A WordNet-Aligned Taxonomy of Logical Fallacies for Low-Resource NLP

Stanislav Penkov
GATE Institute
Sofia University "St. Kliment Ohridski"
stanislav.penkov@gate-ai.eu

Abstract

We present a compact taxonomy of eight logical fallacies designed for the analysis of manipulative reasoning and argumentative discourse, with a specific focus on low-resource settings. The taxonomy consolidates multiple public datasets into a nine-class training target (eight fallacies plus a background label *no fallacy*), with class distributions documented to support future stratified sampling or class-weighted modeling, and organizes these classes within a small parent–child hierarchy (relevance, presumption, causality, reasoning form) used for interpretability only (not as training labels), semantically grounding the target categories in WordNet 3.0 via synset IDs and offsets. From corpus evidence, we extract representative lexical cues using log-odds weighting with lemmatization and bigram detection, and link them to synsets to form a two-layer conceptual–lexical mapping. Crucially, the WordNet anchoring enables cross-lingual projection via Open Multilingual WordNet (OMW), supporting the creation of candidate seed lexicons and potential weak-supervision signals without prior in-language fallacy annotation. The framework thus provides a practical recipe for taxonomy-based lexical-semantic resource development, multilingual bootstrapping, and interpretable, fallacy-aware natural language processing (NLP).

Keywords: logical fallacies, WordNet, Open Multilingual WordNet, lexical-semantic resources, low-resource NLP, argumentation

1 Introduction

Logical fallacies — recurrent patterns of flawed reasoning that undermine the power of argumentation (Tindale, 2007) — are pervasive in public discourse. They are central to manipulative reasoning — a key mechanism through which information manipulation and propaganda operate in public discourse. Logical fallacy detection has thus emerged as an important component of explainable manipulative reasoning and argumentation analysis, where shallow linguistic cues often mask deeper reasoning flaws. However, this task is hindered by a fragmented resource ecosystem. Existing datasets adopt inconsistent label sets, complicating reuse and comparison across studies. Furthermore, prominent resources such as LOGIC (Jin et al., 2022), MAFALDA (Helwe et al., 2024), and CoCoLoFa (Yeh et al., 2024) focus almost exclusively on English and exhibit fragmented label schemes that hinder cross-lingual transfer. To address this, we propose a compact taxonomy of eight fallacy categories (with a separate *no fallacy* background label) semantically grounded in WordNet 3.0 (Princeton University, 2006), providing a stable schema that is portable across languages and resources. For low-resource languages such as Bulgarian — used here as a first test case — where annotated argumentation data is scarce, this semantically grounded taxonomy provides a crucial foundation for future resource building and for aligning large language model (LLM) outputs with interpretable reasoning categories. We will register the taxonomy, mappings, and normalized counts in the LRE Map to facilitate reuse and comparability across languages.

This work forms part of a broader research line on reasoning-aware NLP and manipulative rhetoric detection. Within this program, the present study contributes the linguistic and semantic infrastructure — a compact, WordNet-aligned taxonomy and normalized dataset — intended to support subsequent modules for fallacy detection and explainable argument analysis in low-resource languages.

Contributions. (i) A unified 8-class fallacy schema (plus a *no fallacy* background label) with transparent mappings from five widely used datasets. We selected the eight most stable, cross-source categories with clean WordNet anchors; rarer or source-

specific labels are retained in metadata only;

(ii) a WordNet 3.0-anchored conceptual layer with stable offsets, coupled with a data-driven lexical layer of representative cues;

(iii) a cross-lingual projection recipe that leverages Open Multilingual WordNet (OMW) to derive seed lexicons and weak supervision signals for low-resource languages without prior in-language fallacy annotation, illustrated through Bulgarian projection examples;

and (iv) a reproducible normalization pipeline that others can adopt to extend the resource.

2 Related Work

Research in argument mining and fallacy detection has proposed numerous taxonomies and datasets across diverse discourse domains, ranging from debate platforms (Habernal and Gurevych, 2017) and essays (Stab and Gurevych, 2017) to social-media and news discussions (Yeh et al., 2024). However, their label definitions and taxonomic structures vary substantially, limiting cross-dataset comparability and semantic reuse.

