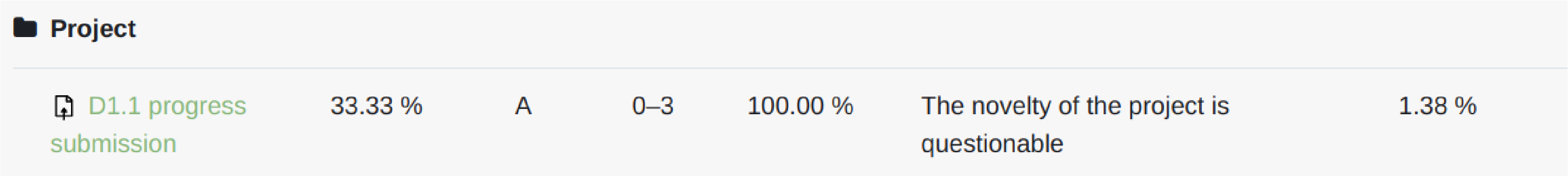
Real or Not? NLP with Disaster Tweets

[F23] PracticalMLandDL Project 2023

Roman Makarov & Adela Krylova



Project						
D1.1 progress submission	33.33 %	A	0–3	100.00 %	The novelty of the project is questionable	1.38 %

Real or Not? NLP with Disaster Tweets

[F23] PracticalMLandDL Project 2023

Roman Makarov & Adela Krylova

Disaster tweet Generation and Detection

[F23] PracticalMLandDL Project 2023

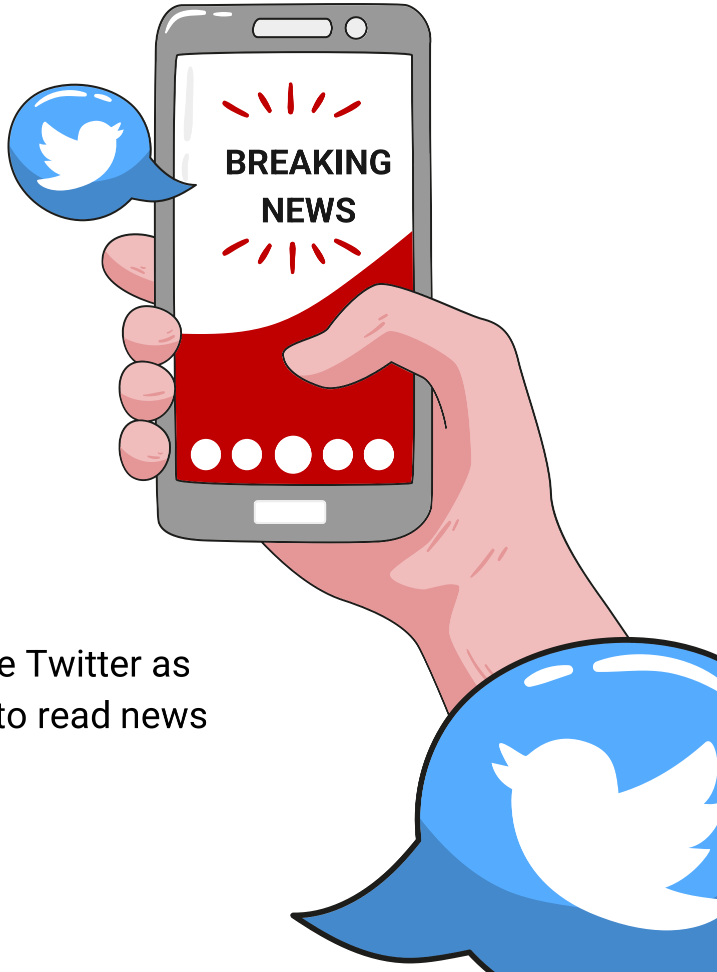
Roman Makarov & Adela Krylova

PROBLEM

19.4% Of active Twitter accounts are fake or spam

38.2% Americans have accidentaly shared fake news

69% Of Americans use Twitter as a main platform to read news



Fake tweets detection on Kaggle

- Competition on Kaggle “Natural Language Processing with Disaster Tweets”
- 10K of real and fake tweets dataset
- Best solution has 96% accuracy
- **However, the dataset is quite limited and it contains low quality fake tweets.**
- **Correct labels were leaked and utilized to mess up leaderboard**

Our solution

Generator

GPT2 pre-trained on a few real tweets and fine-tuned along with Discriminator model

Discriminator

Small Bert fine-tuned on GPT-generated data and 300 real tweets.

