TWEET CLASSIFICATION

REAL DISASTER OR NOT?

Fabulous Tokenizers

01

Isaac Byron

02

Laura Kinya

03

Stella Njeri

04

Precy Mae

Introduction

Problem Statement

There are numerous challenges when considering the use of social media data for emergency response on disasters.

Machine learning can be used to help responders and relief by identifying real and false disasters.

Study Objective

The aim of this project is build a model predict which tweets are real disasters and which tweets are not.

Data Sourcing

Source

The datasets are from the company figure-eight and originally shared on their 'Data For Everyone' website.

Datasets

The train dataset has 7613 rows and 5 columns.
The Test dataset has 3263 rows and 4 columns.

Work Plan

SPRINT 1

Business understanding & Task allocation.
Data cleaning and data exploration.
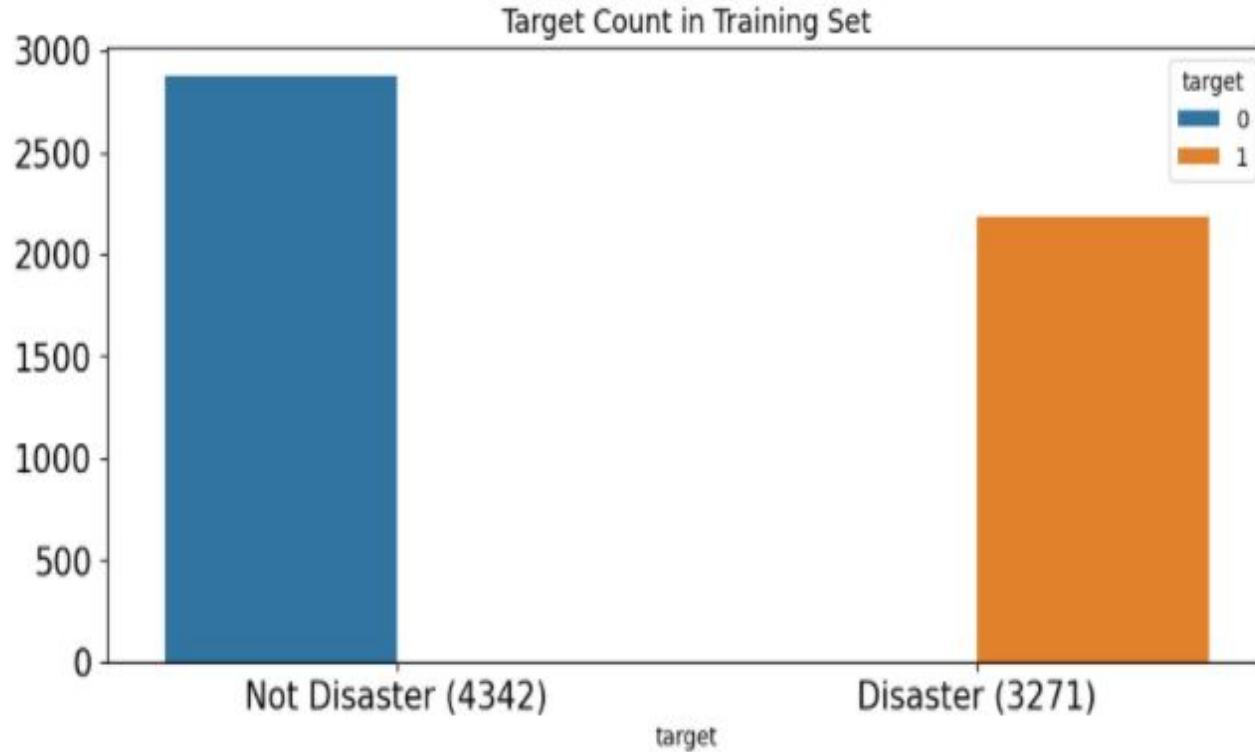
SPRINT 2

Modelling.

