# NATURAL LANGUAGE PROCESSING WITH DISASTER TWEETS

**Final Project Report**

**ISM 6930 - Tech Foundation of AI**

**Submitted by**

| | |
|---|---|
| Uma Srikanth Reddy Koduru | U94125452 |
| Muralidhar Reddy Reddem | U64546777 |

**Major in**

**BUSINESS ANALYTICS AND INFORMATION SYSTEMS**

**Under the guidance of**

**DR. TENGTENG MA**

**MUMA COLLEGE OF BUSINESS**

**UNIVERSITY OF SOUTH FLORIDA**

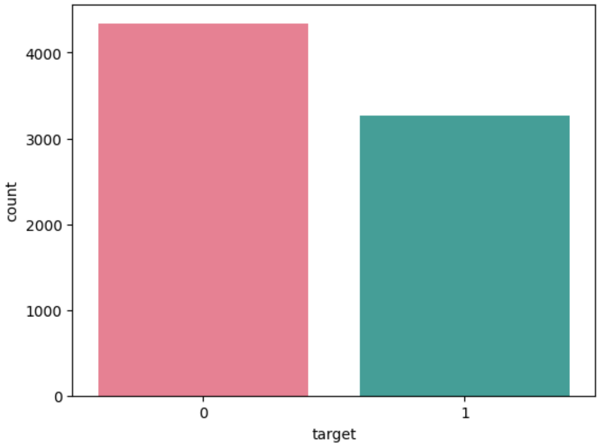
# Table of Contents

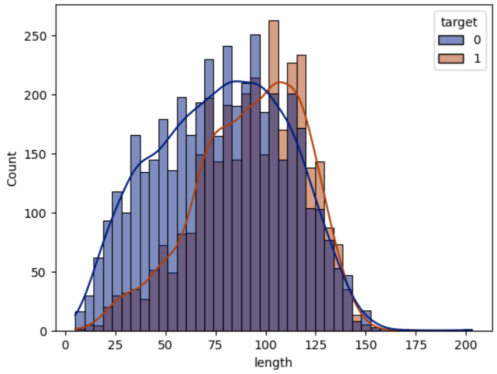
## 1. Project Idea and Competition Rationale

Twitter has emerged as a crucial means of communication during emergency situations. The widespread use of smartphones allows individuals to promptly report emergencies they witness, leading to increased interest from various organizations, such as disaster relief groups and news outlets, in systematically monitoring Twitter. Business Scope: Government Agencies, Insurance Sector

## 2. Data Exploration

Data exploration is a critical phase in the context of text classification within the Kaggle Competition: Natural Language Processing with Disaster Tweets, which attracted participation from 981 teams. The dataset in question comprises a total of 10,000 tweets, encompassing both training and testing data, with each instance characterized by fields such as 'id,' 'keyword,' 'location,' and 'text.' Notably, there are 221 unique keywords associated with the tweets, and 61 instances with null keyword values, while no null values are observed in the 'text' column. These statistics provide an initial overview of the dataset, serving as a foundation for subsequent analytical and modeling efforts in this competition.

