

Four lollipop faces are arranged horizontally. From left to right: a red lollipop with a sad face, a green lollipop with a neutral face, a yellow lollipop with a surprised face, and an orange lollipop with a happy face. The text 'SafeSpace: Cyberbullying Detection' is overlaid in white on the lollipops.

SafeSpace: Cyberbullying Detection

Helen Huang

Vincent Marklynn

Anjali Sebastian

Yong Long Tan

Why?

- 95% of teens in the U.S. are online.
 - 59% of US teens have experience cyberbullying.
 - Twice likely to attempt suicide if experience cyberbullying.
 - 33% of victims experience social anxiety.
- Content moderation prevents cyberbullying.

*Statistics from <https://dataprot.net/statistics/cyberbullying-statistics/>

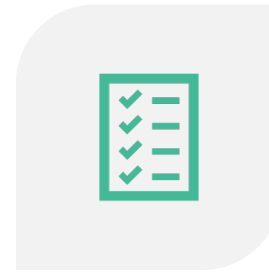
Data



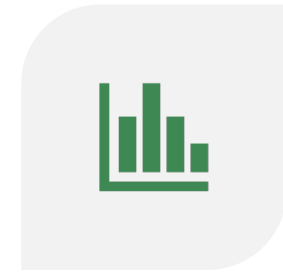
