light materials rather than heavy structural destruction. `Landslides` included mass movements, falling rocks, or large stones lying on roads.

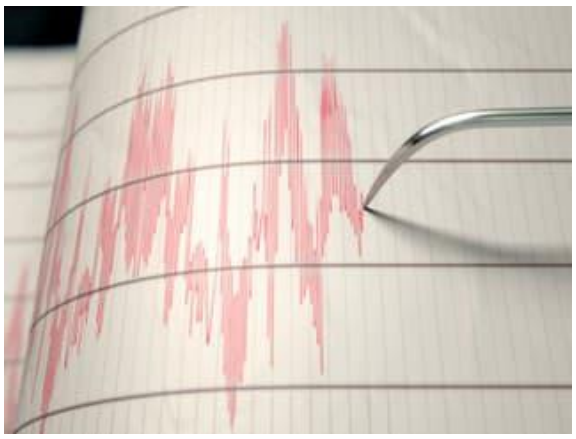


Figure 3 Image from the GDELT dataset excluded in samples with category `Earthquake` depicting a seismograph in a symbolic manner

Symbolic representations of disasters, such as graphs from monitoring devices, were excluded. Obvious manipulations or AI-generated images were also not included. Only images containing a single scene were selected, and duplicates or near-duplicates were removed. Images where the disaster appeared in the background with people or other elements in the foreground were also excluded.

Figure 4 Image from the GDELT dataset excluded in samples with category `Fire` depicting multiple fires in one image

After collection, all images from the individual folders were copied into a single folder from which the models would later draw the images for classification. For subsequent statistical evaluation using ground truth data, a list was created assigning the correct category to each sampled image.

4 Models and Methods

The following section describes the models used for image classification and provides a brief explanation of each. The classification reliability of these models is examined in detail throughout this work.

4.1 Resnet101 and Efficentnet-B1

EfficientNet-B1 (EFF) is a CNN belonging to a family of models ranging from B0 to B7, which differ primarily in the size of the images they can process (Tan and Le, 2019). The main objective of EfficientNet is to achieve high accuracy with a reduced number of parameters, making the model particularly suitable for mobile and resource-constrained applications(Tan and Le, 2019). EfficientNet-B1 takes images of 224×224 pixels as input. To ensure compatibility with this requirement, all images are resized and subsequently normalized before being processed by the model (Tan and Le, 2019). EfficientNet-B1 was pre trained using the Dataset ImageNet.

The other deep convolutional neural network used in this study is ResNet101 (RES). RES is part of the Residual Network (ResNet) family—variations of the same architecture—where the number 101 denotes the total number of layers used (He et al., 2016). RES was designed to overcome optimization challenges in deep networks by introducing residual connections, which help mitigate the problem of vanishing gradients (He et al., 2016).

The model can be applied to a wide range of tasks, including image recognition, feature extraction, image segmentation, and object detection. Similar to EFF, RES uses input images of 224×224 pixels (He et al., 2016). Accordingly, all images are resized to this resolution and normalized prior to processing. RES as well was pre trained using the Dataset ImageNet.

Since both EFF and RES require extensive image transformation and normalization, the MEDIC Dataset was employed in this study to handle preprocessing and model execution.

MEDIC is a multi-task learning dataset for disaster image classification. models like EFF and RES are used to set as input Images and classifies them according to several fixed features.(Alam et al., 2023)

The first feature is `Informativeness`, with the possible categories `Informative` and `Not Informative`. This feature assesses the extent to which an image contributes additional value to the overall informational content. Importantly, the classification is not limited to images depicting destruction; it also includes images showing, for example, the transport of relief goods or the support provided by volunteers. Broadly speaking, the label `Informative` functions as an overarching category, indicating that all images assigned to any of the subsequent categories are considered informative, as they contribute meaningfully to the classification task.

The second feature is `Damage Severity`, which is divided into the categories `Little or none`, `Mild`, and `Severe`. As the name suggests, this feature captures the degree of destruction visible in the image.

The third feature is `Humanitarian`, with the categories `Affected, injured, or dead people`, `Infrastructure and utility damage`, `Not Humanitarian`, and `Rescue, volunteering, or donation effort`. This feature is designed to classify images according to specific disaster-related aspects, all of which are directly connected to humanitarian concerns.

Finally, the feature `Disaster Types` includes the categories `earthquake`, `fire`, `hurricane`, `landslide`, `not disaster`, and `Other disaster`. This feature assigns the image to a particular type of natural disaster. Over the course of this work, this feature has become a central focus, as it demonstrated the greatest overlap with other models, proved to be the most straightforward to classify, and, in light of the GDELT themes provided by the database, offered the highest relevance for the classification task.

MEDIC can be viewed as the source of the data on which the models, such as EfficientNet (EFF) and ResNet (RES), were few-shot prompted for classification. According to CrisisNLP, 2025, MEDIC is the largest multi-task learning dataset related to disasters and represents an extended version of the Crisis Image Benchmark dataset. As stated in CrisisNLP, 2025, *"It consists of data from several sources such as CrisisMMD, data from AIDR, and the Damage Multimodal Dataset (DMD)."*. Consequently, the dataset contains 71,198 images. These images constitute the data on which the models were trained.

For the implementation of both the RES and EFF models, the MEDIC Dataset was utilized. This Dataset was obtained from (Kersten, 2022) whose work relies on (Alam et al., 2023,Alam et al., 2020,Alam et al., 2018,CrisisNLP, 2025, Mozannar et al., 2018) and (Nguyen, Ofli, et al., 2017).

Before the classification process, model-specific transformations were applied to the images using the MEDIC dataset to ensure compatibility with the models. The transformations included resizing the images to 224×224 pixels and normalizing them using the ImageNet mean and standard deviation. Both models, ResNet (RES) and EfficientNet (EFF), were initially trained on ImageNet and subsequently few-shot prompted for disaster detection using MEDIC. Pretrained weights supplied by MEDIC were employed, specifically those optimized for a multi-task learning setting. The classification task was configured accordingly as a multi-task problem, allowing the models to output multiple distinct labels Alam et al., 2023.

Although, in the subsequent analysis, only the results for the disaster types label were considered, the decision was made to retain the multi-task configuration. This choice reflects the fact that the multi-task setting constitutes the most widely adopted and comprehensive configuration within the MEDIC framework and is therefore expected to be

the most relevant for future applications.

4.2 OpenAIs Contrastive Language-Image Pre-training

OpenCLIP is an Open Source version of OPEN AIs CLIP (Contrastive Language-Image Pre-training)(Ilharco et al., 2021a) model which takes terms as input and distributes a total weighting of 1 across the provided terms, depending on how well each term corresponds to a given image (Ilharco et al., 2021a).

To utilize OpenCLIP, the framework provided by (Ilharco et al., 2021a) was employed but overworked with parts from (Hugging Face, 2025) and (Andersson, 2025). A connection to a folder containing the images was established for processing. Initially, relevant terms were selected, including the category labels from MEDIC, as well as a set of key terms extracted from the image sources in GDELT. Those category's were: `No Disaster`; `Blizzard`; `Drought`; `Earthquake`; `Fire`; `Flood`; `Heavy Rains`; `Hurricane`; `Landslides`. Classification was performed using pretrained weights.

During the OpenCLIP classification process, the images were transformed to 224×224 pixels and normalized using the ImageNet mean and standard deviation, similar to the preprocessing applied for EFF and RES. The classification results were exported as a CSV file. In this file, each image is associated with weights for the respective terms, ranging from 0 to 1, with all weights summing to 1. This allows the results to be interpreted as the proportional relevance of each term to the corresponding image. For subsequent data processing, the category with the highest weight for each image was selected.

4.3 Coca: Contrastive Captioners are Image-Text Foundation Models

CoCa also called Contrastive Captioner is according to (Yu et al., 2022) a image-text foundation model designed to combine the strengths of both contrastive learning (like CLIP) and generative methods (like SimVLM). It is an encoder-decoder transformer, but with a unique decoder structure: the first half of the decoder layers process text without looking at the image, while the remaining layers combine image and text information. (Yu et al., 2022)

CoCa also generates captions that correspond to the images (Yu et al., 2022), typically in the form of sentences of around twelve words in length, although variations in length do occur. In the work of (Yu et al., 2022), CoCa promised a very high accuracy in creating captions. In this work, this component of CoCa was evaluated with respect to its accuracy. The fact that the captions produced for the images differed substantially from the outputs of the other models under investigation proved to be a particular challenge. For the use of CoCa, the code provided by (Ilharco et al., 2021b) was used and to build an Implementation. While using the hole Data for Time issues , the code provided from (Raychanan, 2024) was also been used for the Implementation. Similar to the other models under investigation, CoCa processed images by linking to a folder containing the image files. The results produced by CoCa were exported in CSV format as a table.

5 Experiments and Results

This chapter explains the experiments conducted, which involved obtaining statistical measures from the model classifications. The resulting outcomes of these experiments are then presented, accompanied by initial observations.

5.1 Experiments

For the classification results, accuracy, precision, recall, and F1-score were calculated for all models. In addition, a confusion matrix was generated for each model to visualize the classification performance. The Mean Reciprocal Rank (MRR) was computed for the classification results of both CoCa and OpenCLIP. For CoCa, semantic textual similarity was applied to bridge the output format and enable comparability with OpenCLIP. The following section explains these processes in detail.

5.1.1 Mean Reciprocal Rank

The Reciprocal Rank (RR) is defined as

$$RR = \frac{1}{\text{rank of the first element}}$$

The Mean Reciprocal Rank (MRR) is a quality metric commonly used to evaluate ranking and recommendation systems. It measures how effectively a system places relevant results at the top of a sorted list.(Craswell, 2009)

The underlying idea is straightforward: for each query or user, the system produces an ordered list of results—such as products, songs, or categories. Among these, a particular "correct" result is defined as the ground truth. To assess ranking quality, the Reciprocal Rank (RR) is first calculated. This is defined as the inverse of the position at which the first relevant result appears in the ranked list. If the correct answer is ranked first, the RR equals 1. If it appears at rank 2, the RR is 0.5; at rank 3, approximately 0.33. If no relevant element is present within the considered top-K results, the RR equals 0.(Turnbull, 2021),(Craswell, 2009)

The Mean Reciprocal Rank is then the average of these values across all queries or users:

$$MRR = \frac{1}{U} \sum_{u=1}^{U} \frac{1}{\text{rank of the first relevant element for user } u}$$

It thus indicates, on average, how quickly the correct result is retrieved in the rankings.

