Berkeley SCHOOL OF INFORMATION

Comparative Analysis of BERT & ModernBERT in Binary Lexical Complexity Prediction

Jonathan Hernandez

Introduction & Problem Statement

- Lexical Complexity Prediction (LCP) uses language models to identify and simplify complex words in multi-word expressions (MWEs).

- **Research Questions**:
  - Does ModernBERT outperform BERT on classifying spans as complex vs. not-complex, across multiple domains?
  - Do data enrichment strategies outperform raw data?

- **Objective**:
  - Compare the performance and error patterns of BERT and ModernBERT on a newly binarized LCP dataset

Data & Motivation

- SemEval-2021 Task 1 LCP **Dataset:**
  - **10,800** multi-word, annotated, **spans**
  - Likert-averaged **continuous outcome variable**
  - **2 Tasks**: complexity based on unigrams and bigrams
  - **3 Domains**: EU parliament, biomedical, and bible

Single-Task Set			
Quartiles	**Train**	**Validation**	**Test**
Q1 - Q2: Less Complex	3,865	229	476
Median Value:	0.28	0.27	0.27
Q3 - Q4: More Complex	3,797	192	441
Total	7,662	421	917

Multi-Task Set			
Quartiles	**Train**	**Validation**	**Test**
Q1 - Q2: Less Complex	759	51	99
Median Value:	*0.41*	*0.42*	*0.43*
Q3 - Q4: More Complex	758	48	85
Total	1,517	99	184

Data & Motivation

- **Data Engineering**:
  - **Binarize continuous outcome variable** on train medians
  - Feature Engineer **13 new spans** with spaCy and Contractions
  - Re-Balance Dataset

- 2 Data Splits x 2 Y Variables x 13 X Variables x 5 Models = 260 Combinations

Single-Task Set			
Quartiles	**Train**	**Validation**	**Test**
Q1 - Q2	Not Complex		
Q3 - Q4	Complex		
Total:	7,000	1,000	1,000

Multi-Task Set			
Quartiles	**Train**	**Validation**	**Test**
Q1 - Q2	Not Complex		
Q3 - Q4	Complex		
Total:	1,300	250	250

Feature Engineering

- **Example**: "Don't underestimate us, it's more than complicated!"
    - **Eliminate Contractions** → don't → do not
    - **Part-of-Speech Tags** → AUX, PART, VERB, PRON, PUNCT…
    - **Dependency Mapping** → aux, neg, ROOT, dobj, punct, nsubj…
    - **Morphological Complexity** → VerbForm=Fin|Tense=Pres, Polarity=Neg, VerbForm=Inf, Case=Acc|Number=Plur|Person=1

- **Feature Types**:
    - **Solo Feature** → "[CLS] Do not underestimate us, it is more… [SEP]"
    - **Concatenation** → "[CLS] [input sequence] + [feature sequence] [SEP]"
    - **Interleaved** → "[CLS] [word 1] + [word 1 feature] + [word 2] + [word 2 feature] + [word 3] + … [SEP]"

Methods & Models

- **Methods**
  - Predict **complex vs. not-complex**
  - Perform **Naive Bayes Baseline**
  - **Tune hyperparameters** with BERT
  - Train & **Compare BERT vs. ModernBERT** (...and RoBERTA, DeBERTA, XLNet)
  - Perform **ablation study**
  - Evaluate with **F1, Precision, Recall**

Experiments

- 287 total experiments
- **46 baseline results**
- 54 hyperparameter tuning experiments
- **84 BERT vs. ModernBERT** (base and large) comparisons
- 103 reference experiments with RoBERTa, DeBERTa, XLNet

Epochs	Learning Rate	Batch Size	L2 Regularization	Conext Length	Warm-up Steps %	Unfrozen Parameter %
1	5e-6 / 0.00005	128	0.5	Default	10 - 100%	~7%