Early efforts such as *Argotario* (Habernal et al., 2017) gamified the creation of fallacy-annotated arguments through multilingual player interactions, introducing a scalable, crowd-driven approach to fallacy data collection. Related work by Habernal and Gurevych (2017) identified argumentative relations in user-generated web discourse, while Stab and Gurevych (2017) analyzed insufficiently supported claims in essays. Fallacy-specific expansions followed, such as *ad hominem* recognition — the identification of personal attacks directed at the speaker rather than the argument itself — (Habernal et al., 2018) and reasoning comprehension benchmarks (Wachsmuth et al., 2017).

A major step toward standardized evaluation came with Jin et al. (2022), who formalized the detection of logical fallacies as an NLP task and released the `LOGIC` and `LOGIC_CLIMATE` datasets (13 classes, ∼3.5k examples). Their structure-aware classifier masked semantically similar spans to highlight logical form over content, providing the first strong reasoning-based baseline. Sourati et al. (2023) extended this work via case-based reasoning (CBR), retrieving and adapting analogous examples enriched with goals, counterarguments, and structural annotations to improve cross-domain robustness. Lei and Huang (2024) further advanced structural modeling by introducing a Logical Structure Tree representation integrated into large language models through hard and soft prompts, achieving state-of-the-art results.

More recently, Yeh et al. (2024) combined LLM-assisted crowdsourcing and expert review to construct the largest English fallacy dataset (7,706 annotated news comments across eight categories). By integrating GPT-4 (OpenAI, 2023) as an in-task writing assistant, it demonstrated how hybrid human–AI pipelines can generate linguistically rich and contextually realistic fallacious discourse that closely resembles "in-the-wild" arguments. In parallel, Helwe et al. (2024) unified several existing corpora under a harmonized taxonomy, further establishing multi-domain evaluation standards.

Despite these advances, most resources standardize format rather than semantics. Few have linked fallacy or argument-type taxonomies to general lexical-semantic databases such as WordNet (Princeton University, 2006) and FrameNet (Baker et al., 1998). To our knowledge, fallacy categories have rarely been aligned with such lexical-semantic resources at the schema level, constraining semantic interoperability and multilingual reuse. Our work addresses this gap by pursuing taxonomy-driven, multilingual alignment between fallacy categories and general lexical resources, enabling cross-lingual representation within a unified conceptual framework.

3 Taxonomy Construction and Dataset Normalization

Recent work (e.g., Helwe et al. 2024) has reviewed extensive fallacy catalogues, but most lack semantic grounding for computational use. We prioritize interpretability and tractability. We consolidate the empirically represented fallacy types across five public datasets—Argotario, `LOGIC_EDU`, `LOGIC_CLIMATE`, MAFALDA, and CoCoLoFa—normalize them to a common JSONL schema, and merge into 12 531 examples before filtering, yielding a unified schema suitable for multilingual, reasoning-aware NLP.

Selection criteria. The eight fallacy categories were selected according to four practical criteria: (i) cross-source stability, i.e., the category or a close equivalent occurs in more than one existing resource or is clearly represented in a major source dataset; (ii) sufficient empirical support after normalization, so that the class can be meaningfully profiled using corpus-derived lexical evidence; (iii)

semantic anchorability, meaning that the category can be linked to one or more interpretable WordNet concepts without relying only on dataset-specific labels; and (iv) usefulness for low-resource transfer, where broad and recurrent reasoning patterns are preferable to fine-grained labels that cannot be reliably projected across languages. Rarer, highly source-specific, or theoretically ambiguous labels were therefore retained in metadata rather than included in the main training target. This design favours a compact and reusable schema over exhaustive coverage.

Parent–Child structure. Beyond the nine target classes used for modeling (eight fallacies plus `no_fallacy`), we attach a compact set of parent families that explain *why* the reasoning fails: *relevance* (redirection/irrelevant warrant), *presumption* (claim exceeds evidence), *causality* (unwarranted causal structure), and *reasoning form* (self-supporting loop). This hierarchy is used only for interpretability and error analysis and does not define the training label space. The intermediate label `causal_fallacy` consolidates source labels related to false causality and causal oversimplification for taxonomic analysis. It is assigned to the parent family `causality`, but is not included in the nine-class training target. The original labels and their intermediate mapping are retained in metadata rather than silently reassigned to another target class.

