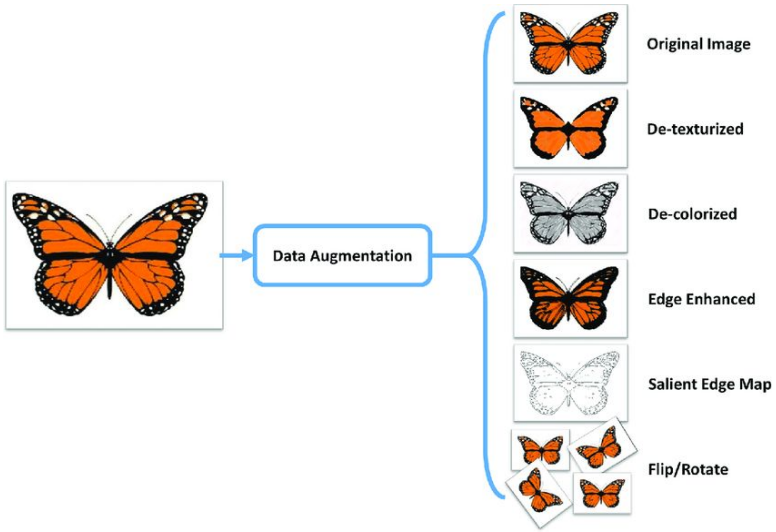
Using LLMs for Generating Tweet Data in Sentiment Analysis

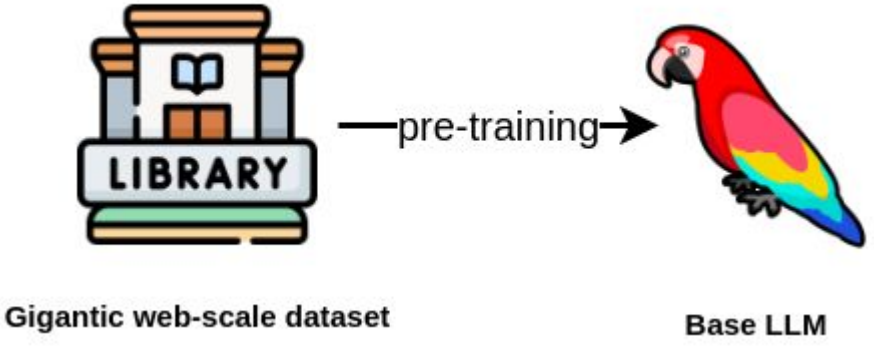
Presentation by Xianle Feng, Xinhan Yang, Haijiao Tao, Haorui Wang

Georgia Tech

Background

- Data augmentation can increase data diversity and overcome data scarcity.
- Several renowned generative models, such as VAEs, GANs, and Diffusion models, require training on authentic datasets.
- Large language models (LLMs) are pre-trained on world knowledge.





Georgia Tech

Motivation

- Sentiment analysis is crucial for identifying sentiment polarities in texts across e-commerce, blogs, and social media, but faces the challenge of costly and time-consuming data labeling.

- LLMs are encoded with extensive world knowledge, offering a valuable tool in data generation. By leveraging LLMs for this purpose, we can reduce human effort and achieve more effective outcomes.

- We want to explore whether we could ask LLMs directly to generate new data for sentiment analysis through different prompting techniques.

Related Work

- LLM as data generators: Generating tabular data [Borisov et al., 2022], relation triplets [Chia et al., 2022] and sentence pairs [Schick & Schütze, 2021].
    - LLM as training data generators: Generating Question Answering chain-of-thought reasoning steps [Ho et at., 2022] and instruction data [Wang et al., 2023].
- AttrPrompt [Yu et al., 2023]: Utilizing attributed prompts to produce diverse and attributed generated data, specifically on topic classification tasks.