Berkeley SCHOOL OF INFORMATION

Key Results

Model Average Performance (All Tasks & Experiments)			
Model	**Precision**	**Recall**	**F1**
Naive Bayes	*0.58956*	*0.58347*	*0.5608*
BERT Base	0.45938	0.53489	0.42765
BERT Large	**0.50057**	0.35012	0.35256
ModernBERT Base	0.49728	**0.94757**	**0.64800**
ModernBERT Large	0.47205	0.57760	0.50285

Per Model Average Performance *Relative to Baseline*			
Model	**Precision**	**Recall**	**F1**
BERT Base	-20.93%	-7.25%	-25.22%
BERT Large	**-14.1%**	-39.81%	-38.31%
ModernBERT Base	-14.1%	**65.04%**	**13.9%**
ModernBERT Large	-17.12%	2.11%	-10.17%

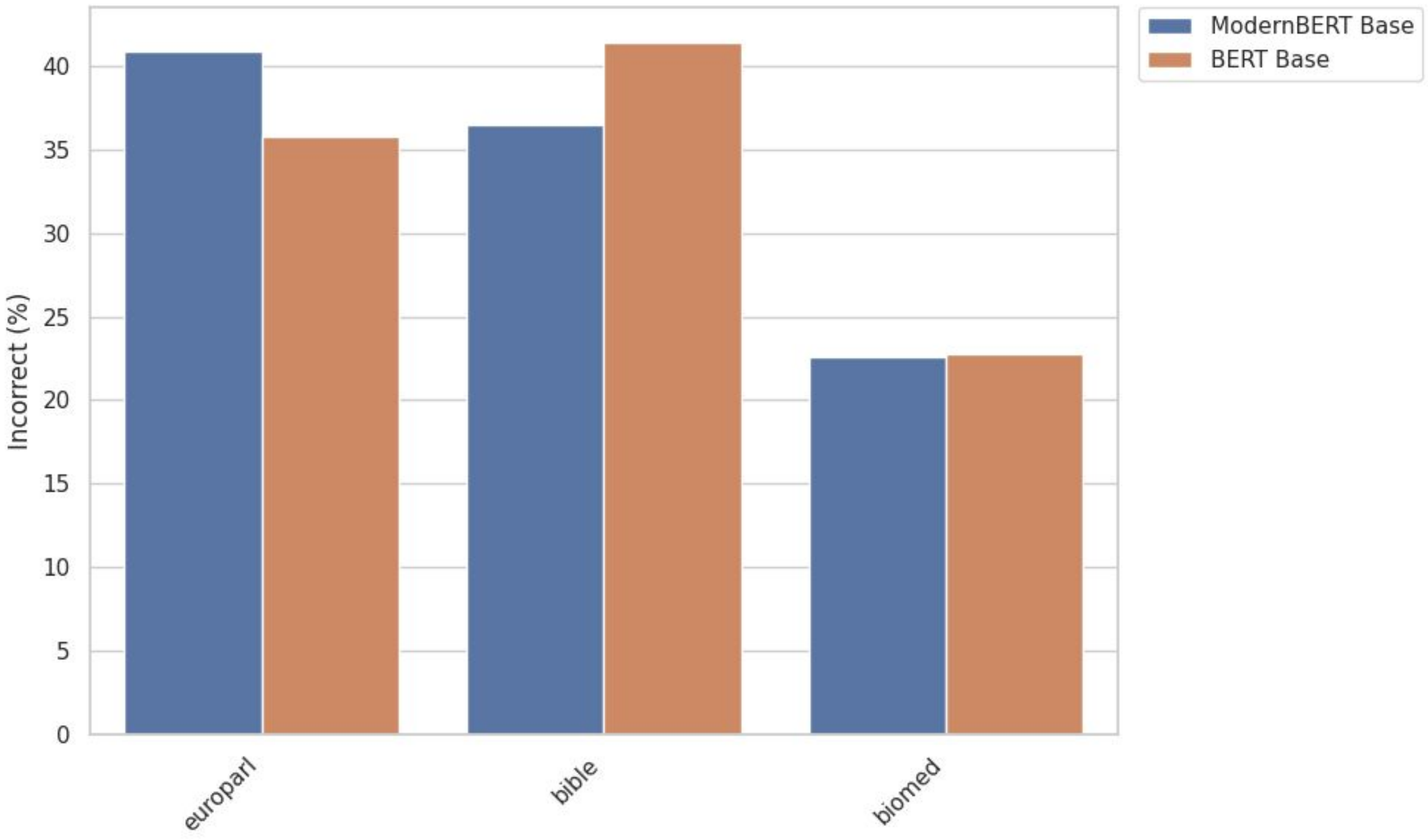
Berkeley SCHOOL OF INFORMATION

Key Results (Ablation Study)

