SemEval-2012 Task 6: Semantic Textual Similarity

A Resource-Light Supervised Approach

Introduction to Human Language Technology (IHLT)

Zoë Finelli & Onat Bitirgen

December 2025

Task & Methodology

Objective → Predict the degree of semantic equivalence between two sentences on a continuous scale (0 to 5).

- 0: Different topics.
- 5: Completely equivalent.

Constraint: "Resource-Light." We avoided pre-trained Deep Learning embeddings (BERT/GloVe) to focus on interpretable linguistic features.

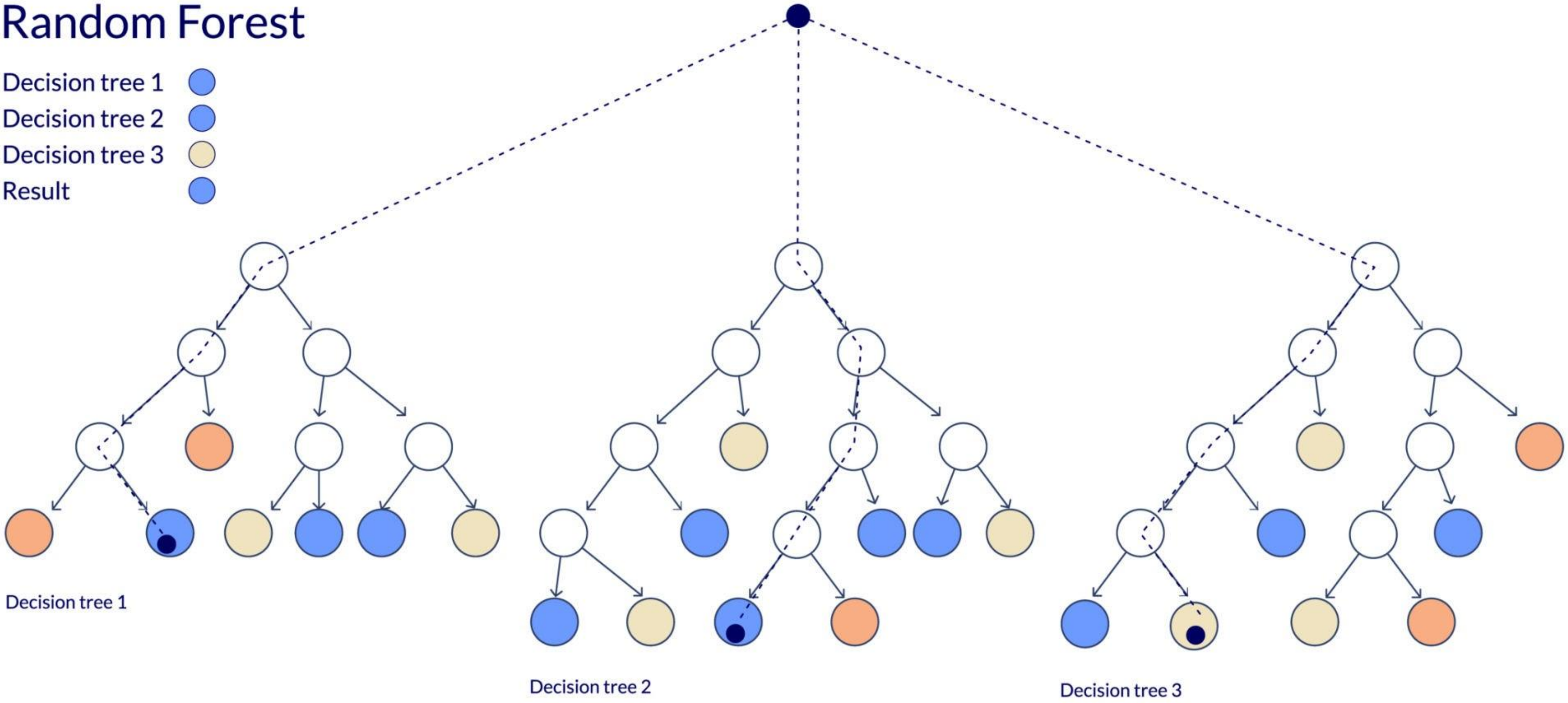
Approach: Supervised Machine Learning (Random Forest) with iterative feature engineering.

