Metaphor Identification in Spanish Oncological Discourse: The Role of Explicit Meaning in Low-Resource Settings

Lucía Pitarch, Jordi Bernad, Gemma Bel-Enguix
Universidad de Zaragoza, Universidad Nacional Autónoma de México
Zaragoza (Spain), Ciudad de México (México)
{lpitarch, jbernad}@unizar.es, gbele@iingen.unam.mx

Abstract

Metaphor identification remains challenging in specialized and low-resource domains, where large annotated datasets are unavailable and general-domain models often fail to transfer effectively. In this paper, we evaluate FLAVORS-AECC, a Spanish dataset of oncological discourse that provides transparent, instance-level annotations of basic meaning (BM) and contextual meaning (CM) following the Metaphor Identification Procedure (MIP). We test the state-of-the-art Contrast-WSD model under two splits: a random split and a lemma-based split to control for lexical memorization. We compare three configurations: (i) a control model with no meaning information, (ii) manually curated basic meanings, and (iii) first dictionary entry as an approximation of basic meaning. Results show that explicitly modeling meaning contrast substantially improves performance in low-resource settings (from below 0.30 to above 0.50 F1). However, contrary to expectations, manually annotated BM does not consistently outperform first dictionary entries, suggesting that definition length rather than theoretical fidelity may introduce noise. We also find that models perform best on cases with high annotator agreement and that verbs remain the most challenging part of speech. Overall, our findings highlight the importance of linguistically grounded modeling for metaphor detection in specialized domains.

Keywords: Metaphor annotation, Spanish resources, oncological discourse, figurative language, low-resource NLP

1. Introduction

Metaphors are a fundamental linguistic device in medical communication, enabling speakers to explain complex or abstract experiences through more concrete domains (Lakoff and Johnson, 1980). In oncology, metaphors such as *cancer as war* or *the body as a battlefield* are pervasive and shape how patients, clinicians, and families conceptualize illness, treatment, and prognosis (Semino et al., 2017). Despite their importance, computational analysis of medical metaphors remains underexplored, particularly for languages other than English.

Traditional research on medical metaphors has largely relied on qualitative linguistic and discourse-analytic methods (Liu et al., 2024). While this work provides rich theoretical insights, it is labor-intensive, difficult to scale, and poorly suited for real-time processing of large volumes of patient narratives. Moreover, most existing annotated resources are English-centric, leaving a substantial gap for Spanish medical discourse.

In parallel, recent advances in automated metaphor identification have achieved strong performance on large general-domain datasets such as the VU Amsterdam Metaphor Corpus (Krennmayr and Steen, 2017). However, these systems struggle in low-resource, domain-specific settings where training data is limited, and language use diverges from everyday contexts.

Metaphor is a complex linguistic phenomenon approached from multiple theoretical perspectives, yielding different annotation methodologies, from explicit and deliberate metaphor frameworks (Dipper et al., 2024) to dynamic approaches such as Cameron's (Cameron, 2007). Any computational system's operationalization of metaphor is therefore inevitably tied to the annotation guidelines of its gold standard. The most widely used benchmark dataset for automated metaphor identification, the VUA corpus, was annotated following the Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007), and so have many of the systems trained and evaluated on it. MIP defines metaphoricity through a contrast between contextual meaning (CM): the sense a word takes in context, and basic meaning (BM): the most concrete, physical, and imaginable sense of a word. This contrast is what defines metaphoricity. We adopt MIP as our working framework because it underlies our gold standard, is among the most replicable and cross-linguistically documented procedures, and enables comparability across studies. A central issue concerns how BM is operationalized computationally. Many systems approximate it using the first dictionary entry or static embeddings (Choi et al., 2021), diverging from MIP's requirement that BM reflect the most concrete sense rather than the most frequent one. We acknowledge that word sense inventories are imperfect proxies, particularly for creative metaphors. Nevertheless, in low-resource, specialized domains, we hypothesize that the manual selection and annotation of BM–CM contrasts

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 31–39
11 May 2026.

provide better guidance for the model's predictions.