SPRINT 3
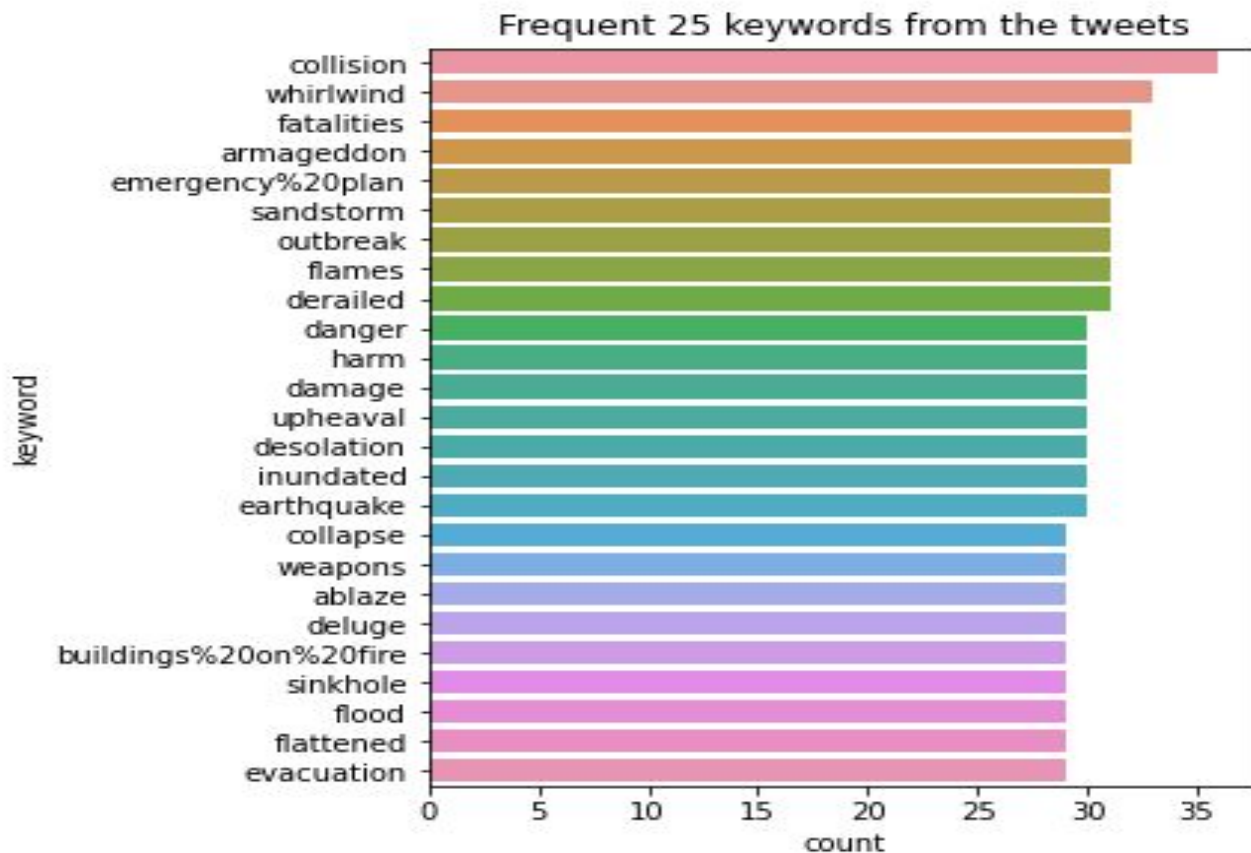
Model deployment.

Exploratory Data Analysis

Distribution of Target Variable

Common Keywords
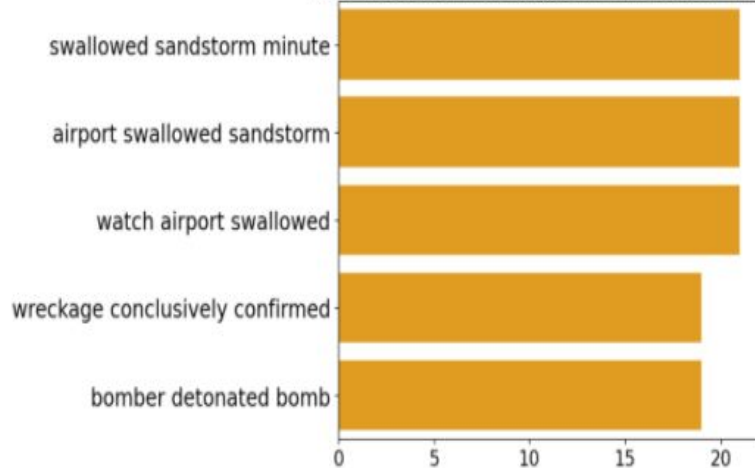

Frequent 25 keywords from the tweets

Common Trigrams

Top 5 most common trigrams in Disaster Tweets

- swallowed sandstorm minute
- airport swallowed sandstorm
- watch airport swallowed
- wreckage conclusively confirmed
- bomber detonated bomb