Examples of generated Tweets

Tweet1

“0Zionist infiltrator may be preparing for Armageddon0Linux-based 'back doors' to Iran's major”

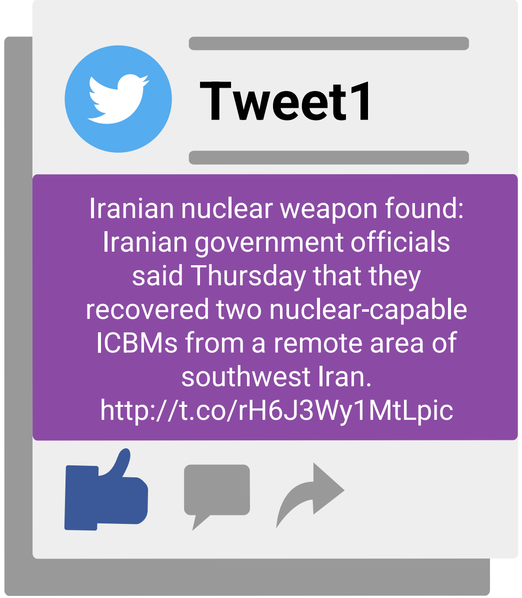
Tweet2

“A helicopter carrying miners from Myanmar to Bangladesh has collided with a boat off the Irian state's west coast.”

Tweet3

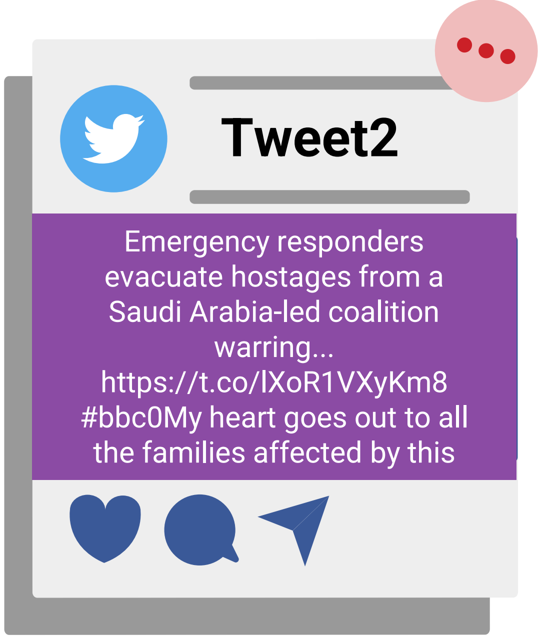
“I was in California and I saw a spaceship land on my front porch. How?? WHAT???”

Examples of tweets which tricked the detector



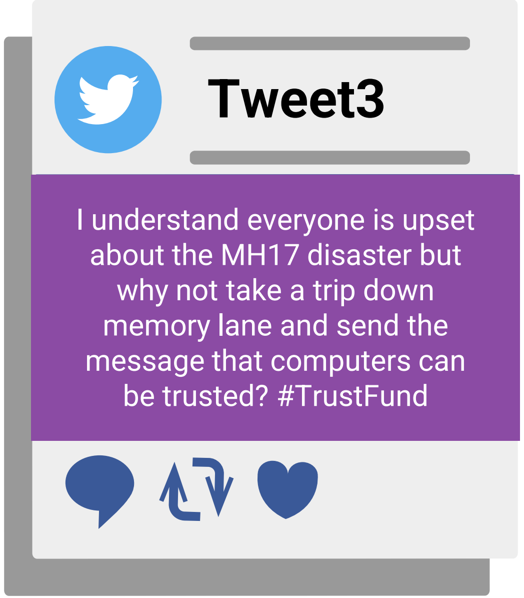
Tweet1

Iranian nuclear weapon found: Iranian government officials said Thursday that they recovered two nuclear-capable ICBMs from a remote area of southwest Iran. http://t.co/rH6J3Wy1MtLpic



Tweet2

Emergency responders evacuate hostages from a Saudi Arabia-led coalition warring... https://t.co/lXoR1VXyKm8 #bbc0My heart goes out to all the families affected by this



Tweet3

I understand everyone is upset about the MH17 disaster but why not take a trip down memory lane and send the message that computers can be trusted? #TrustFund

How good is our solution?

