ISBN 978-9934-619-83-0

Understanding LLM Performance Gaps

Strategic Implications of Stance Detection and Sentiment Analysis in Small Languages

PREPARED AND PUBLISHED BY THE

NATO STRATEGIC COMMUNICATIONS CENTRE OF EXCELLENCE

ISBN: 978-9934-619-83-0

Authors: Jurgita Kapočiūtė-Dzikienė, Mantas Vaškevičius, Tadas Sadzevičius,
Gundars Bergmanis-Korats, Joshua Chia Tee Hiang
Project Manager: Joshua Chia Tee Hiang
Content Editor: Merle Anne Read
Design: Inga Ropša

Cover image generated by Gemini AI.

Riga, April 2026

NATO STRATCOM COE
11b Kalnciema iela,
Riga, LV1048, Latvia
stratcomcoe.org
@stratcomcoe

Understanding LLM Performance Gaps

Strategic Implications of Stance Detection and Sentiment Analysis in Small Languages

Contents

Executive Summary

Large language models (LLMs) have become essential tools for tasks critical to understanding the information environment, including analysing public discourse, detecting information manipulation, and understanding sentiment towards geopolitically sensitive topics. However, their performance varies significantly across languages. Two previous studies by the NATO Strategic Communications Centre of Excellence progressively documented these disparities. The first, *Narrative Detection and Topic Modelling in the Baltics* (Barbu, Banerjee, Isupova, and Zeng, 2024), established that natural language processing (NLP) capabilities for Baltic languages remain significantly underdeveloped, finding that while named entity recognition (NER) was reasonably supported, critical downstream tasks such as relationship extraction and plot discovery remained largely unexplored. The second study, *AI in Support of StratCom: The Use and Evaluation of Large Language Models in Less Widely Used Official EU Languages* (Barbu, Banerjee, Lim, and Zīvere, 2025), expanded the scope considerably by evaluating five contemporary LLMs (*GPT-4.0, Mistral Nemo, Mistral Large, Llama 3.1, and Gemini Pro*) on three strategic NLP tasks: narrative detection, topic modelling, and aspect-based sentiment analysis (ABSA), across both English and Latvian. The report confirmed a persistent performance gap between high-resource and low-resource languages and revealed that while English outputs showed higher fluency, coherence, and structural accuracy, they were still not sufficient to replace human annotators, especially in zero-shot settings. In Latvian, model outputs were generally less complete, with frequent entity misclassification, such as mislabelling geopolitical organisations like NATO as locations rather than actors, and weaker thematic depth across all evaluated layers.

The present report builds directly upon the findings of both preceding studies by shifting the focus to empirically measuring how modern LLMs perform on two additional operationally critical tasks (*stance detection and sentiment analysis*) across English, Lithuanian, and Russian, targeting politically sensitive entities: Ukraine, Russia, NATO, the USA, and China. This report also responds to conclusions from preceding studies by evaluating next-generation adaptation strategies such as fine-tuning and retrieval-augmented generation (RAG) to determine whether targeted model adaptation can close the performance gaps that both earlier studies identified as a structural limitation of working with less-resourced languages.

The findings of this study reveal substantial performance disparities, with direct implications for strategic communication operations in the Baltic region and Eastern Europe. Consistent with the 2025 report's observation that LLM capabilities degrade markedly for less-resourced languages, this study documents that Russian-language analysis underperforms English by up to 9 percentage points, and that fine-tuned lightweight models can outperform larger proprietary systems relying on prompting alone. The 2025 report also noted that models such as Mistral Large and GPT-4.0 performed most consistently in English but showed marked inconsistencies in Latvian, particularly in sentiment analysis where models struggled to align specific sentiments with their respective aspects. This study extends those findings to stance detection, confirming the same pattern across an additional language pair. Together, the three reports establish a clear progression: from identifying what NLP tools and resources exist for Baltic languages (2024), to benchmarking LLM performance on foundational StratCom tasks in English and Latvian (2025), to empirically quantifying how well current LLMs perform on stance detection and sentiment analysis across English, Lithuanian, and Russian, and demonstrating that targeted adaptation strategies can meaningfully close the gaps that all three studies have identified.

