XL-DURel: Finetuning Sentence Transformers for Ordinal Word-in-Context Classification

Master Thesis Presentation

Sachin Yadav

Institute of Natural Language Processing, University of Stuttgart

July 24, 2025

Supervisors: Dr. Dominik Schlechtweg
Lucas Möller

Introduction

Task: Ordinal Graded Word-in-Context (OGWiC) classification.

- Introduced in *CoMedi* shared task (Schlechtweg et al., 2025).
- Each instance:
  - A target word
  - Two sentences with the word in different contexts
- **Objective**: Predict ordinal semantic proximity label between 1 (unrelated) and 4 (identical)

Example:

Sentence 1: and taking a knife from her pocket, she opened a vein in her little **arm.**

Sentence 2: and though he saw her within reach of his **arm.**

Label: 4 (Identical)

Research Gap: Winning submissions do not exploit ordinal nature of labels, they are trained on binary labels.

Contribution: ***XL-DURel*****,** state-of-the-art model exploiting label ranking.

Background

WiC (Word-in-Context)Task

Determine if the word has the same meaning in both contexts

Labels: TRUE (same) or FALSE (different)

Sentence 1: The **bank** closes early on Friday.

Sentence 2: She deposited her paycheck at the **bank**.

Label: **TRUE (same meaning)**

Sentence 1. The **virus** infected his computer.

Sentence 2. The **virus** spread through the population.

Label: **FALSE (different meanings)**

Limitation

- Oversimplifies meaning shifts
- Ignores the nuanced, graded nature of semantic differences

Example:

Sentence 1: She rested her **arm** on the table.

Sentence 2: The statue's **arm** was broken during transport.

Label: **TRUE (Same meaning)**

But, is it really the exact **same meaning**?

GWiC (Graded Word in Context) Task

- Provide graded WiC predictions on an arbitrary scale (0-10).
- Treating the problem as a ranking task.

Example:
These young **men** displayed true Rajput chivalry. Akbar was so impressed with the bravery of these two **warriors** that he commissioned
Similarity Score: 7.88

Limitation

- Different target word.
- Model prediction is number between 0 and 1 but ground truth is between 0 and 10.
- How do they correspond?

OGWiC (Ordinal Graded Word-in-Context) Task

- Introduced as part of the CoMeDi shared task (Schlechtweg et al., 2025)
- Address the limitation of the WiC and GWiC tasks
- Ordinal classification task
- Nuanced and interpretable evaluation

Example:

1. She rested her **arm** on the table.

2.The statue's **arm** was broken during transport.

Label: **3 (Closely Related)**

Research Gap

- Two models excelled in the OGWiC Task
- Both use XLMR as base model.

XL-LEXEME:

- Bi-encoder
- Input: s_1 = ..., <t>, w_i, ..., </t>.....,

  s_2 = ..., <t>, w□, ..., </t>....
- Target word explicitly marked
- Contrastive Loss

DeepMistake:

- Cross-encoder
- Input: [CLS] s_1 [SEP] s_2 [SEP]
- Target word not marked
- Cross-entropy loss

- Train on **binary classification** but test on **ordinal classification** -> mismatch between training and testing task.

Task

- OGWiC task (Schlechtweg et al., 2025).
- Ordinal scale

Example:

Sentence 1: She deposited her paycheck at the **bank** on Friday.