Consolidation rules (child → unified). We harmonize heterogeneous labels in lowercase with underscores. Table 1 shows the child → unified mappings used to normalize sources; arrows indicate the normalized unified label; not all unified labels are members of the nine-class target. Records mapped to `causal_fallacy` are excluded when filtering to the nine-class target but remain available in the normalized pre-filter corpus with their original and intermediate labels preserved in metadata.

Note on `appeal_to_worse_problems`. We group this label under `red_herring` (relevance) as it redirects attention to a different issue rather than supporting the claim; when realized as whataboutism, it functions as a relevance diversion.

Rationale (Popularity/Crowd). We fold `appeal_to_tradition` under `appeal_to_popularity` because its warrant is social acceptance via longevity ("it's always been done") rather than authority or evidence,

Family	**Child labels → Unified**
Popularity/Crowd	`ad_populum`, `appeal_to_majority`, `appeal_to_tradition` → `appeal_to_popularity`
Authority/Credibility	`fallacy_of_credibility`, `irrelevant_authority`, `appeal_to_false_authority`, `appeal_to_authority` → `appeal_to_authority`
Emotion	`appeal_to_fear`, `appeal_to_pity`, `appeal_to_anger`, `appeal_to_ridicule`, `appeal_to_positive_emotion`, `appeal_to_nature` → `appeal_to_emotion`
Relevance/Redirection	`fallacy_of_relevance`, `appeal_to_worse_problems`, `straw_man`, `fallacy_of_extension`, `red_herring` → `red_herring`
Generalization	`faulty_generalization`, `hasty_generalization` → `hasty_generalization`
Forced choice	`false_dilemma` → `false_dilemma`
Causality	`false_causality`, `causal_oversimplification` → `causal_fallacy`; `slippery_slope` kept as-is
Person-based	`guilt_by_association`, `tu_quoque`, `ad_hominem` → `ad_hominem`
Background & other	`none`/`no_fallacy`/`nothing` → `no_fallacy`; `circular_reasoning` retained in metadata (parent-only; not a training label); `intentional` excluded from 9-class training target

Table 1: Label consolidation map used to normalize heterogeneous source labels. Only records whose unified labels belong to the nine-class target, including `no_fallacy`, are retained during target filtering. Intermediate and parent-only labels are preserved for analysis and provenance.

keeping "popular acceptance ⇒ correctness" mechanisms in one class.

Target vs. parent-only. The nine-class modeling target comprises eight fallacies (ad_hominem, appeal_to_authority, appeal_to_emotion, appeal_to_popularity, false_dilemma, hasty_generalization, red_herring, slippery_slope) plus no_fallacy. Labels in the parent-only hierarchy (e.g., relevance, presumption, causality) and intermediate groupings (e.g., circular_reasoning, causal_fallacy) are used for analysis and are not members of the nine-class target. Child labels mapped to these parents are preserved in metadata.

Parents (unified → parent). *relevance:* red_herring, ad_hominem, appeal_to_authority, appeal_to_popularity, appeal_to_emotion; *presumption:* hasty_generalization, false_dilemma; *causality:* causal_fallacy, slippery_slope; *reasoning_form:* circular_reasoning; *background:* no_fallacy; *source_specific:* intentional. These parent families follow standard typologies in argumentation theory (Tindale, 2007); see also Helwe et al. (2024) for recent cross-corpus harmonization.

Source datasets. We integrate: *Argotario* (Habernal et al., 2017) (game-based; multilingual), *LOGIC_EDU* and *LOGIC_CLIMATE* (Jin et al., 2022) (curated 13-class schemes), *MAFALDA* (Helwe et al., 2024) (hierarchical spans), and *CoCoLoFa* (Yeh et al., 2024) (news comments; 8 classes). Each record follows {id, text, label, orig_label, source, domain, meta}. In LOGIC_EDU we adopt updated_label as canonical (preserving old_label in meta); MAFALDA is extracted at span level; Argotario retains consensus-voted rows.

3.1 Balancing and Stratified Splits

The merged corpus contains 12 531 normalized rows. Filtering to the nine target classes yields 11 111 usable examples. If used in downstream modeling, train/dev/test splits (80/10/10) can be produced and stratified by label_final and source; we focus here on the taxonomy and alignment.