To test this hypothesis, we evaluate the FLAVORS-AECC dataset (Pitarch et al., 2026) using the state-of-the-art Contrast-WSD model (Elzohbi and Zhao, 2024). FLAVORS-AECC is the first Spanish dataset that provides transparent annotations of both basic at the instance level. It also includes two evaluation splits: a random split and a lemma-based split designed to reduce lexical memorization.

Our contributions are:

- The first domain-specific evaluation of automated metaphor identification in Spanish oncological discourse.
- A comparison between theoretically aligned models to annotation procedures.
- Analysis of differences between manually annotated BM and dictionary-based approximations.
- An analysis of how annotator agreement and part-of-speech affect model performance.

2. Related Work

The Metaphor Identification Procedure (MIP), introduced by the PRAGGLEJAZ Group in 2007 (Pragglejaz Group, 2007), was a major step toward standardizing metaphor annotation at a time when labeling practices were highly heterogeneous. MIP defines a four-step procedure: (1) segmenting the text into lexical units, (2) identifying the basic meaning of each unit (the most concrete, imaginable, and tangible sense listed in a dictionary), (3) determining its contextual meaning in the given sentence, and (4) marking the unit as metaphorical if the two meanings contrast.

While intentionally minimalist to allow adaptation, MIP's flexibility has also led to substantial variation and subjectivity in annotation practices. Challenges arise in defining lexical units (De Backer et al., 2023), interpreting contextual meaning, and operationalizing basic meaning (Maudslay and Teufel, 2022). Although subjectivity is unavoidable, a key issue is the lack of transparency in annotation decisions. Existing datasets often describe their segmentation criteria (Krennmayr and Steen, 2017; Sanchez-Bayona and Agerri, 2022; Sánchez-Montero et al., 2025), but rarely make explicit, for each instance, how basic and contextual meanings are interpreted. To the best of our knowledge, FLAVORS-AECC is the only dataset that explicitly encodes these meanings via WordNet synset annotations, ensuring transparency and consistency.

Regarding automated metaphor identification, most computational approaches to metaphor identification are inspired by MIP and aim to operationalize basic meaning within neural architectures. MelBERT (Choi et al., 2021) has been particularly influential, inspiring a range of subsequent models (Elzohbi and Zhao, 2024; Babieno et al., 2022; Li et al., 2023). Many of these systems attempt to operationalize basic meaning (BM) within neural architectures. Some approaches assume BM is captured by static or decontextualized embeddings (Song et al., 2021; Choi et al., 2021), while others approximate BM using the first dictionary definition of a word (Su et al., 2021; Babieno et al., 2022). However, both strategies diverge from MIP, as frequency-based meanings or first dictionary senses do not necessarily correspond to the most concrete or physical sense.

Building on these insights, our work advocates for explicit, manually curated basic meaning annotations aligned with MIP and directly provided to the model enhancing transparency, interpretability, and theoretical coherence.

3. Experimental Setup

We conduct our experiments on a filtered subset of the FLAVORS-AECC dataset (Pitarch et al., 2026). To ensure compatibility with the VUA format and the Contrast-WSD model, we retain only single-word metaphor annotations and restrict the data to verbs, nouns, and adjectives. The resulting dataset contains 5,239 instances with annotated basic and contextual meanings, of which 18% are labeled as metaphorical by at least one annotator. The overall inter-annotator agreement reported for the original dataset is 0.49 F1[1]. The part-of-speech distribution is 64% verbs, 24% nouns, and 12% adjectives. A sample of the dataset is shown in Table 1.

We evaluate all experiments under two data splitting strategies: a random and a lemma-based split. In the random split the data is divided into train and test sets using an 80/20 ratio while preserving the proportion of metaphorical and literal instances. The lemma-based split enforces that target lemmas do not overlap between train and test sets, thereby reducing lexical memorization effects.

We use Contrast-WSD (Elzohbi and Zhao, 2024), a state-of-the-art metaphor identification model inspired by the Metaphor Identification Procedure (MIP). Figure 1 presents an overview of the architecture. The model takes as input a sentence, a target word within the sentence, a definition representing the word's basic meaning, and a definition corresponding to the word's contextual meaning.

