Model Evaluation Pipeline

Extensive automatic reporting for Project Detox

Ajinkya Sheth, Ishita Bhandari, Prakirn Kumar

Team Fan-STATS-tic 4

Microsoft

Information School
UNIVERSITY of WASHINGTON

Ajinkya Sheth

Data Engineer

Ishita Bhandari

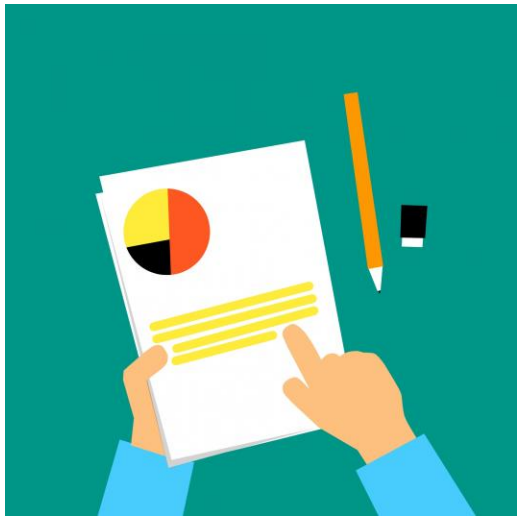
Software Developer

Prakirn Kumar

Data Scientist

Executive Summary

Toxicity is a huge problem in online gaming industry. Toxicity in this context refers to any content posted on social media which may indicate swearing, profanity, bullying, sexual abuse, and racism. In order to combat toxic content and promote pro-social behavior, Data Scientists at Microsoft Xbox Gaming Safety team have deployed state-of-art machine learning models as a part of Project Detox.



However, the models have been tested on limited data and existing process for model evaluation and performance measure is a manual activity. To streamline these activities, Microsoft Gaming Safety Team is collaborating with iSchool at University of Washington as a part of the 2020 iSchool Capstone Event.

Team Fan-STATS-tic 4 of the iSchool is delivering the prototype project – Model Evaluation Pipeline to Microsoft Gaming Safety Team as a part of this Capstone Experience.

Table of Contents

The Sponsor

The capstone project was proposed by Microsoft Corporation. Microsoft Corporation is an American multinational technology company with headquarters in Redmond, Washington. It develops, manufactures, licenses, supports, and sells computer software, consumer electronics, personal computers, and related services.

Within Microsoft, this project is under the Gaming Safety team part of the Xbox group. Xbox is a video gaming brand created and owned by Microsoft. It represents a series of video game consoles developed by Microsoft, with three consoles released in the sixth, seventh, and eighth generations, respectively.

The brand also represents applications (games), streaming services, an online service by the name of Xbox Live, and the development arm by the name of Xbox Game Studios. The brand was first introduced in the United States in November 2001, with the launch of the original Xbox console.

The Prompt

The Microsoft Gaming Safety team was looking for a team to help them pioneer new machine learning models to promote “pro-social” behavior in gaming which is a principal customer promise Microsoft would like to deliver on in 2020. The Gaming Safety team wanted to explore new ways of protecting gamers and their content moderators from toxic content by encouraging pro-social behavior on Xbox Live and in Gaming social channels.

The Gaming Safety team has already created several models to assign a toxicity score for Live messages as an example. Moreover, they also needed a team to be working with other teams including Data in Gaming (DiG) team and the Safety Engineering team to build and put in production enhanced machine learning models.

Why we chose this project?

Being a team of data scientists we wanted to tackle a problem that required core machine learning concepts and enabled us to create a meaningful impact on society. We are also a team of passionate gamers, who've experienced toxic content on gaming forums firsthand.

We understood the depth of the problem and its repercussions and thus, were highly excited when the Xbox Gaming Safety team at Microsoft presented the opportunity to us.

Moreover, on this project we also got a chance to work on the data engineering and reporting aspects of a data science project, which we feel was a great learning experience for all of us to succeed in this field.

Project Objectives

Identifying toxicity on Online gaming forums is an industry-wide problem and, at Microsoft, Data Scientists had developed models to solve the same. But, evaluating the performance of these models was a challenging manual task and this is where our team came in.