KEY TAKEAWAYS

- Russian-language analysis underperforms by up to 9%: **approximately 1 in 3 stance classifications may be incorrect.**
- Fine-tuned lightweight models (7-9B parameters) outperform GPT-4.1 when the latter relies only on prompting.
- Adversaries may deliberately exploit gaps, concentrating operations where detection is weakest.
- Investment in model adaptation closes performance gaps at lower cost than premium API services.

Strategic Context

The information environment in the Baltic region operates across multiple linguistic domains simultaneously. Practitioners monitoring narratives about the Russia–Ukraine conflict, NATO activities, or great power competition must analyse content in English, local Baltic languages, and Russian. The assumption that AI-powered tools perform uniformly across these languages is both common and flawed.

DEFINITION

Stance detection identifies whether a text supports, opposes, or remains neutral towards a specific entity. Unlike **sentiment analysis** (general emotional polarity), stance detection reveals *directed attitudes* towards specific actors, which is critical for tracking narratives about military alliances, governments, or conflicts.

Research Questions

This research is limited to in-target stance detection and text-level sentiment analysis. Within this scope, the study compares the most promising LLMs and LLM-driven adaptation strategies to evaluate their accuracy. The research addresses three core questions:

RQ1 Which LLMs (or their versions) are most suitable for performing stance detection and sentiment analysis on politically sensitive multilingual data?

RQ2 How do different adaptation strategies (RAG and lightweight fine-tuning) affect the performance of LLMs in these tasks?

RQ3 How does stance detection and sentiment analysis accuracy vary across languages (English, Lithuanian, Russian) and politically charged targets (Ukraine, Russia, NATO, the USA, China)?

Research Methodology

Research Framework

The research addresses a joint target extraction and multi-target stance detection problem, where the same text may reference multiple targets. Figure 1 illustrates the analytical framework: input text is processed through stance detection to identify positions towards each target (favour, against, none, or N/A if the target is not present), while simultaneously undergoing sentiment analysis to determine overall emotional polarity (positive, negative, or neutral).

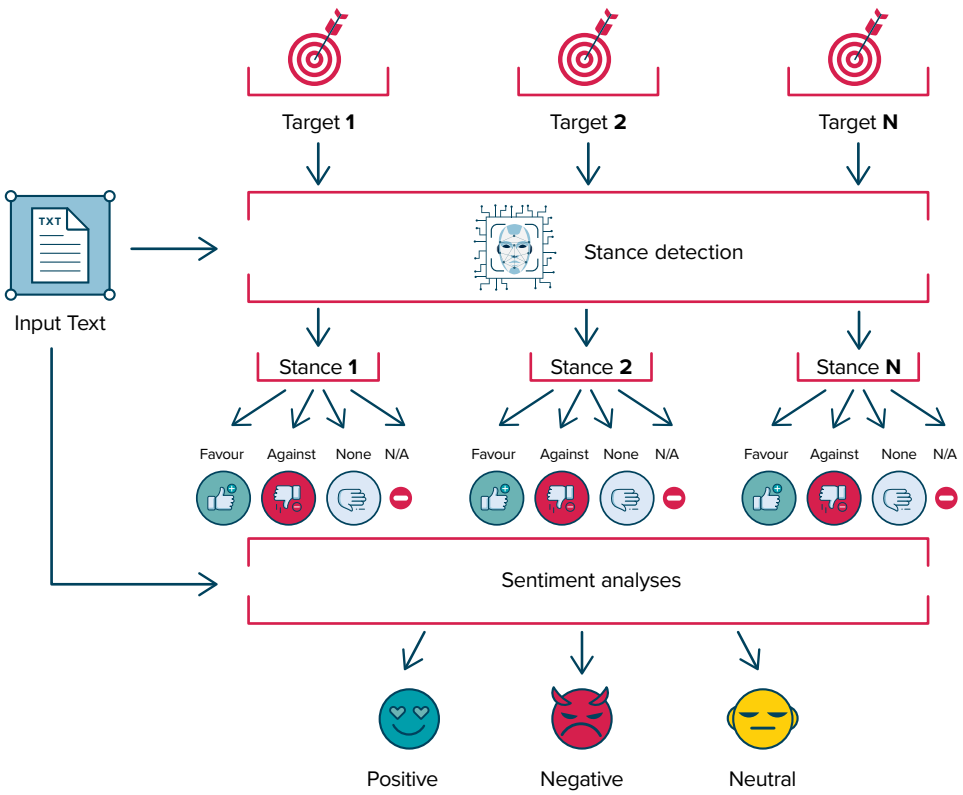

FIGURE 1. Research schema showing in-target, multi-target stance detection with integrated text-level sentiment analysis