A RoBERTa model is used to obtain embeddings: a) of the target word in the full sentence with spe-

[1] F1 score was chosen against the common Kappa score, as suggested by (Boguslav and Cohen, 2017) for flexible span annotations. More details on this choice in the dataset original paper (Pitarch et al., 2026)

ID	w_index	pos	lemma	sentence	agree	label	basic meaning (BM)	contextual meaning (CM)
6_s2_w1	1	N	hermano	Mi hermano con 30 años acaba de ser diagnosticado con un tumor en el pulmón.	2	0	a male with the same parents as someone else	close friend who accompanies his buddies in their activities
6_s3_w0	0	V	tener	Tienen sospecha de que sea maligno y estamos esperando cita con Cirugía	2	0	have or hold in one's hands or grip	have or possess, either in a concrete or an abstract sense
6_s3_w5	5	ADJ	maligno	Tienen sospecha de que sea maligno y estamos esperando cita con Cirugía	2	1	maligno, malvado, maléfico, malévolo	canceroso, maligno

Table 1: AECC-FLAVORS in VUAM format processed sample. Label=1 means metaphoric example, Label=0 means non-metaphoric instance. In this case only *maligno* is annotated as metaphoric.

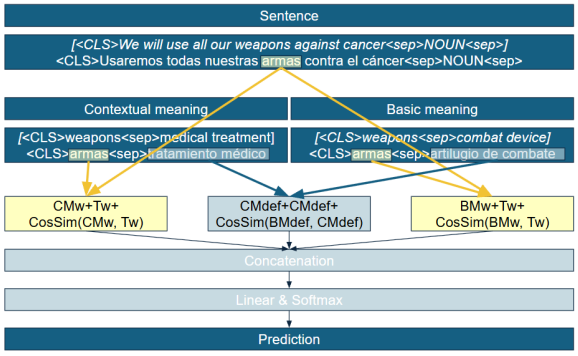

Figure 1: ContrasWSD arquitecture schema. In the original model (Elzohbi and Zhao, 2024) BM definition is the first dictionary entry. '+' signs represent concatenation.

cial tokens marking the target word and POS; b) of the target word in the contextual definition and of the contextual definition; c) of the target word in the basic meaning definition and of the basic meaning. The contextual and basic meaning embeddings are concatenated jointly with its cosine similarity. The same process (concatenate jointly with the cosine similarity) is performed with the target word embeddings in the full sentence and in the contextual meaning definition, and with the target word embeddings in the full sentence and in the basic meaning definition. These three embeddings are concatenated and fed into a final classification layer that predicts whether the target word is used metaphorically or literally.

While we use the same models as Elzohbi and Zhao (2024), we differ in our augmentation of the data. Where they use an additional external model for word sense disambiguation, we directly use Lesk (Lesk, 1986) algorithm. And secondly, while they only use the first Wikitionary dictionary entry as basic meaning, we compare the manual selection of basic meaning with the first wordnet dictionary entry. All definitions are extracted from WordNet using the corresponding synsets and are provided in Spanish when available, and in English otherwise.

We evaluate three experimental configurations:

1. **Control**: only the sentence and target word are provided, without any definitions.
2. **Manual Basic Meaning**: the basic meaning is manually annotated using the most appropriate WordNet synset.
3. **First Dictionary Entry**: the basic meaning is approximated using the first available dictionary definition.

We explore multiple hyperparameter configurations (see Appendix for full details) and select the best-performing setup with class weight[2] $= 5$, learning rate $= 1 \times 10^{-5}$, batch size $= 16$, warm-up epochs $= 2$, and total epochs $= 10$. Since definitions may be provided in both Spanish and English, we use XLM-RoBERTa-base (Conneau et al., 2019) as the encoder. Each experiment is run five times with different random seeds, and we report mean performance along with 95% confidence intervals.

