THE DAUNTING DILEMMA WITH SENTENCE ENCODERS: GLOWING ON STANDARD BENCHMARKS, STRUGGLING WITH CAPTURING BASIC SEMANTIC PROPERTIES

By

Yash Mahajan, Naman Bansal, Shubhra Kanti Karmaker

Research Questions

Do these LLMs really understand the basic semantic in the given text?

How robust and reliable they are?

Evaluation on SentEval Benchmark

Model	MR	CR	SUBJ	MPQA	SSTb	TREC	MRPC	Avg
SBERT	83.95	88.98	93.77	89.51	90.01	84.80	**76.28**	86.90
USE	75.58	81.83	91.87	87.17	85.68	92.20	69.62	83.42
Infersent	81.10	86.30	92.40	90.2	84.60	88.20	76.20	85.57
LASER	56.14	63.89	67.65	72.36	79.85	89.19	75.19	72.04
Doc2Vec	49.76	63.76	49.16	68.77	49.92	19.20	66.49	52.43
Bloom	71.69	80.72	92.09	84.48	84.46	88.80	66.84	81.29
GPTNeo	79.91	83.36	93.48	84.62	88.19	92.40	70.78	84.68
LlaMa-2	83.34	87.15	**95.80**	87.46	91.65	94.00	65.97	86.48
GPT3	**88.36**	**93.08**	95.31	**91.29**	**93.63**	**96.00**	73.97	**90.23**

- MR : Movie Reviews (pos/neg)
- CR : Product Reviews
- SUBJ : Subjective Movie Reviews
- MPQA : Opinion Polarity
- SSTb : Stanford Sentiment Treebank
- TREC : Question-type classification
- MRPC: Paraphrasing dataset

Proposed Criteria

- Five basic semantic criteria*,
  - *Paraphrasing*
  - *Synonym Replacement*
  - *Paraphrase vs Sentence Jumbling*
  - ***Paraphrase vs Antonym Replacement***
  - ***Paraphrase without Negation***

* This list is not an exhaustive list.

Motivation for Negation based Criteria

- Dataset not having enough Negation sentences,
  - *For instance [1],*

Datasets	# of sentences	% of Negation sentence
QQP	1,590,482	8.1
STS-b	17,256	7.1
SST-2	70,042	16.0

[1] Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An Analysis of Negation in Natural Language Understanding Corpora. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 716–723, Dublin, Ireland. Association for Computational Linguistics

Dataset Curation

- **Paraphrasing**
  - *No change in QQP, MRPC and, PAWS dataset*

- For,
  - ***Synonym Replacement***
  - ***Antonym Replacement***
  - ***Sentence Jumbling***

Sentence S was used as the original sentence to generate perturbed S' sentences from QQP*, PAWS*, and MRPC*, forming (S1, S') pairs.

- **Paraphrasing without Negation**
  - *AFIN dataset*

Example of Curated data

Original Sentence: "*Levin's attorney, Bo Hitchcock, declined to comment last Friday*"		
Perturbation	**Example Sentence**	**Expected Encoding**
Paraphrasing	Hitchcock has declined to comment on the case, as has Levin.	Similar to Original
Synonym Replacement	Levin's attorney, Bo Hitchcock, *refused* to comment last Friday.	Similar to Original
Antonym Replacement	Levin's attorney, Bo Hitchcock, *accepted* to comment last Friday.	Diverse from Original
Paraphrase without Negation	Levin's attorney, Bo Hitchcock, remained silent when asked for comment last Friday.	Similar to Original
Sentence Jumbling	Levin's attorney *to* Bo Hitchcock, declined, comment last Friday.	Diverse from Original

Table 1: Example of the five sentence perturbation proposed to evaluate sentence encoders. **Note**: This example in "*Paraphrasing without Negation*" is for illustration purposes only and it hasn't been utilized in our study. It showcases the sentence structure we'd encounter in Afin dataset (Hossain and Blanco, 2022) (see Section 5.1).

