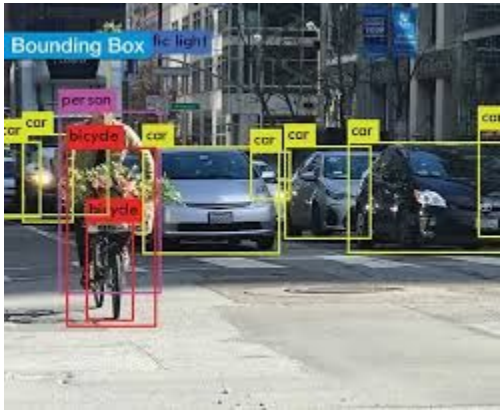
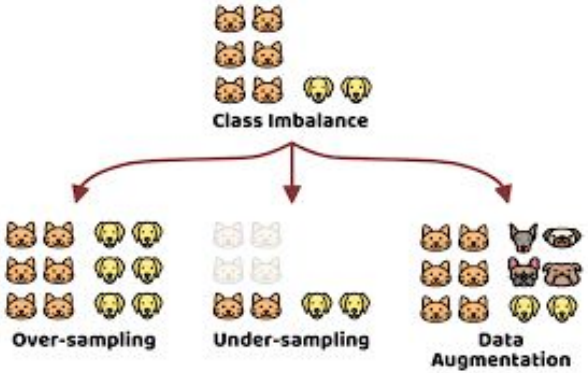
Impact of Synthetic Data on FastText Sentiment Classification

Shri Nidhi, Alice Jung, Edmund Chao, Angela Chen

Problem

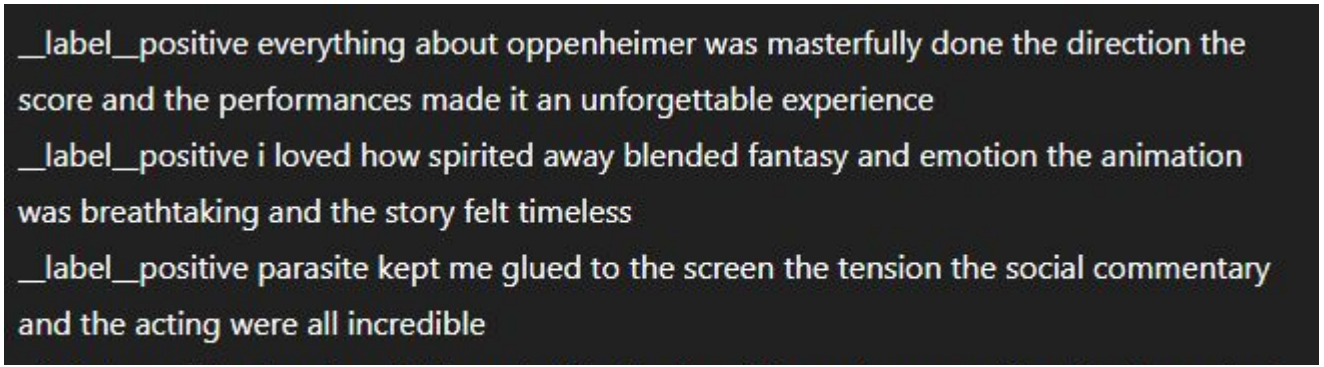
- High-quality training data can sometimes be scarce
  - Privacy and security challenges in fields like finance and healthcare
  - Manual annotation is labor intensive
  - Class imbalance and/or low data diversity





Potential Solution

- Generate synthetic training data
  - Can create as much data as you want
  - Output can be already properly formatted
- Challenges?
  - Bias
  - Limited diversity/redundant data



Our Idea: Augmentation via High-Quality Synthetic Data

- **Core Concept:**
  Use powerful **Large Language Model (GPT-5)** to generate large volume of high-quality, domain-specific synthetic reviews to augment limited real training data.

- **Pipeline:**
  Data: Real IMDB Data/ Synthetic Data (generated by GPT-5) → Clean/Format → Submodular Data Filtering → FastText Training

- **Key Advantage:**
  Synthetic data provides variability and volume that small real dataset lacks, potentially improving the model's ability to generalize to new, unseen domains (Amazon reviews).

Dataset Setup

Dataset Name	Composition	Total Size	Purpose
1. Real 5k	**5,000** Real IMDB Reviews (50/50 split)	5,000	Baseline performance model
2. Synthetic 5k	**5,000** Synthetic Reviews (50/50 split)	5,000	Performance of pure augmentation
3. Hybrid 5k	**2,500** Real + **2,500** Synthetic	5,000	Balanced augmentation test

This setup directly isolates the variable we want: the **effect of synthetic data composition** on the final model's performance.

Data Generation

- Accessed the GPT-5 model via the ChatGPT webapp
- Generated IMDB style reviews in batches of 50
- Encouraged the model to generate diverse output

```
Generate 10 IMDB-style movie reviews with a positive sentiment rating.
Vary the length and content of the reviews below. Randomly
incorporate references to actual movie content. Each review must
follow these rules:
- Use lowercase letters only
- Remove all punctuation (no . , ! ? ; : " ' or any symbols)
- Normalize spaces (no double spaces)
- Each review must be realistic and coherent
- Output each example in FastText format:
  __label__positive <review text>
  or
  __label__negative <review text>
```

Data Generation Cont.

