NLP Investigation

Content Moderation at Meta and Twitter

Meta

- Started by Mark Zuckerberg in 2004
- 1.9 Billion active daily users
- $85.9 Billion revenue in 2020
- 98% of revenue from advertising

Company Profile

Meta Platforms, Inc. develops products that enable people to connect and share with friends and family through mobile devices, personal computers, virtual reality headsets, and in-home devices worldwide. It operates in two segments, Family of Apps and Facebook Reality Labs. The Family of Apps segment's products include Facebook, which enables people to connect, share, discover, and communicate with each other on mobile devices and personal computers; Instagram, a community for sharing photos, videos, and private messages; Messenger, a messaging application for people to connect with friends, family, groups, and businesses across platforms and devices; and WhatsApp, a messaging application that is used by people and businesses to communicate in a private way, as well as other services. The Facebook Reality Labs segment provides augmented and virtual reality related consumer hardware, software, and content that help people feel connected, anytime, and anywhere. The company was formerly known as Facebook, Inc. and changed its name to Meta Platforms, Inc. in October 2021. Meta Platforms, Inc. was founded in 2004 and is headquartered in Menlo Park, California.

https://digital.fidelity.com/prgw/digital/research/quote/dashboard/summary?symbol=FB

Twitter

- Started by Jack Dorsey in 2006
- 396 Million daily active users
- $3.7 Billion revenue in 2020
- 86% of revenue from advertising

Company Profile

Twitter, Inc. operates as a platform for public self-expression and conversation in real time United States, Japan, and internationally. The company offers Twitter, a platform that allows users to consume, create, distribute, and discover content. It also provides promoted products and services, such as promoted tweets, promoted accounts, and promoted trends, which enable its advertisers to promote their brands, products, and services. In addition, the company offers MoPub, a mobile-focused advertising exchange that combines ad serving, ad network mediation, and a real-time bidding exchange into one monetization platform; Twitter Audience platform, an advertising offering that enables advertisers to extend advertising campaigns; Developer and Enterprise solutions, a software-as-a-service platform that enables developers to build products on Twitter; and paid enterprise access for its public data streams. Twitter, Inc. was founded in 2006 and is headquartered in San Francisco, California.

https://digital.fidelity.com/prgw/digital/research/quote/dashboard?symbol=TWTR

Problem Statement

- Delinquent acts on social media can include but are not limited to cyber bullying, spreading misinformation, fake news, hateful speech, inciting violence, terrorist propaganda

- Some of the most notable tribulations:
  - The War on Hate Speech in Myanmar (Burmese language)
  - Tsunami of Hate Posts in Mumbai (Indian language)
  - Abuse in Fiji (Fijian language)
  - Deadly Ethnic Clashes in Ethiopia (Amharic language)
  - Easter Sunday Bombings in Sri Lanka (Sinhala language)
  - The Riots at the Capitol in Washington DC (English language)

- Facebook supports 111 languages while the community standards are translated in 40 languages, artificial intelligence tools can be used for 30 languages

- Twitter supports 47 languages while the community standards are translated in 37, likely does not have AI support for all of them as Twitter just started implementing AI in 2018

- Regulations are becoming increasingly strict, for example Australia, Singapore, and the UK have imposed harsh penalties including fines and potential jail time for executives if a company does not promptly remove objectionable posts

- Content moderators suffer from psychological distress because of the content that they are exposed to on the job

"Machine learning requires massive volumes of data to train computers, and a scarcity of text in other languages presents a challenge in rapidly growing the tools"

- Guy Rosen VP Integrity at Meta

"A heavy lift to translate into all those different languages"

- Monika Bickert Head of Global Policy Management at Meta

"The result has been more division, more harm, more lies, more threats and more combat. In some cases, this dangerous online talk has led to actual violence that harms and even kills people"

- Frances Haugen Product Manager at Meta

"In light of my recent experiences I am choosing to take a step back, of sorts, from Twitter"