Models Evaluated

- Classical Model
  - *USE*
  - *Sentence-Bert*
  - *LASER*
  - *InferSent*
  - *Doc2Vec*
- Emergent Models
  - *GPT3-ada-text embedding*
  - *LlaMa2*
  - *Bloom*
  - *GPTNeo*

Model Comparison

	Classical Models					Emergent Models			
	SBert	USE	LASER	InferSent	Doc2Vec	Bloom	LlaMa-2	GPT3.5	GPTNeo
Developed By	Sent-Transformer	Google	Facebook	FAIR	Google + Stanford	BigScience	Facebook	OpenAI	EleutherAI
Embedding Dimension	768	512	1024	4096	100-300	2048	4096	1536	2048
Parameter	~110M	~110 M	~93M	~24M	Variable	~560M, ~1B, ~7B, ~176B	~7B, 13B, 70B	~175B	~1.3B, ~2.8B
Size in GB	~0.4	~0.4	~0.4	~0.8	Variable	Variable (in 100s)	Variable (in 100s)	Variable	Variable (in 100s)
GPU Req.	✗	✗	✗	✗	✗	✓	✓	✗	✓
Open-source	✓	✓	✓	✓	✓	✓	✓*	✗	✓

Results

- Criterion 1: Paraphrasing
  - *Expectation : "A good sentence encoder should generate similar embeddings for two sentences which are paraphrases of each other"*

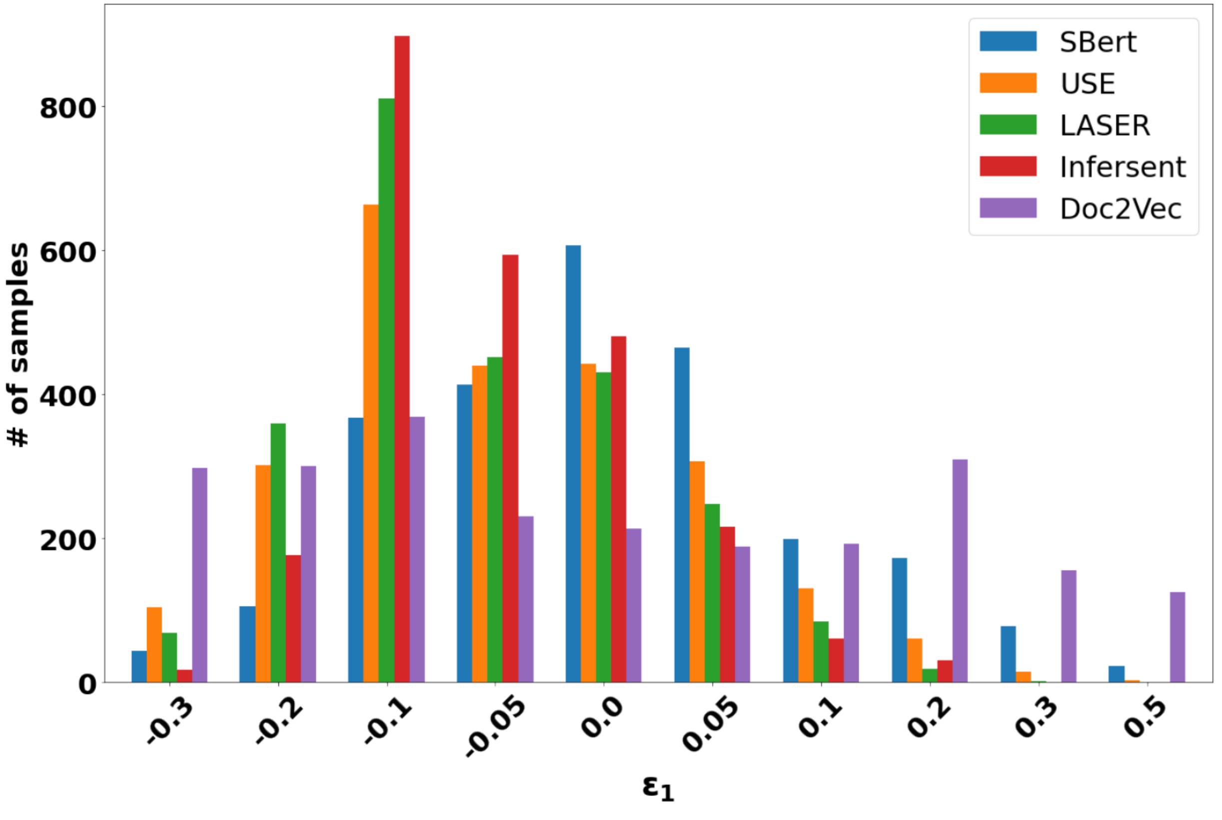
Model		USE	SBERT	Infer-Sent	LASER	D2V	Bloom	GPTNeo	GPT3-Ada	LlaMa-2
QQP	Pos	0.7553	0.8526	0.3182	0.3652	0.2516	0.0059	0.2669	0.2609	0.4277
	Neg	0.5278	0.5488	0.2849	0.3124	0.2368	0.0059	0.2512	0.2367	0.3734
	Diff	**0.2275**	**0.3038**	0.0333	0.0528	0.0148	0.0001	0.0157	0.0242	0.0543
WIKI	Pos	0.8645	0.9506	0.3552	0.4268	0.5180	0.0059	0.2767	0.2719	0.4646
	Neg	0.8554	0.9408	0.3552	0.4136	0.5402	0.0059	0.2750	0.2703	0.4568
	Diff	**0.0091**	**0.0098**	0.0000	0.0132	-0.0222	0.0000	0.0016	0.0016	0.0077
MRPC	Pos	0.7098	0.8134	0.3367	0.3828	0.4440	0.0059	0.2706	0.2634	0.4442
	Neg	0.6097	0.5488	0.3256	0.3564	0.3700	0.0059	0.2652	0.2549	0.4243
	Diff	**0.1001**	**0.2646**	0.0111	0.0264	0.0740	0.0001	0.0053	0.0085	0.0198