Data Collection and Annotation

The research employed a rigorously constructed multilingual dataset comprising approximately 1000 texts per language (English, Lithuanian, and Russian), harvested from Lithuanian media sources including social media comments and news outlets. These texts were selected for their political relevance and subjective nature, capturing authentic discourse on geopolitically sensitive topics.

Dataset component	Details
Languages	English, Lithuanian, Russian (~1000 texts each)
Targets	Ukraine, Russia, NATO, the USA, China
Stance labels	Favour, against, none (per target)
Sentiment labels	Positive, negative, neutral (per text)
Annotators	Seven trained annotators (three per language), with reconciliation
Sources	Lithuanian media: social media comments, news outlets

TABLE 1. Research Parameters

To facilitate consistent annotation, target-related keywords were automatically highlighted using GPT-5, capturing both morphological variations and semantic dependencies (e.g. 'Kremlin' and 'Putin' as proxies for Russia, 'AFU' for Ukraine's Armed Forces). Cases where annotators disagreed underwent additional review and discussion to reach consensus via a multi-annotator approach with reconciliation to ensure label quality while acknowledging the inherent subjectivity of stance and sentiment classification.

Models and Techniques Evaluated

Models evaluated

- Gemma-2-9B-Instruct (Google)
- Llama-3-8B-Instruct (Meta)
- DeepSeek-LLM-7B-Chat
- Qwen2.5-7B-Instruct (Alibaba)
- GPT-4.1 (OpenAI) proprietary

Techniques compared

- Zero-shot prompting: Direct task instructions only
- RAG (retrieval-augmented): Five similar labelled examples via Language Agnostic BERT Sentence Embeddings (LaBSE)
- Fine-tuning (LoRA): Parameter-efficient weight adaptation

Experimental Design

Three training-testing scenarios were employed to assess cross-linguistic generalisation: **in-language** (train and test on the same language: EN→EN, LT→LT, RU→RU); **cross-lingual** (train on one language, test on another: EN→LT, EN→RU, etc.); and multilingual (combine training data from all three languages, test on each separately). Each fine-tuning experiment was conducted three times with different random seeds to calculate 95% confidence intervals, ensuring statistical robustness.

Performance was measured using **weighted-F1** for stance detection (accounting

for class imbalance across targets) and **macro-F1** for sentiment analysis (treating all classes equally). These metrics provide more reliable assessment than simple accuracy, particularly for imbalanced datasets where majority-class prediction would artificially inflate scores.

Key Research Findings

RQ1: Model Performance Comparison

FINDING: Gemma 2:9B emerged as the dominant performer across most experimental configurations, consistently achieving the highest scores for stance detection regardless of language or training scenario. For sentiment analysis, results showed slightly more variation, with DeepSeek-7B achieving the best English results and Llama-3.1-8B excelling on Lithuanian, though Gemma 2:9B remained the top performer for Russian.

Performance by language (best achieved scores):

English	Lithuanian	Russian
Stance: 0.760 ± 0.007 Sentiment: 0.703 ± 0.132 ~1 in 4 errors	Stance: 0.717 ± 0.046 Sentiment: 0.787 ± 0.098 ~1 in 4 errors	Stance: 0.695 ± 0.138 Sentiment: 0.676 ± 0.040 ~1 in 3 errors ⚠

TABLE 2. Model Performance Comparison by language

RQ2: Impact of Adaptation Strategies

FINDING: Fine-tuning consistently outperformed both RAG and prompting, with improvements ranging from 0.02 to 0.38 F1 points over RAG. Crucially, fine-tuned lightweight open-weight models (7-9B parameters) outperformed the much larger proprietary GPT-4.1 when the latter relied on prompting or RAG approaches.