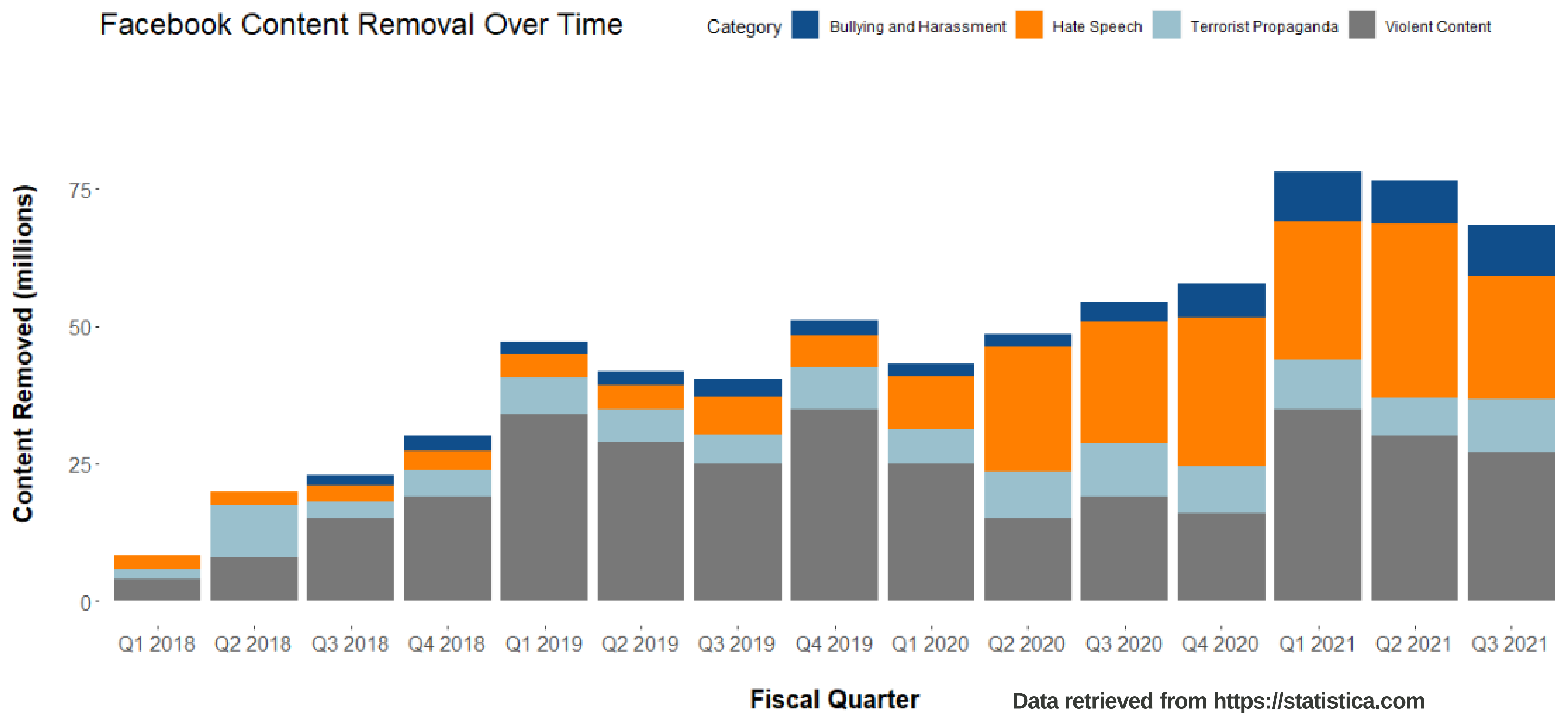
- Will Poulter English Actor

Meta Content Moderation

Key Events in NLP Advancements at Meta

- 2015 Q4 – Hate speech detection with AI begins
- 2016 Q1 - Deep text – text understanding engine
- 2018 Q1 - Scaling NLP across many languages
- 2019 Q2 - Bi-Transformer in production for hate speech
- 2019 Q3 - XLM in production for hate speech
- 2020 Q1 - Self supervision improves hate speech in English
- 2020 Q3 - XLM-R RoBERTa for hate speech across 15 languages
- 2020 Q3 - Linformer – new text encoder
- 2021 Q1 - Implemented cross-problem system that tackles hate speech, bullying and harassment, and violence and incitement