Class	Count
no_fallacy	3248
appeal_to_popularity	1463
hasty_generalization	1265
red_herring	1163
appeal_to_emotion	1138
appeal_to_authority	836
false_dilemma	780
slippery_slope	695
ad_hominem	523

Table 2: Original class distribution (9-class target).

Source	Count (9-class)
CoCoLoFa	7,706
LOGIC_EDU	1,718
LOGIC_CLIMATE	813
MAFALDA	457
Argotario	417
Total	11,111

Table 3: Counts by source after filtering to the 9-class schema.

Domain	Count (9-class)
news_comments	7,706
education	1,718
climate	813
forum	457
game	417
Total	11,111

Table 4: Counts by domain after filtering to the 9-class schema.

3.2 Normalization Pipeline

All preprocessing is implemented in Python 3.10 using pandas and orjson. Modules: (1) per-dataset normalization to the canonical schema; (2) merge and mechanical label normalization; (3) consolidation to label_unified and parent; (4) filtering to the 9-class target; (5) integrity checks. Artifacts include mapping CSVs (child→unified, unified→parent), split manifests, and per-class count logs. All outputs are stored under data/interim/.

3.3 Rationale and Design Considerations

Normalization ensures semantic comparability across sources so that each class can be analyzed on equal conceptual and lexical footing. For lexical profiling, we compute log-odds ratios with informative Dirichlet priors (Monroe et al., 2008) over lemmatized unigrams and bigrams, identifying empirically distinctive cues per class. These cues form

the empirical lexical layer. A subset of the highest-ranked, semantically interpretable tokens is subsequently linked to WordNet 3.0 synsets (via domain-filtered mapping), creating the symbolic layer used in the taxonomy. For public tables we apply a minimal blocklist to remove non-scholarly profanity and single-instance named entities from display, while retaining the full ranked lists in the release artifacts for transparency. We compute log-odds with informative Dirichlet priors using the pooled background ($\alpha_k = 0.01 \times c_k^{\text{bg}}$; Monroe et al. 2008). Bigrams are selected with $PMI > 5$ and minimum count ≥ 5, then lemmatized. We use log-odds with informative Dirichlet priors because it is well suited for identifying class-distinctive terms under uneven class distributions, while reducing the instability of raw frequency ratios in small or imbalanced classes. Bigrams are included because several fallacy-related cues are relational or constructional rather than purely lexical, e.g., `vote_majority`, `expert_testimony`, or `easily_lead`. The extracted cues are not treated as universal markers of a fallacy; rather, they provide candidate lexical evidence that can be inspected, filtered, and linked to semantic anchors.

Methodological alternatives. Log-odds was selected over raw frequency and TF–IDF because the objective is to identify class-distinctive cues under unequal class sizes rather than document-level relevance. PMI is used only to identify recurrent bigram collocations. The head-based WordNet lookup is adopted as a transparent baseline; contextual word-sense disambiguation and embedding-based synset alignment may better address polysemy and should be compared systematically in future work.

4 WordNet Alignment

We enrich each fallacy class with two complementary semantic layers derived from the previous analysis. **(i) Conceptual anchors:** each class is linked to a primary WordNet synset representing the core reasoning concept (e.g., *false dilemma* → `dilemma.n.01`). **(ii) Lexical evidence:** we extract class-distinctive lemmatized unigrams and bigrams using log-odds with informative Dirichlet priors (Monroe et al., 2008). The ranked lists reflect corpus usage and therefore include both reasoning-related terms and corpus-specific artifacts.

Mapping policy. We map only a subset of semantically interpretable cues via automatic lookup with lexical-domain filtering, retaining `noun.cognition`, `noun.communication`, `noun.person`, `noun.act`, `noun.group`, `noun.quantity`, `noun.event`, `verb.communication`, `verb.cognition`, `verb.change` and excluding adjectives and attribute senses. For multiword cues (e.g., `vote_majority`), we perform a head-lemma lookup (*right head* by default: `majority`) and fall back to the left or the full bigram if no head match is found.

For polysemous lemmas (e.g., *authority*), we select one sense per cue by manual inspection of glosses and usage; ambiguous cases are omitted from the presented table but retained in release artifacts with an 'ambiguous' flag.

