Task 1: Topic Modeling: Use **LDA** and **BERTopic** to perform topic modeling on the dataset and determine the optimal number of topics by testing the number of topics ranging from 1 to 10.

1. Test different numbers of topics (K) ranging from 1 to 10 for both models.
2. Use the **coherence score** to evaluate the quality of the topics and find the best number of topics for each model.

   Compare the coherence scores from LDA and BERTopic to assess which model provides more interpretable topics at its optimal number of topics. Explain why model approach is performing better than the other.

   Generate word clouds for the best topic modeling approach with the optimal number of topics.

Task 2: Named Entity and Affective Analysis: Analyze how the two entities **Hamas** and **Israel** are portrayed in the text. To do so, we check the surrounding words of each entity.

1. **NER Extraction:**
   - Use an NER tool like **Stanford parser or spaCy** to retrieve sentences mentioning the above two entities.
2. **Affective Analysis:**
   - For each type of entity, extract the following tokens:

     If the entity is used as a subject, extract its direct verbs and direct modifiers (using POS tagging) of that entity. Calculate the average sentiment score of extracted words using **Valence** lexicons (https://saifmohammad.com/WebPages/nrc-vad.html). Compare and analyze these average sentiment scores for both entities. Discuss your findings on sentiment used to describe each entity in the text.

   - Check if the entity is described to be more powerful or showing **Dominance**, according to the text.

     Possible solution: For each entity, extract related verbs, nouns, or adjectives around that entity. You can use dependency or constituency parsing for extraction. Calculate **Dominance** scores for extracted words (https://saifmohammad.com/WebPages/nrc-vad.html). Feel free to refine the extraction pipeline if required (add or subtract types of surrounding words). Discuss your findings on **dominance** used to describe each entity in the text.

     Use word clouds to visualize related surrounding words for each entity. Compare these word clouds for two entities and discuss your findings.

Task 3: Predictive Model: Build models to predict the column "score" for posts based on their text. This column represents the number of upvotes minus downvotes for each Reddit post.:

- ○ Fine-tune one of the models: BERT, RoBERTa, or XLNet for the prediction task.

  You can use Simple Transformers library for fine tuning (https://simpletransformers.ai/)

  (https://huggingface.co/docs/transformers/en/index)

- ○ Evaluate performance using MSE (Mean Squared Error) on 10-fold cross validation.