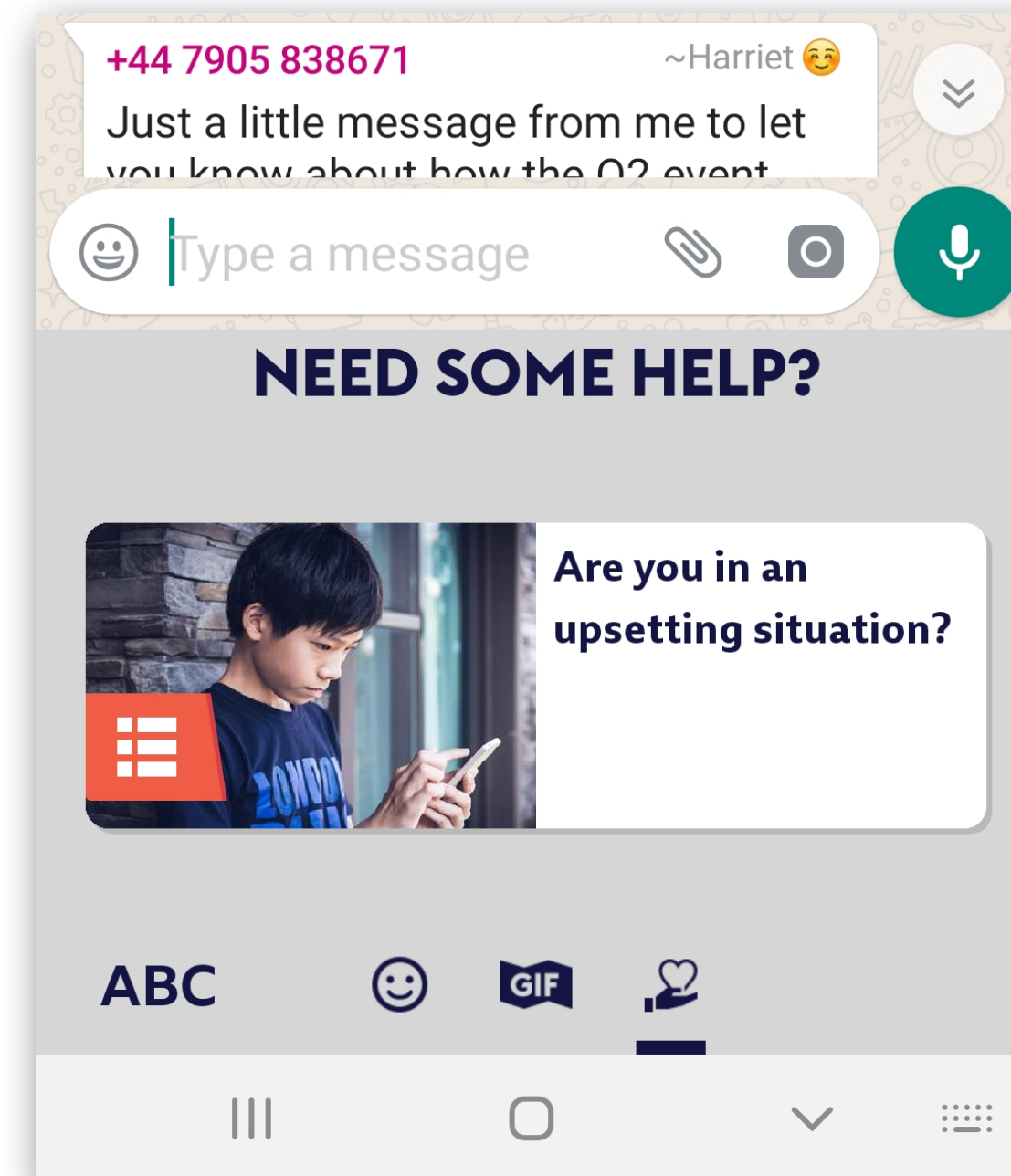
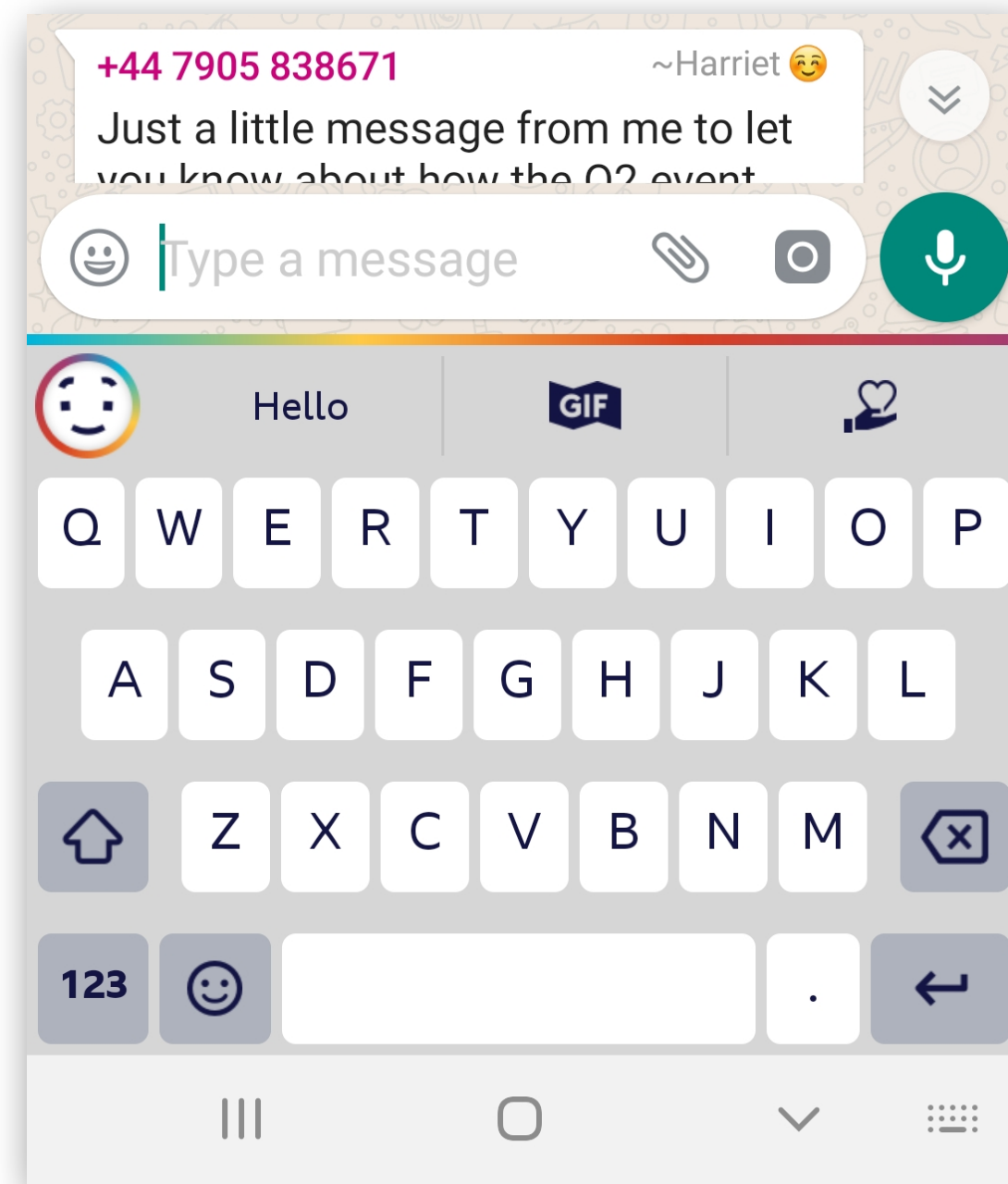