Sentence 2: The **bank** closes early on weekends

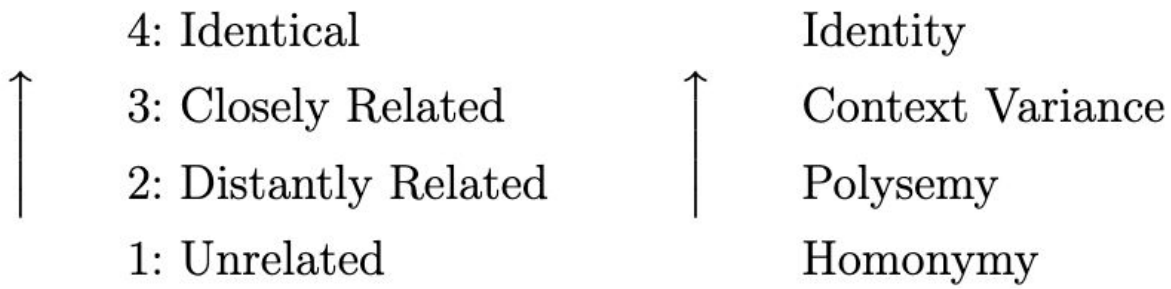
Label**: 4 (Identical)**

Sentence 1: The **virus** infected his computer and deleted files

Sentence 2: The **virus** spread rapidly through the population

Label**: 2 (Distantly Related)**

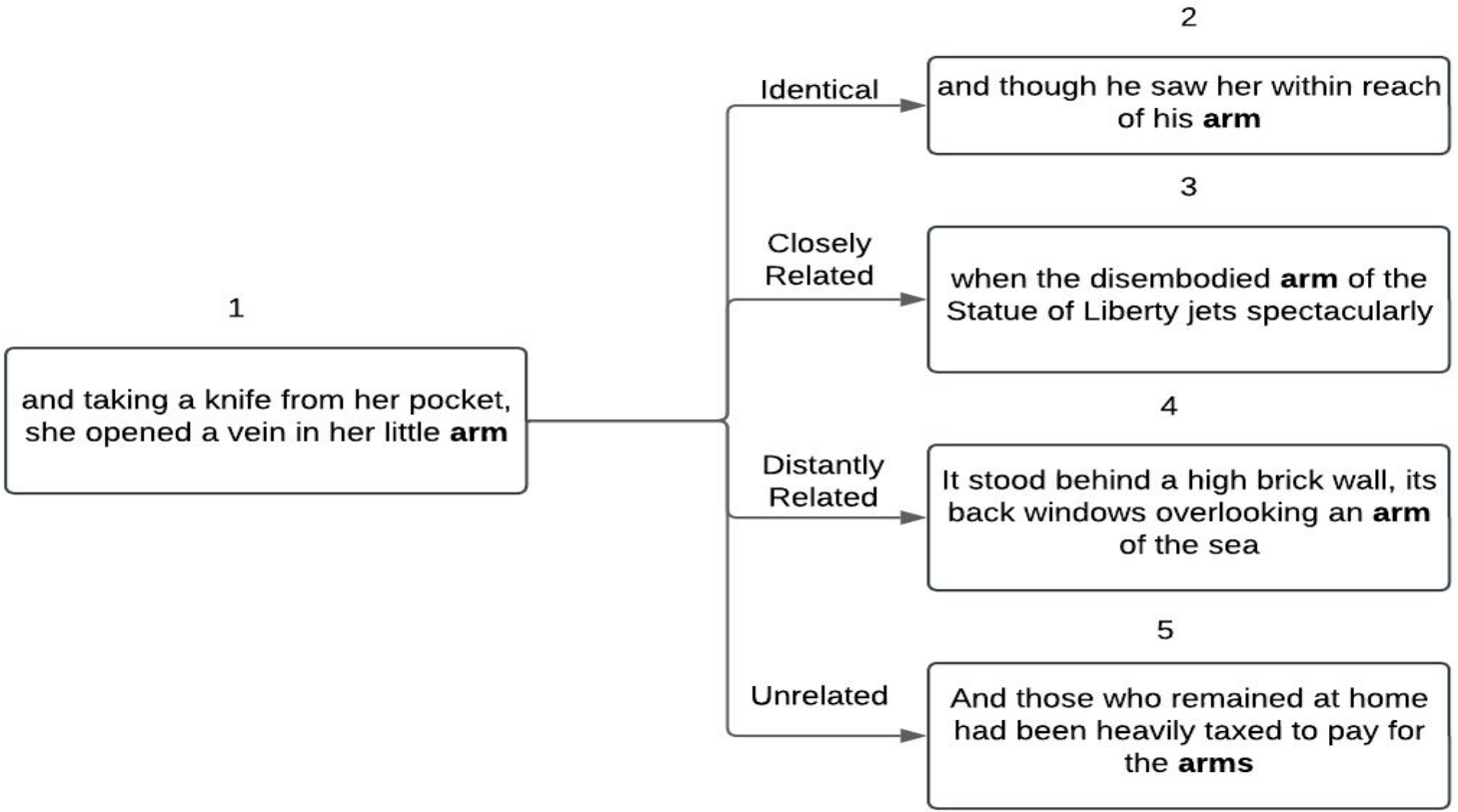
DURel Annotation Scale



The DURel relatedness scale (Schlechtweg et al., 2018).

