Improving Neural Argumentative Stance Classification in Controversial Topics with Emotion-Lexicon Features

Mohammad Yeghaneh Abkenar[1,3] Weixing Wang[2]
Manfred Stede[3] Davide Picca[4] Mark A. Finlayson[5] Panagiotis Ioannidis[6]
[1]Innovations Department, Bundesdruckerei GmbH, Berlin, Germany
[2]Hasso-Plattner-Institut, University of Potsdam, Germany
[3]Department of Linguistics, University of Potsdam, Germany
[4]University of Lausanne, Switzerland, [5]Knight Foundation School of Computing & Information Sciences, Florida International University, USA, [6]Pisquared, Germany

yeghanehabkenar@uni-potsdam.de, weixing.wang@hpi.de
stede@uni-potsdam.de, davide.picca@unil.ch, markaf@fiu.edu, pioannidis@pisquared.io

Abstract

Argumentation mining comprises several subtasks, among which stance classification focuses on identifying the standpoint expressed in an argumentative text toward a specific target topic. While arguments—especially about controversial topics—often appeal to emotions, most prior work has not systematically incorporated explicit, fine-grained emotion analysis to improve performance on this task. In particular, prior research on stance classification has predominantly utilized non-argumentative texts and has been restricted to specific domains or topics, limiting generalizability. We work on five datasets from diverse domains encompassing a range of controversial topics and present an approach for expanding the Bias-Corrected NRC Emotion Lexicon using DistilBERT embeddings, which we feed into a Neural Argumentative Stance Classification model. Our method systematically expands the emotion lexicon through contextualized embeddings to identify emotionally charged terms not previously captured in the lexicon. Our expanded NRC lexicon (eNRC) improves over the baseline across all five datasets (up to +6.2 percentage points in F_1 score), outperforms the original NRC on four datasets (up to +3.0), and surpasses the LLM-based approach on nearly all corpora. We provide all resources—including eNRC, the adapted corpora, and model architecture—to enable other researchers to build upon our work.

Keywords: Argumentation Mining, Stance Classification, Emotion Lexicon

1. Introduction

Argumentation has long been central to human discourse. In ancient Greece (4th century BCE), Aristotle systematically analyzed rhetoric and persuasion. He developed modal logic theory which described three modes of persuasion: *Pathos* (emotions and values), *Ethos* (credibility and authority), and *Logos* (logical reasoning; Schiappa, 2010). Meanwhile, in China (5th century BCE), Confucius developed theories of and approaches to ethical reasoning (Lu, 1998). Centuries later, Persian polymaths Al-Farabi (9th–10th century CE) and Avicenna (Ibn Sina) (10th–11th century CE) refined and expanded Aristotelian logic, creating comprehensive theories of dialectics, syllogistic reasoning, and rhetoric (Street and Germann, 2021). More recently, in the 20th century, Toulmin, in his famous book *The Uses of Argument* (1958), defined arguments as comprising six components—*data* (premise), *claim*, and *warrant* being required, while *backing*, *qualifier* and *rebuttal* being optional. Freeman's approach (2011) integrates the standard model with Toulmin's framework, using five elements: premise, conclusion, modality, rebuttal, and counter-rebuttal. Premise and conclusion (Claim) are the basic elements. For practical argumentation mining applications, this is typically simplified to premises and claims. For example, *"Climate change threatens coastal cities"* (premise) supports *"We must reduce carbon emissions"* (claim).

Related to argumentation, discussions about controversial topics have taken many forms throughout history: from Munazara (formal debate) in madrasas and the Athenian agora, to European parliamentary debates shaping European democracies, town hall meetings in early America, 20th century mass media debates, and today's online micro blogging forums like *Reddit* and *Quora* alongside televised presidential debates. Controversies serve a productive function when contributions successfully delineate proponent and opponent standpoints, facilitating the organization of discourse into identifiable argumentative clusters (Vecchi et al., 2021). Together, these forms demonstrate how argumentation about controversial topics remains vital to human communication and decision-making.

The continuing interest in argumentation about controversial topics motivates our focus, but there are also significant practical applications. For example, robust stance classification can contribute to computational approaches to analyze public discourse at scale, identify majority and minority per-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9678–9691
11–16 May 2026.

spectives, distinguish genuine sentiment from amplified extremes, and detect misinformation campaigns that manipulate public opinion on controversial topics. Such capabilities are becoming increasingly valuable in a highly polarized media ecosystem.

1.1. Contributions

We focus here on the use and expansion of emotion lexicons for one of the less addressed subtasks: argumentative stance classification on controversial topics. Controversial issues involve fundamental value conflicts that evoke affective responses—emotional framing (e.g., fear, anger, hope, disgust) often reveals underlying stances more reliably than propositional content alone. Empirically, emotion lexicons enable knowledge transfer across controversial targets (Zhang et al., 2020), improving stance detection on diverse contemporary issues (Hosseinia et al., 2020).