Our project objectives were:
First, Analysis to capture and process data, score them with different models and work with Monica to develop metrics for evaluation.

Second was Reporting to display visually appealing analysis of the models output using Plotly in HTML format for our stakeholders.

And the last piece was the Pipeline developed using Azure Data Factory and Databricks to automate our evaluation framework and generate reports on the Azure Cloud Portal.

Analysis

Reporting

Pipelining

Analysis

Model Performance Evaluation

Static Data: Benchmark & External Labelled Data

Our model performance evaluation is divided into two parts. First is the evaluation of static data which consists of true toxicity labels acting as a ground truth for our model evaluation metrics. For this piece we used Xbox benchmark data and an external data which includes Wikipedia comments dataset. We used this external data to see how our models' prediction compare with Wikipedia's pre-existing toxicity labels based on a completely different policy.

We are comparing these datasets using 2 different Detox API. The Detox API 1.0 is currently in production and scores Xbox live messaging data. On the other hand, Detox API 1.1 is currently in development phase and is an iteration over Detox API 1.0, with improvements in misspellings & internet slangs.

After initial research and analysis ,we decided to use metrics like Accuracy, F1-score, ROC and Precision-Recall Curves and we'll delve deeper into them in our reports.

Xbox Benchmark Data

Wikipedia Comments Data

Detox API 1.1

Detox API 1.0

Model Performance Evaluation

Live Data: Twitch and Mixer

The second part is the evaluation of live data which lacks true toxicity labels. For this we scraped chat data from Mixer and Twitch which are among the top video game streaming platforms.

We used this data because it is live and easy to gather. This chat data is public and hence free from the shackles of privacy violations.

Since, we do not have true labels in this data, we are plotting distribution charts, histograms and word clouds to understand the underlying toxicity patterns.