- Each label of scale has linguistic interpretation
- That’s why it is interesting to exactly reproduce the label

Example:


1
and taking a knife from her pocket, she opened a vein in her little **arm**
Identical
2
and though he saw her within reach of his **arm**
Closely Related
3
when the disembodied **arm** of the Statue of Liberty jets spectacularly
Distantly Related
4
It stood behind a high brick wall, its back windows overlooking an **arm** of the sea
Unrelated
5
And those who remained at home had been heavily taxed to pay for the **arms**

Datasets

Datasets	Train	Dev	Test
CoMeDi	48k	8k	15k
WiC	251k	8k	6k
WiC + CoMeDi	299k	16k	21k

Table: Dataset Statistics

Evaluation Metrics

Krippendorff alpha

- Ordinal classification
- Penalizes larger prediction errors
- controls for chance

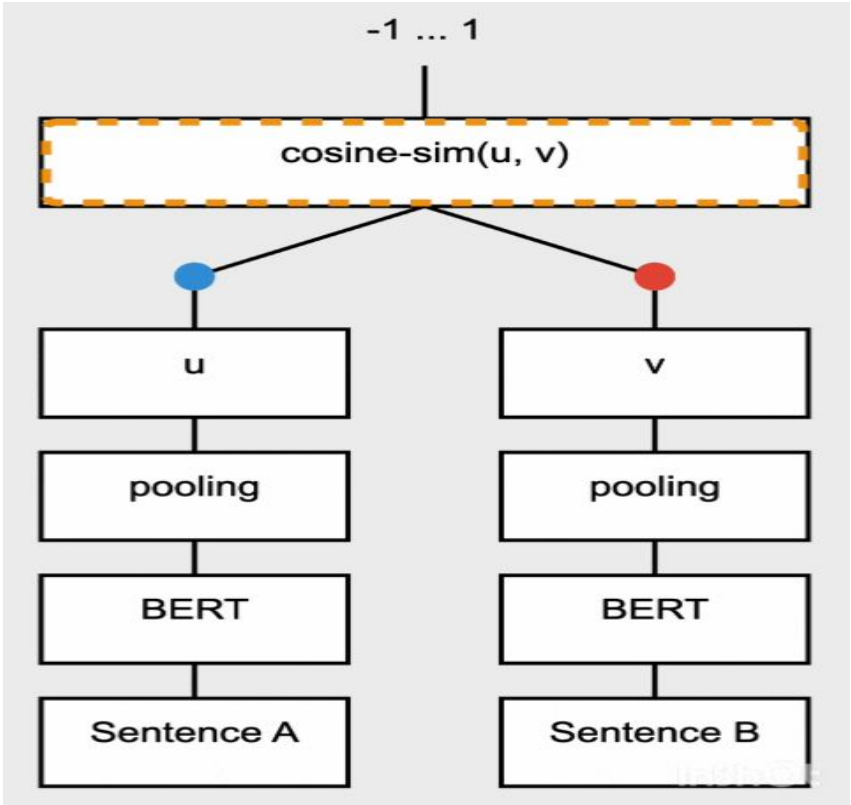
Spearman Ranking

- Ranking evaluation
- Used for validation on dev

Model

Sentence Transformer

- XLM-RoBERTa-large
- Based on Siamese or Triplet Network Architecture
- Dual BERT Encoders Run Simultaneously
- Embeddings are pooled on sentence level
- Cosine Similarity is optimized in finetuning



Loss Function

Contrastive Loss (Hadsell et al., 2006b)