# Keeping Children Safe Online, A Dive into Social Media Post Analysis

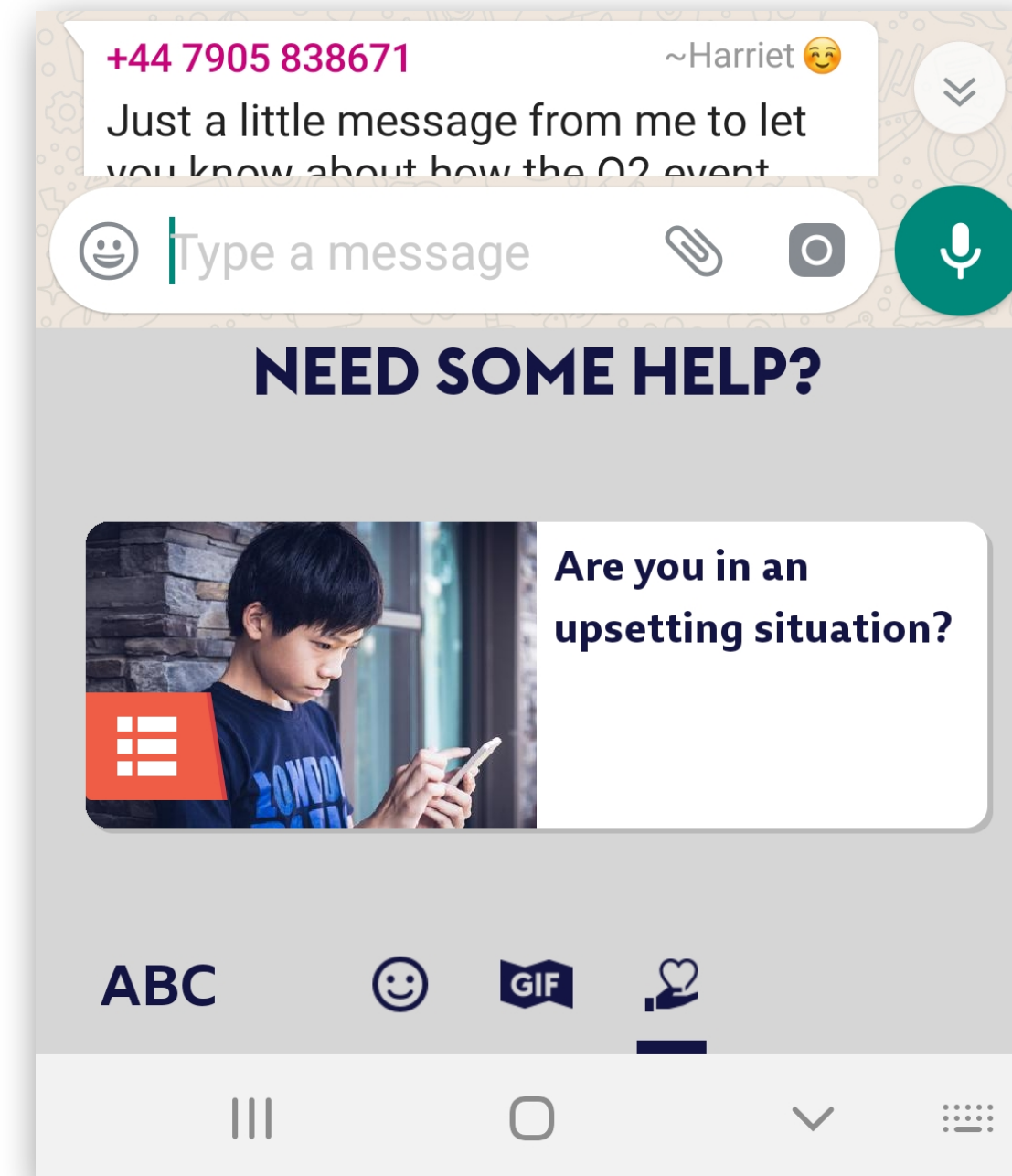
