Boosting up the Sentiment Analysis Models' Accuracy by Blending Multi-label Learning with a Large Sentiment Lexicon

Dimitrios Kokkinakis[1]

[1] *University of Gothenburg, Box 200, 405 30, Gothenburg, Sweden*

Abstract

This study compares sentiment analysis approaches for Swedish texts using a manually annotated gold-standard dataset. Two methods were examined: i) a multi-label sentiment classifier trained for Swedish, and ii) the Swedish version of VADER, a lexicon-based tool that computes sentiment scores from a vocabulary of polarity-weighted words. The analysis also examined agreement and disagreement between the two methods, with a focus on mixed or context-dependent sentiment. Results indicate that the multi-label classifier aligns more closely with human judgments, especially for medium- or long-text segments with complex or subtle emotional tones. VADER, while prone to errors in idiomatic or nuanced expressions, performs reliably on short, informal utterances, offering computational efficiency and transparency. A hybrid approach combining classifier predictions with lexicon-based scores was investigated to leverage their complementary strengths. Findings underscore the value of rigorous evaluation against human annotations and highlight strategies to improve sentiment analysis in under-resourced languages such as Swedish.

Keywords

sentiment analysis, multi-label classifier, multi-class model, lexicon-based method (VADER)

1. Introduction

Sentiment analysis is a central task in Natural Language Processing (NLP) with broad applications in social media, customer feedback, and automated content evaluation. Despite recent advances driven by machine learning and large language models, lexicon-based methods remain relevant, especially for analyzing short texts or when annotated data is limited. The reason that lexicon-based sentiment methods remain valuable is because they contribute to *transparency* (each word's contribution is explicit, unlike LLMs' opaque decision-making), *adaptability* (they can be easily adapted with custom lexicons for specific domains without retraining or large datasets), and *practicality* (suitable for real-time or resource-constrained settings) — making them complementary to large language models (LLMs) rather than obsolete. This study addresses sentiment analysis for Swedish texts by evaluating two complementary approaches: a multi-label sentiment classifier and a lexicon-based tool (VADER; "Valence Aware Dictionary and sEntiment Reasoner"; [1]). Using a manually annotated gold-standard dataset, we assess their performance in terms of, among other metrics, accuracy, precision, recall, and F1-score, while also leveraging lexicon scores to enhance classifier predictions. Results indicate that the classifier is more effective for longer or syntactically complex sentences, whereas the lexicon-based method better captures mixed or context-dependent sentiments and shorter sentences.

Building on these complementary strengths, we propose a hybrid strategy that integrates both approaches, leading to improved robustness and accuracy for sentiment analysis in under-resourced languages. Previous research in English has explored combining lexicon-based sentiment analysis with machine learning or transformer-based models [2]; such studies demonstrate that hybrid methods can mitigate individual model limitations and improve robustness. The present work builds on this idea, applying and evaluating a similar approach in Swedish, a language with more limited NLP resources.

The rest of the paper is organized as follows: Section 2 describes the dataset and resources; Section 3 details the methodology; Section 4 presents the results and discusses future research directions.

Huminfra Conference 2025, Stockholm, 12-13 November 2025.

dimitrios.kokkinakis@svenska.gu.se (Dimitrios. Kokkinakis)

2. Dataset and linguistic resources

To enable a rigorous evaluation of sentiment analysis methods for Swedish texts, we employed a manually annotated gold-standard dataset as the primary benchmark. This dataset consists of short textual units [*n=2017*], such as social media posts, blogs, passages from Swedish newspapers, and some user-generated comments, each labelled with the standard sentiment categories *positive*, *neutral* and *negative*. The original dataset and annotation process can be found in [3], some minor corrections [n≈20] and adjustments were imposed, after manual inspection, to increase the reliability of the dataset[2]. For instance, sentiment annotation was changed for some data entries: *Hur fina vänner är inte det?* eng. "*What great friends, aren't they?*" from negative in the original gold standard, to positive; *Det blev en pizza och en god vattenmelon-juice istället men det var trevligt.* eng. "*It ended up being a pizza and a tasty watermelon juice instead, but it was nice.*" from neutral in the gold standard, to positive. Moreover, several duplicate entries were removed, ensuring that only unique records remained, e.g. *Kaos på stand up-klubben igår*. eng. "*Chaos at the stand up club yesterday*". The dataset is rather balanced, and Figure 1 (left) shows the distribution of the entries with respect to the three sentiment classes.