5.1.2 Accuracy, Precission, Recall and F1 Score

In the evaluation of classification models in machine learning, it is essential to assess how well a model distinguishes between different classes. Several quality metrics are used for this purpose, each highlighting a different aspect of model performance. The most important among these are Accuracy, Precision, Recall, and the derived F1-score. (Klein and Mitchinson, 2023)

Accuracy measures how often a classification model makes a correct prediction overall. It is calculated as the ratio of correct predictions to the total number of predictions. A value of 1 (or 100%) indicates perfect accuracy, meaning every prediction was correct. Accuracy is easy to interpret and works well for balanced datasets in which all classes occur with similar frequency. However, this metric becomes problematic in the case of imbalanced datasets. In such cases, a model may achieve high accuracy while barely recognizing the minority class, which is often the more important one. (Klein and Mitchinson, 2023),(Sokolova et al., 2006)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures how often the model is correct when it predicts the positive class. It answers the question: How reliable are the positive predictions? High precision means that when the model predicts "positive," the prediction is usually correct. This metric is especially useful when false positives are costly or harmful. (Klein and Mitchinson, 2023),(Sokolova et al., 2006),(Bynke, 2022)

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall, also referred to as sensitivity, measures how well the model identifies all actual positive cases. It answers the question: How many of the truly positive examples are detected by the model? High recall means that very few positive cases are missed. This metric is critical when false negatives carry severe consequences. (Klein and Mitchinson, 2023),(Sokolova et al., 2006),(Bynke, 2022)

$$\text{Recall} = \frac{TP}{TP + FN}$$

The F1-score combines precision and recall into a single metric, defined as the harmonic mean of the two. A high F1-score can only be achieved if both precision and recall are strong simultaneously. This measure is particularly valuable for imbalanced datasets, as it balances the trade-off between false positives and false negatives. (Klein and Mitchinson, 2023),(Sokolova et al., 2006)

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

5.1.3 Konfusionmatrix

A confusion matrix is a table used to compare the predictions of a classification model with the actual classes (ground truth). The rows represent the true categories, while the columns correspond to the categories predicted by the model. This allows for a quick assessment of which instances were correctly classified and where the model made errors.(Susmaga, 2004)

5.1.4 Mapping Coca captions with Semantic Textual Similarity

To enable comparison with the outputs of the other models, a Semantic Textual Similarity from (SBERT.net, 2025), created from people like (Reimers and Gurevych, 2019), was used, which takes words as input and measures their similarity with the captions generated by CoCa. Semantic Textual Similarity from (SBERT.net, 2025) uses Aarsen, 2025 wich pre-trained the version from Reimers, 2025 what is just a 6 Layered version of the model original issued by microsoft.

For this analysis using semantic textual similarity, the same terms were used as those previously applied in OpenCLIP as well as the Classification was performed using pre-trained weights here as well.

$$\text{cosine_similarity}(\mathbf{A}, \mathbf{B}) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Cosine similarity is computed as part of the semantic textual similarity process. It measures the cosine of the angle between two vectors of arbitrary dimensionality. A value of –1 indicates that the vectors point in exactly opposite directions, while a value of 0 signifies that they are orthogonal (i.e., perpendicular) to each other. A value of 1 means that the vectors are perfectly aligned and point in the same direction. Later is here as well for subsequent data processing; the category with the highest weight for each image then determined.(Bynke, 2022)

5.2 Results

The classification results of the models are presented here in the form of statistical evaluations, including tables and matrices. Observations and interpretations of these statistical analyses and their outcomes are also discussed.

5.2.1 MEDIC models Efficentnet-b1 and Resnet101

The EfficientNet-B1 model was evaluated on the Medic dataset, achieving an overall precision of 0.604, a recall of 0.741, and an F1-score of 0.658 across 320 samples. When considering only the categories actually represented in the predictions, the model attains a category-level precision of 0.745, a recall of 0.912, and an F1-score of 0.845, indicating strong performance for the included categories.

At the category level, performance varies notably. High-performing categories include `Fire` (F1 = 0.86), Landslide (F1 = 0.89), `Earthquake` (F1 = 0.83), and `No Disaster` (F1 = 0.85), all of which demonstrate both high precision and recall. These results suggest that the model reliably identifies these events while minimizing false positives. `Flood` (F1 = 0.68) represents a moderately performing category, exhibiting solid recall but slightly lower precision, indicative of some misclassifications. Categories not present in the evaluated subset—`Blizzard`, `Drought`, and `Heavy Rains`—naturally exhibit precision, recall, and F1-scores of 0 because there obviously wasn´t any Classifikation for them with this model.

In the confusion matrix, the Y-axis represents the ground-truth classes, while the X-axis shows the predicted classes. This allows a comparison between the true category of each image and the number of images assigned to each predicted class. In the case of perfect classification, all images would lie along the diagonal from the top left to the bottom right, with no off-diagonal entries, indicating that every image was correctly classified into its ground-truth category.

For `Earthquake` and `No Disaster`, classification performance was near-perfect. All 20 `Earthquake` images were correctly identified, and 158 out of 160 `No Disaster` images were correctly classified. In contrast, the `Hurricane` category performed poorly, with only 7 out of 20 images correctly identified. Misclassifications were broadly distributed across other classes.

While the model demonstrates strong classification confidence for `No Disaster`, this performance comes at a cost. Nine images from four different categories were incorrectly classified as `No Disaster`, indicating that EfficientNet-B1, although highly accurate in recognizing `No Disaster` images, tends to misassign images from other classes to this category. In other words, the high recognition rate for `No Disaster` partially occurs at the expense of the classification accuracy of other categories.

Overall, these results indicate that EfficientNet-B1 performs robustly on the Medic dataset for categories that are included and frequent or visually distinct, while its performance is inherently limited for categories absent from the subset. The model demonstrates particularly strong predictive capabilities for `Earthquake` and `No Disaster` events. The model's strong performance in accurately identifying `No Disaster` images substantially contributed to its high overall classification accuracy. But `No Disaster` does also Tend to over classify.

Table 1 Category-wise classification metrics for the EfficientNet-B1 model on the Medic dataset. `Blizzard`, `Drought` and `Heavy Rains` were not part of MEDIC Categories and are therefore not included

Category	Precision	Recall	F1-Score	Support
Earthquake	0.71	1.00	0.83	20
Fire	0.82	0.90	0.86	20
Flood	0.57	0.85	0.68	20
Hurricane	0.64	0.35	0.45	20
Landslide	0.94	0.85	0.89	20
No Disaster	0.75	0.99	0.85	160
Overall / Weighted	0.74	0.91	0.84	260
Overall / Weighted Ex.	0.60	0.74	0.66	320

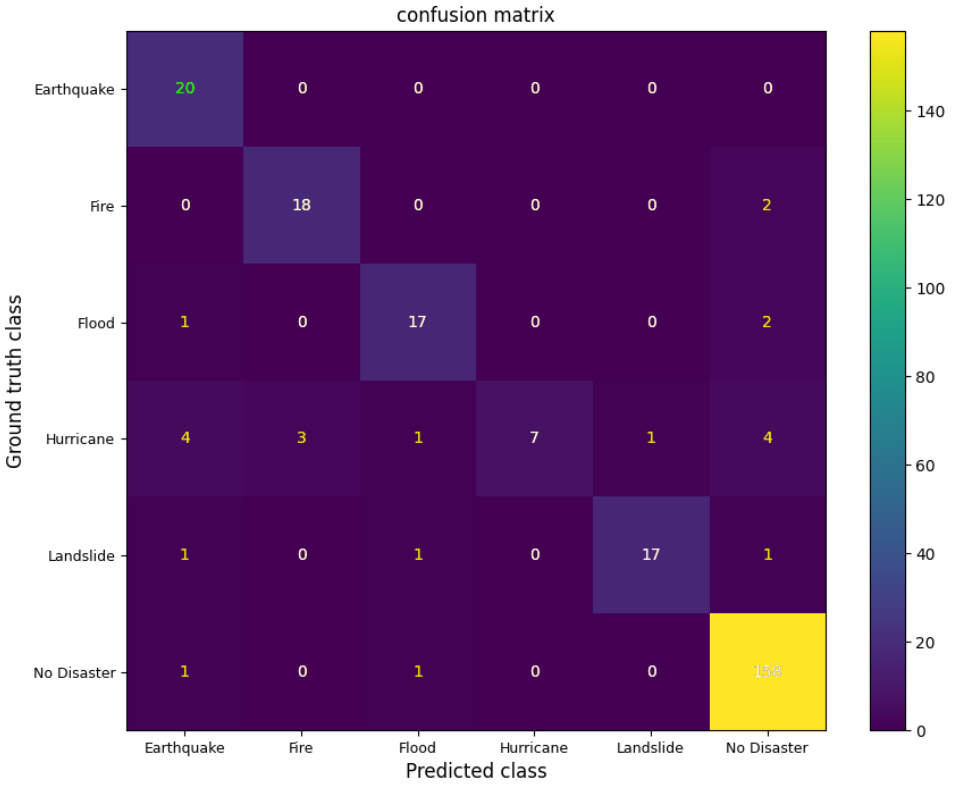


Figure 5 A confusion matrix showing the ground truth classes versus the classes predicted by Efficient-Net B1

The ResNet-101 model evaluated on the Medic dataset achieves an overall precision of 0.611, recall of 0.744, and an F1-score of 0.665 across 320 samples. When restricting the analysis to only the categories represented in the predictions (i.e., excluding `Blizzard`, `Drought`, and `Heavy Rains`), the model attains a category-level precision of 0.755, recall of 0.912, and an F1-score of 0.822, demonstrating strong performance for included categories.

