Disaster Tweets Classification

Group #2

Manasa Hari (013855075)
Feiyu Cai (008879234)
Tushar Sharma (009803885)
Amit Kamboj (013827489)

Introduction

- 500 million tweets everyday
- Disaster tweets often get lost in the flood of tweets
- Irrevocable loss if disaster news didn’t reach out to people in the right time

Problem Statement

- To classify the tweets into real disaster tweets or not
- Highlight the disaster emergency news
- Users are notified on the emergency issues out of millions of other tweets

Data

1) train.csv - 7613 samples
2) test.csv - 3263 samples

Columns:

id, text, location, keyword, target (train.csv only)

Key Challenge

- Build a machine learning model
- Use NLP
- Sentiment analysis

Solution - Six steps

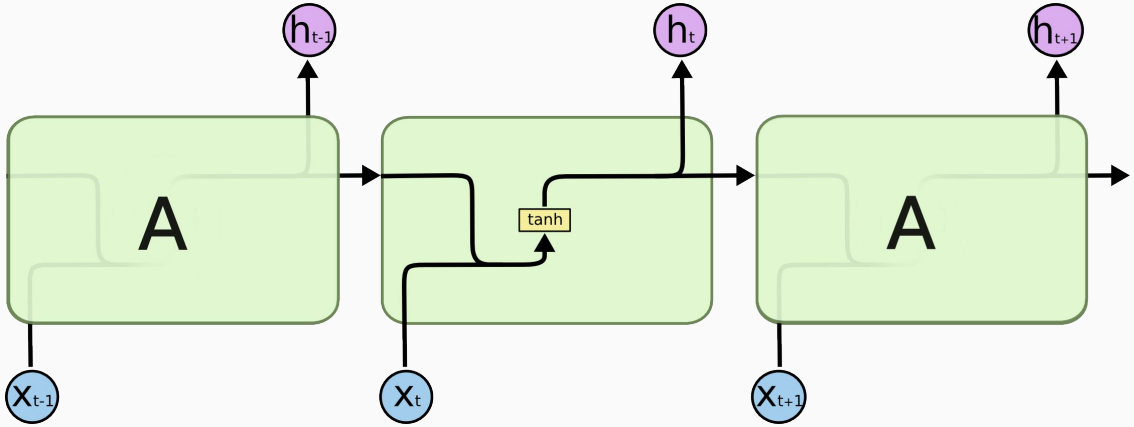
1) Study the data
2) Data pre-processing & data cleaning
3) Develop a model
4) Train the model
5) Test the model (make predictions)
6) We will fine-tune the model

State-of-art

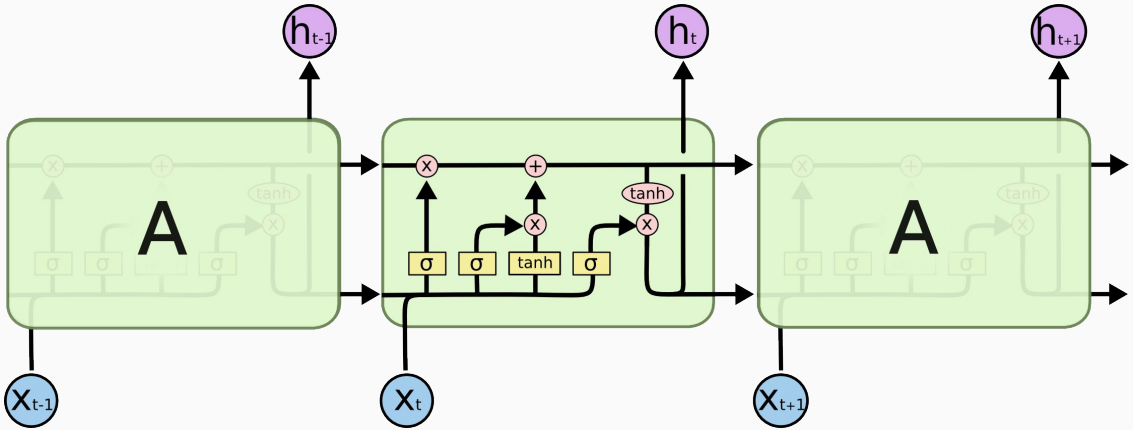
Why Recurrent Neural Network (RNN)?

