Final Project CIS 5300

Gokul Nair, Rohan Saraogi, Shivani Prasad Bondapalli, Yash Agrawal

Penn Engineering
University of Pennsylvania

Introduction

Our Final Project focuses on developing a multimodal classifier for analyzing disaster-related tweets. Social media, particularly X, formerly Twitter, has become a key platform for sharing real-time updates during disasters. Our goal is to create a tool that distinguishes between informative and non-informative tweets for humanitarian aid purposes.

The core of our project is to leverage the CrisisMMD: Multimodal Crisis Dataset, featuring thousands of annotated tweets and images from major 2017 natural disasters. This dataset provides a comprehensive basis for training our classifier.

Our approach involves integrating techniques in machine learning and multimodal data analysis, such as Logistic Regression, CLIP, Salesforce BLIP, to process and categorize tweets. The classifier is designed to evaluate both textual and visual content, ensuring a robust analysis of each tweet's relevance to disaster response efforts.

We anticipate our project will offer valuable insights into the application of multimodal classification in real-world scenarios, specifically in the context of emergency management and humanitarian aid.

Motivation

Social media platforms are vital for disseminating information during disaster events. Through platforms like Twitter, individuals share multimedia content on various aspects of the unfolding situation, such as updates on casualties, infrastructure damage, the status of missing or found individuals etc. Numerous studies have highlighted the value of this online information for humanitarian organizations, as when processed swiftly and efficiently, it significantly enhances their situational awareness, enabling them to plan and execute relief operations more effectively. To this end, our goal in this project is to build a multi-modal classifier to determine if a given tweet and/or image is useful for humanitarian aid or not.

Motivation

Opportunity to …

- Apply our learnings from this course on a multimodal (text, image) problem
- Explore the interplay of text and image data such as fusion of text and image embeddings, augmenting text with image captions etc.

Problem Statement

Binary classification task to determine whether a given tweet (text, image) pair is useful for humanitarian aid purposes ("Informative") or not ("Not informative").

(The authors consider a tweet/image as "Informative" if it reports/shows one or more of the following: cautions, advice, and warnings, injured, dead, or affected people, rescue, volunteering, or donation request or effort, damaged houses, damaged roads, damaged buildings; flooded houses, flooded streets; blocked roads, blocked bridges, blocked pathways; any built structure affected by earthquake, fire, heavy rain, strong winds, gust, etc., disaster area maps. Images showing banners, logos, and cartoons are not considered as "Informative".)

Related Work

- **CrisisMMD**: Uses CNNs for both text and images from social media in disaster response, employing early fusion for integrating features from both modalities. Achieves F1-scores of 0.842 in informativeness task.

- **Image4Act**: Develops a pipeline for classifying Twitter images aiding disaster response, combining deep neural networks and human computation for high accuracy. Demonstrated precision between 0.67-0.92 in cyclone image filtering.

- **COVID-19 Tweets**: Focuses on identifying informative COVID-19 tweets with limited data. Applies data augmentation and transformer-based models like CT-BERT. Achieves an impressive F1-score of 0.912 with only a fraction of the dataset.

Dataset

CrisisMMD 2.0: Multimodal Crisis Dataset

Thousands of tweets and images collected during seven major natural disasters.

Goal: Informative vs Not informative

	Informative	Non Informative
Train	8341	5267
Dev	1407	830
Test	1373	864

Class distribution

(a) Informative image

(b) Not informative image

Figure 1: Examples of informative and not informative images

Evaluation Metric

Standard classification metrics:

- Accuracy
- Precision
- Recall
- F1 Score
- ROC curve/AUC Score

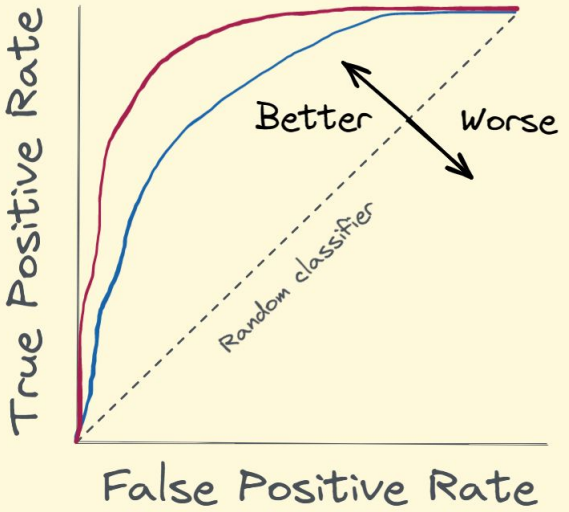
$$\text{Accuracy} = \frac{TP + TN}{\text{Total Samples}}$$

