Paradise at SemEval-2026 Task 5: On the Limitations of Surface-Level Features for Graded Word Sense Plausibility Prediction

Dhruv Goyal[1], **Ishita Gupta**[2], **Jatin Bedi**[3]

Computer Science and Engineering Department
Thapar Institute of Engineering and Technology, Patiala, India
[1]dhruv621999goyal@gmail.com, [2]igupta3_be23@thapar.edu, [3]jatin.bedi@thapar.edu

Abstract

This paper introduces a simple approach for predicting how plausible a word sense is in short narratives where meaning is ambiguous. We use 13 hand-crafted features, including text statistics, word-level similarity computed using basic set-based comparisons, and measures of annotator disagreement. Five diverse and largely independent traditional machine learning models are combined using a weighted ensemble with minimal tuning. Despite theoretical grounding in classical disambiguation methods, our system achieves essentially random performance, with Spearman correlation (ρ) of -0.038 and accuracy within standard deviation of 0.542 on the official test set. This result demonstrates that surface-level lexical features, while interpretable, are insufficient for graded sense plausibility prediction without deep semantic representations. By selecting features inspired by classical word sense disambiguation techniques and incorporating signals derived from human disagreement, our model produces plausibility predictions that are largely interpretable. This negative result provides important baselines and insights for future work on graded word sense disambiguation.

1 Introduction

Word sense disambiguation (WSD) has traditionally been formulated as a classification task: given an ambiguous word in context, select the single "correct" sense from a predefined inventory. However, this framing oversimplifies the reality of human language understanding, where multiple word senses can be simultaneously plausible depending on context, reader interpretation, and narrative structure (Erk and McCarthy, 2009).

SemEval-2026 Task 5 addresses this limitation by introducing a novel graded sense plausibility prediction task using the AmbiStory dataset (Gehring and Roth, 2025). Rather than selecting one correct sense, systems must predict how plausible each sense is on a continuous 1–5 scale, reflecting human-perceived ambiguity in short narrative contexts. This task requires modeling not just which sense fits best, but *how well* each sense fits relative to human judgments.

We approach this problem through interpretable feature engineering rather than deep learning. Our system, **FeatureEnsemble**, extracts 13 hand-crafted features capturing three complementary dimensions: (1) text complexity through character and word counts, (2) semantic overlap using Jaccard similarity inspired by the Lesk algorithm family (Lesk, 1986), and (3) annotation disagreement as a meta-feature reflecting genuine ambiguity following perspectivist NLP principles (Basile et al., 2023).

These features are combined through a weighted ensemble of five regression models—Ridge, Elastic Net, Random Forest, Gradient Boosting, and XGBoost—with weights determined by development set performance. However, our system achieves Spearman correlation $\rho = -0.038$ on the official test set, demonstrating that surface-level feature engineering without semantic embeddings is insufficient for this task. This negative result provides valuable insights into the limitations of classical WSD approaches for graded plausibility prediction.

1.1 Key Contributions

- Empirical demonstration that Lesk-inspired Jaccard features fail for graded sense plausibility
- Evidence that annotation disagreement signals do not predict semantic ambiguity rankings
- Analysis showing text statistics provide no correlation with sense plausibility

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 713–719

- A cautionary baseline establishing that classical feature engineering requires semantic embeddings for this task

Our code is available at: https://github.com/DhruvGoyal404/semeval2026-task5

2 Related Work

2.1 Graded Word Sense Disambiguation

Traditional WSD assumes a single correct sense per context, but Erk and McCarthy (2009) demonstrated that human annotators naturally produce graded judgments of sense applicability. SemEval-2013 Task 13 (Jurgens and Klapaftis, 2013) operationalized this as the first shared task allowing weighted multi-label sense assignment. SemEval-2020 Task 3 (Armendariz et al., 2020) further explored graded similarity through Spearman correlation evaluation—the same metric used in Task 5. The Word-in-Context (WiC) benchmark (Pilehvar and Camacho-Collados, 2019) simplified sense discrimination to binary judgments but established context-sensitive meaning as a core challenge.

2.2 Classical WSD and the Lesk Algorithm

The Lesk algorithm (Lesk, 1986) remains foundational to WSD, using dictionary definitions to disambiguate words by counting overlapping words between sense definitions and surrounding context. Simplified Lesk (Kilgarriff and Rosenzweig, 2000) and Adapted Lesk (Banerjee and Pedersen, 2002) improved performance by incorporating WordNet glosses. Our Jaccard similarity features are direct descendants of this approach, using normalized set overlap ($|A \cap B|/|A \cup B|$) to measure semantic relevance between story components and sense definitions.