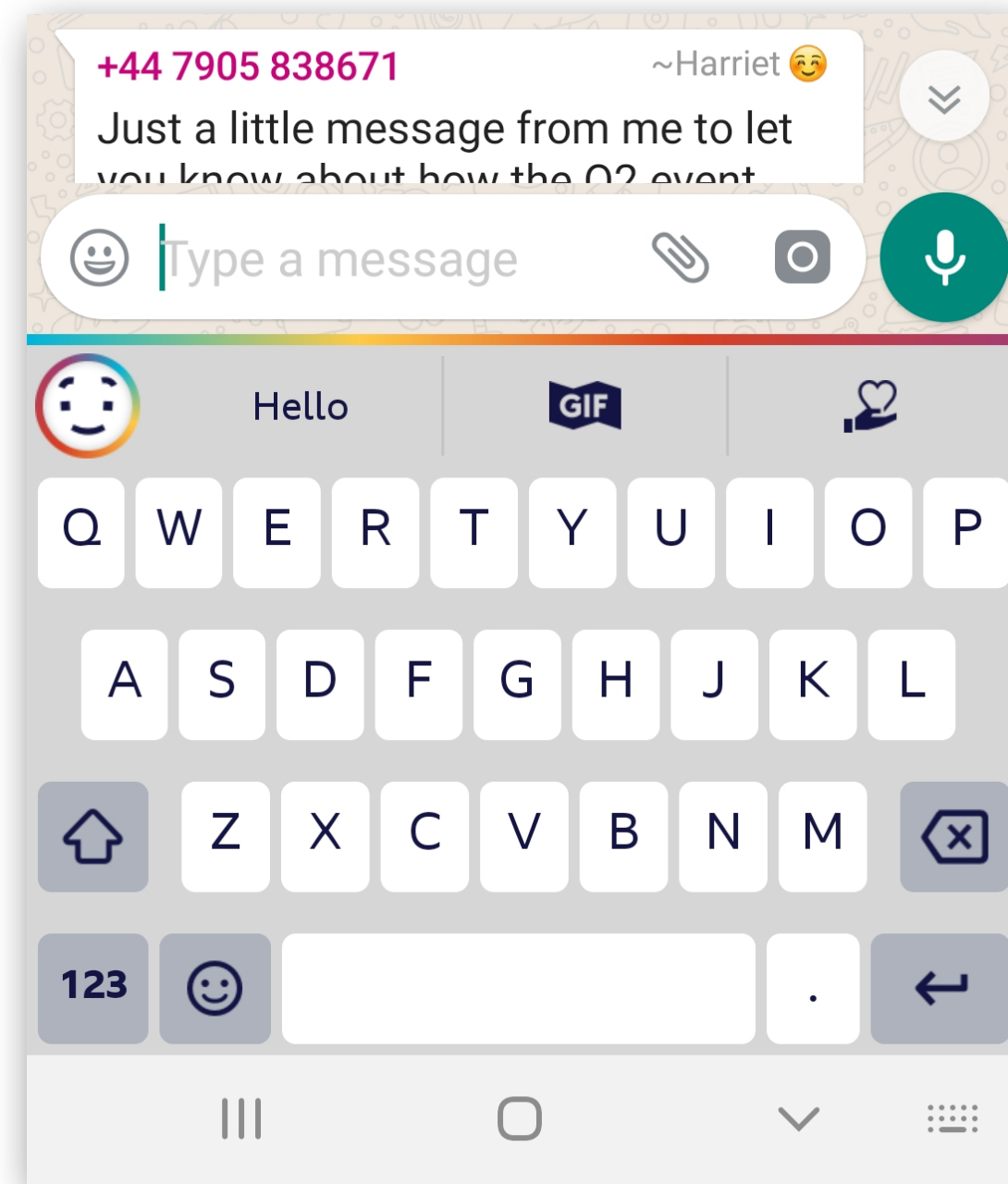
Applied Machine Learning Days 2020  
François Helg - 28.01.2020