At the category level, performance varies across disaster types. High-performing categories include `Fire` (F1 = 0.93), `No Disaster` (F1 = 0.85), and `Earthquake` (F1 = 0.79), which combine high precision and recall, indicating reliable identification and minimal false positives. Moderately performing categories include `Flood` (F1 = 0.72) and `Hurricane` (F1 = 0.67), where recall is relatively high but precision is slightly lower, suggesting occasional misclassifications. Landslide exhibits good precision but slightly lower recall (F1 = 0.78), reflecting a balance of strengths and weaknesses. Categories absent from the subset—`Blizzard`, `Drought`, and `Heavy Rains`—naturally of ground of lack from classification, exhibit precision, recall, and F1-scores of 0.

In the confusion matrix, the ground-truth classes are listed along the Y-axis, while the predicted classes are displayed along the X-axis. Ideally, correct classifications align along the diagonal from the top left to the bottom right, indicating perfect performance. This

ideal alignment was achieved most effectively by RES within the MEDIC framework.

The model performed particularly well in classifying non-relevant data, i.e., images belonging to the `No Disaster` category. Out of 160 images, 156 were correctly identified as `No Disaster`, demonstrating a high level of accuracy in distinguishing non-disaster content.

Regarding the disaster categories, the classes `Earthquake` and `Fire` performed strongly, with 19 out of 20 images correctly classified in each case. The categories `Flood`, `Hurricane`, and Landslide, however, showed lower to moderate classification reliability, with misclassification's distributed across several other categories.

Overall, these results indicate that RES performs robustly on the Medic dataset for frequent or visually distinct categories while showing inherent limitations for excluded or rare categories. The model's strong performance in accurately identifying `No Disaster` images substantially contributed to its high overall classification accuracy.

Table 2 Category-wise classification metrics for the RES model on the Medic dataset. `Blizzard`, `Drought` and `Heavy Rains` were not part of MEDIC Categories and are therefore not included

Category	Precision	Recall	F1-Score	Support
Earthquake	0.68	0.95	0.79	20
Fire	0.90	0.95	0.93	20
Flood	0.60	0.90	0.72	20
Hurricane	0.75	0.60	0.67	20
Landslide	0.88	0.70	0.78	20
No Disaster	0.75	0.97	0.85	160
Overall /Weighted	0.75	0.91	0.82	260
Overall / Weighted Ex.	0.61	0.74	0.67	320

The models EFF and RES both using the MEDIC Data set are showing both very similar results. This raises the question: Is the Difference of the Data Set more Important than the model it self ?

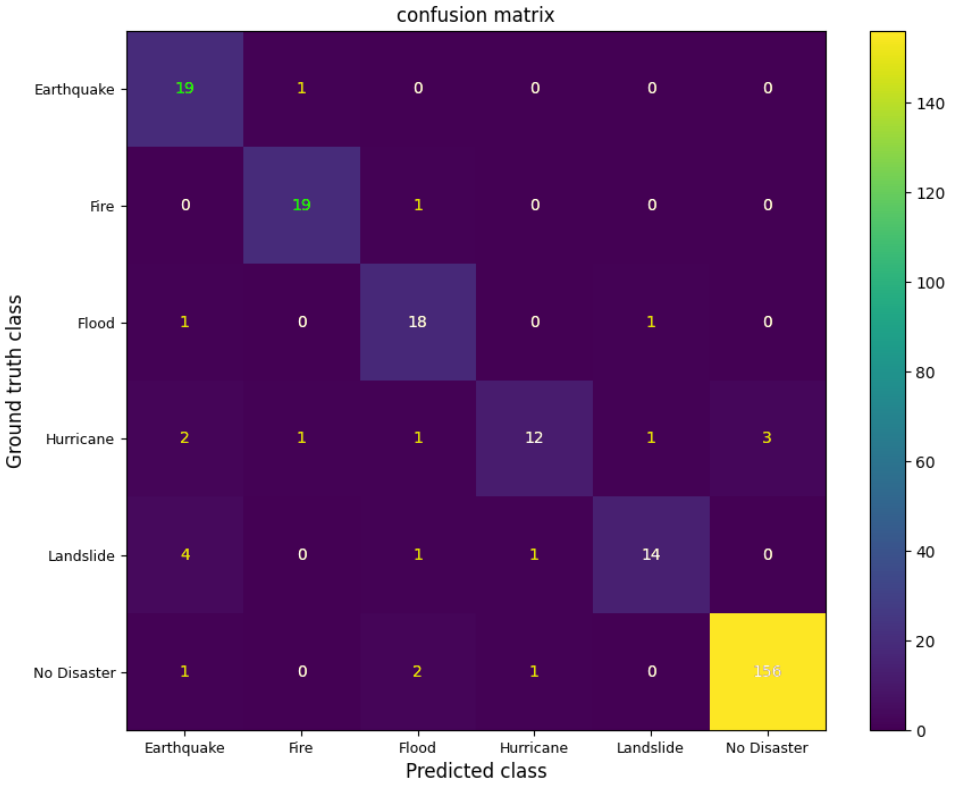


Figure 6 A confusion matrix showing the ground truth classes versus the classes predicted by ResNet 101

5.2.2 OpenClip

The OpenCLIP dataset presents the results of a ranking evaluation for categories based on a weighted model. Each row records the rank of the ground-truth category among the computed weights. The rank values were converted using 1/rank, so that rank 1 corresponds to 1, and rank 9 to 0.1111. Higher values indicate better model performance in predicting the correct category.

A total of 320 rank values were collected, with an average rank value of $\bar{R} = 0.553$. Compared to previous evaluations of CoCa, this indicates that OpenCLIP tends to rank ground-truth categories higher on average, reflecting improved performance.

The frequency distributions reveal notable differences between categories. The `No Disaster` category achieved consistently low ranks across the spectrum, being ranked first only twice. This indicates that No Disaster is, as in CoCa, not effectively recognized. In contrast, categories such as `Heavy Rains`, Drought, `Flood`, `Earthquake`, `Fire`, and `Blizzard` frequently occupy rank 1, indicating very strong performance for these events. Less frequent low or mid-range ranks are observed for `Hurricane` and `Landslide`, reflecting more variability in model performance.

Overall, the best-performing categories are `Fire` and `Blizzard`—with `Fire` achieving all 20 of its instances at rank 1—whereas categories with more dispersed rank distributions, such as `Hurricane`, show comparatively weaker performance. These results suggest that OpenCLIP is highly effective for clearly defined categories but still faces challenges with rarer or more complex events. The model exhibits more pronounced extremes, tending to favor certain categories strongly.

Complementing the rank-based evaluation, the OpenCLIP model achieves a precision of 0.796, a recall of 0.400, and an overall F1-score of 0.325, with an accuracy of 0.40 across the 320 samples.

A detailed category-level analysis reveals notable variability in performance across disaster types. Among the high-performing categories, `Blizzard` (F1 = 0.84), `Drought` (F1 = 0.86), and `Heavy Rains` (F1 = 0.85) demonstrate both high precision and recall, indicating that the model effectively captures the majority of instances while minimizing false positives. `Earthquake` (F1 = 0.74) and `Landslide` (F1 = 0.70) represent moderately performing categories, characterized by reasonably high recall but slightly lower precision, suggesting some misclassification.

In contrast, categories such as `Fire` (F1 = 0.37), `Flood` (F1 = 0.27), `Hurricane` (F1 = 0.36), and `No Disaster` (F1 = 0.02) exhibit comparatively poor performance. Notably, the `No Disaster` category achieves perfect precision but extremely low recall, reflecting that almost all instances are not identified despite correct predictions when made. This discrepancy may be influenced by the number of disaster-related images available across categories, potentially introducing a Inclination in model behavior.

In the confusion matrix, the Y-axis lists the ground truth classes, while the X-axis represents the predicted classes—i.e., how OpenCLIP ultimately classified each image. Examination of the confusion matrix reveals that the classes `Fire` and `Blizzard` performed exceptionally well, with 19/20 and 20/20 correctly identified images, respectively. This indicates a high level of confidence and accuracy in OpenCLIP's ability to correctly recognize these categories.

Unlike CoCa, OpenCLIP does not exhibit a broadly distributed misclassification pattern but rather a few distinct outliers. The most significant misclassification occurs with the `No Disaster` category, where the model performed poorly. However, misclassified `No Disaster` images were mostly reassigned to `Fire` and `Flood`, with 62 and 75 out of 160 images respectively—both objectively short and simple labels. Another notable outlier is the `Hurricane` category, in which only 6 out of 20 images were correctly classified, while 7 were instead labeled as `Flood`.

Overall, it is evident that the `Fire` and `Flood` classes dominate not only their own predictions but also absorb a substantial portion of misclassified samples from other categories. For instance, within the `Drought` class, 16 out of 20 images were correctly identified (ap-

proximately the average accuracy across all classes), while the four misclassified instances were redistributed to `Fire` (1) and `Flood` (3).

The confusion matrix demonstrates that OpenCLIP achieved good to excellent classification performance across most categories, with the exceptions of `No Disaster` and `Hurricane`.

In the binary classification task distinguishing Disaster from `No Disaster`, the results tend toward moderate to slightly above-average performance. The overall accuracy achieved was 0.59, with a precision of 0.6161, a recall of 0.5906, and an F1-score of 0.5668.

The largest difference between the two categories is observed in the recall values: while the Disaster class reached a recall of 0.82, the `No Disaster` class achieved only 0.36. This indicates that, even when all disaster-related categories are combined into a single overarching class, the model still fails to correctly identify the majority of `No Disaster` images.

However, the inclusion of the Disaster class as a unified category led to a reduction in classification uncertainty for the `No Disaster` class, suggesting that the binary framing contributed to slightly more stable decision boundaries.

