The Magic of Synthetic Data

By
Dewayne
Whitfield

LARGE DATASET

+

COMPUTE POWER

+

TALENT

LARGE DATASET

What if you don’t have enough data?

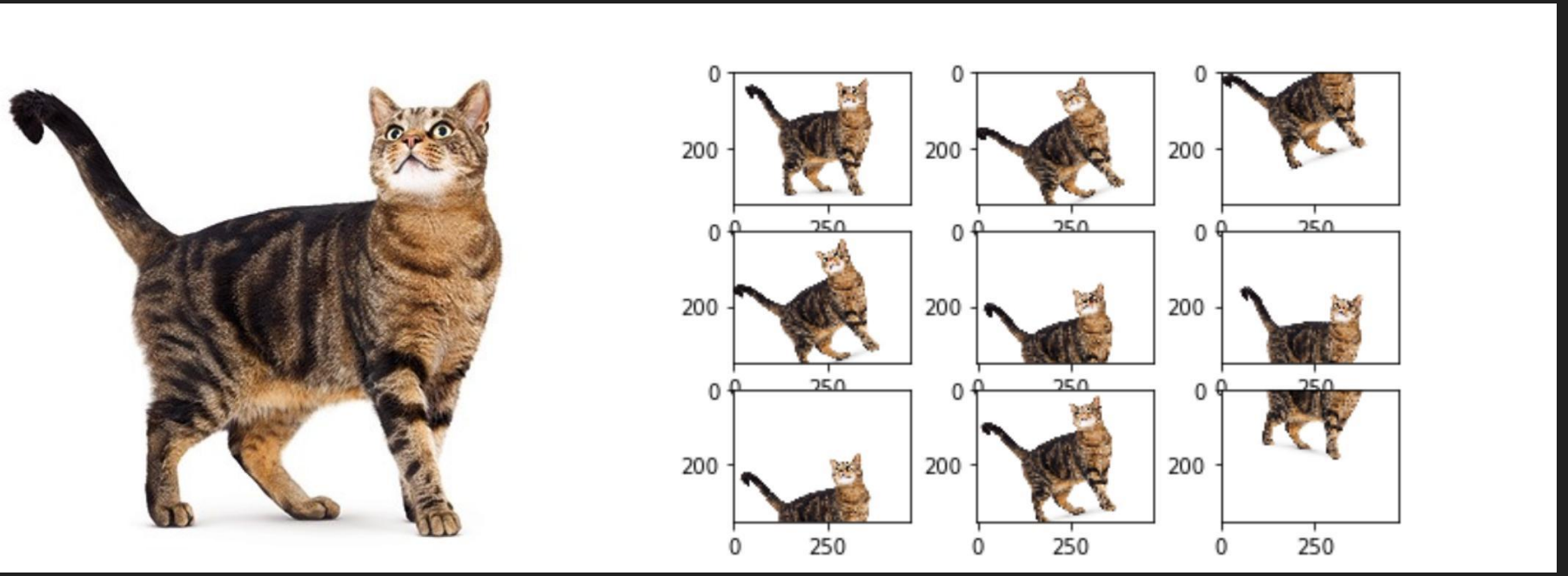
Data Augmentation

↓

Genuine Data

↓

Synthetic Data

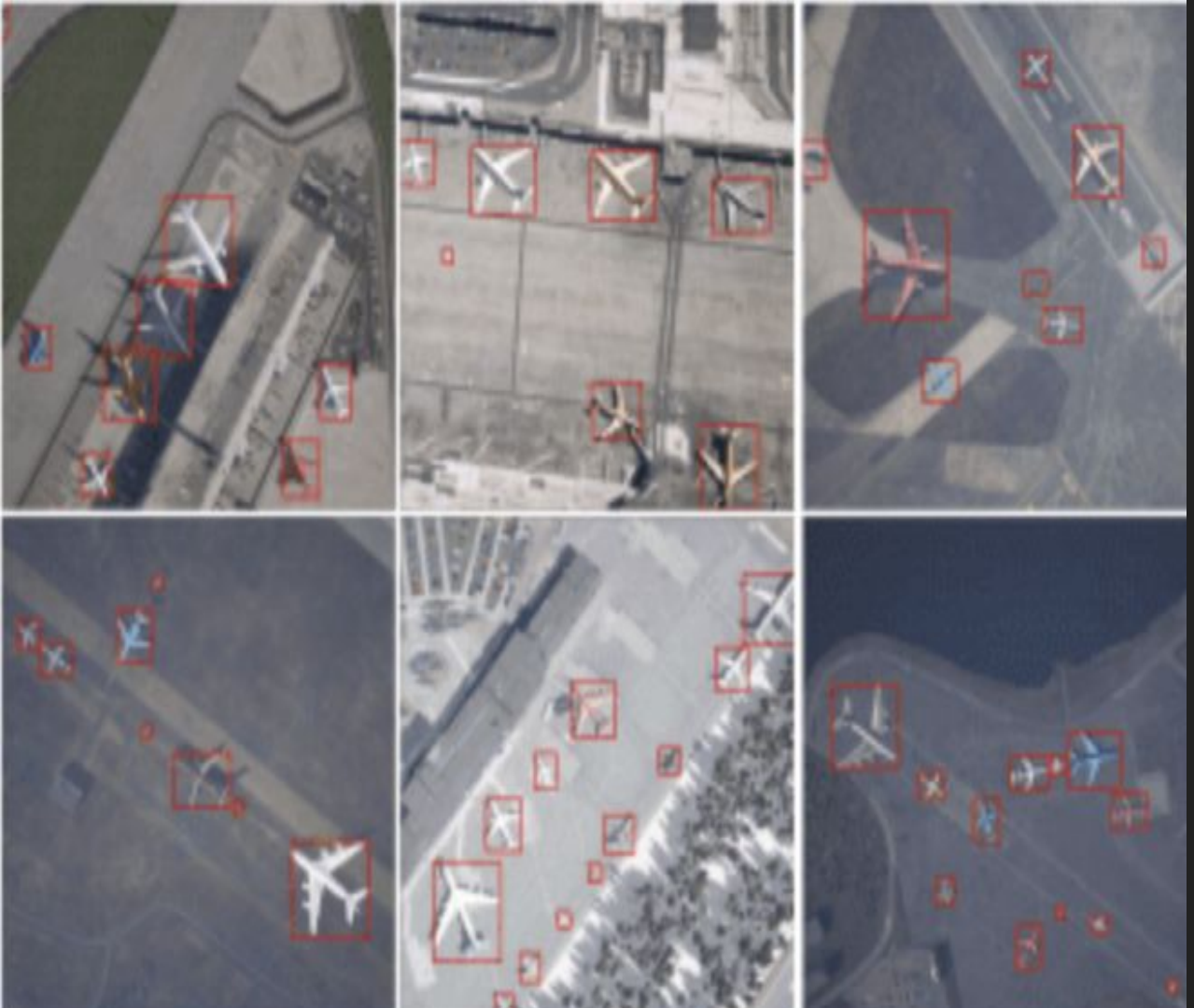


Data Scientists Crop, Zoom, and Rotate Images to generate Synthetic Image Data

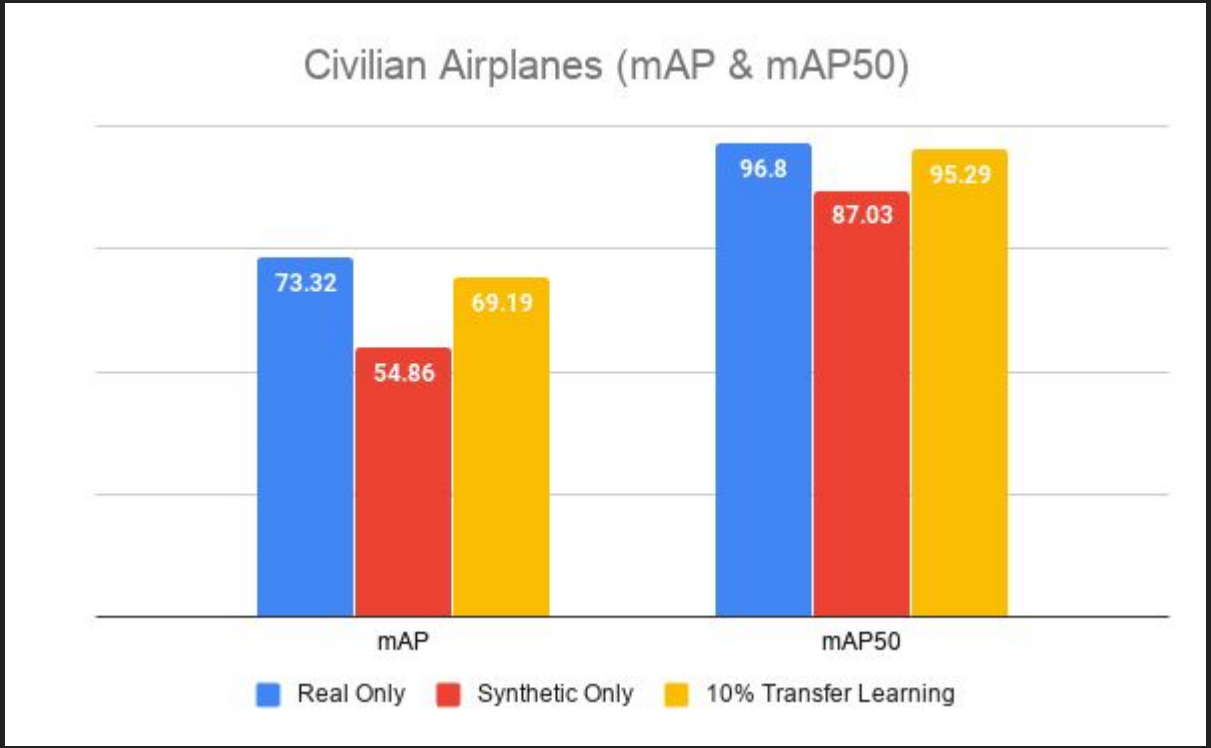
Shell is using synthetic data to identify rare problems

Deteriorating Oil Lines

Customers smoking at the gas pump



AI.REVERIE



When the Synthetic Data was fine-tuned with 10% of genuine data, it came within 4% of the accuracy of the Genuine Data Only

How might we leverage Synthetic Data to improve the performance of NLP Classification Models?

yelp®

Auto Repair Reviews

Pizza Reviews

	stars	rating	text
1	5	Positive	first time eating there and everything was so yummy! great pizza and salad, my son had the meatball sub he said it was very good, must have been because he wouldn't share. highly recommend.

