A Hybrid Lexicon-Machine Learning Approach to Macedonian Sentiment Analysis

Sofija Kochovska*
kochovskasofija@gmail.com
University of Primorska, UP
FAMNIT
Koper, Slovenia

Branko Kavšek*
branko.kavsek@upr.si
University of Primorska, UP
FAMNIT
Koper, Slovenia
Jožef Stefan Institute
Ljubljana, Slovenia

Jernej Vičič*
jernej.vicic@upr.si
University of Primorska, UP
FAMNIT
Koper, Slovenia

Abstract

This work extends our previous study presented at the ITAT conference, where we introduced a rule-based sentiment analysis system for Macedonian text [7] using handcrafted lexicons and linguistic rules. Building on that foundation, we now integrate these rule-based features with supervised machine learning classifiers, specifically Logistic Regression and Support Vector Machines (SVM), to evaluate their effectiveness in improving sentiment classification. By combining lexicon-derived features such as polarity, intensifiers, diminishers, and negation handling with statistical models, we aim to enhance performance beyond purely rule-based methods. Experimental results show that the hybrid approach improves mean F1 scores from 73.6% (our rule-based baseline) [7] to 86.7% (SVM) and 86.4% (LR), providing a more robust framework for sentiment analysis in under-resourced languages like Macedonian.

Keywords

Sentiment Analysis, Macedonian, Rule-based Approach, Machine Learning, Hybrid Model, Natural Language Processing, Support Vector Machine, Logistic Regression, Low-resource Languages

ACM Reference Format:
Sofija Kochovska, Branko Kavšek, and Jernej Vičič. 2025. A Hybrid Lexicon-Machine Learning Approach to Macedonian Sentiment Analysis. In *Proceedings of Information Society 2024: 27th international multiconference (Information Society 2025).* ACM, New York, NY, USA, 4 pages.

1 Introduction

Sentiment analysis is a core task in natural language processing (NLP), commonly applied to social media, reviews, and feedback analysis. While progress has been substantial for high-resource languages such as English, low-resource languages like Macedonian still face limited availability of annotated corpora, sentiment lexicons, and reliable tools. Macedonian, an Eastern South Slavic language spoken by around 1.6 million people as the official language of North Macedonia, remains under-explored in computational linguistics despite its close relation to Bulgarian and Serbo-Croatian.

In this study, we build on our earlier work presented at the ITAT conference (WAFNL workshop) [7], where we developed a rule-based sentiment analysis system for Macedonian. That work focused on lexicon construction and the integration of modifiers such as intensifiers, diminishers, and polarity shifters. Here, we extend the approach by implementing a hybrid framework that combines rule-based linguistic features with supervised machine learning classifiers. Specifically, we evaluate *Logistic Regression (LR)* and *Support Vector Machines (SVMs)*, using features derived from sentiment lexicons and rule-based weighting schemes.

Our contributions are twofold: (i) we demonstrate how rule-based features enhance the performance of statistical classifiers in a low-resource setting, and (ii) we provide a systematic evaluation of the hybrid approach on Macedonian sentiment data. This study highlights the effectiveness of combining linguistic knowledge with machine learning to improve sentiment detection for under-resourced languages.

2 Related Work

Sentiment analysis has been widely studied, from lexicon-based methods [10, 3] to supervised machine learning and deep learning [9, 1]. Lexicon-based systems use predefined dictionaries and modifiers like intensifiers, diminishers, and negations; they are interpretable and require no large datasets but have limited coverage. Machine learning models achieve higher accuracy with sufficient data but often act as "black boxes." In low-resource languages, limited data and tools pose challenges. Hybrid approaches combining lexicons with statistical learning improve robustness [8, 12]. For the Macedonian language, Jovanoski et al. (2015) [6] conducted early work by compiling sentiment lexicons and manually annotating Twitter datasets for evaluation. They also analysed how different seed lists affect the construction of induced sentiment lexicons, using Macedonian data throughout. This study provides a useful foundation for building or extending sentiment lexicons for Macedonian. Jovanoski et al. (2016) [5] studies how different seed lists affect induced sentiment lexicons and discusses/uses Macedonian within the analysis. More recently, lexicon-based and rule-based systems tailored for Macedonian have been proposed, though their scalability and integration with machine learning remain limited. Uzunova and Kulakov [11] present supervised classification of Macedonian movie reviews; an early non-Twitter benchmark for MK SA. Gajduk and Kocarev [2] classify forum posts from the Kajgana[1] portal into positive/negative/neutral, and report 92 % accuracy with analysis of preprocessing choices (stemming and stop-words). SADEmma 1.0 [4] are an open, 3-class sentiment labels for news across several languages, including Macedonian—useful for domain transfer beyond tweets. Our previous study [7] introduced a curated lexicon of approximately 4,000 sentiment-bearing words,