Taken together, these findings indicate that OpenCLIP performs particularly well for frequent or visually distinct disaster categories—especially `Blizzard`, `Drought`, and `Heavy Rains`—while continuing to face substantial challenges with less frequent or visually ambiguous events.

Table 3 Rank Frequencies per Category (OpenCLIP)

Category	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9
No Disaster	2	7	23	41	43	20	13	4	7
Heavy Rains	17	1	1	0	1	0	0	0	0
Drought	16	1	1	2	0	0	0	0	0
Flood	17	3	0	0	0	0	0	0	0
Earthquake	16	3	4	1	0	2	0	0	0
Fire	20	0	0	0	0	0	0	0	0
Blizzard	19	0	0	0	0	0	0	1	0
Hurricane	6	3	4	2	3	2	0	0	0
Landslide	15	5	0	0	0	0	0	0	0

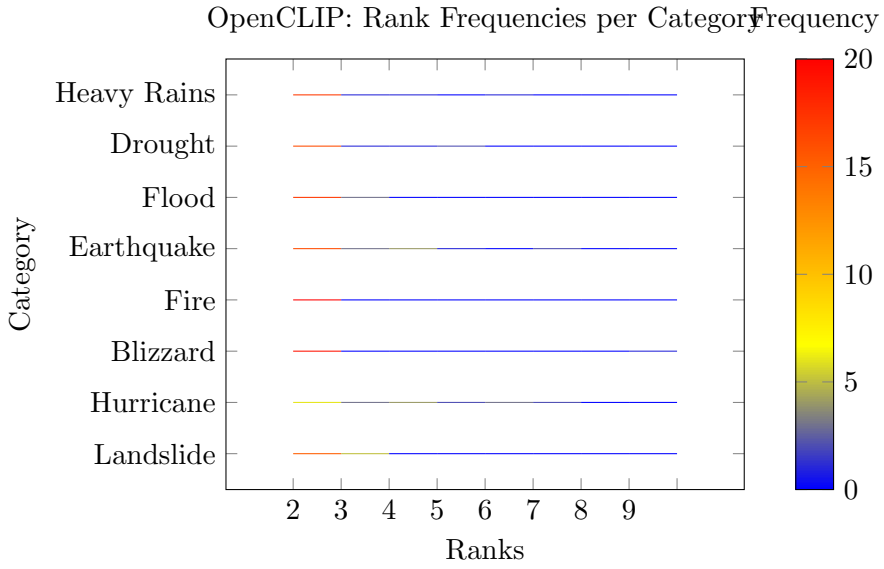


Figure 7 OpenCLIP heatmap of rank frequencies per category. Darker colors indicate higher frequencies.

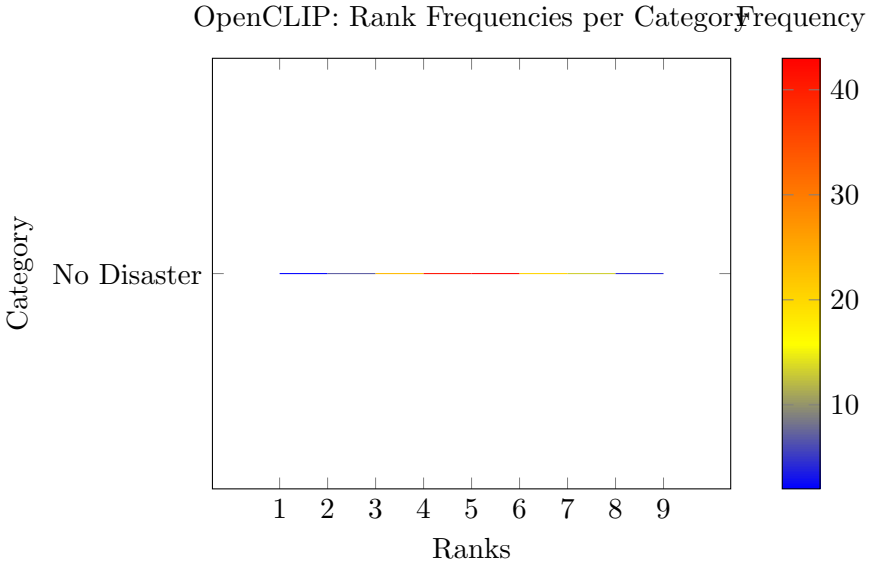


Figure 8 OpenCLIP heatmap of rank frequencies per category. Darker colors indicate higher frequencies.

Table 4 Category-wise classification metrics for the OpenCLIP model. Best-performing categories are highlighted in bold.

Category	Precision	Recall	F1-Score	Support
Blizzard	0.76	0.95	0.84	20
Drought	0.94	0.80	0.86	20
Earthquake	0.70	0.80	0.74	20
Fire	0.22	1.00	0.37	20
Flood	0.16	0.85	0.27	20
Heavy Rains	0.85	0.85	0.85	20
Hurricane	0.46	0.30	0.36	20
Landslide	0.65	0.75	0.70	20
No Disaster	1.00	0.01	0.02	160
Overall / Weighted	0.80	0.40	0.32	320

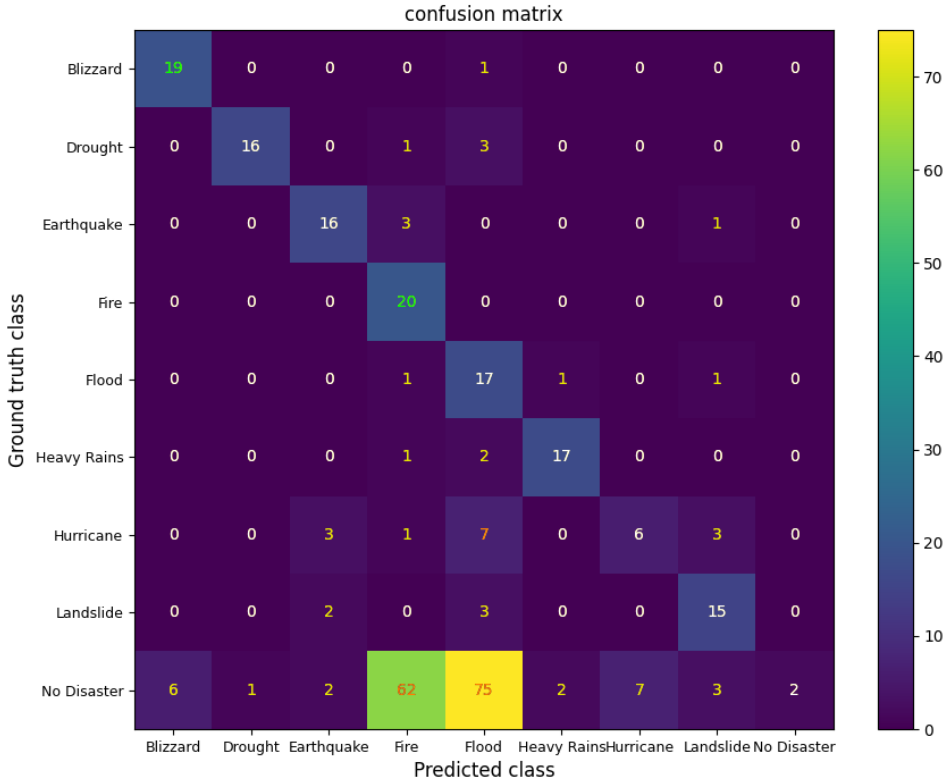


Figure 9 A confusion matrix showing the ground truth classes versus the classes predicted by OpenCLIP

5.2.3 CoCa

The dataset contains 320 evaluation instances used to assess the performance of the CoCa model. Two complementary approaches were employed: a rank-based evaluation using the Mean Reciprocal Rank (MRR) and a classification evaluation based on accuracy, precision, recall, and F1-score. As well as Konfusion Matrix were Made. Together, these analyses provide a detailed view of the model's strengths and weaknesses across categories. Because

Table 5 Disaster/No Disaster classification metrics for the OpenCLIP model. Best-performing categories are highlighted in bold.

Category	precision	recall	f1-score	support
disaster	0.56	0.82	0.67	160
no disaster	0.67	0.36	0.47	160
Overall / Weighted	0.6161	0.5906	0.5668	320

- Accuracy: 0.59
- Macro Avg: 0.62 / 0.59 / 0.57
- Weighted Avg: 0.62 / 0.59 / 0.57

Coca had captions as outputs semantic texual similarity was edit to detect. Every Mention of CoCa Detecting or Classifying something is meant as Semantic Texual Similarity done that base on the caption CoCa generated.

For each instance, the position of the ground-truth category in the ranked list was recorded. Reciprocal ranks were computed as 1/rank 1/rank, where rank 1 corresponds to a score of 1.0 and rank 9 corresponds to 0.1111. Higher values indicate better model performance in identifying the correct category.

Across all instances, the average reciprocal rank was $\bar{R} = 0.446$. This suggests that, on average, the ground-truth category tends to appear near the second rank. While this reflects reasonably strong performance, there is still room for improvement.

The distribution of ranks reveals marked differences between categories. The `No Disaster` category most frequently appeared at lower ranks (8 and 9), reflecting poor predictive performance. By contrast, categories such as `Drought`, `Flood`, and `Fire` exhibited greater variability, with many instances achieving higher ranks. Notably, `Heavy Rains` was most often ranked first and rarely appeared at intermediate positions, indicating strong predictive accuracy.

Overall, the best-performing categories are `Heavy Rains`, `Fire`, and `Flood`. In contrast, `Blizzard` and `Hurricane` show weaker results, with broader and more dispersed rank distributions. These findings suggest that the model performs well for clear and dominant event categories, while struggling with rarer or more complex events.

Complementing the MRR analysis, the CoCa model achieves an overall precision of 0.629, recall of 0.297, and an F1-score of 0.202, with an accuracy of 0.30 across the 320 samples. These aggregate values highlight limited predictive performance, with precision substantially exceeding recall.

A closer inspection at the category level reveals considerable disparities. The best-performing categories include `Heavy Rains` (F1 = 0.63), `Flood` (F1 = 0.59), and `Fire` (F1 = 0.47),

each demonstrating a favorable balance between precision and recall. This indicates that the model can both correctly identify these events and capture a substantial proportion of their occurrences.