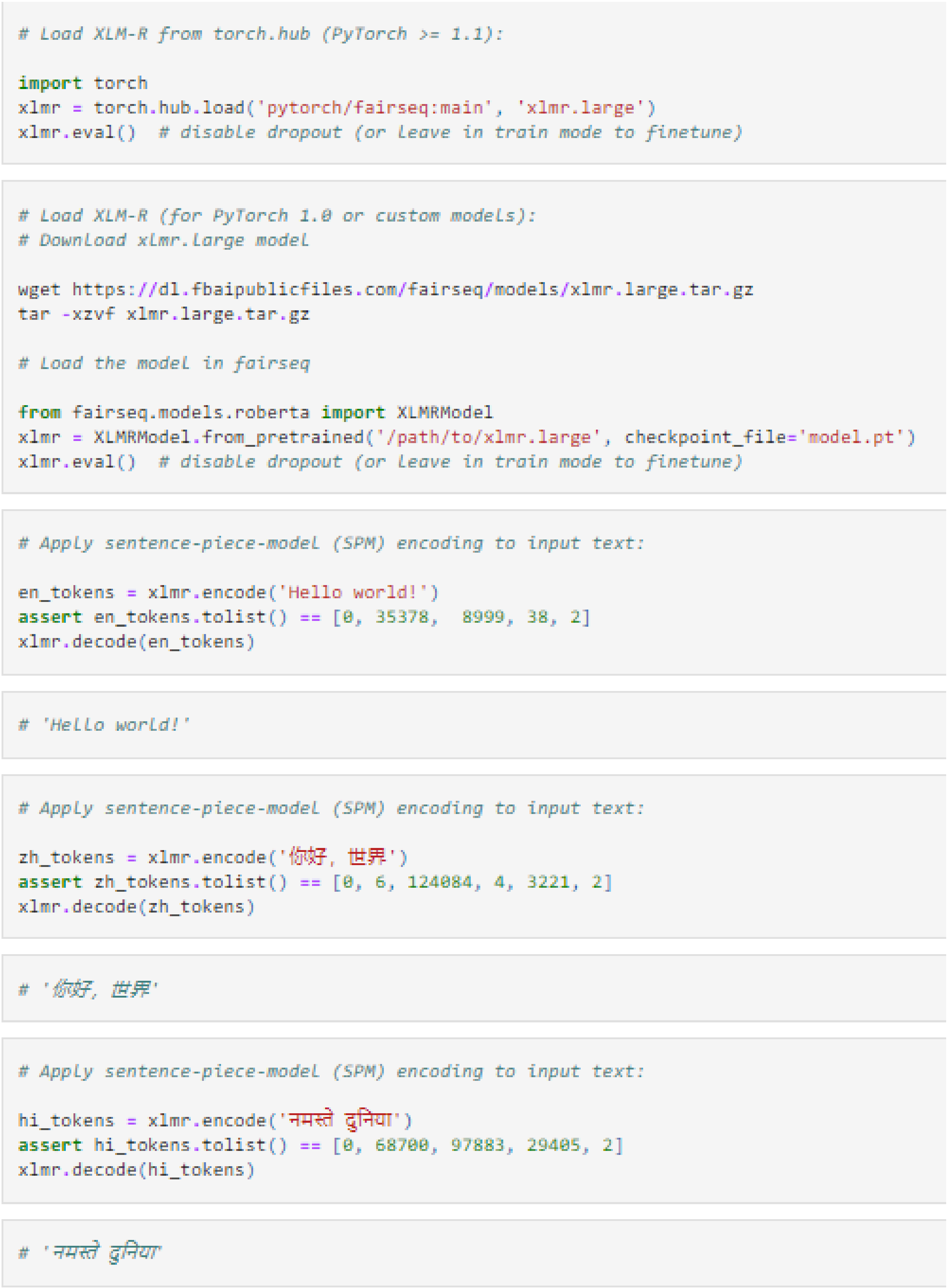
Facebook Content Removal Over Time



Data retrieved from https://statistica.com

XLM-R AI Model at Meta

- A transformer based multilingual masked language model which is pretrained on about 100 languages with a mass of 2.5 Terabytes of text data from CommonCrawl
- The most reliable model available for flagging harmful content across languages
- Uses self supervision, it trains on one language and can be used with other languages without additional training data
- Shown to outperform traditional models such as mBERT or XLM by several percentage points in accuracy and F1 score
- All the code is publically available on GitHub



```
# Load XLM-R from torch.hub (PyTorch >= 1.1):

import torch
xlmr = torch.hub.load('pytorch/fairseq:main', 'xlmr.large')
xlmr.eval()  # disable dropout (or leave in train mode to finetune)
```

```
# Load XLM-R (for PyTorch 1.0 or custom models):
# Download xlmr.large model

wget https://dl.fbaipublicfiles.com/fairseq/models/xlmr.large.tar.gz
tar -xzvf xlmr.large.tar.gz

# Load the model in fairseq

from fairseq.models.roberta import XLMRModel
xlmr = XLMRModel.from_pretrained('/path/to/xlmr.large', checkpoint_file='model.pt')
xlmr.eval()  # disable dropout (or leave in train mode to finetune)
```

```
# Apply sentence-piece-model (SPM) encoding to input text:

en_tokens = xlmr.encode('Hello world!')
assert en_tokens.tolist() == [0, 35378,  8999, 38, 2]
xlmr.decode(en_tokens)
```

```
# 'Hello world!'
```

```
# Apply sentence-piece-model (SPM) encoding to input text:

zh_tokens = xlmr.encode('你好，世界')
assert zh_tokens.tolist() == [0, 6, 124084, 4, 3221, 2]
xlmr.decode(zh_tokens)
```

```
# '你好，世界'
```

```
# Apply sentence-piece-model (SPM) encoding to input text:

hi_tokens = xlmr.encode('नमस्ते दुनिया')
assert hi_tokens.tolist() == [0, 68700, 97883, 29405, 2]
xlmr.decode(hi_tokens)
```

```
# 'नमस्ते दुनिया'
```