51K + LABELLED TWEETS



39K CYBERBULLYING
COMMENTS, 12K NOT
CYBERBULLYING COMMENTS

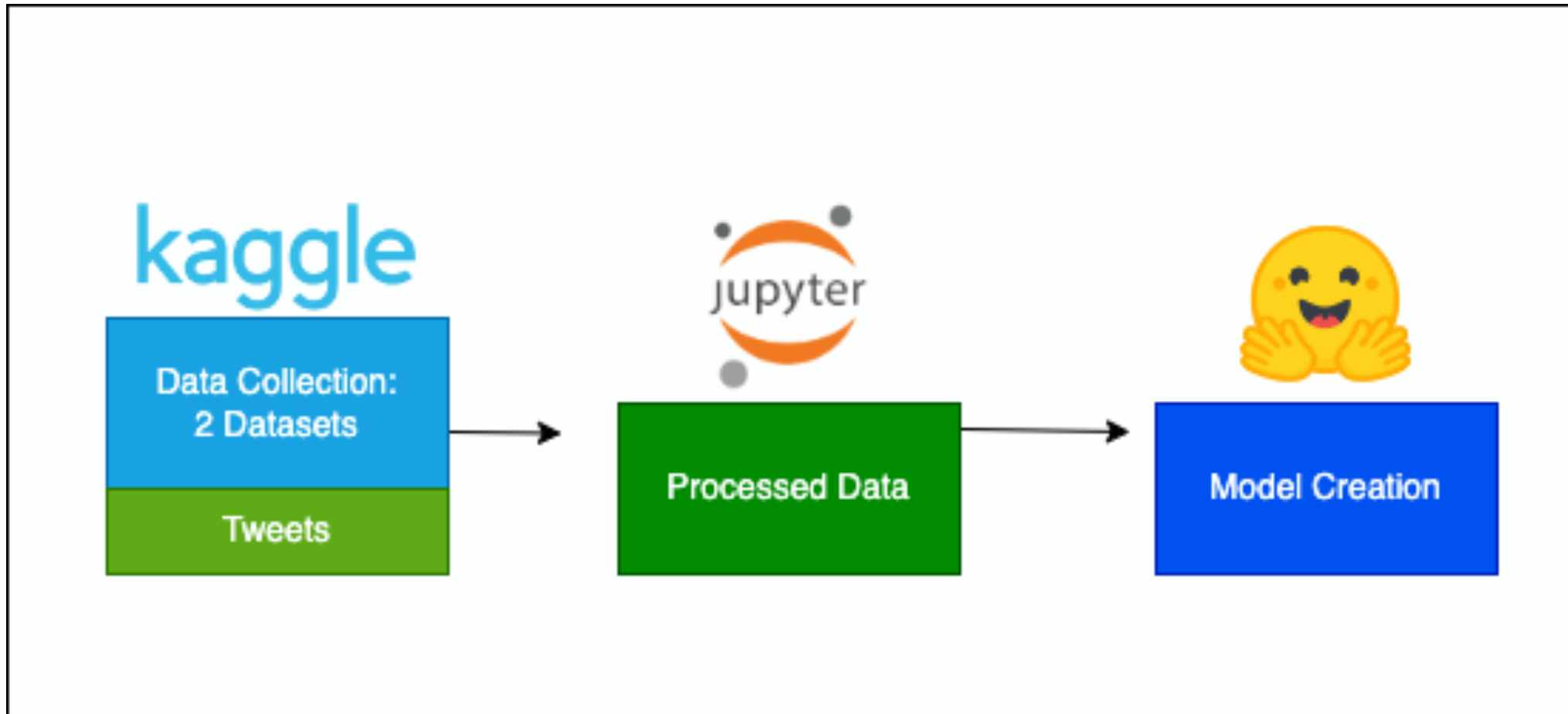


CLEAN AND PROCESS TWEETS
BEFORE TRANSFER LEARNING



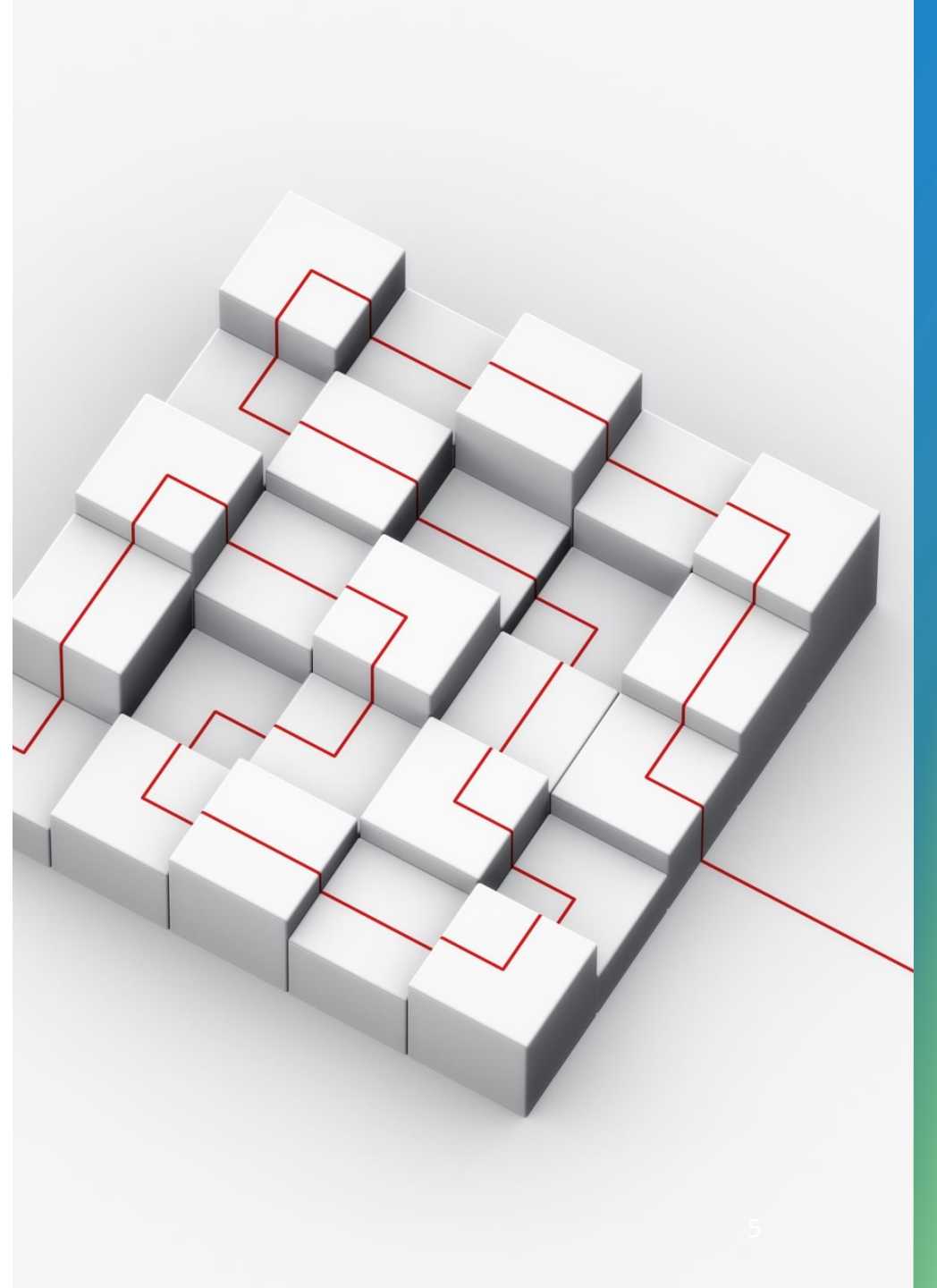
ASSESS RESULTS

How we built it?



Model

- DistilBERT
- Uses the Transformer architecture.
- Fine Tuning
- 3 Epochs



Metrics and Results

Training results

Training Loss	Epoch	Step	Validation Loss	Recall
0.2714	1.0	2599	0.2467	0.9293
0.2194	2.0	5198	0.2673	0.9475
0.1633	3.0	7797	0.3108	0.9242

Demo

- https://huggingface.co/vmarklynn/cyberbully_test_recall_v1

Git Hub Repository

<https://github.com/anjumorris/detect-cyberbullying.git>

Future Work and Impact

OpenSource moderation tool for smaller communities

Expandable to audio and video content

Try other LLMs



Thank You