Moderate performance is observed for `Earthquake` (F1 = 0.42) and Landslide (F1 = 0.32). These categories are characterized by relatively high recall but lower precision, suggesting that the model frequently detects such events but also generates a notable number of false positives.

By contrast, `Blizzard` (F1 = 0.00), `Hurricane` (F1 = 0.04), and `No Disaster` (F1 = 0.06) fall into the low-performing group. Particularly, the `No Disaster` category shows perfect precision but extremely low recall, indicating that while predictions are correct when made, the majority of relevant instances are missed. This discrepancy may be influenced by the imbalanced number of images available across categories.

The dataset presents the results of a ranking evaluation for categories based on a weighted model. For each row, the rank of the ground-truth category among the computed weights was recorded. The rank values were then converted using 1/rank, such that rank 1 receives a value of 1, while rank 9 corresponds to 0.1111. Values closer to 1 indicate better model performance in identifying the correct category.

In total, 320 rank values were collected, with an average rank value of $\bar{R} = 0.446$. This suggests that, on average, the ground-truth categories tend to appear at the second rank, that's relatively high in the rankings, though there remains room for improvement.

The frequency distributions across ranks reveal notable differences between categories. The `No Disaster` category achieved the lowest ranks most frequently (ranks 8 and 9), indicating a bad model performance for this Category. In contrast, categories such as `Drought`, `Flood`, and `Fire` show greater variability, with many higher-ranked occurrences. Particularly, `Heavy Rains` predominantly appears in rank 1 and is rarely found in intermediate ranks, indicating therefore, a strong predictive accuracy.

In the confusion matrix, the ground truth categories are listed along the y-axis, indicating the correct category for each image, while the x-axis represents the categories assigned by the models. For CoCa, 36 out of 320 images—most of them belonging to the `No Disaster` class, i.e., images not depicting natural disasters—were misclassified as `Fire`. In general, misclassification is particularly pronounced for the `No Disaster` images, which were widely distributed across all disaster categories, with an average of around 19 images per disaster type. Out of 160 `No Disaster` images, only 5 were correctly identified as such. However, as already observed in the precision analysis, only `No Disaster` images were ever classified as `No Disaster`, which indicates that the model does not erroneously assign images to this category without justification.

In contrast, `Heavy Rains` achieved 19 out of 20 correct classifications, though 21 ad-

ditional images—10 of which were `No Disaster` images—were incorrectly classified as `Heavy Rains`. For `Hurricane`, `Blizzard`, and `Drought`, the classification performance was notably poor, with only a quarter or fewer of the respective images being correctly identified. The remaining disaster categories generally achieved around a three-quarter accuracy rate, although the persistent misclassification of `No Disaster` images substantially contributed to the error distribution across categories.

In the binary classification task distinguishing Disaster from `No Disaster` for CoCa the model achieved an overall accuracy of 0.54. As shown in Table 8, performance varied notably between classes. For the disaster class, the model obtained a precision of 0.52, recall of 0.89, and F1-score of 0.66, indicating strong sensitivity but moderate precision. In contrast, the `No Disaster` class achieved a precision of 0.62, recall of 0.19, and F1-score of 0.29, suggesting that many Captions of non-disaster Pictures were misclassified as disasters. The macro-averaged F1-score was 0.47, while the weighted average F1-score was also 0.47, reflecting an overall moderate but imbalanced performance, with the model biased toward detecting disaster-related pictures.

Overall, the best-performing categories are `Heavy Rains`, `Fire` and `Flood`, while `Blizzard` and `Hurricane` exhibit the weakest performance, as evidenced by their broader and more dispersed rank distributions such like `No Disaster`s. It seemed that there isnt quite the compatibility for `No Disaster` with CoCa. These findings highlight that the model performs well for clear and dominant categories but struggles with rarer or more complex events.

Table 6 Rank Frequencies per Category

Category	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9
No Disaster	5	7	5	15	12	14	22	37	43
Heavy Rains	19	1	0	0	0	0	0	0	0
Drought	5	7	3	1	2	1	1	0	0
Flood	18	1	0	0	0	1	0	0	0
Earthquake	15	3	2	1	1	0	0	0	0
Fire	18	1	0	0	0	1	0	0	0
Blizzard	0	2	1	1	5	5	5	0	1
Hurricane	1	5	2	6	4	2	0	0	0
Landslide	14	2	2	1	1	0	0	0	0

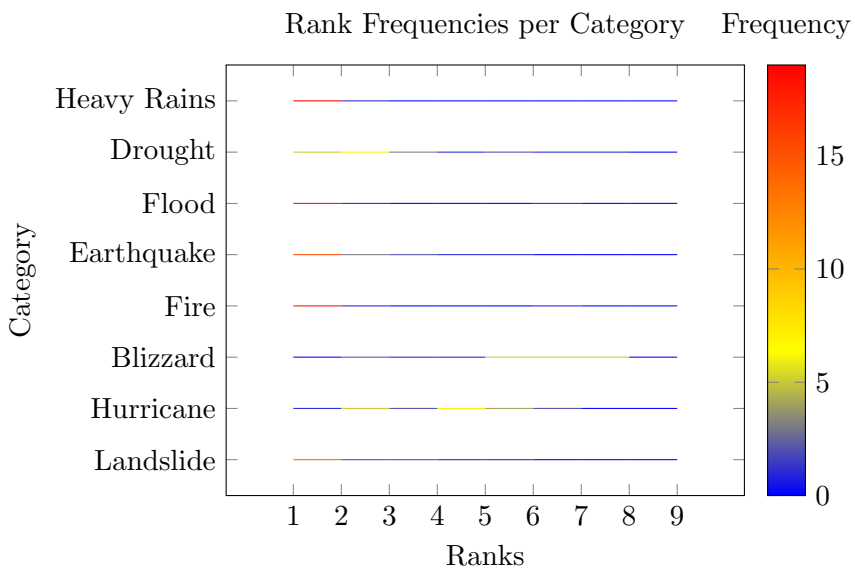


Figure 10 Heatmap of rank frequencies per category. Darker colors indicate higher frequencies.

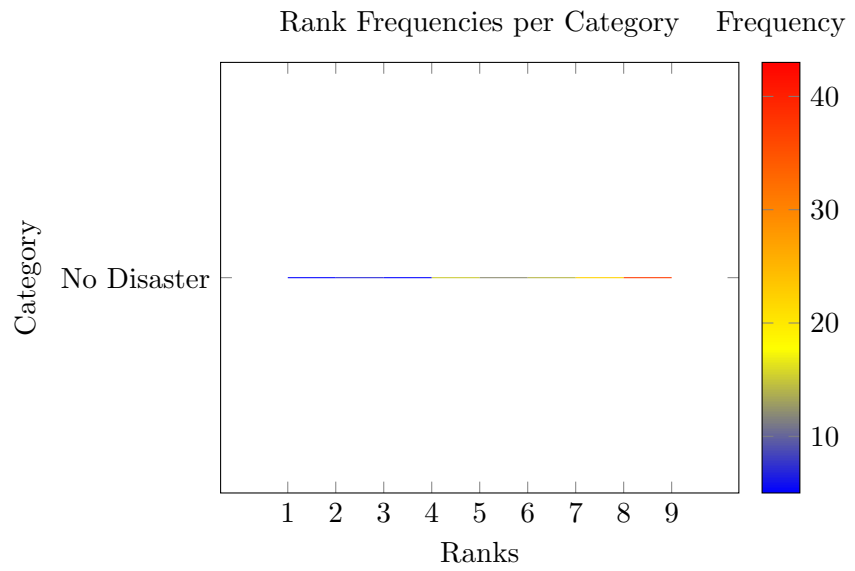


Figure 11 Heatmap of rank frequencies per category. Red colors indicate higher frequencies.

Table 7 Category-wise classification metrics for the CoCa model. Best-performing categories are highlighted in bold.

Category	Precision	Recall	F1-Score	Support
Blizzard	0.00	0.00	0.00	20
Drought	0.29	0.25	0.27	20
Earthquake	0.29	0.75	0.42	20
Fire	0.32	0.90	0.47	20
Flood	0.44	0.90	0.59	20
Heavy Rains	0.47	0.95	0.63	20
Hurricane	0.04	0.05	0.04	20
Landslide	0.21	0.70	0.32	20
No Disaster	1.00	0.03	0.06	160
Overall / Weighted	0.63	0.30	0.20	320

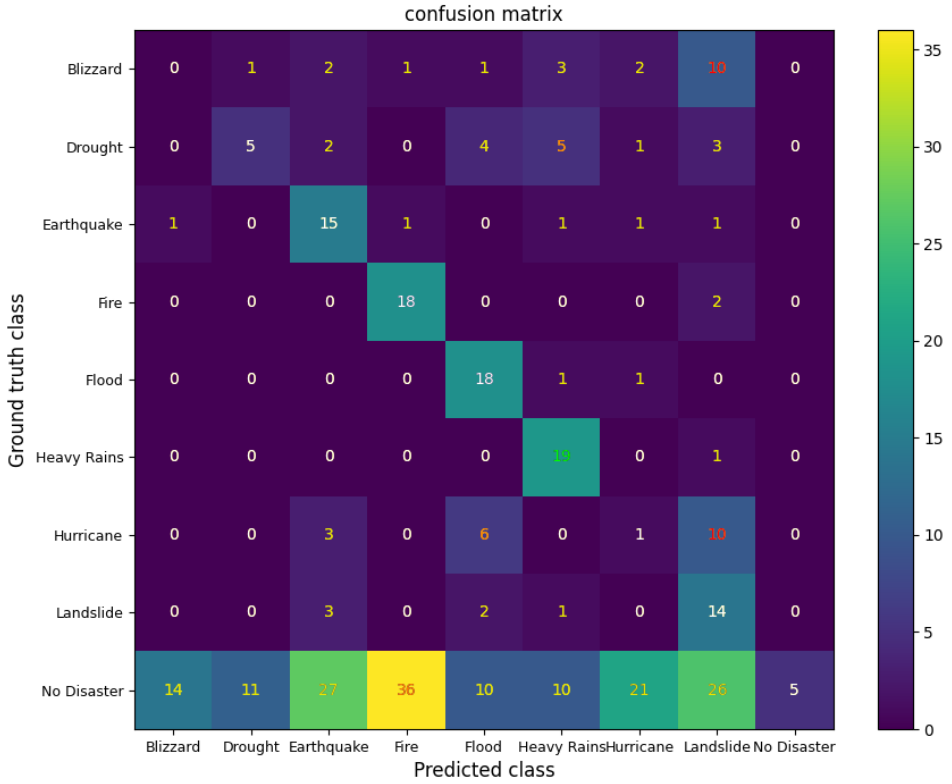


Figure 12 A confusion matrix showing the ground truth classes versus the classes predicted by CoCa

6 Discussion

In the following section, the results of the models are examined in greater detail. The findings are interpreted to assess their significance and compared to those reported in related studies in order to evaluate whether the models used in this work can achieve competitive performance. Subsequently, all potential sources of uncertainty that may have arisen during the study and could have influenced the results are discussed.