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

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

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

$$\text{Log Loss} = -\sum_{k}^{K} y^{(k)} \log(p^{(k)})$$



Models

1. **Weak Baseline** - Majority classifier

2. **Strong Baseline** - Unimodal (text-only and image-only) baselines

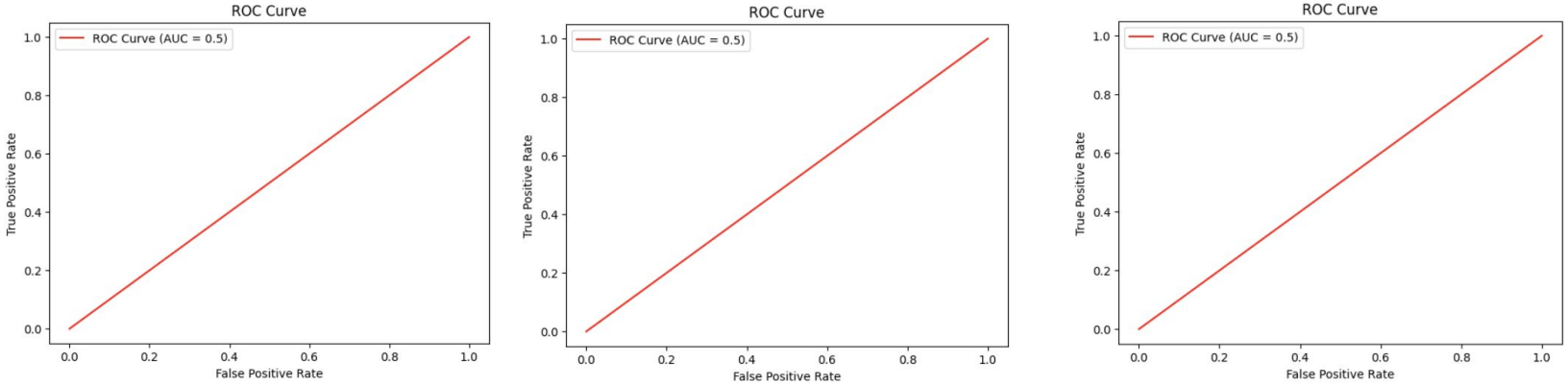
3. **Extension I** - Comparison of 2 embedding strategies :
   a. **CLIP** (Contrastive Language-Image Pre-Training) image and text embeddings followed by elastic net
   b. **BLIP** (Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation) generated image captions appended to tweets followed by sentence-transformer embeddings and elastic net

4. **Extension II** - Attention model with CLIP embeddings and disaster event information

Weak Baseline: Majority Class Classifier

The weak baseline relies on the majority class classifier, predicting labels based solely on data distribution. Among the 13608 examples, 8341 were informative tweets, resulting in an accuracy of approximately 61%.

Metric	Train	Dev	Test
Accuracy	61%	63%	61%
Macro-F1 Score	0.38	0.39	0.38



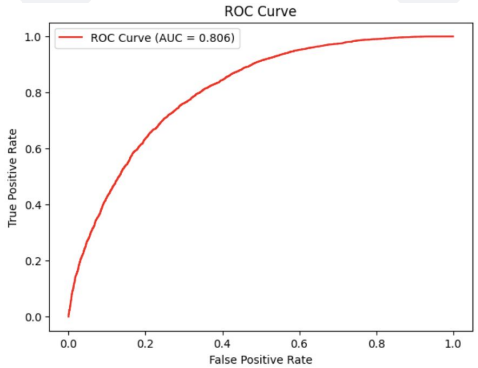
Penn Engineering

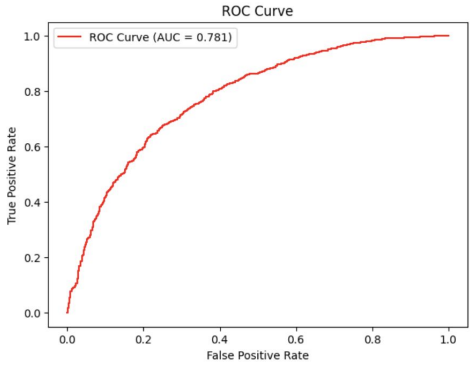
Figure 1: Simple baseline classification report and ROC curve for train (left), dev (middle), and test (right) data

