D-Neg: Syntax-Aware Graph Reasoning for Negation Detection

Leon Hammerla, Andy Lücking, Carolin Reinert, Alexander Mehler
Goethe University Frankfurt am Main
{hammerla,luecking,reinert,mehler}@em.uni-frankfurt.de

Abstract

Despite the communicative importance of negation, its detection remains challenging. Previous approaches perform poorly in out-of-domain scenarios, and progress outside of English has been slow due to a lack of resources and robust models. To address this gap, we present D-Neg: a syntax-aware graph reasoning model based on a transformer that incorporates syntactic embeddings by attention-gating. D-Neg uses graph attention to represent syntactic structures, emulating the effectiveness of rule-based dependency approaches for negation detection. We train D-Neg using 7 English resources and their translations into 10 languages, all aligned at the annotation level. We conduct an evaluation of all these datasets in in-domain and out-of-domain settings. Our work represents a significant advance in negation detection, enabling more effective cross-lingual research.

1 Introduction

Negation is present in all natural languages (Horn, 1989). Despite their crucial role in understanding, negative statements are often more challenging to process than affirmative ones (Dudschig et al., 2021; Tian and Breheny, 2019). This is due to their semantic and morphosyntactic complexity, which makes them difficult to comprehend (Just and Carpenter, 1971; Givón, 1978; Sho, 2007; Horn and Wansing, 2025). Large language models (LLMs) demonstrate this difficulty as they struggle with accurately interpreting negations (Truong et al., 2023; Kassner and Schütze, 2020). One reason is the underrepresentation of negation, which is rare in linguistic resources and evaluation benchmarks. This limits model exposure and training opportunities (Hossain and Blanco, 2022; Hossain et al., 2020).

NLP mostly follows the model of "standard negation", that is "the non-emphatic negation of a lexical main verb in a declarative main clause" (van der Auwera and Krasnoukhova, 2020, p. 91). It usually involves a *cue* (e.g., *not*) and a *scope*, such that the cue scopes over the proposition expressed by the affirmative declarative sentence projected from the main verb and reverses its truth value (Frege, 1892). Other kinds of negation follow this pattern, e.g., constituent negation regarding NPs (*not many people*), adjectives (*not happy*), or affixal negations (*un-happy*). Negation detection then involves two subtasks (Jiménez-Zafra et al., 2020b): cue detection and scope resolution (Szarvas et al., 2008). This approach has several issues, as the cue can be realized by several negators. An example is French *ne pas*: (1) two or more cue parts can express a single, possibly emphasized negation (*negative concord*); (2) the notion of scope becomes unclear: *do both cues have scope? Only one? Which one?* (3) A cross-linguistic, diachronic perspective is revealed, since not every language has multiple cue exponance. Its existence in a language seems to be a historic contingency known as the *Jespersen Cycle*, a potential universal of language change. It says that a pre-verbal exponent cue can develop into multiple exponents, in which case the pre-verbal part may be lost later on. The universality claim still needs to be verified by cross-linguistic empirical research. From this we can conclude that a linguistically valid negation detection ought to be 1. multiple exponent, 2. cross-linguistic, 3. diachronic. Due to the limited number of diachronic resources covering an adequate time period, we focus on the first two.

This paper focuses on cue detection and scope resolution starting from 6 corpora: (1) the *SOCC* dataset (Kolhatkar et al., 2019), consisting of user comments on online news; (2) the *BioScope* corpus (Vincze et al., 2008), containing biomedical abstracts and papers; (3) the *Conan* dataset (Morante and Daelemans, 2012), composed of sentences from two short stories by Arthur Conan Doyle; (4) the *DT-Neg* dataset (Banjade and Rus, 2016), including student-tutor dialogue interactions; (5) the *SFU* corpus (Konstantinova et al., 2012), contain-

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1432–1454
December 20-24, 2025 ©2025 Association for Computational Linguistics

ing product reviews; and (6) the *ProbBank (Focus)* dataset (Blanco and Moldovan, 2011), containing Penn Treebank sentences. The datasets use different annotation schemes and guidelines (Jiménez-Zafra et al., 2020b): different types of negation (syntactic, lexical and morphological) are annotated, with the cue sometimes included and sometimes excluded within the annotated scope. These inconsistencies, combined with diverse domain settings, led to poor cross-dataset generalization of models trained on them (Khandelwal and Sawant, 2020; Truong et al., 2022). Our goal is to enhance in-domain (ID) and out-of-domain (OOD) performance of negation detection on all datasets. We leverage syntactic features, including dependency trees, syntactic labels, and PoS tags, for fine-tuning (Khandelwal and Sawant, 2020). We expand our approach to include 11 languages, including Germanic and Romance languages, as well as Chinese, Russian and Hindi. We achieve this by translating and aligning datasets and annotations to address the limited availability of negation resources (Jiménez-Zafra et al., 2020b). Our contributions are threefold:

1. We develop a syntax-aware approach for token classification exploring dependency relations. It augments a transformer with a graph attention network (GAT) (Veličković et al., 2018), i.e., the GATv2 mechanism (Brody et al., 2022), to process dependency trees. We incorporate learned embeddings of dependency relations and PoS tags, linked by cross-attention gating.
2. We conduct experiments for 11 languages, training 14 models for cue detection and 12 models for scope resolution per language, for a total of 286 models. We evaluate the models in all possible ID and OOD settings, thereby providing a novel negation detector in a multilingual setting.
3. We publish our models and the corpus for cue and scope detection on Hugging Face[1]. We offer a Python library[2] to facilitate the application of our models across use cases.

2 Related Work

2.1 Cue Detection and Scope Resolution

Early approaches use rule-based methods for cue detection (Mutalik et al., 2001), e.g., NegEx (Chapman et al., 2001), a regular expressions-based system that remains popular in the clinical domain. For scope resolution, systems rely on constituency trees and cue position using rules for selecting subtrees manifesting the scope (Huang and Lowe, 2007; Carrillo de Albornoz et al., 2012). More recent systems incorporate syntactic features through dependency patterns into their rule system (Mehrabi et al., 2015; Sohn et al., 2012). E.g., NegBio (Peng et al., 2017) improves upon NegEx by integrating dependency structure. Although they generalize poorly to other domains, rule-based systems achieve strong performance in detecting cues in the biomedical domain.

As these methods rely on hand-crafted rules, ML approaches gained traction, which initially focused on extracting features from input sentences to perform token classification for scope resolution, employing models such as CRFs and SVMs (Read et al., 2012; Cruz et al., 2016; Ou and Patrick, 2015). Recently, there has been a focus on applying deep learning to scope resolution. A first wave of models employed recurrent neural networks, specifically long short-term memory networks and their bidirectional variant (BiLSTMs) (Fancellu et al., 2016; Lazib et al., 2016; Fancellu et al., 2017, 2018; Gautam et al., 2018). A second wave employs transfer learning. (Taylor and Harabagiu, 2018) frame cue and scope detection as a joint task using a BiLSTM to label both cue and scope tokens, while incorporating pretrained word2vec embeddings.

Starting with BERT (Devlin et al., 2019), transformer architectures led to a new SOTA for negation detection: NegBERT (Khandelwal and Sawant, 2020) frames detection as a token classification task, finetuning BERT models with classification heads for cues and scopes. To incorporate information on cues in scope resolution, (Khandelwal and Sawant, 2020) experiment with encoding strategies: replacing the cue with a special token [CUE], or augmenting the input sequence by inserting [CUE] before the cue token. (Truong et al., 2022) improve this approach by introducing negation-focused pretraining. The idea is that pretraining LLMs on domain-specific text (e.g., text rich in negation) enables more effective learning of representations of the target domain. Similarly, (Qian et al., 2024) explore data augmentation strategies to address the issue of limited annotated data for negation detection. They apply lexical masking, replacing tokens with either a generic mask token, their PoS or named entity (NE) tag. As a transformer architecture, they

[1] https://huggingface.co/D-NEG
[2] https://github.com/texttechnologylab/dneg

employ RoBERTa (Liu et al., 2019), similar to the approach of (Truong et al., 2022).

2.2 PoS Labels for Transformer Finetuning

Several studies incorporate PoS tags into fine-tuning bidirectional, encoder-based pretrained language models (PLMs). One approach is to inject PoS tags into the input sequence as additional tokens or features, allowing the model to use syntactic cues for training (Wang et al., 2023; Benamar et al., 2021). An alternative is to fuse embeddings of PoS tags with those produced by the PLM, e.g. through concatenation or gating (Liu et al., 2022). These strategies enhance the models' perception of linguistic structure, which has been shown to improve performance in downstream tasks.

2.3 Dependency-aware Finetuning

Several studies investigate the integration of graph-based information into transformers. One approach is to use Graph Neural Networks (GNNs) to incorporate syntactic or semantic dependencies. E.g., (Wu et al., 2021) integrate BERT with a Graph Convolutional Network (GCN) built over semantic dependency graphs. BERT4GCN (Xiao et al., 2021) improves upon aspect-based sentiment classification by combining BERT's intermediate-layer outputs with syntactic dependency graphs via GCNs. Regarding semantic role labeling, (Fei et al., 2021) utilize both dependency structures and their labels via GCNs to improve performance.

Due to their ability to assign dynamic weights to neighboring nodes, Graph Attention Networks (GATs) attract attention, as they allow for encoding local graph structures. Recent studies apply GATs in transformer-based fine-tuning to improve downstream tasks (Verma et al., 2024; Zhou et al., 2023). More recently, the GATv2 operator (Brody et al., 2022) was introduced to overcome the limitations of GAT, enabling more flexible and expressive attention over graphs. Though its use with PLMs is in its early stages, it improves the structural awareness of transformer architectures.

2.4 Non-English Negation Detection

The task of non-English negation detection is relatively unexplored. Progress has been made in training negation detection models, particularly for Chinese (Qian et al., 2024; Kang et al., 2017; He et al., 2017) and Spanish (Jiménez-Zafra et al., 2020a; Fabregat et al., 2019). Similar efforts have been made for French and Portuguese (Dalloux et al., 2021, 2019), and for globally significant languages, such as Arabic (Al-Khawaldeh, 2019) and Hindi (Shah and Pareek, 2024). As noted by (Jiménez-Zafra et al., 2020b), resources for non-English languages are scarce and suffer from inconsistent annotation schemas. In fact, there is an entire lack of parallel multilingual corpora for negation.