1 and 2 Star = “Negative”
4 and 5 Star = “Positive”

Sentiment Analysis

Auto Repair Reviews

1210 Genuine Observations

12,050 Synthetic Observations

13,260 Combined Observations

Pizza Reviews

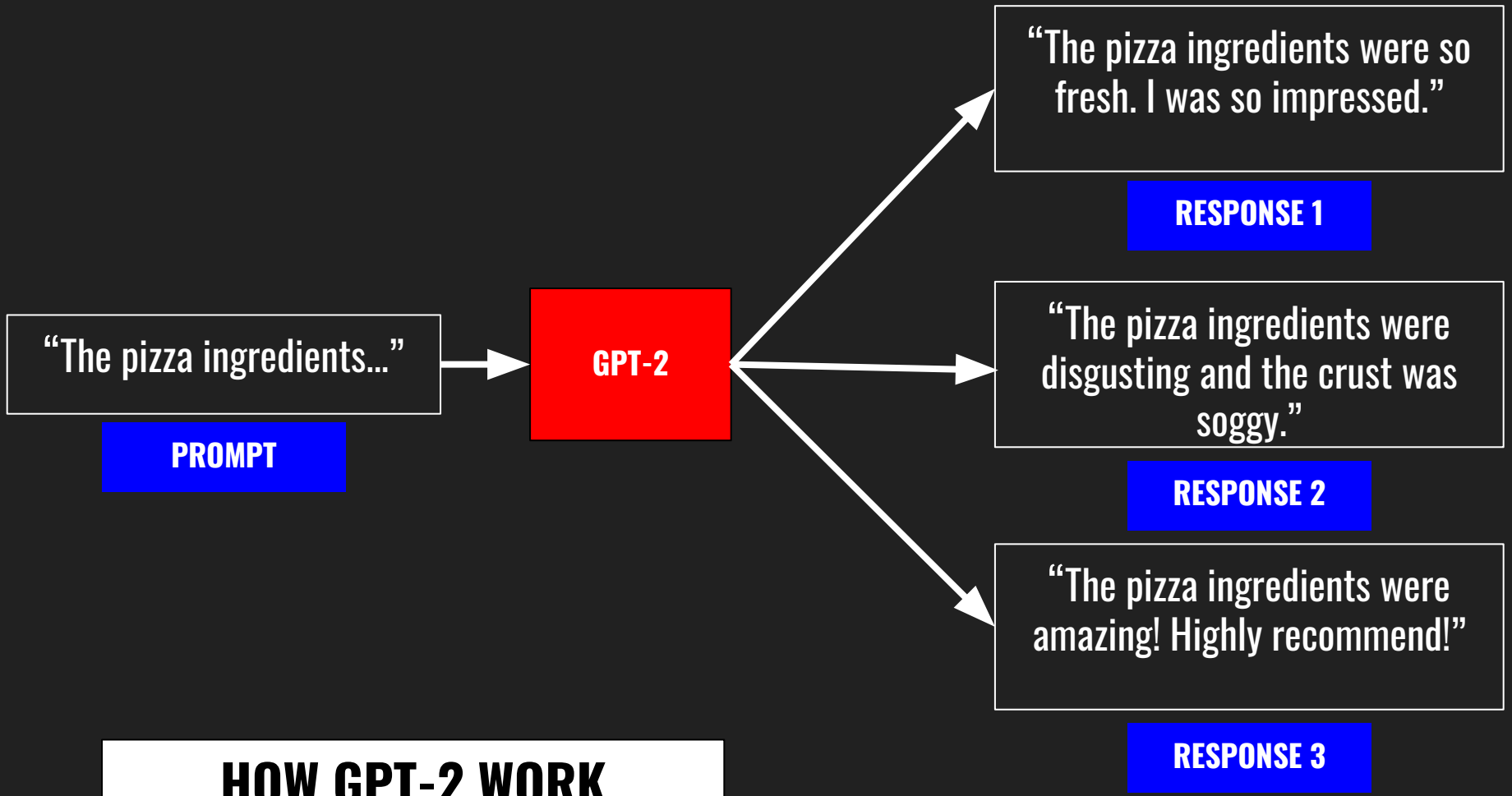
450 Genuine Observations

10,930 Synthetic Observations

11,380 Combined Observations

GPT-2 model was used to create synthetic reviews

GPT-2 is a transformer model trained on **8 million** web pages that predicts text.



HOW GPT-2 WORK

Prompt Design Principle:

Expand the Data

Don't Distort the Data

Most Used Trigrams, Bigrams, Words

“Very professional”

“Transmission Flush”

“Pepperoni Pizza”

“Ingredients”

“Never Coming Back”

“Not Very Happy”

“Highly Recommend”

“Repair”

“Sauce”

“Leaky Oil”

“Great Service”

“Amazing Crust”

“Disappointed”

“Transmission”

“Happy”

“Soggy Crust”

Words	80% - 100%	50% - 80%	25% - 50%
Trigrams	"Was very gross", "Bad customer service", "Pizza was cold"	"Not coming here", "Ingredients were stale", "Would not eat"	"Not very happy", "small portion size", "very salty taste"
Bigrams	"Soggy Crust", "Late delivery", "not happy"	"Dirty Table", "Money back",	"Very Greasy", "Never again", "was burned"
Words	"Pizza", "Angry", "Fresh", "Cheese"	"Incompetent", "Sauce", "Crust", "Nasty", "Delivery"," Service"	"Manager", "disgusting", "Stale"