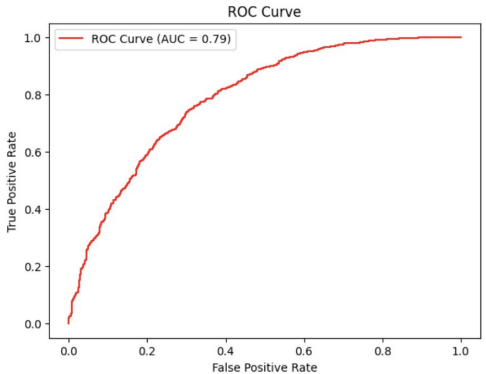
Strong Baselines: Text-Only Sentence-Transformer

The text-only strong baseline employs the 'all-MiniLM-L6-v2' Sentence Transformer for generating embeddings from tweet texts, leading to significant enhancements in accuracy and F1 scores. A logistic regression classifier is trained on these embeddings.

Metric	Train	Dev	Test
Accuracy	75.4%	73.6%	74.3%
Macro-F1 Score	0.73	0.70	0.71







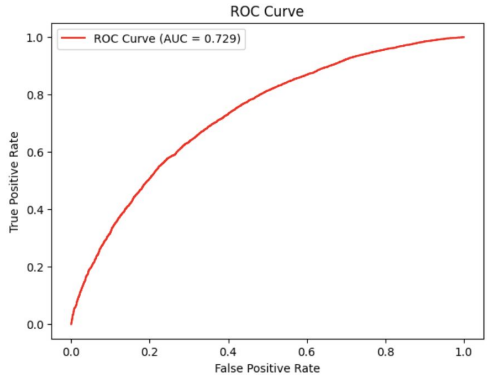
Penn Engineering

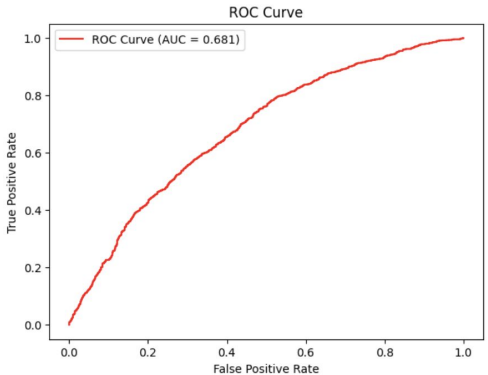
Figure 2: Strong baseline classification report and ROC curve for train (left), dev (middle), and test (right) data

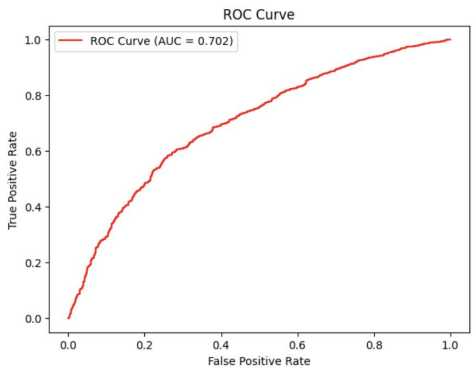
Strong Baseline: Image-Only VGG16 Model

Image-only baseline involves fine-tuning the classification head of a pre-trained VGG16 model with ImageNet weights. This model, augmented with an additional fully-connected layer and ReLU activation, is optimized using cross-entropy loss and an Adam optimizer. While its performance trails the text-only baseline, it still represents a notable improvement over simpler models.

Metric	Train	Dev	Test
Accuracy	68.2%	67.3%	66.2%
Macro-F1 Score	0.60	0.58	0.58





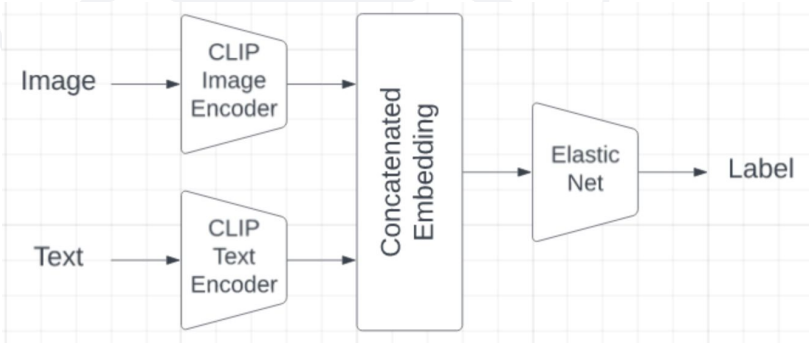


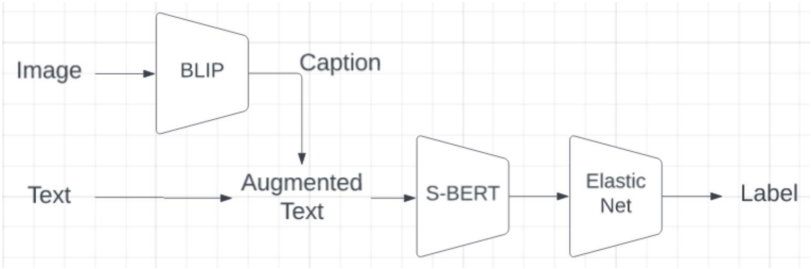
Penn Engineering

Figure 2: Strong baseline classification report and ROC curve for train (left), dev (middle), and test (right) data