Table 8 Disaster/No Disaster classification metrics for the CoCa model. Best-performing categories are highlighted in bold.

Category	precision	recall	f1-score	support
disaster	0.52	0.89	0.66	160
no disaster	0.62	0.19	0.29	160
Overall / Weighted	0.5735	0.5375	0.4729	320

- Accuracy: 0.54
- Macro Avg: 0.57 / 0.54 / 0.47
- Weighted Avg: 0.57 / 0.54 / 0.47

6.1 Interpretaition of Results

6.1.1 Similarity Between Efficient Net B1 and ResNet-101

The two models used within the MEDIC Dataset, EFF and RES, exhibit very similar results. Both distinguish themselves through their strong classification capability for the `No Disaster` category, which enables them to effectively identify irrelevant images, setting them apart from OpenCLIP and CoCa.

This similarity may partly be attributed to the fact that EFF and RES are not fundamentally different architectures. More importantly, however, both models benefit from the MEDIC framework, which employs a two-stage classification process (Alam et al., 2021;Alam et al., 2023). In this approach, images are first classified as either Disaster or `No Disaster`, and only subsequently are disaster images further categorized into specific disaster classes in a second step. This binary pre-classification likely explains why the `No Disaster` category is classified with such high accuracy by the models implemented within the MEDIC framework.

When comparing the models, EFF and RES within the MEDIC framework are well-suited for classifying `No Disaster` images. While EFF achieves a slightly higher recall, it experiences a comparable drop in precision, albeit for different reasons: in EFF, the reduction is caused by over-classification of other categories as `No Disaster`, whereas in RES, it is primarily due to a slightly lower recall.

In contrast, CoCa and OpenCLIP continue to face significant challenges in accurately identifying and classifying `No Disaster` images, demonstrating lower performance on this category compared to the MEDIC-based models.

6.1.2 Difference in Klassifikation from CoCa and OpenCLIP

The results of CoCa and OpenCLIP differ considerably, even though both models share the same weakness in classifying `No Disaster` images. However, the distribution of `No Disaster` rankings differs between the two: in OpenCLIP, the distribution follows a Gaus-

sian (normal) curve, whereas in CoCa, it forms a linearly increasing trend. Despite both models being generally unsuitable for accurately classifying `No Disaster` images, OpenCLIP demonstrates relatively more stable and balanced performance and is therefore preferable between the two.

When evaluating the results concerning the classification of Disaster categories, it is noteworthy that OpenCLIP performs better in most categories and significantly better in some, such as `Drought` and `Blizzard`. Consequently, based on the results of this study, OpenCLIP demonstrates a higher level of classification reliability for Disaster categories compared to CoCa.

However, both models still exhibit only moderate overall performance in the classification of Natural Disaster images. In contrast, for the classification of `No Disaster` images, both models prove to be unsuitable, performing poorly in accurately identifying non-disaster content.

The models trained with MEDIC data produced results that differ substantially from those of CoCa and OpenCLIP. While the MEDIC-based results are more accurate, they do not include the categories `Blizzard`, `Drought`, and `Heavy Rains`, which indicates a limited breadth of classification. The results of OpenCLIP and CoCa may also have lost some reliability due to the effort of aligning them with the MEDIC results in order to ensure comparability. However, this adjustment introduces a different form of incomparability.

As evident from the results, there is a considerable discrepancy in the confidence with which the class `No Disaster` was classified across the models. For EFF and RES, both trained on the MEDIC dataset, the `No Disaster` class was identified with high precision and accuracy (see Table 1 and Table 2). In contrast, both the CoCa and OpenCLIP models show significant drops in performance when classifying the `No Disaster` category. These models were trained on a different dataset and thus rely on a different training basis.

An additional analysis in which `No Disaster` was paired with each disaster category to form a binary classification task indicates that precision and accuracy for `No Disaster` improved substantially (see Table 5 and Table 8). Assuming that these improvements are not solely due to the reduced number of possible misclassifications—i.e., a higher likelihood of correct assignments occurring by chance—the results suggest that the discrepancy is primarily attributable to differences in the training datasets.

This hypothesis is further supported by the observation that the two models, EFF and RES, both of which share the same training data, yield almost identical results with respect to the classification confidence of the `No Disaster` class. Conversely, CoCa and OpenCLIP, which were trained on different datasets, exhibit distinct patterns in their classification accuracy.

This becomes particularly evident in Tables 3 and 6 as well as Figure 8 and 11: for CoCa, the frequency of images classified as `No Disaster` increases linearly with lower ranks, meaning that the number of such images decreases as the rank rises. In contrast, for OpenCLIP, this decline does not follow a linear pattern but rather resembles a Gaussian distribution. As a result, `No Disaster` classifications in OpenCLIP are concentrated more around the middle ranks rather than being skewed toward the lowest ranks.

6.1.3 Comparison with Existing Research

The results obtained for the four models show varying levels of performance, with each model outperforming the others in certain areas. However, how do these models perform in comparison to previous studies that employed different models and methodologies?

(Didur et al., 2025) reported substantially higher classification performance in the multi-class identification of remnants of destroyed buildings, achieving precision and recall values of 0.97. This performance was obtained using the YOLO model trained on a larger dataset with a greater number of classes (ten in total). However, it is important to note that (Didur et al., 2025)'s model was specifically trained for this particular classification task, which likely contributed to the improved results.

In (Colombo, 2018)'s study, the application of Gaussian Naive Bayes demonstrated a similar or, in some cases, even less precise filtering performance for the `Earthquake` category compared to the classification results of OpenCLIP. With a recall of 95% for Gaussian Naive Bayes applied to English-language texts containing relevant and potentially relevant data, (Colombo, 2018)'s results are comparable to—and in some cases exceed—the outcomes achieved by EFF and RES in this work. Using a Random Forest approach for filtering, (Colombo, 2018) reported precision and recall values of 86% and 80%, respectively. While these results are significantly higher than those obtained by CoCa and OpenCLIP, they are comparable to the results achieved by EFF and RES, where precision was notably lower but recall significantly higher. Therefore, the results of EFF and RES trained on the MEDIC dataset can be considered highly competitive with other commonly applied methods. In some cases, these two models even demonstrate greater classification reliability than other pre-trained models, such as those used in (Colombo, 2018)'s study.

(Nguyen, Alam, et al., 2017) also applied filtering to social media images to distinguish between disaster-relevant and non-disaster-relevant content. In this study, (Nguyen, Alam, et al., 2017) achieved a precision of 0.99, a recall of 0.97, and an F1-score of 0.98 using a combination of deep learning and perceptual hashing. The recall obtained by (Nguyen, Alam, et al., 2017) is comparable to that achieved by EFF and RES in this work; however, the precision reported by (Nguyen, Alam, et al., 2017) is significantly higher.

(LeoNila et al., 2024) achieved strong results in forest fire detection using camera imagery, obtaining an accuracy of 89% with artificial neural networks and 95% with convolutional neural networks. The results for the `Fire` category in this study—82% accuracy for

OpenCLIP and 89% for CoCa—are therefore slightly less reliable or roughly comparable to those reported by (LeoNila et al., 2024) using ANNs. However, when excluding the `No Disaster` class, the accuracy values increase to 96% for OpenCLIP and 98% for CoCa, indicating that both models perform even better than (LeoNila et al., 2024)'s approach when distinguishing fire events from other disaster types.

Overall, it shows that the models employed in this study achieve performance levels that are competitive with, most still lacking behind and in several cases could surpass, those reported in prior research. These results highlight the robustness and adaptability of the proposed approach across diverse disaster-related classification tasks.

6.2 Possible Sources of Error

6.2.1 Possible Errors Made in the Sampling Prozess

Potential errors in the sampling process may include the following: in the `Drought` category, some images may actually depict desert landscapes rather than drought phenomena per se. While such scenes are dry, they may not directly correspond to the concept of drought. Similarly, images of fields with varying levels of vegetation blur the boundary between normal soil conditions and drought-induced dryness.

Figure 13 Image form the GDELT Data included in Samples depicting a rather green drought faced landscape

For `Earthquake`, images of destroyed buildings may, in fact, originate from war-related destruction rather than Earthquakes. The accuracy of these images with respect to the true cause of destruction could not be fully verified. In the case of `Flood`, some images may simply depict rivers or people within rivers, independent of flooding (e.g., images of cleaning activities).

For `Heavy Rains`, the division of the category into two words may itself introduce problems, and the category may overlap visually with `Blizzard` or `Flood`. For `Hurricane`, phenomena not specifically attributable to hurricanes—such as tornadoes or other disasters—may have been sampled into the category. In addition, the images sometimes show visual similarities to `Flood`.

For `Landslides`, stones lying on roads may be interpreted not as evidence of mass movements but rather as rock debris similar to that associated with Earthquakes. Cracks in roads, which were sampled under `Landslides`, could also plausibly be classified as earthquake-related damage.