Inspiration: Built on top SemEval-2012 systems:

- **UKP Lab (Rank 1):** String similarity & log-linear models.
- **TakeLab (Rank 2):** Syntactic dependencies & SVR.

Random Forest

Decision tree 1	🔵
Decision tree 2	🔵
Decision tree 3	🟡
Result	🔵



Feature Engineering

Core Layers:

- **Lexical (Surface):** Jaccard Similarity, Overlap Coefficient, Length Difference. Based on UKP Lab's n-gram approach.
- **Semantic (Meaning):** WordNet Path Similarity. Captures synonyms (e.g., "car" $\approx$ "automobile"). Based on TakeLab.

Syntactic (Structure): spaCy Dependency Parsing. Checks if Subject, Object, and Root Verb align. **Domain Boosters:**

- **Negation:** Critical for News ("did" vs "did not").
- **Numbers:** Quantifies digit overlap ("$5" vs "$50"). ***MT Metrics:** BLEU & LCS (suited for Machine Translation datasets).

Phase 1 – The Baseline

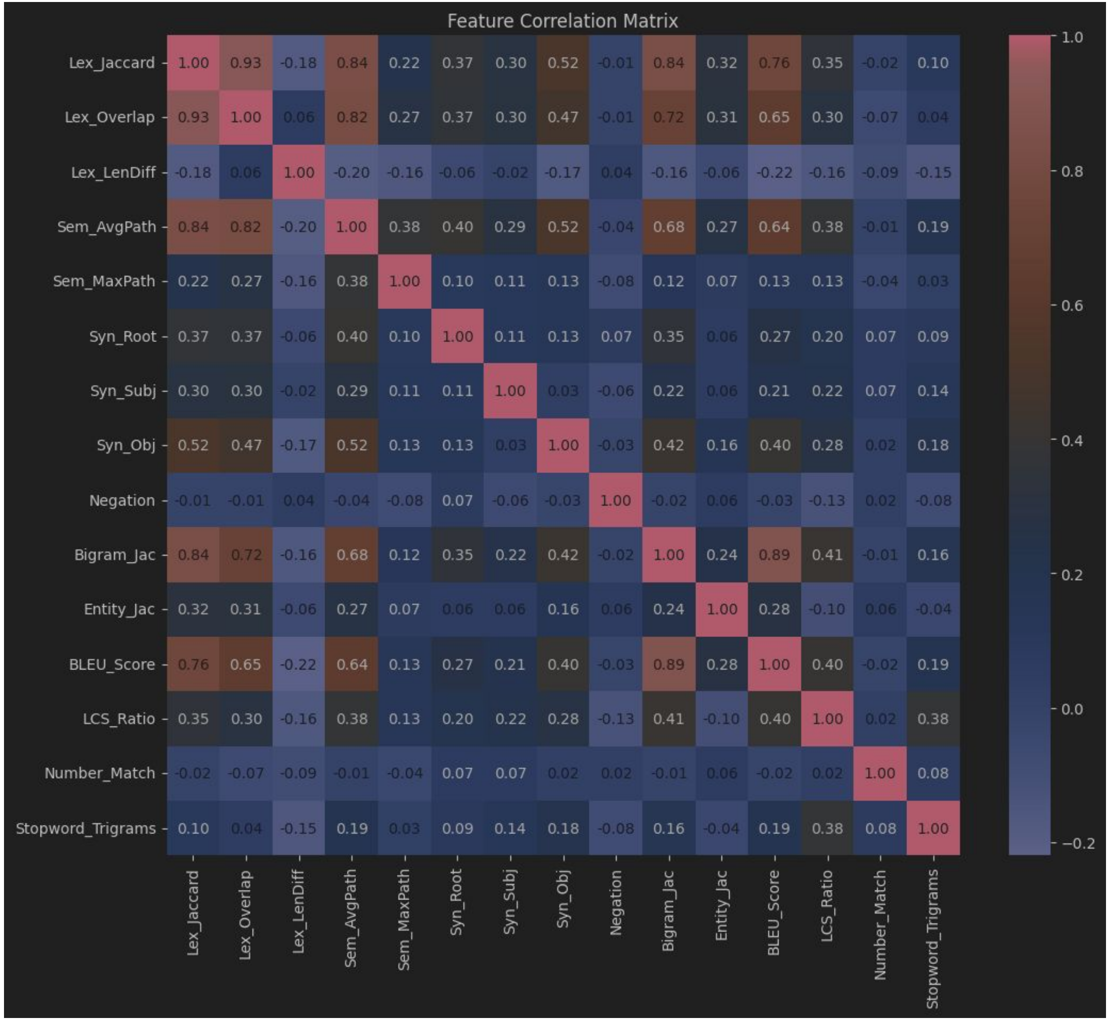
Hypothesis: Simple lexical overlap + WordNet synonyms should capture basic similarity.