This produces a sparse candidate symbolic layer that requires filtering and human validation. In our presented run (after blocklisting table-only artifacts and excluding adjective/attribute synsets), the automatic mapping yielded 401 token–synset correspondences across the nine classes, with per-class distinct token counts: `appeal_to_authority=48`, `false_dilemma=46`, `slippery_slope=45`, `appeal_to_emotion=45`, `appeal_to_popularity=45`, `red_herring=44`, `no_fallacy=44`, `ad_hominem=42`, `hasty_generalization=42`. Counts reflect the filtered set of mappings reported here; the full mapping is released for reproducibility.

We keep the full empirical ranked lists for transparency and use the filtered synset layer for interpretability analyses. Table 5 summarizes representative cues for each class.

Preliminary quality audit of WordNet alignment. The present work does not claim a full downstream evaluation of fallacy detection in Bulgarian. Instead, we report a preliminary quality audit of the English cue-to-synset mapping used to construct the symbolic layer. We manually inspected a random sample of 50 cue→synset pairs from the automatically generated mapping. Of these, 26 (52%) were judged *acceptable* (correct or meaningfully related), 11 (22%) *borderline* (semantically plausible but overly broad/vague), and 13 (26%) *incorrect*. These results indicate that the heuristic linker is useful for generating candidate

Fallacy	Representative lexical cues (examples)
Ad Hominem	hypocrite, opponent_raise, move_town
Appeal to Authority	listen_expert, citation_expert, expert_testimony
Appeal to Emotion	mother_nature, power_line, natural_flow (*climate-domain artifacts*)
Appeal to Popularity	vote_majority, want_majority, human_tradition
False Dilemma	half_measure, meaningful_alternative, love_hate
Hasty Generalization	class_hard, college_good, philosophy_class
Red Herring	issue_poverty, cut_military, pale_comparison
Slippery Slope	easily_lead, censorship_start, lawlessness

Table 5: **Representative lexical cues.** Class-distinctive cues from the log-odds analysis (exported from `top_terms_per_class.presented.json`). Items may include corpus-specific artifacts; full ranked lists with provenance are released. Only an interpretable subset is used for WordNet alignment.

mappings, but not reliable enough to be used without filtering or human review. We therefore treat the WordNet-linked lexical layer as a bootstrapping resource rather than a finished annotation or classification system. A full evaluation will require downstream experiments, including Bulgarian seed lexicon validation, manual review of projected cues, and classification or retrieval experiments on Bulgarian argumentative texts.

The mappings should therefore not be interpreted as fully automatic fallacy annotations. Their value lies in narrowing the search space for annotation, prompting, and lexicon construction by producing inspectable candidate anchors. Human validation is part of the intended workflow: the automatic layer supports bootstrapping and prioritization, but is not sufficient as a standalone classification resource.

Cross-lingual projection. Anchored by WordNet synsets, we project to Bulgarian through Open Multilingual WordNet (OMW) (Bond and Foster, 2013) by retrieving aligned Bulgarian lemmas. We treat these lemmas strictly as seeds or weak constraints rather than inferred fallacy labels. Language-specific resources such as BulNet (Koeva, 2021) provide a complementary basis for future validation and extension. We use OMW offsets to collect aligned lemmas (not to infer labels) and treat them strictly as seeds or weak constraints (e.g., label propagation, pattern matching, lexicon-guided prompting) in low-resource settings. Table 7 illustrates how selected English WordNet anchors can be projected into Bulgarian seed cues. The projected lemmas are not used as gold labels or direct classifiers; they serve as candidate lexical material for weak supervision, lexicon-guided prompting, or manual annotation support.

Illustrative chain. *Text:* "If we allow this exception, it could **easily lead** to allowing everything."

The bigram `easily_lead` is extracted as class-distinctive lexical evidence for `slippery_slope`. At the conceptual layer, the class is associated with `consequence.n.01`, representing the anticipated outcome expressed by the argument. This is a class-level conceptual association, not a claim that the lemma `lead` is synonymous with `consequence`.