Prompt 1: Pizza was Nasty
Prompt 2: Not very happy with service
Prompt 3: Soggy Crust was very gross

Table 1: Sample of GPT-2 Prompt Aid Tool

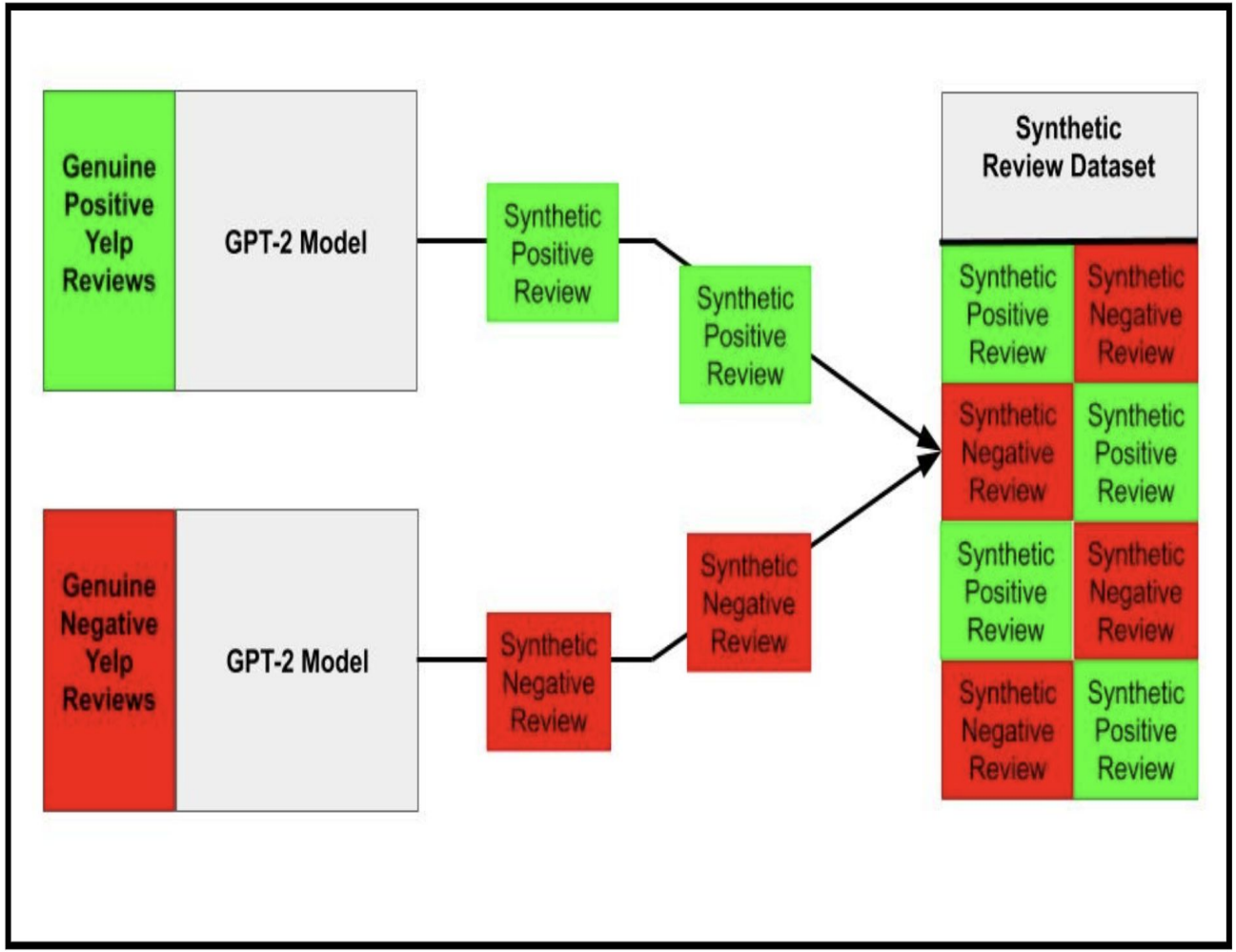