Initial Features:

- Lexical: Jaccard Similarity, Overlap Coefficient.
- Semantic: WordNet Path Similarity (capturing "car" ≈ "automobile").

Initial Results:

- 🟢 MSRvid: 0.763 (Success: Works well on simple descriptions).
- 🔴 MSRpar: 0.402 (Failure: Cannot distinguish complex news headlines).



Feature Correlation Matrix

	Lex_Jaccard	Lex_Overlap	Lex_LenDiff	Sem_AvgPath	Sem_MaxPath	Syn_Root	Syn_Subj	Syn_Obj	Negation	Bigram_Jac	Entity_Jac	BLEU_Score	LCS_Ratio	Number_Match	Stopword_Trigrams
Lex_Jaccard	1.00	0.93	-0.18	0.84	0.22	0.37	0.30	0.52	-0.01	0.84	0.32	0.76	0.35	-0.02	0.10
Lex_Overlap	0.93	1.00	0.06	0.82	0.27	0.37	0.30	0.47	-0.01	0.72	0.31	0.65	0.30	-0.07	0.04
Lex_LenDiff	-0.18	0.06	1.00	-0.20	-0.16	-0.06	-0.02	-0.17	0.04	-0.16	-0.06	-0.22	-0.16	-0.09	-0.15
Sem_AvgPath	0.84	0.82	-0.20	1.00	0.38	0.40	0.29	0.52	-0.04	0.68	0.27	0.64	0.38	-0.01	0.19
Sem_MaxPath	0.22	0.27	-0.16	0.38	1.00	0.10	0.11	0.13	-0.08	0.12	0.07	0.13	0.13	-0.04	0.03
Syn_Root	0.37	0.37	-0.06	0.40	0.10	1.00	0.11	0.13	0.07	0.35	0.06	0.27	0.20	0.07	0.09
Syn_Subj	0.30	0.30	-0.02	0.29	0.11	0.11	1.00	0.03	-0.06	0.22	0.06	0.21	0.22	0.07	0.14
Syn_Obj	0.52	0.47	-0.17	0.52	0.13	0.13	0.03	1.00	-0.03	0.42	0.16	0.40	0.28	0.02	0.18
Negation	-0.01	-0.01	0.04	-0.04	-0.08	0.07	-0.06	-0.03	1.00	-0.02	0.06	-0.03	-0.13	0.02	-0.08
Bigram_Jac	0.84	0.72	-0.16	0.68	0.12	0.35	0.22	0.42	-0.02	1.00	0.24	0.89	0.41	-0.01	0.16
Entity_Jac	0.32	0.31	-0.06	0.27	0.07	0.06	0.06	0.16	0.06	0.24	1.00	0.28	-0.10	0.06	-0.04
BLEU_Score	0.76	0.65	-0.22	0.64	0.13	0.27	0.21	0.40	-0.03	0.89	0.28	1.00	0.40	-0.02	0.19
LCS_Ratio	0.35	0.30	-0.16	0.38	0.13	0.20	0.22	0.28	-0.13	0.41	-0.10	0.40	1.00	0.02	0.38
Number_Match	-0.02	-0.07	-0.09	-0.01	-0.04	0.07	0.07	0.02	0.02	-0.01	0.06	-0.02	0.02	1.00	0.08
Stopword_Trigrams	0.10	0.04	-0.15	0.19	0.03	0.09	0.14	0.18	-0.08	0.16	-0.04	0.19	0.38	0.08	1.00

Phase 2 – Context & Debugging

Diagnosis: The low MSRpar score suggested the model missed "logical" differences or context.

