CIS 530 Computational Lingusitics: Sentiment Analysis on Movie Reviews

Xiling Zou
Xinyi Pang
Hui Sui
Jingyi Lu

Penn Engineering

Introduction

- Online reviews for products and comments on social media -> Sentiment Analysis
- Application/Social Impact:

  - improve recommendation system of social apps

  - quickly gather attitudes towards products
- Goal: text information -> sentiment/attitudes

  - eg: predict sentiment score from reviews

Introduction

- Models:

  - simple baseline: logistic regression

  - strong baseline: vanilla Bert

  - extension1: fine tuning Bert

  - extension 2: BiLSTM-CNN

Data

- Available Data: Amazon Product Data, Tweets, IMDB movie reviews, ect.

- Our choice: Movie Reviews from Rotten Tomatoes
  - sentiment score from 0 - 4

Evaluation Metric

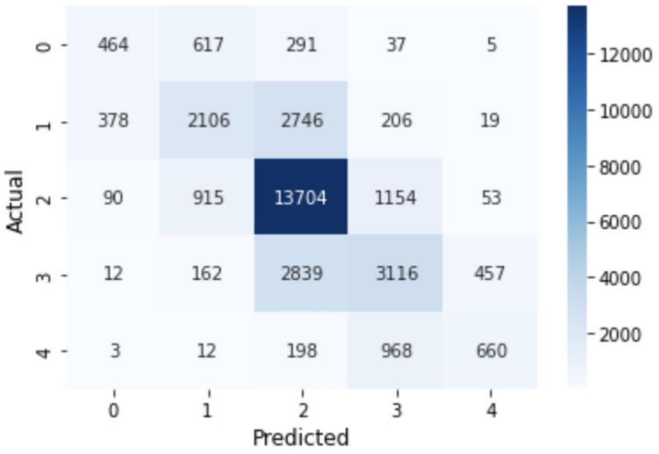
- F1 score:

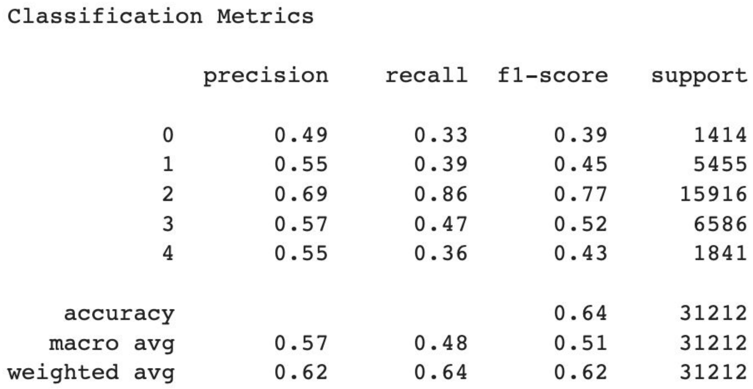
$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{tp}}{2\mathrm{tp} + \mathrm{fp} + \mathrm{fn}}$$

- Macro F1 score: sum(F1 scores)/ # classes

Simple Baseline

- Bag of Words + Logistic Regression
- Results:

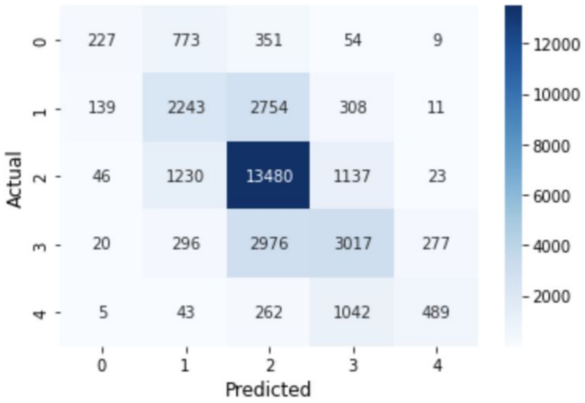


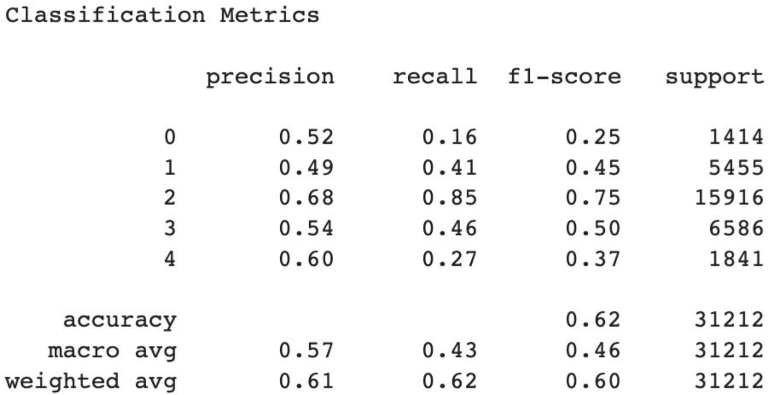


```
Classification Metrics

              precision    recall  f1-score   support

           0       0.49      0.33      0.39      1414
           1       0.55      0.39      0.45      5455
           2       0.69      0.86      0.77     15916
           3       0.57      0.47      0.52      6586
           4       0.55      0.36      0.43      1841

    accuracy                           0.64     31212
   macro avg       0.57      0.48      0.51     31212
weighted avg       0.62      0.64      0.62     31212
```

Strong Baseline

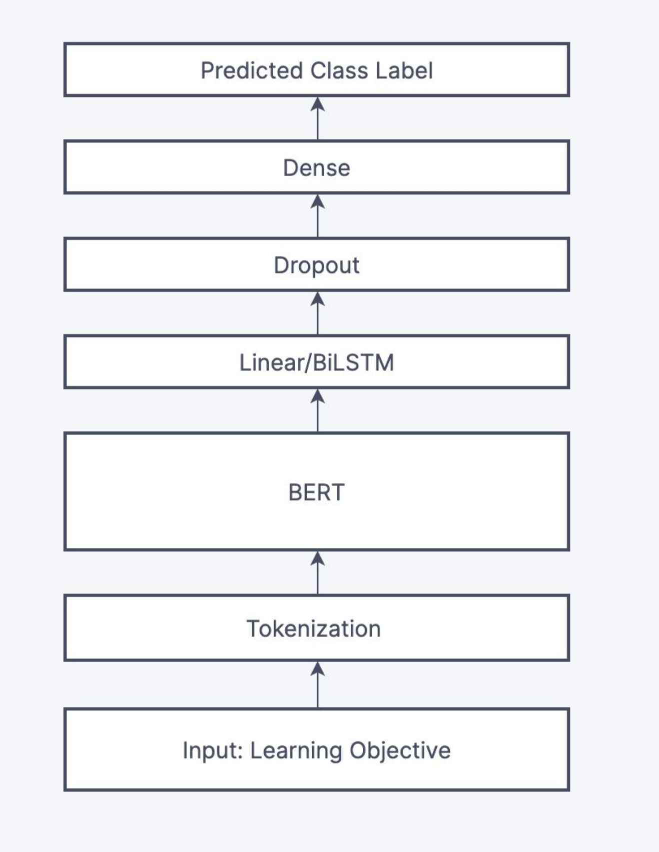
- BERT: context information
- BERT encoder + Logistic Regression
- Results:





```
Classification Metrics

              precision    recall  f1-score   support

           0       0.52      0.16      0.25      1414
           1       0.49      0.41      0.45      5455
           2       0.68      0.85      0.75     15916
           3       0.54      0.46      0.50      6586
           4       0.60      0.27      0.37      1841

    accuracy                           0.62     31212
   macro avg       0.57      0.43      0.46     31212
weighted avg       0.61      0.62      0.60     31212
```

