Report CommonShare Assignment

Advanced NLP (Natural Language Processing)

By: BENHIMA Mohemed-amine

I. Sentiment Analysis

1. Methodology

I fine-tuned DistillBERT on a subset of (5000 rows) the Twitter Sentiment Analysis dataset in kaggle

Steps:

- Loading the Data (3 labels: negative, neutral, positive)
- Cleaning the Data for text classification task
  - Remove duplicates
  - Remove nulls
  - Remove unused columns
  - Remove URLs
  - Remove html tags
  - Handle spaces
  - Convert to lowercase.
- Tokenization
  - Using padding and truncation
  - Fix the sentence to the default max_length = 512
- Define the evaluation metrics to use
  - Accuracy
  - Recall ⇒ macro (average of all classes recalls)
  - Precision ⇒ macro
  - F1 ⇒ macro
- Define my Training Arguments object
  - L2 regularization = 0.01
  - Train and Validation batch size = 16
  - Learning rate scheduler is Linear for faster convergence
  - Warmup is 20% of total steps (batches)
  - Load at the end the best model that has the best F1 score
  - Use half-precision (fp16) for increasing speed training and memory efficiency
  - Others hyperparameters were left as default
- Training
  - For 5 epochs (compute constraints)

2. challenges faced

- Compute constraints (I already used my colab quota and close to use my kaggle quota also)
- Struggling to find a real world dataset

3. Results

- Accuracy, precision, recall, and F1 are 78%
- For 5 epochs training on 5000 dataset, it’s a good results

- Using a complex model like bert base can increase the results
- Cleaning the dataset more can help increase the results
- Training for more epochs can increase the results

4. Technologies

- pandas
- HuggingFace Ecosystem
    - Transformers
    - Datasets
    - Evaluate
    - HuggingFace Hub
- ClearML
- Kaggle

II. Topic Modeling

1. Methodology

I used the Latent Dirichlet allocation (LDA) for topic modeling
Steps

- Pre-process the dataset for topic modeling
    - Remove URLs
    - Remove mentions
    - Remove hashtags
    - Remove punctuation
    - Lowercase
    - Lemmatization
- Convert the data to a Bag of Words corpus
- Define the different hyper-paramters ranges of values
- Apply grid search to find the best hyper-parameters
- Use Coherence to select the best combination
- Train the LDA model with the best hyper-parameters
- Display some topics
- Then visualize using pyLDAvis

2. challenges faced

Since i am using a twitter dataset, that is messy, I pre-processed it carefully

3. Results

The best coherence I got is 40%, which is good for a messy Twitter dataset.

4. Technologies

- Nltk
- Spacy
- Gensim
- pyLDAvis

III. Named Entity Recognition (NER)

1. Methodology

I used Spacy pre-trained NER on the Twitter Sentiment Analysis dataset

2. challenges faced

No challenges

3. Results

Without training, the results are good. But we can achieve better results using fine-tuned NER transformer model like BERT. Overall without any training the results are good

4. Technologies

- Spacy

IV. Text Summarization

1. Methodology

We utilized a BERT-based model fine-tuned for extractive summarization.
Each sentence is represented by the embedding of its first token (CLS token).
Added a simple linear classifier on top of BERT to predict sentence importance scores.
No training or fine-tuning done; model is used as-is for extractive summarization.

2. Challenges Faced

Loading pre-trained weights properly without fine-tuning. Because HuggingFace don’t support a pipeline for extractive summarization

3. Results

Without fine-tuning we got very good results

4. Technologies

- HuggingFace BERT
- Nltk