Top 5 most common trigrams in Non-disaster Tweets

- liked youtube video
- reddit quarantine offensive
- cross body bag
- quarantine offensive content
- wind neighbour ass
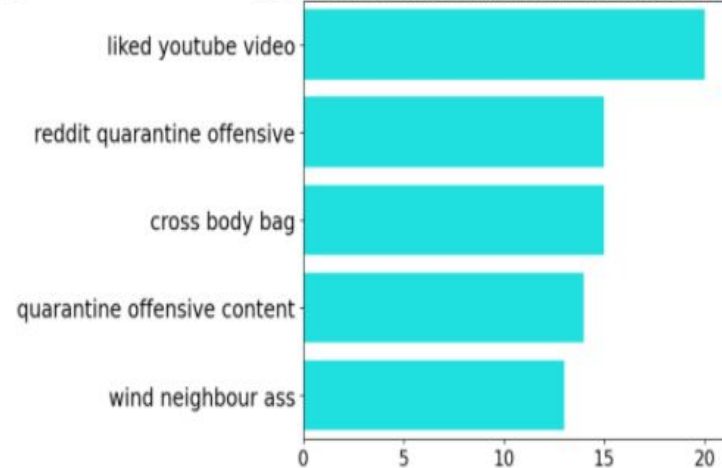
Common Bi-grams

Top 5 most common bigrams in Disaster Tweets

bigram	
suicide bomber	
oil spill	
burning buildings	
sandstorm minute	
swallowed sandstorm	
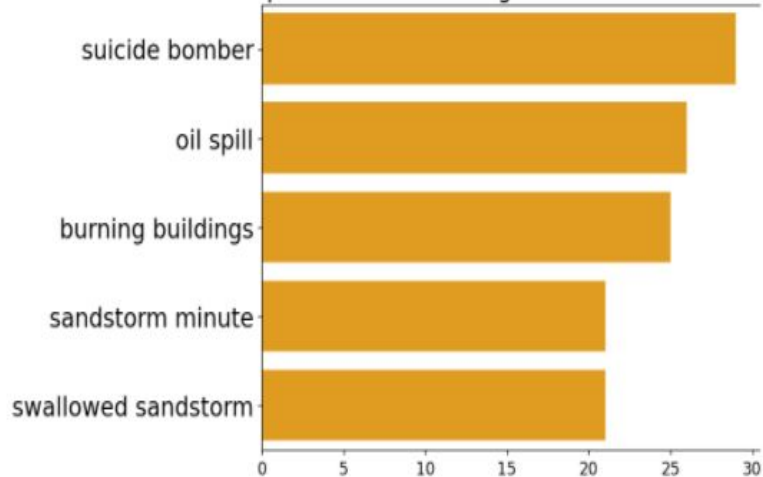
Top 5 most common bigrams in Non-disaster Tweets

bigram	
cross body	
full read	
youtube video	
liked youtube	
body bags	
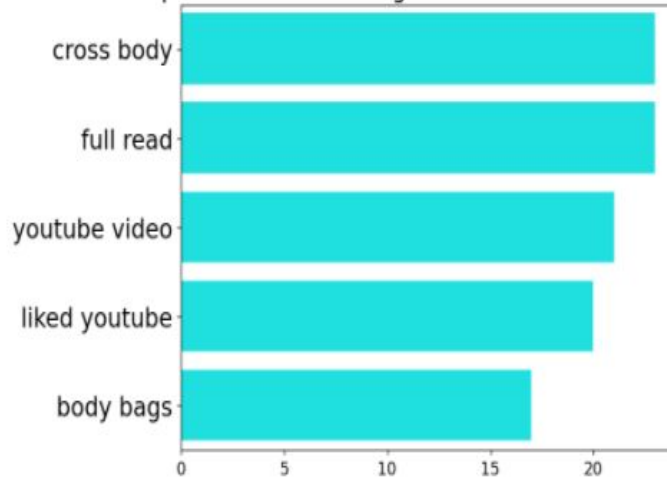
Word Cloud