*These authors contributed equally.

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

[1] Kajgana: https://kajgana.com/

later expanded to a merged lexicon of about 8,000 entries, and evaluated its performance on Macedonian Twitter data.

This work extends the existing literature by systematically integrating lexicon-based features into machine learning classifiers, comparing Logistic Regression and Support Vector Machines (SVMs) for Macedonian sentiment classification. To the best of our knowledge, this is the first study to combine rule-based linguistic insights with supervised classifiers for this language.

3 Methodology

Our approach builds on the framework presented in Kochovska et al. [7], combining lexicon-based rule features with supervised machine learning classifiers. The methodology is designed to handle the challenges of sentiment analysis in Macedonian, a low-resource language, by leveraging linguistic insights alongside statistical learning.

3.1 Lexicon-Based Feature Extraction

We use a set of manually-checked Macedonian lexicons:

- **Positive and Negative Sentiment Lexicons:** Lists of sentiment-bearing words that indicate positive or negative polarity.
- **Intensifiers and Diminishers:** Words that amplify or attenuate the sentiment of neighbouring words (e.g., *very*, *slightly*).
- **Polarity Shifters (Negations):** Words that invert sentiment, such as *not* or *never*, applied within a small context window.
- **Stop-words:** Common Macedonian words that bare minimum meaning are removed to improve feature quality.

Texts are preprocessed to normalize repetitions, remove URLs, mentions, punctuation, and filter stop-words. Each token is then analysed for sentiment, taking into account intensifiers, diminishers, and polarity shifters. The extracted features include:

- Normalized lexicon score
- Counts of positive and negative words
- Counts of intensifiers, diminishers, and negations

These features provide a compact numerical representation of the textual sentiment, which can be effectively used as input to supervised learning models.

3.2 Machine Learning Models

The rule-based features (lexicon score, counts of positive/negative words, intensifiers, diminishers, and negations) serve as input to two classifiers:

- **Logistic Regression (LR):** A linear classifier trained on the rule-based features. The hyper-parameters for intensifier weight (1.5), diminisher weight (0.7), and negation window size (2) were carried over from our previous ITAT study, where we tested 108 different combinations to identify the optimal configuration.
- **Support Vector Machine (SVM):** A linear-kernel SVM trained using the same feature set. In this work, the C parameter was tuned specifically using grid search over values from 0.1 to 5. The best performance was achieved with $C = 0.15$.

The best rule-based configuration used for both models is: intensifier weight = 1.5, diminisher weight = 0.7, negation window = 2, and $\epsilon = 0.30$.

These values control the contribution of linguistic modifiers to the overall sentiment score of a text.

3.3 Dataset Splitting

The Macedonian sentiment dataset used in this study is identical to that from our previous ITAT/WAFNL paper [7]. For machine learning evaluation, we employ stratified 5-fold cross-validation. In each fold, 80% of the data is used for training and 20% for testing, ensuring that the class distribution is preserved across folds. This approach allows robust evaluation of both Logistic Regression and SVM models while leveraging all available data for training and testing across different folds.

3.4 Evaluation Procedure

Both classifiers were evaluated using stratified 5-fold cross-validation, following the methodology presented in [7]. Metrics include F1 scores for positive and negative classes, confusion matrices, and classification reports for all classes. We focus on positive and negative F1 to enable direct comparison with Jovanoski et al. [6], while still monitoring neutral classification.

This approach combines the interpretability of lexicon-based features with the adaptability of machine learning, providing a robust framework for Macedonian sentiment analysis.