Twitch

Mixer

Detox API
1.1

Detox API
1.0

Pipeline

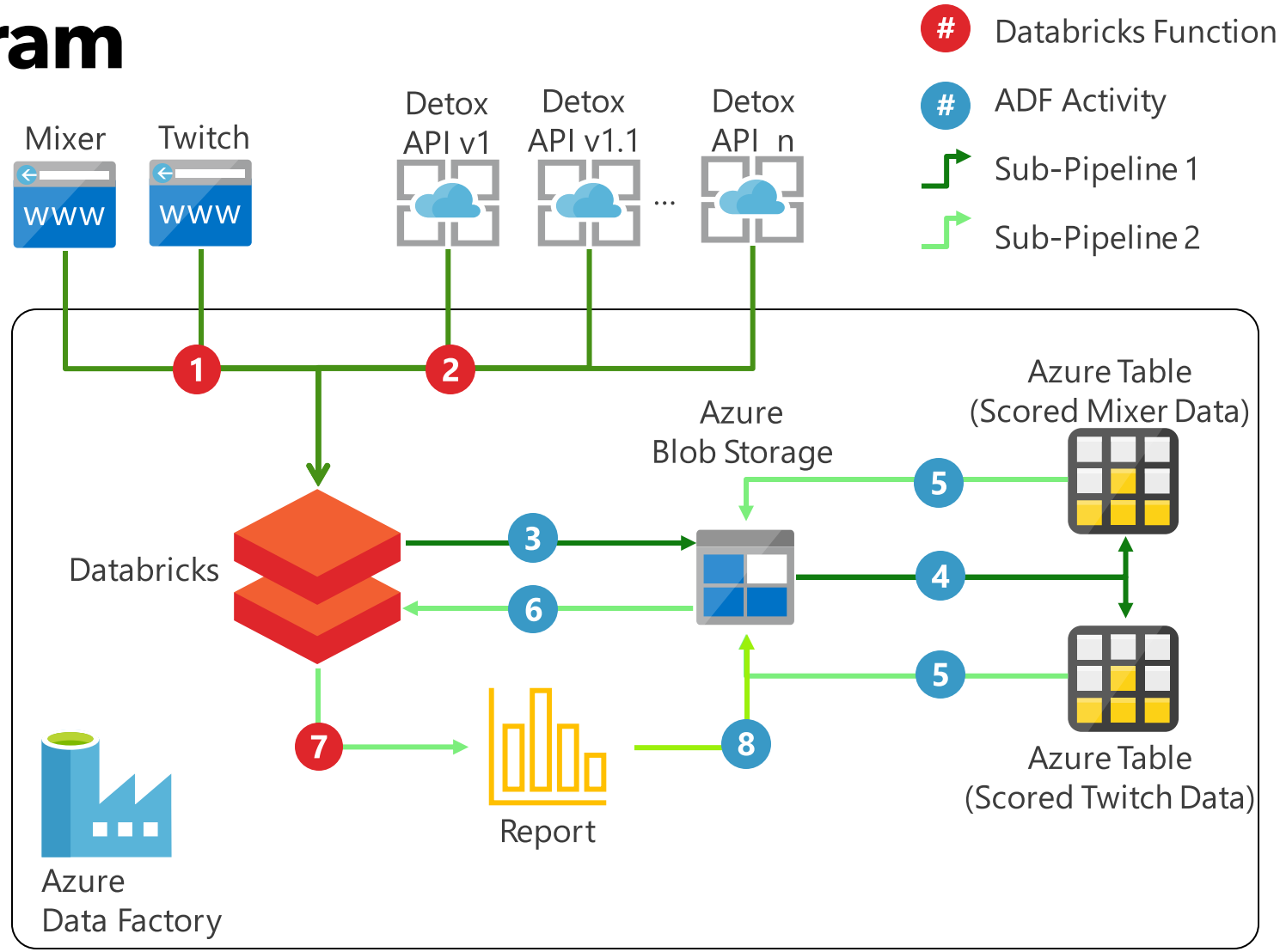
Pipeline Flow Diagram

New / Incremental Data

1. Web Scraping
2. Batch Scoring
3. Databricks to Blob
4. Blob to Azure Table

Existing Data

5. Azure Table to Blob
6. Blob to Databricks
7. Report Generation
8. Report Storage



Pipeline Explanation

The pipeline has been designed in Azure Data Factory. Databricks (DB) Cluster is the compute in the pipeline, and we are using Azure Blob storage and Table storage for storing the data. ADF Activities available in Azure Data Factory to perform data movement between compute and storage.

Sub-pipeline 1 captures live or newly generated chat data from game streaming websites - Twitch and Mixer through web scraping. The scraped data is scored with Detox API v1 and v 1.1 via Batch Scoring Process. Azure Blob storage is mounted on the Databricks cluster to provide temporary storage for newly captured and scored data. The data is incrementally maintained in Azure Table.

In Sub-pipeline 2, data in Azure table is first copied to the blob storage which is mounted to Databricks cluster. The DB cluster reads this data and generates interactive JS/HTML reports which are stored in the Blob for easy access.

Web Scraping Process

Obtain publicly available data on demand for model evaluation

Twitch

Scraping Protocol : IRC
(get chat stream)

1. Choose a channel
2. Set capture duration
3. Log chat streams
4. Store the data

Mixer

Scraping Protocol : Rest API
(get chat history)

1. Choose a channel
2. Set capture Duration (d)
3. Set number of Hits (n)
4. Hit Mixer API n times in d duration
5. Store the data

Web Scraping

Web Scraping in this context is about collecting newly generated live data. Since this model APIs have been tested on limited data, it is imperative for us to collect data from popular live streaming websites like Twitch and Mixer. Twitch has a huge user base and viewership and has been around for years.

Meanwhile, Mixer is Microsoft's in-house game streaming website launched recently. Moreover, Twitch has been considered more Toxic as compared to Mixer. By collecting publicly available data, we also attempt to measure model performance on competitors.

Web-scraping process is different for Mixer and Twitch as outlined in the previous slide. However, both the processes grant the user enough control on the amount of data to be scraped and from which channel.