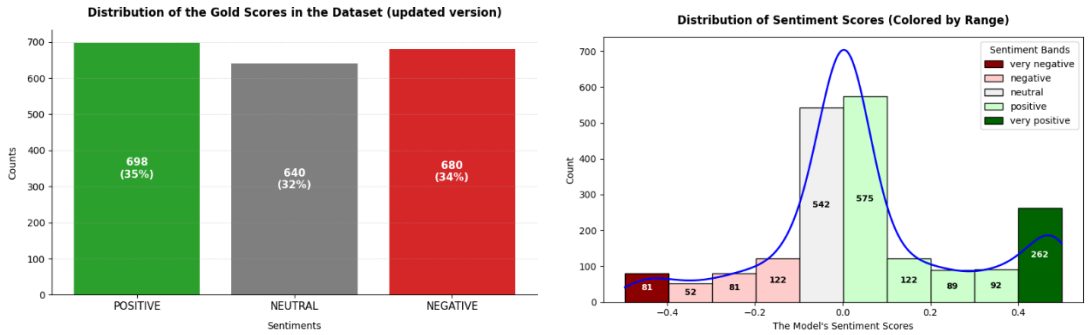

Figure 1: Distribution of the gold dataset's sentiments – the updated version (left) and the distrbution of the KBLab's model output in this same dataset (right) – in this plot, the "very positive" (>= 0.48) and "very negative" (<= -0.48) values are based basically on the obtained scores – *not* highlighted explicitely by the model's output as "very positive" or "very negative".

For the initial annotation we use the '`robust-swedish-sentiment-multiclass`' model from the National Library of Sweden (KBLab[3]). According to the creators of the model [4], the model is a release of a robust, multi-label sentiment classifier finetuned on Megatron-BERT-large-165K. The model was trained on approximately 75K Swedish texts from multiple linguistic domains and datasets. The model addresses gaps in Swedish sentiment analysis by including a neutral category and training across diverse datasets beyond reviews. Using data from reviews, Twitter, news, immigration discourse, and translated texts, it achieves strong generalization and accuracy (0.80 multiclass, 0.88 binary), making it more robust than earlier Swedish sentiment models. In addition, we applied a lexicon-based approach on the output of the previous model. Specifically, the VADER sentiment lexicon, adapted and extended for Swedish (svVADER; [5]), containing polarity scores associated with words and multiword expressions, was used. VADER ignores word context, especially when word order or distant lexical items intervene in multi-word expressions. Nonetheless, VADER has a negation identification

[2] The column *sentimentannotation.csv* from the original gold standard file: https://raw.githubusercontent.com/richi1998/Evaluating-Lexicon-Based-Models-versus-BERT-for-Sentence-Level-Sentiment-Ana-lysis-in-Swedish/refs/heads/main/koden/**Data_svm.csv** was used for the evaluation exercise. It was manually reviewed, slightly updated, and subsequently used in the experiment. The resulting gold file, renamed *updatedGoldDataset.csv* can be found here: https://github.com/DimitrisKokkinakis/swedish-notebooks/blob/main/textual-resources/HiC-2025/**updatedGoldDataset.csv**.

[3] KBLab's blog post (https://kb-labb.github.io/posts/2023-06-16-a-robust-multi-label-sentiment-classifier-for-swedish/) describe the model as "multi-label" because the underlying architecture could assign multiple sentiment labels simultaneously. However, the training data and released checkpoints use single-label annotations, and the model outputs one dominant class per text (i.e., the highest-probability label). Therefore, the classifier is a multi-class model, predicting one sentiment label per sentence (positive, negative, or neutral), rather than a true multi-label setup.

mechanism to shift polarity under certain circumstances. The enhanced Swedish lexicon[4] includes more than 50,000 entries, encompassing several thousand multi-word expressions, compared to 5,501 entries in the original translation[5] provided in [7].

Together, these resources provide the foundation for both the learning approach (multi-label classifier) and the knowledge-based strategy (lexicon-driven sentiment scoring). They also allow us to explore hybrid methods that combine lexical information with model-based predictions, thereby addressing the limitations of working with under-resourced languages.

3. Methodology, experimental design and results

Performance was quantified using various standard metrics for sentiment analysis evaluation. These metrics collectively provide a comprehensive assessment of classification performance. Accuracy offers an overall measure of correctness, while the Matthews Correlation Coefficient (MCC) captures the balance between true and false classifications, making it particularly informative under class imbalance; however, this metric is less informative in the present case, as the dataset is relatively balanced, but it was included for completeness and comparability with related studies. Precision and Recall quantify, respectively, the proportion of correctly identified positive predictions and the ability to retrieve all relevant instances, with their harmonic mean expressed as the F1 Score. The macro-averaged metrics treat each class equally, reflecting performance across categories regardless of frequency, whereas the micro-averaged metrics aggregate all instances to emphasize overall system performance relative to the gold-standard annotations (see Figure 3).