- **Binary classification**
- Trained to push dissimilar pairs apart, pull similar pairs together
- Not ideal for ranking or ordinal classification

Cosine Similarity Loss

- **Regression**
- Minimizes squared error between predicted cosine similarity and gold label
- Not ideal for classification—assumes normal noise distribution

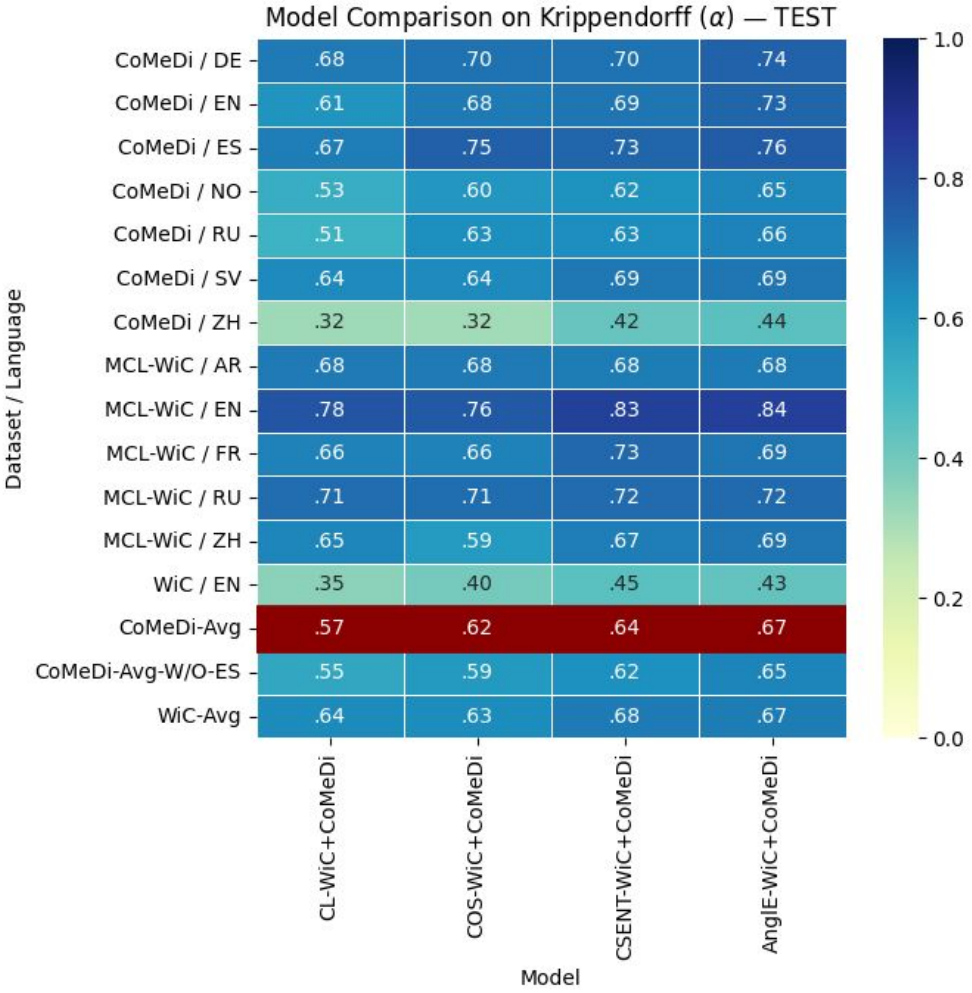
Consent Loss (Huang et al. (2024)