4. Results

Table 2 presents the main quantitative results of our experiments for the three modeling configurations (Control, Manual Basic Meaning, and First Dictionary Entry) under both evaluation splits (random by label and lemma-based). Scores correspond to the mean over five runs, using the best-performing hyperparameter configuration.

trainingSplit	BMAugmentation	Precission	Recall	F1
lemma	Control	0.311	0.243	0.273
lemma	Control	0.308	0.280	0.292
lemma	ManualBM	0.259	0.507	0.342
lemma	ManualBM	0.270	0.526	0.357
lemma	1stDicEntry	0.277	0.508	0.360
lemma	1stDicEntry	0.296	0.541	0.383
random	1stDicEntry	0.414	0.617	0.495
random	ManualBM	0.426	0.607	0.501
random	1stDicEntry	0.455	0.568	0.505
random	ManualBM	0.462	0.580	0.514

Table 2: Results ordered by F1 (lowest to highest) to asses the best configuration of three model augmentation settings (manual BM selection, 1st dictionary entry as BM, and control: no added basic nor contextual meaning information). Both random and lemma splits are displayed.

[2]Class weight is a parameter in Elzohbi and Zhao (2024) arquitecture which weights more metaphoric instances than non metaphoric ones to balance the difference between non metaphoric sentences and metaphoric ones.

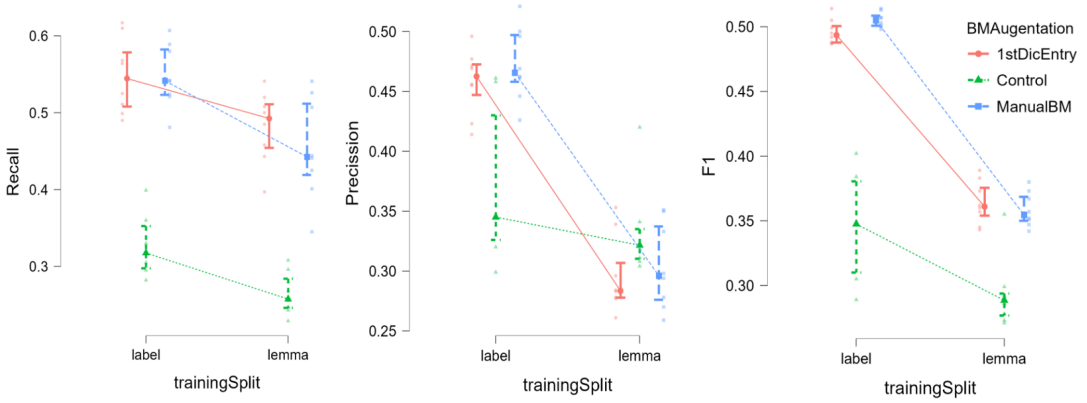

Figure 2: Performance metrics across different configurations and splits

Overall, the results reveal three central patterns that align with our research questions. First, models that explicitly incorporate information about basic and contextual meaning substantially outperform the control condition that relies solely on sentence context and the target word. Under the lemma-based split, the control model achieves only 0.27 F1, whereas the two meaning-aware configurations reach up to 0.38 F1. The effect is even more pronounced under the random split, where performance increases from below 0.30 to above 0.50 F1. This confirms our main hypothesis: in low-resource, domain-specific settings, providing linguistically grounded meaning representations is more beneficial than relying on data quantity alone.

Second, contrary to our expectations, manually annotated basic meanings do not consistently outperform the first dictionary entry approximation. Under the random split, Manual BM yields the best result (0.51 F1), but the difference with the dictionary-based approach is marginal (0.50 F1). More strikingly, under the lemma-based split, the first dictionary entry slightly outperforms the manual annotations. It is worth noting that the competitive performance of the first-sense heuristic is itself a well-established finding in Word Sense Disambiguation (McCarthy et al., 2007), where it has long served as a strong baseline. Our surprise, however, stems not from its general effectiveness but from its performance in this specific setting: MIP explicitly instructs annotators not to rely on the first dictionary entry when identifying basic meaning, emphasizing instead the most concrete, physical, and imaginable sense, which may differ from the most frequent one. The marginal gains of manual BM annotations thus suggest that theoretical fidelity to MIP's definition does not automatically translate into empirical gains in this context. One plausible explanation is that manual definitions tend to be longer and more detailed, potentially introducing noise into the model, whereas shorter dictionary definitions provide a more stable and compact semantic signal, one that, despite diverging from MIP's theoretical intent, proves empirically competitive.