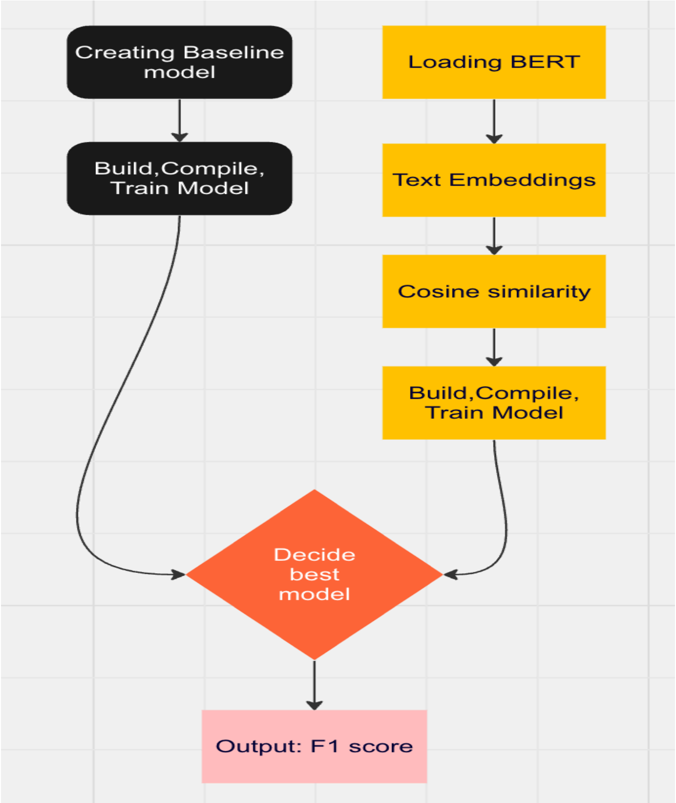




## 3. Model Selection

Methods used:

- LSTM – Considered for baseline.
- RoBERTa- chose the base model.



## LSTM

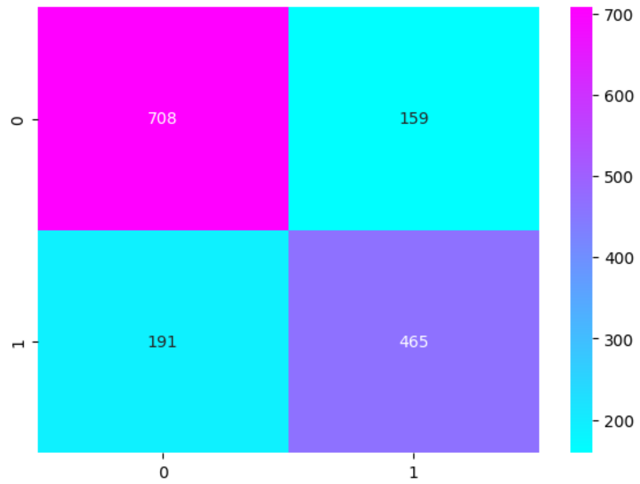
- Applied different techniques to remove unwanted content in tweets
- Converted tweets to tokens to word2vec [embeddings]
- Applied LSTM model with act=sigmoid, drop=0.2, loss=binary cross entropy,
- Got f1 score .79

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

|  |  | Actual Values |  |
|---|---|---|---|
|  |  | Positive (1) | Negative (0) |
| Predicted Values | Positive (1) | TP | FP |
|  | Negative (0) | FN | TN |

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

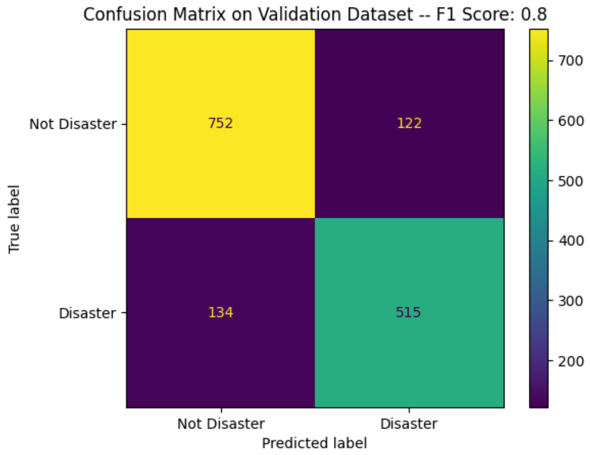
$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