- Criterion 2: Synonym Replacement
  - *Expectation:* *"If we replace n words (where n is small) from sentence S with their respective synonyms to create another sentence S_P', a good sentence encoder will yield similar embeddings for S and S_P'.*

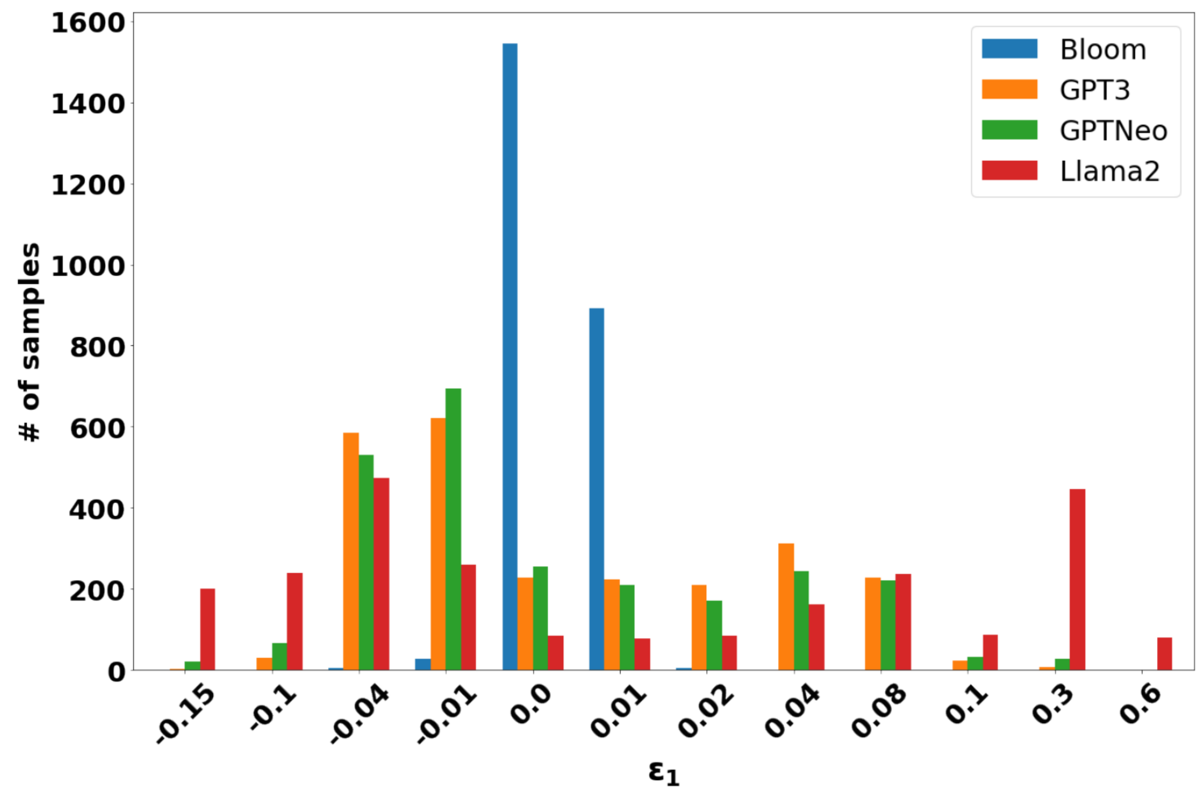
Models	QQP			WIKI.			MPRC		
	n=1	n=2	n=3	n=1	n=2	n=3	n=1	n=2	n=3
SBERT	0.898	0.831	0.775	0.945	0.909	0.874	0.929	0.879	0.829
USE	0.814	0.736	0.672	0.865	0.821	0.78	0.864	0.819	0.774
Infer-Sent	0.347	0.331	0.32	0.359	0.349	0.34	0.361	0.353	0.346
LASER	0.417	0.399	0.387	0.432	0.425	0.418	0.43	0.423	0.415
D2V	0.506	0.434	0.391	0.569	0.517	0.496	0.588	0.497	0.432
Bloom	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006
GPTNeo	0.273	0.266	0.259	0.277	0.272	0.267	0.278	0.274	0.269
GPT3 Ada	0.894	0.869	0.851	0.915	0.904	0.894	0.916	0.905	0.895
LlaMa-2	0.443	0.393	0.347	0.462	0.433	0.398	0.463	0.43	0.388