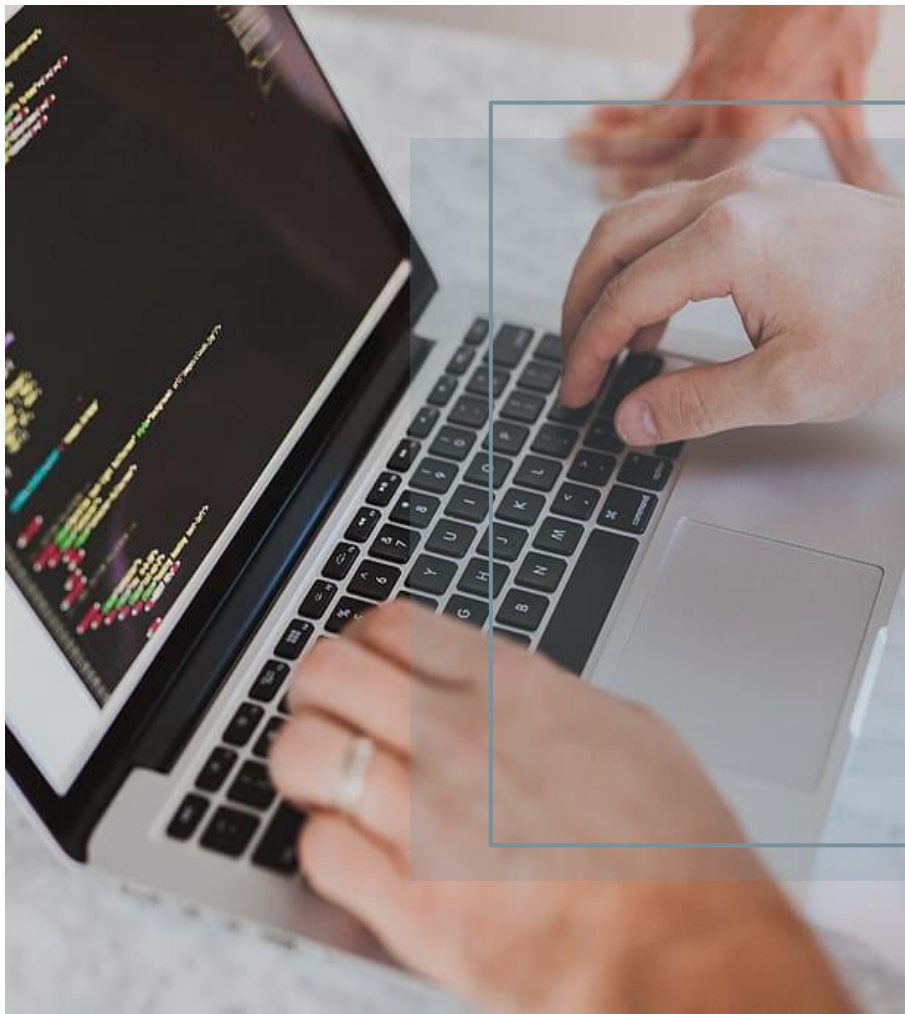## Word Cloud Disaster

Word cloud not disaster

Modelling

Modelling

Naive Bayes Model

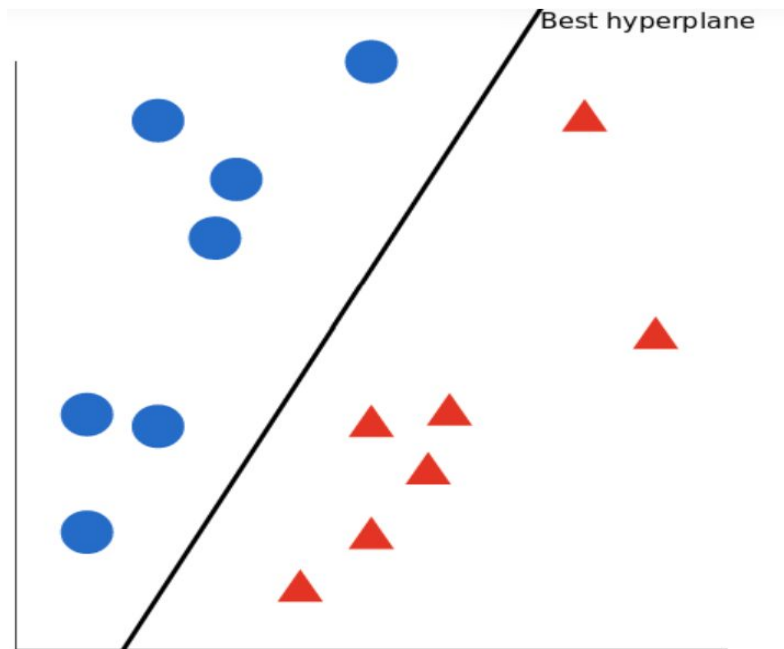Linear SVM Model