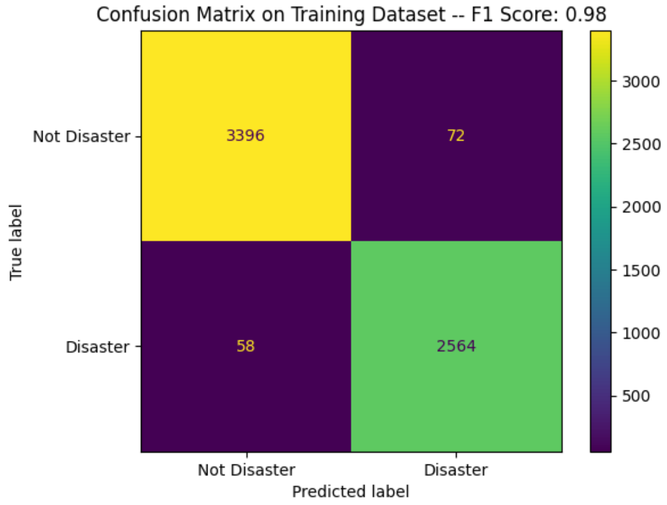


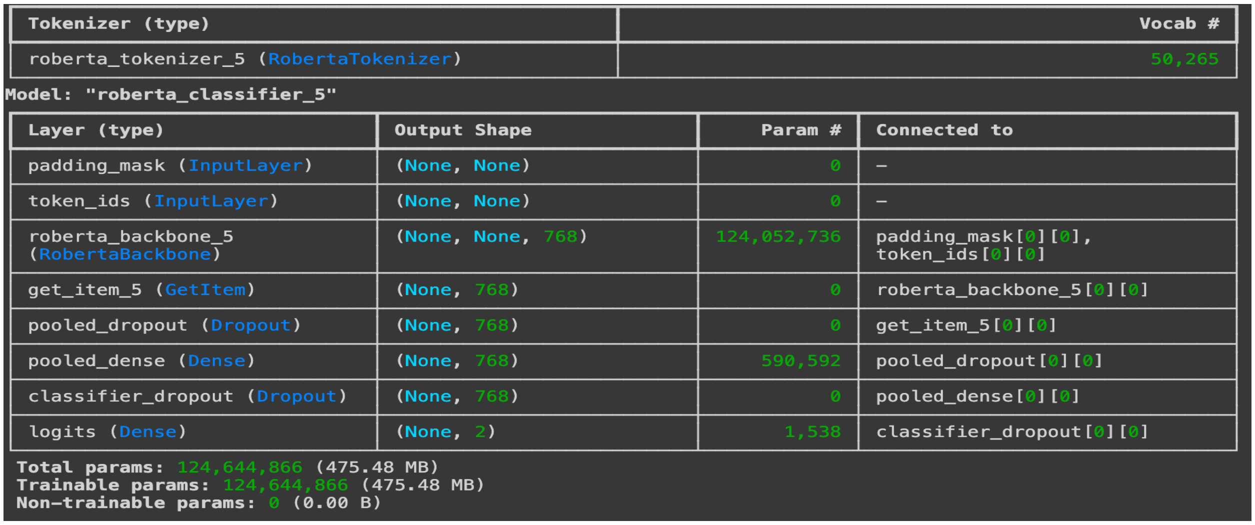
## RoBERTa – BASE

RoBERTa-Base is a variant of the RoBERTa (A Robustly Optimized BERT Pretraining Approach) model, specifically fine-tuned for English language processing tasks. This model is built upon the Transformer architecture, a state-of-the-art deep learning model for natural language processing (NLP). It utilizes a combination of pre-training objectives, including the Masked Language Model (MLM), where it learns to predict masked words in sentences, and the Next Sequence Prediction (NSP), which involves predicting the likelihood of the next sentence in a given text. These dual pre-training tasks help RoBERTa-Base develop a comprehensive understanding of contextual language representations.

RoBERTa-Base has achieved a validation accuracy of 0.84, indicating its high performance in various NLP tasks. This level of accuracy demonstrates its ability to capture intricate language nuances and provide reliable results in applications such as text classification, sentiment analysis, and text generation. It serves as a robust foundation for English language understanding, thanks to its extensive training on a substantial corpus of text data, enabling it to handle a wide range of linguistic challenges.