Our work makes two main contributions. First, we demonstrate a new method for expanding existing emotion lexicons through a clustering-based method using DistilBERT embeddings and similarity metrics. We apply this method to the Bias-Corrected version of the NRC lexicon (Zad et al., 2021) to produce the expanded NRC lexicon (eNRC). Second, we present a novel neural argumentative stance classification framework for leveraging the emotion lexicon to improve classifying stance in argumentative text on controversial topics. This framework involves a redefinition of the stance classification task for the argumentation mining corpora, achieved through extensive pre-processing of five well-known datasets and the integration of an emotion lexicon adapter. We release all code and resources to ensure full reproducibility.[1]

2. Background & Related work

2.1. Background

According to Bentahar et al. (2010), there are three types of models of argumentation: *monological* models, *dialogical* models, and *rhetorical* models. Monological models focus on the internal structure of arguments and relations among components (microstructure). Dialogical models focus on relationships between arguments while ignoring internal structures. Finally, rhetorical models consider rhetorical structure and patterns. Given that stance classification primarily concerns the internal structure of arguments and relationships among components, our work is situated primarily within work on monological models. All our corpora either have microstructure annotations or are adapted accordingly. Argumentation microstructures are closely tied to discourse analysis and frameworks such as RST (Mann and Thompson, 1987; Hewett et al., 2019), which models discourse structure as a tree, PDTB (Prasad et al., 2017), which focuses more on shallow structure, and SDRT (Asher and Lascarides, 2003), which encompasses both logical and rhetorical structural elements. Despite this focus, argumentation inherently has dialogical aspects (Freeman, 2011), making it suitable for analyzing controversial topics, a characteristic partially reflected in the structure of our corpora.

We organized our review of the background into three sections: *argumentation mining*, *emotion lexicons*, and *stance classification*, since these are closest to the focus of this paper.

Argumentation Mining Argumentation mining as an topic of research has evolved rapidly in recent years with the widespread use of language models resulting in numerous workshops at major NLP conferences (Chistova et al., 2025). The term "argument mining" is an umbrella term for several subtasks, such as *argument component type classification* (ACTC), *argumentative relation classification* (ARC), *argument structure identification*, *argument stance classification*, and *argument quality assessment*. However, despite the important connections among these subtasks, each is inherently complex and poses distinct challenges. Therefore, it is reasonable to address each of these subtasks separately. Moreover, as discussed in Feger et al. (2025), the goal is to learn the task itself rather than merely memorizing the dataset. This paper focuses on one of the less-addressed subtasks, namely argumentative stance classification (identifying an author's standpoint on a topic), and evaluates the approach on five well-known corpora.

Emotion Lexicons An emotion lexicon is a specialized linguistic resource that connects the emotional words in a language to predefined emotion categories such as Plutchik's eight basic emotions (Plutchik, 1965). Each word in the lexicon is linked to one or more emotion labels, or sometimes none at all (Mohammad, 2023). Prominent emotion lexicons include the General Inquirer (Stone et al., 1966), ANEW (Nielsen, 2011), the Pittsburgh Subjectivity Lexicon (Wilson et al., 2005), the NRC Emotion Lexicon and it's updated versions (Mohammad and Turney, 2013, 2010; Zad et al., 2021), the NRC Valence, Arousal, and Dominance (VAD) Lexicon (Mohammad, 2025), and SenticNet (Cambria et al., 2016), all of which were created through manual annotation by experts or crowdsourcing.

[1] Our expansion of the NRC lexicon **eNRC** can be found online, while the code for the experiments and model can be found at **ArgStanceNRC**. We will archive these objects also in a permanent institutional repository.

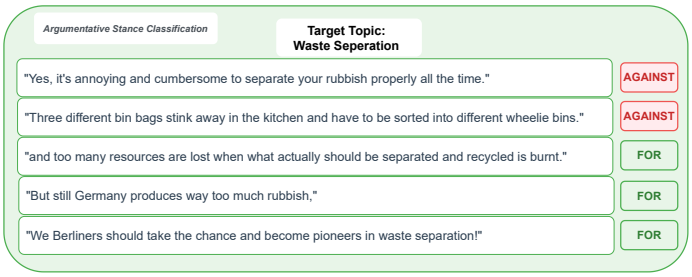

Figure 1: An example with five segments from the Argumentative Microtext corpus (Part 1), showing stance labels (For or Against) toward a controversial target topic.

For our stance classification model, we use the eight emotion categories from the NRC Emotion Lexicon (*joy*, *trust*, *fear*, *surprise*, *sadness*, *disgust*, *anger*, and *anticipation*) as features and propose an embedding-based expansion (eNRC) that extends emotion labels to semantically similar words while preserving categorical structure.

Stance Classification One of the earliest works addressed detecting support or opposition to controversial legislation from Congressional floor debates. Interestingly, they framed this as document-level sentiment-polarity classification without using the term "stance" (Thomas et al., 2006). Stance can be defined as being for or against a defined target, such as a controversial topic, e.g., being *for* or *against* "waste separation" (a case illustrated in figure 1), "charge for plastic bags", "technology makes children even more creative", "multiculturalism", or "death penalty".

Another relevant genre for stance classification is argumentative student essays responding to pro/con prompts. Faulkner (2014) estimated essay-level stance using on-topicness scoring via Wikipedia link similarity and stance-oriented dependency parsing to detect pro/con subtrees. The task of stance detection was later popularized and standardized by Mohammad et al. (2016), which focused on Twitter data. Stance classification (detection) is a well-established task, often studied separately from argument mining more broadly (Küçük and Can, 2020). In the next subsection on related work, we focus primarily on approaches that explicitly incorporate arguments into their stance models. The emphasis is on the intersection of argumentation mining, stance classification, and emotion analysis.