- *Example:* "The bird is flying" vs "The bird is not flying."

Features Added:

- **Negation Detection:** Explicit check for "not", "no", "never", "n't".
- **Named Entities:** Matching people, organizations, and locations via spaCy.

Result → 🟢 MSRpar improved to 0.512 (+0.11 gain)

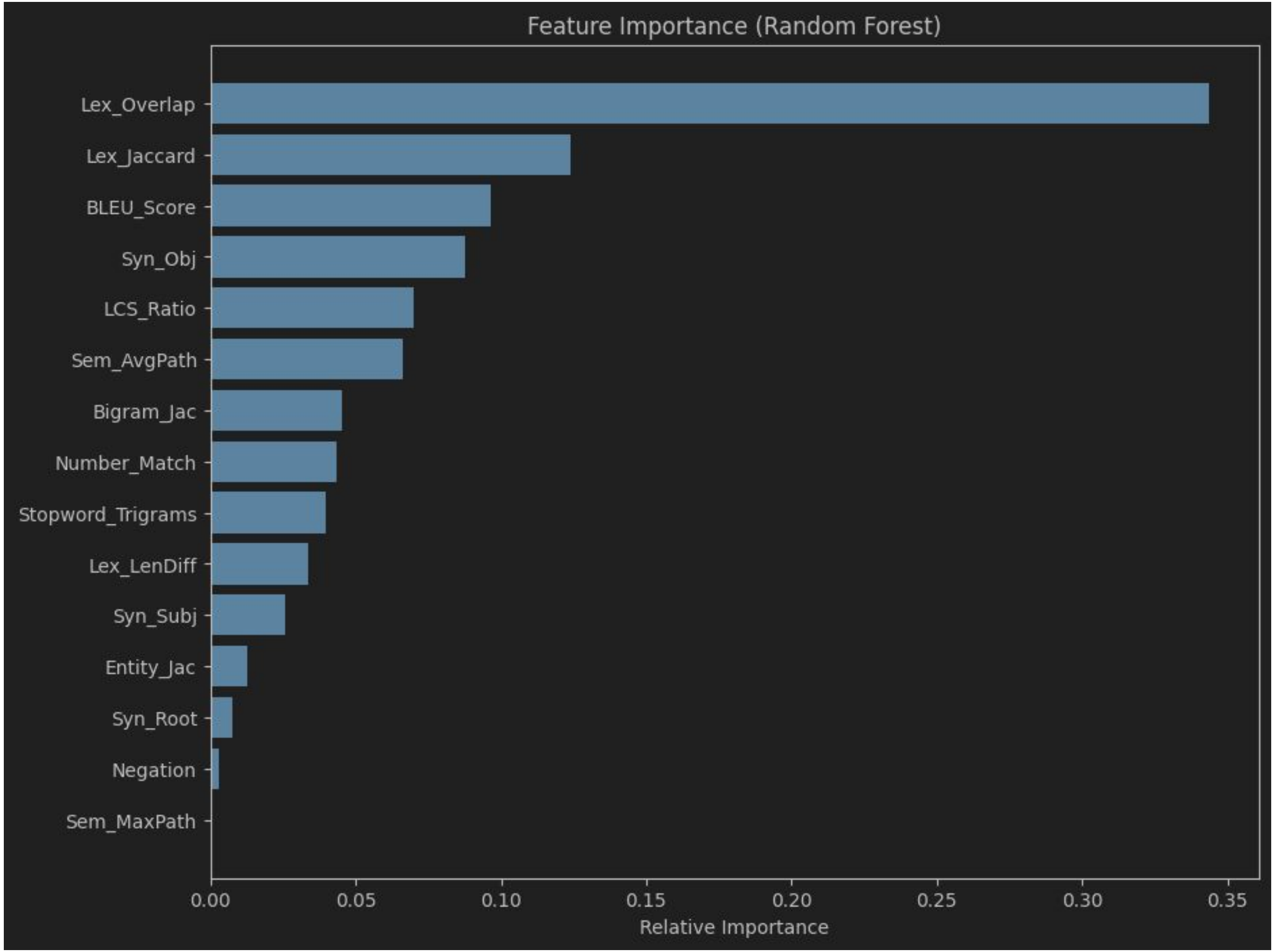
Phase 3 – Domain Adaptation

The News Problem: News datasets rely heavily on specific dates, money, and quantities. Standard similarity metrics ignore these "details."

Domain Boosters:

- **Number Matching:** Quantifies overlap of digits/years ("$5" vs "$50").
- **BLEU Score:** Standard metric for Machine Translation evaluation.
- **LCS Ratio:** Longest Common Subsequence to check word order.

Result → 🟢 MSRpar exploded to 0.661. Number matching was the key differentiator.


Feature Importance (Random Forest)
Lex_Overlap
Lex_Jaccard
BLEU_Score
Syn_Obj
LCS_Ratio
Sem_AvgPath
Bigram_Jac
Number_Match
Stopword_Trigrams
Lex_LenDiff
Syn_Subj
Entity_Jac
Syn_Root
Negation
Sem_MaxPath