⚡ TECHNIQUE PERFORMANCE HIERARCHY

Fine-tuning ▶ Best performance across all configurations

RAG ▶ Moderate improvement (+0.07 to +0.20 F1 over prompting)

Prompting ▶ Baseline performance only

RAG showed minimal benefit for sentiment analysis compared to stance detection as sentiment classification is already well represented in LLM pre-training data. Stance detection towards specific geopolitical entities, being more specialised and context dependent, benefits more substantially from retrieved examples.

RQ3: Cross-Language and Target Variation

FINDING: Russian consistently showed the weakest performance, with both lower absolute scores and higher variability (±0.138 confidence interval vs ±0.007 for English). English maintained the most stable performance across configurations. Notably, Lithuanian achieved surprisingly strong sentiment analysis results, actually outperforming English.

⚠ OPERATIONAL INTERPRETATION

For a team processing 1000 Russian-language posts daily, the performance gap translates to 300+ additional misclassified items compared to English, which is significant enough to obscure emerging narratives or generate substantial analytical noise. The wider confidence interval further indicates that Russian-language results are less predictable across different model configurations.

Analysis of stance correlations across targets revealed predictable alignment patterns: texts favouring Ukraine typically opposed Russia, while pro-Russian stances often correlated with negative positions towards NATO. These correlation patterns varied by language community, reflecting different geopolitical perspectives in English, Lithuanian, and Russian discourse.

Operational Application within a Hypothetical Scenario

⚠ THE COST OF CAPABILITY GAPS

A monitoring cell tracks discourse during a military exercise. Its LLM-powered dashboard flags 847 English posts expressing a negative stance towards NATO, triggering analyst review and rapid response.

Simultaneously 1200 Russian-language posts expressing anti-NATO sentiment circulate in Russian-speaking communities. Due to the 9% performance gap, only 680 are correctly classified. The remaining 520, nearly half, are coded neutral or missed entirely. The coordinated nature goes undetected.

Two weeks later, polling shows a measurable shift in attitudes. Post hoc analysis reveals the missed campaign, but the information window has closed.

Strategic Implications

Adversary Exploitation

The performance disparities documented in this research reflect broader patterns in AI development that sophisticated adversaries understand. **Hostile actors may deliberately concentrate influence operations in linguistic spaces where automated detection performs worst.** Russian-language content targeting Baltic Russian-speaking minorities represents precisely this vulnerability.

Resource Allocation

The finding that fine-tuned smaller models outperform larger commercial models suggests a cost-effective path. Rather than premium API access, organisations may achieve better results investing in fine-tuning open-weight model – providing greater control over data handling and operational security.

Recommendations

IMMEDIATE Within weeks
→ **Audit current tool performance** by language against labelled samples
→ **Implement differentiated verification** such as 50% higher review rate for Russian content
→ **Adjust alert thresholds** for Russian to compensate for under-detection

NEAR-TERM Within months
→ **Develop language-specific fine-tuned models** (Gemma 2:9B recommended)
→ **Build labelled training datasets** specific to regional targets and entities
→ **Establish performance benchmarks** for ongoing evaluation

STRATEGIC Ongoing
→ **Pursue regional data-sharing partnerships** with Baltic and NATO partners
→ **Integrate capability gaps into threat modelling** to war-game adversary exploitation
→ **Advocate for multilingual AI research priorities** in policy forums

Conclusion

The promise of LLM-powered analysis must be tempered by recognition of capability gaps. Current LLMs represent a substantial advancement over previous generations as they are more capable, multilingual, and more accessible than ever before. Yet they still fall short of their promise when applied to small languages and politically sensitive content. The performance disparities documented in this research demonstrate that out-of-the-box deployment, even of state-of-the-art models, delivers unreliable results in precisely the linguistic environments most relevant to Eastern European security.

These gaps create both operational challenges and exploitable vulnerabilities. However, they are not immutable. Investment in careful data engineering and model fine-tuning can substantially close performance differences, as evidenced by the 38% improvement achieved through LoRA adaptation in this study. The path forward is not to wait for better foundation models, but to actively adapt existing ones.

Critically this research underscores the necessity of extensive per-language and per-geography performance comparison. Assumptions that a model performing well in English would transfer to Lithuanian or Russian are demonstrably false. Organisations deploying LLM-powered tools for strategic communication must conduct rigorous, language-specific benchmarking before operational use, and repeat such evaluations as models and information environments evolve.