5 Discussion

The proposed two-layer taxonomy — combining conceptual WordNet anchors and lexical evidence — supports several research directions:

- **Taxonomy and resource development:** The present contribution is a taxonomy with lexical-semantic anchors, not a full ontology with axioms, formal relations, or constraints. The WordNet-linked nodes may nevertheless support future argumentation or reasoning ontologies by providing stable conceptual entry points for fallacy categories.
- **Cross-lingual mapping:** WordNet-based identifiers allow alignment with multilingual lexical resources such as BulNet (Koeva, 2021) for Bulgarian and related WordNets for other low-resource languages. This facilitates the creation of comparable reasoning datasets and seed lexicons in settings where manual fallacy annotation is limited or infeasible.
- **Interpretability:** The lexical layer clarifies how abstract reasoning flaws manifest linguistically, supporting transparent and explainable fallacy detection in manipulative or persuasive discourse.
- **Dataset integration:** The use of stable WordNet offsets (`offset--pos`) supports alignment where corresponding multilingual synset mappings are available to equivalent synsets in other languages, providing a bridge for future bilingual and multilingual reasoning corpora.

Beyond linguistic alignment, the taxonomy is designed to interface with cognitive and applied models of manipulative reasoning that distinguish between structural and content-based fallacies. By

Fallacy	Term	Synset	ID	Gloss (shortened)
Ad Hominem	hypocrite	hypocrite.n.01	10195593-n	professes beliefs not actually held
	liar	liar.n.01	10256537-n	a person who lies repeatedly
Appeal to Authority	authority	authority.n.01	05196582-n	power or right to give orders / decide
	expert	expert.n.01	09617867-n	person with special knowledge
	testimony	testimony.n.01	07184149-n	a solemn statement to establish a fact
Appeal to Emotion	pity	pity.n.01	07517737-n	a feeling of sympathy and sorrow
	fear	fear.n.01	07519253-n	an emotion experienced in anticipation of danger
Appeal to Popularity	majority	majority.n.02	13581067-n	more than half of the votes
	tradition	tradition.n.02	06685456-n	cultural continuity transmitted across generations
False Dilemma	dilemma	dilemma.n.01	05907857-n	uncertainty between two unfavorable options
	alternative	alternative.n.01	05790944-n	one of several possibilities
Hasty Generalization	generalization	generalization.n.01	05774415-n	reasoning from detailed facts to general principles
	assumption	assumption.n.02	05924488-n	a hypothesis that is taken for granted
Red Herring	diversion	diversion.n.02	00350380-n	turning aside of attention or concern
	irrelevance	irrelevance.n.01	04720284-n	lack of relation to the matter at hand
Slippery Slope	escalation	escalation.n.01	00290257-n	increase to a higher level of intensity
	consequence	consequence.n.01	11410625-n	outcome of an event or action
Background (`no_fallacy`)	balance	balance.n.01	14002279-n	state of equilibrium
	neutrality	neutrality.n.01	01240850-n	nonparticipation in a dispute

Table 6: Conservative WordNet 3.0 anchors per class after domain filtering. Shown items are semantically central, not exhaustive.

Fallacy	Synset	BG seed
appeal_to_popularity	majority.n.02	*mnozinstvo*
appeal_to_emotion	fear.n.01	*strah*
appeal_to_authority	expert.n.01	*ekspert*
slippery_slope	consequence.n.01	*sledstvie*

Table 7: Illustrative Bulgarian seed cues retrieved through OMW-aligned synsets. These are candidate cues for bootstrapping, not validated Bulgarian fallacy labels.

grounding each category in explicit WordNet concepts, these theoretical distinctions can be operationalized in computational form, facilitating integration with broader studies of argument quality and reasoning transparency.

The taxonomy further offers a structured target space for supervised and hybrid model training. Because each class includes both conceptual and lexical anchors, it supports multi-task learning setups where language models predict not only fallacy labels but also their semantic relations. This approach is particularly beneficial for low-resource languages such as Bulgarian, where OMW-based projection can provide initial seed cues for annotation, retrieval, and weak supervision, while full downstream validation remains necessary.

6 Reproducibility and Tools

All preprocessing and alignment were conducted in Python 3.10 within an isolated virtual environment. Key libraries include `spaCy 3.8.7` for tokenization and lemmatization, `NLTK 3.9` with Princeton WordNet 3.0 for English synset access, and `omw-1.4` for multilingual lemma retrieval. All operations use deterministic random seeds to ensure replicability.