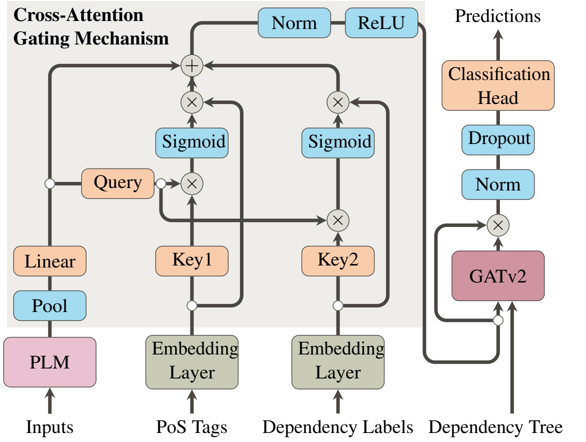

Figure 1: D-Neg's architecture highlighting its cross-attention gating mechanism and the GAT component.

We go beyond this research by introducing D-Neg: a syntax-aware graph reasoning model that incorporates syntactic information for negation detection. We train NegDep on 11 languages and create a parallel negation corpus to support cross-linguistic evaluation and resource development.

3 Methodology of D-Neg

D-Neg is based on the idea of extending a pre-trained transformer by incorporating syntactic embeddings through an attention-like gating mechanism that is conditioned on contextualized representations from the PLM. We expand on this architecture by using graph-based reasoning to assist with fine-grained, token-level classifications, such as cue detection and scope resolution. Our model is implemented using PyTorch (Paszke et al., 2019), PyTorch Geometric (Fey and Lenssen, 2019), and the Huggingface library (Wolf et al., 2020).

3.1 Model Architecture

Transformer Backbone We use DeBERTa (He et al., 2021) as an encoder for English and EuroBERT (Boizard et al., 2025) for all other languages. Both PLMs consistently outperform earlier models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), commonly used in architectures for negation detection (Khandelwal and

Sawant, 2020; Truong et al., 2022), across a range of benchmark tasks (e.g., GLUE, SQuAD, and SuperGLUE). Given a token sequence $\mathbf{x}$ of padded length n, the encoder outputs contextualized subword embeddings $\mathbf{H}_{\text{sub}} \in \mathbb{R}^{n \times d'}$, where d' is the hidden size of the encoder. Since the PLMs operate at the subword level, we apply a pooling operation to the subword embeddings corresponding to each word, resulting in word-level representations $\mathbf{H}_{\text{word}} \in \mathbb{R}^{n \times d'}$. Following (Devlin et al., 2019), we use the embedding of the first subword token to represent the entire word. We project the word-level embeddings into a lower-dimensional space using a learned linear transformation to align them with the syntactic feature embedding space:

$$\mathbf{H} = \mathbf{W}_{\text{proj}} \mathbf{H}_{\text{word}}, \quad \mathbf{H} \in \mathbb{R}^{n \times d},$$

where $d < d'$ is the shared hidden size used in subsequent components of the model, including the syntactic fusion and classification layers.

Syntactic Embeddings To incorporate syntactic information, we learn task-specific embeddings for PoS tags and dependency relation labels:

$$\mathbf{P} = \text{Emb}(p_{seq}), \quad \mathbf{D} = \text{Emb}(d_{seq}),$$

p_{seq} and d_{seq} are sequences of *P*oS and *d*ependency labels; $\mathbf{P}, \mathbf{D} \in \mathbb{R}^{n \times d}$ are the embedding matrices learned using an embedding layer $\text{Emb}(\cdot)$.

Cross-Attention like Gating Inspired by the attention gating mechanism of (Oktay et al., 2018), we design a cross-attention-like gating module that selectively integrates syntactic information into the contextualized embeddings $\mathbf{H}$. Thus, $\mathbf{H}$ is projected to queries $\mathbf{Q}$, while the PoS and dependency embeddings $\mathbf{P}$ and $\mathbf{D}$ are mapped to keys $\mathbf{K}_P$ and $\mathbf{K}_D$:

$$\mathbf{Q} = \mathbf{W}_Q \mathbf{H}, \quad \mathbf{K}_P = \mathbf{W}_{K_P} \mathbf{P}, \quad \mathbf{K}_D = \mathbf{W}_{K_D} \mathbf{D},$$

$\mathbf{W}_Q$, $\mathbf{W}_{K_P}$, and $\mathbf{W}_{K_D}$ are learnable projection matrices. Unlike the standard scaled dot-product, we compute attention-like scores by applying element-wise multiplication between the query vectors and each set of key vectors. Then, a sigmoid activation is applied to produce feature-based gating signals instead of token-wise attention scores:

$$\mathbf{g}_P = \sigma(\mathbf{Q} \odot \mathbf{K}_P), \quad \mathbf{g}_D = \sigma(\mathbf{Q} \odot \mathbf{K}_D).$$

where $\sigma(\cdot)$ is the sigmoid function. Next, we apply the gating signals to the syntactic label embeddings, modulating their contribution and producing contextually relevant representations. These gated embeddings are subsequently fused with the main embedding by element-wise addition:

$$\mathbf{H}' = \text{ReLU}(\text{LayerNorm}(\mathbf{H} + g_P \odot \mathbf{P} + g_D \odot \mathbf{D}))$$

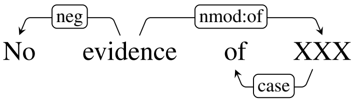

Figure 2: Dependency graph for "no evidence of XXX".

Graph Reasoning with GATv2. To account for structural token dependencies, we use a GATv2 layer (Brody et al., 2022) on the word embeddings $\mathbf{H}'$, using the dependency tree as the input graph. Residual connections are included to further stabilize training. The GATv2 operator performs multi-head attention over adjacent nodes in the graph:

$$\mathbf{H}_{\text{GAT}} = \text{GATv2Conv}(\mathbf{H}', \mathscr{E}) + \mathbf{H}',$$

$\mathscr{E}$ is the edge set of the dependency tree.

Classification Layer. The output of GATv2 is passed through layer normalization and dropout, followed by a linear classification head $\mathbf{W}_c$:

$$\mathbf{H}'_{\text{GAT}} = \text{Dropout}(\text{LayerNorm}(\mathbf{H}_{\text{GAT}}))$$
$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_c \mathbf{H}'_{\text{GAT}})$$

With $\mathbf{H}'_{\text{GAT}}$ D-Neg implements a negation model that uses transformer-based embeddings to feed a graph reasoner, combining hierarchical and sequence information at the level of lexical, PoS, and syntactic features: while positional information maps the linearity of token sequences, the graph reasoner provides hierarchical information that correlates with context-free languages. In doing so, D-Neg approximates rule-based approaches that operate over dependency structures, such as those in NegBio (Peng et al., 2017), which detect negation by traversing dependency paths and leveraging lemma-based pattern matching. E.g., the phrase "no evidence of XXX" is captured by the rule:

$$\{\} < \{\text{dep}_{\texttt{nmod:of}}\}\{\text{lemma}_{\texttt{evidence}}\} > \{\text{dep}_{\texttt{neg}}\}\{\},$$

$\{\}$ is a wildcard matching any lemma (s. Figure 2). D-Neg learns to generalize such patterns from data, using sequence representations as a fallback when syntactic structure is ambiguous or unavailable.

3.2 Training and Task Setup

D-Neg uses a two-stage approach for cue detection and scope resolution, similar to NegBERT (Khandelwal and Sawant, 2020), where both tasks are framed as token-level classification problems. For cue detection, we classify each input token as "cue" or "non-cue". Unlike (Khandelwal and Sawant, 2020), we do not distinguish between types of cues. We use the same strategy for scope resolution, performing it as a binary token classification in which each token is classified as "scope" or "non-scope." We opt for the replacement strategy of (Khandelwal and Sawant, 2020) because it aligns better with our word-level approach and since experiments using this strategy produced promising results. In the replace method, the cue token (or multiple tokens, if the cue is multi-token) is replaced by [CUE]. This allows the model to identify the position of the target cue. If a sentence contains multiple negations, [CUE] helps the model determine the scope of the current target cue. For training on languages other than English, we adopt the CLAT approach (Schäfer et al., 2022), which combines machine translation with annotation projection. It relies on word alignments generated by SimAlign (Jalili Sabet et al., 2020) to transfer annotations from the source to the target language. For translation, we use the models of the Helsinki OPUS project (Tiedemann and Thottingal, 2020).

4 Experiments

For cue detection, we train on *SOCC*, *BioScope (Full)*, *BioScope (Abstracts)*, *Conan*, *DT-Neg*, *SFU*, and *ProbBank (Focus)*. For scope resolution, we exclude *ProbBank (Focus)* as it lacks scope annotations. To ensure cross-dataset consistency, all datasets are augmented so that the cue is consistently removed from the annotated scope. This preprocessing step harmonizes differences in annotation guidelines across corpora. All models are trained for six epochs using the AdamW optimizer (learning rate: 2e−5, weight decay: 0.01) with cross-entropy loss. The checkpoint with the highest validation F1 score is selected as the final model. We evaluate the resulting models on both the ID and OOD test sets. Each dataset is split into 80% training, 10% validation, and 10% test portions. We repeat this procedure for each target language and apply it to both our proposed method and the NegBERT baseline. For fair comparison, we replace the original BERT backbone in NegBERT with a DeBERTa encoder for English and EuroBERT for all other languages, following the same backbone configuration used for training D-Neg. All backbone models use their base configurations: DeBERTa-V3 Base ($\sim$ 184M parameters) and EuroBERT Base ($\sim$ 210M parameters). To extract syntactic information, we rely on the small models from the spaCy (Honnibal et al., 2020) family (s. Appendix A). For languages not supported by spaCy (Hindi and Arabic), we employ corresponding models from the UDPipe (Straka et al., 2016) library. For dataset translation, we use bilingual OPUS-MT models from the Helsinki-NLP project (Tiedemann and Thottingal, 2020), each based on a standard Transformer architecture with six self-attention layers in both the encoder and decoder, totaling approximately 75M parameters per model (s. Appendix B). For dataset alignment, we adopt the SimAlign (Jalili Sabet et al., 2020) library, using its IterMax matching strategy and the default multilingual BERT-base-cased model for embedding similarity computation. Finally, we conduct a small-scale in-context learning (ICL) few-shot experiment with a compact contemporary LLM to illustrate the necessity of fine-tuning PLMs for negation detection.[3] For this experiment, we use Gemma3-1B (GemmaTeam et al., 2025), which is comparable in order of magnitude to our PLM backbones, though still roughly five times larger.