Third, the best overall performance (0.51 F1) is comparable to the reported inter-annotator agreement of the dataset (0.49 F1). This suggests that the task is inherently difficult and that further improvements may require richer linguistic modeling rather than more data alone. Matching human agreement is expected in this setting; exceeding it would raise concerns about overfitting or unintended annotation leakage.

Figure 2 visualizes the relative gains across conditions, highlighting that the primary performance jump occurs when any form of basic and contextual meaning is introduced. This pattern contrasts with findings in the original Contrast-WSD study, where meaning information yielded only modest improvements on larger, general-domain datasets (0.74 vs. 0.72 F1). Our results indicate that explicit semantic modeling becomes crucial precisely in the most challenging scenarios: small datasets, specialized domains, and less-resourced languages.

4.1. Error Analysis

To better understand model behavior, we conducted a targeted error analysis focusing on annotator agreement and part-of-speech (POS) effects.

Figure 3: Error analysis in relation to annotator agreement.

Figure 3 shows performance stratified by in-

stances with full annotator agreement versus disagreement. As expected, cases where both annotators labeled a token as metaphorical are significantly easier for the model to predict. This effect is especially pronounced under the lemma-based split, suggesting that when lexical memorization is minimized, the model relies more on genuinely prototypical metaphoric patterns. Conversely, instances with annotator disagreement remain particularly challenging, indicating that these cases are ambiguous even for humans.

Figure 4 reports performance by part of speech. Adjectives emerge as the most robust category, showing relatively stable performance across splits and modeling conditions. Verbs, in contrast, are consistently the most difficult to classify. We attribute this to their greater semantic complexity: verbs encode events, relations, and argument structures that are not fully captured by simple definition-based representations. We therefore propose that future work should incorporate additional linguistic features for verbs, such as valence, semantic roles, or event structure, which may help the model better distinguish literal from metaphorical uses.

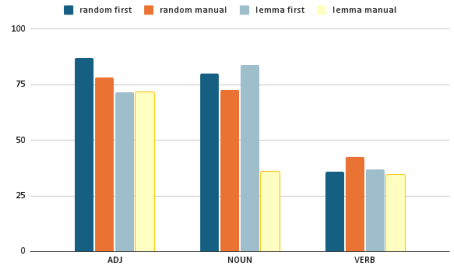

Figure 4: Error analysis in relation to POS

5. Conclusions

Taken together, these results demonstrate that (1) meaning-aware architectures are crucial in low-resource metaphor identification, (2) simpler approximations of basic meaning can be competitive with manual annotations, and (3) substantial gains will likely require more linguistically informed modeling rather than larger datasets alone.

Beyond the technical contributions, these findings carry direct implications for the medical domain that originally motivated this work. The Spanish oncological forum of the Asociación Española Contra el Cáncer was manually annotated at considerable cost — over 88,000 words, six annotators, two years of effort, and consultation across multiple disciplinary experts, yielding the gold-standard resource used throughout this study. The methods evaluated here open a path toward extending this coverage to the full forum, comprising over five million words, without incurring equivalent annotation costs. Such large-scale metaphor identification would enable the study of metaphor as a dynamic, longitudinal process: how figurative language shifts across disease phases, how patients adapt their expression over time, and what linguistic patterns may signal changes in emotional state or coping strategies. This connects directly to the agenda outlined in (Pitarch and Bel-Enguix, 2026), where metaphor is framed not as a static lexical phenomenon but as an evolving communicative resource shaped by the patient's trajectory. We therefore see the present work not as a closed contribution, but as a methodological stepping stone toward a richer, data-driven understanding of illness narratives in online health communities.

6. Acknowledgements

This paper has been supported by PAPIIT project IG-400325, by the I+D+i projects PID2024-159530OB-I00 (funded by MCIN/AEI/10.13039/501100011033), the EU research and innovation program HORIZON Europe in the "4D PICTURE" project under grant agreement 101057332 and UZ2024-IyA-02 (funded by Univ. Zaragoza), and by DGA Government predoctoral fellowship.

7. Bibliographical References

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. Miss roberta wilde: Metaphor identification using masked language model with wiktionary lexical definitions. *Applied Sciences*, 12(4).