4 Results and Evaluation

The hybrid sentiment analysis framework was evaluated on the Macedonian test dataset that we also used for evaluation of the rule-based only approach discussed in the ITAT/WAFNL paper [7], however this time using Logistic Regression (LR) and Support Vector Machine (SVM) classifiers. Both models leveraged the rule-based features described in section 3, with hyper-parameters tuned based on our previous ITAT study for LR and specifically tested on this dataset for SVM.

4.1 Logistic Regression (LR)

Logistic Regression trained on rule-based features demonstrates consistently strong performance, achieving a mean F1 score of 0.864 on positive and negative classes. The per-fold results indicate stable performance across folds, suggesting robustness to variations in the training data (Figure 1).

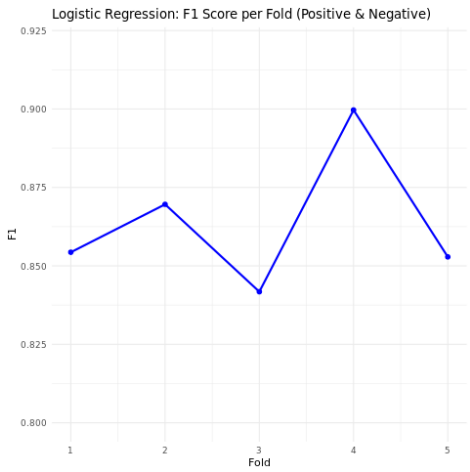

Figure 1: Logistic Regression: F1 score per fold for positive and negative classes.

The confusion matrix (Figure 2) shows that most misclassifications involve neutral and negative instances. Specifically, 43 neutral examples were predicted as negative, and 29 negative examples were labelled as neutral. Positive instances are generally well-separated, with minimal confusion, reflecting the effectiveness of the lexicon-based features. These patterns suggest that LR captures polarized sentiment effectively but struggles with subtle neutral expressions.

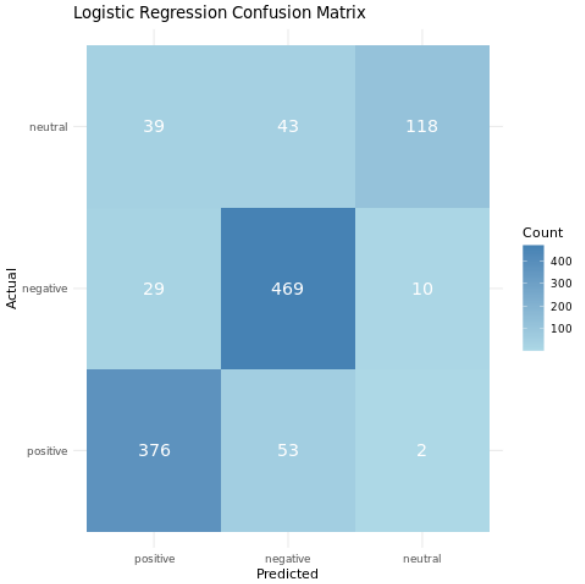

Figure 2: Logistic Regression confusion matrix for all classes.

Overall classification metrics confirm high precision and recall for positive and negative classes (Precision = 0.847 / 0.830, Recall = 0.872 / 0.923, F1 = 0.859 / 0.874), while neutral sentiment remains more challenging (F1 = 0.715). Figure 5 presents these metrics visually, highlighting the differences between classes.

4.2 Support Vector Machine (SVM)

SVM, also trained on the same rule-based features, achieves a slightly higher mean F1 score of 0.867 for positive and negative classes and shows stable per-fold performance (Figure 3). The hyper-parameter $C = 0.15$, selected after testing a range from 0.1 to 5, provided optimal regularization for this dataset.

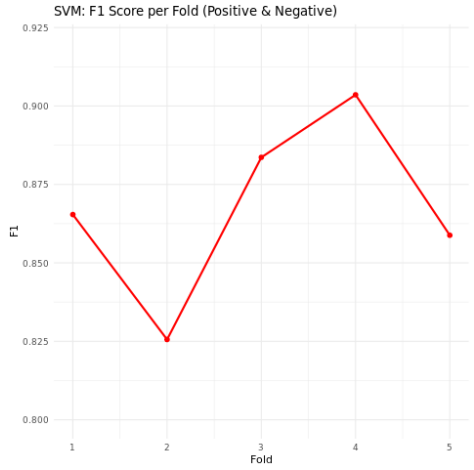

Figure 3: SVM: F1 score per fold for positive and negative classes.