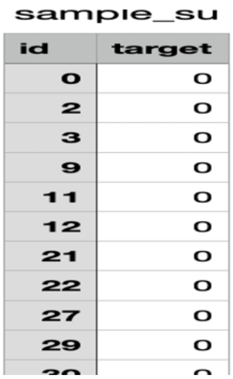




```
┏━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Tokenizer (type)                                        ┃                                                   Vocab # ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│ roberta_tokenizer_5 (RobertaTokenizer)                  │                                                    50,265 │
└─────────────────────────────────────────────────────────┴───────────────────────────────────────────────────────────┘
Model: "roberta_classifier_5"
┏━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Layer (type)                     ┃ Output Shape                 ┃      Param # ┃ Connected to                       ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│ padding_mask (InputLayer)        │ (None, None)                 │            0 │ -                                  │
├──────────────────────────────────┼──────────────────────────────┼──────────────┼────────────────────────────────────┤
│ token_ids (InputLayer)           │ (None, None)                 │            0 │ -                                  │
├──────────────────────────────────┼──────────────────────────────┼──────────────┼────────────────────────────────────┤
│ roberta_backbone_5               │ (None, None, 768)            │  124,052,736 │ padding_mask[0][0],                │
│ (RobertaBackbone)                │                              │              │ token_ids[0][0]                    │
├──────────────────────────────────┼──────────────────────────────┼──────────────┼────────────────────────────────────┤
│ get_item_5 (GetItem)             │ (None, 768)                  │            0 │ roberta_backbone_5[0][0]           │
├──────────────────────────────────┼──────────────────────────────┼──────────────┼────────────────────────────────────┤
│ pooled_dropout (Dropout)         │ (None, 768)                  │            0 │ get_item_5[0][0]                   │
├──────────────────────────────────┼──────────────────────────────┼──────────────┼────────────────────────────────────┤
│ pooled_dense (Dense)             │ (None, 768)                  │      590,592 │ pooled_dropout[0][0]               │
├──────────────────────────────────┼──────────────────────────────┼──────────────┼────────────────────────────────────┤
│ classifier_dropout (Dropout)     │ (None, 768)                  │            0 │ pooled_dense[0][0]                 │
├──────────────────────────────────┼──────────────────────────────┼──────────────┼────────────────────────────────────┤
│ logits (Dense)                   │ (None, 2)                    │        1,538 │ classifier_dropout[0][0]           │
└──────────────────────────────────┴──────────────────────────────┴──────────────┴────────────────────────────────────┘
 Total params: 124,644,866 (475.48 MB)
 Trainable params: 124,644,866 (475.48 MB)
 Non-trainable params: 0 (0.00 B)
```

## 4. Findings

Findings from the project include the primary output, which involves predicting whether a given tweet pertains to a real disaster (1) or not (0) in the target column using an evaluation metric. The model achieved a commendable F1 score of 0.837, indicating its effectiveness in distinguishing between real disaster-related tweets and those that are not, underlining its strong performance in this classification task.



sample_su

| id | target |
|---|---|
| 0 | 0 |
| 2 | 0 |
| 3 | 0 |
| 9 | 0 |
| 11 | 0 |
| 12 | 0 |
| 21 | 0 |
| 22 | 0 |
| 27 | 0 |
| 29 | 0 |
| 30 | 0 |

## 5. Future Scope

The future scope of this project includes two key aspects: Model Deployment on AWS, facilitating scalable and cost-effective utilization of RoBERTa-Base for NLP tasks. Additionally, UI Implementation will create an intuitive interface for user-friendly interactions, making the model accessible to a broader audience, enhancing its usability. These initiatives collectively aim to extend the model's reach and utility, benefiting businesses and individuals seeking advanced NLP capabilities.

## 6. Challenges

Our findings have revealed substantial opportunities for enhancing the system's capabilities. These prospects encompass broadening its scope to accommodate multiple food chains, in addition to the current coffee chains. This strategic expansion is expected to yield considerable cost and time savings for business owners, while concurrently mitigating apprehensions related to data size and overfitting. Furthermore, we have identified a critical need for comprehensive data cleaning procedures to ensure the quality and reliability of our dataset. Additionally, the utilization of BERT Large, a model endowed with a formidable 340 million parameters, has proven to be instrumental in enhancing the system's natural language processing capabilities. Finally, we have explored solutions to effectively manage the computational challenges associated with such a substantial model, ensuring the system's efficiency and scalability.

## 7. Conclusion

In summary, the RoBERTa language model offers significant potential for conducting sentiment analysis on Twitter. It delivers precise and dependable outcomes, enabling businesses and researchers to develop a more profound insight into public sentiment across various subjects. **Got 85th rank in Kaggle competition out 1123 teams.**

## 8. Lessons learnt

While sentiment analysis is indeed a potent method, one of its primary limitations lies in its proclivity to generate inconsistent or potentially prejudiced outcomes. This issue may arise when the model's training data is imbalanced or carries inherent biases, resulting in a distorted comprehension of what qualifies as positive or negative sentiment.

## 9. References

- Francois Chollet. Deep learning with Python. Manning Publications Co., 2017.
- An Introduction to the Unified Modeling Language © Laurie Williams 2004