Mayla Boguslav and Kevin Bretonnel Cohen. 2017. Inter-annotator agreement and the upper limit on machine performance: evidence from biomedical natural language processing. In *MEDINFO 2017: Precision Healthcare through Informatics*, pages 298–302. IOS Press.

Lynne Cameron. 2007. The affective discourse dynamics of metaphor clustering. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies*, (53):041–062.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Laurence De Backer, Renata Enghels, and Patrick Goethals. 2023. Metaphor analysis meets lexical strings: finetuning the metaphor identification procedure for quantitative semantic analyses. *Frontiers in Psychology*, 14:1214699.

Stefanie Dipper, Adam Roussel, Alexandra Wiemann, Won Kim, and Tra-My Nguyen. 2024. Guidelines for the annotation of deliberate linguistic metaphor. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 53–58, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Mohamad Elzohbi and Richard Zhao. 2024. ContrastWSD: Enhancing metaphor detection with word sense disambiguation following the metaphor identification procedure. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3907–3915, Torino, Italia. ELRA and ICCL.

Tina Krennmayr and Gerard Steen. 2017. *VU Amsterdam Metaphor Corpus*, pages 1053–1071. Springer Netherlands, Dordrecht.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. The University of Chicago Press.

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*, SIGDOC '86, page 24–26, New York, NY, USA. Association for Computing Machinery.

Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.

Yufeng Liu, Elena Semino, Judith Rietjens, and Sheila Payne. 2024. Cancer experience in metaphors: patients, carers, professionals, students – a scoping review. *BMJ Supportive & Palliative Care*, 14(e3):e2366–e2376.

Rowan Hall Maudslay and Simone Teufel. 2022. Metaphorical polysemy detection: Conventional metaphor meets word sense disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 65–77.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590.

Lucía Pitarch and Gemma Bel-Enguix. 2026. Modeling metaphor evolution on cancer online narratives. In *Proceedings of the 16th International Conference on the Evolution of Language (EVOLANG XVI)*.

Lucía Pitarch, Jordi Bernad, Sergio-Luis Ojeda-Trueba, Alec Sánchez-Montero, Max Ionov, Emma Anglés-Herrero, Ángel Óscar Corona Beomont, and Gemma Bel-Enguix. 2026. Medical-flavors-aecc: Spanish oncological metaphors dataset. In Press. Accepted at CL4H@LREC 2026.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra. 2025. Disagreement in metaphor annotation of Mexican Spanish science tweets. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 155–164, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Elena Semino, Zsófia Demjén, Andrew Hardie, Sheila Payne, and Paul Rayson. 2017. *Metaphor, cancer and the end of life: A corpus-based study*. Routledge.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. Verb metaphor detection via contextual relation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4240–4251, Online. Association for Computational Linguistics.

Chang Su, Kechun Wu, and Yijiang Chen. 2021. Enhanced metaphor detection via incorporation of external knowledge based on linguistic theories. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1280–1287, Online. Association for Computational Linguistics.