The SVM confusion matrix (Figure 4) exhibits a similar trend to LR: neutral instances are most frequently misclassified, with 54 neutral examples predicted as negative and 38 predicted as positive. SVM shows improved recall for negative instances, correctly identifying 481 of 508 examples, indicating enhanced sensitivity to strong negative cues.

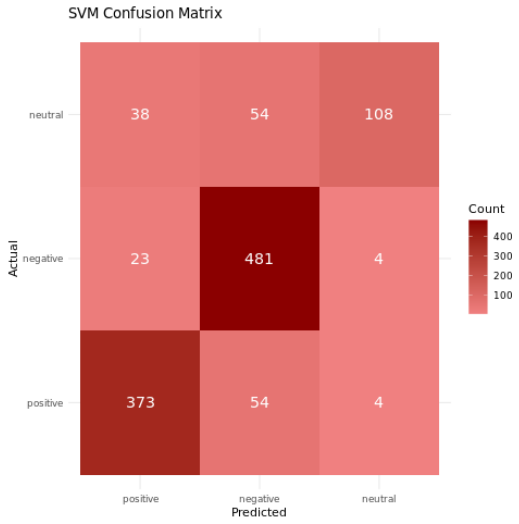

Figure 4: SVM confusion matrix for all classes.

Classification metrics (Figure 5) reinforce these observations: SVM maintains high precision for positive and neutral classes and slightly higher F1 scores for polarized sentiment compared to LR (Positive: F1 = 0.862, Negative: F1 = 0.877, Neutral: F1 = 0.684). This demonstrates that combining rule-based features with SVM improves detection of nuanced sentiment in Macedonian text.

4.3 Discussion

The evaluation shows that our hybrid sentiment analysis framework significantly improves performance over the rule-based system alone. The purely rule-based approach, as reported in our previous ITAT/WAFNL study, achieved a mean F1 score of 73.6% when considering positive and negative classes. By incorporating machine learning classifiers on top of the rule-based features,

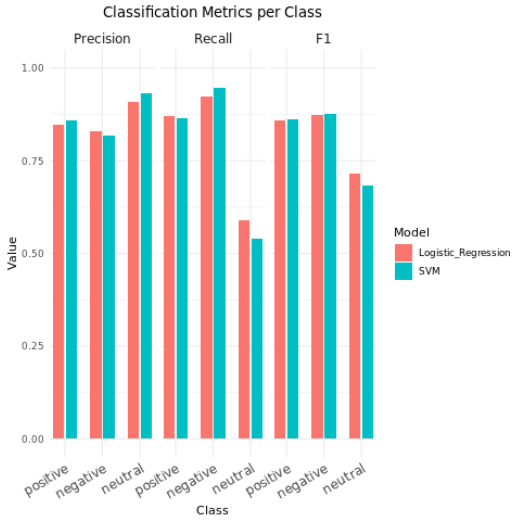

Figure 5: Overall precision, recall, and F1 scores for Logistic Regression and SVM.

both Logistic Regression and SVM achieved substantially higher mean F1 scores, consistently outperforming the rule-based baseline.

The confusion matrices reveal that most misclassifications still occur in the neutral class, which remains inherently difficult due to subtle or context-dependent sentiment cues. Positive and negative instances are effectively distinguished, demonstrating that using lexical knowledge as structured features for statistical learning enhances polarity detection. Furthermore, the per-fold F1 scores for both classifiers indicate stable performance across different data splits, suggesting that the hybrid features generalize well to unseen Macedonian text.

These results confirm that hybrid models can leverage the interpretability of rule-based features while benefiting from the adaptability of machine learning. Future work could explore syntactic and semantic embeddings, as well as attention-based models, to further improve neutral sentiment detection and overall predictive accuracy.