Figure 2: Synthetic Review Generation and Dataflow

Combining the Data

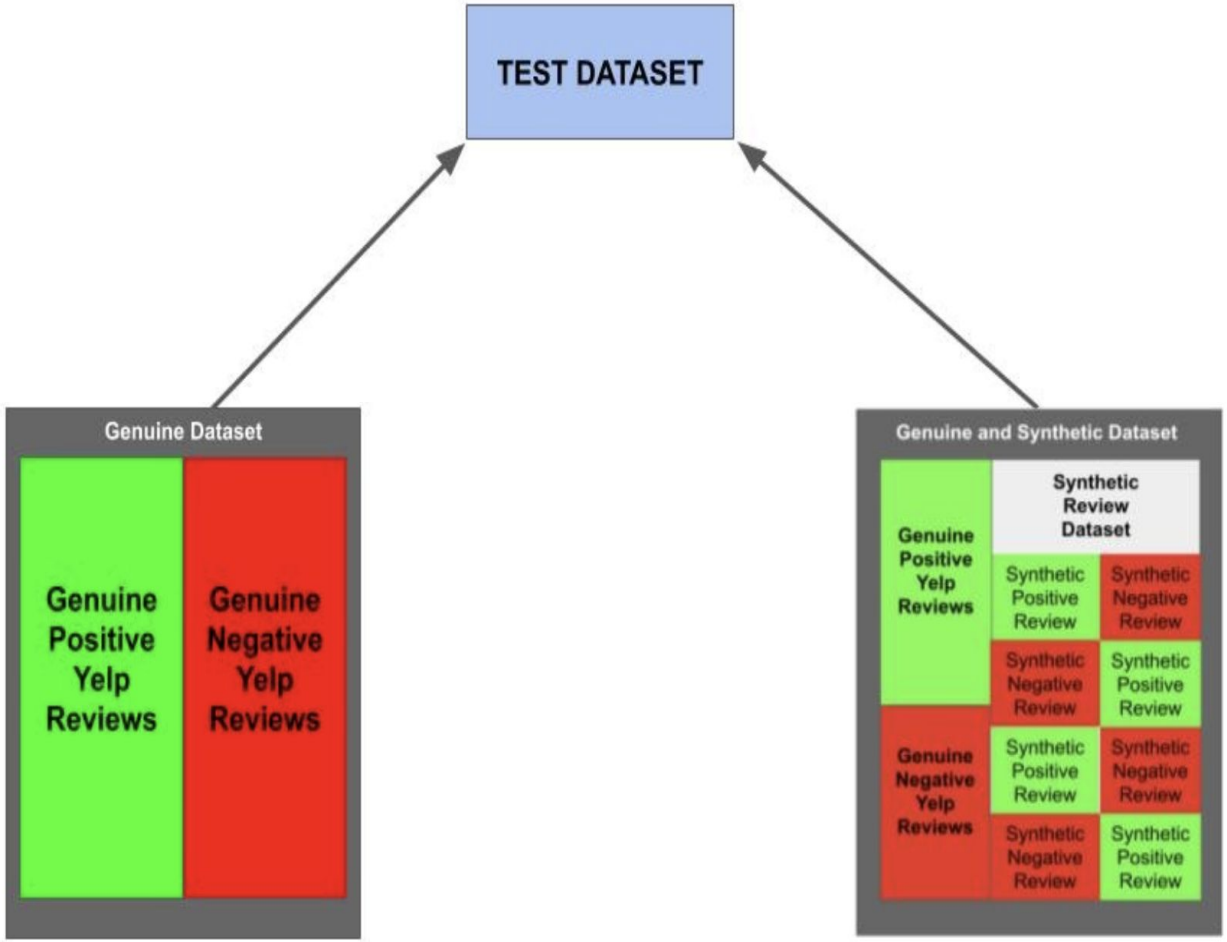
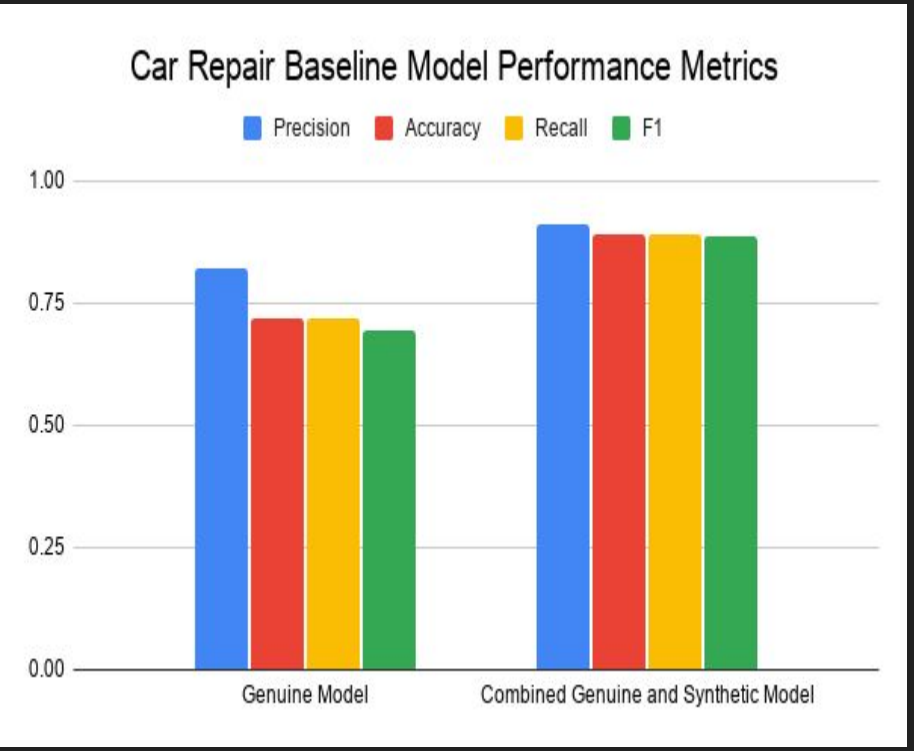


