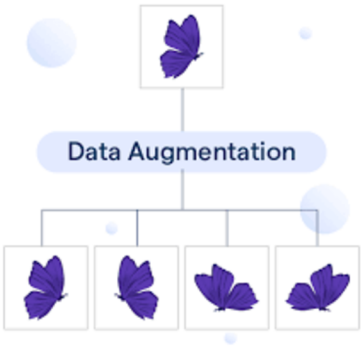
DATA AUGMENTATION FOR TEXT CLASSIFICATION USING LARGE LANGUAGE MODELS

SOUMYA BHARATHI VETUKURI - 016668964

01 Dec, 2024



My Presentation is based on the following research paper:

Data Generation Using Large Language Models for Text Classification: An Empirical Case Study

By: **Yinheng Li, Rogerio Bonatti, Sara Abdali, Justin Wagle, Kazuhito Koishida**

Link: https://arxiv.org/pdf/2407.12813

Other References:
https://medium.com/nlplanet/two-minutes-nlp-a-taxonomy-of-data-augmentation-for-text-classification-52c96f332bad

A Survey on Data Augmentation for Text Classification (https://arxiv.org/pdf/2107.03158)

AGENDA OVERVIEW

1 DATA AUGMENTATION

2 DATA AUGMENTATION PIPELINE

3 PROMPTS FOR TOPIC GENERATION

4 EXPERIMENTS

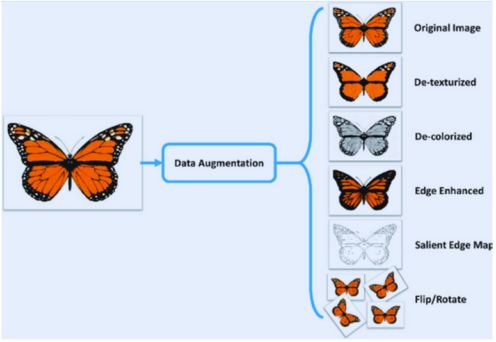
5 RECOMMENDATIONS & FINDINGS

6 EXAMPLE PROMPTS

7 DATA GENERATION TECHNIQUES IN PRACTICE

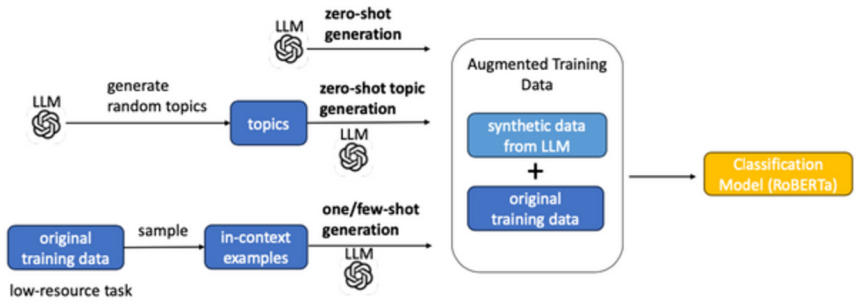
8 CONCLUSION

DATA AUGMENTATION



- Method that utilizes existing data to generate additional training data without collecting more data.
- Cost-efficient alternative to human-labeled data.
- Goal: Increase diversity of existing data by exposing the model to unseen data.
- Tasks use natural language understanding models with transformer encoder architecture.
- LLMs when prompted properly can generate better quality data similar to human text.
- Data augmentation in text classification leads to better models as the models see more linguistic patterns during training

PIPELINE FOR DATA AUGMENTATION USING LLM



ZERO-SHOT IN-CONTEXT

Provide the task description in the prompt and ask the LLM to generate a similar example.

ONE-SHOT IN-CONTEXT

Provide Task description and an example, prompting LLM to generate similar example

FEW-SHOT IN-CONTEXT

Provide Task description and few examples, prompting LLM to generate similar example.

ZERO-SHOT TOPIC IN-CONTEXT

Provide Task description and one sample topic, prompting LLM to generate similar example

PROMPTS FOR TOPIC GENERATION

Task	Role	Message
BoolQ, RET, NYT, SST-2, Emo	System	You are an AI assistant that generates random topics. There is no limit on the number of topics you can generate.
BoolQ, RET, NYT	User	Please generate 500 topics
BoolQ, RET, NYT	LLM	Output example: The world's most beautiful sculptures, The role of technology in modern education ...
SST-2, Emo	User	Please generate 500 twitter post topics
SST-2, Emo	LLM	Output example: Lunch break, Online dating ...
Review	System	You are an AI assistant that knows Amazon product categories. The user will ask you to generate a list of categories. It is your responsibility to generate the entire list of categories.
Review	User	Please generate 500 amazon different product categories
Review	LLM	Output example: Baby Products, Clothing, Jewelry ...