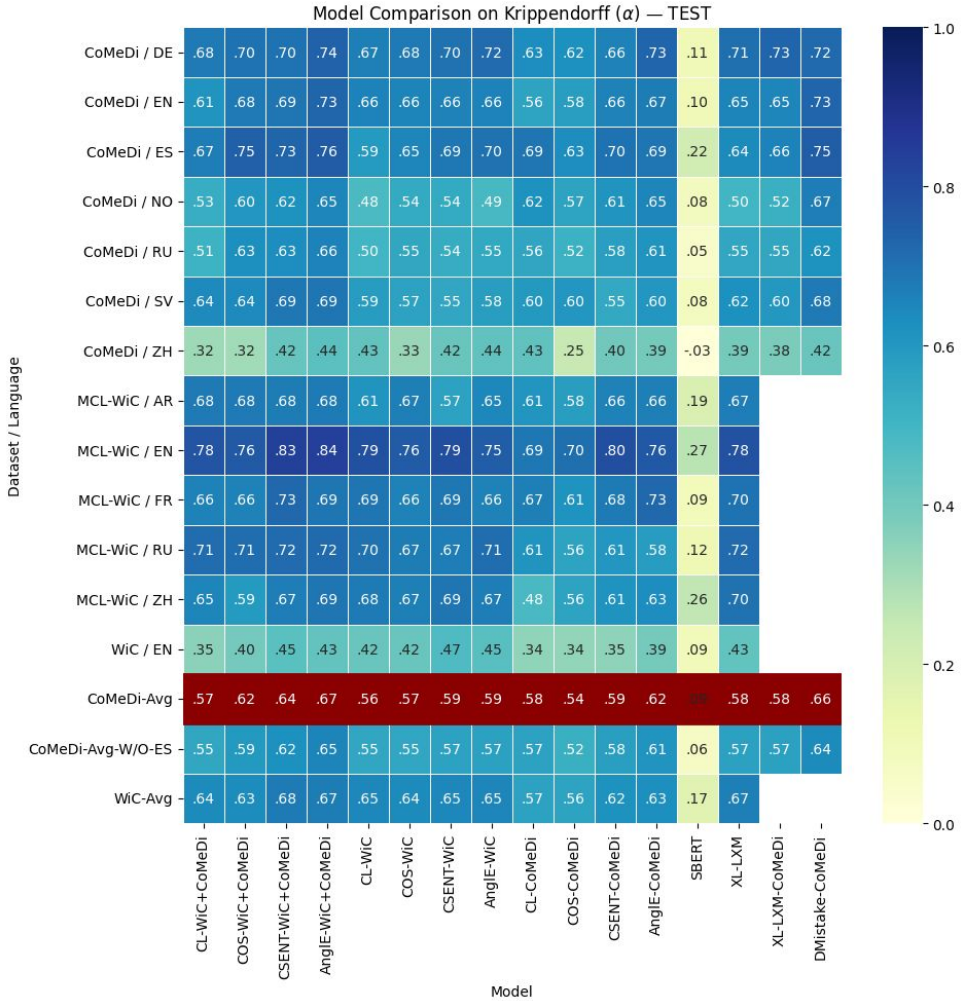
- **Ranking Objective**
- Optimizes pairs with higher labels should have higher similarity
- Better alignment with ordinal ranking
- Cosine gradient can vanish as similarity nears 1

Angle Loss (Li and Li, 2023)

- **Ranking objective**
- Same structure as CoSENT, but uses angle distance in complex space
- Fixes gradient issues of cosine similarity
- Better learning dynamics near extrema
- **Achieved top performance in the thesis**

Results


Model Comparison on Krippendorff (α) — TEST
Dataset / Language
CoMeDi / DE .68 .70 .70 .74
CoMeDi / EN .61 .68 .69 .73
CoMeDi / ES .67 .75 .73 .76
CoMeDi / NO .53 .60 .62 .65
CoMeDi / RU .51 .63 .63 .66
CoMeDi / SV .64 .64 .69 .69
CoMeDi / ZH .32 .32 .42 .44
MCL-WiC / AR .68 .68 .68 .68
MCL-WiC / EN .78 .76 .83 .84
MCL-WiC / FR .66 .66 .73 .69
MCL-WiC / RU .71 .71 .72 .72
MCL-WiC / ZH .65 .59 .67 .69
WiC / EN .35 .40 .45 .43
CoMeDi-Avg .57 .62 .64 .67
CoMeDi-Avg-W/O-ES .55 .59 .62 .65
WiC-Avg .64 .63 .68 .67
CL-WiC+CoMeDi
COS-WiC+CoMeDi
CSENT-WiC+CoMeDi
AngIE-WiC+CoMeDi
Model
1.0
0.8
0.6
0.4
0.2
0.0