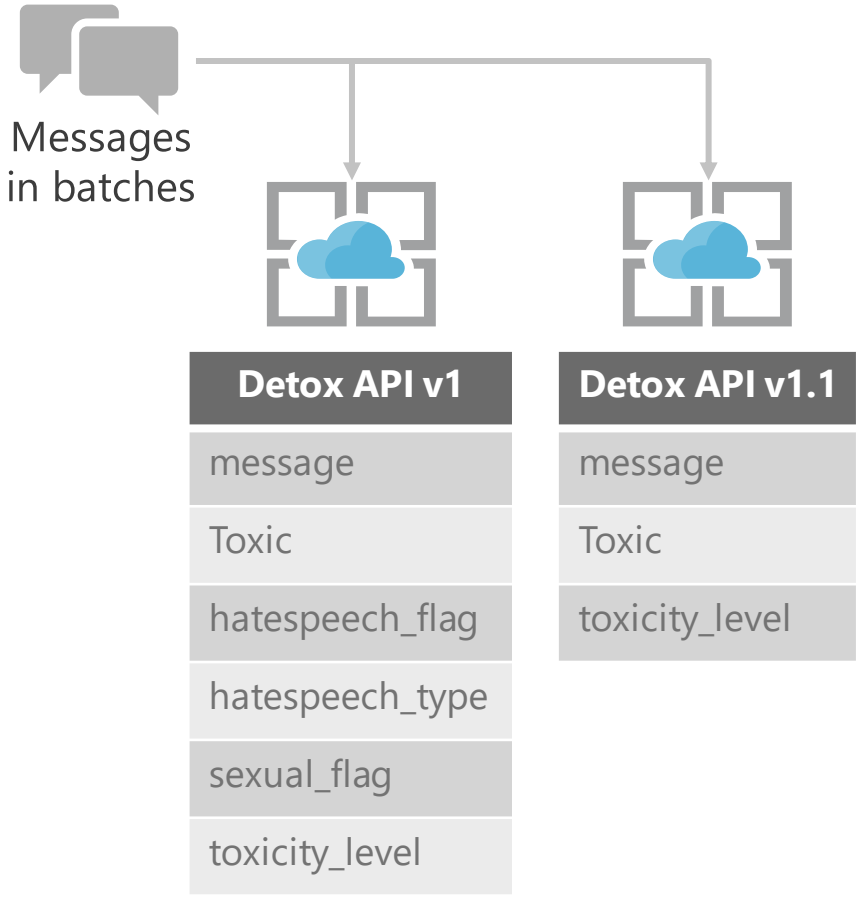
Batch Scoring

Score the scraped data in batches

Gamer chat contains a large number of small messages. Detox API are deployed on cloud and hence for the scoring to happen the API needs to be hit via the internet. Hitting the API multiple times through the network, would result into a bottleneck.

Hence, an optimal way is to score the chat text in batches. We figured out, a decent batch size is 2000. The Batch Scoring process sends messages in batches of 2000 and gets the scored data through REST API.

The scored data is then stored in Azure table incrementally for further processing.