Flair Model

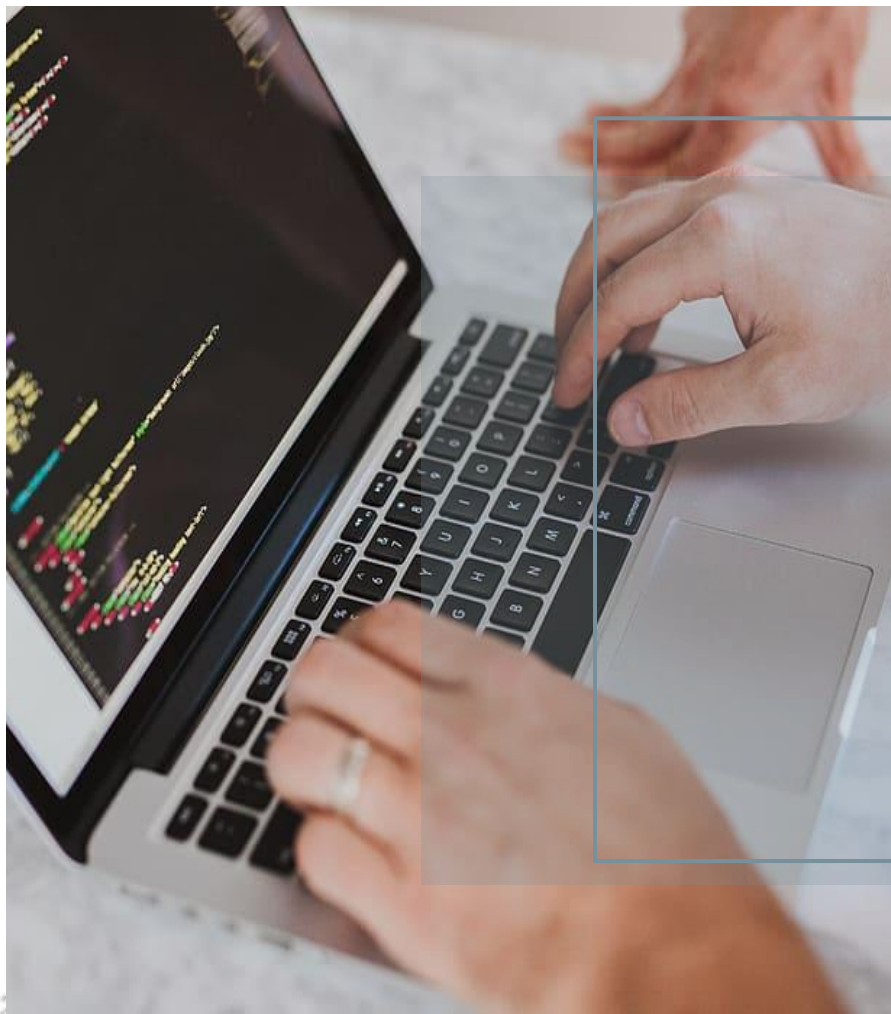BERT Model

Naive Bayes Model

Our Naive Bayes model classified tweets with an accuracy of 79.25% as disaster and non- disaster tweets.