Table 3: Normalized Average Cosine Similarity between the Original and the Synonym Replaced Sentence pairs. Columns are grouped by dataset and subdivided by the number of word replacements, $n = \{1, 2, 3\}$. The **blue** and **purple** indicate the best and second-best performer.

- Criterion 3: Paraphrase vs Antonym Replacement:
  - ***Expectation:*** *Given a sentence S, its paraphrase S_P' and an antonym-replaced sentence S_A', created by replacing exactly one word (verb or adjective) with its antonym, S_P' should be semantically more similar to S than S_A' to S by some clear margin, i.e.,* ***$Sim(S, S_P') - Sim(S, S_A') > \epsilon 1$****, where $\epsilon 1$ denotes the expected minimum margin*



(a) Classical Model - Antonym Replacement on QQP



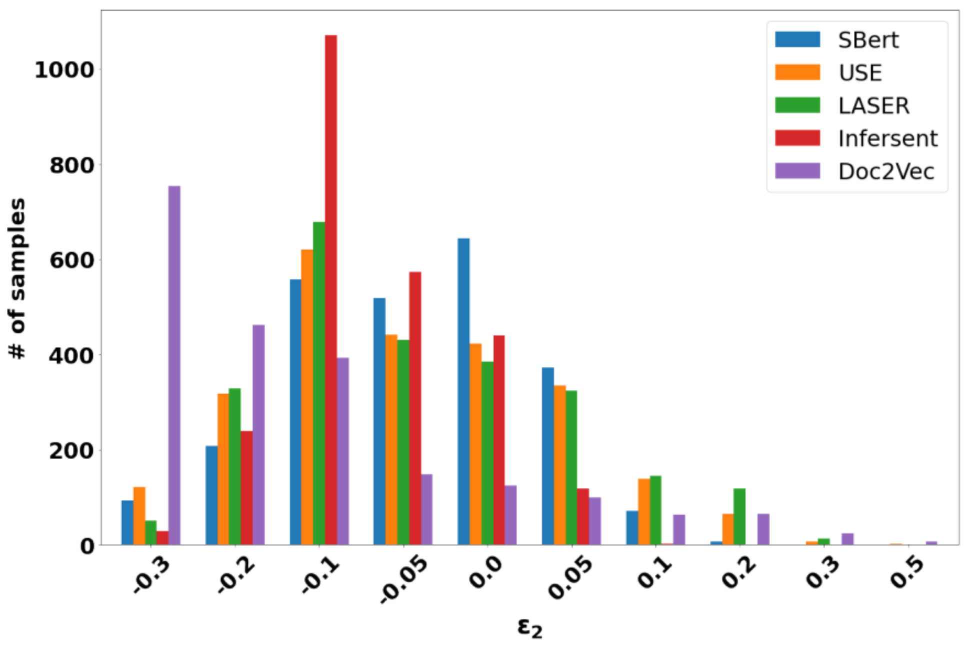
(b) Emergent Model - Antonym Replacement on QQP