Reporting

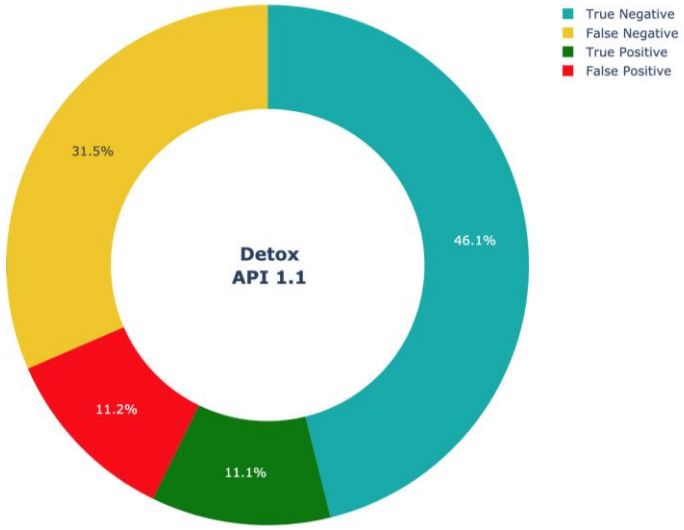
Reporting

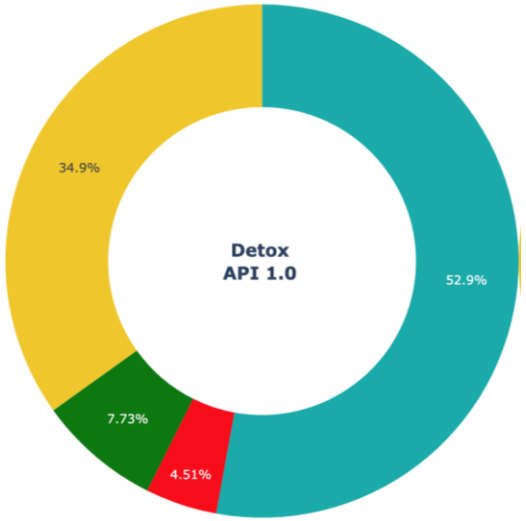
The reports are generating meaningful insights and evaluating the models against each other and on different datasets. This will help our stakeholders in making decisions on which models to deploy on production and understanding how their performance change over time as they add more data.

To generate our reports we are using the Plotly library because it creates interactive and visually appealing visualizations within Python notebooks. Hence, it is easy to plug in with our analysis in Databricks. The initial plan was to do the reporting in Power BI and but given the time we had it was easier to generate interactive visualizations in Plotly than in Power BI.

Our pipeline generates three reports – the first two analyses results from Detox API 1 and API 1.1 respectively on both Twitch and Mixer and the last one compares the performance of both the APIs to each other.

Confusion Matrix on Benchmark Data using both Detox APIs





Report 1: Performance of Detox API 1.0 on Twitch and Mixer

The first two reports analyze the performance of each API Detox 1.0 and Detox 1.1. The first graph (Fig. 1) gives the user a good overview of the data on which performances of our models is evaluated.

As we don't have true labels in the Mixer and Twitch data the, first two reports – Performance of Detox API 1.0 on Twitch and Mixer and Performance of Detox API 1.1 on Twitch and Mixer - are more on the exploratory side and so we have bar plots and histograms which try to unearth the underlying patterns in the data.