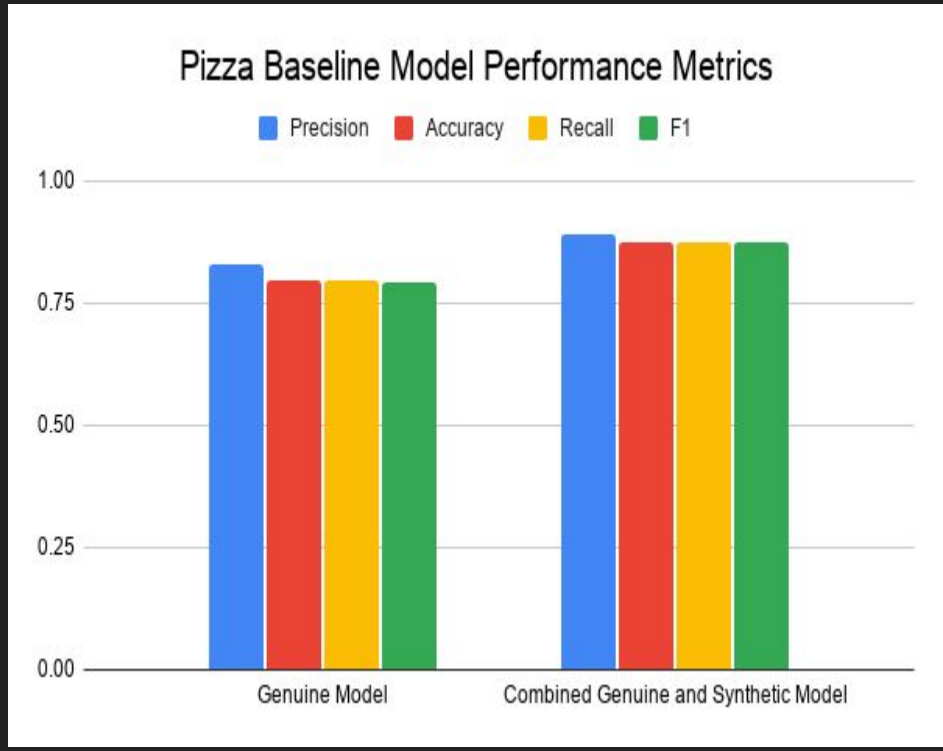
Figure 7: Baseline Model Testing

Data Leakage prevention

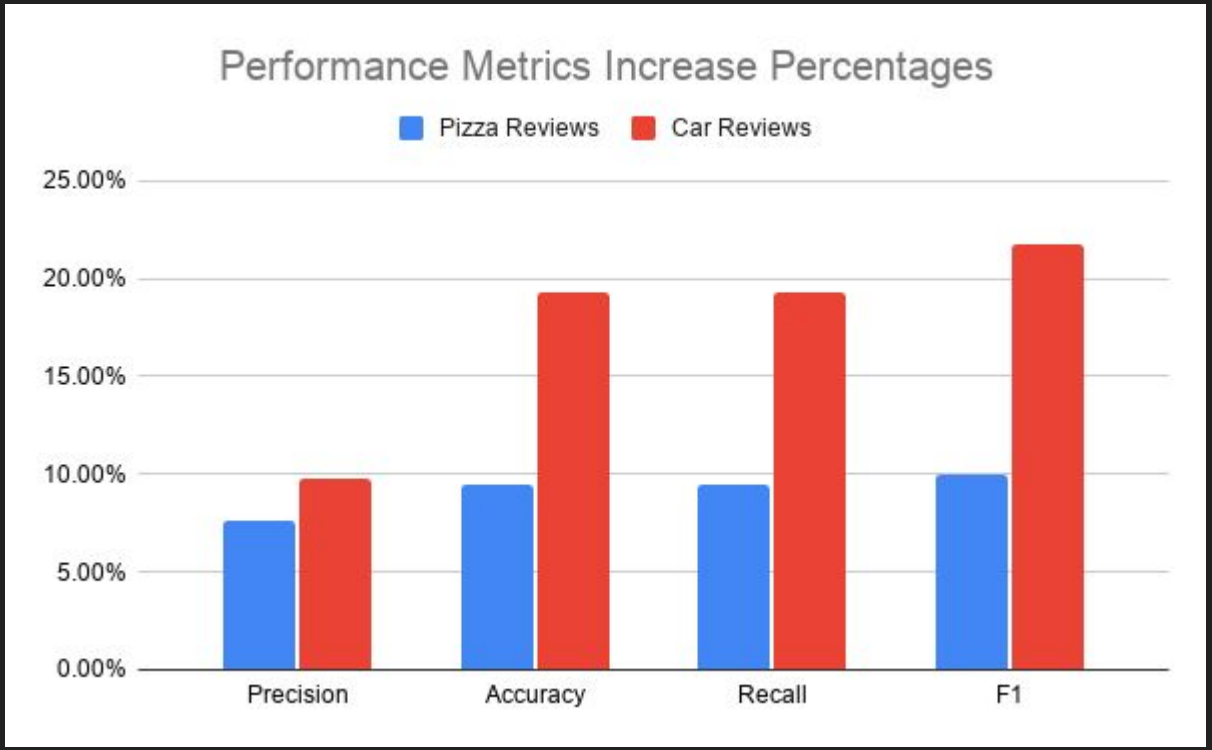
Took the test data directly from the Yelp Dataset