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Relative Importance

Phase 4 – Stylistic Refinement

The SMT Problem: Syntactic parsers punish "garbled" output from Machine Translation datasets (SMTeuroparl/SMTnews), lowering scores.

The Fix:

- **Stopword n-Grams:** Captures structural/stylistic similarity (e.g., "of the...", "in the...") without demanding perfect grammar.

Final System:

- **Total Features:** 15 Dimensions.
- **Model:** Random Forest (Handles boolean and float features robustly).

Analysis – Model & Feature Selection

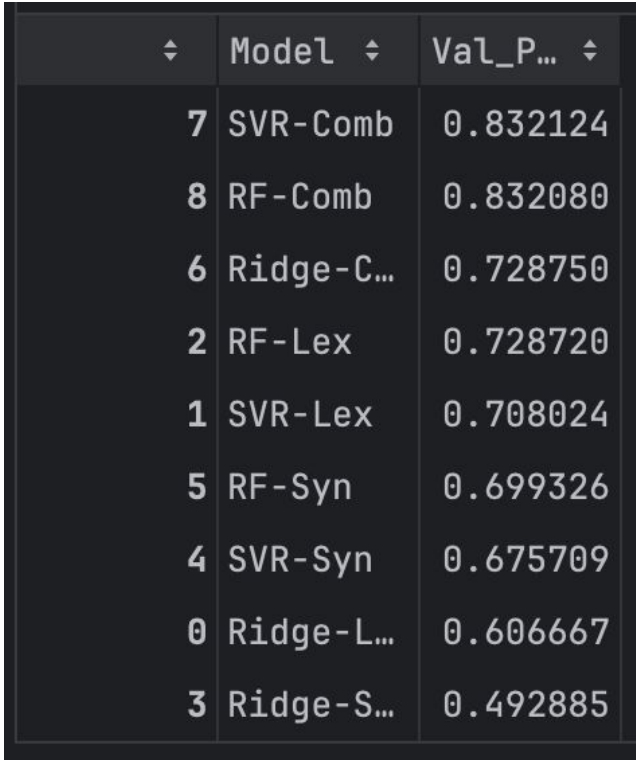
Ablation Study (Validation): We compared models on isolated feature sets (Lexical-Only vs. Syntactic-Only).

- **Result:** The Combined Ensemble outperformed the individual baselines by ~0.10 Pearson *r.*
- **Conclusion:** This suggests that syntax acts as a multiplier for lexical features rather than a replacement.

Parsimony Check: We tested retraining the model on subsets of features (Top-5 vs. All-15).