5 Results

The results of our English experiments are presented in Tables 1 and Table 2. In most cases, we observe a significant improvement in macro F1 scores for both ID and OOD performance, particularly in scope resolution. For cue detection, there is a slight decrease in ID performance on the *ProbBank (Focus)* and *SOCC* datasets. However, the overall trend shows an average improvement of 1.03 absolute percentage points for ID performance and 1.26 percentage points for OOD performance (s. Table 3), which corresponds to 21% ID and 10% OOD error reduction, demonstrating a meaningful leap in an already high-performance regime. For scope detection, the only notable drop in performance occurs on the *BioScope (Full)* corpus. The overall trend shows a substantial average improvement of 4.17 absolute percentage points for ID performance and 4.90 percentage points for OOD performance (s. Table 3). The same result, viewed through the lens of error reduction, corresponds to 47% ID and 30% OOD error reduction. The strongest model for cue detection is the one trained with D-Neg on the *SOCC* dataset, achieving an average F1 score of 92.97 across all test sets. In contrast, the overall average F1 score across all cue

[3] The few-shot prompts for cue and scope detection are provided in Appendix C.

	Neg-Bert (Cue)							D-Neg (Cue)							G3:1b
	DT-Neg	Conan	Bioscope (A)	Bioscope (F)	Prob-Bank (Focus)	SOCC	SFU	DT-Neg	Conan	Bioscope (A)	Bioscope (F)	Prob-Bank (Focus)	SOCC	SFU	
DT-Neg	97.26	82.64	77.85	77.11	93.13	92.44	92.14	97.59	83.39	92.65	82.29	93.22	91.21	93.69	49.26
Conan	83.85	96.08	79.72	81.45	80.82	90.9	91.35	85.49	97.88	82.19	88.86	84.64	91.01	91.52	56.18
Bioscope (A)	88.85	86.56	93.29	93.26	88.1	91.37	91.61	91.0	83.77	93.45	92.42	88.28	92.09	91.03	50.26
Bioscope (F)	85.33	84.16	93.21	87.44	86.78	91.44	89.27	85.01	83.47	91.39	91.51	84.13	90.99	90.6	49.83
Prob-Bank	94.12	94.0	90.94	90.49	99.72	95.45	95.35	93.68	92.9	92.83	94.55	98.95	95.45	95.49	52.36
SOCC	82.05	83.86	73.71	71.94	80.37	98.91	92.63	78.04	84.22	74.91	74.91	80.73	98.48	89.57	49.48
SFU	78.01	86.92	73.98	71.67	80.08	88.65	93.51	81.36	87.89	77.01	78.87	83.85	91.58	95.57	49.68

Table 1: **Cue** Detection Results for **EN**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	Neg-Bert (Scope)						D-Neg (Scope)						G3:1b
	DT-Neg	Conan	Bioscope (A)	Bioscope (F)	SOCC	SFU	DT-Neg	Conan	Bioscope (A)	Bioscope (F)	SOCC	SFU	
DT-Neg	90.96	85.09	82.99	85.79	81.23	81.89	97.42	93.57	84.43	87.53	81.04	83.33	52.93
Conan	90.11	79.83	83.76	84.47	83.5	84.33	91.07	95.18	86.86	85.04	86.71	81.49	49.14
Bioscope (A)	77.04	55.85	96.69	95.65	93.33	89.96	87.57	87.91	97.67	94.49	93.94	89.24	49.54
Bioscope (F)	76.28	61.08	94.74	94.89	90.09	87.68	84.36	86.98	93.71	93.76	89.58	89.4	51.02
SOCC	73.98	74.11	91.52	89.29	93.8	89.14	79.32	86.92	91.93	89.39	96.76	87.47	49.71
SFU	76.0	69.04	81.15	81.58	81.93	90.75	82.84	81.99	89.19	88.05	88.88	91.18	52.85

Table 2: **Scope** Detection Results for **EN**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	in-domain results			out-of-domain results		
	Neg-Bert	D-Neg	Diff.	Neg-Bert	D-Neg	Diff.
Cue	95.17	96.20	1.03	87.42	88.69	1.26
Scope	91.15	95.33	4.18	83.88	88.78	4.91

Table 3: Aggregated results for **EN**: average F1 score of models on in-domain and out-of-domain datasets.

detection models is significantly lower, at 88.05. For scope resolution, the best-performing model also results from training with D-Neg on the *BioScope (Abstracts)* dataset, reaching an average F1 score of 90.63, compared to the average of 86.33 across all scope resolution models. The ICL few-shot experiments for both cue and scope detection on the English datasets yielded substantially lower results compared to the PLM-based baselines. Cue detection performance peaked on the Conan dataset with a macro-F1 score of 56.18, while scope detection performed best on the DT-Neg dataset with 52.93. On average, the ICL approach achieved a macro-F1 slightly above 50 for both tasks, in contrast to the PLM-based models, which consistently scored above 80. Moreover, ICL inference was considerably more time-consuming than the training and inference of D-Neg. For reference, D-Neg was trained on the SFU corpus (the largest dataset) for six epochs with a batch size of 16, taking 1980s for cue detection and 2143s for scope detection. Average inference times on the SFU test sets were 21s (cue) and 23s (scope). In contrast, Gemma3-1B required 3180s per cue test set and 43,532s per scope test set, measured on the same GPU, and also consumed significantly more memory due to its larger model size. Figure 3 shows our results for the multilingual setting, with per-language F-Scores provided in the Appendix. For cue detection, the aggregated results indicate substantial performance improvements with D-Neg over NegBERT for French, Italian, Spanish, and Russian. An exception is Hindi, where D-Neg exhibits a significant performance drop. For scope resolution, performance differences between D-Neg and NegBERT are less pronounced. Nevertheless, D-Neg achieves measurable gains for German, Japanese, Dutch, and Arabic. In contrast, we observe performance declines for Italian, French, Spanish, and Chinese. Overall, most models trained on the translated and aligned datasets achieve strong performance, with F-scores above 0.8 for cue and scope detection. However, models for Chinese and Japanese exhibit significantly lower performance (s. Figure 3).

5.1 Discussion

In our experiment on English, D-Neg significantly improves scope resolution in both ID and OOD settings compared to NegBERT serving as a baseline. This suggests that D-Neg more effectively captures linguistic structures that generalize across datasets. In contrast, improvements in cue detection are more modest and dataset-dependent. Notably, the datasets showing slight performance drops in cue detection already achieve very high

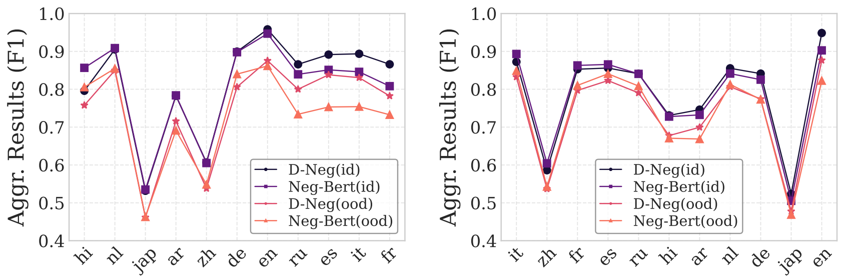

Figure 3: Aggregated results across all eleven languages. Left: cue detection, right: scope resolution.

Task	Dataset	Neg-Bert*	CueNB	SN	D-Neg
Cue	Conan	92.94	91.17	-	97.88
	BioScope (Abstracts)	95.65	97.09	95.98	93.45
	BioScope (Full)	92.42	91.67	91.89	91.51
	SFU	87.08	87.99	82.42	95.57
Scope	Conan	92.36	91.24	-	95.18
	BioScope (Abstracts)	95.68	95.81	96.22	97.67
	BioScope (Full)	91.24	92.57	93.52	93.76
	SFU	90.95	91.03	92.42	91.18

*Scores from the original NegBERT model as presented in its paper, not from our retrained DeBERTa-based variant.

Table 4: Negation detection results of transformer-based approaches for different studies reported for NegBERT (Khandelwal and Sawant, 2020), CueNB (Truong et al., 2022), and SN (Khandelwal and Britto, 2020). Training/test splits and cue definitions may vary across works.

scores (above 98%), indicating that these variations may be due to random fluctuations, annotation inconsistencies, or the limited margin for further improvement. We compare D-Neg with key transformer-based approaches for cue and scope detection, that is, NegBERT (Khandelwal and Sawant, 2020), CueNB (Truong et al., 2022), and the multitask model SN (Khandelwal and Britto, 2020). We exclude models such as (Qian et al., 2024) due to differences in classification schemes and F1 score calculation, which render a direct comparison unreliable. Furthermore, we note that the split between training and testing may vary across studies and that our method does not distinguish between different types of cues. Therefore, the comparisons provided are for general guidance only and should not be interpreted as strict performance ratings. Despite these limitations, our model outperforms all other methods in cue detection on the *Conan* and *SFU* datasets and in scope resolution on *Conan*, *BioScope (Full)*, and *BioScope (Abstracts)* (s. Table 4). The remaining datasets used in our study, namely *DT-Neg*, *SOCC*, and *ProbBank (Focus)*, are rarely employed in related work and are therefore excluded from this comparison. While these results are not strictly comparable due to potential differences in experimental setups, they nonetheless suggest that our approach is highly competitive. To further explore alternative approaches, we also conducted an ICL few-shot experiment using a compact contemporary LLM (Gemma3-1B). The substantially lower performance of this approach, combined with its markedly higher computational cost and runtime, highlights the continued need for specialized systems built on PLM backbones in the era of general-purpose LLMs, thereby further reinforcing the motivation for our approach. In the multilingual setting, D-Neg's performance gains are more moderate: improvements are observed for a subset of languages and tasks. An interesting pattern emerges: gains in cue detection occur for Romance languages (IT, FR, and ES), while the same languages exhibit the largest performance drops in scope resolution. This suggests that, for Romance languages, negation cue detection benefits significantly from syntactic patterns and predictable dependency attachments. In contrast, negation scope boundaries appear to be more diffuse and less effectively captured by dependency-based features. In any case, with D-Neg, we provide a negation detector for 11 languages, making certain language comparisons possible for the first time.