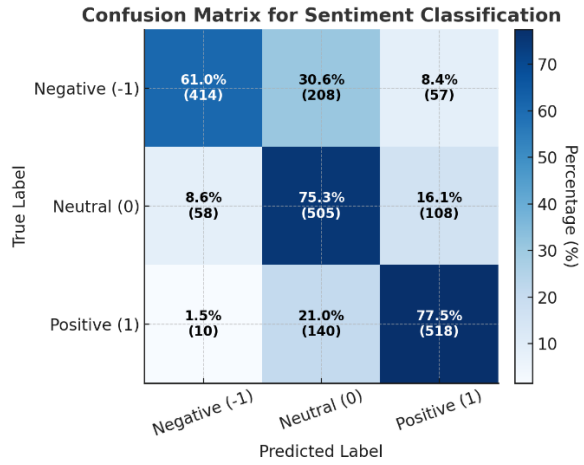

Figure 2: The confusion matrix of the KBLab's sentiment classifier

Figure 2 on the left shows clearly how the KBLab sentiment classifier performed across the three sentiment categories. The darker diagonal cells indicate stronger agreement with the gold standard, while the lighter off-diagonal cells highlight areas of confusion (especially between neutral and neighboring classes). The system achieves roughly 72% accuracy with a macro-F_1, indicating a balanced though moderate performance across the three sentiment classes. Precision is highest for negative (0.86) and positive (0.76) sentiments, while neutral (0.59) lags behind, suggesting that the model tends to confuse neutral expressions with polar ones—a typical challenge in multi-class sentiment analysis.

The two resources were applied sequentially, first the KBLab model followed by the svVADER on the dataset entries (rows) where the model did not agree with the gold standard. On the specific gold standard dataset, the model's evaluation metrics were *accuracy* 72% and *MCC* 59.04%. svVADER was applied on the 565 mismatches (rows), i.e. the cases in which the model and gold standard did not agree, and the results on this subset's evaluation metrics were *accuracy* 52.38% and *MCC* 29.69%. From the number of mismatches, 296 rows were assigned the correct sentiment label and, while 269 were assigned an erroneous sentiment. The combined, global accuracy was 86.67% and the MCC 80.18%.

[4] Selected subsets of the svVADER's lexicon have been evaluated using LLMs (ChatGPT) with manual follow-up. Consistent with similar studies [6], ChatGPT performed quite well, suggesting LLMs can effectively support initial annotations and accelerate lexicon development. One of such subset evaluations focused on entries containing the substring 'under', such as ***under****bart* eng. wonderful; *väl****under****byggd* eng. well-founded and ***under****känd* eng. failed (n=250) can be found here: https://github.com/DimitrisKokkinakis/swedish-notebooks/blob/main/textual-resources/HiC-2025/svVADER-vs-LLM-proofOfConcept.xlsx.

[5] https://github.com/marcusgsta/vaderSentiment/tree/master/vaderSentiment (visited 2025-10-27).