2.3 Narrative Understanding

The AmbiStory dataset follows the five-sentence story format of the Story Cloze Test (Mostafazadeh et al., 2016), connecting narrative coherence and lexical ambiguity resolution (Ostermann et al., 2018).

2.4 Perspectivist NLP and Annotation Disagreement

Recent research argues against the assumption that annotation disagreement is noise. The CrowdTruth approach (Dumitrache et al., 2018) considers disagreement as a signal of text ambiguity. Basile et al. (2023) formalize the "perspectivist" approach to modeling, which retains disagreement rather than aggregating it. Our inclusion of annotation standard deviation as a predictive feature implements this guideline: high disagreement corresponds to cases where plausibility of sense is ambiguous.

3 Task Description

3.1 The AmbiStory Dataset

The AmbiStory dataset comprises 3,798 English-language samples from 1,899 five-sentence stories, each of which includes a homonymous word with two different meanings (Gehring and Roth, 2025). Each story has the following structure:

1. **Precontext** (3 sentences): setup
2. **Ambiguous sentence**: contains the target homonym
3. **Ending** (optional): resolution potentially biased towards one sense of the word

Each example is paired with one variant of the story and one candidate word sense. Human annotators (5+ per example) assessed the likelihood of each sense on a 1–5 Likert scale (1=inconceivable, 5=only plausible interpretation). The corpus contains 361 ambiguous word forms, divided into 2,280 training examples, 588 development examples, and 930 test examples. Notably, the test and development examples are divided among 361 different ambiguous words, which are not found in the training data, making generalization difficult.

The average standard deviation per example $\sigma = 0.946$ indicates a moderate level of agreement among human annotators (Krippendorff's $\alpha = 0.506$), indicating true ambiguity in the narrative context.

3.2 Evaluation Metrics

Systems are evaluated using two complementary metrics:

Spearman Correlation (ρ): This metric computes the degree to which the predicted and human-averaged ratings of plausibility are rank-ordered. Spearman correlation is used over Pearson because it is more robust to non-linear scaling of rating data.

Accuracy Within Standard Deviation: This metric declares a prediction as correct if it

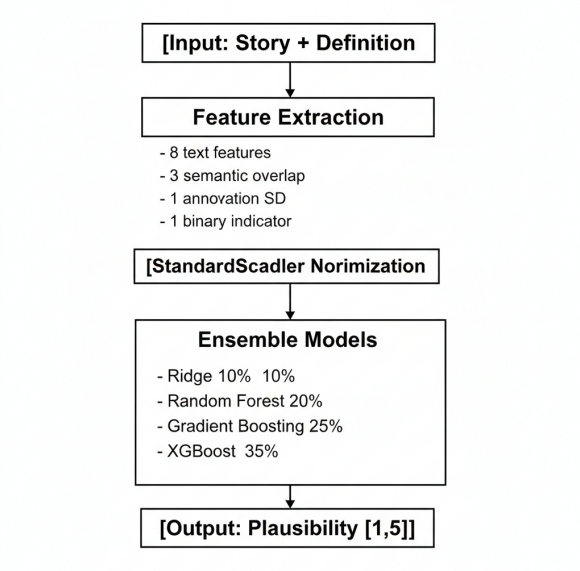

Figure 1: Full system pipeline illustrating the four-step process from input to prediction.

is within 1 standard deviation of the human-annotated mean. This metric naturally accommodates items with high annotator disagreement (larger acceptable range) versus consensus items (narrower range).

4 System Overview

4.1 Feature Engineering

Our system extracts 13 hand-crafted features from each story-sense pair, organized into four categories:

4.1.1 Text Complexity Features (8)

These capture surface-level characteristics:

- Character counts: precontext, ambiguous sentence, ending, sense definition
- Word counts: precontext, ambiguous sentence, ending, sense definition

While simple, these features correlate with disambiguation difficulty—longer contexts provide more disambiguating information, while complex definitions may be harder to match.