2.2. Related Work

There has been limited work on the intersection of argumentative stance classification, emotions, and controversial topics. For example, Sobhani et al. (2015) found that using automatically-extracted arguments as features for stance classification yielded promising results. However, this work was also relying on TF-IDF features and predicted argument tags with a linear SVM classifier. They compared this to a TF-IDF-only linear SVM and a majority-class baseline. In contrast, we developed a sophisticated end-to-end neural framework for argumentative stance classification, evaluated across multiple domains. Although we did not use explicit argument features for the stance classification task, our approach implicitly captures argumentative and emotional cues through its context-aware neural architecture. Furthermore, Bar-Haim et al. (2017) presented pivotal work in stance classification using the IBM dataset Aharoni et al. (2014). However, their approach focused on a single corpus and employed sentiment rather than emotion features. Schaefer and Stede (2019) utilized the Atheism Stance Corpus (ASC) (Wojatzki and Zesch, 2016), comprising 715 tweets from the atheism portion of SemEval dataset. They demonstrated that word and sentence embeddings improved task performance, though not all variants performed equally. Comparing fastText, GloVe, and Universal Sentence Encoder (USE), they found USE achieved state-of-the-art results. However, they noted limited generalization assumptions given the corpus's focus on a single topic. In contrast, we evaluated our approach on multiple corpora spanning various topics, providing a broader assessment of its generalizability across domains. Stede (2020) discussed the role of stance in argumentation mining, but not emotion or controversy. He focused more on theory, however, with some interesting examples which reflected how the task of finding a stance standpoint on a specific target topic can be complicated. He discussed how argumentation mining links to sentiment and stance detection—two popular tasks in computational linguistics often used as features in argumentation mining systems. He noted that while stance is closely related to argument, computing it accurately can be challenging. Meanwhile, straightforward sentiment systems relying only on pre-stored lexical polarities offer limited value and can often mislead. Therefore, we explicitly avoided using simple sentiment-based fea-

tures, which can lead to misleading interpretations (Stede, 2020, Section 7). Instead, we employed a context-aware emotion lexicon, designed to capture emotional nuances beyond simple lexical polarity. We demonstrated the initial potential of this flexible emotion lexicon for improving performance on the argumentative stance classification subtask.

3. Development of eNRC: An Expansion of NRC

To improve emotion coverage in argumentative and stance-related text, we developed an approach to expand the NRC Emotion Lexicon. Although widely used, NRC misses many emotionally charged words common in modern language. We therefore propose an embedding-based expansion (eNRC) that groups semantically similar words and extends emotion labels while preserving categorical structure. The clustering step identifies semantic regions of emotional vocabulary, enabling us to normalize similarity scores and prevent category overlap. This ensures that the extended lexicon remains both comprehensive and interpretable.

Our experimental framework uses a batch-processing pipeline to compute DistilBERT embeddings and evaluate similarity scores. Cluster statistics (mean and standard deviation) for normalized similarity are precomputed. We use Hamming distance to measure the average difference between binary emotion vectors and entropy to quantify the diversity of emotion assignments. These metrics provide insights into the coherence and diversity of the expanded lexicon.

3.1. Clustering and Embedding Distribution

Trimodal Similarity Patterns We used DistilBERT to embed each word into a 768-dimensional space and to calculate cosine similarity. Figure 2 shows the similarity distribution between each NRC word embedding and all other words in the lexicon (including self-comparisons). The distributions are not uniform: each emotion exhibits a trimodal pattern, with the tall bar at a similarity of one corresponding to self-comparisons. This non-canonical shape suggests the presence of multiple density regions in the embedding space, which we explore next through dimensionality reduction and clustering.

Dimensionality Reduction and Clustering After observing a similar trimodal distribution in the similarity scores for each emotion, we reduced the embeddings to three dimensions using PCA (Hotelling, 1933) to capture the dominant variance

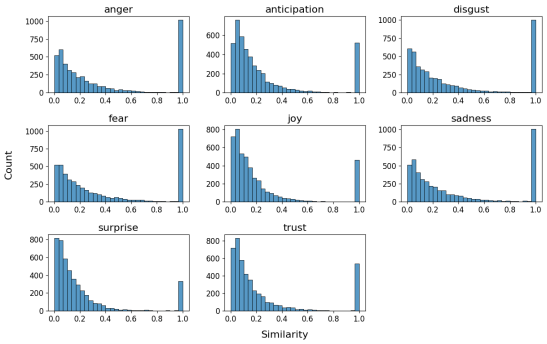

Figure 2: Trimodal similarity distributions across emotion categories in the original NRC lexicon.