# The Use Case



Provide instant feed-back within a keyboard

# The Use Case



Spot Hate + Toxicity + Emotion on Texts

# The Challenge

@rev\_er OMG ILU so muuuuuuuuuuuuch 🍷🍷🍷

→ How to analyse that! 🤔💥🤔

# Customised language resources for social media content

@rev\_er OMG ILU so  
muuuuuuuuuuuuch  
💕💕💕



**Custom pre-processor to handle Social Media syntax**

Removing url, mentions, html, numbers, caps, repetition

# Customised language resources for social media content

@rev\_er OMG ILU so  
muuuuuuuuuuuuch  
💕💕💕



**Custom Byte-Pair Encodings**  
trained on a large Twitter Corpus  
gave the best results.

Optimised for **mobile deployment**

# Customised language resources for social media content

@rev\_er OMG ILU so  
muuuuuuuuuuuuch  
💕💕💕



**Custom Sub-word Embeddings**  
based on **Byte-Pair-Encodings**  
trained on a large Twitter corpus

Optimised for **mobile deployment**

# Customised language resources for social media content

@rev\_er OMG ILU so  
muuuuuuuuuuuuch  
💕💕💕



**Ensembling of DPCNN** models gave the best results (*and were compatible with TF Lite*)

**Distant Supervision** strategy for data acquisition



# AI-Driven Data Labeling

Emotion Problematic Tokens			
token	count	↓ score	normalized_score
😬	21	2.7	0.1
__scary	6	2.6	0.4
__scared	7	2.5	0.4
smile	11	1.9	0.2
__not	25	1.9	0.1
__hurt	4	1.9	0.5

1-6 of 50

Rows per page:

### Selection Method

- data at random
- data confusing the classifier
- data labeled by others

**Labeled Data**

Select a token to filter the labeled data (max of 20 items is visible at a time).

2.7 😬 2.6 \_\_scary

2.5 \_\_scared

1.9 smile 1.9 \_\_not

1.9 \_\_hurt 1.8 !

1.6 😬 1.5 😬

1.5 > 1.4 ?

1.3 \_\_actually

1.2 \_\_why 1.2 😬

1 how's this shit **scary**

emotion classifier

11 anger 66 fear 1 joy 2 love 0 neutral 7 sadness

13 surprise

2 ok fine first i thought it was a **scary** short film after watching it i started laughing it was so funny

emotion classifier

11 anger 32 fear 8 joy 18 love 2 neutral 12 sadness

17 surprise

3 @ don't be a **bitch**. get two.

emotion classifier

71 anger 5 fear 10 joy 3 love 1 neutral 9 sadness 1 surprise

hate classifier

1 clean 99 hate

toxicity classifier

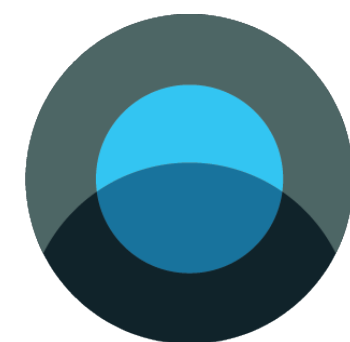
0 safe 100 toxic

# Interested to learn more ?



**Come and talk to me at the break**

François Helg - CTO Privately [francois@privately.eu](mailto:francois@privately.eu)



P R I V A T E L Y