Linear SVM

The Linear Support Vector Machine Model classified tweets with an accuracy of 79.77% as disaster and non-disaster tweets.
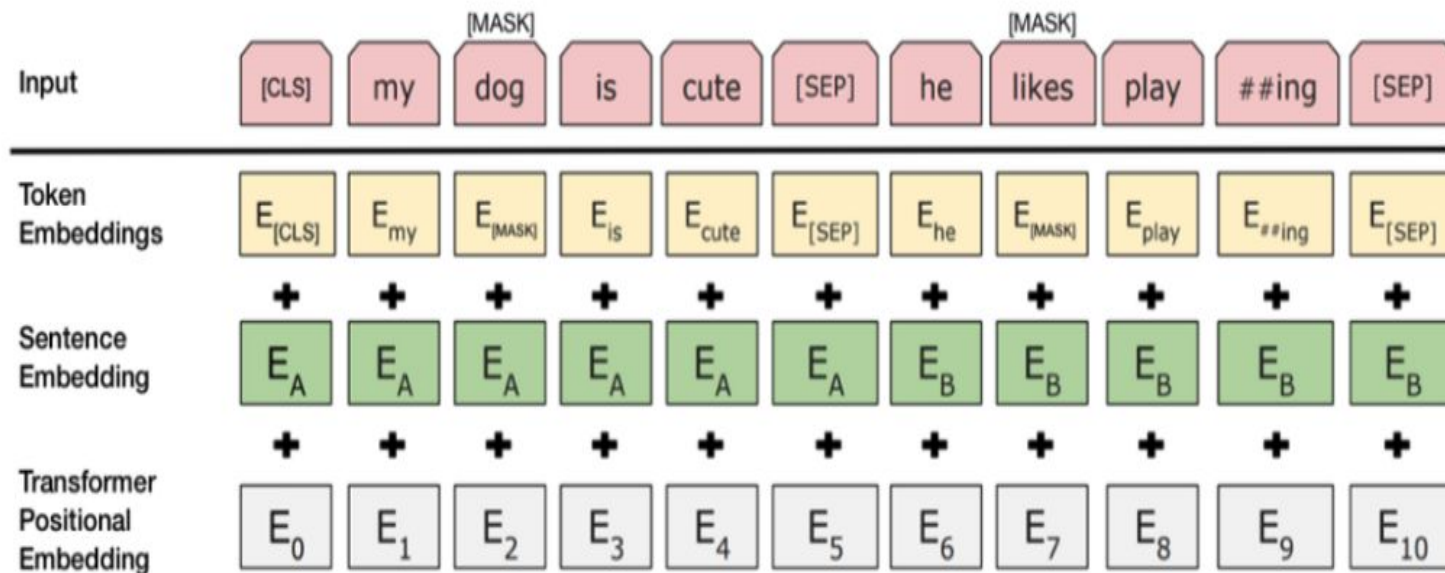

Best hyperplane

BERT Model

BERT is a transformers model pre-trained on a large corpus of English data in a self-supervised fashion.

It classified tweets in the test data with an accuracy 83%.
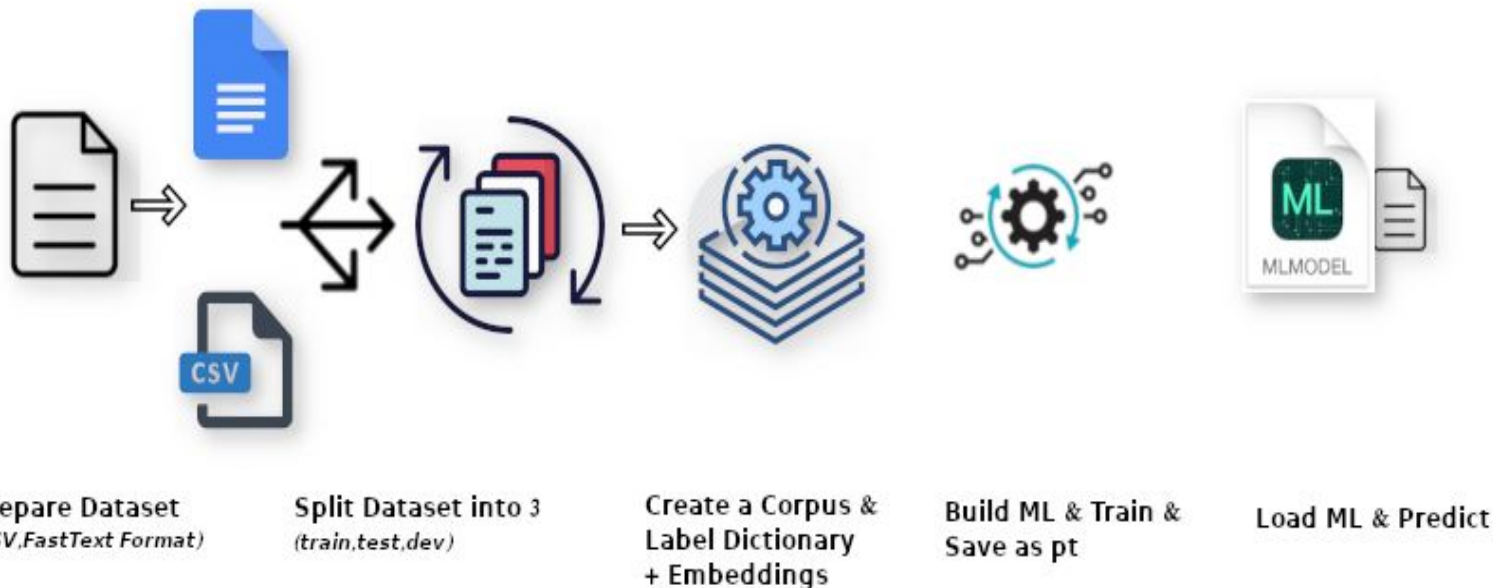
BERT Model

Flair Model

Our model was able to classify tweets in the test data with an accuracy of 79.19%.