6 Error Analysis

Our approach is so effective that it is difficult to improve upon. In the ID scenario, we achieve average F1 scores above 95% for cue detection and scope resolution. In the OOD setting, average performance remains above 88% (s. Table 3). The models are thus already highly effective. We aim to explore the remaining error margins through the lens of our syntactic framework to detect the scenarios in which the models still struggle.

6.1 Cue Detection

For cue detection, we split the positive and negative classes into positive and negative subclasses for each PoS tag and dependency label. This allows for a more detailed analysis of scenarios where the

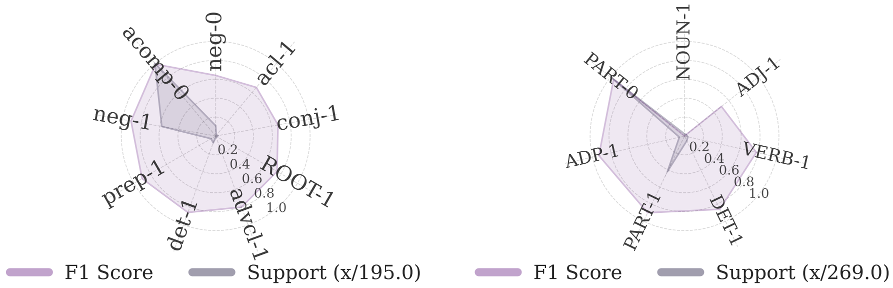

Figure 4: Left: Dependency labels, right: PoS tags. F1 scores and sample support for each class (where F1 score < 1.0) across different linguistic features.

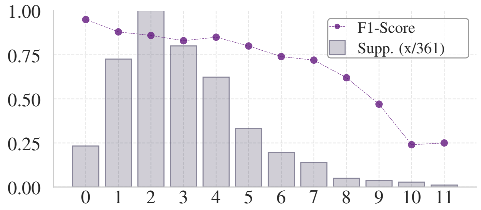

Figure 5: Positive class-specific F1 scores stratified by the depth (x-axis) of scope tokens in dependency tree.

model performs poorly. We conduct this analysis on the BioScope (Abstract) corpus, where our model showed comparatively weaker performance and which is of reasonable size for a focused evaluation. In the PoS-based evaluation, we observe that only seven subclasses have an F1 score below 1.0. Notably, the model performs poorly when the annotated negation cue is a determiner, adjective, noun, or verb. However, these cases represent only a very small fraction of the data, suggesting that their limited presence in the training set likely contributes to the lower performance. Due to their low frequency, they have minimal impact on the overall performance drop (s. Figure 4).

The dependency-based evaluation offers more insights than the PoS-based one. Although 14 classes fall below the threshold of F1=1, only three have substantial support in the training and test datasets making their results more reliable and meaningful. The model performs very well on the two most well-supported classes: it reliably detects cues when the word functions as a negation modifier, and identifies non-cues when the word functions as an adjectival complement. However, it struggles in cases where the word has the syntactic function of a negation modifier but is not a cue. Though this class has the third-highest sample count among those with F1 < 1, the model only achieves a class-specific F1 score of 0.64 in this scenario (s. Figure 4).

6.2 Scope Resolution

Regarding scope resolution, Fancellu et al. (2017) note that models trained to predict scope often rely on surface-level patterns, particularly punctuation, since many scopes are delimited thereby. Khandelwal and Sawant (2020) conduct an in-depth error analysis and demonstrate that this claim does not hold for transformer-based approaches. They show that models like theirs can learn scope boundaries beyond simple surface cues, and that performance drops are likely due to the underrepresentation of edge cases in corpora such as *SFU*. Our goal is not to replicate these studies, but to test the hypothesis that scope tokens become more difficult to predict as their surrounding dependency structure becomes more complex. To this end, we use the depth of a token in the dependency tree to approximate structural complexity. Similar to our cue analysis, we split F1 scores according to the depth of the tokens in the dependency tree. When focusing on positive classes, we observe a strong correlation between greater depth and lower class-specific F1 scores (s. Figure 5). This suggests that prediction errors are more likely to occur for scope tokens deeper in the tree, making these tokens significantly more difficult for the model to detect correctly.

6.3 Multilingual Experiment

In the multilingual setting, we identify three primary sources of error. First, translation quality varies substantially across language pairs. Translations from English into typologically distant languages such as Chinese, Japanese, and Arabic are particularly prone to errors, including untranslated segments, placeholder characters, or literal word-by-word output that fails to preserve meaning. Second, the reliability of dependency and PoS parsers differs across languages. Since D-Neg relies on syntactic features, lower parsing quality directly degrades its performance. Third, inaccuracies in source–target word alignment can result in anno-

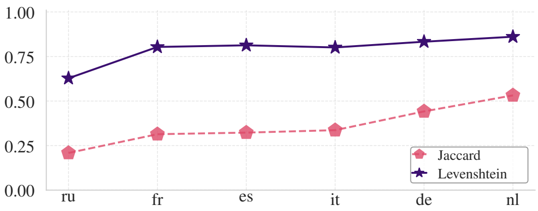

Figure 6: Jaccard for alignment, Levenshtein for translation quality (Language on the x-axis).

tations being incorrectly transferred or omitted. This issue is especially pronounced in Chinese and Japanese, where many examples lack usable transferred annotations. As a result, D-Neg tends to default to predicting the negative class, yielding macro F-scores around 0.5 (with high scores for the negative class and near-zero scores for the positive class).[4] Errors originating at these stages propagate throughout the system and substantially reduce downstream performance. We also analyze the role of negative concord (s. Section 1) in translation and alignment. We observe a decline in performance when moving from non-concord languages (e.g., German, Dutch) to weak concord languages (e.g., Spanish, Italian, French), and a further drop for Russian, which exhibits strict concord (s. Figure 6). Our evaluation is conducted on the BioScope (Abstract) corpus using round-trip translation and alignment. We measure alignment quality using Jaccard similarity between forward and reverse alignment pairs, and translation quality using Levenshtein distance. We additionally performed a small-scale human evaluation. Human annotators assessed both alignment and translation quality from the original English into each target language[5]. For each language, we randomly sampled 100 examples. Translations received a score of 1 if they were intelligible and alignable with the source (allowing minor grammatical errors), and a score of 0 if the translation was grammatically incorrect or semantically incoherent. The human evaluation confirms that translation quality decreases for languages that are typologically more distant from English. German translations performed best, followed by the Romance languages, while Arabic and Japanese showed the lowest quality; they were the only languages with a satisfactory rate below 0.8. Arabic translations often failed to convey full contextual meaning, while Japanese outputs hallucinated unrelated content, making reliable alignment evaluation infeasible. The alignment evaluation yields results consistent with the automatic evaluation: we again observe negative concord effects. Italian and French perform worse than German, and Russian shows a further decline. Spanish alignment performed unexpectedly well, with near-perfect Kappa scores (0.95 and 0.97). Aside from Japanese, Chinese was the only language with cue and scope agreement below 0.5, due to suboptimal tokenization (s. Figure 7).

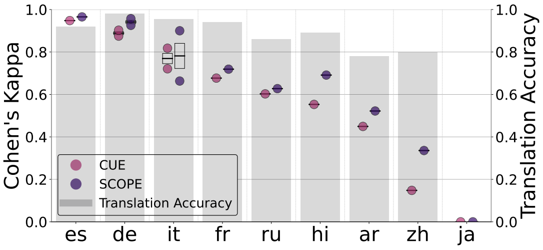

Figure 7: Pairwise (Auto–Human) agreement (Cohen's Kappa) for the alignment of negation cues and scopes from English datasets to other languages, along with human-evaluated automatic translation accuracy. For German and Italian, we report results from two human annotators and present their average.

7 Conclusion

We presented D-Neg to address the shortage of negation detection resources and models for languages other than English. D-Neg combines a transformer (using PoS and dependency features, as well as syntax label embeddings gated by attention-like mechanisms) with a graph reasoning unit informed by dependency structure. D-Neg emulates rule-based systems consisting of manually combined syntactic constraints and lexical patterns. When evaluated against NegBERT on English data and their annotation-aligned translations, D-Neg shows strong performance in both ID and OOD settings. In high-performing English setups, D-Neg reduces ID cue detection error by up to 21% and ID scope resolution error by up to 47%. Our multilingual experiments reveal that D-Neg adapts differently across languages. The most notable improvement is in cue detection for Romance languages, while the biggest challenge lies in resolving the scope of these languages. These results highlight the importance of incorporating syntactic signals, and demonstrate D-Neg's potential for detecting and studying negation across languages.

[4] For datasets that contain no positive samples, F1 is 1.

[5] Dutch was excluded from the human evaluation due to the unavailability of a qualified annotator.

Limitations

An alleged limitation of D-Neg is that it implements a syntactic notion of scope, whereas one would actually like to have a semantic notion. For instance, the negator *no* in (1) is syntactically connected to the noun *good* (indicated by brackets), but semantically takes scope over the entire embedded modal clause (indicated by square brackets) (McKenna and Steedman, 2020, p. 137). If this is right, one would expect a negation cue and scope detector to capture this.

(1) I thought no [(good) would come from it].

However, the notion of semantic scope rests on the implicit assumption that negation is an operator that enters into syntactic movements on so-called logical form (Reinhart, 1997). But this is just an artifact of a particular formal approach; it is not required on type shifting approaches (Partee and Rooth, 1983) or on recent non-GQT theories (Lücking and Ginzburg, 2022).

Acknowledgements

We thank Mercedes Martínez Bruera (Spanish/Italian), Tommaso Mattiuzzi (Italian), Svetlana Mnogogreshnova (Russian), Bhuvanesh Verma (Hindi), an anonymous contributor (Arabic), Julius Schwarzwälder (French), Longwei Cong (Chinese), and Patryk Czerwiński (Japanese) for their valuable contributions to the human evaluation of our multilingual dataset. This work was supported by the German Research Foundation (DFG) as part of Project INF within CRC 1679 "Negation in Language and Beyond" (NegLaB[6]).

References

2007. *Bibliography*, page 382–410. Cambridge University Press.

Fatima T Al-Khawaldeh. 2019. Speculation and negation detection for arabic biomedical texts. *World of Computer Science and Information Technology Journal (WCSIT)*, 9(3):12–16.

Rajendra Banjade and Vasile Rus. 2016. DT-neg: Tutorial dialogues annotated for negation scope and focus in context. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3768–3771, Portorož, Slovenia. European Language Resources Association (ELRA).