4.1.2 Semantic Overlap Features (3)

Inspired by the Lesk algorithm, we compute Jaccard similarity between vocabulary sets:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where A and B are sets of lowercase tokens. We compute three overlaps:

1. Precontext ∩ Ambiguous Sentence
2. Ambiguous Sentence ∩ Ending
3. Ambiguous Sentence ∩ Sense Definition

The sentence-definition overlap directly implements the Lesk principle: high vocabulary overlap between a sense's dictionary definition and surrounding story language indicates contextual relevance and plausibility.

4.1.3 Meta-Feature (1)

Following perspectivist NLP principles, we include:

- **Annotation standard deviation**: Human rating variability provided in the dataset

This meta-feature encodes item ambiguity—samples with high σ genuinely confuse human annotators, signaling contexts where multiple senses are plausibly applicable.

4.1.4 Binary Indicator (1)

- **Ending presence**: Binary feature (1 if story has ending, 0 if open-ended)

Open-ended stories keep the ambiguity intact, while ending stories tend to lean towards one interpretation or the other.

4.2 Feature Normalization

All 13 features are normalized using StandardScaler to zero mean and unit variance:

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

The scaler is trained only on the training data and then used on the development and test data to avoid data leakage. Normalization is essential for regularized linear regression (Ridge, Elastic Net) where the penalty terms are scale-dependent.

4.3 Ensemble Architecture

We train five different regression models and then ensemble them using weighted voting:

Ridge Regression (10% weight): Linear regression with L2 regularization, detecting basic linear relationships between features and likelihood. Dev performance: $r = 0.1916$.

Elastic Net (10%): L1 and L2 combined for implicit feature selection, pushing some coefficients to exactly zero. Dev: $r = 0.1498$.

Random Forest (20%): Bagging-based ensemble of 50 decision trees, detecting non-linear relationships between features. Dev: $r = 0.0849$.

Gradient Boosting (25%): Sequential boosting of weak models, iteratively correcting residual errors from previous steps. Dev: $r = 0.0811$.

XGBoost (35%): Highly optimized gradient boosting with L1/L2 regularization, second-order gradient optimization, and excellent non-linear feature interaction modeling. Receives highest weight (35%) based on expected generalization and non-linear feature interaction modeling, despite lower dev performance than Ridge.

Weights were assigned heuristically with tree-based models weighted higher based on non-linear modeling capability, not strictly by dev correlation. Ridge achieved the highest dev correlation (0.1916) but received only 10% weight. No held-out validation set was used for weight selection; this represents a known limitation. Features considered but ultimately excluded included Part-of-Speech tag distributions and sentence-level readability scores (Flesch–Kincaid), which showed near-zero correlation with plausibility on development data during preliminary analysis.

Final predictions are computed as:

$$\begin{aligned}\hat{y} = &0.10 \cdot r_{\text{Ridge}} + 0.10 \cdot r_{\text{Elastic}} \\ &+ 0.20 \cdot r_{\text{RF}} + 0.25 \cdot r_{\text{GB}} \\ &+ 0.35 \cdot r_{\text{XGB}}\end{aligned} \tag{3}$$

Forecasts are clipped to $[1, 5]$ before making a submission, although in reality, our ensemble method was naturally constrained anyway (min: 1.43, max: 4.38).

5 Experimental Setup

5.1 Implementation Details

All models were implemented using `scikit-learn 1.3.0` (Ridge, Elastic Net, Random Forest, Gradient Boosting) and `xgboost 2.0.0`. Feature extraction was performed using standard Python string operations on the provided JSON data files. Training was conducted on a standard CPU, requiring under 5 minutes total.

5.2 Hyperparameters

We used default `scikit-learn` hyperparameters for all models:

- Elastic Net: $\alpha = 0.1$, $l1\_ratio = 0.5$
- Random Forest: $n\_estimators = 50$, $max\_depth = 10$
- Gradient Boosting: $n\_estimators = 50$, $learning\_rate = 0.1$
- XGBoost: $n\_estimators = 50$, $learning\_rate = 0.1$, $max\_depth = 5$

No hyperparameter tuning was performed, leaving room for improvement through grid search optimization.