5 Conclusion and Future Work

In this study, we presented a hybrid sentiment analysis framework for Macedonian, combining rule-based lexical features with machine learning classifiers, specifically Logistic Regression and Support Vector Machines. Our results show that the hybrid models substantially outperform the purely rule-based system, which achieved a mean F1 score of 73.6% on positive and negative classes. Both SVM and Logistic Regression consistently improved classification performance, particularly in capturing polarized sentiment, while the rule-based features provided interpretability and robustness. Both classifiers relied exclusively on lexicon-derived rule-based features, ensuring a consistent comparison between LR and SVM.

The evaluation demonstrates that integrating linguistic knowledge with statistical learning is highly effective for under-resourced languages such as Macedonian, where large annotated datasets are scarce. The rule-based component captures explicit and context-modified cues, and the ML models generalize well across folds, ensuring stable performance.

For future work, several directions are planned:

- Incorporating syntactic and semantic embeddings to enhance the representation of context and word meaning, which may improve detection of subtle neutral sentiment.
- Experimenting with attention-based or transformer models to capture long-range dependencies and richer contextual information in Macedonian text.
- Expanding the size and diversity of annotated datasets, including social media posts, reviews, and other user-generated content, to improve coverage and robustness.
- Investigating domain adaptation techniques to allow the models to generalize across different types of Macedonian text, such as formal news articles versus informal social media.
- Integrating additional linguistic cues such as part-of-speech tags or dependency relations to further enhance interpretability and accuracy.

Overall, this work provides a useful foundation for Macedonian sentiment analysis, demonstrating the value of hybrid approaches and opening avenues for more advanced models and richer linguistic feature integration.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*. Jill Burstein, Christy Doran, and Thamar Solorio, editors. Association for Computational Linguistics, Minneapolis, Minnesota, (June 2019), 4171–4186. DOI: 10.18653/v1/N19-1423.

[2] Andrej Gajduk and Ljupco Kocarev. 2014. Opinion mining of text documents written in macedonian language. *arXiv preprint arXiv:1411.4472*. https://arxiv.org/abs/1411.4472 arXiv: 1411.4472 [cs.CL].

[3] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '04). Association for Computing Machinery, Seattle, WA, USA, 168–177. ISBN: 1581138881. DOI: 10.1145/1014052.1014073.

[4] Nikola Ivačič, Andraž Pelicon, Boshko Koloski, Senja Pollak, and Matthew Purver. 2024. News sentiment analysis datasets for serbian, bosnian, macedonian, albanian and estonian (sademma 1.0). CLARIN.SI repository. Version 1.0. (2024). http://hdl.handle.net/11356/1987.

[5] Dame Jovanoski, Veno Pachovski, and Preslav Nakov. 2016. On the impact of seed words on sentiment polarity lexicon induction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, (Dec. 2016), 1557–1567. https://aclanthology.org/C16-1147/.

[6] Dame Jovanoski, Veno Pachovski, and Preslav Nakov. 2015. Sentiment analysis in Twitter for Macedonian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. Ruslan Mitkov, Galia Angelova, and Kalina Bontcheva, editors. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, (Sept. 2015), 249–257. https://aclanthology.org/R15-1034/.

[7] Sofija Kochovska, Branko Kavšek, and Jernej Vičič. 2025. Rule-based sentiment analysis of Macedonian. In *Proceedings of the ITAT 2025: Information Technologies – Applications and Theory* (CEUR Workshop Proceedings). Telgárt, Slovakia.

[8] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: a survey. *Ain Shams Engineering Journal*, 5, 4, 1093–1113. DOI: https://doi.org/10.1016/j.asej.2014.04.011.

[9] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors. Association for Computational Linguistics, Seattle, Washington, USA, (Oct. 2013), 1631–1642. https://aclanthology.org/D13-1170/.

[10] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37, 2, (June 2011), 267–307. DOI: 10.1162/COLI_a_00049.

[11] Vasilija Uzunova and Andrea Kulakov. 2015. Sentiment analysis of movie reviews written in macedonian language. In *ICT Innovations 2014*. Advances in Intelligent Systems and Computing. Vol. 311. Ana Madevska Bogdanova and Dejan Gjorgjevikj, editors. Springer, Cham, 279–288. DOI: 10.1007/978-3-319-09879-1_28.

[12] Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis : a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8, (Jan. 2018). DOI: 10.1002/widm.1253.