Alexandra Benamar, Meryl Bothua, Cyril Grouin, and Anne Vilnat. 2021. Easy-to-use combination of pos and bert model for domain-specific and misspelled terms. In *NL4IA Workshop Proceedings*.

Eduardo Blanco and Dan Moldovan. 2011. Semantic representation of negation using focus detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–589, Portland, Oregon, USA. Association for Computational Linguistics.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. Eurobert: Scaling multilingual encoders for european languages.

Shaked Brody, Uri Alon, and Eran Yahav. 2022. How attentive are graph attention networks?

Jorge Carrillo de Albornoz, Laura Plaza, Alberto Díaz, and Miguel Ballesteros. 2012. UCM-I: A rule-based syntactic approach for resolving the scope of negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 282–287, Montréal, Canada. Association for Computational Linguistics.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Noa P. Cruz, Maite Taboada, and Ruslan Mitkov. 2016. A machine-learning approach to negation and speculation detection for sentiment analysis. *Journal of the Association for Information Science and Technology*, 67(9):2118–2136.

Clément Dalloux, Vincent Claveau, and Natalia Grabar. 2019. Speculation and negation detection in french biomedical corpora. In *RANLP 2019 - Recent Advances in Natural Language Processing*, pages 1–10, Varna, Bulgaria.

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Emanuel Silva Oliveira, Claudia Maria Cabral Moro, Yohan Bonescki Gumiel, and Deborah Ribeiro Carvalho. 2021. Supervised learning for the detection of negation and of its scope in french and brazilian portuguese biomedical corpora. *Natural Language Engineering*, 27(2):181–201.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

[6] https://www.neglab.de/

bidirectional transformers for language understanding.

Carolin Dudschig, Barbara Kaup, Mingya Liu, and Juliane Schwab. 2021. The processing of negation and polarity: An overview. *Journal of psycholinguistic research*, 50(6):1199–1213.

Hermenegildo Fabregat, Andrés Duque, Juan Martinez-Romo, and Lourdes Araujo. 2019. Extending a deep learning approach for negation cues detection in spanish. In *IberLEF@ SEPLN*, pages 369–377.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2016. Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 495–504, Berlin, Germany. Association for Computational Linguistics.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2018. Neural networks for cross-lingual negation scope detection.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. Detecting negation scope is easy, except when it isn't. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 58–63, Valencia, Spain. Association for Computational Linguistics.

Hao Fei, Fei Li, Bobo Li, and Donghong Ji. 2021. Encoder-decoder based unified semantic role labeling with label-aware syntax. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12794–12802.

Matthias Fey and Jan E. Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Gottlob Frege. 1892. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Dipesh Gautam, Nabin Maharjan, Rajendra Banjade, Lasang Jimba Tamang, and Vasile Rus. 2018. Long short term memory based models for negation handling in tutorial dialogues. In *FLAIRS*, pages 14–19.

GemmaTeam, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Thomas Givón. 1978. Negation in language: Pragmatics, function, ontology. *On understanding grammar*, 18:59–116.

Hangfeng He, Federico Fancellu, and Bonnie Webber. 2017. Neural networks for negation cue detection in Chinese. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages

59–63, Valencia, Spain. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Laurence R. Horn. 1989. *A natural history of negation*. University of Chicago Press, Chicago.

Laurence R. Horn and Heinrich Wansing. 2025. Negation. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2025 edition. Metaphysics Research Lab, Stanford University.

Md Mosharaf Hossain and Eduardo Blanco. 2022. Leveraging affirmative interpretations from negation improves natural language understanding.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Yang Huang and Henry J. Lowe. 2007. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, 14(3):304–311.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Salud María Jiménez-Zafra, Roser Morante, Eduardo Blanco, María Teresa Martín Valdivia, and L. Alfonso Ureña López. 2020a. Detecting negation cues and scopes in Spanish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6902–6911, Marseille, France. European Language Resources Association.

Salud María Jiménez-Zafra, Roser Morante, María Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2020b. Corpora annotated with negation: An overview. *Computational Linguistics*, 46(1):1–52.

Marcel Adam Just and Patricia Ann Carpenter. 1971. Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, 10(3):244–253.

Tian Kang, Shaodian Zhang, Nanfang Xu, Dong Wen, Xingting Zhang, and Jianbo Lei. 2017. Detecting negation and scope in chinese clinical notes using character and word embedding. *Computer Methods and Programs in Biomedicine*, 140:53–59.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Aditya Khandelwal and Benita Kathleen Britto. 2020. Multitask learning of negation and speculation using transformers. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 79–87, Online. Association for Computational Linguistics.

Aditya Khandelwal and Suraj Sawant. 2020. Negbert: A transfer learning approach for negation detection and scope resolution.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2019. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4(2):155–190.

Natalia Konstantinova, Sheila CM De Sousa, Noa P Cruz Díaz, Manuel J Mana López, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Lrec*, pages 3190–3195.

Lydia Lazib, Yanyan Zhao, Bing Qin, and Ting Liu. 2016. Negation scope detection with recurrent neural networks models in review texts. In *Social Computing*, pages 494–508, Singapore. Springer Singapore.

Wenxiao Liu, Shuyuan Lin, Boyu Gao, Kai Huang, Weilin Liu, Zhongcai Huang, Junjie Feng, Xinhong Chen, and Feiran Huang. 2022. Bert-pos: Sentiment analysis of mooc reviews based on bert with part-of-speech information. In *International Conference on Artificial Intelligence in Education*, pages 371–374. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Andy Lücking and Jonathan Ginzburg. 2022. Referential transparency as the proper treatment of quantification. *Semantics and Pragmatics*, 15(4):1–58.

Nick McKenna and Mark Steedman. 2020. Learning negation scope from syntactic structure. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, *SEM, pages 137–142.

Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M. Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C. Max Schmidt, Hongfang Liu, and Mathew Palakal. 2015. Deepen: A negation detection system for clinical text incorporating dependency relation into negex. *Journal of Biomedical Informatics*, 54:213–219.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation cues and their scope in conan doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1563–1568, Istanbul, Turkey. European Language Resources Association (ELRA).

Pradeep G. Mutalik, Aniruddha M. Deshpande, and Prakash M. Nadkarni. 2001. Research paper: Use of general-purpose negation detection to augment concept indexing of medical documents: A quantitative study using the umls. *Journal of the American Medical Informatics Association : JAMIA*, 8 6:598–609.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. 2018. Attention u-net: Learning where to look for the pancreas.

Ying Ou and Jon Patrick. 2015. Automatic negation detection in narrative pathology reports. *Artificial Intelligence in Medicine*, 64(1):41–50.

Barbara H. Partee and Mats Rooth. 1983. Generalized conjunction and type ambiguity. In Rainer Bäuerle, Christoph Schwarze, and Arnim von Stechow, editors, *Meaning, Use, and Interpretation of Language*, Grundlagen der Kommunikation und Kognition / Foundations of Communication and Cognition), pages 361–383. De Gruyter, Berlin.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2017. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports.

Zhong Qian, Ting Zou, Zihao Zhang, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2024. Speculation and negation identification via unified machine reading comprehension frameworks with lexical and syntactic data augmentation. *Engineering Applications of Artificial Intelligence*, 131:107806.

Jonathon Read, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2012. UiO1: Constituent-based discriminative ranking for negation resolution. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 310–318, Montréal, Canada. Association for Computational Linguistics.

Tanya Reinhart. 1997. Quantifier scope: How labor is divided between QR and Choice Functions. *Linguistics and Philosophy*, 20(4):335–397.

Henning Schäfer, Ahmad Idrissi-Yaghir, Peter Horn, and Christoph Friedrich. 2022. Cross-language transfer of high-quality annotations: Combining neural machine translation with cross-linguistic span alignment to apply NER to clinical texts in a low-resource language. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 53–62, Seattle, WA. Association for Computational Linguistics.

Nirja Shah and Jyoti Pareek. 2024. Optimized hindi negation detection using a hybrid rule-based and bert model. In *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, pages 544–550.

Sunghwan Sohn, Stephen T Wu, and Christopher G. Chute. 2012. Dependency parser-based negation detection in clinical narratives. *AMIA Summits on Translational Science Proceedings*, 2012:1 – 8.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Milan Straka and Jana Straková. 2019. Universal dependencies 2.5 models for UDPipe (2019-12-06). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45, Columbus, Ohio. Association for Computational Linguistics.

Stuart J Taylor and Sanda M Harabagiu. 2018. The role of a deep-learning method for negation detection in patient cohort identification from electroencephalography reports. *AMIA Annu. Symp. Proc.*, 2018:1018–1027.

Ye Tian and Richard Breheny. 2019. Negation. In *The Oxford Handbook of Experimental Semantics* *and Pragmatics*, chapter 12, pages 195–207. Oxford University Press.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Thinh Hung Truong, Timothy Baldwin, Trevor Cohn, and Karin Verspoor. 2022. Improving negation detection with negation-focused pre-training.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: An analysis of language models on negation benchmarks.

Johan van der Auwera and Olga Krasnoukhova. 2020. The typology of negation. In Viviane Déprez and M. Teresa Espinal, editors, *The Oxford Handbook of Negation*, Oxford Handbooks in Linguistics, chapter 7, pages 91–116. Oxford University Press, Oxford, UK.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks.

Sharad Verma, Ashish Kumar, and Aditi Sharan. 2024. Wrgat-ptbert: weighted relational graph attention network over post-trained bert for aspect based sentiment analysis. *Applied Intelligence*, 55(3).

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S11).

Junlang Wang, Xia Li, Junyi He, Yongqiang Zheng, and Junteng Ma. 2023. Enhancing implicit sentiment learning via the incorporation of part-of-speech for aspect-based sentiment analysis. In *Chinese Computational Linguistics*, pages 382–399, Singapore. Springer Nature Singapore.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Zhaofeng Wu, Hao Peng, and Noah A. Smith. 2021. Infusing finetuning with semantic dependencies. *Transactions of the Association for Computational Linguistics*, 9:226–242.

Zeguan Xiao, Jiarun Wu, Qingliang Chen, and Congjian Deng. 2021. Bert4gcn: Using bert intermediate layers to augment gcn for aspect-based sentiment classification.

Xiaotang Zhou, Tao Zhang, Chao Cheng, and Shinan Song. 2023. Dynamic multichannel fusion mechanism based on a graph attention network and bert for aspect-based sentiment classification. *Applied Intelligence*, 53(6):6800–6813.