Model Comparison on Krippendorff (α) — TEST
Dataset / Language
Model
CoMeDi / DE .68 .70 .70 .74 .67 .68 .70 .72 .63 .62 .66 .73 .11 .71 .73 .72
CoMeDi / EN .61 .68 .69 .73 .66 .66 .66 .66 .56 .58 .66 .67 .10 .65 .65 .73
CoMeDi / ES .67 .75 .73 .76 .59 .65 .69 .70 .69 .63 .70 .69 .22 .64 .66 .75
CoMeDi / NO .53 .60 .62 .65 .48 .54 .54 .49 .62 .57 .61 .65 .08 .50 .52 .67
CoMeDi / RU .51 .63 .63 .66 .50 .55 .54 .55 .56 .52 .58 .61 .05 .55 .55 .62
CoMeDi / SV .64 .64 .69 .69 .59 .57 .55 .58 .60 .60 .55 .60 .08 .62 .60 .68
CoMeDi / ZH .32 .32 .42 .44 .43 .33 .42 .44 .43 .25 .40 .39 -.03 .39 .38 .42
MCL-WiC / AR .68 .68 .68 .68 .61 .67 .57 .65 .61 .58 .66 .66 .19 .67
MCL-WiC / EN .78 .76 .83 .84 .79 .76 .79 .75 .69 .70 .80 .76 .27 .78
MCL-WiC / FR .66 .66 .73 .69 .69 .66 .69 .66 .67 .61 .68 .73 .09 .70
MCL-WiC / RU .71 .71 .72 .72 .70 .67 .67 .71 .61 .56 .61 .58 .12 .72
MCL-WiC / ZH .65 .59 .67 .69 .68 .67 .69 .67 .48 .56 .61 .63 .26 .70
WiC / EN .35 .40 .45 .43 .42 .42 .47 .45 .34 .34 .35 .39 .09 .43
CoMeDi-Avg .57 .62 .64 .67 .56 .57 .59 .59 .58 .54 .59 .62 .58 .58 .66
CoMeDi-Avg-W/O-ES .55 .59 .62 .65 .55 .55 .57 .57 .57 .52 .58 .61 .06 .57 .57 .64
WiC-Avg .64 .63 .68 .67 .65 .64 .65 .65 .57 .56 .62 .63 .17 .67
CL-WiC+CoMeDi
COS-WiC+CoMeDi
CSENT-WiC+CoMeDi
AnglE-WiC+CoMeDi
CL-WiC
COS-WiC
CSENT-WiC
AnglE-WiC
CL-CoMeDi
COS-CoMeDi
CSENT-CoMeDi
AnglE-CoMeDi
SBERT
XL-LXM
XL-LXM-CoMeDi
DMistake-CoMeDi
1.0
0.8
0.6
0.4
0.2
0.0

Conclusion

- AnglE Loss performs best across the board.
- Mixed training (CoMeDi+ WiC) gives best results.
- Optimizing for ordinal improves binary too.
- Outperforms DeepMistake benchmarks.
- Binary WiC = special case of Ordinal WiC.
- Ordinal WiC is the future.

Future Work

- Ordinal classification directly as loss function.
- Influence on other datasets.

Thank You

Any Questions?

References

Schlechtweg, D., Choppa, T., Zhao, W., & Roth, M. (2025). *CoMeDi Shared Task: Median judgment classification & mean disagreement ranking with ordinal word-in-context judgments*. In *Proc. of Context and Meaning: Navigating Disagreements in NLP Annotation* (pp. 33–47). Abu Dhabi, UAE: ICCL.

Schlechtweg, D., Schulte im Walde, S., & Eckmann, S. (2018). *DURel: A framework for annotating lexical semantic change*. NAACL-HLT, 169–174. https://www.aclweb.org/anthology/N18-2027/

Sakai, T. (2021). *Evaluating evaluation measures for ordinal classification and ordinal quantification*. In *Proc. of ACL-IJCNLP 2021 (Vol. 1: Long Papers)* (pp. 2759–2769).