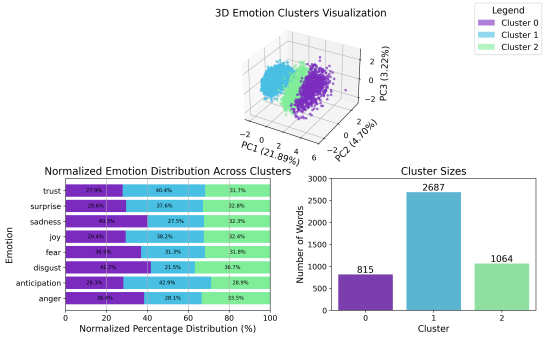

Figure 3: Three-cluster structure of NRC word embeddings in PCA-reduced space

directions and facilitate clustering. Figure 3 visualizes the embedding space after PCA reduction. We see three clusters that form well-separated, dense regions of semantically related NRC words. A Gaussian Mixture Model (GMM) (Dempster et al., 1977) with three components partitions the space into clusters (violet, blue, and green), assigning each word to one of three dense regions representing distinct semantic distributions within the lexicon.

Cluster-based Normalization To account for varying density across clusters, we normalized similarity scores by first assigning each candidate and lexicon word to a cluster using PCA+GMM. We then computed similarity only among words within the same cluster—penalizing similarities between words from different clusters—and optionally rescale these scores to a standardized range within each cluster.

Our normalization follows the equation below. Let s denote the raw similarity between two embeddings. For each cluster i, let μ_i and σ_i be the precomputed mean and standard deviation of raw similarity scores, respectively. The normalized sim-

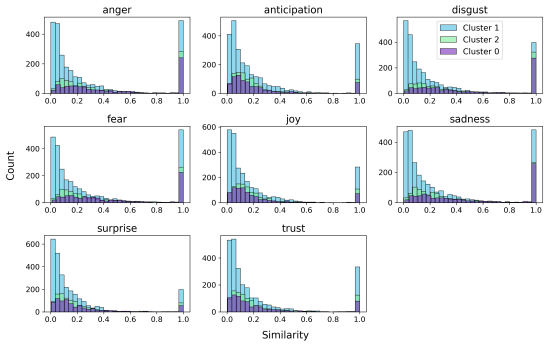

Figure 4: Aligned similarity distributions after cluster-based normalization.

ilarity for cluster i is computed as

$$s^i_{\text{norm}} = \text{clip}\left(\frac{\frac{s-\mu_i}{\sigma_i} + \sigma_c}{2\sigma_c}, 0, 1\right),$$

where $\sigma_c = 3$ is a fixed scaling constant derived from the three-sigma rule. This mapping constrains values within ± 3 standard deviations of each cluster's mean to the [0,1] range, clipping only extreme outliers.

Finally, weighting by the cluster probabilities p_i, the overall normalized similarity is given by

$$s_{\text{final}} = \text{clip}\left(\sum_{i=1}^{K} p_i \, s^i_{\text{norm}}, 0, 1\right),$$

with $K = 3$ being the number of clusters.

This normalization ensures that the similarity threshold used for lexicon expansion is applied consistently across different density regions.

Figure 4 illustrates the similarity distributions after cluster normalization. The per-cluster curves now align in peak and shape, indicating that normalization harmonizes density differences across clusters and preserves emotion-category separability during expansion.

3.2. Threshold-based Lexicon Expansion

For each candidate word, we identified the nearest lexicon word for each emotion. If a calibrated similarity score exceeded a threshold θ, we assigned the corresponding emotion to the candidate word. We varied θ from 0.05 to 0.95 in steps of 0.05 and recorded the number of new emotion assignments, the number of unique words expanded, diversity metrics (such as Hamming distance and entropy), and emotion-specific expansion counts.

Figure 5 shows the impact of varying the similarity threshold θ on lexicon expansion. As θ decreases, more emotion assignments occur, resulting in broader lexicon expansion. Crucially, however, lower thresholds risk introducing false positives by assigning emotions to words with only

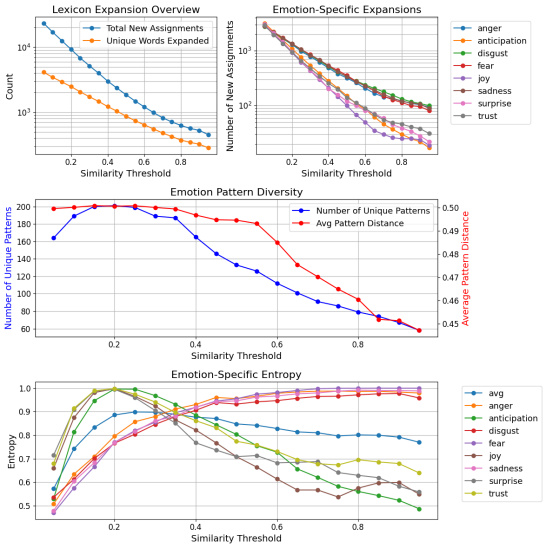

Figure 5: Analysis of lexicon expansion under varying similarity thresholds.

marginal similarity. Conversely, higher thresholds ensure more conservative expansion but might miss subtle emotion associations. The plots display total and unique new assignments (on a log scale), the number of unique emotion patterns, the average Hamming distance between binary emotion vectors, and trends in emotion-specific entropy. Importantly, the process is run on 9-dimensional feeling vectors (with 0 for absence and 1 for presence of an emotion), which change with each threshold.