A SpaCy & UDPipe Models

Language	Model	Link
English	en_core_web_sm	https://spacy.io/models
French	fr_core_news_sm	https://spacy.io/models
German	de_core_news_sm	https://spacy.io/models
Italian	it_core_news_sm	https://spacy.io/models
Spanish	es_core_news_sm	https://spacy.io/models
Dutch	nl_core_news_sm	https://spacy.io/models
Russian	ru_core_news_sm	https://spacy.io/models
Chinese	zh_core_web_sm	https://spacy.io/models
Japanese	ja_core_news_sm	https://spacy.io/models
Hindi	hindi-hdtb-ud-2.5-191206	https://lindat.mff.cuni.cz/repository/items/41f05304-629f-4313-b9cf-9eeb0a2ca7c6
Arabic	arabic-padt-ud-2.5-191206	https://lindat.mff.cuni.cz/repository/items/41f05304-629f-4313-b9cf-9eeb0a2ca7c6

Table 5: Overview of syntax tagging models from spaCy (Honnibal et al., 2020) and UDPipe (Straka and Straková, 2019) used in our experiments. The Link column provides the URLs to the pretrained models.

B OPUS-MT Models

Language Pair	Link
en→de	https://huggingface.co/Helsinki-NLP/opus-mt-en-de
de→en	https://huggingface.co/Helsinki-NLP/opus-mt-de-en
en→fr	https://huggingface.co/Helsinki-NLP/opus-mt-en-fr
fr→en	https://huggingface.co/Helsinki-NLP/opus-mt-fr-en
en→es	https://huggingface.co/Helsinki-NLP/opus-mt-en-es
es→en	https://huggingface.co/Helsinki-NLP/opus-mt-es-en
en→it	https://huggingface.co/Helsinki-NLP/opus-mt-en-it
it→en	https://huggingface.co/Helsinki-NLP/opus-mt-it-en
en→nl	https://huggingface.co/Helsinki-NLP/opus-mt-en-nl
nl→en	https://huggingface.co/Helsinki-NLP/opus-mt-nl-en
en→ru	https://huggingface.co/Helsinki-NLP/opus-mt-en-ru
ru→en	https://huggingface.co/Helsinki-NLP/opus-mt-ru-en
en→hi	https://huggingface.co/Helsinki-NLP/opus-mt-en-hi
hi→en	https://huggingface.co/Helsinki-NLP/opus-mt-hi-en
en→ar	https://huggingface.co/Helsinki-NLP/opus-mt-en-ar
ar→en	https://huggingface.co/Helsinki-NLP/opus-mt-ar-en
en→zh	https://huggingface.co/Helsinki-NLP/opus-mt-en-zh
zh→en	https://huggingface.co/Helsinki-NLP/opus-mt-zh-en
en→ja	https://huggingface.co/Helsinki-NLP/opus-mt-en-jap
ja→en	https://huggingface.co/Helsinki-NLP/opus-mt-jap-en

Table 6: Overview of bilingual translation models from the OPUS-MT project (Tiedemann and Thottingal, 2020) used for dataset translation. The Link column provides the URLs to the corresponding pretrained models on the HuggingFace Hub.

C ICL Few-Shot Prompts

```
### Task:
Detect **negation cues** in a tokenized input sentence.

### Instructions:
- The input is a JSON object with one field:
  - "sent": a list of tokens (strings) representing the sentence.
- The output must be a JSON object with one field:
  - "cue_mask": a list of integers (0 or 1) of the same length as "sent".
    - 1 → the token is a negation cue.
    - 0 → the token is not a negation cue.
- A sentence can contain multiple, single, or no negation cues.

### Definition of Negation Cues:
A **negation cue** is any word or phrase that explicitly or implicitly indicates
denial, absence, opposition, or non-existence of an event, property, or entity.

Negation cues include (but are not limited to):

1. **Explicit negators:**
   not, n't, never, no, none, nothing, nobody, neither, nor, nowhere, without

2. **Weak or limiting negators:**
   hardly, barely, scarcely

3. **Negative adjectives/adverbs:**
   Words that begin with negative prefixes such as im-, in-, un-, il-, ir-, dis-, non-
   and convey negation, e.g. improper, invalid, unacceptable, incorrect, disallowed, nonexistent

4. **Negative verbs:**
   Verbs expressing refusal, denial, or absence, e.g. deny, lack, fail, forbid, prohibit, exclude

5. **Context-dependent negators:**
   Words or multi-word expressions that function as negation cues depending on context
   (e.g. free from, absent from, opposed to).

### Examples:

Input:
{"sent": ["<TOKEN_1>", "<TOKEN_2>", "..."]}
Output:
{"cue_mask": [0, 1, 0, ...]}

...
```

Figure 8: ICL few-shot cue detection prompt.

```
### Task:
Identify the **negation scope** in a tokenized input sentence.

### Instructions:
- The input is a JSON object with:
  - "sent": a list of tokens (strings) representing the sentence.
  - "cue_mask": a binary mask indicating which token(s) are negation cues.
- The output must be a JSON object with:
  - "scope_mask": a list of integers (0 or 1) of the same length as "sent".
    - 1 → the token is part of the negation scope.
    - 0 → the token is not part of the negation scope.

### Rules:
- The **negation cue token(s)** are always **excluded** from the scope.
- A negation scope may include **multiple tokens** and can be **discontinuous**.
- Scope tokens may occur **before or after** the cue, depending on sentence structure.
- The scope represents the part ofthe sentence affected by the negation.

### Examples:

Input:
{"sent": ["<TOKEN_1>", "<TOKEN_2>", "..."], "cue_mask": [0,1,0,...]}
Output:
{"scope_mask": [1,1,0,...]}

...
```

Figure 9: ICL few-shot scope detection prompt.

D Multilingual Results

D.1 Germanic Languages

	Neg-Bert (Cue)							D-Neg (Cue)						
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU
DT-Neg	95.05	75.44	78.38	69.63	79.25	91.38	80.77	94.94	80.35	79.23	70.71	79.76	89.51	78.04
Conan	85.83	90.47	87.93	84.18	80.36	88.62	81.95	54.09	93.08	91.35	79.18	82.22	87.8	87.34
Bioscope	86.96	81.18	91.69	91.51	86.02	88.33	89.06	66.77	87.51	90.52	84.77	84.92	88.54	88.74
Bioscope (Full)	79.05	74.68	87.09	82.26	83.64	82.13	82.13	62.07	82.87	79.75	81.85	81.95	82.88	83.21
Prob-Bank	93.37	89.38	93.71	91.98	98.13	95.36	94.18	69.55	94.38	91.08	89.36	97.52	92.35	95.03
SOCC	81.91	84.24	85.11	77.69	84.34	89.33	88.42	55.37	89.14	79.99	73.53	82.19	90.01	89.06
SFU	81.62	80.82	82.65	80.02	85.11	86.55	88.05	51.03	87.08	79.73	77.55	83.82	85.96	88.13

Table 7: **Cue** Detection Results for **DE**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	Neg-Bert (Scope)						D-Neg (Scope)					
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU
DT-Neg	85.99	83.06	77.29	74.73	83.04	84.91	89.33	83.59	82.07	79.19	74.55	75.29
Conan	78.61	81.04	82.66	75.38	78.02	81.86	82.08	86.22	77.7	79.28	76.84	73.22
Bioscope (Abstracts)	74.89	75.36	92.12	82.64	84.09	83.16	77.23	79.81	85.43	82.39	83.07	79.75
Bioscope (Full)	71.23	71.89	86.82	75.66	78.37	72.35	72.35	78.04	74.95	76.5	76.63	76.67
SOCC	68.81	74.63	78.13	77.92	82.69	80.23	73.14	77.78	74.73	78.64	82.28	79.71
SFU	70.06	74.98	81.5	80.42	85.64	87.24	74.69	75.77	80.45	80.89	84.39	85.24

Table 8: **Scope** Detection Results for **DE**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	in-domain results			out-of-domain results		
	Neg-Bert	D-Neg	Diff.	Neg-Bert	D-Neg	Diff.
Cue Detection	90.71	90.86	0.15	85.45	82.36	-3.08
Scope Detection	84.12	84.17	0.04	79.37	79.16	-0.21

Table 9: Aggregated results for **DE**: Showing the average F1 score of models on in-domain and out-of-domain datasets.

	Neg-Bert (Cue)							D-Neg (Cue)						
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU
DT-Neg	97.46	80.27	85.0	73.06	85.62	89.06	88.56	96.88	77.1	80.57	80.24	79.32	79.62	80.3
Conan	79.53	90.03	84.96	81.83	81.01	89.18	81.01	84.79	91.34	89.38	87.33	84.0	90.66	87.92
Bioscope	84.82	86.23	92.39	92.1	87.33	88.34	90.04	81.98	84.81	92.01	87.47	86.64	87.9	89.13
Bioscope (Full)	79.47	80.51	88.75	82.68	84.61	85.4	84.69	80.47	78.43	87.15	81.18	83.23	85.22	83.86
Prob-Bank	92.23	94.82	93.8	92.2	97.72	92.5	92.79	92.74	95.19	91.88	92.11	97.65	91.52	94.58
SOCC	80.74	90.66	86.41	81.87	87.18	92.95	89.21	83.31	87.56	84.03	82.55	85.67	91.96	88.73
SFU	81.22	85.63	85.2	82.39	83.87	87.36	88.66	82.45	84.13	83.24	84.11	82.67	87.18	88.63

Table 10: **Cue** Detection Results for **NL**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	Neg-Bert (Scope)						D-Neg (Scope)					
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU
DT-Neg	89.61	87.23	87.42	85.09	88.36	84.88	92.22	87.17	82.45	79.05	80.93	83.22
Conan	82.62	82.45	88.11	84.98	83.85	84.2	83.15	85.68	81.52	80.56	82.98	83.83
Bioscope (Abstracts)	80.42	83.88	87.44	83.64	86.82	84.22	77.83	80.21	90.33	84.44	87.8	84.28
Bioscope (Full)	73.33	80.3	86.52	80.84	85.02	78.81	74.15	78.77	84.18	81.3	85.87	81.08
SOCC	72.8	80.18	79.82	77.84	86.95	80.86	71.99	77.07	79.84	82.4	84.86	80.81
SFU	78.4	82.61	78.23	76.48	86.78	86.2	76.08	80.24	84.09	83.71	86.16	86.71

Table 11: **Scope** Detection Results for **NL**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	in-domain results			out-of-domain results		
	Neg-Bert	D-Neg	Diff.	Neg-Bert	D-Neg	Diff.
Cue Detection	91.70	91.38	-0.32	86.80	86.34	-0.46
Scope Detection	85.58	86.85	1.27	82.98	82.42	-0.56

Table 12: Aggregated results for **NL**: Showing the average F1 score of models on in-domain and out-of-domain datasets.