Code retrieved from https://github.com/pytorch/fairseq/tree/main/examples/xlmr

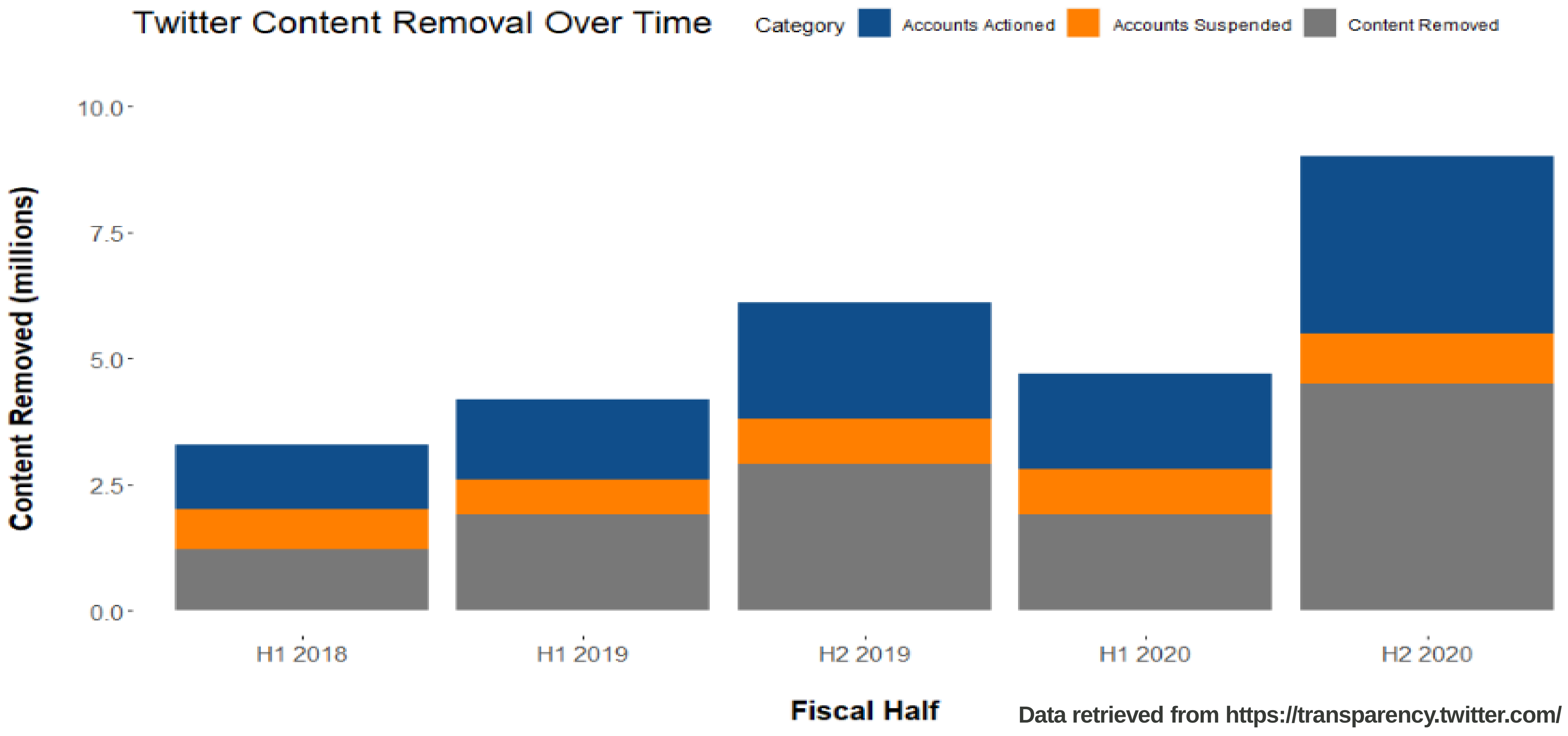
Twitter Content Moderation

- Before 2019, Twitter was not using any AI for content moderation, all misconduct had to be reported from the community

- as of 2021, Twitters AI detection is able to identify 65% of tweets that are in violation of policies proactively before being reported by a user