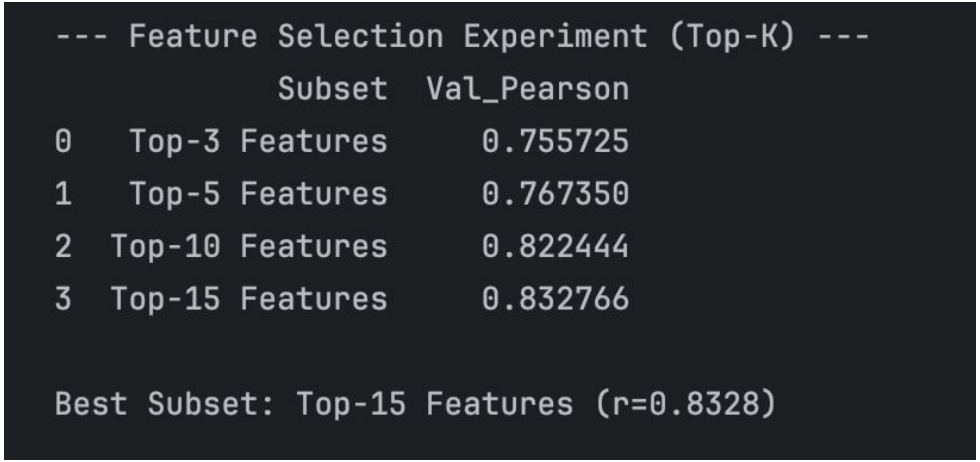
- **Top-5 Features:** $r = 0.767$
- **All 15 Features:** $r = 0.833$
- **Finding:** Pruning features caused a notable drop in accuracy. We retained the full 15-feature set for the final system.

Cell 3 Output: Validation Leaderboard Table



	Model	Val_P…
7	SVR-Comb	0.832124
8	RF-Comb	0.832080
6	Ridge-C…	0.728750
2	RF-Lex	0.728720
1	SVR-Lex	0.708024
5	RF-Syn	0.699326
4	SVR-Syn	0.675709
0	Ridge-L…	0.606667
3	Ridge-S…	0.492885

Cell 5 Output: Feature Selection Table



```
--- Feature Selection Experiment (Top-K) ---
             Subset  Val_Pearson
0    Top-3 Features     0.755725
1    Top-5 Features     0.767350
2   Top-10 Features     0.822444
3   Top-15 Features     0.832766

Best Subset: Top-15 Features (r=0.8328)
```

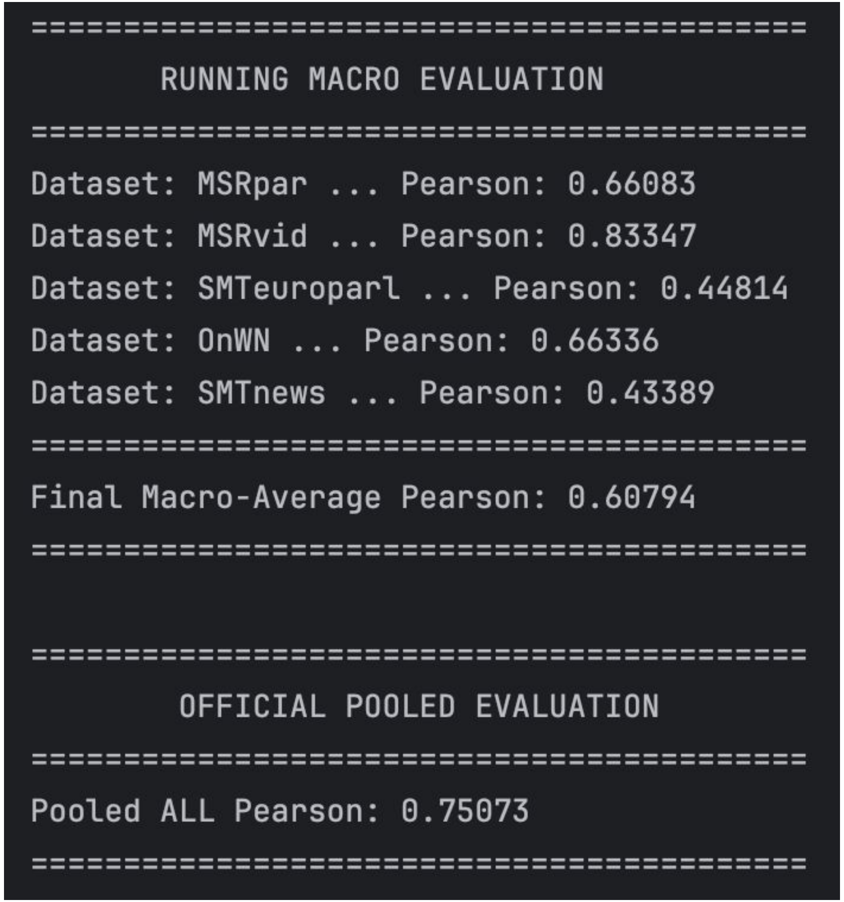
Final Evaluation (Test Set)