D.2 Romance Languages

	Neg-Bert (Cue)							D-Neg (Cue)						
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU
DT-Neg	93.92	72.24	69.57	52.32	72.75	83.36	69.18	96.24	78.65	90.81	73.14	89.37	88.79	80.47
Conan	70.24	82.65	61.95	53.45	64.0	77.98	74.27	77.9	84.09	79.05	75.06	71.29	82.98	80.54
Bioscope	81.79	79.21	90.62	73.61	79.21	84.92	85.57	84.9	86.9	91.77	86.57	83.89	87.66	87.57
Bioscope (Full)	75.56	75.24	85.38	62.67	73.16	83.72	80.77	78.28	80.21	84.75	84.71	78.12	81.55	82.32
Prob-Bank	87.92	91.97	89.44	62.67	97.71	85.89	92.81	89.11	91.26	89.71	87.16	97.58	93.01	93.52
SOCC	78.39	84.96	74.52	56.53	80.8	89.73	83.66	79.95	87.77	82.5	82.13	84.19	91.34	88.46
SFU	74.99	81.17	77.76	58.67	79.76	81.11	84.73	80.44	83.65	81.19	80.33	82.99	85.1	86.61

Table 13: **Cue** Detection Results for **IT**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	Neg-Bert (Scope)						D-Neg (Scope)					
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU
DT-Neg	90.59	87.05	88.35	86.31	85.38	85.63	89.8	84.6	79.28	85.11	81.85	83.22
Conan	83.62	85.95	90.64	86.64	82.83	87.27	87.05	87.15	85.51	85.46	83.08	84.07
Bioscope (Abstracts)	82.93	83.57	93.0	89.28	88.11	87.01	81.24	83.67	89.15	87.02	85.81	86.84
Bioscope (Full)	83.38	81.9	88.39	91.48	86.99	86.93	79.72	82.27	84.57	87.78	87.01	88.33
SOCC	77.98	81.03	86.55	84.89	91.11	88.52	76.74	81.64	84.67	85.24	89.53	88.63
SFU	80.78	81.4	87.11	85.01	88.03	89.62	81.79	80.12	84.42	85.39	87.08	86.98

Table 14: **Scope** Detection Results for **IT**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	in-domain results			out-of-domain results		
	Neg-Bert	D-Neg	Diff.	Neg-Bert	D-Neg	Diff.
Cue Detection	86.00	90.33	4.33	77.64	84.60	6.96
Scope Detection	90.29	88.40	-1.89	86.26	84.77	-1.48

Table 15: Aggregated results for **IT**: Showing the average F1 score of models on in-domain and out-of-domain datasets.

	Neg-Bert (Cue)							D-Neg (Cue)						
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU
DT-Neg	92.89	67.28	71.34	58.74	72.63	78.53	63.61	93.8	72.1	85.66	75.75	84.71	86.89	87.21
Conan	61.18	83.06	72.0	62.54	74.27	80.4	68.95	76.84	87.51	81.13	77.36	75.22	89.08	83.67
Bioscope	71.39	78.15	90.06	82.47	82.27	86.09	75.62	84.54	82.99	92.46	85.97	85.16	85.66	86.6
Bioscope (Full)	68.2	76.35	82.98	72.58	77.25	85.54	73.56	80.24	78.9	87.51	81.4	80.73	83.45	83.2
Prob-Bank	83.25	90.29	87.52	77.81	96.1	86.47	83.94	93.1	90.28	89.43	91.5	97.2	91.76	93.39
SOCC	68.63	82.7	72.71	65.12	79.91	88.45	76.4	87.12	84.01	86.43	86.04	84.2	90.94	88.79
SFU	74.0	82.51	76.44	67.55	84.24	84.2	82.16	82.01	85.73	80.93	81.78	84.38	86.3	87.64

Table 16: **Cue** Detection Results for **ES**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	Neg-Bert (Scope)						D-Neg (Scope)					
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU
DT-Neg	88.57	87.59	82.09	85.98	85.22	86.0	87.42	67.6	84.03	81.24	78.66	77.94
Conan	83.26	84.28	87.66	84.22	83.43	85.34	77.59	65.07	84.97	85.57	87.34	81.94
Bioscope (Abstracts)	83.58	83.55	90.99	86.62	83.56	87.98	68.53	61.75	85.24	86.68	86.45	84.21
Bioscope (Full)	80.71	84.29	90.46	87.11	84.75	85.78	68.94	62.32	84.69	84.1	86.16	84.62
SOCC	78.23	84.3	84.93	84.24	85.18	87.26	71.3	65.23	79.86	84.39	86.69	84.48
SFU	82.87	87.04	86.69	86.62	88.27	90.53	72.1	65.85	82.99	86.71	89.24	88.74

Table 17: **Scope** Detection Results for **ES**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	in-domain results			out-of-domain results		
	Neg-Bert	D-Neg	Diff.	Neg-Bert	D-Neg	Diff.
Cue Detection	86.47	90.14	3.66	77.56	85.28	7.72
Scope Detection	87.78	82.88	-4.90	85.53	79.46	-6.07

Table 18: Aggregated results for **ES**: Showing the average F1 score of models on in-domain and out-of-domain datasets.

	Neg-Bert (Cue)							D-Neg (Cue)						
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU
DT-Neg	89.53	68.46	72.95	65.1	75.37	72.93	60.21	93.24	66.68	66.53	68.78	70.67	67.97	65.94
Conan	72.48	75.61	71.52	61.85	73.96	78.97	67.34	75.51	86.85	80.78	77.93	71.32	85.66	76.68
Bioscope	76.65	73.33	87.33	79.06	81.13	83.94	76.43	81.49	78.93	88.75	83.7	80.55	85.3	84.74
Bioscope (Full)	75.99	74.83	79.76	72.88	76.88	80.76	73.61	74.59	73.91	83.38	79.97	80.37	79.35	80.16
Prob-Bank	80.6	83.6	82.77	80.91	92.52	83.33	73.39	84.44	86.16	86.9	86.04	92.9	86.85	86.59
SOCC	67.11	74.99	69.7	64.56	76.89	84.82	65.87	71.13	83.03	81.48	79.62	77.75	88.82	83.25
SFU	70.62	78.62	73.43	68.66	79.81	82.63	75.23	74.8	81.37	81.53	78.42	80.28	84.95	84.15

Table 19: **Cue** Detection Results for **FR**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	Neg-Bert (Scope)						D-Neg (Scope)					
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU
DT-Neg	89.64	84.74	72.79	78.73	82.64	80.23	89.22	74.87	75.5	77.76	83.89	81.3
Conan	81.9	85.09	77.25	85.5	80.53	81.46	82.92	82.69	83.48	80.84	84.33	82.01
Bioscope (Abstracts)	75.88	82.01	88.88	88.81	85.05	85.39	73.5	78.84	87.61	86.02	85.32	82.36
Bioscope (Full)	76.49	83.83	85.82	88.86	84.29	85.54	72.05	78.24	83.85	86.52	85.85	81.53
SOCC	72.57	80.04	81.3	82.65	85.84	86.79	68.69	76.21	83.45	82.55	87.55	84.17
SFU	77.61	81.72	80.36	84.52	87.11	86.9	75.81	78.01	82.7	83.24	86.17	86.21

Table 20: **Scope** Detection Results for **FR**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	in-domain results			out-of-domain results		
	Neg-Bert	D-Neg	Diff.	Neg-Bert	D-Neg	Diff.
Cue Detection	82.56	87.81	5.25	75.69	80.21	4.52
Scope Detection	87.54	86.63	-0.90	82.74	81.53	-1.21

Table 21: Aggregated results for **FR**: Showing the average F1 score of models on in-domain and out-of-domain datasets.

D.3 Global Languages

	Neg-Bert (Cue)							D-Neg (Cue)						
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU
DT-Neg	90.67	61.16	52.73	67.2	64.08	69.67	72.0	89.13	64.89	62.75	51.04	69.08	72.43	66.59
Conan	69.36	64.54	49.82	49.78	68.61	73.65	72.93	70.31	62.81	64.01	54.5	77.41	74.81	76.76
Bioscope	74.14	65.81	76.27	80.44	79.92	77.78	81.94	72.43	74.18	84.06	68.34	77.68	82.56	81.48
Bioscope (Full)	66.79	66.58	62.9	72.24	72.91	73.74	76.56	71.35	70.81	79.08	68.6	77.37	80.79	80.56
Prob-Bank	82.15	76.68	59.86	72.2	92.44	86.55	88.27	85.43	75.83	78.75	62.27	93.79	85.96	81.5
SOCC	78.76	72.68	55.05	63.08	77.63	83.46	84.81	77.18	78.61	75.65	54.18	84.52	86.15	79.95
SFU	71.75	69.01	55.28	65.95	75.26	76.02	82.84	71.61	76.17	73.08	53.31	77.98	78.77	77.2

Table 22: **Cue** Detection Results for **AR**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	Neg-Bert (Scope)						D-Neg (Scope)					
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU
DT-Neg	84.56	52.87	67.75	71.16	69.94	67.88	85.72	68.64	74.36	72.12	80.04	76.19
Conan	66.89	51.54	74.03	73.85	74.08	73.38	70.53	73.6	75.8	69.49	75.59	71.36
Bioscope (Abstracts)	65.33	49.29	84.28	80.43	78.69	80.68	63.64	67.61	83.56	68.6	80.23	72.83
Bioscope (Full)	65.02	48.95	77.91	78.43	77.3	77.78	64.22	69.04	81.26	69.92	79.37	76.13
SOCC	62.82	49.54	70.13	76.12	78.73	77.45	64.38	65.71	77.95	67.45	74.77	72.71
SFU	56.83	51.57	71.75	76.18	75.85	76.43	61.65	67.76	74.71	69.6	75.93	73.96

Table 23: **Scope** Detection Results for **AR**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	in-domain results			out-of-domain results		
	Neg-Bert	D-Neg	Diff.	Neg-Bert	D-Neg	Diff.
Cue Detection	80.35	80.25	-0.10	71.92	74.16	2.24
Scope Detection	75.66	76.92	1.26	69.87	72.68	2.81

Table 24: Aggregated results for **AR**: Showing the average F1 score of models on in-domain and out-of-domain datasets.