Text Classification with Flair

Workflow

Prepare Dataset
(CSV,FastText Format)

Split Dataset into 3
(train,test,dev)

**Create a Corpus &
Label Dictionary
+ Embeddings**

**Build ML & Train &
Save as pt**

Load ML & Predict

Disaster Tweet Monitor

Are you wondering if A tweet is Real Disaster Tweet Or Not ???
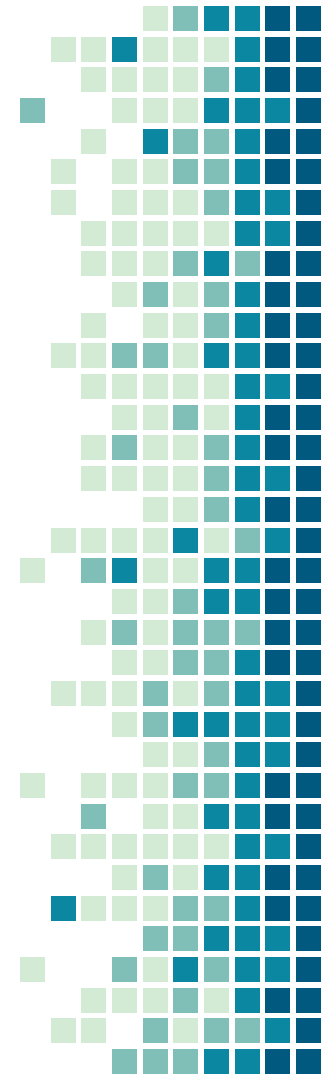
CLASSIFY SINGLE TWEET or #HASH TAGS

Single Tweet ▾

Single tweet classification

Tweet:

wreckage conclusively confirmed and many people were injured

Prediction:

Disaster with 85.95282435417175 % confidence

Disaster Tweet Monitor

Are you wondering if A tweet is Real Disaster Tweet Or Not ???

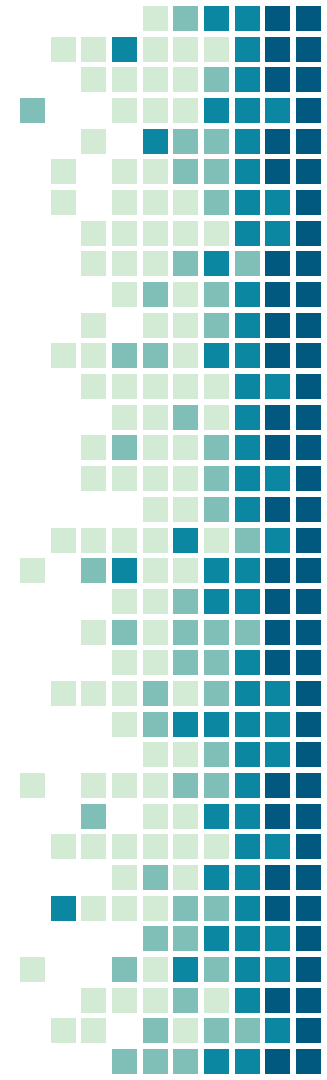CLASSIFY SINGLE TWEET or #HASH TAGS

Single Tweet

Single tweet classification

Tweet:

your hair and nails are a disaster, The teacher will kill you

Prediction:

No_Disaster with 80.29524683952332 % confidence

Conclusion

Conclusion

- Bert transformer model and Flair model have performed better than the ordinary classification models.

- The non- disaster tweets are more than disaster tweets.

Use Case

- Our work can be used by disaster relief organizations such as Red Cross and World Relief.
- News agencies can use the model to report real disaster in real time.
- Classifying other tweet types like real/fake news.

Challenges Faced

- Model size.

- Training Time.

Thanks!

ANY QUESTIONS?