Performance Results





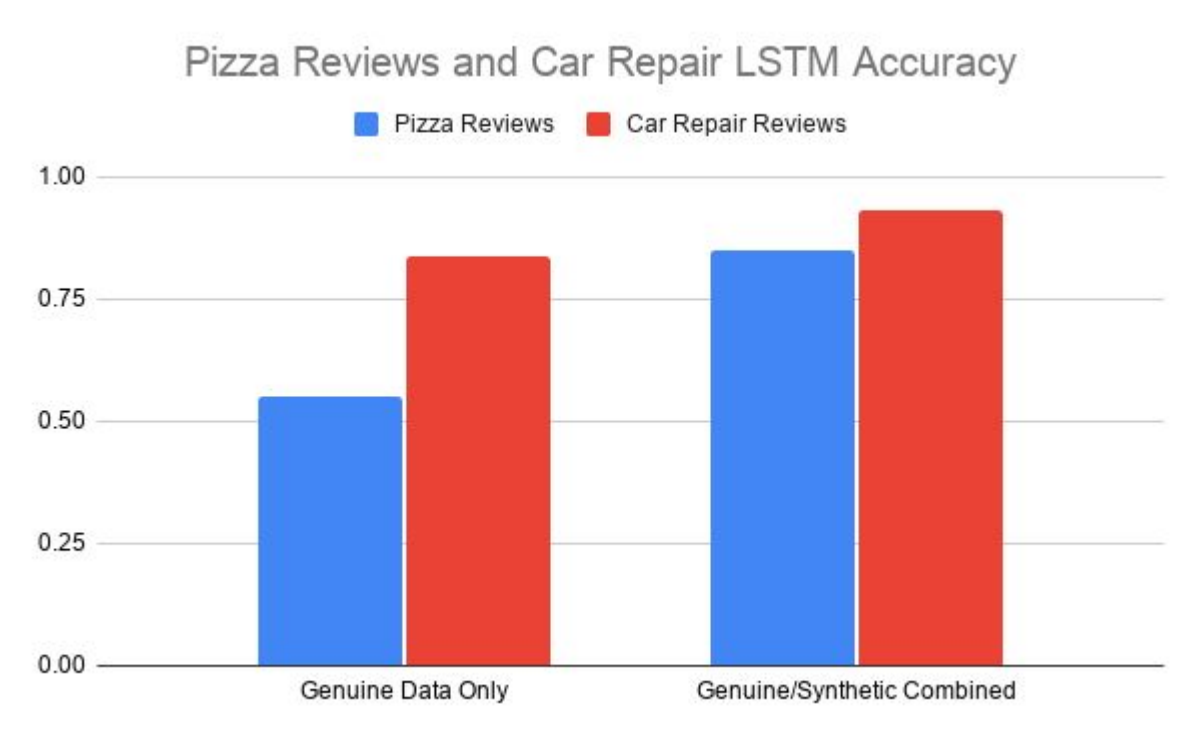
Naive Bayes Car Repair and Pizza Models



Performance Improvement of Models by Percentage

Key Observations

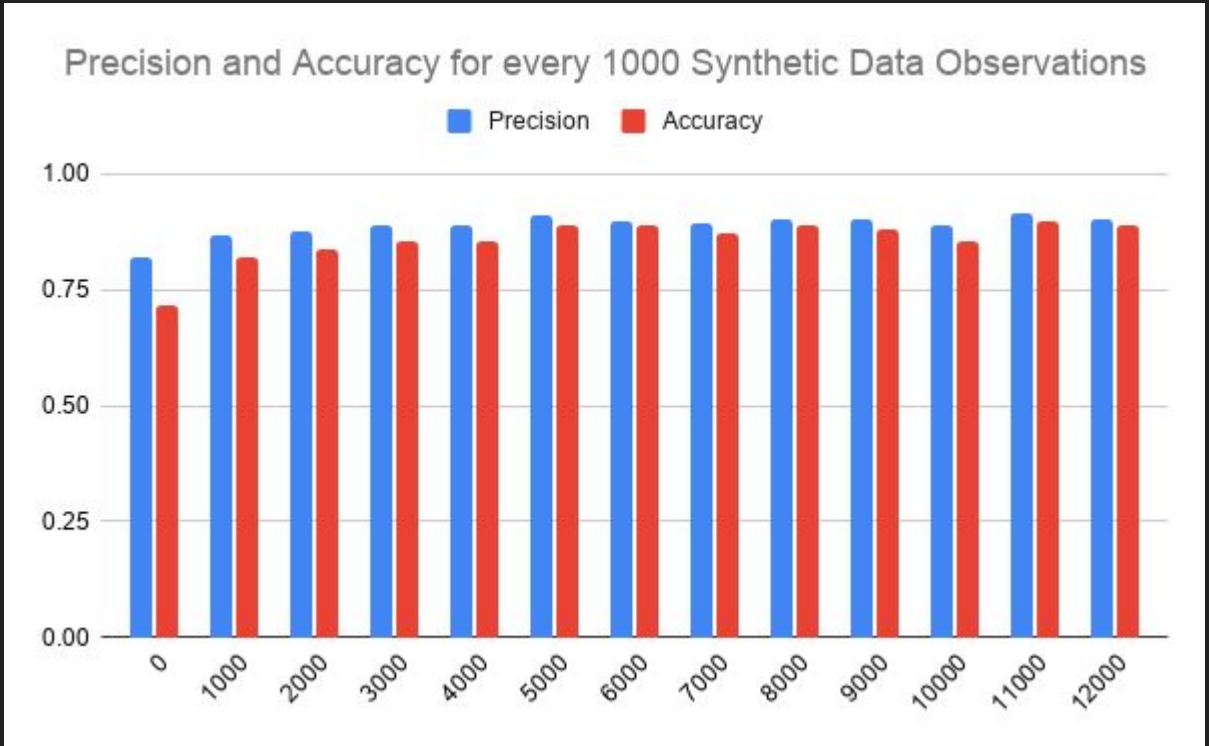
- The Car Repair Model's Accuracy improved by almost 20%
- The Pizza Model's Accuracy improved by almost 10%



LSTM Models

Key Observations

- **Both models improved** with synthetic data.
- The Car Repair LSTM Model's accuracy was 93%. **The highest accuracy of all the models.**



Car Repair Model Incremented by 1000

Key Observation

- The model stopped consistently increasing at 5000 synthetic observations

Why You May Want To Use Synthetic Data

Save Money

In 2018,
Companies spent
$19.2 Billion
on Data
Acquisition
Activities

Save Time

GPT-2 was trained on **40 GB** of data. **How long would it take you to gather that much data?**

Simulate Sensitive Model Data

Synthetic Data can be used to prototype Models that will handle Sensitive or Classified Data.