- Used few-shot prompting to guide generation
- 1 real positive, 1 real negative, 1 synthetic positive, 1 synthetic negative example

```
Examples of reviews in the proper formatting:
__label__positive this movie was absolutely amazing and touching
__label__negative i honestly can't believe i wasted two hours on this
everything about it was awful
```

Data Generation Cont.

- Just looking at the generated reviews by eye, the synthetic reviews are nowhere near as diverse as real data

__label__positive i thought everything everywhere all at once was such a creative and emotional film the performances were outstanding and the story felt fresh and original
__label__positive the shawshank redemption continues to inspire me every time i watch it the friendship between andy and red is one of the best in cinema
__label__positive i really enjoyed top gun maverick it captured the spirit of the original while delivering even better action scenes and heartfelt moments
__label__positive inside out was such a clever and touching movie the idea of emotions having their own personalities was beautifully done and made me tear up
__label__positive the lord of the rings trilogy still feels timeless the visuals and music are stunning and the journey of frodo and sam never gets old

Data Selection - Submodular Function (Facility Location)

Process:

1. TF-IDF Vectorization (5,000 features)

   → Numerical representation of review text

2. Similarity Matrix (5,000 × 5,000)

   → Precomputed cosine similarity for speed

3. Greedy Selection (3,000 iterations)

   → Each iteration: pick example with max coverage gain

   → 1,500 positive + 1,500 negative

Data Setup - Submodular Function (Facility Location)

Dataset Name	Source (5k → 3k via submodular)	Total Size	Diversity Retention
1. Real 3k	Real IMDB Reviews (1.5k pos/1.5k neg)	3,000	73.2%
2. Synthetic 3k	Synthetic Reviews (1.5k pos/1.5k neg)	3,000	99.5%
3. Hybrid 3k	Real + Synthetic Mix (1.5k pos/1.5k neg)	3,000	91.1%

Experiments & Ablation Studies

Overview

- Two main research questions:
  - Can synthetic data replace real data?
  - Can submodular data selection improve or maintain model performance?

Objective

- Evaluate whether synthetic dataset can maintain competitive performance
- Evaluate whether a 3k subset selected from a 5k dataset can maintain competitive performance

Evaluation

Test Dataset

- 500 samples from the IMDB Reviews dataset (non-overlapping with training data)
- 500 samples from the Amazon Reviews dataset (out-of-distribution)
- Evaluate generalization across domains

Dataset Name	F1 Score (IMDB)	F1 Score (Amazon)
Real 5k	0.88	0.764
Synthetic 5k	0.63	0.598
Hybrid 5k	0.688	0.686

After submodular selected examples

Dataset Name	F1 Score (IMDB)	F1 Score (Amazon)
Real 3k	0.848	0.79
Synthetic 3k	0.628	0.588
Hybrid 3k	0.85	0.78

Evaluation: Real Data

- w/o filtering, performed the best across all test sets
- w/ filtering, performed on par w/ hybrid data
- Good data diversity -> better performance, even with OOD data
- However, it overfits more than synthetic/hybrid data

Dataset Name	F1 Score (IMDB)	F1 Score (Amazon)
Real 5k	**0.88**	**0.764**
Synthetic 5k	0.63	0.598
Hybrid 5k	0.688	0.686

After submodular selected examples

Dataset Name	F1 Score (IMDB)	F1 Score (Amazon)
Real 3k	0.848	**0.79**
Synthetic 3k	0.628	0.588
Hybrid 3k	**0.85**	**0.78**

Evaluation: Synthetic

- The Synthetic-only dataset underperforms compared to the other datasets, both before/after submodular
- High redundancy, Low diversity, Poor Generalization
  - Some performance drop on OOD data

Dataset Name	F1 Score (IMDB)	F1 Score (Amazon)
Real 5k	0.88	0.764
Synthetic 5k	0.63	0.598
Hybrid 5k	0.688	0.686

After submodular selected examples

Dataset Name	F1 Score (IMDB)	F1 Score (Amazon)
Real 3k	0.848	0.79
Synthetic 3k	0.628	0.588
Hybrid 3k	0.85	0.78

Evaluation: Hybrid

- w/o filtering, it performs in between the real and synthetic dataset
  - Negligible performance drop on OOD data
- w/ filtering, it shows notable improvement
  - Competitive with Real 3k
  - Submodular function chose the real examples, as they are more diverse than the synthetic examples

Dataset Name	F1 Score (IMDB)	F1 Score (Amazon)
Real 5k	0.88	0.764
Synthetic 5k	0.63	0.598
Hybrid 5k	0.688	0.686

After submodular selected examples

Dataset Name	F1 Score (IMDB)	F1 Score (Amazon)
Real 3k	0.848	0.79
Synthetic 3k	0.628	0.588
Hybrid 3k	**0.85**	**0.78**

Conclusion

Q: Does synthetic data impact the performance of sentiment classification models?
A: Yes. Synthetic data does affect classifier performance. But synthetic data alone is insufficient.

- Combining with real data+submodular selection can improve robustness and OOD generalization

- Our results suggest that this mixture, when appropriately filtered, does not harm model performance and can even improve it by introducing beneficial diversity