Dataset

- Twitter Sentiment Dataset
  - Categorize tweets by sentiment
  - Labels: positive, neutral, negative
  - example:
    THANK YYYYYYYYYOOOOOOOOOOUUUUU! – positive
    what interview! leave me alone – negative

- Covid-19 Twitter Dataset
  - Categorize tweets related to covid-19 by sentiment
  - Labels: Extremely positive, Positive, Neutral, Negative, Extremely negative
  - example:

    if i don't die bc of covid-19, I'll die from the lack of food bc YALL DONE TOOK EVERYTHING IN THE SUPERMARKET – extremely negative

    Covid-19: No need for panic buying, food available at all times ¬ñ PM assures – neutral

    Hi, COVID-19. Thanks for making me do more online shopping. – positive

Georgia Tech

Initial exploration

- AttrPrompt is not suitable for tweets generation.
  - These attributes are abstract and vague.

Which attribute dimensions do you consider vital in determining the sentiment of a tweet in COVID-19?

1. Emotion: Understanding the emotional tone of the tweet is crucial. Tweets can express a range of emotions such as fear, anxiety, anger, sadness, hope, or optimism.
2. Context: Consider the context of the tweet. Is it providing information, sharing personal experiences, expressing opinions, or asking questions? The context can influence the sentiment interpretation.
3. Keywords and Phrases: Analyze the specific words and phrases used in the tweet. Certain words or hashtags may indicate positive or negative sentiment. For example, words like "vaccine," "recovery," "lockdown," or "pandemic" can carry sentiment implications.
……

- Tweets are much shorter than news articles. They lack consistent elements such as style or location.

- Short phrases or single sentences in tweets might be highly ambiguous, so tweets often lack sufficient context for accurately determining the sentiment.

Georgia Tech

Method

- Fact-imagination prompts:
  - Instead of analyzing the attributes, we ask LLMs to imagine facts or events that might lead people to compose such tweets.
  - Let LLMs analyze and reflect on the **reasons** and **events** behind these specific labels. This step could be also done **by human**.
  - Instructing the LLMs to generate new data based on those reasons and events.

Please analyze the dataset I upload. Then reflect on the reasons behind these specific labels. For example, people often post on Twitter in response to certain events or situations. Please imagine a variety of potential events that could lead to such posts for {sentiment class}.

Personal Loss or Grief: Tweets expressing sorrow over the loss of a loved one, a pet, or mourning significant life changes.
Health Issues: Sharing struggles with physical or mental health, or expressing frustration over a health system or treatment.
Work-related Stress: Complaints about workload, challenging work environments, unsatisfactory job conditions, or conflicts with colleagues or superiors.
……

Suppose you post a tweet on Twitter. Then you play the role of a sentiment analyst and analyze the sentiment of the tweets you publish.
Here is some example data: {100 random sample of original dataset of {sentiment cass}}
Here are some potential event that might lead to such post, please generate tweets base on these situations:
{situations given by LLM or human}
Here are your requirements:
Please generate 100 new samples.
Please generate only {sentiment class} samples
Please follow the same format of example data

textID,text,sentiment
1,My computer keeps crashing and it's driving me crazy!,negative
2,The social event I was looking forward to got canceled, and now I'm feeling really disappointed.,negative
3,I'm so stressed with work deadlines, it's overwhelming.,negative
……

Baseline Prompting Method

- Simple prompt
  - directly ask GPT to generate new data

Suppose you post a tweet on Twitter. Then you play the role of a sentiment analyst and analyze the sentiment of the tweets you publish.
Here is some example data: {100 random sample of original dataset}
Here are your requirements:
Please generate 100 new samples.
Please follow the same format of example data

textID,text,sentiment
1,I can't believe this happened,neutral
2,Feeling really down today, everything seems hopeless",negative
3,This movie made me laugh so hard,positive
……