- Criterion 4: Paraphrase without Negation
  - *Expectation:* *A "good" sentence encoder will recognize the semantic equivalence despite negation being present in S but not in S ′ , and thus produce high similarity scores*

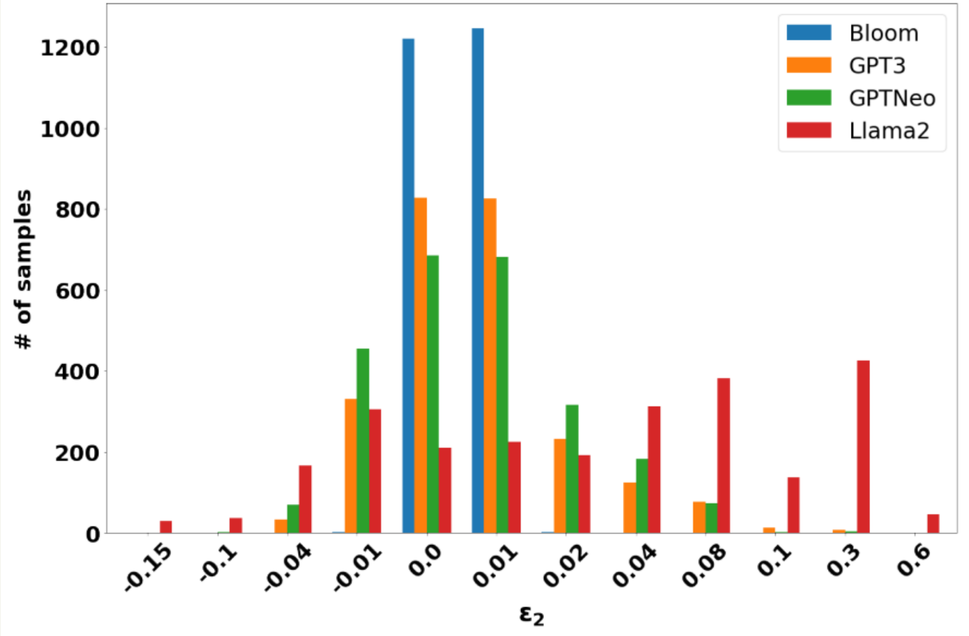
Model	USE	SBERT	Infer-sent	LASER	D2V	Bloom	GPTNeo	GPT3-Ada	LlaMa2
Avg. Sim. score	**0.695**	**0.779**	0.325	0.387	-0.001	0.006	0.267	0.260	0.423

Table 4: Criterion-4: Normalized Avg. similarity score of negation-affirmative sentence pair sentences from the AFIN dataset. The **blue** and **purple** indicate the best and second-best performer.

