Sentiment Analysis

Alisa Khuu & Richard Tran

Overview

01 **Background**

What is sentiment analysis, NLP, sentiment140 dataset

02 **Methods**

NLTK Vader & BERT Model

03 **Results & Analysis**

Outcomes and analysis of performance

04 **Conclusion & Future Work**

RoBERTa & other models, tokenization/encoding strategies, stop-word removal

Background 01

Sentiment Analysis

- **Sentiment Analysis**
  - NLP technique
  - Focused on understanding polarity of data
    - how positive or negative (or neutral)
  - Often used in business to gauge customer interest / satisfaction
- **Natural Language Processing (NLP)**
  - The ability of a program to understand a human language (natural language) in a written or spoken context.
  - Uses Artificial Intelligence

Dataset

- Sentiment140
  - 1.6 mil tweets annotated (0 = negative, 4 = positive)
  - Fields
    - target
    - ids
    - date
    - user
    - text
  - 50/50 split of negative & positive sentiment

Methods 02

NLTK Sentiment Vader

- **Valence Aware Dictionary and sEntiment Reasoner**
- **Lexicon and rule-based model**
  - Key words are given a value
  - VADER provides estimates based on key words that appear
- **Downside:**
  - Difficult to upkeep
  - If no keywords appear, VADER will evaluate data as neutral
- Process:
  - Vader is pretrained
  - Using sentiment140 dataset, used Vader to calculate Polarity Scores

Bidirectional Encoder Representations from Transformers (BERT)

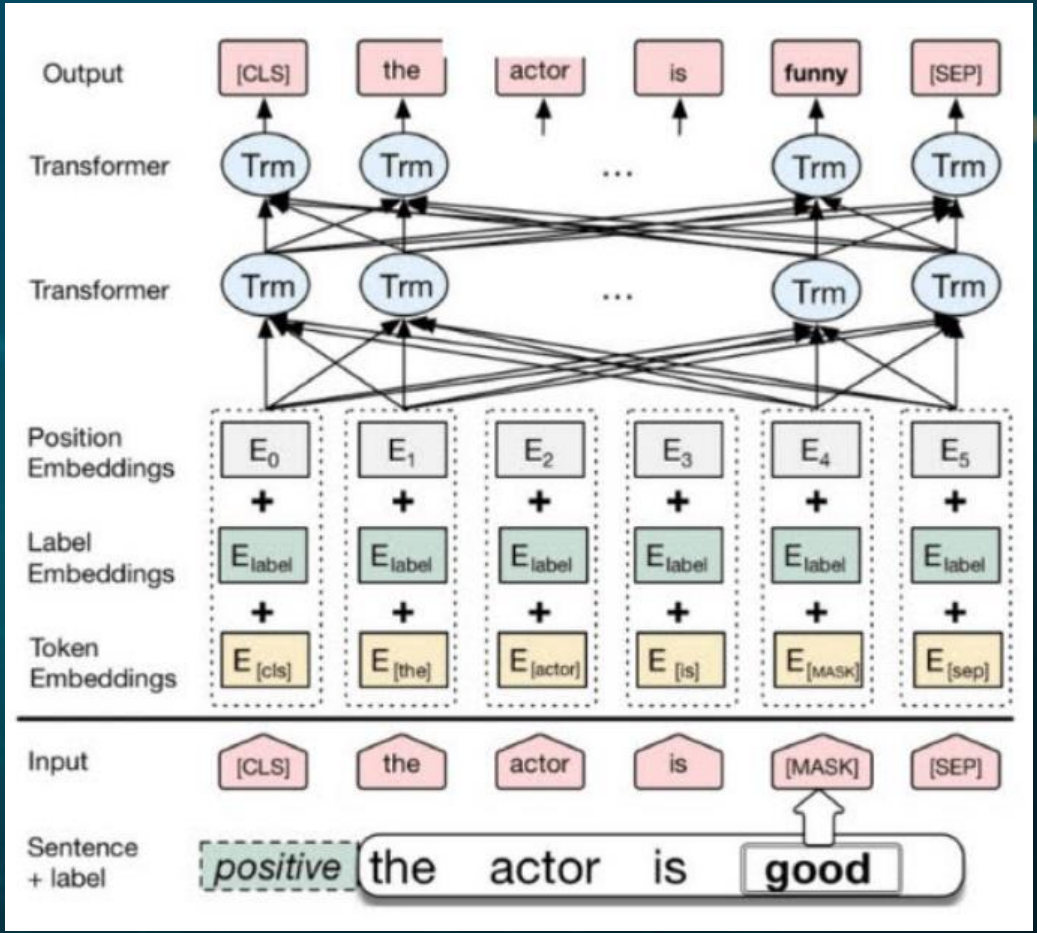
- **Transformer:**
  - every output element connected to every input element, weightings btwn. them dynamically calculated during training
  - self-attention mechanism to **figure out how important all the other "words" in the sentence are** w.r.t. to the aforementioned "word"
    - e.g. "I am hungry."
- **BERT:**
  - Pre-trained with text from Wikipedia (Masked Language Modeling & Next Sentence Prediction
  - Can "read both directions" at once