3.3. Threshold Selection and Category Separation

The threshold calibration analysis revealed a clear trade-off between expansion coverage and structural coherence. Lower similarity thresholds led to rapid growth of the lexicon by assigning more emotion labels, but this also introduced semantic drift and weakened the separation between emotion categories. When the threshold drops below approximately 0.15, the extended lexicon reaches near-total corruption, with substantial cross-category contamination and loss of identifiable emotion boundaries. Higher thresholds, in contrast, restrict expansion and maintain strong category separation but limit lexical coverage. The optimal range lies where coverage is maximized without compromising the structural integrity of the emotion space. Within this range, emotion clusters remain distinct, preserving the categorical topology and ensuring that the expanded lexicon stays semantically consistent and empirically stable across clusters.

4. Corpora and Statistics

4.1. Corpora

Argumentative Microtexts (Part 1) The AMT1 corpus, developed by (Peldszus and Stede, 2015). Originally written in German, the texts have been professionally translated into English, as well as Russian (Fishcheva and Kotelnikov, 2019), and more recently, Persian Abkenar and Stede (2024). The corpus is annotated with complete argumentation tree structures.

Argumentative Microtexts (Part 2) The second part of the AMT2 corpus, developed by Skeppstedt et al. (2018) through crowdsourcing. It follows the same annotation approach as the original corpus, ensuring consistency. A notable difference in this corpus is the inclusion of implicit claims.

Preprocessing Details: In order to adapt the task for stance classification on the Microtext corpora, we projected the *Pro* label in the XML files to *For* and the *Opp* label to *Against*. Since the number of samples was limited, we concatenated both corpora to create a combined dataset.

UKP The UKP dataset comprises comments on various controversial topics (Stab et al., 2018). Compared to other datasets, it contains the largest number of argumentative segments. The UKP comments are generally longer and often exhibit more complex argumentative structures, which increases the difficulty of stance classification.

Preprocessing Details: To maintain consistency with other corpora, we focused on two labeling steps: support or oppose segment, labeled as *For* or *Against*, and we ignore the *no argument* category.

Persusasive Essay The PE corpus contains 402 argumentative essays written by English learners in response to specific prompts. Collected by (Stab and Gurevych, 2016) from an online source, each essay is annotated with an argumentation graph. Essays begin with a guiding question and present a major claim, typically at the end, supported by evidence that may include sub-arguments. Some sentences serve a non-argumentative function, offering background or minor elaboration.

Preprocessing Details: The stance classification model is designed to predict the argumentative stance of each segment. The stance of a claim is specified through its stance attribute, as provided in the original corpus. We treat the major claim as the target topic and focus only on the stance of individual claims toward this target.

IBM Argumentative Structure Dataset The IBM Argumentative Structure Dataset, originally introduced by Aharoni et al. (2014), includes claims about controversial topics. Each claim expresses support or opposition to its topic, and stance labels (*Pro*/*Con*) Bar-Haim et al. (2017).

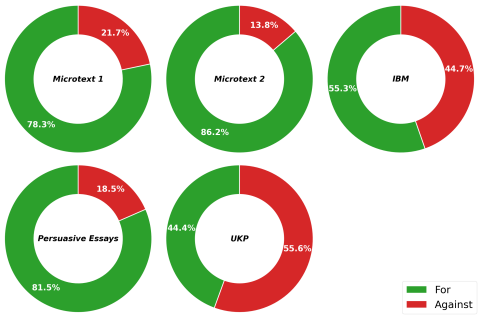

Figure 6: Distribution of stance labels in the preprocessed corpora.

Preprocessing Details: We preprocessed the IBM Debater dataset by extracting, for each debate, topics, claims, and stances, converting *pro* and *con* labels to *For* and *Against*.

4.2. Corpora Statistics

Table 1 presents the statistics for each corpus, including domain, size and average number of tokens and Figure 6 presents the distribution of their stance labels.

5. Experiments and Discussion

5.1. Model Architecture

Let $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ denote an input argumentative text segment consisting of n tokens, and let t represent the associated controversial topic. Our model aims to predict the stance $y \in \{0, 1\}$ (*For* or *Against*) of the text toward the topic.

We employ a pre-trained BERT encoder to obtain contextualized representations of the input. Following standard practice, we encode the text with explicit topic information by constructing the input as: "[CLS] Topic: t [SEP] Argument: $\mathbf{x}$ [SEP]". This concatenation allows the model to jointly encode both the argumentative content and its target topic within the same semantic space.

The BERT encoder produces a sequence of hidden states $\mathbf{H} \in \mathbb{R}^{n \times d}$, where $d = 768$ is the hidden dimension. We extract the [CLS] token representation $\mathbf{h}_{\text{CLS}} \in \mathbb{R}^d$ as the aggregate sentence embedding.

When additional emotion NRC features are available, we concatenate them with the [CLS] representation. Let $\mathbf{f}_{\text{emo}} \in \mathbb{R}^{d_e}$ denote the d_e-dimensional emotion feature vector derived from the extended NRC lexicon (eNRC). The combined representation $\mathbf{h}_{\text{combined}} = [\mathbf{h}_{\text{CLS}}; \mathbf{f}_{\text{emo}}] \in \mathbb{R}^{d+d_e}$ is then passed through the classifier for binary classification.