Appendix

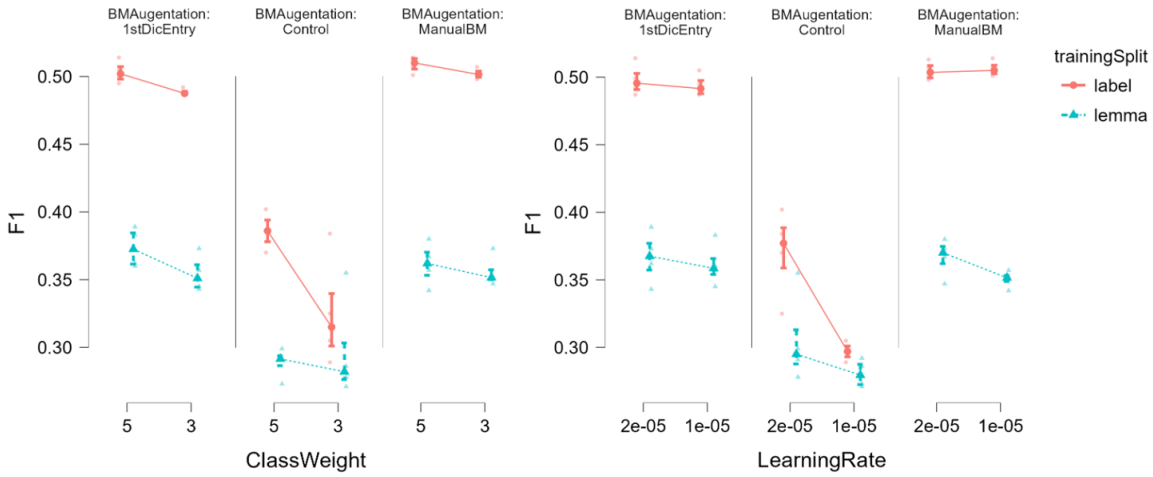

Figure 5: Learning Rate and Class Weight Plot

trainingSplit	BMAugentation	ClassWeight	LearningRate	Precission	PrecCI	Recall	RecCI	F1	F1CI
lemma	Control	3	1,00E-05	0.333	(0.290–0.375)	0.229	(0.195–0.263)	0.271	(0.236–0.305)
lemma	Control	5	1,00E-05	0.311	(0.270–0.352)	0.243	(0.208–0.278)	0.273	(0.240–0.307)
lemma	Control	3	2,00E-05	0.322	(0.283–0.361)	0.247	(0.215–0.279)	0.278	(0.245–0.312)
lemma	Control	3	1,00E-05	0.341	(0.300–0.386)	0.247	(0.215–0.279)	0.286	(0.250–0.322)
label	Control	3	1,00E-05	0.299	(0.245–0.360)	0.282	(0.230–0.339)	0.289	(0.239–0.345)
lemma	Control	5	2,00E-05	0.321	(0.283–0.364)	0.268	(0.234–0.304)	0.291	(0.257–0.327)
lemma	Control	5	1,00E-05	0.308	(0.272–0.345)	0.280	(0.245–0.313)	0.292	(0.259–0.325)
lemma	Control	5	2,00E-05	0.304	(0.266–0.340)	0.296	(0.260–0.333)	0.299	(0.266–0.332)
label	Control	3	1,00E-05	0.320	(0.260–0.376)	0.295	(0.244–0.348)	0.305	(0.249–0.353)
label	Control	3	2,00E-05	0.346	(0.303–0.391)	0.305	(0.268–0.345)	0.325	(0.288–0.361)
lemma	ManualBM	5	1,00E-05	0.259	(0.237–0.284)	0.507	(0.467–0.549)	0.342	(0.316–0.368)
lemma	1stDicEntry	3	2,00E-05	0.261	(0.237–0.289)	0.500	(0.462–0.539)	0.343	(0.314–0.370)
lemma	1stDicEntry	3	1,00E-05	0.284	(0.257–0.312)	0.443	(0.403–0.481)	0.345	(0.314–0.376)
lemma	ManualBM	3	2,00E-05	0.350	(0.314–0.386)	0.345	(0.309–0.381)	0.347	(0.314–0.383)
lemma	ManualBM	3	1,00E-05	0.298	(0.269–0.327)	0.425	(0.388–0.463)	0.351	(0.320–0.380)
lemma	ManualBM	3	1,00E-05	0.294	(0.266–0.325)	0.441	(0.403–0.481)	0.352	(0.323–0.383)
lemma	Control	3	2,00E-05	0.420	(0.323–0.526)	0.308	(0.228–0.390)	0.355	(0.272–0.435)
lemma	1stDicEntry	3	1,00E-05	0.283	(0.254–0.309)	0.485	(0.446–0.525)	0.357	(0.330–0.386)
lemma	ManualBM	5	1,00E-05	0.270	(0.245–0.296)	0.526	(0.484–0.565)	0.357	(0.329–0.384)
lemma	1stDicEntry	5	1,00E-05	0.277	(0.252–0.301)	0.508	(0.470–0.547)	0.360	(0.333–0.387)
lemma	1stDicEntry	5	2,00E-05	0.278	(0.249–0.304)	0.520	(0.481–0.556)	0.362	(0.335–0.388)
lemma	ManualBM	5	2,00E-05	0.278	(0.255–0.304)	0.541	(0.502–0.577)	0.367	(0.340–0.396)
label	Control	5	2,00E-05	0.344	(0.314–0.377)	0.399	(0.364–0.436)	0.370	(0.340–0.404)
lemma	1stDicEntry	3	2,00E-05	0.353	(0.318–0.387)	0.397	(0.358–0.433)	0.373	(0.342–0.405)