10%

Accuracy Increase

Was achieved after fine-tuning Discriminator and Generator

How good is our solution?

10%

Accuracy Increase

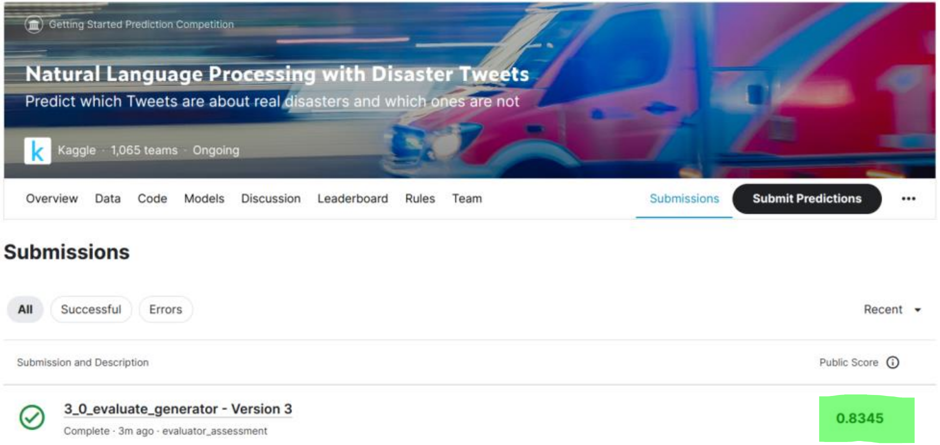
Was achieved after fine-tuning Discriminator and Generator

DistilBERT trained on 10000 samples: 0.84 accuracy

Discriminator trained on 300 real and 100 generated samples: 0.68 accuracy

With only 3% of data!

How good is our solution?



Getting Started Prediction Competition

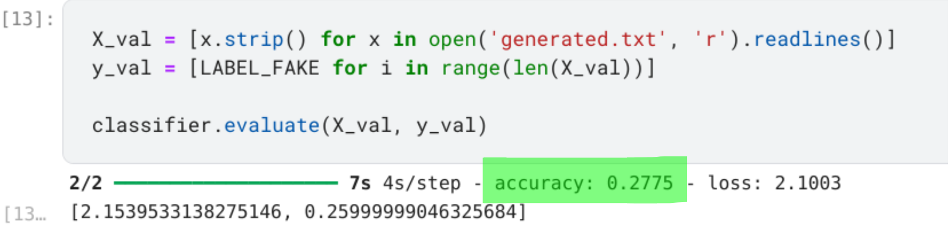
Natural Language Processing with Disaster Tweets

Predict which Tweets are about real disasters and which ones are not

Kaggle · 1,065 teams · Ongoing

Submissions

Submission and Description	Public Score
✓ 3_0_evaluate_generator - Version 3 Complete · 3m ago · evaluator_assessment	0.8345



```
[13]:  X_val = [x.strip() for x in open('generated.txt', 'r').readlines()]
       y_val = [LABEL_FAKE for i in range(len(X_val))]

       classifier.evaluate(X_val, y_val)

       2/2 ━━━━━━━━━━━━━━━━━━━━ 7s 4s/step - accuracy: 0.2775 - loss: 2.1003
[13…   [2.1539533138275146, 0.25999999046325684]
```

72%

Generations

Were able to fool one of the top Kaggle competition model

Limitations

Link generation

All generated links are not working; hence, the fakeness could be check by the links

Merged Tweets

Some generated Tweets comprise information of several disasters at the same time. This tricks the discriminator

Unstable convergence

Time and memory

The solution consume quite a large number of memory and time

Thank you for your attention

"Times change! Before, no one believed the weather forecast.
Today in the news only the weather forecast can be trusted."

— Ljupka Cvetanova, Yet Another New Land