Corpora	Full Name	Domain	Size	Argument Types	Stance Labels	#Topics	Average #Tokens
AMT1	Microtext Corpus 1	Short argumentative texts	112 texts, 576 segments (ADUs)	Premise, Claim	Pro/Opp	19	14
AMT2	Microtext Corpus 2	Short argumentative texts	116 texts, 614 segments (ADUs)	Premise, Claim	Pro/Opp	34	13
PE	Persuasive Essays	Student essays	402 essays, 7,116 segments(ADUs), (1506 used)	Premise, Claim, Major Claim	For/Against	402	16
IBM	IBM Debater Claim Stance	Debate arguments	2,394 segments	Claim	Pro/Con	55	12
UKP	UKP Sentential Argument Mining	Web discourse	25,492 segments	Argumentative segment	For/Against/No	8	24

Table 1: Overview of original datasets before preprocessing and conversion. Actual subset sizes used in experiments are indicated where applicable.

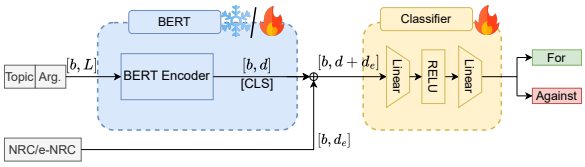

Figure 7: Illustration of the NASCA model. The *Topic* and *Arg.* are composed into a single prompt and fed into the BERT model. Arg. stands for Argumentative segment.

5.2. Experimental Setup

We implement three variants of NASCA. **noNRC** refers to not using the extra NRC features. **NRC** refers to using the conventional NRC features. **eNRC** refers to using the expanded NRC features proposed by us. The classifier consists of two hidden layers. Each hidden layer applies ReLU activation followed by dropout regularization with rate $p = 0.1$. The final output layer produces a single logit for binary classification. As baselines we choose majority class (**MajC**) and a state-of-the-art LLM **Qwen2.5-7B** (Qwen et al., 2025). For Qwen2.5-7B, we ask the model with the prompt: *Please classify the following argument as "support" or "agains" to the given topic.* We split each dataset into train, test, and validation, and report the F_1 score on the test split using the best checkpoint on the validation split.

5.3. Results and Discussion

Table 2 presents the macro F_1 scores for stance classification across four datasets. The results demonstrate that incorporating emotion-lexicon features consistently improves performance, with eNRC achieving the best results across all datasets. Notably, both the majority class baseline (MajC) and the state-of-the-art LLM Qwen2.5-7B show limited effectiveness, with Qwen2.5-7B performing particularly poorly on UKP and PE, highlighting the challenge of stance classification in argumentative contexts even for large language models.

Our expanded emotion lexicon (eNRC) shows substantial improvements over both baseline approaches. The gains are particularly pronounced on the Persuasive Essays (PE) dataset, where eNRC achieves 62.9% F_1, representing a 6.0 percentage point improvement over the noNRC baseline and a 2.6 percentage point gain over the original NRC. This significant boost reflects PE's longer essay format, which provides richer emotional content and more opportunities for the expanded lexicon to capture subtle emotional expressions. Similarly, the combined Microtext corpus (AMT1+2) shows strong improvement (+4.4 points over noNRC), suggesting that even in shorter argumentative texts, expanded emotion coverage enhances stance detection. The UKP dataset shows the highest absolute performance (68.4% F_1 with eNRC), benefiting from both its larger size and the systematic emotion signal enhancement provided by the expanded lexicon. The superior performance of eNRC over NRC stems from its ability to address fundamental coverage limitations in manually curated emotion lexicons. By leveraging DistilBERT embeddings and cluster-based normalization, our expansion method systematically identifies semantically similar words that carry similar emotional connotations. This approach proves particularly effective in the argumentation domain, where authors employ diverse emotional vocabulary to persuade and influence readers.

While our primary goal was not to surpass state-of-the-art (SOTA) results, but to ensure consistent performance across corpora and assess the effect of NRC lexicons on argumentative stance classification, our NASCA model with eNRC still achieved competitive results. As shown in Table 2, it reached macro F_1 scores of 65.9% on IBM (prev. SOTA: 64.5% Bar-Haim et al., 2017) and 68.4% on UKP (prev. SOTA: 63.2% Reimers et al., 2019). Results marked with † indicate cases where our model exceeded prior SOTA. For AMT1+2, no prior study evaluated both corpora jointly, and for PE, previous work used different task formulations.

	IBM	UKP	PE	AMT1+2
MajC	35.6	35.7	44.9	45.3
Qwen2.5-7B	36.2	30.7	38.9	62.3
noNRC	64.7	67.3	56.9	64.1
NRC	64.8	65.4	60.3	63.4
eNRC	**65.9**†	**68.4**†	**62.9**	**68.5**

Table 2: Comparison of eNRC with other methods, majority class baseline (MajC), and Qwen2.5-7B on our corpora. We report the macro F_1 scores as the evaluation metric.