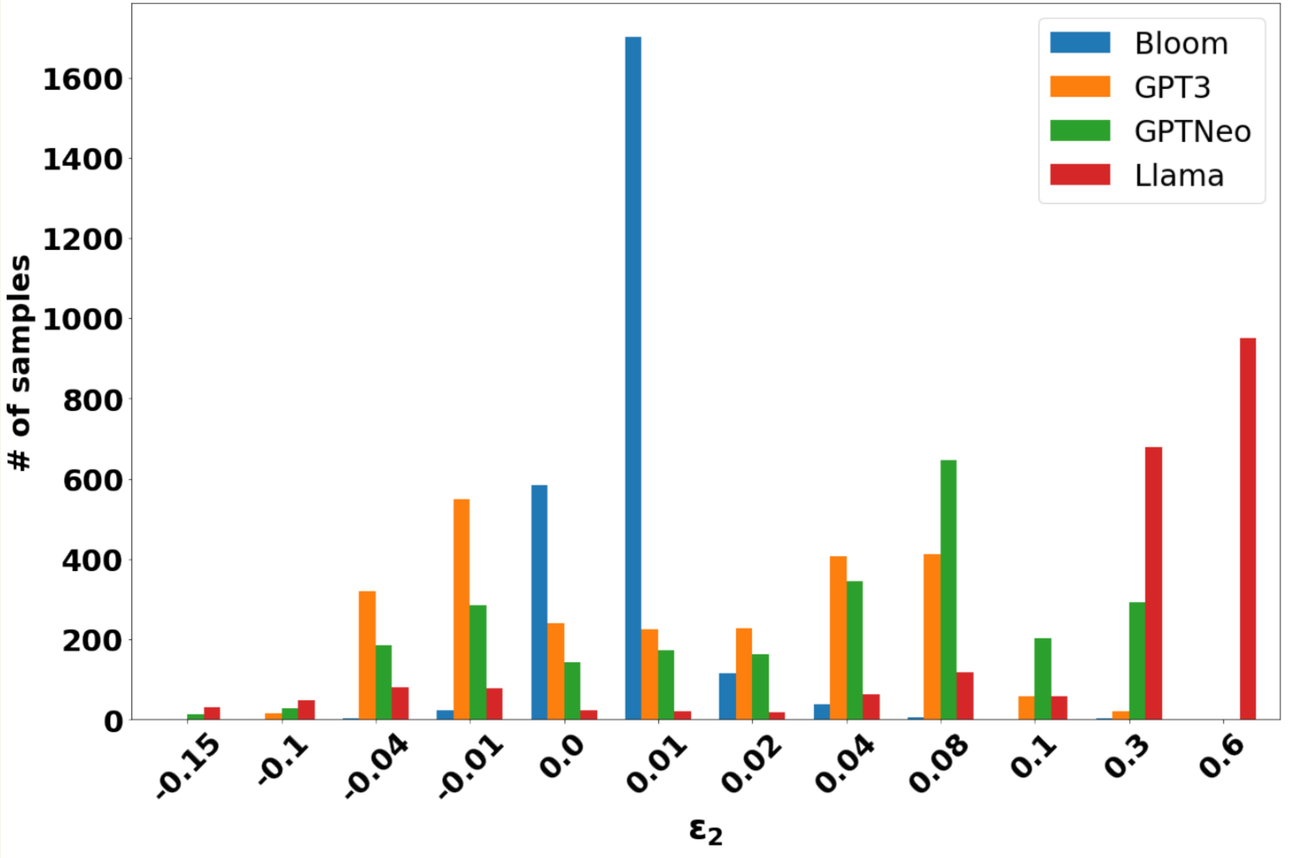
- **Criterion 5: Paraphrase vs Sentence Jumbling**
  - *Expectation: Given a sentence S, its paraphrase S_P' and a jumbled sentence S_J', S_P' should be semantically more similar to S compared to S_J' by some clear margin, i.e,* ***$Sim(S, S_P') - Sim(S, S_J') > \epsilon 2$****, where $\epsilon 2$ denotes the expected minimum margin*



(a) Classical Models- Sentence Jumb. on QQP for $n = 1$



(b) Emergent Models- Sentence Jumb. on QQP for $n = 1$



(c) Emergent Models- Sentence Jumb. on QQP for $n = 3$

Conclusion

1

Criteria demonstrated the struggle of LLMs on basic foundational language properties.

2

We need more robust benchmark datasets which also include granule semantic understanding, negation focused data.

3

Similarity metric like cosine similarity might be inadequate to capture granule semantic in high dimensional vector space.

Limitation

The study is limited to English language.

Evaluated under unsupervised semantic understanding.

Thank You!!!

Any Questions