- Words are related in each other
- Internal memory
- Shortage: Gradient vanishing
- Solution: Long Short-Term Memory (LSTM)

RNN - Diagrammatic representation



LSTM - Diagrammatic representation



Demo - in Python notebook

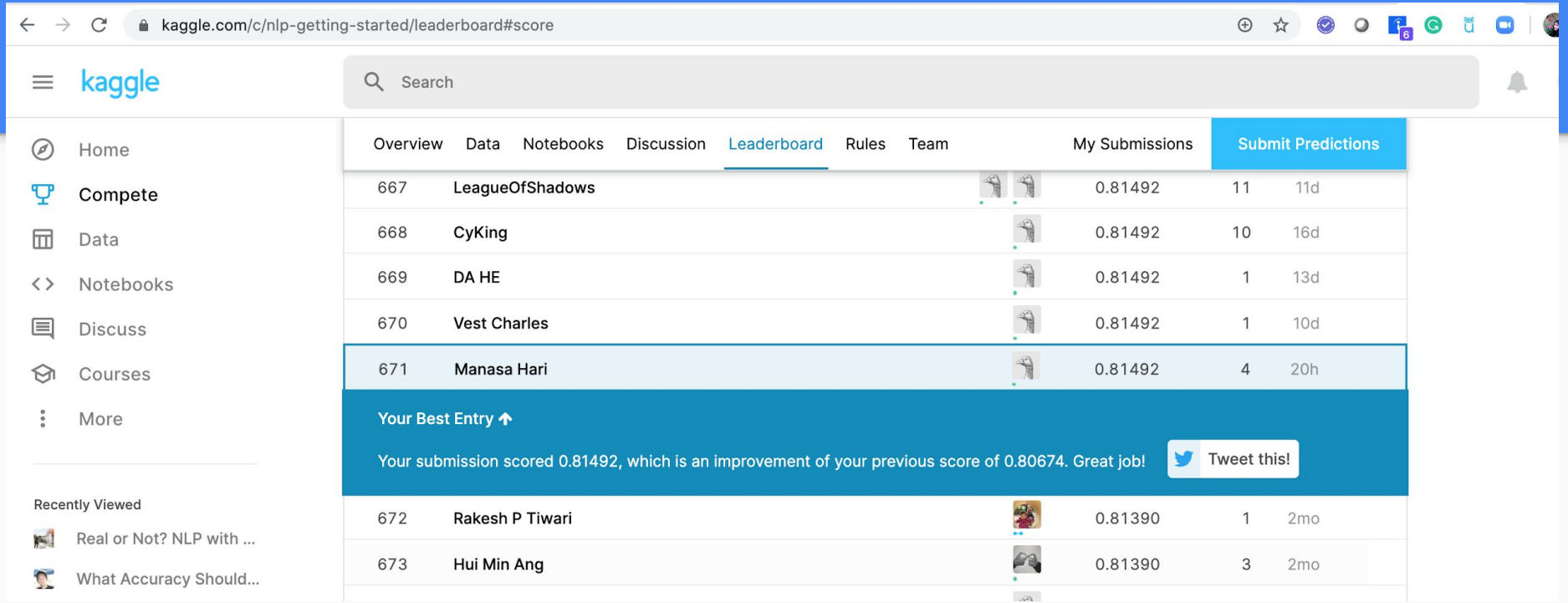
In the attached python notebook, we covered:

- Step by step process of this project
- Code (right from data pre-processing till predictions in Kaggle)
- Process of choosing optimal parameters for the LSTM model
- Performance analysis
- Visualization of the performance and comparison
- Kaggle scores at each step

Results

- Achieved 0.81492 accuracy in the ongoing Kaggle competition, by 4th submission itself.
- Standing at 671th position in the leaderboard out of a total of 2299 submissions.

Kaggle competition (ongoing)



Kaggle link: https://www.kaggle.com/c/nlp-getting-started/leaderboard#score

Thank you