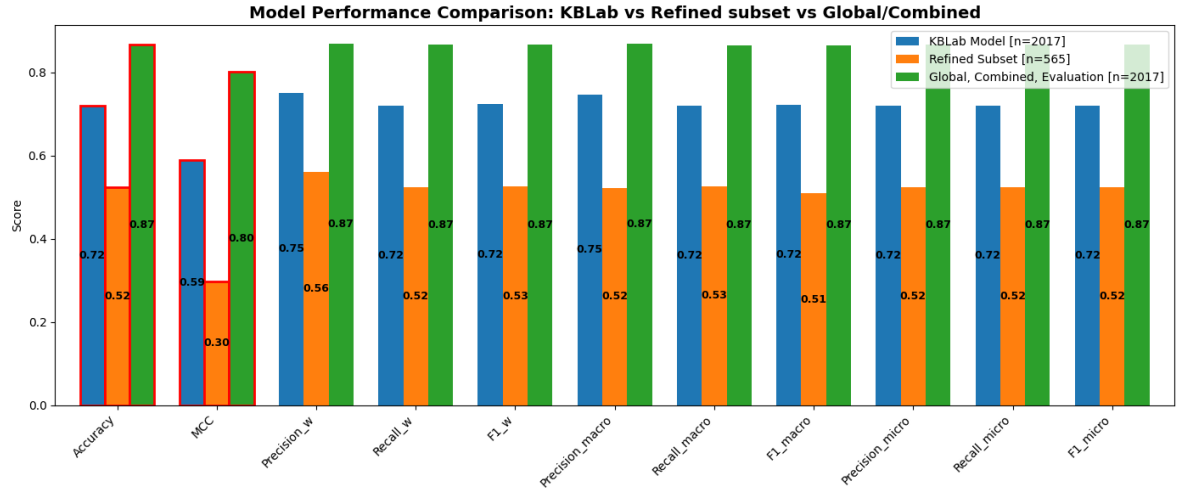

Figure 3: The various metrics used for the evaluation: the KBLab model (the left bar in each bar group), the refined subset, that is the datapoints erroneously annotated by the KBLab model (the middle bar in each bar group) and the combination scores of the two (the right bar in each bar group).

The final mismatched cases (n = 269) predominantly originated from sentences containing long, coordinated constructions anchored by a negative lexical element, which often resulted in scope-related interpretation errors. A further source of misclassification involved brief sentences whose correct sentiment interpretation depended on contextual cues or world knowledge beyond the textual input. Additional inconsistencies were observed in instances featuring figurative, metaphoric, or ironic expressions, where the intended evaluative meaning was challenging for the model to discern. Representative examples of these categories are provided in Table 1.

Table 1
Examples of Sentiment Annotation Challenges

Category	Swedish sentence	English glossing	Description/ Challenge	Gold Annotations
Long coordinated constructions (scope errors)	Bristen på värme, medmänsklighet, trygghet och tröst är total.	The lack of warmth, humanity, security, and comfort is total.	Sentence anchored by a negative word ('bristen på'). Complex coordination leads to potential scope interpretation errors.	–1 (Negative)
Very short input sentences requiring world knowledge & contextual disambiguation	År 1958 kom Sverige tvåa. *or* Det blev ett brons för henne i big air med skidor.	In 1958, Sweden came second. *or* She won a bronze medal in big air with skis.	Short sentences needing external or contextual understanding (e.g., sports results).	+1 (Positive)
Metaphoric sentences	Se hur lång tid det tar innan maskinen äter ditt kort.	See how long it takes before the machine eats your card.	Figurative use of 'eats' (metaphor); potential for literal misinterpretation.	–1 (Negative)
Sarcastic or ironic sentences	Det var någon slags alkoholist-bingo typ.	It was some kind of alcoholic bingo type.	Sarcastic or ironic tone is usually difficult for models to detect.	–1 (Negative)

The motivation for applying the lexicon-based model *only* to sentences where the first model fails is to explore their complementary strengths. This approach helps reveal where each method performs better—for instance, the lexicon model may handle explicit polarity words more effectively, while the machine learning model captures contextual nuances. Although such an approach is not directly applicable in real-world settings where true labels are unknown, it serves as a *proof of concept* demonstrating the potential of selective combination. In practice, this insight could inform confidence-

based or ensemble strategies, where the lexicon model acts as a fallback when the main model's confidence is low.

4. Conclusions and future work

The paper evaluates Swedish sentiment analysis tools — svVADER and the KBLab sentiment model. The results are valuable given the limited Swedish NLP resources. It also proposes a hybrid approach combining both methods to offset their individual weaknesses, a promising idea for other low-resource languages. The hybrid accuracy scores were calculated by sequentially combining the outputs of two sentiment analysis systems—a transformer-based model and a lexicon-based method—and then evaluating their cumulative performance relative to a gold-standard dataset. In essence, the hybrid accuracy scores represent a cumulative metric reflecting the complementary strengths of both models—the contextual robustness of the transformer-based classifier and the lexical sensitivity of the rule-based system—applied in a corrective, sequential manner.

This study has shown that sentiment analysis for Swedish texts can be substantially improved by combining multi-label classification with lexicon-based methods. The multi-label classifier aligned more closely with human annotations, particularly for complex or ambivalent sentences, while the lexicon-based approach contributed transparency and efficiency, capturing nuances in short ones. By integrating both approaches, we achieved a hybrid system with notably higher accuracy (86.67%) than either method independently. These findings confirm that leveraging complementary strengths is especially valuable in under-resourced language contexts, where annotated data remains limited.