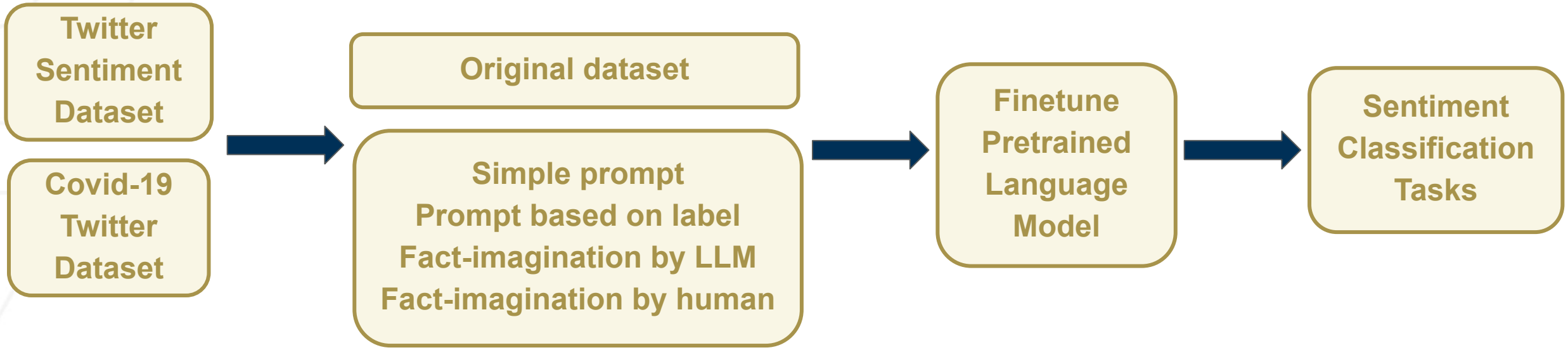
- Prompt based on label
  - directly ask GPT to generate new data for one class

Suppose you post a tweet on Twitter. Then you play the role of a sentiment analyst and analyze the sentiment of the tweets you publish.
Here is some example data: {100 random sample of original dataset of {sentiment class}}
Here are your requirements:
Please generate 100 new samples.
Please generate only {sentiment class} samples
Please follow the same format of example data

textID,text,sentiment
1,Celebrating my graduation with friends and family,positive
2,Having a great time at a music festival.
#goodvibes,positive
3,Spent the day volunteering and making a difference.
#givingback,positive
……

Georgia Tech

Experiments



LLM for Data Augmentation: GPT 3.5-turbo 16k

Pretrained Language Model: distilBERT

Dataset: 500 sample from original dataset, augmented with 500 sample for each prompt method respectively

Georgia Tech

Dataset Diversity

Measuring the diversity of datasets: Average Pairwise Similarity (APS)

Method	twitter		covid-19	
	In-label APS	APS	In-label APS	APS
Original	*0.127*	*0.119*	*0.313*	*0.308*
Simple Prompt	0.181	**0.141**	0.358	0.342
Label-Based Prompt	0.195	0.156	0.383	0.363
Fact-Imagination by LLM	**0.180**	0.145	0.353	0.335
Fact-Imagination by Human	0.185	0.142	**0.332**	**0.306**