For strategic communication practitioners, the imperative is clear: acknowledge current limitations, implement compensating verification measures, invest in language-specific fine-tuning, and build evaluation frameworks that match the multilingual reality of modern information competition.

Bibliography

Abercrombie, G., & Batista-Navarro, R. (2022). *Policy-focused stance detection in parliamentary debate speeches*. Northern European Journal of Language Technology 8, 45–54. Linköping University Electronic Press.

Ahn, H., Jeong, D., & Park, E. (2025). *SABER: Integrating sentiment and stance detection for climate change discourse on social media*. SSRN.

Alahmadi, K., Alharbi, S., Chen, J., et al. (2025). *Generalizing sentiment analysis: A review of progress, challenges, and emerging directions*. Social Network Analysis and Mining 15, 45.

Ali, M., & Hassan, N. (2022). *A survey of computational framing analysis approaches*. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 9335–48). Association for Computational Linguistics.

Barbu, Banerjee, Isupova, and Zeng. (2024). *Narrative Detection and Topic Modelling in the Baltics*.

Barbu, Banerjee, Lim, and Zīvere (2025). *AI in Support of StratCom: The Use and Evaluation of Large Language Models in Less Widely Used Official EU Languages*.

Brown, T., Mann, B., Ryder, N., et al. (2020). *Language models are few-shot learners*. Advances in Neural Information Processing Systems 33, 1877–1901.

Burnham M. (2025). *Stance detection: A practical guide to classifying political beliefs in text*. Political Science Research and Methods 13(3), 611–28.

Chen, X., Liu, B., Hu, H., et al. (2025). *Integrating graph neural networks and large language models for stance detection via heterogeneous stance networks*. Applied Sciences 15(11), 5809.

Chetviorkin, I., & Loukachevitch, N. (2013). *Evaluating sentiment analysis systems in Russian*. In J. Piskorski, L. Pivovarova, H. Tanev, & R. Yangarber (Eds.), Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (pp. 12–17). Association for Computational Linguistics.

Chuang, Y.-S. (2023). *Tutorials on stance detection using pre-trained language models: Fine-tuning BERT and prompting large language models*. arXiv preprint arXiv:2307.15331.

DeepSeek-AI, Bi, X., Chen, D., et al. (2024). *DeepSeek LLM: Scaling open-source language models with longtermism*. arXiv.

DeepSeek-AI, Guo, D., Yang, D., et al. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv*.

Derczynski, L., Bontcheva, K., Liakata, M., et al. (2017). *SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours*. In S. Bethard, M. Carpuat, M. Apidianaki, et al. (Eds.), Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 69–76). Association for Computational Linguistics.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). *QLoRA: Efficient finetuning of quantized LLMs*. In Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023) (Article 441, pp. 1–28). Curran Associates Inc.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers) (pp. 4171–86). Association for Computational Linguistics.

Feng, F., Yang, Y., Cer, D., et al. (2022). *Language-agnostic BERT sentence embedding*. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 878–91). Association for Computational Linguistics.

Gambini, M., Senette, C., Fagni, T., & Tesconi, M. (2024). *Evaluating large language models for user stance detection on X (Twitter)*. Machine Learning 113(10), 7243–66.

Gemma Team, Kamath, A., Ferret, J., et al. (2025). *Gemma 3 technical report*. arXiv.

Gemma Team, Riviere, M., Pathak, S., et al. (2024). *Gemma 2: Improving open language models at a practical size*. arXiv.

Gera, P., & Neal, T. (2025). *Deep learning in stance detection: A survey*. ACM Computing Surveys 58(1), Article 26.

Glandt, K., Khanal, S., Li, Y., et al. (2021). *Stance detection in COVID-19 tweets*. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), 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) (pp. 1596–1611). Association for Computational Linguistics.

Grattafiori, A., Dubey, A., Jauhri, A., et al. (2024). *The Llama 3 herd of models*. arXiv.

Gül, İ., Lebret, R., & Aberer, K. (2024). *Stance detection on social media with fine-tuned large language models*. arXiv.

Hanselowski, A., PVS, A., Schiller, B., et al. (2018). *A retrospective analysis of the Fake News Challenge stance-detection task*. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics (pp. 1859–74). Association for Computational Linguistics.