Despite the promising results, several challenges remain. The analysis revealed recurrent difficulties in handling coordinated constructions, context-dependent expressions, sarcasm, and irony—phenomena that continue to challenge both statistical and lexicon-driven approaches. Furthermore, reliance on static lexical resources makes it difficult to adapt to emerging vocabulary and evolving usage in social media and digital communication. Future work will address these limitations in several directions. First, expanding the gold-standard dataset with broader domain coverage and richer annotations will improve both training and evaluation. Second, integrating contextual embeddings from large-scale transformer models could enhance the detection of subtle sentiment cues, such as sarcasm [8; 9], irony [10] or metaphors [11]. Third, adaptive or dynamically updated lexicons may mitigate the rigidity of current dictionary-based resources. Finally, applying the hybrid strategy to other under-resourced languages will test its generalizability and contribute to cross-linguistic sentiment analysis research. In sum, this work provides both methodological insights and practical contributions toward more robust, accurate, and interpretable sentiment analysis systems for Swedish and beyond.

Acknowledgements

Work on the article has been supported by The National Language Bank of Sweden (Nationella Språkbanken) and HUMINFRA, the Swedish national infrastructure supporting digital and experimental research in the Humanities and their participating partner institutions, both funded by the Swedish Research Council (2018–2024, contract 2017-00626; 2022–2024, contract 2021-00176).

References

[1] C. Hutto, E. Gilbert. Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI Conference on Web and Social Media, vol. 8, 2014, pp. 216–225.

[2] L. Barros, A. Trifan, and J. L. Oliveira. VADER meets BERT: sentiment analysis for early detection of signs of self-harm through social mining. In: CLEF 2021 – Conference and Labs of the Evaluation Forum, Bucharest, Romania, 2021. https://ceur-ws.org/Vol-2936/.

[3] R. Mansour, E. Nilsson. Evaluating Lexicon-Based Models versus Bert for Sentence Level Sentiment Analysis in Swedish, 2024, https://github.com/richi1998/Evaluating-Lexicon-Based-Models-versus-BERT-for-Sentence-Level-Sentiment-Ana-lysis-in-Swedish

[4] H. Hägglöf, A Robust, Multi-Label Sentiment Classifier for Swedish. June 16, 2023. https://huggingface.co/KBLab/robust-swedish-sentiment-multiclass
[5] D. Kokkinakis, R. Muñoz Sánchez, and Mia-Marie Hammarlin. Scaling-up the Resources for a Freely Available Swedish VADER (svVADER) in: Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), Tórshavn, Faroe Islands. University of Tartu Library, 2023, pp. 667–672.
[6] F.S. Marcondes, A. Gala, M. Rodrigues, J.J. Almeida, P. Novais. Lexicon Annotation with LLM: A Proof of Concept with ChatGPT. In: Quintián, H., et al. Hybrid Artificial Intelligent Systems. HAIS 2024. Lecture Notes in Computer Science(), vol 14858. Springer, Cham. https://doi.org/10.1007/978-3-031-74186-9_16
[7] M. Gustafsson. Sentiment analysis for tweets in Swedish: Using a sentiment lexicon with syntactic rules. Bachelor's thesis. [Online]. Linnaeus University, Sweden. 2020. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1391359&dswid=-8277
[8] E. Riloff, A. Qadir, P. Surve, L. de Silva, N. Gilbert, and R. Huang. Sarcasm as Contrast between a Positive Sentiment and Negative Situation in: Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Seattle, Washington, USA Association for Computational Linguistics, 2013, pp. 704–714. https://aclanthology.org/D13-1066/
[9] Q. Li, Z. Li, W. Liu, X. He, and Y. Pan. Sarcasm-GPT: advancing sarcasm detection with large language models, The Computer Journal, 2025; bxaf055, https://doi.org/10.1093/comjnl/bxaf055
[10] B. Ghanem, J. Karoui, F. Benamara, P. Rosso, and V. Moriceau. Irony Detection in a Multilingual Context. Advances in Information Retrieval. 2020 Mar 24;12036:141–9. https://dl.acm.org/doi/10.1007/978-3-030-45442-5_18.
[11] S. Yang, D. Zhang, J. Ren, Z. Xu, X. Zhang, Y. Song, H. Lin, and F. Xia. Cultural Bias Matters: A Cross-Cultural Benchmark Dataset and Sentiment-Enriched Model for Understanding Multimodal Metaphors, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Vienna, Austria. Association for Computational Linguistics. 2025, pp. 26301–26317. https://aclanthology.org/2025.acl-long.1275/.