lemma	ManualBM	3	2,00E-05	0.351	(0.318–0.386)	0.401	(0.365–0.440)	0.373	(0.345–0.405)
lemma	ManualBM	5	2,00E-05	0.333	(0.300–0.367)	0.444	(0.407–0.482)	0.380	(0.349–0.412)
lemma	1stDicEntry	5	1,00E-05	0.296	(0.270–0.321)	0.541	(0.503–0.581)	0.383	(0.355–0.413)
label	Control	3	2,00E-05	0.461	(0.418–0.503)	0.330	(0.294–0.365)	0.384	(0.350–0.420)
lemma	1stDicEntry	5	2,00E-05	0.339	(0.305–0.371)	0.458	(0.420–0.495)	0.389	(0.356–0.422)
label	Control	5	2,00E-05	0.458	(0.416–0.500)	0.360	(0.327–0.394)	0.402	(0.369–0.438)
label	1stDicEntry	3	2,00E-05	0.456	(0.422–0.492)	0.525	(0.488–0.563)	0.487	(0.456–0.518)
label	1stDicEntry	3	1,00E-05	0.469	(0.435–0.503)	0.511	(0.474–0.550)	0.487	(0.455–0.521)
label	1stDicEntry	3	1,00E-05	0.477	(0.443–0.510)	0.499	(0.463–0.535)	0.488	(0.458–0.518)
label	1stDicEntry	3	2,00E-05	0.496	(0.458–0.530)	0.490	(0.452–0.528)	0.492	(0.460–0.526)
label	1stDicEntry	5	1,00E-05	0.414	(0.382–0.442)	0.617	(0.580–0.653)	0.495	(0.468–0.523)
label	ManualBM	3	2,00E-05	0.462	(0.428–0.497)	0.541	(0.504–0.580)	0.498	(0.470–0.530)
label	1stDicEntry	5	2,00E-05	0.423	(0.393–0.453)	0.610	(0.576–0.644)	0.499	(0.468–0.528)
label	ManualBM	3	2,00E-05	0.521	(0.482–0.557)	0.481	(0.443–0.521)	0.500	(0.468–0.537)
label	ManualBM	5	1,00E-05	0.426	(0.396–0.456)	0.607	(0.570–0.644)	0.501	(0.471–0.532)
label	ManualBM	3	1,00E-05	0.469	(0.435–0.502)	0.542	(0.503–0.580)	0.503	(0.471–0.535)
label	1stDicEntry	5	1,00E-05	0.455	(0.423–0.488)	0.568	(0.535–0.604)	0.505	(0.473–0.533)
label	ManualBM	3	1,00E-05	0.496	(0.458–0.533)	0.521	(0.482–0.559)	0.507	(0.474–0.538)
label	ManualBM	5	2,00E-05	0.446	(0.414–0.480)	0.589	(0.552–0.624)	0.507	(0.479–0.536)
label	ManualBM	5	2,00E-05	0.500	(0.464–0.536)	0.524	(0.486–0.562)	0.513	(0.484–0.542)
label	1stDicEntry	5	2,00E-05	0.471	(0.438–0.505)	0.564	(0.526–0.598)	0.514	(0.482–0.547)
label	ManualBM	5	1,00E-05	0.462	(0.429–0.493)	0.580	(0.543–0.617)	0.514	(0.485–0.545)

Table 3: Extended version of Table 2 including confidence intervals for each evaluation metric and the different hyperparameter configurations (class weight and learning rate). Results are ordered by F1 (lowest to highest) and display the three model augmentation settings (manual BM selection, 1st dictionary entry as BM, and control: no added basic nor contextual meaning information) under both random and lemma splits.