Data augmentation by LLM reduces the diversity of datasets

Fact-Imagination prompts usually yield higher diversity

Label-based prompt has low sample diversity

Georgia Tech

Dataset Diversity

APS per label in the covid-19 dataset

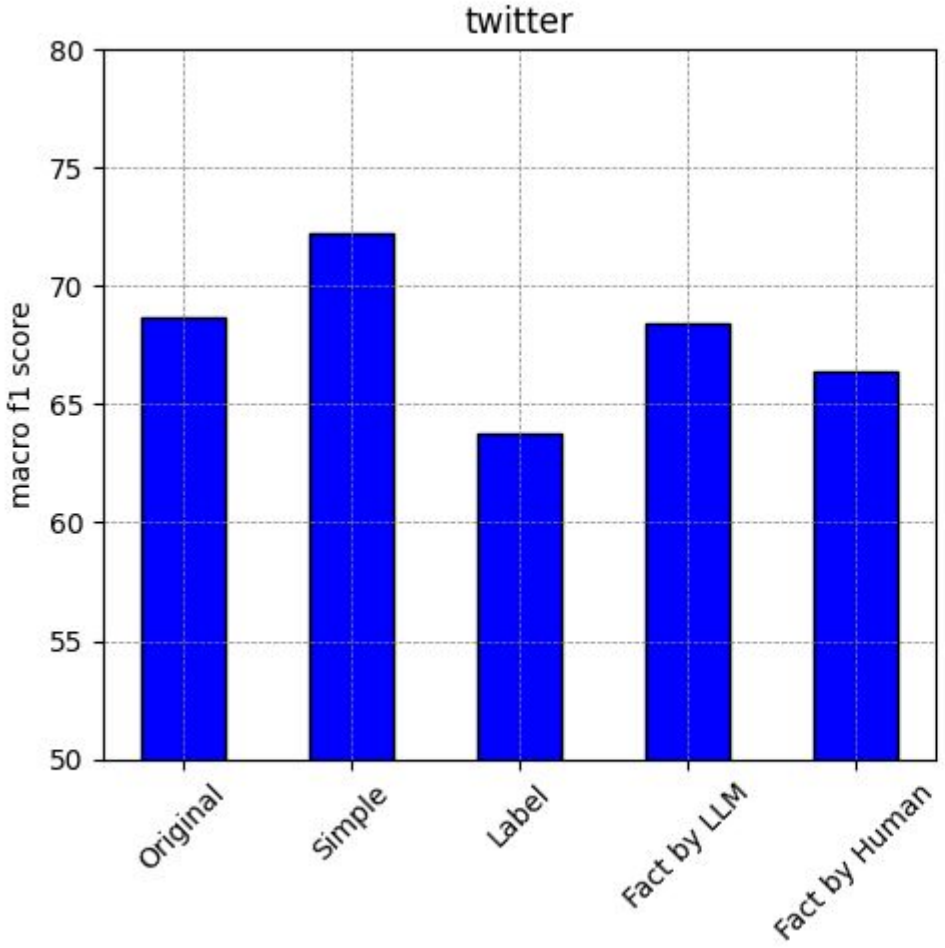
Method	Extremely Negative	Negative	Neutral	Positive	Extremely Positive
Original	*0.329*	*0.292*	*0.341*	*0.318*	*0.289*
Simple Prompt	**0.335**	0.318	0.38	0.338	0.356
Label-Based Prompt	0.405	0.396	0.418	0.36	0.335
Fact-Imagination by LLM	0.411	0.342	**0.362**	**0.333**	0.327
Fact-Imagination by Human	0.395	**0.271**	0.387	0.335	**0.276**