- Twitter has developed a framework called DeepBird using the Tensorflow package in Python which is a deep learning solution

- According to a statement by Jack Dorsey in February 2021, Twitter plants to make their content moderation algorithms more transparent



Twitter Content Removal Over Time

Data retrieved from https://transparency.twitter.com/

Resources

- https://backlinko.com/facebook-users
- https://backlinko.com/twitter-users
- https://backlinko.com/snapchat-users
- https://www.reuters.com/technology/facebook-knew-about-failed-police-abusive-content-globally-documents-2021-10-25/
- https://www.reuters.com/investigates/special-report/myanmar-facebook-hate
- https://www.reuters.com/article/us-facebook-india-content/facebook-a-megaphone-for-hate-against-indian-minorities-idUSKBN1X929F
- https://www.reuters.com/article/us-facebook-languages-insight-idUSKCN1RZ0DW
- https://about.fb.com/news/2021/11/how-meta-addresses-bullying-harassment/
- https://www.reuters.com/technology/australia-regulator-says-concerned-about-facebook-approach-media-licencing-law-2021-10-25/
- https://www.npr.org/2021/10/05/1043377310/facebook-whistleblower-frances-haugen-congress
- https://ai.facebook.com/blog/the-shift-to-generalized-ai-to-better-identify-violating-content
- https://ai.facebook.com/blog/-xlm-r-state-of-the-art-cross-lingual-understanding-through-self-supervision/
- https://arxiv.org/abs/1911.02116
- https://github.com/facebookresearch/cc_net
- https://paperswithcode.com/paper/unsupervised-cross-lingual-representation-1
- https://github.com/pytorch/fairseq
- https://github.com/facebookresearch/pytext
- https://github.com/facebookresearch/XLM
- https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/
- https://github.com/pytorch/fairseq/tree/main/examples/roberta
- https://gluebenchmark.com/leaderboard
- https://arxiv.org/abs/1907.11692
- https://ai.facebook.com/blog/pytorch-builds-the-future-of-ai-and-machine-learning-at-facebook/
- https://ai.facebook.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech/
- https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/
- https://ronan.collobert.com/pub/matos/2008_nlp_icml.pdf
- https://arxiv.org/pdf/1301.3781.pdf
- https://nlp.stanford.edu/pubs/glove.pdf
- https://arxiv.org/pdf/2103.01988.pdf?fbclid=IwAR2pqhYda6MV9r2b3Afx_0eKUiZhX-Es6Pa_FbLOqH8fglQzO2kY3yKxZE8
- https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/
- https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/
- https://tech.fb.com/how-ai-is-learning-to-see-the-bigger-picture/
- https://www.fastcompany.com/90539275/facebooks-ai-for-detecting-hate-speech-is-facing-its-biggest-challenge-yet
- https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/facebook-says-hate-content-on-its-platform-continued-to-decline-in-q3-67552325
- https://medium.com/@aman.anand54321/cross-lingual-models-xlm-r-7d557302698b
- https://github.com/pytorch/fairseq/tree/main/examples/xlmr
- https://techcrunch.com/2021/09/01/twitter-safety-mode-harassment/
- https://www.popsci.com/artificial-intelligence-identify-real-news-on-twitter-facebook/https://www.reutersagency.com/en/reuters-community/reuters-news-tracer-filtering-through-the-noise-of-social-media/https://www.niemanlab.org/2016/11/reuters-built-its-own-algorithmic-prediction-tool-to-help-it-spot-and-verify-breaking-news-on-twitter/
- https://www.bloomberg.com/opinion/articles/2021-02-19/facebook-and-twitter-content-moderation-is-failing
- https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/
- https://journals.sagepub.com/doi/full/10.1177/2053951720943234
- https://www.ted.com/talks/jack_dorsey_how_twitter_needs_to_change/transcript#t-181509
- https://www.tandfonline.com/doi/abs/10.1080/1369118X.2016.1153700
- https://blog.twitter.com/en_us/topics/company/2019/addressing-the-abuse-of-tech-to-spread-terrorist-and-extremist-c
- https://fortune.com/2019/10/24/twitter-abuse-tweets/
- https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jul-dec
- https://www.techdirt.com/articles/20200925/14414345379/content-moderation-case-study-twitters-algorithm-misidentifies-harmless-tweet-as-sensitive-content-april-2018.shtml
- https://blog.twitter.com/engineering/en_us/topics/insights/2018/twittertensorflow
- https://economictimes.indiatimes.com/tech/tech-bytes/twitter-intends-to-make-its-content-moderation-practices-more-transparent-jack-dorsey/articleshow/81223668.cms

Data retrieved from https://transparency.twitter.com/