6.2.2 Possible errors Made in the Usage of the Models

During the application of the models, careful attention was paid to their correct use. However, when processing the full dataset of over 300,000 images, divided into subsets of 50,000, corrupted image files repeatedly occurred. This issue was resolved by conducting the analysis on a smaller sample of only 320 images drawn from the dataset. Consequently, the reported results of the models were not influenced by any technical issues arising during their execution. But the exclusion of influence trought this error could not be proven.

6.2.3 Possible errors Made in the Statistical valuation of the Results

During the statistical evaluation, rows of results from the published tables were manually extracted. As a consequence, a minimal residual risk remains that values corresponding to individual images may have been inadvertently transposed. Several manual validations were carried out to ensure that such errors were avoided. Although this issue occurred on multiple occasions throughout the study, the likelihood of any error having affected the final reported results is considered to be extremely low. Nevertheless, for the sake of completeness, it cannot be entirely ruled out.

6.2.4 Inclination trought difference in Training Data

The training datasets of OpenCLIP and CoCa were designed for general-purpose classification tasks and therefore appear less capable of effectively handling the category `No Disaster`. The hypothesis that CoCa and OpenCLIP struggle simply because they cannot adequately account for the absence of an event can, however, be rejected. This is supported by an analysis showing that, when `No Disaster` was contrasted solely with the single category Disaster, classification accuracy for `No Disaster` improved significantly, while the reliability of Disaster classification remained largely unaffected.

This suggests that the weaker performance is more likely attributable to the broad and heterogeneous definition of the Disaster category itself, which introduces considerable variability. Overall, the results of both models indicate that more prominent and well-known disaster types were classified with greater accuracy—likely because their higher public salience provided more extensive representation within the training data.

One factor that may contribute to the advantage of the MEDIC models is the preliminary binary classification step that precedes the subsequent multi-class classification. This assumption is supported by the fact that `No Disaster` is among the best-classified categories in the MEDIC models. In contrast, when a binary classification of Disaster versus `No Disaster` was applied to CoCa and OpenCLIP, the classification performance for `No Disaster` deteriorated considerably.

6.2.5 Difficulty in compairing CoCa Results

This difference in output format particularly affected CoCa, as the Contrastive Captioner produces textual captions for images rather than single-word classifications or numerical values. To make use of these captions and ensure comparability of results, semantic textual similarity between the predefined categories and the captions generated by CoCa was applied. However, the introduction of an additional AI model inevitably introduced further uncertainty into the results.

In the manual comparison of captions with their corresponding images, two types of errors on the part of CoCa can be identified.

1. The caption describes elements that are not present in the image, thereby rendering the caption inaccurate.

2. The caption omits obvious and salient elements of the image, which makes the caption incomplete.

In CoCa, the largest deviations occurred for images in the `No Disaster` category that were misclassified as `Fire`. Specifically, 36 out of 160 `No Disaster` images were assigned to this category. For one such image, CoCa generated the caption: "two men standing next to each other on a field .". As shown in Figure 14, the generated caption accurately describes the visual content of the image. This suggests that the issue lies in the application of semantic textual similarity rather than in CoCa's captioning itself. How-

Figure 14 Image form the GDELT Data included in Samples with Category `No Disaster` who was Classified as `Fire`

ever, when evaluated within the binary classification task of Disaster versus `No Disaster`, the same image was classified as `No Disaster`, albeit by a narrow margin (–0.0281 for Disaster and –0.0178 for `No Disaster`). The negative values indicate that, according to semantic textual similarity, neither category matched the caption well. This implies that the misclassification may stem from the number and diversity of the disaster categories. Alternatively, it could also suggest that semantic textual similarity struggles to adequately capture the meaning of broad, abstract umbrella terms.

Figure 15 One of 5 Images from the GDELT Data included in Samples with Category `No Disaster` who was Classified as `No Disaster`

It should be noted, however, that not all images in CoCa were misclassified. Out of the 160 images, five were correctly identified as `No Disaster`. Nevertheless, this does not reflect strong model accuracy on the part of semantic textual similarity. In these cases, CoCa first produced the caption: "a young girl is eating a donut in a kitchen ." that were a nearly accurate description of the image content (Figure 15). Semantic textual similarity subsequently classified the images correctly, but the category scores were weighted very closely together (ranging from 0.035 to –0.1131), which indicates a high degree of classification uncertainty within the semantic textual similarity model itself. In CoCa, the

Figure 16 Image form the GDELT Data included in Samples with Category Blizzard who was Classified as Landslide

categories `Blizzard` and `Hurricane` also exhibited high error rates. The results show that, for `Blizzard`, not a single image was classified correctly. Instead, a large proportion of images from this category were assigned to `Landslides`, with roughly half of the captions generated for `Blizzard` images being classified as such. An illustrative example is shown in Figure 16: CoCa generated the caption "a group of people walking down a street in the snow .", which accurately describes the image content and even explicitly includes the word "snow". This indicates that CoCa clearly recognized snow-related features, which in this case should have corresponded to `Blizzard`. However, the fact that semantic textual similarity assigned the category `Landslides` instead strengthens the assumption that the uncertainty observed in CoCa's results is, to a significant extent, attributable to the inherent uncertainty of the subsequent processing step using the semantic textual similarity

model.

Similar to the case of `Blizzard`, the `Hurricane` category also exhibited a classification shift toward `Landslides`. Half of the `Hurricane` images were incorrectly classified as `Landslides`. In Figure 17, CoCa generated the caption "a tree that has fallen over a car and a bus .", which provides a reasonably accurate description of the image. The only inaccuracies are that the vehicle depicted is a fire truck rather than a bus, and that it is located behind the fallen tree. Notably, the caption does more than simply list recognizable objects; it interprets the scene as an event, stating that the tree has fallen onto the vehicle rather than merely lying across the road. This suggests that CoCa may be capable

Figure 17 Image form the GDELT Data included in Samples with Category `Hurricane` who was Classified as `Landslides`

of representing complex visual relationships in an abstract form. Ideally, such an interpretation could have bridged the fallen tree to wind damage and thereby to the category `Hurricane`. However, a tree may also fall due to ground movement. Thus, the misclassification appears less attributable to errors in semantic textual similarity and more to the complexity of distinguishing between causal contexts (e.g., wind versus ground movement), compounded by potentially broad sampling of the image data and resulting visual similarities with `Landslides`.

Within the `Hurricane` category, there was only one image for which the CoCa-generated caption was also correctly classified as `Hurricane` by semantic textual similarity. The image in Figure 18 received the caption "a car that is sitting in the snow next to a tree .". While this caption satisfies the completeness criterion by mentioning all major objects, it misrepresents their relations: it states that the car is next to a tree, whereas the image clearly shows the tree lying across the car. This constitutes an error in CoCa's caption generation.

Figure 18 Only Image form the GDELT Data included in Samples with Category `Hurricane` who was Classified as `Hurricane`

Moreover, although the caption correctly referenced snow, this did not trigger a misclassification into `Blizzard` within semantic textual similarity. Consequently, this was the only one out of 20 `Hurricane` images that was classified correctly. From this, it can be concluded that while many hypotheses can be formulated based on the observed results, the underlying mechanisms remain opaque. Since the models are CNN's with unsupervised components, they effectively operate as a "black box," making it unclear which specific factors contributed to a given weighting or classification outcome. For this reason, the interpretations presented here must be regarded with caution.

7 Conclusion and Future Work

7.1 Conclusion

The aim of this work was to investigate the reliability of classification accuracy across different models. Initially, the large volume of data provided by GDELT and the absence of valid ground-truth classification data limited the extent to which the GDELT dataset could be fully utilized. Instead, a smaller subset of manually validated and sampled images from this dataset was employed. These curated images were assumed to possess a sufficient degree of correctness to serve as ground truth for a quality analysis of classification performance. In this way, an assessment of classification reliability could be carried out, and although minor errors may cast some uncertainty, the overall objective of the study is considered to have been successfully achieved.

Although this study was influenced by numerous imperfections and external factors, it nevertheless produced valuable results. The findings indicate that the classifications achieved by EFF and RES are comparable to existing studies employing similar models—largely due to the use of MEDIC's pretrained weights for classification.

In contrast, OpenCLIP and especially CoCa performed significantly worse than reported in previous research. Both EfficientNet-B1 and ResNet101 can be considered moderately reliable for the classification of Disaster images and highly reliable for the classification of `No Disaster` images. Between the two, RES is regarded as the more reliable model, as a high number of false negatives in disaster detection—as observed in EFF—could have critical implications.

CoCa and OpenCLIP, on the other hand, achieved moderate classification reliability for disaster categories, with OpenCLIP outperforming CoCa in this regard. However, both models demonstrated very low reliability in recognizing `No Disaster` images. This suggests that, without specialized fine-tuning, CoCa and OpenCLIP are more suited to object recognition tasks than to the detection of complex, abstract phenomena such as natural disasters

7.2 Future Work

Future research could focus on examining whether the performance gap between CoCa and OpenCLIP can be reduced through adjustments to their pretraining processes or by employing techniques such as few-shot prompting. Additionally, it would be of particular interest to investigate whether the observed differences in performance are due to the models' inherent suitability for object recognition tasks or whether they primarily result from the specific training data and methodologies used in their development.