Summary:

- **Final Pooled ALL:** 0.751 (Substantially higher than the official SemEval baseline of 0.311)

Dataset	Pearson (*r*)	Performance
MSRvid	0.833	🟢 Excellent (Simple descriptions)
OnWN	0.663	🟡 Good (Definitions)
MSRpar	0.661	🟡 Good (News Paraphrases)
SMTnews	0.433	🔴 Fair (Noisy Input)
SMTeuroparl	0.448	🔴 Fair (Noisy Input)

Cell 7 Output: *evaluate.sh* Evaluation



```
==========================================
       RUNNING MACRO EVALUATION
==========================================
Dataset: MSRpar ... Pearson: 0.66083
Dataset: MSRvid ... Pearson: 0.83347
Dataset: SMTeuroparl ... Pearson: 0.44814
Dataset: OnWN ... Pearson: 0.66336
Dataset: SMTnews ... Pearson: 0.43389
==========================================
Final Macro-Average Pearson: 0.60794
==========================================

==========================================
        OFFICIAL POOLED EVALUATION
==========================================
Pooled ALL Pearson: 0.75073
==========================================
```

Comparison with State-of-the-Art – Complexity vs. Efficiency

UKP Lab (Rank 1): Utilized massive external knowledge bases (Wikipedia, Wiktionary) and 300+ features (mostly n-grams).

TakeLab (Rank 2): Heavily relied on "Distributional Semantics" (LSA) trained on large corpora (NYT/Wikipedia) and syntactic dependencies.

- **Our Approach: "Resource-Light."** We used 0 external corpora and only 15 features.
- **Conclusion:** We achieved comparable performance on MSRvid with <5% of computational complexity.

System	**Pooled ALL**	**MSRpar**	**MSRvid**	**SMTeuroparl**	**OnWN**	**SMTnews**
UKP Lab (Rank 1)	**0.823**	0.724	**0.868**	**0.528**	**0.679**	**0.398**
TakeLab (Rank 2)	0.813	**0.734**	0.880	0.477	**0.679**	0.400
Official Baseline	0.311	0.433	0.299	0.454	0.586	0.390
Our System	**0.751**	**0.661**	**0.833**	**0.448**	**0.663**	**0.433***

Conclusion & Future Work

Success: Achieved >0.8 correlation on validation/video data using only classical NLP (no Neural Networks).

Key Insight: No "one-size-fits-all" feature exists. The Ensemble allows the model to switch strategies (e.g., using **Negation** for News vs. **Jaccard** for Videos).

Future Work:

- **Distributional Semantics:** Incorporate LSA/LDA to handle idioms where word overlap is zero.
- **Knowledge Graphs:** Better entity linking for proper nouns.

References

[1] D. Bär, C. Biemann, I. Gurevych, and T. Zesch, "UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM 2012), Montréal, Canada, 2012, pp. 435–440.

[2] F. Šarić, G. Glavaš, M. Karan, J. Šnajder, and B. Dalbelo Bašić, "TakeLab: Systems for Measuring Semantic Text Similarity," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM 2012), Montréal, Canada, 2012, pp. 441–448.