Applying TOPSIS to Find the Best Pre-trained Text Generation Model

Aditya Vashishta (102103546)

Objective:

The goal is to compare and evaluate different pre-trained text generation models and identify the best one based on specific evaluation criteria.

Models used:

1. GPT-2
2. BART
3. T5-small
4. XLNet
5. GPT-Neo 125M

These are the pre-trained text generation models that are being compared and evaluated in the given scenario.

Evaluation Criteria:

1. Semantic Similarity:

- **Definition:** Semantic similarity measures how closely the meaning of two texts align. It is crucial for ensuring that the generated text captures the intended semantics of the prompt.

- **Evaluation Metric:** Cosine Similarity using TF-IDF vectors.

- **Aim:** Maximization (Higher similarity indicates better performance).

- **Weight:** 1 (Equal weight assigned to each criterion).

2. ROUGE Score:

- **Definition:** ROUGE (Recall-Oriented Understudy for Gisting Evaluation) evaluates the quality of summaries by comparing n-grams, word overlap, and other measures between the generated text and reference summaries.

- **Evaluation Metric:** ROUGE metrics (e.g., ROUGE-1, ROUGE-2, ROUGE-L).

- **Aim:** Maximization (Higher ROUGE scores indicate better quality).

- **Weight:** 1 (Equal weight assigned to each criterion).

3. Diversity:

- **Definition:** Diversity measures the variation and uniqueness of the generated text. It assesses the model's ability to produce diverse and novel outputs.

- **Evaluation Metric:** Metrics such as uniqueness, diversity of vocabulary, or the use of rare words.

- **Aim:** Maximization (Higher diversity indicates better performance).

- **Weight:** 1 (Equal weight assigned to each criterion).

TOPSIS Methodology:

TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) is a multi-criteria decision-making method that helps in ranking alternatives based on their proximity to the ideal solution and distance from the negative solution.

1. Normalization:

- Normalize the scores of each criterion for each model to ensure that all criteria are on the same scale.

2. Weighted Normalization:

- Multiply the normalized scores by their respective weights.

3. Ideal and Negative Ideal Solutions:

- Identify the ideal and negative ideal solutions for each criterion. The ideal solution represents the maximum values for each criterion, while the negative ideal solution represents the minimum values.

4. Euclidean Distances:

- Calculate the Euclidean distances between each model and the ideal and negative ideal solutions.

5. Similarity to Ideal Solution:

- Calculate the similarity to the ideal solution for each model using the formula: Similarity = Negative Ideal Distance / (Negative Ideal Distance + Ideal Distance).

6. Ranking:

- Rank the models based on their similarity to the ideal solution. A higher similarity score indicates a better ranking.

Prompts:

The prompts serve as input stimuli to the text generation models, guiding them to produce coherent and relevant text in response to each prompt's context.
Each set of prompts is tailored to a specific domain (political, sports, or science), allowing for targeted evaluation of the models' performance in generating content within those domains.
The diversity in the prompts within each category helps assess the models' ability to handle various subtopics and nuances within the broader domain.
These prompts can be used for testing and benchmarking the text generation models across different subject areas, providing insights into their proficiency and versatility.

List of Prompts -

1. **Political Prompts:**
    1.1 "Politics will be a game-changer for the world because"
    1.2 "Worst President of USA was"
    1.3 "As compared to monarchy, democracy is"
    1.4 "2024 India elections are"
    1.5 ”Congress Party served for over 20 years, but still”

2. **Sports Prompts:**
    2.1 "Virat Kohli is the best batsman because"
    2.2 "Football was the most"
    2.3 "Ronnie Coleman is the most awarded"
    2.4 "Max Verstappen is a very talented F1 driver which is evident from"
    2.5 "One of the lesser-known sports is"

3. **Science Prompts:**
    3.1 "An atom consists of"
    3.2 "Rocket thrusts upwards by"
    3.3 "Full form of VIBGYOR is"
    3.4 "When the apple fell on Newton's head, he discovered"
    3.5 "Black holes were considered to be a myth until”

Result:

A) Political prompts -

Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.3832372670142710	0.339622641509434	0.813953488372093	0.40865155124141400	4.0
BART	0.4955148952547200	0.4615384615384620	0.7	0.45269199667521200	3.0
T5-small	0.832624272096304	0.339622641509434	0.2727272727272730	0.5225058481096630	1.0
XLNet	0.5248470263769430	0.3333333333333330	0.8409090909090910	0.48788847797768700	2.0
GPT-Neo 125M	0.4950804386204100	0.3461538461538460	0.6046511627906980	0.29048499850303	5.0
Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.0780114138840906	0.0425531914893617	0.918918918918919	1.0	1.0
BART	0.0	0.0	0.875	0.2674291605135710	3.0
T5-small	0.0	0.0	0.1041666666666670	0.0	5.0
XLNet	0.0424035882220659	0.0416666666666666	0.7368421052631580	0.5128104627927850	2.0
GPT-Neo 125M	0.0	0.0	0.2619047619047620	0.02549758983222990	4.0
Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.0770975839482976	0.0769230769230769	0.7045454545454550	0.8231666160592380	2.0
BART	0.0060931941029141	0.0377358490566037	0.4042553191489360	0.13841700963291400	4.0
T5-small	0.0	0.0	0.3870967741935480	0.0	5.0
XLNet	0.0725223333603399	0.08	0.8461538461538460	0.9339435519297630	1.0
GPT-Neo 125M	0.0649992429935402	0.037037037037037	0.5434782608695650	0.41565095368858100	3.0
Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.0874050629038427	0.0363636363636363	0.8260869565217390	0.6516149938328320	2.0
BART	0.0	0.0	0.4090909090909090	0.11413962240788600	4.0
T5-small	0.0	0.0	0.1219512195121950	0.0	5.0
XLNet	0.0544212272416112	0.037037037037037	0.6888888888888890	0.471002811781028	3.0
GPT-Neo 125M	0.1281585538974640	0.0384615384615384	0.4186046511627910	0.6799417662069710	1.0
Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.1090151469792600	0.0833333333333333	0.9230769230769230	1.0	1.0
BART	0.0314265405583347	0.0714285714285714	0.85	0.44603132126569000	4.0
T5-small	0.0262821737810469	0.0384615384615384	0.2727272727272730	0.0	5.0
XLNet	0.0972945110804333	0.0754716981132075	0.8888888888888890	0.8164830724476470	2.0
GPT-Neo 125M	0.1042291340116050	0.08	0.6428571428571430	0.7036545140693840	3.0

Best Model: GPT-2

B) Sports prompts -

Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.1571492464355600	0.1176470588235290	0.7380952380952380	0.465338683790747	2.0
BART	0.0799568401944539	0.1666666666666670	0.8666666666666670	0.5631084475081510	1.0
T5-small	0.2486862051690650	0.0888888888888888	0.2222222222222220	0.4469326399227570	4.0
XLNet	0.0964069703600352	0.0784313725490196	0.8205128205128210	0.3903030719553530	5.0
GPT-Neo 125M	0.2312047744980590	0.1132075471698110	0.2954545454545450	0.45093691969247600	3.0
Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.1907681304606670	0.0769230769230769	0.8333333333333330	0.8578894436470800	1.0
BART	0.028342565491219	0.074074074074074	0.7777777777777780	0.44988127645127300	4.0
T5-small	0.0036910239314198	0.0338983050847457	0.08	0.0	5.0
XLNet	0.1544148273979980	0.0851063829787234	0.7368421052631580	0.6743302387934430	2.0
GPT-Neo 125M	0.2089822028221100	0.0784313725490196	0.4047619047619050	0.6357174393115430	3.0
Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.0881059979303853	0.0434782608695652	0.8285714285714290	0.4164028745565970	3.0
BART	0.0128134336066701	0.0408163265306122	0.5384615384615380	0.1646067508133590	4.0
T5-small	0.0358969160353761	0.0377358490566037	0.2045454545454550	0.03657609224243810	5.0
XLNet	0.1192872696658400	0.0350877192982456	0.7708333333333330	0.4563078369278300	2.0
GPT-Neo 125M	0.174182776348803	0.0714285714285714	0.5957446808510640	0.8004518192777370	1.0
Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.0527312791151431	0.0816326530612245	0.9	0.7991801725762770	1.0
BART	0.0	0.0	0.9230769230769230	0.23507813262750800	4.0
T5-small	0.0162209528300455	0.0454545454545454	0.4	0.2070805118506530	5.0
XLNet	0.0260183932727257	0.0425531914893617	0.8333333333333330	0.2947190524015710	3.0
GPT-Neo 125M	0.0643139715338522	0.04	0.6666666666666670	0.5564952697567900	2.0
Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.0829262882934382	0.037037037037037	0.7555555555555560	0.4355145009508120	2.0
BART	0.0137436504747494	0.0526315789473684	0.5517241379310350	0.24935168143525300	4.0
T5-small	0.0548725827621821	0.0338983050847457	0.18	0.12115916654626400	5.0
XLNet	0.0795900695513905	0.0377358490566037	0.7727272727272730	0.43449136924910700	3.0
GPT-Neo 125M	0.1221884541159970	0.0833333333333333	0.5135135135135140	0.7743012462403270	1.0