Simulate Rare Situations

Synthetic Data can be used to build models on rare situations.

Potential Use Cases

Contract Reviews

Models can be trained to identify rare situations to flag for reviewers.

Specific Security Threats

Government Cybersecurity experts can train models to identify unique and specific online threats.

Advanced Brand Management

Brand Managers can train models to identify specific events that may threaten the brand.

Artificial Entity Sentiment Analysis

Asking 1000 people
20 Questions

<

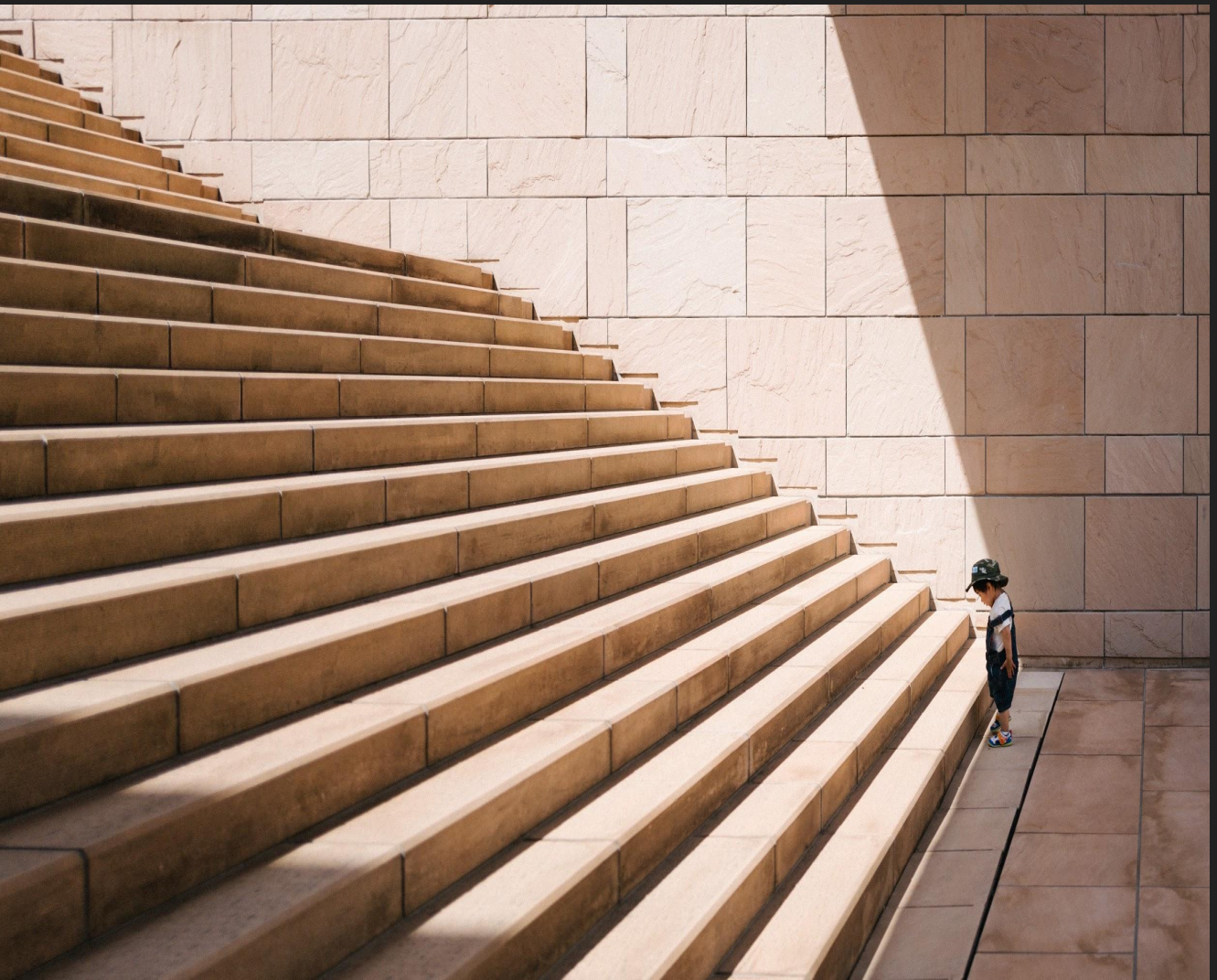
Asking 5 Artificial people
500k Questions

Rapid Government AI Prototyping

Government Technologist working in Intelligence, Defense, and Various Government Agencies can quickly prototype AI concepts without compromising data security or going through expensive procurement processes.

Healthcare AI Innovation

Government Agencies like the Veterans Affairs can test various ML models without compromising patient information.



Next Steps:

1. Develop out prototypes of the Use Cases

2. Collaborate with other researchers, companies, or Government Agencies

Questions?

linkedin.com/in/dewaynewhitfield/

@Dewayne_W