EXPERIMENT

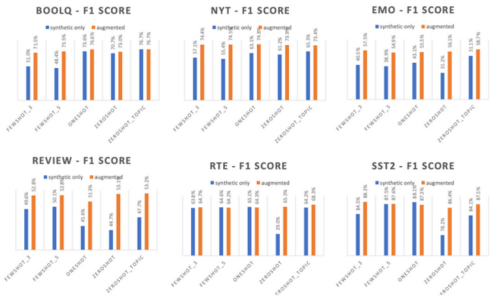
Corpus	Training Size	Test Size	Task	Metrics	Domain
SST-2	67k	1.8k	Binary Classification	F1	Movie Reviews
EMO	16k	2k	Multi-class Classification	Macro-F1	Twitter
NYT	256k	3k[2]	Multi-class Classification	Macro-F1	News
Review	200k	5k	Multi-class, Ordinal Regression	Macro-F1	Amazon Review
RTE	2.5k	3k	Pair Classification, Question Answering	Macro-F1	News, Wikipedia
BoolQ	16k	3.2k	Pair Classification, Question Answering	Macro-F1	News, Wikipedia, Web Query

TASKS SST-2 (sentiment analysis), Twitter Emotional Classification(EMO), New York Times News Classification(NYT), Review(Amazon Review Classification), Recognizing Textual Entailment(RTE) and BoolQ (binary question answering)

METRICS Accuracy, F1 and Macro F1 Scores

MODEL ROBERTa, Transformer Based Language Model

FINDINGS



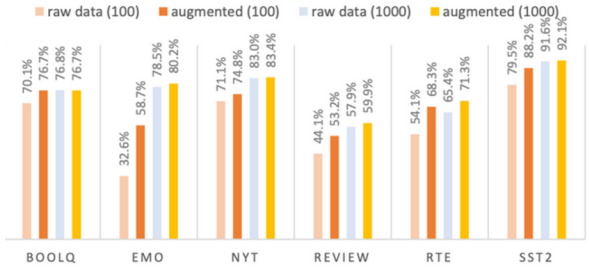
1. **Data Augmentation = Raw Data + Synthetic Data**

2. **Trivial Questions are the Trick**

Did the Mars Exploration Rover mission **only** involve one rover? – False
Did scientists in the 20th century make **no significant** discoveries or advancements? – False

	Trivial Q. Count		F1 Score			
	Raw	Rephrased	Raw (SD)	Raw (SD)	Raw (AD)	Rephrased (AD)
Zero-Shot Topic	230	208	0.19	**0.77**	0.75	**0.77**
One-Shot	131	74	0.38	0.74	0.76	**0.77**
Few-Shot (3 ex.)	90	30	0.55	0.51	0.70	**0.72**
Few-Shot (5 ex.)	57	28	0.53	0.48	0.75	0.73
Zero-Shot	11	-	0.71	-	0.73	-
Raw Data	31	-	-	0.768	-	-

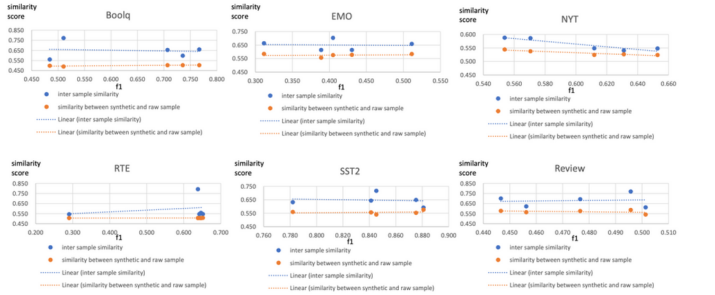
FINDINGS



3. Synthetic Data is best in Low-Resource Settings

4. Data Diversity Matters!!

5. Quality Over Quantity



PROMPTS USED FOR DATA GENERATION FOR EACH TASK:

Task	Prompt Type	Prompt
BoolQ	zero-shot	**Step 1** Please generate a random short passage. Passage: **Step 2** Please generate a True or False question based on the passage. The answer to the question must be [random([True, False])] Passage: [passage from step 1] Question:
BoolQ	zero-shot topic	**Step 1** Please generate a short passage about this topic: [topic sampled from a topic list] Passage: **Step 2** Please generate a True or False question based on the passage. The answer to the question must be [random([True, False])] Passage: [passage from step 1] Question:
BoolQ	one-shot	**Step 1** Please generate a Passage, a Question and the Label to the question following this example: [example from raw data: Passage, Question, Label] Please generate a similar passage. Passage: **Step 2** Please generate a True or False question based on the passage. The answer to the question must be [label from example in Step 1] Passage: [passage generated in Step 1] Question:
BoolQ	few-shot (3 or 5)	**Step 1** Please generate a Passage, a Question and the Label to the question. Here are some examples: [examples from raw data: Passage, Question, Label] Please generate a similar example. Make sure the question is a True or False question and the answer to the question is [random([True, False])]. Passage:
EMO	zero-shot	**Step 1** Please generate a twitter post with the emotion of [random(label)]. Text:
EMO	zero-shot topic	**Step 1** Please consider this topic for generation: [topic sampled from a topic list]. Please generate a twitter post with the emotion of [random(label)]. Text:
EMO	one-shot	**Step 1** The task is to predict the emotion of a twitter post. The emotion contains six categories: sadness, joy, love, anger, fear, surprise. Here is an example. Text: [example from raw data] Emotion: [example label from raw data] Please generate another example for the same emotion. Text:
EMO	few-shot (3 or 5)	**Step 1** The task is to predict the emotion of a twitter post. The emotion contains six categories: sadness, joy, love, anger, fear, surprise. Here are some examples: [examples: Text, Emotion] Please generate a twitter post with the emotion of [first label from examples]. Text:

Task	Prompt Type	Prompt
RTE	zero-shot	**Step 1** Given a premise and a hypothesis, a model needs to predict whether the hypothesis can be logically inferred from the premise. The response should be either True if the hypothesis can be inferred from the premise, or False if it cannot be inferred. Here is the output format: Premise: Hypothesis: Label: True or False Please generate an example where the Label is [random(label)]. Premise:
RTE	zero-shot topic	**Step 1** Given a premise and a hypothesis, a model needs to predict whether the hypothesis can be logically inferred from the premise. The response should be either True if the hypothesis can be inferred from the premise, or False if it cannot be inferred. Here is the output format: Premise: Hypothesis: Label: True or False Please generate an example about [premise] where the Label is [random(label)]. Premise:
RTE	one-shot	**Step 1** Given a premise and a hypothesis, a model needs to predict whether the hypothesis can be logically inferred from the premise. The response should be either True if the hypothesis can be inferred from the premise, or False if it cannot be inferred. Here is an example: Premise: [example premise] Hypothesis: [example hypothesis] Label: [example label] Please generate another similar example where the Label is [example label]. Premise:
RTE	few-shot (3 or 5)	**Step 1** Given a premise and a hypothesis, a model needs to predict whether the hypothesis can be logically inferred from the premise. The response should be either True if the hypothesis can be inferred from the premise, or False if it cannot be inferred. Here are some examples: [examples: Premise, Hypothesis, Label] Please generate a similar example. Make sure the label is [first label from examples]. Premise:
SST-2	zero-shot	**Step 1** Please generate a sentence that contains a [random(label)] sentiment. Sentence:
SST-2	zero-shot topic	**Step 1** Please consider this topic for generation: [topic from the topic list]. Please generate a sentence that contains a [random(label)] sentiment. Sentence:
SST-2	one-shot	**Step 1** The task is to predict whether the following sentence is positive or negative sentiment. Sentence: [example sentence] Label:[example label] Please generate a similar example on the same topic, including a Sentence and a Label. Sentence:
SST-2	few-shot (3 or 5)	**Step 1** The task is to predict whether the following sentence is positive or negative sentiment. [examples: Sentence, Label] Please generate a similar example, including a Sentence and a Label. Sentence:

PROMPTS USED FOR DATA GENERATION FOR EACH TASK:

Task	Prompt Type	Prompt
RTE	zero-shot	**Step 1** Given a premise and a hypothesis, a model needs to predict whether the hypothesis can be logically inferred from the premise. The response should be either True if the hypothesis can be inferred from the premise, or False if it cannot be inferred. Here is the output format: Premise: Hypothesis: Label: True or False Please generate an example where the Label is [random(label)]. Premise:
RTE	zero-shot topic	**Step 1** Given a premise and a hypothesis, a model needs to predict whether the hypothesis can be logically inferred from the premise. The response should be either True if the hypothesis can be inferred from the premise, or False if it cannot be inferred. Here is the output format: Premise: Hypothesis: Label: True or False Please generate an example about [premise] where the Label is [random(label)]. Premise:
RTE	one-shot	**Step 1** Given a premise and a hypothesis, a model needs to predict whether the hypothesis can be logically inferred from the premise. The response should be either True if the hypothesis can be inferred from the premise, or False if it cannot be inferred. Here is an example: Premise: [example premise] Hypothesis: [example hypothesis] Label: [example label] Please generate another similar example where the Label is [example label]. Premise:
RTE	few-shot (3 or 5)	**Step 1** Given a premise and a hypothesis, a model needs to predict whether the hypothesis can be logically inferred from the premise. The response should be either True if the hypothesis can be inferred from the premise, or False if it cannot be inferred. Here are some examples: [examples: Premise, Hypothesis, Label] Please generate a similar example. Make sure the label is [first label from examples]. Premise:
SST-2	zero-shot	**Step 1** Please generate a sentence that contains a [random(label)] sentiment. Sentence:
SST-2	zero-shot topic	**Step 1** Please consider this topic for generation: [topic from the topic list]. Please generate a sentence that contains a [random(label)] sentiment. Sentence:
SST-2	one-shot	**Step 1** The task is to predict whether the following sentence is positive or negative sentiment. Sentence: [example sentence] Label:[example label] Please generate a similar example on the same topic, including a Sentence and a Label. Sentence:
SST-2	few-shot (3 or 5)	**Step 1** The task is to predict whether the following sentence is positive or negative sentiment. [examples: Sentence, Label] Please generate a similar example, including a Sentence and a Label. Sentence:

Evaluation

The following prompts were used to evaluate the performance of LLMs on each task..

Task	Prompt Type	Prompt
RTE	zero-shot	**Step 1** Given a premise and a hypothesis, a model needs to predict whether the hypothesis can be logically inferred from the premise. The response should be either True if the hypothesis can be inferred from the premise, or False if it cannot be inferred. Premise: [premise], Hypothesis: [hypothesis], Label:
RTE	0/1/3/5-shot	**Step 1** Given a premise and a hypothesis, a model needs to predict whether the hypothesis can be logically inferred from the premise. The response should be either True if the hypothesis can be inferred from the premise, or False if it cannot be inferred. Here are some examples: [example premise, hypothesis, label] Premise: [premise], Hypothesis: [hypothesis], Label:
BoolQ	zero-shot	**Step 1** The task is to answer a question which is solely based on the content provided. Passage: [passage] , Question: [question], Label:
BoolQ	0/1/3/5-shot	**Step 1** The task is to answer a question which is solely based on the content provided. Here are some examples: [example passage, question, label] Passage: [passage], Question: [question], Label:
Review	zero-shot	**Step 1** The task is to predict the rating of an Amazon customer review based on the content. The rating ranges from 1 to 5, with 1 being the lowest and 5 being the highest. Text: [text] , Label:
Review	0/1/3/5-shot	**Step 1** The task is to predict the rating of an Amazon customer review based on the content. The rating ranges from 1 to 5, with 1 being the lowest and 5 being the highest. Here are some examples: [example text, label] Text: [text], Label:
NYT	zero-shot	**Step 1** The task is to predict the topic of a news headline. The topics include: 'sports', 'arts, culture and entertainment', 'business and finance', 'health and wellness', 'lifestyle and fashion', 'science and technology', 'politics', 'crime'. Text:[text], Label:
NYT	0/1/3/5-shot	**Step 1** The task is to predict the topic of a news headline. The topics include: 'sports', 'arts, culture and entertainment', 'business and finance', 'health and wellness', 'lifestyle and fashion', 'science and technology', 'politics', 'crime'. Here are some examples: [example text, label] Text: [text], Label:
EMO	zero-shot	**Step 1** The task is to predict the emotion of a Twitter text. The emotions include six categories: sadness, joy, love, anger, fear, surprise. Text: [text], Label:
EMO	0/1/3/5-shot	**Step 1** The task is to predict the emotion of a Twitter text. The emotions include six categories: sadness, joy, love, anger, fear, surprise. Here are some examples: [example text, label] Text: [text], Label:
SST-2	zero-shot	**Step 1** The task is to predict whether the given sentence has a positive or negative sentiment. Sentence: [sentence], Label:
SST-2	0/1/3/5-shot	**Step 1** The task is to predict whether the given sentence has a positive or negative sentiment. Here are some examples: [example sentence, label], Sentence: [sentence], Label:

USEFUL DATA GENERATION TECHNIQUES CURRENTLY IN PRACTICE

1.Condition on Label

- Saves effort in parsing the label
- User has control over label distribution in Synthetic data.

2.Target Corpus Generation:

- Provide topics or descriptions closely related to usecase for generating examples.
- Better performance in Classification Tasks

3.Iterative Data Generation, Prompt Refinement:

- Generate few examples and evaluate quality initially.
- Refine the prompts based on the quality of initial generation.

CONCLUSION

- Experiment can be more refined with used of different Large Language Models and versions.
- This study highlights the potential of LLMs in generating synthetic data, it also underscores the complexities involved.
- Future research could explore advanced prompting techniques, such as Chain-of-Thought prompting and evaluate their impact on data quality.
- Studying structural and lexical biases in synthetic data could unlock even greater potential for this approach.
- Synthetic data generation is no silver bullet, but in the hands of skilled practitioners, it can be a powerful tool to overcome data scarcity, reduce costs and push the boundaries of what's possible in NLP.

THANK YOU

Soumya Bharathi Vetukuri (016668964)

01 Dec, 2024