BERT Tokenization

```
>>> from transformers import BertTokenizer

>>> tokenizer = BertTokenizer.from_pretrained("bert-base-uncased")
>>> tokenizer.tokenize("I have a new GPU!")
["i", "have", "a", "new", "gp", "##u", "!"]
```

BERT Architecture



Results & Analysis

03

NLTK Sentiment Vader vs. BERT

- **Precision:**
  - Measures exactness
  - VADER:
    - Positive: 0.64
    - Negative: 0.78
  - BERT:
    - Positive: 0.89
    - Negative: 0.89

- **Recall:**
  - Measures completeness
  - VADER:
    - Positive: 0.89
    - Negative: 0.44
  - BERT:
    - Positive: 0.89
    - Negative: 0.89

NLTK Sentiment Vader vs. BERT

- **F1 Score:**
  - Harmonic mean of Precision & Recall
  - VADER:
    - Positive: 0.75
    - Negative: 0.56
  - BERT:
    - Positive: 0.89
    - Negative: 0.89

- **Accuracy:**
  - VADER: 0.68
  - BERT: 0.89

Analysis

- **Evaluation of Performance:**
  - recall: BERT
  - precision: BERT
  - f1-score: BERT
  - accuracy: BERT
- Due to its bidirectional transformer encoder architecture and its overall intricate deep learning design (better suited for bigger datasets, i.e. sentiment140), theoretically speaking it should perform better

Conclusion & Future Work

04

Conclusion

- **Conclusion:**
  - Bigger datasets should generally employ/utilize deep learning models
  - Advancement of NLP time & effort -> BERT, RoBERTa, DistilBERT, XLNet
- **Future Work:**
  - Limited due to lack of GPU
  - Tokenizers & Models (RoBERTa, other HuggingFace models)
  - Stop Word removal (“the”, “by”, etc.)
  - Different Hyperparameters (learning rates, epochs)
  - Different Optimization & Error/Loss Functions (AdamW & Cross Entropy Loss)

References

- https://www.nltk.org/howto/sentiment.html
- https://monkeylearn.com/sentiment-analysis/
- https://www.analyticsvidhya.com/blog/2021/06/vader-for-sentiment-analysis/
- https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP
- https://arxiv.org/abs/1810.04805
- https://www.securekloud.com/blog/benchmarking-sentiment-analysis-systems/
- https://huggingface.co/docs/transformers/tokenizer_summary
- https://i.ytimg.com/vi/gUJFq9hNFrw/maxresdefault.jpg
- https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model#:~:text=BERT%20is%20an%20open%20source,surrounding%20text%20to%20establish%20context.

Q/A