Model Extension 1: Fine-Tuning BERT

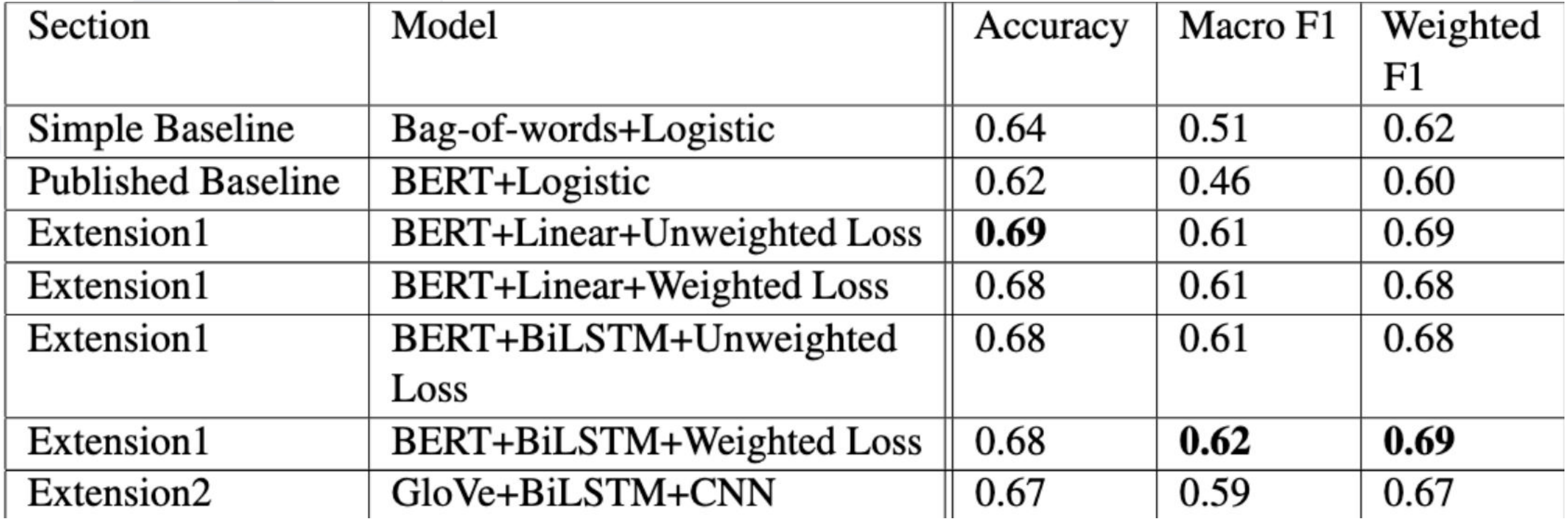


Model Extension 1: Training Details

- Epoch: 31 epochs
  1st epoch: train on weights of Bert and Fine-tuning
  Remaining epochs: only train on weight of Fine-tuning

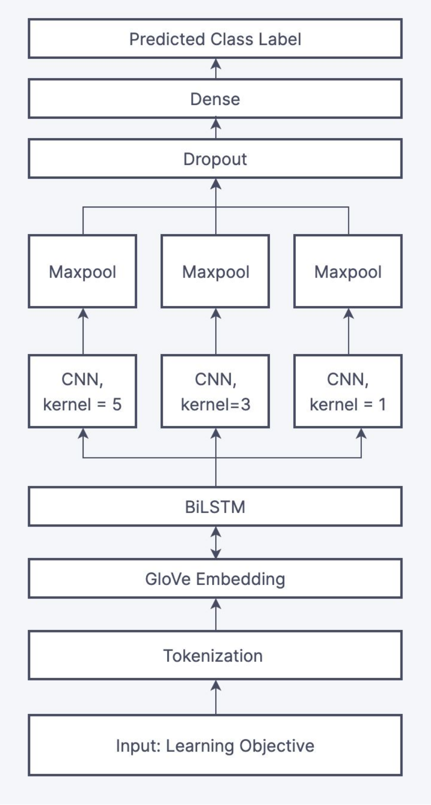
- 2 Loss functions:
  - Cross Entropy Loss
  - Weighted Cross Entropy Loss

