Assignment 2: Reddit Data Analysis

Deadline: November 12th
Submission Format: Jupyter Notebook (.ipynb) along with Report (pdf)
Total Marks: 100

Overview:

In this assignment, you will analyze social media data from Reddit focusing on discussions related to the Israel-Hamas conflict. Your tasks include performing topic modeling, named entity recognition (NER), and developing predictive models. The goal is to explore how key entities are portrayed emotionally, extract meaningful themes from the dataset, and build models to predict scores from post content.

Dataset: linkDownload link

Task 1: Topic Modeling

Objective: Discover underlying themes within Reddit posts using topic modeling.

Steps to Complete:

1. **Preprocess the data (Optional)**:
    - Remove stopwords, special characters, and unnecessary whitespace.
    - Perform lemmatization to normalize words.
2. **Apply LDA (Latent Dirichlet Allocation) for Topic Modeling:**
    - Experiment with different numbers of topics (1–10).
    - Use the coherence score to evaluate the LDA model for each topic configuration.
    - Select the best number of topics based on the coherence score.
3. **Label the Topics:**
    - Manually interpret and assign meaningful labels to the topics.
    - Create **Word Clouds** to visualize the themes.

Deliverables:

- A **table or brief discussion** explaining the chosen number of topics and the themes they represent.
- **Visualizations** (e.g., word clouds) to summarize key topics.

Task 2: Named Entity Recognition (NER)

Objective: Identify how key entities are portrayed (positive or negative) in the dataset.

Entities to Analyze:

1. **Israel**
2. **Hamas**
3. **Antony Blinken**
4. **Benjamin Netanyahu**
5. **IDF (Israeli Defense Forces)**

Steps to Complete:

1. **NER Extraction:**
    - Use an NER tool like **Stanford parser or spaCy** to retrieve sentences mentioning the above entities. Entities such as **Benjamin Netanyahu** may sometimes be mentioned just by their last names, i.e **Netanyahu** in this case.
2. **Interpretation of Results:**
    - For each entity, calculate the average sentiment score of sentences mentioning it, using the VADER tool in python. Compare and analyze these sentiment scores across entities.
    - For each entity, extract verbs, nouns, and adjectives (using POS tagging in NLTK) coming in the same sentence as that of the entity. Use **Word Clouds** to visualize these surrounding words. Compare these word clouds and discuss your findings.

Task 3: Predicting Score for Each Reddit Post

Objective: Build models to predict the column "score" for posts based on their text and title. This column represents the number of upvotes minus downvotes for each Reddit post.

Steps to Complete:

1. **Feature Selection:**
    - Use **at least one** of the following features:
        - **Topics**
        - **NER**
        - **Lexicon-based scores (Valence, Arousal, Dominance)**
            1. **Valence** (pleasantness of the emotion, positive or negative)
            2. **Arousal** (intensity or energy of the emotion)

3. **Dominance** (degree of control or influence)

-
    -
        -
        -
        - **Lexicon**
          Reference: https://saifmohammad.com/WebPages/nrc-vad.htmlLinks to an external site.

1.
    -
        - You must use word or sentence embeddings (options: Word2Vec, GloVe, Sentence-BERT, and USE).
        - You are free to explore other text based features as POS tags (optional).

2. **Model Training:**
    - Train models using **SVM** and **XGBoost** to predict scores.
    - Evaluate models using **MSE (Mean Squared Error)**.
3. **Deliverables:**
    - Include model performance comparisons with evaluation metrics (e.g., MSE).
    - Discuss the effectiveness of different features in improving predictions.

Submission Guidelines

- **Upload a Jupyter Notebook (.ipynb) on Canvas**.
- Upload a report having all your findings along with the necessary steps you conducted during your experimentation.
- Ensure that **all code cells execute correctly** with visible outputs.
- Use **Markdown cells** to document your analysis and explain your findings.
- Provide **clear explanations** of your steps and reasoning in the notebook.