Total Count	No. of Users	No. of Scraped Channels
692	537	3

Fig. 1: Overview of Twitch scraped data

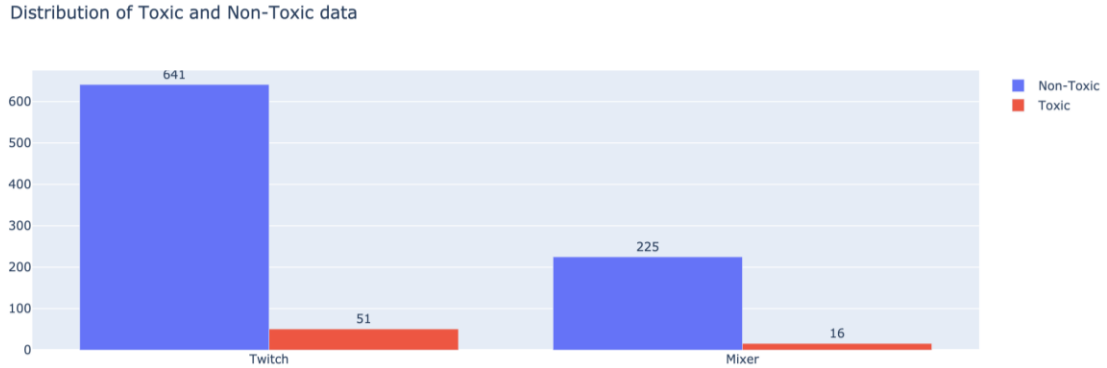


Fig. 2: Distribution of class

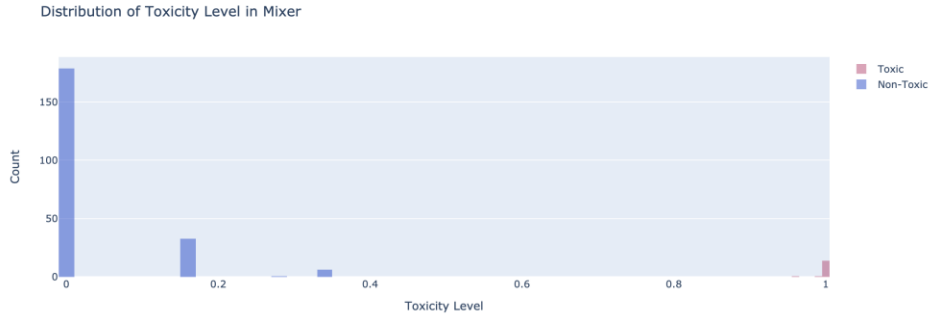


Fig. 3 Histogram to Toxicity levels in Mixer

Report 2: Performance of Detox API 1.1 on Twitch and Mixer

The second report is pretty much the same as the first one and evaluates performance of Detox API 1.1 on the scraped data.

In order to understand the predictions made by the two models we have also created word clouds of both Toxic and Non-Toxic labeled class. Shown here is a word cloud generated by one of our reports.

Word cloud of Non-Toxic Mixer data

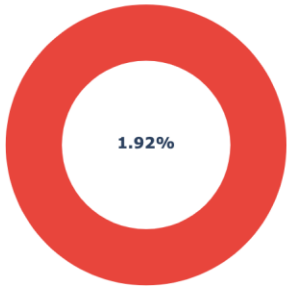
Report 3: Comparison of Detox API 1.0 and 1.1 on Twitch and Mixer

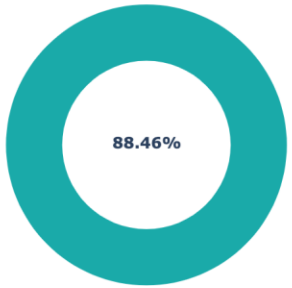
The last report is for comparing the results from both the APIs – Detox 1.0 and Detox 1.1.

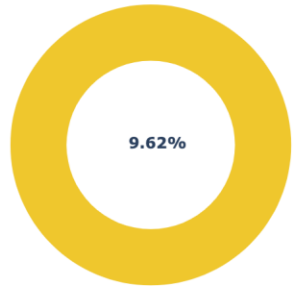
As Detox 1.0 is already in production we want our stakeholders to be able to measure the performance of any other API with respect to Detox 1.0.

The figure shown here gives a detailed view of on what kind of comments did the two APIs agree and on what kind of comments did they disagree.

Overlap Percentage between results from Detox 1 and Detox 1.1 on Mixer Data







Toxic Overlap
Non-Toxic Overlap
Non-Overlap

Sample Toxic Comments
[redacted]

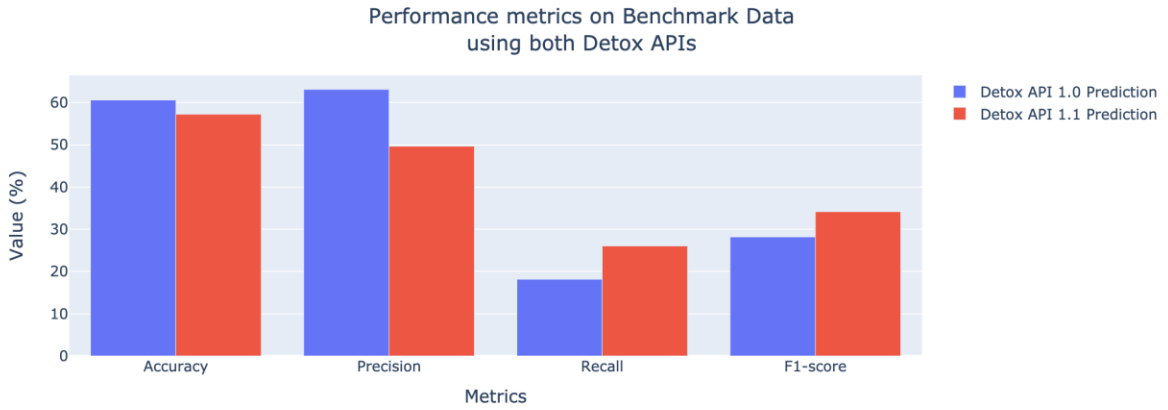
Sample Non-Toxic Comments
Bro 800k watching space x live stream and 1.1 mil watching on NASA's channel
It's almost a fully autonomous rocket
Immortal 1 for the 5th time.
😂😂
two people are on it
What if those 2 people hate each other?
going to the space station
Cya chat