Best Model: GPT-2

C) Science prompts -

Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.0908766610226937	0.109090909090909	0.8478260869565220	1.0	1.0
BART	0.0	0.0	0.4210526315789470	0.08668205115588170	4.0
T5-small	0.0	0.0	0.1666666666666670	0.0	5.0
XLNet	0.0211260826086319	0.0816326530612245	0.675	0.3727369877527460	2.0
GPT-Neo 125M	0.0572459468964195	0.0357142857142857	0.4893617021276600	0.2952222035567540	3.0
Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.0	0.0	0.8947368421052630	0.19238102396676200	2.0
BART	0.0	0.0	0.8235294117647060	0.1675581329856670	4.0
T5-small	0.0	0.0	0.0909090909090909	0.0	5.0
XLNet	0.0	0.0	0.8620689655172410	0.180945323449607	3.0
GPT-Neo 125M	0.098337412387225	0.0384615384615384	0.3488372093023260	0.8305733066551610	1.0
Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.0930139716226402	0.1224489795918370	0.8	0.5950011681835640	1.0
BART	0.0	0.0	0.1914893617021280	0.012982258606952600	4.0
T5-small	0.0	0.0	0.1363636363636360	0.0	5.0
XLNet	0.0724038008081121	0.1176470588235290	0.8571428571428570	0.5513933433188500	3.0
GPT-Neo 125M	0.1601591468594460	0.074074074074074	0.2888888888888890	0.5667484692937020	2.0
Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.0476739022667647	0.0377358490566037	0.8636363636363640	0.3126486724708220	3.0
BART	0.0	0.0	1.0	0.27773757855693200	4.0
T5-small	0.0	0.0	0.1836734693877550	0.0	5.0
XLNet	0.1004449953675050	0.0377358490566037	0.7045454545454550	0.5406883412512090	2.0
GPT-Neo 125M	0.0569669373269969	0.0816326530612245	0.6	0.5965676476534660	1.0
Unnamed: 0	Semantic_Similarity	ROUGE	Diversity	Topsis Score	Rank
GPT-2	0.0923166058783102	0.074074074074074	0.8222222222222220	0.3926583532160910	3.0
BART	0.0070950686705188	0.0357142857142857	0.3829787234042550	0.05435955255536330	4.0
T5-small	0.0042241527122866	0.0338983050847457	0.2	0.0	5.0
XLNet	0.1014042554171430	0.1276595744680850	0.8918918918918920	0.6192050313057270	2.0
GPT-Neo 125M	0.1718546622777230	0.0754716981132075	0.7441860465116280	0.6436353443930230	1.0

Best Model: GPT-Neo 125M

Outcome

	Best Model
Political prompts	GPT-2
Sports prompts	GPT-2
Science prompts	GPT-Neo 125m

Conclusion:

In conclusion, **GPT-2** emerges as the most effective pre-trained text generation model, offering robust performance in semantic similarity, ROUGE scores, and diversity across various prompt categories. Its consistent excellence positions GPT-2 as the preferred choice for tasks requiring diverse and contextually accurate text generation.