Natural Language Processing with Disaster Tweets

MGMT 59000 Machine learning Final Project

Present by Hsueh-Ning Chao, Shih-Yu Wang

Data Overview

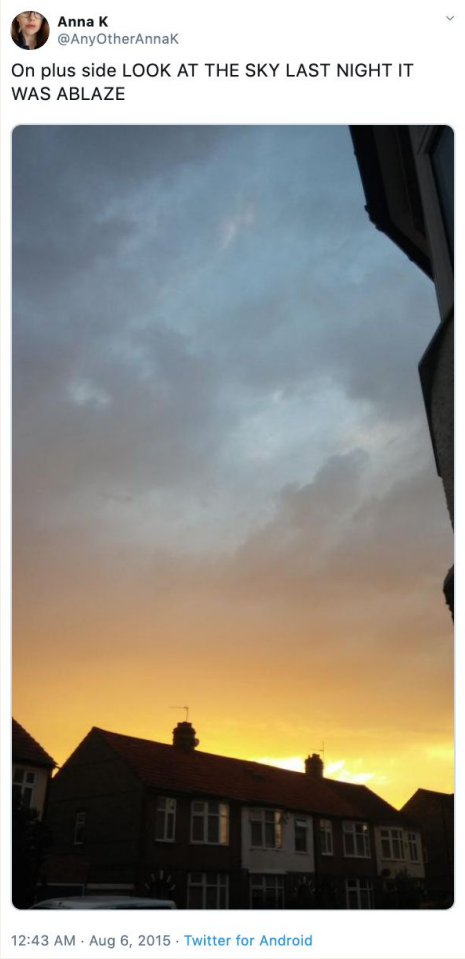
Average Text Length: 14.9 vocabularies

Missing Values:

- ID: 0
- Keyword: 61 [object]
- Location: 2534 [object]
- text: 0 [object]
- target: 0 [int]

Goal

Predicts which Tweets are about **real disasters** and which one's aren't.



Text Processing Methods

- BoW: Ignores word meaning, order, and context
- TF-IDF: Still ignores word order and meaning
- ✅ **Word Embeddings(Word2Vec)**: Captures semantic meaning, relationships between words
- ✅ **Tokenization & Embedding**: Vocabulary-dependent

Data Engineering

Using Multiple inputs **did not significantly improve the performance:**

- text + location + keywords: Accuracy 0.699
- text + keywords: Accuracy 0.57
- **ONLY TEXT: 0.79 !**

>> Text itself is enough for prediction, other informations provide noises

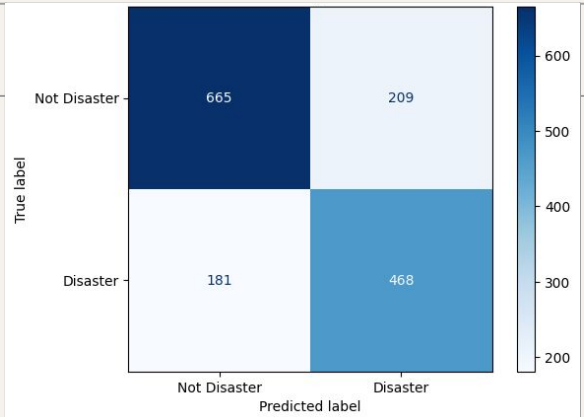
Model Comparison

Model	Features	Advantages	Acc
LSTM	• Help capture long-term dependencies	• Suitable for time series data • Reduces vanishing gradient issue.	0.767
BiLSTM	• Process data both forward and backward directions	• Captures richer contextual information	0.793
DistilBERT	• Use knowledge distillation • Reduce BERT's weights by 40%	• Faster training and inference • Lightweight for limited resources	0.831
RoBERTa	• Use dynamic mask • Removes Next Sentence Prediction (NSP) • Train with more batch	• More powerful • state-of-the-art performance in many NLP tasks	0.841

Confusion Matrix

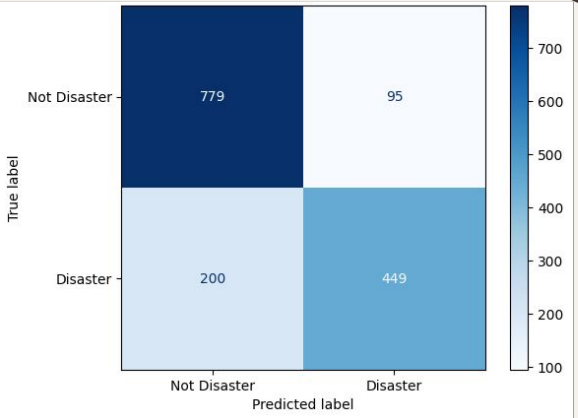
LSTM

F1: 0.71



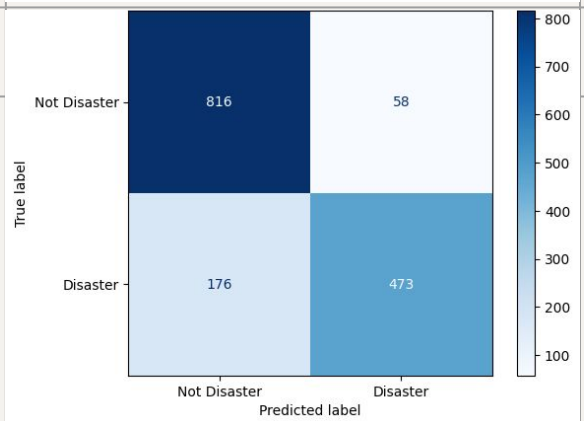
BiLSTM

F1: 0.75



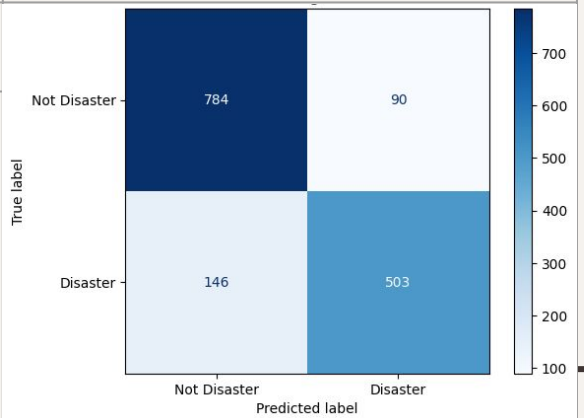
DistilBERT

F1: 0.8



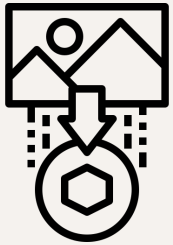
RoBERTa

F1: 0.81

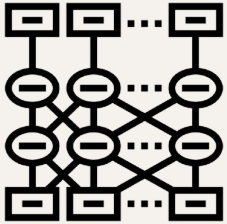


Lesson Learned

More training data does not necessarily mean better performance



Tokenizer + Embedding performs better than other text transformation methods



Transformer Models outperform Traditional RNNs

Thanks