December 10th, 2020

Amazon Reviews Ratings

Abdelalim Yakoub
Ashit Patel

Overview

- Introduction
- Dataset & Data Preprocessing
- Base Model & Results
- BERT Model & Results
- Conclusion

Introduction

- Build a software that predicts customer rating based on customer's review
  - 5-Star rating system
- Multiclass classification problem
- Useful application
  - Provide numerical values to people's feedbacks

Dataset

- Amazon product data
  - Clothing, Shoes, and Jewelry
    - ~300,000 reviews
    - Training: 243,000
    - Validation: 27,000
    - Testing: 30,000


```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the piano.
He is having a wonderful time playing these old hymns.  The music  is
at times hard to read because we think the book was published for
singing from more than playing from.  Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```

Dataset Preprocessing

- Data Tokenization
  - BERT Tokenizer
    - Tokenizes words in each review text
    - Adds start and end tokens
    - Sets maximum length to 512
- Labels Format
  - 5-Star rating to 1-Hot vector
    - 4-Stars === [0, 0, 0, 1, 0]

Base Model

- Initial Model
  - Logistic Regression Model
    - Only used training set
    - Did not use validation set
  - Training Set
    - Accuracy: 69.39%
  - Testing Set
    - Accuracy: 69.37%

	Precision	Recall	F1-Score
1	0.53	0.49	0.51
2	0.36	0.15	0.21
3	0.41	0.28	0.33
4	0.50	0.26	0.34
5	0.76	0.95	0.84

BERT Model

- BERT Model
  - Used BERT Tokenizer
  - Truncated to 512 tokens
  - Set attention masks
  - Batch Size
    - 16
  - 12-Layers BERT model'
    - 'bert-base-uncased'
    - 5 Labels
  - Adam Optimizer
  - 2 Epochs

Avg. Training Loss	0.57
Avg. Training Accuracy	78%

Validation Accuracy	75%

Testing Accuracy	74%

Conclusion

- Comparison
  - Training accuracy +10%
  - Testing accuracy +6%
- Difficulties
  - Long training periods
  - Memory issues
- Future Improvements
  - Different BERT model
  - Use a larger dataset

Thank You!