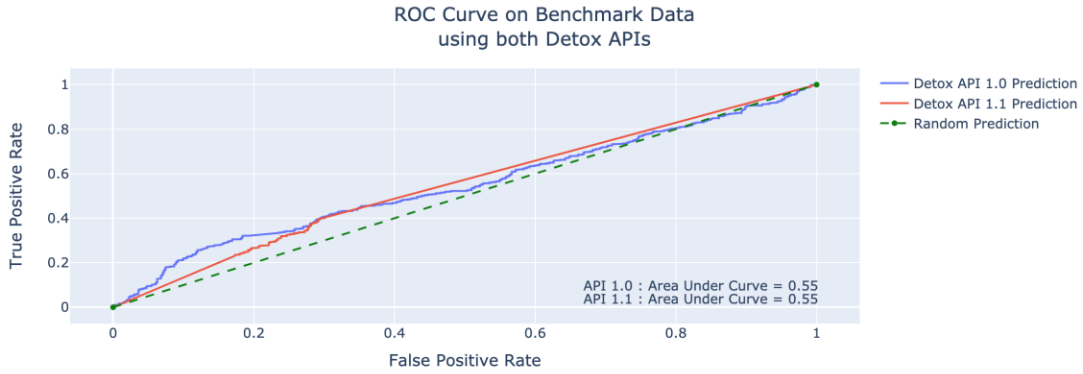
Sample Non-Overlapping Comments
Sometimes tik toks are hit or miss
man feels bad you could of chose the number 1 dad in fortnite
Ninj r u gonna watch the spacex launch nasa has 1.1 million views
going to space station
They are going to the ISS

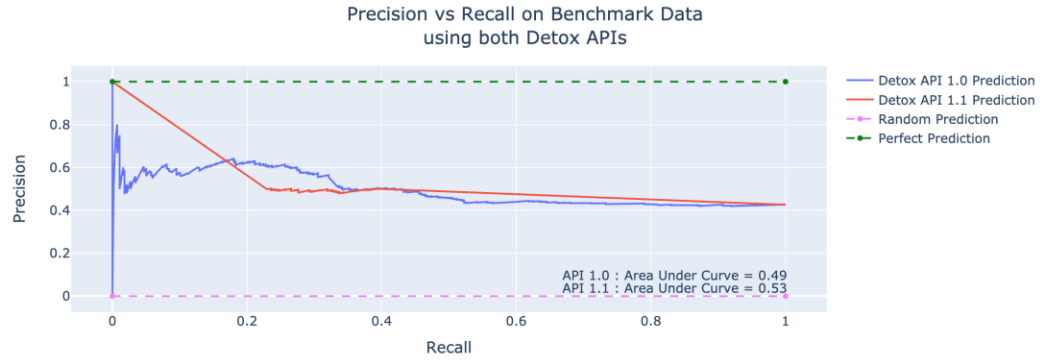
Report on Xbox Benchmark Data

We were also given a benchmark data by the Xbox team which had comments and were labeled by humans.

Though this isn't part of our pipeline, we generated reports on this data as well. Since we had true labels for this data, we had metrics like accuracy, precision, and recall and plotted more sophisticated graphs like Receiver Operating Characteristic curve to evaluate our models.







THANK YOU