Hu, E. J., Shen, Y., Wallis, P., et al. (2022). *LoRA: Low-rank adaptation of large language models*. In Proceedings of the International Conference on Learning Representations (ICLR). OpenReview.net.

Huang, W., & Yang, J. (2024). *A multi-stance detection method by fusing sentiment features*. Applied Sciences 14(9), 3916.

Jamadi Khiabani, P., & Zubiaga, A. (2025). *Cross-target stance detection: A survey of techniques, datasets, and challenges*. Expert Systems with Applications 283, Article 127790.

Kang, L., Yao, J., Du, R., et al. (2025). *A stance detection model based on sentiment analysis and toxic language detection*. Electronics 14(11), 2126.

Kapočiūtė-Dzikienė, J., Damaševičius, R., & Woźniak, M. (2019). *Sentiment analysis of Lithuanian texts using traditional and deep learning approaches*. Computers 8(1), 4.

Kapočiūtė-Dzikienė, J., Krupavičius, A., & Krilavičius, T. (2013). *A comparison of approaches for sentiment classification on Lithuanian internet comments*. In J. Piskorski, L. Pivovarova, H. Tanev, & R. Yangarber (Eds.), Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (pp. 2–11). Association for Computational Linguistics.

Kim, Y. (2014). *Convolutional neural networks for sentence classification*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746–51). Association for Computational Linguistics.

Kochkina, E., Liakata, M., & Augenstein, I. (2017). *Turing at SemEval-2017 Task 8: Sequential approach to rumour stance classification with Branch-LSTM*. In S. Bethard, M. Carpuat, M. Apidianaki, et al. (Eds.), Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 475–80). Association for Computational Linguistics.

Lan, X., Gao, C., Jin, D., & Li, Y. (2024). *Stance detection with collaborative role-infused LLM-based agents*. Proceedings of the International AAAI Conference on Web and Social Media 18(1), 891–903.

Lavrouk, A., Ligon, I., Zheng, J., et al. (2024). *Stanceosaurus 2.0 – Classifying stance towards Russian and Spanish misinformation*. In R. van der Goot, J. Bak, M. Müller-Eberstein, et al. (Eds.), Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024) (pp. 31–43). Association for Computational Linguistics.

Lewis, P., Perez, E., Piktus, A., et al. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. Advances in Neural Information Processing Systems 33, 9459–74.

Li, A., Liang, B., Zhao, J., et al. (2023). *Stance detection on social media with background knowledge*. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023) (pp. 15703–17). Association for Computational Linguistics.

Li, A., Zhao, J., Liang, B., et al. (2025). *Mitigating biases of large language models in stance detection with counterfactual augmented*

calibration. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 7075–92). Association for Computational Linguistics.

Li, Y., Vasilakes, J., Zhao, Z., & Scarton, C. (2025). *SCRum-9: Multilingual stance classification over rumours on social media*. arXiv preprint arXiv:2505.18916.

Liu, S., Guo, L., Mays, K., et al. (2019). *Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence*. In M. Bansal & A. Villavicencio (Eds.), Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL) (pp. 504–14). Association for Computational Linguistics.

Lozhnikov, N., Derczynski, L., & Mazzara, M. (2018). *Stance prediction for Russian: Data and analysis*. In International Conference in Software Engineering for Defence Applications (pp. 176–86). Springer.

Lüüsi, L., Kangur, U., Chakraborty, R., & Sharma, R. (2024). *Political stance detection in Estonian news media*. In Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages (IWCLUL 2024) (pp. 12–28). Association for Computational Linguistics.

Mets, M., Karjus, A., Ibrus, I., & Schich, M. (2024). *Automated stance detection in complex topics and small languages: The challenging case of immigration in polarizing news media*. PLOS One 19(4), e0302380.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., et al. (2016). *SemEval-2016 Task 6: Detecting stance in tweets*. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 31–41. Association for Computational Linguistics.

Motyka, D., & Piasecki, M. (2024). *Target-phrase zero-shot stance detection: Where do we stand?* In V. V. Kravets, J. M. S. Vázquez, & V. V. Kalyuzhny (Eds.), Computational Science – ICCS 2024: 24th International Conference, Malaga, Spain, July 2–4, 2024, Proceedings, Part II (pp. 34–49). Springer-Verlag.