The project follows a reproducible structure:

- `data/` — raw and normalized datasets
- `splits/` — stratified train/dev/test splits (not pre-balanced)
- `taxonomy_paper/` — scripts for lexical cue extraction and WordNet lookup
- `cache_json/` — hashed intermediate results
- `analysis/` — evaluation and inspection notebooks

All scripts and configuration files are implemented in Python and documented for reusability. The repository is currently private and will be released under a **CC BY 4.0** license upon publication, permitting reuse with attribution. All mapping files and scripts are versioned; random seeds are fixed for reproducibility.

7 Conclusion

We present a compact, WordNet-aligned taxonomy of logical fallacies designed as a semantic backbone for reasoning-aware NLP. The framework

unifies conceptual definitions with corpus-derived lexical cues, supporting taxonomy extension, multilingual resource development, and interpretable model evaluation. Future work will include release of the taxonomy and mappings, systematic Bulgarian seed-lexicon validation, and integration within hybrid fallacy-detection pipelines. The release includes both the parent–child hierarchy for interpretability and the 9-class training target (the eight fallacies plus `no_fallacy` as a background label) for practical modeling.

The resource is intended to interoperate with broader research on argumentative reasoning and manipulative discourse. Its WordNet-based structure enables seamless inclusion in forthcoming multilingual systems and cognitively informed studies, ensuring that linguistic, semantic, and reasoning layers evolve within a unified, transparent framework.

8 Ethics Statement and Limitations

This work analyzes reasoning structures in text, not individuals. Fallacy labels may reflect culturally specific reasoning norms; caution is required when transferring them across languages or social contexts. WordNet mappings for rhetorical phenomena are necessarily approximate, and the current manually evaluated alignment is limited to English. Future iterations will expand coverage and refine sense disambiguation. Projected lexicons may carry sense drift across languages; we recommend human review of seed lists before deployment. In particular, the Bulgarian examples in this paper demonstrate the projection mechanism, but they do not constitute a validated Bulgarian fallacy dataset or a downstream Bulgarian classifier. Their intended role is to support future annotation, retrieval, and weak supervision experiments.

This study forms part of a broader research effort on reasoning and manipulation across languages. The taxonomy was designed to remain conceptually transparent while avoiding culturally prescriptive definitions of "fallacious" reasoning (Tindale, 2007). Ongoing collaborative work will further examine how cognitive and cultural factors influence the perception and annotation of fallacies in diverse linguistic settings.

Lexical cue artifacts. The log-odds procedure can surface topic-specific or named-entity cues (e.g., items highly specific to a single example). We retain these in the released lists for transparency, but they may not generalize and should be interpreted as corpus artifacts rather than universal markers of a fallacy's form.

Acknowledgments

This research is part of the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 Programme under grant agreement no. 857155, the Programme "Research, Innovation and Digitalization for Smart Transformation" 2021–2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01, and the project BROD (Bulgarian-Romanian Observatory of Digital Media), funded by the Digital Europe Programme of the European Union under contract number 101226153.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 86–90, Montreal, Quebec, Canada.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2024. MAFALDA: A benchmark and comprehensive study of fallacy detection and classification. In *Proceedings of the 2024*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4810–4845, Mexico City, Mexico. Association for Computational Linguistics.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Svetla Koeva. 2021. The bulgarian WordNet: Structure and specific features. *Papers of the Bulgarian Academy of Sciences: Humanities and Social Sciences*, 8(1):47–70.

Yuanyuan Lei and Ruihong Huang. 2024. Boosting logical fallacy reasoning in LLMs via logical structure tree. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13157–13173, Miami, Florida, USA. Association for Computational Linguistics.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Princeton University. 2006. WordNet 3.0. `https://wordnet.princeton.edu/`. Accessed: 2025-10-20.

Zhivar Sourati, Filip Ilievski, Hông-Ân Sandlin, and Alain Mermoud. 2023. Case-based reasoning with language models for classification of logical fallacies. *arXiv preprint arXiv:2301.11879*.

Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.

Christopher W. Tindale. 2007. *Fallacies and Argument Appraisal*. Cambridge University Press, Cambridge, UK.

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59, Copenhagen, Denmark. Association for Computational Linguistics.

Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao Kenneth Huang. 2024. CoCoLoFa: A dataset of news comments with common logical fallacies written by LLM-assisted crowds. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 660–677, Miami, Florida, USA. Association for Computational Linguistics.