Extension 1: Embedding Generation Strategies

Embedding Strategies: We experimented with CLIP-Concat (left) and Text+Captions+S-BERT (right) for generating embeddings.



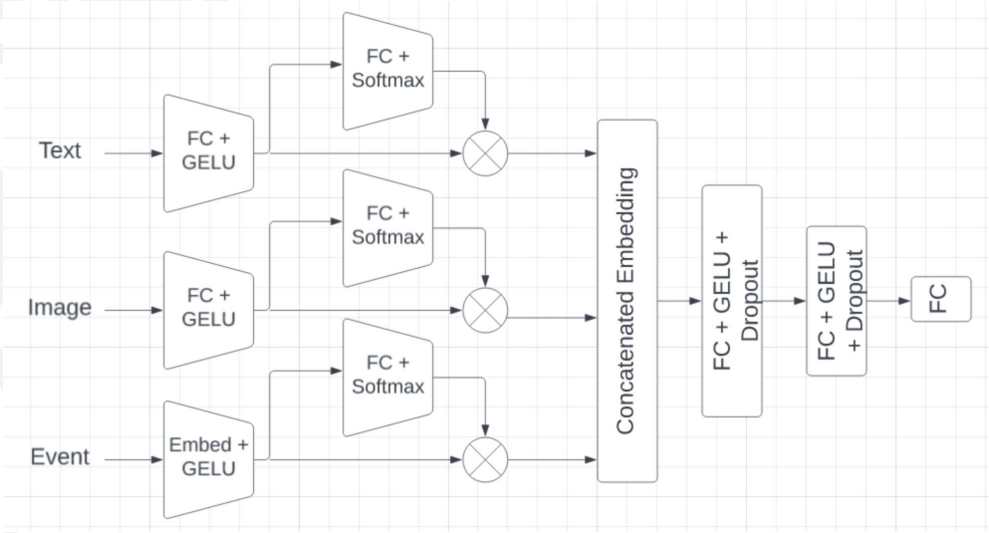


Extension 1: Embedding Generation Strategies

Embeddings	Data	Accuracy	Precision	Recall	F-Score	AUC
CLIP-Concat	Train	0.829	0.82	0.81	0.82	0.909
CLIP-Concat	Dev	0.806	0.79	0.79	0.79	0.869
CLIP-Concat	Test	0.785	0.77	0.77	0.77	0.860
Text + Captions + S-BERT	Train	0.783	0.78	0.76	0.76	0.850
Text + Captions + S-BERT	Dev	0.759	0.75	0.72	0.73	0.829
Text + Captions + S-BERT	Test	0.766	0.76	0.74	0.74	0.837

- **Performance Dominance:** CLIP-Concat consistently outperforms Text + Captions + S-BERT.
- **Baseline Surpassing:** Both models surpass the simple and strong baseline, emphasizing the effectiveness of embeddings in the classification task.

Extension 2: Attention Model + Events



Metric	Train	Dev	Test
Accuracy	82.4%	80.3%	78.7%
Macro-F1 Score	0.81	0.79	0.77

Conclusion & Future Work

Experimented with unimodal and multimodal approaches to disaster tweets classification. Our best performing model was an attention model with a test accuracy of 0.787 and an F-score of 0.773

In the future, could…

- Experiment with different data augmentation strategies to increase dataset size and reduce overfitting
- Explore alternative approaches to attention

References

Firoj Alam, Muhammad Imran, and Ferda Ofli. 2017. Image4act: Online social media image processing for disaster response. In 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 601–604. IEEE/ACM.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In Proceedings of the 12th International AAAI Conference on Web and Social Media(ICWSM).

Ferda Ofli, Firoj Alam, and Muhammad Imran. 2020. Analysis of social media data using multimodal deep learning for disaster response. In 17th International Conference on Information Systems for Crisis Response and Management. ISCRAM, ISCRAM.

A. Santhanavijayan Srinivasulu Kothuru. 2023. Identifying covid -19 english informative tweets using limited labelled data. In Social Network Analysis and Mining.