Mou, Y., Morstatter, F., Ferrara, E., & Liu, H. (2022). *A two stage adaptation framework for frame detection*. In Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022) (pp. 2981–93). International Committee on Computational Linguistics.

OpenAI, Achiam, J., Adler, S., et al. (2024). *GPT-4 technical report*. arXiv.

OpenAI, Hurst, A., Lerer, A., et al. (2024). *GPT-4o system card*. arXiv.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques*. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002) (pp. 79–86). Association for Computational Linguistics.

Pu, Q., Huang, F., Li, F., et al. (2025). *Integrating emotional features for stance detection aimed at social network security: A multi-task learning approach*. Electronics 14(1), 186.

Qin, L., Chen, Q., Feng, X., et al. (2024). *Large language models meet NLP: A survey*. arXiv preprint arXiv:2405.12819.

Rahman Jim, J., Talukder, M. A. R., Malakar, P., et al. (2024). *Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review*. Natural Language Processing Journal 6, 100059.

Rusnachenko, N., Golubev, A., & Loukachevitch, N. (2024). *Large language models in targeted sentiment analysis*. arXiv preprint arXiv:2404.12342.

Singh, V. K., Mohankumar, P., & Kamal, A. (2023). *Fin-STance: A novel deep learning-based multi-task model for detecting financial stance and sentiment*. In Proceedings of the 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1–6). IEEE.

Socher, R., Perelygin, A., Wu, J., et al. (2013). *Recursive deep models for semantic compositionality over a sentiment treebank*. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1631–42). Association for Computational Linguistics.

Somasundaran, S., & Wiebe, J. (2009). *Recognizing stances in online debates*. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on

Natural Language Processing of the AFNLP (pp. 226–34). Association for Computational Linguistics.

Student. (1908). *The probable error of a mean*. Biometrika 6(1), 1–25.

Sun, K., Luo, X., & Luo, M. Y. (2022). *A survey of sentiment analysis based on pretrained language models*. In 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 1239–44). IEEE.

Vileikytė, B., Lukoševičius, M., & Stankevičius, L. (2024). *Sentiment analysis of Lithuanian online reviews using large language models*. arXiv preprint arXiv:2407.19914.

Wagner, S. S., Behrendt, M., Ziegele, M., & Harmeling, S. (2024). *The power of LLM-generated synthetic data for stance detection in online political discussions*. arXiv preprint arXiv:2406.12480.

Wang, S., Zhang, Y., Li, J., & Teng, C. L. (2025). *VSDQ: A comprehensive vaccine stance detection quadruple dataset for analyzing vaccine discussions on social media*. In Y. Zhang et al. (Eds.), Health Information Processing (CHIP 2024) (Communications in Computer and Information Science, Volume 2432). Springer.

Wei, W., Zhang, X., Liu, X., et al. (2016). *pkudblab at SemEval-2016 Task 6: A specific convolutional neural network system for effective stance detection*. In S. Bethard, M. Carpuat, D. Cer, et al. (Eds.), Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 384–88). Association for Computational Linguistics.

Yan, Y., Sun, S., Tang, Z., et al. (2025). *Collaborative stance detection via small-large language model consistency verification*. arXiv preprint arXiv:2502.19954.

Yang, A., Li, A., Yang, B., et al. (2025). *Qwen3 technical report*. arXiv.

Yang, R., Ma, J., Gao, W., & Lin, H. (2025). *LLM-enhanced multiple instance learning for joint rumor and stance detection with social context information*. ACM Transactions on Intelligent Systems and Technology 16(3), Article 58.

Zarrella, G., & Marsh, A. (2016). *MITRE at SemEval-2016 Task 6: Transfer learning for stance detection*. In S. Bethard, M. Carpuat, D. Cer, et al. (Eds.), Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 458–63). Association for Computational Linguistics.

Zhang, Z., Li, Y., Zhang, J., & Xu, H. (2024). *LLM-driven knowledge injection advances zero-shot and cross-target stance detection*. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), 371–78. Association for Computational Linguistics.

Research Source: Vytautas Magnus University (Lithuania) & NATO Strategic Communications Centre of Excellence (Latvia)

Prepared and published by the
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel. Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org