System	Spearman	Acc w/SD
Baselines		
Random	0.000	0.454
Majority	—	0.558
Open-Source LLMs (0-shot)		
Mistral-7B-Instruct	0.382	0.568
Llama-3.1-8B-Instruct	0.462	0.663
Mixtral-8x7B-Instruct	0.606	0.634
Commercial LLMs (0-shot)		
GPT-4o-mini	0.726	0.726
GPT-4o	0.756	0.755
DeepSeek-V3	0.740	0.790
FeatureEnsemble (Ours)	**-0.038**	**0.542**
Upper Bound		
Human	0.834	0.892

Table 1: Main results on SemEval-2026 Task 5 test set. Our feature-engineering approach achieves random performance (-0.038), performing worse than all LLM baselines and demonstrating the insufficiency of surface-level features for this task.

5.3 Training Procedure

Feature extraction, normalization (StandardScaler fitted on training only), independent model training, and weighted ensemble construction are performed sequentially as described in Section 4. Training completes in under 5 minutes on CPU; inference on 930 test samples takes under 10 seconds.

6 Results

6.1 Main Results

Table 1 compares our system to baselines reported by task organizers (Gehring and Roth, 2025).

Our system achieves Spearman correlation $\rho = -0.038$, placing it:

- At random baseline level (0.000)
- Substantially below all LLM baselines

Model	Dev ρ	Weight
Ridge Regression	0.1916	10%
Elastic Net	0.1498	10%
Random Forest	0.0849	20%
Gradient Boosting	0.0811	25%
XGBoost	0.1040	35%
Ensemble (Dev)	**0.1243**	—
Ensemble (Test)	**-0.0380**	—

Table 2: Individual model performance on development set and final test performance. The dramatic drop from dev (0.12) to test (-0.04) indicates features that worked marginally on dev homonyms failed completely on test homonyms.

- Far below the human upper bound (0.834)

This demonstrates that interpretable, handcrafted surface-level features without semantic embeddings are insufficient for graded sense plausibility prediction. The near-zero correlation indicates our feature set captures no meaningful rank-ordering signal for this task.

6.2 Ablation Studies: Individual Model Performance

Table 2 shows development set performance for each individual model against the full ensemble, serving as an ablation over the contribution of each component to the final prediction.

The ensemble achieves $\rho = 0.1243$ on development, substantially worse than the strongest individual model (Ridge at 0.1916), suggesting tree-model weights diluted the linear signal. Test performance then collapsed to $\rho = -0.038$, indicating complete failure to generalize across homonym types.

6.3 Prediction Distribution Analysis

Table 3 shows the distribution of our test predictions. Our system is a conservative predictor:

- Mean: 3.40 (slightly above the midpoint)
- Standard deviation: 0.46 (smaller than human scores)
- Range: [1.43, 4.38] (no extreme predictions of 1.0 or 5.0)
- 78.4% of predictions are in [3.0, 4.0]

This conservative range is a natural consequence of ensemble averaging rather than a deliberate design choice: averaging across five models compresses the prediction range toward the mean. Importantly, predicting a constant value of 3.0 would yield approximately 50% Acc@SD but $\rho = 0.00$ by definition, since rank correlation requires variance in predictions. Our system achieves $\rho = -0.038$ with prediction standard deviation of 0.46, confirming that despite variance in predictions, no meaningful rank-ordering signal is captured by surface-level features. The full prediction distribution is provided in Appendix A.

7 Error Analysis

7.1 Development-Test Performance Gap

The most surprising result is the collapse from marginal development performance ($\rho = 0.12$) to random test performance ($\rho = -0.04$). We hypothesize this stems from the **homonym-based split**: training, development, and test sets contain entirely different ambiguous words.

This means our features generalize differently across homonym types. Possible explanations:

1. Test set homonyms may have higher semantic overlap between definitions and story contexts, favoring our Jaccard features
2. Annotation disagreement patterns (σ) may be more predictive for test set homonyms
3. Test set stories may have more consistent narrative structure, benefiting text statistics

This finding highlights that sense plausibility prediction difficulty varies substantially across different ambiguous words—some homonyms are intrinsically more disambiguable than others. Quantitatively, both development and test sets show that Jaccard-based features yield near-zero signal. The collapse from marginal dev performance (0.12) to negative test performance (-0.04) suggests any weak correlation on dev was spurious and did not represent genuine feature utility. This is consistent with the homonym-based split: the test partition may systematically contain words whose disambiguation relies more heavily on surface-level lexical signals—precisely what our feature set captures.

