Comparative Analysis: Evaluating the Performance of Language Models on various NLP Tasks across different domain mid-size datasets

- Abhinav Asthana, Arshia Joshi, Sushumna Ch., Sai Vardhan

Problem Statement

- Evaluating the performance of state-of-the-art language models on four NLP tasks and analyze the impact of model size, training data, and computational resources on task performance.

Literature Review and Research gap

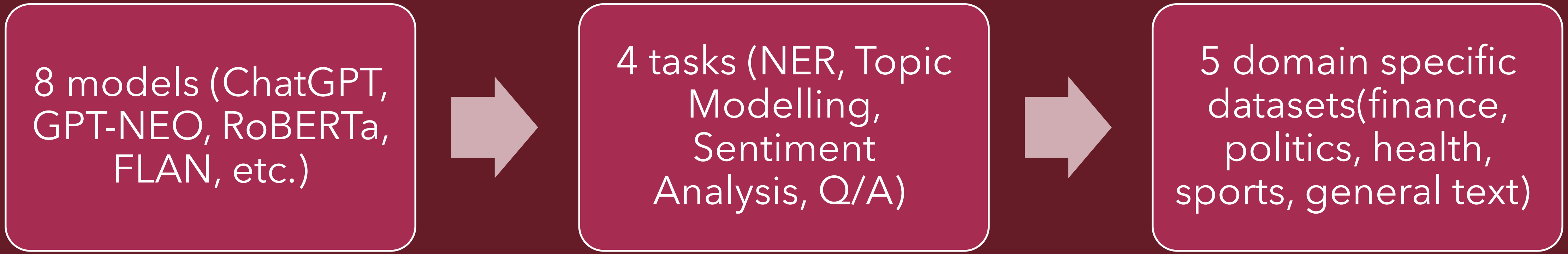
There are various studies which have evaluated performance of language models pertaining to a specific task and domain-specific dataset. Eg. BioBERT, FinBERT, BioNER etc.

Research gap: A One-stop solution that evaluates the performance of latest Language models for various NLP tasks such as Topic Modelling, Sentiment Analysis, NER, and Q/A for multi-domain small to mid-size datasets. This meta-analysis will serve as a reference point for researchers/domain experts to quickly build applications using the most suitable model

Scope

- The study includes researchers, academics, and practitioners in the field of natural language processing (NLP), machine learning, and data science working improving LLM across various tasks and domain.
- The impact we are trying to create:
  1. Provide insights into strength and weakness of each model
  2. Study Applicability of these models on diverse topics
  3. Study emerging models like Flan and Galactica offering glimpse into future of NLP.

Problem Statement : Models, Tasks, Dataset



8 models (ChatGPT, GPT-NEO, RoBERTa, FLAN, etc.)

4 tasks (NER, Topic Modelling, Sentiment Analysis, Q/A)

5 domain specific datasets(finance, politics, health, sports, general text)

Problem Statement: Tasks

- Topic Modelling
- Sentiment Analysis
- Question-Answering
- Named Entity Recognition

Problem Statement: Language Models

- CHATGPT API
- GPT3 API
- BLOOM
- OPT175B
- GPT NEO
- GLM130B
- ROBERTA
- FLAN
- GALECTICA 120B

Datasets Selected

THE GLUE DATASET, OR GENERAL LANGUAGE UNDERSTANDING EVALUATION

FINANCIAL PHRASEBANK (FPB) DATASET (SENTIMENT ANALYSIS)

20 NEWSGROUPS DATA SET (TOPIC MODELLING)

NCBI DISEASE DATASET *(NER)*

THE STANFORD QUESTION ANSWERING DATASET (SQUAD)

Dataset Description: General Language Understanding Evaluation dataset (GLUE)

- Nine different tasks including sentiment analysis, question answering, and natural language inference
- 33,000 examples across nine different tasks.
- The columns in the dataset are as follows:
  - "index": A unique identifier for each data instance.
  - "sentence1": The first sentence of the data instance.
  - "sentence2": The second sentence of the data instance (if applicable).
  - "label": The ground truth label or target value for the data instance.
  - "label_num": The label encoded as a numerical value.
  - "score": A continuous score indicating the similarity or relatedness of the two sentences (if applicable).

Dataset Description: Financial PhraseBank (FPB) dataset

- **Task – Sentiment Analysis**
- Collection of financial phrases that have been annotated with their sentiment polarity.
- 4,840 annotated phrases in total, with 2,200 from news articles and 2,640 from social media.
- Column names for both the training and test files:
  - "Sentence": The text of the financial news article or statement.
  - "Label": A binary label indicating whether a specific financial phrase or expression occurs in the text or not. The label is 1 if the phrase is present, and 0 if it is not.

Dataset Description: 20 Newsgroups data set (Topic Modelling)

- **Task - Topic Modelling**
- *Collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups.*
- *Feature description:*
  - *"target": A numerical label indicating the category or newsgroup of the document. The possible labels are integers ranging from 0 to 19, with each integer corresponding to one of the 20 newsgroups in the dataset.*
  - *"data": The text content of the document.*

Dataset Description: NCBI Disease dataset

Task- Named Entity Recognition

This dataset contains the disease name and concept annotations of the NCBI disease corpus, a collection of 793 PubMed abstracts.

Train (5433 instances), Validation (924 instances) and Test set (941 instances).

The training and testing TSV files have six columns:

- *"PMID": The unique identifier of the PubMed article.*
- *"Sent_ID": The unique identifier of the sentence in which the disease mention occurs.*
- *"Start_Offset": The starting character offset of the disease mention in the sentence.*
- *"End_Offset": The ending character offset of the disease mention in the sentence.*
- *"Disease Mention": The text of the disease mention.*
- *"Negation": A binary label indicating whether the disease mention is negated in the sentence. The label is 1 if the disease mention is negated, and 0 if it is not.*

Dataset Description: Stanford Question Answering Dataset (SQuAD)

- **Task – Question/Answering**
- Popular benchmark for machine comprehension tasks. It consists of over 100,000 question-answer pairs, with each question associated with a short passage of text from a set of Wikipedia articles.
- Three columns in the SQuAD dataset:
  1. "paragraphs": This column contains a list of paragraphs. Each paragraph is a dictionary with two keys:
  2. "qas": This column contains a list of questions and answers. Each question-answer pair is represented as a dictionary with several keys.
  3. "version": This column is a string indicating the version of the SQuAD dataset.

- Dataset Preprocessing and Exploration.
- Fine-tuning the models as per use case.
- Evaluating models on different tasks and dataset.

Goals to achieve

Thank You!

Feel Free to ask any Questions