Model Extension 1: Evaluation

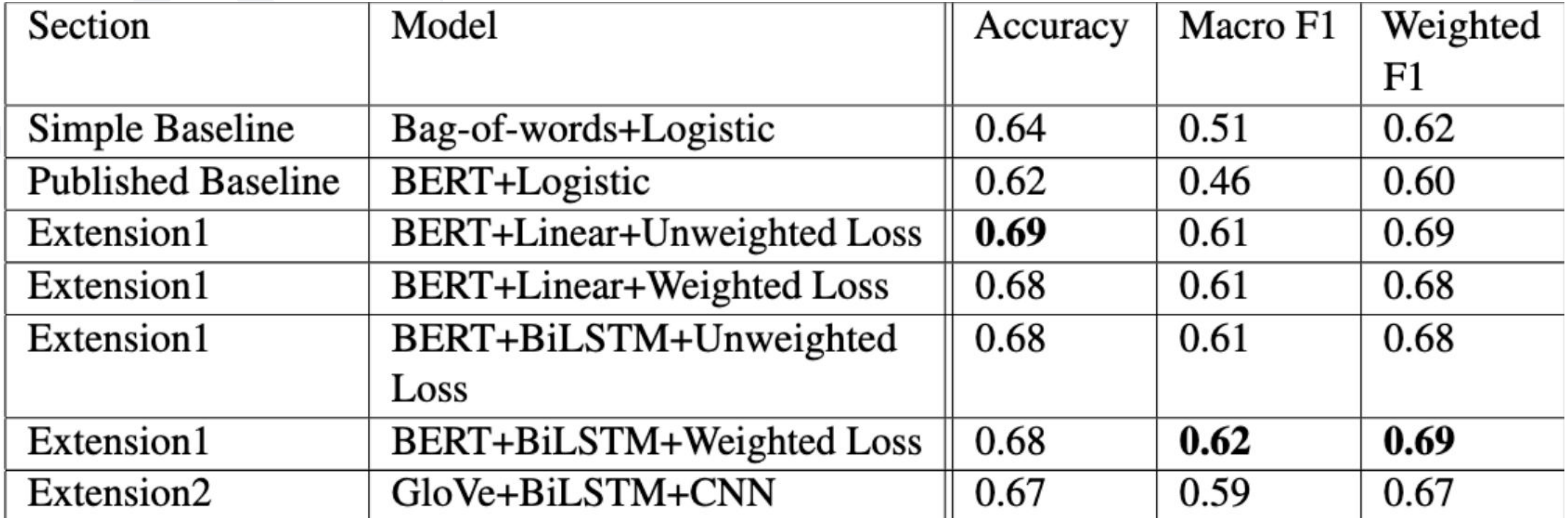


Section	Model	Accuracy	Macro F1	Weighted F1
Simple Baseline	Bag-of-words+Logistic	0.64	0.51	0.62
Published Baseline	BERT+Logistic	0.62	0.46	0.60
Extension1	BERT+Linear+Unweighted Loss	**0.69**	0.61	0.69
Extension1	BERT+Linear+Weighted Loss	0.68	0.61	0.68
Extension1	BERT+BiLSTM+Unweighted Loss	0.68	0.61	0.68
Extension1	BERT+BiLSTM+Weighted Loss	0.68	**0.62**	**0.69**
Extension2	GloVe+BiLSTM+CNN	0.67	0.59	0.67

Model Extension 2: BiLSTM-CNN



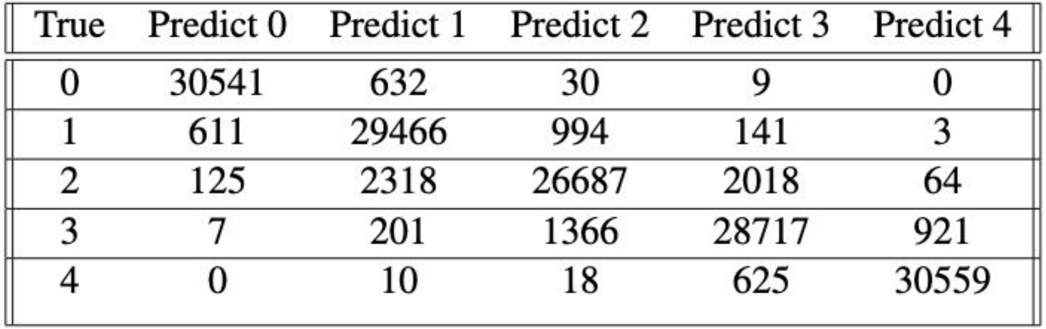
Model Extension 2: BiLSTM-CNN Detail



Section	Model	Accuracy	Macro F1	Weighted F1
Simple Baseline	Bag-of-words+Logistic	0.64	0.51	0.62
Published Baseline	BERT+Logistic	0.62	0.46	0.60
Extension1	BERT+Linear+Unweighted Loss	**0.69**	0.61	0.69
Extension1	BERT+Linear+Weighted Loss	0.68	0.61	0.68
Extension1	BERT+BiLSTM+Unweighted Loss	0.68	0.61	0.68
Extension1	BERT+BiLSTM+Weighted Loss	0.68	**0.62**	**0.69**
Extension2	GloVe+BiLSTM+CNN	0.67	0.59	0.67

Error Analysis

- Does well in predicting extremely negative (class 0) and extremely positive (class 4)
- As feeling becomes neutral, gets more incorrect predictions - class 2 worst
- Mostly incorrectly predicted as the nearby class(es)



True	Predict 0	Predict 1	Predict 2	Predict 3	Predict 4
0	30541	632	30	9	0
1	611	29466	994	141	3
2	125	2318	26687	2018	64
3	7	201	1366	28717	921
4	0	10	18	625	30559

Conclusion

- Simple baseline: bag of words + logistic regression
- Strong baseline: BERT + logistic regression
- Extension 1: BERT + neural network (MLP & BiLSTM) + weighted/unweighted loss function
- Extension 2: GloVe.6B.300d + BiLSTM-CNN
- Best performance achieved by BERT + BiLSTM + weighted cross entropy loss - **accuracy 0.68, F1-score 0.69**

Thank You

- Thank you for listening!