References

Aarsen, T. (2025). All-minilm-l6-v2: Sentence embeddings for semantic search and clustering [Accessed: October 7, 2025]. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Alam, F., Alam, T., Hasan, M. A., Hasnat, A., Imran, M., & Ofli, F. (2023). Medic: A multi-task learning dataset for disaster image classification. *Neural Computing and Applications*, *35*(3), 2609–2632. https://doi.org/10.1007/s00521-022-07717-0

Alam, F., Ofli, F., & Imran, M. (2018). Crisismmd: Multimodal twitter datasets from natural disasters. *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*. https://doi.org/10.1609/icwsm.v12i1.14983

Alam, F., Ofli, F., Imran, M., Alam, T., & Qazi, U. (2020). Deep learning benchmarks and datasets for social media image classification for disaster response. *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. https://doi.org/10.1109/ASONAM49781.2020.9381294

Alam, F., Alam, T., Hasan, M. A., Hasnat, A., Imran, M., & Ofli, F. (2021). Medic: A multi-task learning dataset for disaster image classification. *arXiv preprint*. https://doi.org/10.48550/arXiv.2108.12828

Andersson, J. (2025). Openclip — deep learning project repository [Accessed: October 7, 2025]. https://gitlab.liu.se/wiler441/Deep_Learning_Project/-/blob/b8d6c51ff7038772733b740938ac0a5d1f7cfd9d/openclip

Ankith, I., & Akshaya, H. P. (2021). Convolutional neural network for image recognition. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, *9*(11), 1619–1624. https://doi.org/0.22214/ijraset.2021.39061

Bynke, M. (2022). *Multi-label image classification with language-image models: An approach for a fine-grained domain-specific dataset* [Master's thesis]. NTNU. https://hdl.handle.net/11250/3025437

Colombo, A. (2018). Design of a classification model to filter relevant social media posts in the emergency context [Accessed: October 8, 2025]. https://www.politesi.polimi.it/handle/10589/164393

Craswell, N. (2009). Mean reciprocal rank. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems*. Springer. https://doi.org/10.1007/978-0-387-39940-9_488

CrisisNLP. (2025). Medic [Accessed: October 8, 2025]. https://crisisnlp.qcri.org/medic/index.html

Didur, V., Molchanova, M., & Mazurets, O. (2025). Research on the effectiveness of neural network detection of plots with the destroyed buildings remains. https://elar.khmnu.edu.ua/handle/123456789/18411

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [Accessed: October 8, 2025]. MIT Press. http://www.deeplearningbook.org

Hamadain, F. A., Osman, A. A., Abdelrahman, A., & Hamed, M. (2023). Deep convolutional neural network (dcnn) models for image recognition: A review. Conference proceedings. https://api.semanticscholar.org/CorpusID:262029699

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recog-*

nition (CVPR), 770–778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

Hugging Face. (2025). Using openclip at hugging face [Accessed: October 7, 2025]. https://huggingface.co/docs/hub/open_clip

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., & Schmidt, L. (2021a, July). *Openclip* (Version v0.1). https://doi.org/10.5281/zenodo.5143773

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., & Schmidt, L. (2021b, July). *Openclip* (Version v0.1) [Alternative usage notebook: https://colab.research.google.com/github/mlfoundations/open_clip/blob/master/docs/Interacting_with_open_coca.ipynb]. https://doi.org/10.5281/zenodo.5143773

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Comput. Surv.*, *47*(4), 67. https://doi.org/10.1145/2771588

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2018). Processing social media messages in mass emergency: Survey summary. *Companion Proceedings of the The Web Conference 2018*, 507–511. https://doi.org/10.1145/3184558.3186242

Kersten, J. (2022). Sm-disaster-image-classification [Accessed: October 8, 2025]. https://gitlab.dlr.de/dw-bws/sm-disaster-image-classification

Klein, B., & Mitchinson, D. (2023). Metriken zu evaluation [Accessed: October 8, 2025]. https://www.python-kurs.eu/metriken.php

Leetaru, K. (2013–2022a). Gdelt global knowledge graph (gkg) files [Accessed: October 5, 2025]. http://data.gdeltproject.org/gkg/index.html

Leetaru, K. (2013–2022b). The global database of events, language, and tone (gdelt) [Accessed: October 5, 2025]. https://www.gdeltproject.org/about.html

LeoNila, T., Senthil, G. A., Geerthik, S., Saravanan, K., Kumar, G. A., & Saranya, M. K. (2024). Natural disaster sustainability forest fire detection system based on computer vision using smart iot sensor networks and deep learning techniques. *2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, *1*, 194–199. https://doi.org/10.1109/ICAICCIT64383.2024.10912099

Liu, Y. H. (2018). Feature extraction and image recognition with convolutional neural networks. *Journal of Physics: Conference Series*, *1087*, 062032. https://doi.org/10.1088/1742-6596/1087/6/062032

Mozannar, H., Rizk, Y., & Awad, M. (2018). Damage identification in social media posts using multimodal deep learning. *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, 529–543. https://idl.iscram.org/files/husseinmouzannar/2018/2129_HusseinMouzannar_etal2018.pdf

Nguyen, D. T., Ofli, F., Imran, M., & Mitra, P. (2017). Damage assessment from social media imagery data during disasters. *Proceedings of the IEEE/ACM International*

Conference on Advances in Social Networks Analysis and Mining (ASONAM), 1–8. https://doi.org/10.1145/3110025.3110109

Nguyen, D. T., Alam, F., Ofli, F., & Imran, M. (2017). Automatic image filtering on social networks using deep learning and perceptual hashing during crises. *arXiv preprint*. https://doi.org/10.48550/arXiv.1704.02602

Pi, Y., Nath, N. D., & Behzadan, A. H. (2020). Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Adv. Eng. Informatics*, *43*, 101009. https://doi.org/10.1016/j.aei.2019.101009

Raychanan, _. (2024, January). *How to perform batch inference to accelerate the process - coca*. https://github.com/mlfoundations/open_clip/issues/781

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. https://arxiv.org/abs/1908.10084

Reimers, N. (2025). Minilm-l6-h384-uncased: A 6-layer version of minilm [Accessed: October 7, 2025]. https://huggingface.co/nreimers/MiniLM-L6-H384-uncased

Sangameswar, M. V., Nagabhushana Rao, M., & Satyanarayana, S. (2017). An algorithm for identification of natural disaster affected area. *Journal of Big Data*, *4*(1), 39. https://doi.org/10.1186/s40537-017-0096-1

SBERT.net. (2025). Semantic textual similarity [Accessed: October 8, 2025]. https://sbert.net/docs/sentence_transformer/usage/semantic_textual_similarity.html

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In A. Sattar & B. H. Kang (Eds.), *Ai 2006: Advances in artificial intelligence* (pp. 1015–1021, Vol. 4304). Springer. https://doi.org/10.1007/11941439_114

Susmaga, R. (2004). Confusion matrix visualization. In M. A. Kłopotek, S. T. Wierzchoń, & K. Trojanowski (Eds.), *Intelligent information processing and web mining* (Vol. 25). Springer. https://doi.org/10.1007/978-3-540-39985-8_12

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning*, 6105–6114. https://proceedings.mlr.press/v97/tan19a.html?ref=ji

Toan, N. T., Phan, T. C., Hung, N. Q. V., & Jo, J. (2019). A deep learning approach for early wildfire detection from hyperspectral satellite images. *2019 7th International Conference on Robot Intelligence Technology and Applications (RiTA)*, 38–45. https://doi.org/10.1109/RITAPP.2019.8932740

Turnbull, D. (2021). Compute mean reciprocal rank (mrr) using pandas [Accessed: October 8, 2025]. https://softwaredoug.com/blog/2021/04/21/compute-mrr-using-pandas

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). Coca: Contrastive captioners are image-text foundation models [Accessed: October 8, 2025]. *arXiv preprint arXiv:2205.01917*. http://arxiv.org/pdf/2205.01917v2

Tools

- **ChatGPT-5 mini, OpenAI:** https://chatgpt.com
  - Translation of Phrases and single Words.
  - Summarization of Papers (no Text used in Work)
  - Grammar and Spelling assistance.
  - Fine-tune text appearance in LaTeX.
- **Elicit – The AI Research Assistant, version September 4 2025, Ought:** https://elicit.com
  - Literature research and creation of summaries.
- **Open WebUI, chat-kratos:** chat.kratos.dlr.de
  - Bug Fixing and light Coding Assistance (no generated code lines)
  - Summarization of Papers (no Text used in Work)

Declaration of Academic Integrity (last edited: January 2024)

1. I hereby confirm that this work — or in case of group work, the contribution for which I am responsible and which I have clearly identified as such — is my own work and that I have not used any sources or resources other than those referenced.

   I take responsibility for the quality of this text and its content and have ensured that all information and arguments provided are substantiated with or supported by appropriate academic sources. I have clearly identified and fully referenced any material such as text passages, thoughts, concepts or graphics that I have directly or indirectly copied from the work of others or my own previous work. Except where stated otherwise by reference or acknowledgement, the work presented is my own in terms of copyright.

2. I understand that this declaration also applies to generative AI tools which cannot be cited (hereinafter referred to as 'generative AI').

   I understand that the use of generative AI is not permitted unless the examiner has explicitly authorized its use (Declaration of Permitted Resources). Where the use of generative AI was permitted, I confirm that I have only used it as a resource and that this work is largely my own original work. I take full responsibility for any AI-generated content I included in my work.

   Where the use of generative AI was permitted to compose this work, I have acknowledged its use in a separate appendix. This appendix includes information about which AI tool was used or a detailed description of how it was used in accordance with the requirements specified in the examiner's Declaration of Permitted Resources.

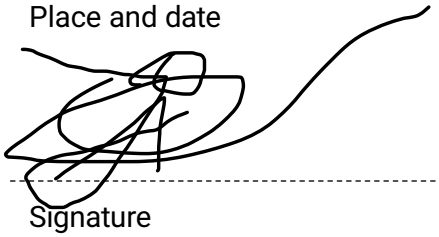
   I have read and understood the requirements contained therein and any use of generative AI in this work has been acknowledged accordingly (e.g. type, purpose and scope as well as specific instructions on how to acknowledge its use).

3. I also confirm that this work has not been previously submitted in an identical or similar form to any other examination authority in Germany or abroad, and that it has not been previously published in German or any other language.

4. I am aware that any failure to observe the aforementioned points may lead to the imposition of penalties in accordance with the relevant examination regulations. In particular, this may include that my work will be classified as deception and marked as failed. Repeated or severe attempts to deceive may also lead to a temporary or permanent exclusion from further assessments in my degree programme.

Jena, Germany 09.10.2025

Place and date



Signature