Average Best & Worst Performing Experiment Designs, By Model

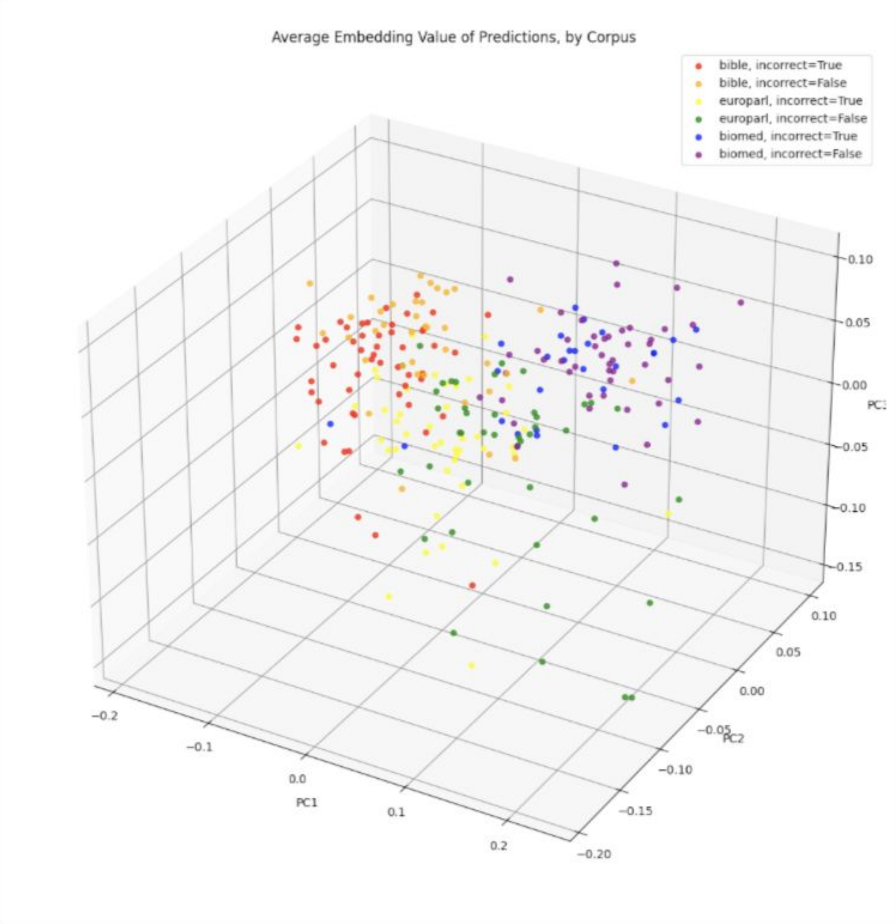
Model		Single Task	Multi Task	Avg Precision	Avg Recall	Avg F1
ModernBERT-base	Best	snc_pos_seq	sentence_no_contractions	**0.51198**	0.99896	**0.67647**
	Worst	sentence_no_contractions	snc_morph_alt	0.42717	0.87317	0.54788
ModernBERT-large	Best	pos_sequence	snc_dep_seq	0.50878	0.87812	0.64050
	Worst	snc_dep_seq	pos_sequence	0.24415	0.17324	0.20267
bert-base-cased	Best	*sentence*	snc_morph_seq	0.50948	**1.00000**	0.67457
	Worst	snc_pos_alt	snc_morph_alt	*0.00000*	*0.00000*	*0.00000*
bert-large-cased	Best	snc_pos_alt	snc_morph_alt	0.50698	0.85270	0.63150
	Worst	pos_sequence	snc_morph_complexity_value	0.25652	0.12755	0.13166