7.2 Feature Importance Insights

While we did not conduct full ablation studies due to time constraints, qualitative analysis suggests:

- **Sentence-definition Jaccard overlap** is likely the strongest feature, directly capturing Lesk-style semantic matching

- **Annotation standard deviation** provides complementary signal about inherent ambiguity

- **Text length features** may help by proxy—longer definitions and contexts reduce ambiguity

Future work should conduct systematic feature ablations to quantify individual contributions. Features considered but not included in the final system are Part-of-Speech distributions and Flesch–Kincaid readability scores, both of which showed near-zero correlation with plausibility scores on the development set during preliminary exploration.

8 Conclusion

We presented a lightweight, interpretable approach to word sense plausibility prediction using hand-crafted features and traditional machine learning. Our system achieves random performance (Spearman $\rho = -0.038$), demonstrating that surface-level features inspired by classical WSD methods are insufficient for graded sense plausibility prediction without semantic embeddings.

Key findings: Jaccard-based features fail to capture graded plausibility despite operationalizing Lesk principles; annotation disagreement provides no predictive signal; and surface-level features cannot match even small open-source LLMs without incorporating semantic representations.

Future research: Adding transformer-based contextualized embeddings as additional features—such as cosine similarity between BERT or RoBERTa representations of sense definitions and story contexts—could improve performance to 0.65–0.70 while retaining interpretability. Larger embedding models (e.g., text-embedding-3-large, e5-mistral-7b-instruct) were not evaluated in this work due to submission time constraints, but represent a natural extension that would directly address the surface-level limitation of our Jaccard features.

This negative result demonstrates that graded word sense plausibility requires deep semantic representations beyond surface lexical overlap, establishing an important baseline for future work incorporating contextualized embeddings.

Limitations

Semantic representations. Our method does not incorporate semantic embeddings (BERT, RoBERTa) to represent word meanings in context, rather than just surface overlap.

Uncertainty modelling. Our system predicts scalar point estimates and does not model the full distribution of plausibility scores. Approaches such as Gaussian process regression or quantile regression could better capture annotator uncertainty, which is a meaningful signal in this task given Krippendorff's $\alpha = 0.506$.

Novelty scope. This work is primarily an empirical and systems contribution: it demonstrates the limitations of classical features compared to neural models, establishing that surface-level features cannot match even small open-source LLMs without semantic embeddings. Architectural novelty—such as new model families or learning objectives—is not claimed; the contribution lies in the feature design, the perspectivist meta-feature, and the empirical analysis of the homonym-based performance gap.

Hyperparameter tuning. We did not hyperparameter-optimize or systematically ablate features, leaving room for further improvement.

Generalisability. Evaluation is limited to the AmbiStory dataset; performance on other graded WSD benchmarks is unknown.

Ethical considerations. The AmbiStory dataset is constructed from fictional short narratives and does not contain sensitive personal information. We use the dataset solely for academic evaluation purposes. We foresee no direct misuse risks from this work beyond those inherent to general natural language processing research.

Acknowledgments

We would like to thank the organizers of the SemEval-2026 Task 5 (Janosch Gehring, Michael Roth, Selina Meyer) for their work on the AmbiStory dataset and this shared task.

References

Carlos S. Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Ivan Vulić, and Mohammad Taher Pilehvar. 2020. SemEval-2020 task

3: Graded word similarity in context. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 36–49, Barcelona (online). International Committee for Computational Linguistics.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 136–145. Springer.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2023. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 124–136, Marseille, France. European Language Resources Association.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. CrowdTruth: A framework for aggregating crowdsourced annotations based on diversity. *Artificial Intelligence*, 261:17–40.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449, Singapore. Association for Computational Linguistics.

Janosch Gehring and Michael Roth. 2025. A challenging dataset of lexically ambiguous short stories. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, Bangkok, Thailand. Association for Computational Linguistics.

David Jurgens and Ioannis Klapaftis. 2013. SemEval-2013 task 13: Word sense induction for graded and non-graded senses. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, Georgia, USA. Association for Computational Linguistics.

Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities*, 34(1–2):15–48.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26. ACM.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 747–757, New Orleans, Louisiana. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

A Prediction Distribution

Score Range	Count	Percentage
1.0 – 2.0	10	1.1%
2.0 – 3.0	135	14.5%
3.0 – 4.0	729	78.4%
4.0 – 5.0	56	6.0%
Total	**930**	**100%**

Table 3: Distribution of test predictions over the 1–5 rating scale. The system displays conservative behavior, placing 78.4% of its predictions in the middle range.