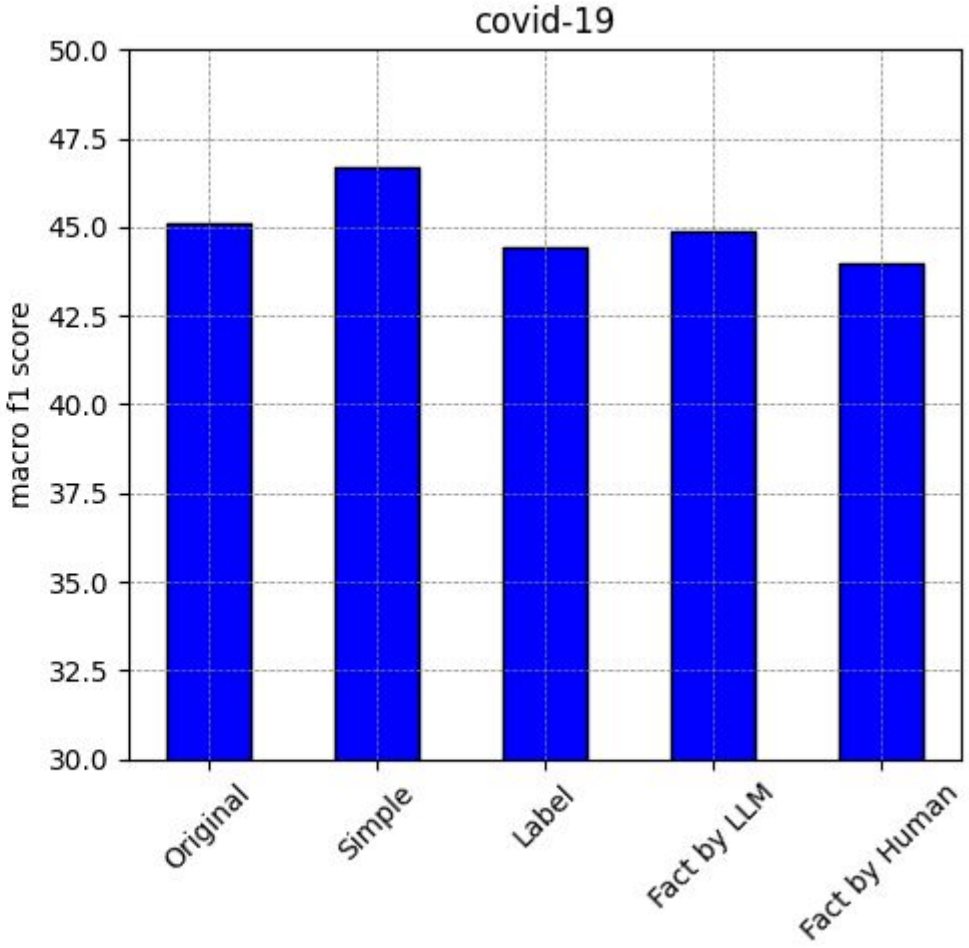
Fact-Imagination prompts yield higher sample diversity for each label

With the help of human knowledge and expertise, the diversity on some labels are higher than the original dataset

Georgia Tech

Classification Result





Georgia Tech

Classification Result

Method	twitter		covid-19	
	Accuracy	F1	Accuracy	F1
Simple Prompt	71.90 +3.40	72.26 +3.58	46.29 −1.31	46.71 +1.59
Label-Based Prompt	63.40 −5.10	63.76 −4.92	45.47 −2.13	44.41 −0.71
Fact-Imagination by LLM	68.10 −0.40	68.39 −0.29	45.87 −1.73	44.90 −0.22
Fact-Imagination by Human	65.90 −2.60	66.37 −2.31	46.39 −1.21	43.98 −1.14

possible reasons:

- Fact-imagination prompts make LLM generate out-of-scope data.
- LLM may generate data with incorrect labels.

Georgia Tech

Conclusion

We explore a new prompting method, fact imagination prompt, on twitter data augmentation for sentiment analysis tasks.

Only increasing the diversity of data may not be sufficient for improving the fine-tuned model, as it overlooks data accuracy.

When implementing data augmentation using LLMs, it's crucial to balance the diversity and accuracy of the generated data.

Future work:
Evaluate the effect of data accuracy, and whether filtering out low-accuracy data would help model performance.

Questions

Georgia Tech

References

Borisov V, Seßler K, Leemann T, et al. Language models are realistic tabular data generators[J]. arXiv preprint arXiv:2210.06280, 2022.

Chia Y K, Bing L, Poria S, et al. RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction[J]. arXiv preprint arXiv:2203.09101, 2022.

Schick T, Schütze H. Generating datasets with pretrained language models[J]. arXiv preprint arXiv:2104.07540, 2021.

Wang Y, Kordi Y, Mishra S, et al. Self-instruct: Aligning language model with self generated instructions[J]. arXiv preprint arXiv:2212.10560, 2022.

Yu Y, Zhuang Y, Zhang J, et al. Large language model as attributed training data generator: A tale of diversity and bias[J]. arXiv preprint arXiv:2306.15895, 2023.