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. EMNLP-IJCNLP, 3982–3992. https://doi.org/10.18653/v1/D19-1410

Arefyev, N., Fedoseev, M., Protasov, V., Homskiy, D., Davletov, A., & Panchenko, A. (2021). *DeepMistake: Which senses are hard to distinguish for a word-in-context model*. Dialog-21, 2021-June, 16–30.

Cassotti, P., Siciliani, L., de Gemmis, M., Semeraro, G., & Basile, P. (2023a). *XL-Lexeme: WIC pretrained model for cross-lingual lexical semantic change*. In Proceedings of the 61st Annual Meeting of the ACL (online).

Hadsell, R., Chopra, S., & Y.LeCun (2006). *Dimensionality reduction by learning an invariant mapping*. CVPR, 2, 1735–1742. https://doi.org/10.1109/CVPR.2006.100

Huang, X., Peng, H., Zou, D., Liu, Z., Li, J., Liu, K., Wu, J., Su, J., & Yu, P. S. (2024). *CoSENT: Consistent sentence embedding via similarity ranking*. IEEE/ACM Trans. Audio, Speech, Lang. Process., 32, 2800–2813. https://doi.org/10.1109/TASLP.2024.3402087

Kuklin, M., & Arefyev, N. (2025). *Deep-Change at CoMeDi: The cross-entropy loss is not all you need*. In *Proc. of the 1st Workshop on Context and Meaning – Navigating Disagreements in NLP Annotations*, Abu Dhabi.

Li, X., & Li, J. (2023). *Angle-optimized text embeddings*. arXiv:2309.12871.

Appendix: Cosine Similarity

$$\cos(u, v) = \frac{u \cdot v}{\|u\| \ \|v\|}$$

$$d(u, v) = 1 - \cos(u, v)$$

Appendix: ContrastiveLoss

$$\mathcal{L} = \frac{1}{2}\left[y \cdot d(u,v)^2 + (1-y) \cdot \max(0, m - d(u,v))^2\right]$$

where:

- d(u, v) is the distance between the embeddings, in our case the cosine distance
- m is the margin hyperparameter, which specifies the minimum required distance between dissimilar pairs

Appendix: Cosine Similarity Loss

$$\mathcal{L} = \|\cos(u, v) - y\|_2$$

where

- $\|\cdot\|_2$ is the L2 norm.

Appendix: CosentLoss

$$\mathcal{L} = \log\left(1 + \sum_{y(u,v)>y(k,l)} \exp\left(\lambda(s(k,l) - s(u,v))\right)\right)$$

where

- s(u, v) is the similarity between the embeddings, in our case the cosine simiarity.
- y(u, v) > y(k, l) defines the set of embedding pairs (k, l) for which the ground truth label y(k, l) is smaller than y(u, v) and
- λ is a hyperparameter for amplification

Appendix: Data Statistics

Language	Train Instances	Dev Instances	Test Instances
German	8,279	1,663	3,141
English	5,910	863	2,444
Swedish	5,457	871	1,245
Chinese	10,833	2,532	3,240
Spanish	4,821	621	1,497
Russian	8,029	1,126	2,285
Norwegian	4,504	611	1380
Total	**47,833**	**8,287**	**15,332**

Appendix: Label Mapping

Dataset	Train	Dev	Test	Cosine	Binary	Ordinal
CoMeDi	47,833	8,287	15,332	$1 \rightarrow 0.0$ $2 \rightarrow \frac{1}{3}$ $3 \rightarrow \frac{2}{3}$ $4 \rightarrow 1.0$	$1 \rightarrow 0$ $2 \rightarrow 0$ $3 \rightarrow 1$ $4 \rightarrow 1$	none
WiC	251,972	8,380	6400	$0 \rightarrow \frac{1}{3}$ $1 \rightarrow 1.0$	none	$0 \rightarrow 2$ $1 \rightarrow 4$
WiC+CoMeDi	299,805	16,667	21,732	as above	as above	as above

Training Parameters

Learning Rate: 1e-05

Batch Size: 32

Epoch: 10

Checkpoint was selected based on highest spearman score on validation data