	Neg-Bert (Cue)							D-Neg (Cue)						
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU
DT-Neg	99.34	94.96	47.97	47.97	47.97	50.07	47.97	100.0	49.06	49.06	49.06	49.06	49.06	49.06
Conan	49.5	49.5	49.5	49.5	49.5	49.42	49.5	49.73	49.73	49.73	49.73	49.73	49.73	49.73
Bioscope	50.0	75.0	50.0	50.0	50.0	61.76	50.0	49.99	49.99	49.99	49.99	49.99	49.99	49.99
Bioscope (Full)	100.0	49.99	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Prob-Bank	49.97	51.88	49.97	49.97	49.97	52.67	49.97	49.97	49.97	49.97	49.97	49.97	49.97	49.97
SOCC	54.7	49.93	49.94	49.94	49.94	49.91	49.94	49.91	49.93	49.93	49.91	49.91	49.93	49.93
SFU	52.41	56.63	49.98	49.98	49.98	51.57	49.98	49.96	49.96	49.96	49.96	49.96	49.96	49.96

Table 25: **Cue** Detection Results for **ZH**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	Neg-Bert (Scope)						D-Neg (Scope)					
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU
DT-Neg	86.67	67.77	56.81	63.39	57.01	50.33	81.04	50.79	48.43	63.93	61.34	48.59
Conan	57.57	50.41	68.69	49.6	54.83	47.23	51.99	51.99	51.36	56.3	54.35	48.23
Bioscope (Abstracts)	56.4	52.16	73.86	53.42	55.18	50.39	49.31	50.31	61.38	60.45	50.74	49.33
Bioscope (Full)	59.0	56.21	64.6	57.84	52.7	49.69	51.31	55.55	64.39	70.18	57.62	49.43
SOCC	59.68	61.81	66.93	56.6	60.83	49.82	50.7	51.57	60.63	62.17	60.89	48.96
SFU	55.31	61.29	66.27	58.54	60.53	54.52	50.89	55.28	58.27	64.89	61.53	49.47

Table 26: **Scope** Detection Results for **ZH**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	in-domain results			out-of-domain results		
	Neg-Bert	D-Neg	Diff.	Neg-Bert	D-Neg	Diff.
Cue Detection	64.10	64.23	0.13	58.95	57.99	-0.97
Scope Detection	64.02	62.49	-1.53	58.44	55.93	-2.51

Table 27: Aggregated results for **ZH**: Showing the average F1 score of models on in-domain and out-of-domain datasets.

	Neg-Bert (Cue)							D-Neg (Cue)						
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU
DT-Neg	94.16	75.39	74.79	69.54	78.74	80.48	72.38	93.19	82.97	82.16	49.19	77.29	81.97	90.21
Conan	70.01	80.27	77.83	70.77	72.69	77.64	72.25	72.04	74.1	76.22	49.55	73.54	79.11	78.19
Bioscope	83.77	87.23	88.75	85.57	86.43	87.08	84.84	71.08	83.54	88.43	49.89	85.97	85.45	88.76
Bioscope (Full)	75.95	83.06	85.21	82.18	82.55	83.65	77.59	70.72	81.7	85.71	49.87	81.73	84.08	84.08
Prob-Bank	87.42	94.79	93.4	92.52	96.21	93.48	93.18	84.33	90.35	92.1	49.62	96.0	90.81	93.86
SOCC	75.77	85.71	86.27	76.27	81.48	86.72	78.43	73.11	83.44	85.76	49.52	81.07	85.33	86.88
SFU	77.81	83.33	83.2	79.99	81.8	84.62	80.4	71.83	78.92	82.05	49.73	80.43	81.03	83.22

Table 28: **Cue** Detection Results for **HI**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	Neg-Bert (Scope)						D-Neg (Scope)					
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU
DT-Neg	84.92	73.03	80.12	68.66	73.81	69.54	81.09	76.3	78.02	72.6	68.23	66.8
Conan	79.51	79.95	77.06	63.26	72.49	67.66	80.32	77.28	78.65	71.8	68.49	71.99
Bioscope (Abstracts)	70.15	61.11	76.88	64.38	70.12	64.18	68.72	64.71	71.84	65.84	61.91	63.81
Bioscope (Full)	69.93	56.51	70.51	63.72	65.14	59.9	68.29	61.86	67.22	63.71	62.62	60.56
SOCC	66.86	67.83	71.73	62.94	73.82	67.93	65.54	66.78	69.27	62.54	64.52	68.28
SFU	69.98	69.73	74.39	66.66	75.65	72.22	69.84	69.89	71.35	68.53	69.08	74.0

Table 29: **Scope** Detection Results for **HI**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	in-domain results			out-of-domain results		
	Neg-Bert	D-Neg	Diff.	Neg-Bert	D-Neg	Diff.
Cue Detection	86.96	81.45	-5.51	82.32	77.96	-4.36
Scope Detection	75.25	72.07	-3.18	70.06	69.23	-0.83

Table 30: Aggregated results for **HI**: Showing the average F1 score of models on in-domain and out-of-domain datasets.

	Neg-Bert (Cue)							D-Neg (Cue)						
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU
DT-Neg	91.48	49.76	49.76	49.76	49.76	49.76	49.76	93.98	49.67	49.67	49.67	49.67	49.67	49.67
Conan	49.86	49.86	49.86	49.86	49.86	49.86	49.86	49.89	49.89	49.89	49.89	49.89	49.89	49.89
Bioscope	49.99	49.99	49.99	52.77	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99
Bioscope (Full)	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99	49.99
Prob-Bank	49.98	49.98	49.98	49.98	63.03	49.98	49.98	49.98	49.98	49.98	49.98	58.32	49.98	49.98
SOCC	51.7	49.91	49.91	49.91	49.91	49.91	49.91	51.95	49.91	49.91	49.91	49.91	49.91	49.91
SFU	51.39	49.95	49.95	49.95	49.95	49.95	49.95	50.73	49.95	49.95	49.95	49.95	49.95	49.95

Table 31: **Cue** Detection Results for **JAP**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	Neg-Bert (Scope)						D-Neg (Scope)					
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU
DT-Neg	75.18	57.75	49.1	49.1	54.81	49.75	63.82	48.94	48.94	48.94	53.46	48.94
Conan	57.43	54.58	49.62	48.83	48.83	49.6	52.83	48.92	48.92	48.92	50.58	48.92
Bioscope (Abstracts)	54.28	49.64	50.56	49.51	49.51	49.64	50.15	49.67	49.67	49.67	50.31	49.67
Bioscope (Full)	54.25	50.6	52.05	49.55	49.55	49.55	50.14	49.64	49.64	49.64	49.64	49.64
SOCC	54.41	51.87	50.19	49.28	49.28	49.27	49.77	49.39	49.39	49.39	51.24	49.39
SFU	51.93	51.04	50.63	49.6	49.6	50.98	50.53	49.64	49.64	49.64	54.07	49.64

Table 32: **Scope** Detection Results for **JAP**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	in-domain results			out-of-domain results		
	Neg-Bert	D-Neg	Diff.	Neg-Bert	D-Neg	Diff.
Cue Detection	57.74	57.43	-0.31	51.16	51.04	-0.12
Scope Detection	55.02	52.16	-2.87	51.70	50.31	-1.39

Table 33: Aggregated results for **JAP**: Showing the average F1 score of models on in-domain and out-of-domain datasets.

	Neg-Bert (Cue)							D-Neg (Cue)						
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	Prob-Bank (Focus)	SOCC	SFU
DT-Neg	92.71	62.21	67.9	65.25	71.24	85.29	70.04	93.55	82.19	76.77	72.66	77.07	87.94	89.83
Conan	72.76	78.99	68.39	66.9	71.55	78.3	72.4	71.0	85.42	73.32	77.9	78.55	84.33	81.61
Bioscope	72.96	60.18	88.32	80.51	79.41	80.72	80.61	72.6	81.65	88.74	85.09	84.75	86.43	82.77
Bioscope (Full)	65.83	57.04	76.8	77.28	74.18	74.67	77.19	69.3	75.97	78.43	79.25	78.78	80.62	80.79
Prob-Bank	84.71	70.37	87.19	85.18	95.31	83.21	90.12	87.51	91.11	84.66	89.29	94.94	90.22	88.67
SOCC	72.18	65.29	72.79	71.4	76.65	82.93	82.07	73.7	81.77	75.23	77.64	79.86	87.55	85.26
SFU	73.64	67.48	74.37	71.8	78.87	77.44	82.28	77.52	80.89	76.62	78.85	82.18	83.39	85.13

Table 34: **Cue** Detection Results for **RU**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	Neg-Bert (Scope)						D-Neg (Scope)					
	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU	DT-Neg	Conan	Bioscope (Abstracts)	Bioscope (Full)	SOCC	SFU
DT-Neg	85.28	84.07	82.49	77.67	85.26	84.11	85.43	85.8	75.02	69.09	77.91	73.54
Conan	81.19	80.25	84.21	82.1	85.45	81.11	83.21	82.69	82.13	75.02	83.25	77.25
Bioscope (Abstracts)	81.15	83.8	90.83	86.41	86.32	86.13	80.37	81.19	90.14	84.21	85.69	83.26
Bioscope (Full)	72.88	81.62	90.25	85.02	82.89	78.87	73.3	80.29	81.32	76.65	81.78	75.41
SOCC	69.41	80.08	81.07	79.97	85.3	84.27	69.93	77.13	74.09	73.62	85.31	80.85
SFU	76.8	80.41	83.32	81.09	85.9	86.37	74.77	77.03	79.82	81.13	84.38	84.92

Table 35: **Scope** Detection Results for **RU**: Columns represent training datasets, and rows represent test datasets. Scores indicate macro F1 scores. Green highlights the best overall performance of models on target test datasets.

	in-domain results			out-of-domain results		
	Neg-Bert	D-Neg	Diff.	Neg-Bert	D-Neg	Diff.
Cue Detection	85.40	87.80	2.39	75.81	81.82	6.01
Scope Detection	85.51	84.19	-1.32	82.59	79.64	-2.96

Table 36: Aggregated results for **RU**: Showing the average F1 score of models on in-domain and out-of-domain datasets.