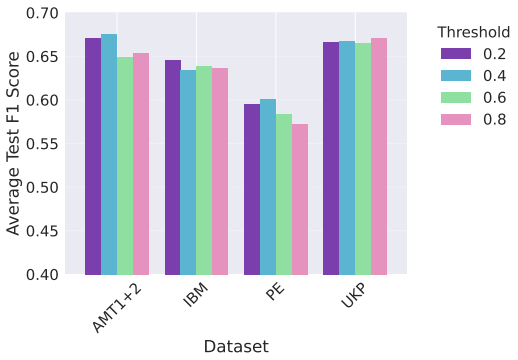

Figure 8: Threshold analysis of eNRC. We report the macro F_1 scores using eNRC features with different thresholds.

Sensitivity on eNRC Threshold. Figure 8 reveals how the similarity threshold θ used during eNRC construction affects stance classification performance across all five datasets. This threshold determines the aggressiveness of lexicon expansion, with lower values assigning emotions to words with weaker similarity to NRC entries and higher values maintaining stricter semantic criteria. The results demonstrate that $\theta = 0.4$ emerges as the most robust choice across datasets. This threshold achieves an effective balance between expanding emotion vocabulary coverage and maintaining semantic precision.

Different datasets exhibit varying degrees of sensitivity to threshold selection, reflecting their distinct linguistic characteristics. The Persuasive Essays dataset shows pronounced sensitivity, with performance peaking sharply at $\theta = 0.4$ and declining more steeply at both extremes. This pattern aligns with PE's rich emotional language—the dataset requires expanded coverage to capture diverse emotional expressions, yet precision remains critical to avoid diluting the signal with false emotional associations. In contrast, IBM and UKP maintain relatively stable performance across a wider threshold range (0.2–0.8), suggesting these more formal argumentative texts rely on core emotional vocabulary that is already well-represented in the original NRC.

Figure 9: Performance improvements after unfreezing the BERT model. We report the change in macro F_1 score.

The Argumentative Microtexts datasets (AMT1+2) display moderate threshold sensitivity with optimal performance around $\theta = 0.4$, suggesting that the balanced expansion can effectively capture emotional signals in short texts where every word can carry substantial information.

Impact of BERT Weights on the Performance. Figure 9 presents the performance difference between models trained with frozen versus unfrozen BERT encoders across all datasets. While unfreezing BERT parameters typically yields performance gains in stance classification, our results reveal that this benefit diminishes substantially when incorporating eNRC features. For the eNRC variant, the performance differential remains within ±2.2 F_1 points across all datasets, compared to larger gaps observed with baseline approaches.

This reduced sensitivity to BERT fine-tuning suggests that the expanded emotion lexicon provides sufficiently discriminative features for stance classification, allowing the model to achieve competitive performance through classifier-level adaptation alone. The eNRC features appear to capture complementary information that partially compensates for the representational limitations of a frozen encoder. Notably, on the PE dataset, eNRC shows only +2.2 points improvement when unfreezing BERT, while noNRC and NRC show -1.3 and +3.6 points respectively, indicating that emotion features stabilize performance and reduce dependence on full model fine-tuning.

This finding has practical implications for deployment scenarios where computational resources are constrained, as it demonstrates that effective stance classification can be achieved with frozen BERT when augmented with high-quality emotion features. The ability to maintain performance without fine-tuning the entire encoder reduces training time, memory requirements, and computational costs while preserving model effectiveness.

6. Conclusion

In this article, we presented eNRC, an expansion of the NRC lexicon, integrated into an end-to-end approach for argumentative stance classification on controversial topics. Previous studies suffered from several limitations: they often focused on non-argumentative texts, were restricted to a single domain, or addressed only one corpus. In contrast, we reformulated five well-known corpora covering a range of controversial topics to evaluate our approach comprehensively.

To the best of our knowledge, this was the first study to systematically incorporate an emotion lexicon into this subtask, demonstrating its potential to improve performance. Additionally, we introduced this lexicon expansion through a rigorous and reproducible process, making it useful for other researchers. We also proposed a neural stance classification framework that could be easily adapted to other topics and domains with minimal modification.

Both the expanded NRC lexicon and the argumentative stance classification framework are publicly available to the research community.

7. Acknowledgements

We thank our colleagues in Innovations department of the Bundesdruckerei GmbH and the Hasso Plattner Institute for providing us with the opportunity to freely work on our research topics. Thank you for fostering an environment that encourages innovation and academic growth. The authors also sincerely thank the anonymous reviewers for their thoughtful recommendations that significantly improved this paper.

8. Bibliographical References

Mohammad Yeghaneh Abkenar and Manfred Stede. 2024. Neural mining of persian short argumentative texts. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)@ LREC-COLING 2024*, pages 30–35.

Mohammad Yeghaneh Abkenar, Manfred Stede, and Stephan Oepen. 2021. Neural argumentation mining on essays and microtexts with contextualized word embeddings.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Basit Ali, Sachin Pawar, Girish Palshikar, and Rituraj Singh. 2022. Constructing a dataset of support and attack relations in legal arguments in court judgements using linguistic rules. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 491–500.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261.

Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz, and Noam Slonim. 2021. Project debater apis: Decomposing the ai grand challenge. *arXiv preprint arXiv:2110.01029*.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. 2016. Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2023. Exploring the potential of large language models in computational argumentation. *arXiv preprint arXiv:2311.09022*.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

Elena Chistova, Philipp Cimiano, Shohreh Haddadan, Gabriella Lapesa, and Ramon Ruiz-Dolz, editors. 2025. *Proceedings of the 12th Argument mining Workshop*. Association for Computational Linguistics, Vienna, Austria.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Adrian de Wynter and Tommy Yuan. 2023. I wish to have an argument: Argumentative reasoning in large language models. *arXiv preprint arXiv:2309.16938*.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and et al. Angela Fan. 2024a. The llama 3 herd of models.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024b. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Adam Robert Faulkner. 2014. *Automated classification of argument stance in student essays: A linguistically motivated approach with an application for supporting argument summarization*. City University of New York.

Marc Feger, Katarina Boland, and Stefan Dietze. 2025. Limited generalizability in argument mining: State-of-the-art models learn datasets, not arguments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23900–23915, Vienna, Austria. Association for Computational Linguistics.

Irina Fishcheva and Evgeny Kotelnikov. 2019. Cross-lingual argumentation mining for russian texts. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 134–144. Springer.

James B Freeman. 2011. *Argument Structure:: Representation and Theory*, volume 18. Springer Science & Business Media.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Freya Hewett, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. 2019. The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103, Florence, Italy. Association for Computational Linguistics.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*. Columbia Univ., New York, NY (United States).

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2020. Stance prediction for contemporary issues: Data and experiments. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 32–40, Online. Association for Computational Linguistics.

Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

Dilek Küçük and Fazli Can. 2020. Stance detection. *ACM Computing Surveys (CSUR)*, 53:1 – 37.

John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024a. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Zhenyu Li, Sunqi Fan, Yu Gu, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang. 2024b. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18608–18616.

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to do with an airport? mining arguments in the german online participation project tempelhofer feld. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 144–153.

Xing Lu. 1998. *Rhetoric in ancient China, fifth to third century, BCE: A comparison with classical Greek rhetoric*. Univ of South Carolina Press.

William C Mann and Sandra A Thompson. 1987. Rhetorical structure theory: Description and construction of text structures. In *Natural language generation: New results in artificial intelligence, psychology and linguistics*, pages 85–95. Springer.

Saif Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836, Dubrovnik, Croatia. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34.

Saif M Mohammad. 2025. Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. *arXiv preprint arXiv:2503.23547*.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

Ivan Namor and Manfred Stede. 2019. Mining italian short argumentative texts. In *Proceedings of the 5th Workshop on Argument Mining*.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Marco Passon, Marco Lippi, Giuseppe Serra, and Carlo Tasso. 2018. Predicting the usefulness of amazon reviews using off-the-shelf argumentation mining. *arXiv preprint arXiv:1809.08145*.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, volume 2, pages 801–815.

Robert Plutchik. 1965. What is an emotion? *The Journal of psychology*, 61(2):295–303.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2017. The penn discourse treebank: An annotated corpus of discourse relations. In *Handbook of linguistic annotation*, pages 1197–1217. Springer.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Gil Rocha, Christian Stab, Henrique Lopes Cardoso, and Iryna Gurevych. 2018. Cross-lingual argumentative relation identification: from english to portuguese. In *Proceedings of the 5th Workshop on Argument Mining, 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*.

Robin Schaefer and Manfred Stede. 2019. Improving implicit stance classification in tweets using word and sentence embeddings. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 299–307. Springer.

Edward Schiappa. 2010. Keeping faith with reason.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Neelabh Sinha, Vinija Jain, and Aman Chadha. 2024. Evaluating open language models across task types, application domains, and reasoning types: An in-depth experimental analysis. *arXiv preprint arXiv:2406.11402*.

Maria Skeppstedt, Andreas Peldszus, and Manfred Stede. 2018. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 155–163.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77.

Christian Stab and Iryna Gurevych. 2016. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43:619–659.

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *arXiv preprint arXiv:1802.05758*.

Manfred Stede. 2020. Automatic argumentation mining and the role of stance and sentiment. *Journal of Argumentation in Context*, 9(1):19–41.

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*, volume 40 of *Synthesis Lectures in Human Language Technology*. Morgan & Claypool.

Manfred Stede, Jodi Schneider, and Graeme Hirst. 2019. *Argumentation mining*. Springer.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Tony Street and Nadja Germann. 2021. Arabic and Islamic Philosophy of Language and Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2021 edition. Metaphysics Research Lab, Stanford University.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. 2024. Hermes 3 technical report. *arXiv preprint arXiv:2408.11857*.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia. Association for Computational Linguistics.

Stephen E. Toulmin. 1958. The uses of argument. *Philosophy*, 34(130):244–245.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. 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 1338–1352, Online. Association for Computational Linguistics.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: Technical papers*, pages 1680–1691.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 interactive demonstrations*, pages 34–35.

Michael Wojatzki and Torsten Zesch. 2016. ltl. uni-due at semeval-2016 task 6: Stance detection in social media using stacked classifiers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 428–433.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Samira Zad, Joshuan Jimenez, and Mark Finlayson. 2021. Hell hath no fury? correcting bias in the nrc emotion lexicon. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 102–113.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197, Online. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

9. Language Resource References

Saif M. Mohammad and Peter D. Turney. 2013. *NRC Emotion Lexicon*.

ELRA. PID https://catalogue.elra.info/en-us/repository/search/?q=NRC+emotion+lexicon. Language Resource.