Error Analysis (By Domain)

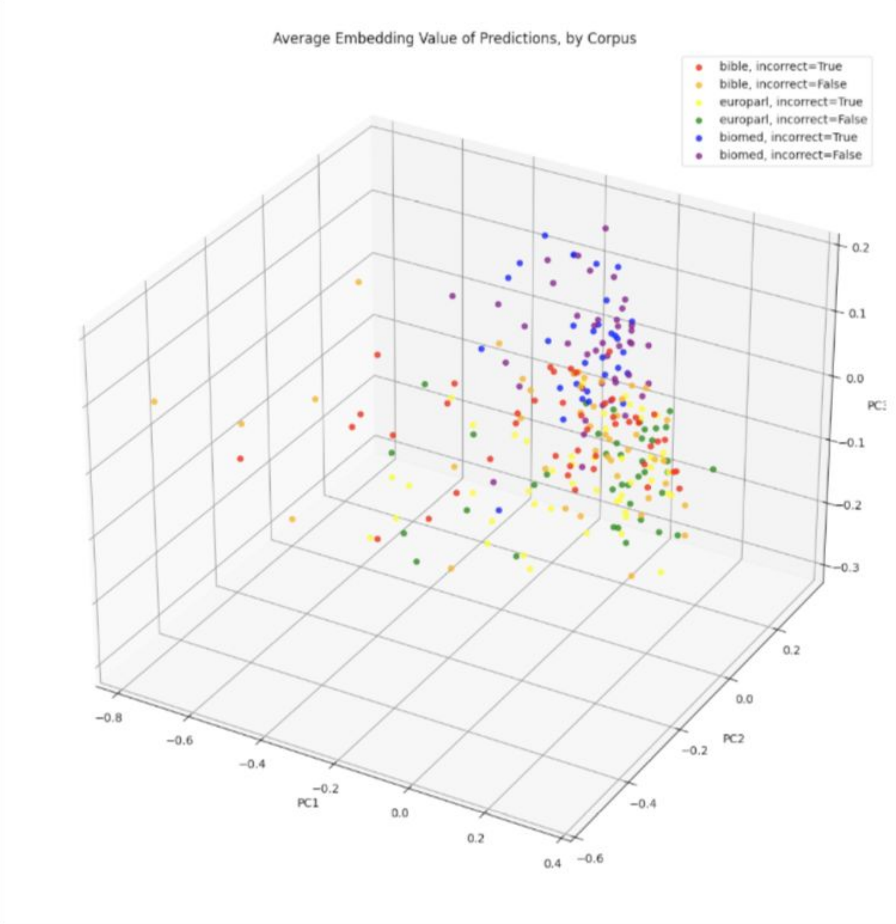


Average Symmetric KL Divergence of Predictions

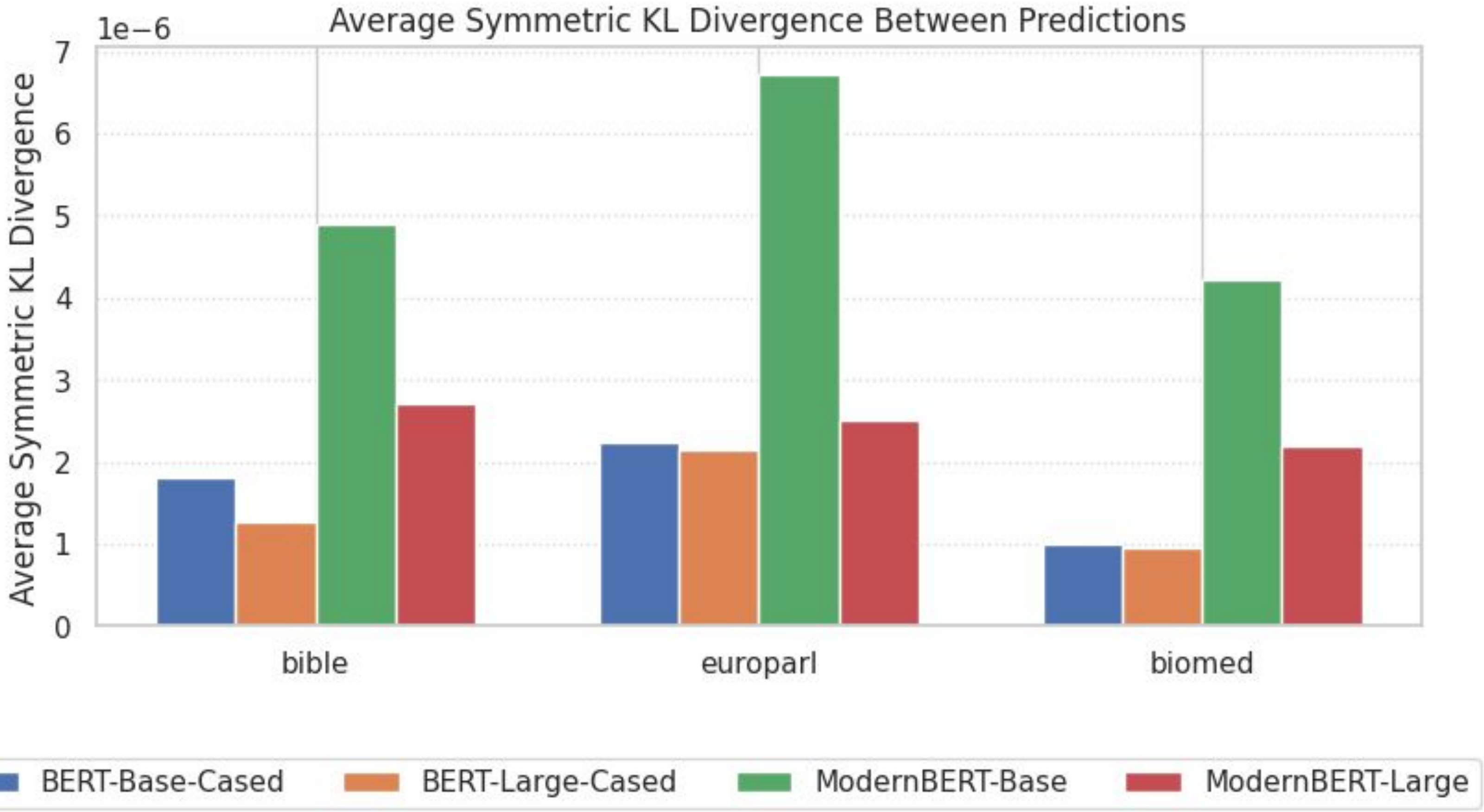
BERT-base-cased



ModernBERT-base



Average Symmetric KL Divergence of Predictions



Future Directions

- Data Quality:
  - Join the Datasets from both tasks
  - Use Cosine Similarity to drop additional training data, where complex unigram and bigram tokens are above a threshold that don't appear related to the sequence
- Model Training:
  - Perform Multi-Stage Fine-Tuning on Bible and European Parliament
  - Add token-level binary LCP task
  - Add text-generation task