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 study extends our previous work on a rule-based sentiment analysis system for Macedonian text [10], which relied on handcrafted lexicons and linguistic rules. We now investigate the integration of these rule-based features with supervised machine learning classifiers, specifically Logistic Regression (LR) and Support Vector Machines (SVM), to improve sentiment classification performance. Lexicon-derived features, including polarity, intensifiers, diminishers, and negation handling, are combined with statistical models to evaluate their impact. Experimental results show that the hybrid approach substantially outperforms the rule-based baseline, increasing the mean F1 score from 73.5% to 86.7% for SVM and 86.4% for LR. Paired t-tests confirm that these improvements are statistically significant ($p < 0.001$), while Wilcoxon tests indicate a strong trend ($p = 0.0625$). These findings demonstrate that integrating rule-based linguistic features with machine learning classifiers provides a robust framework for sentiment analysis in under-resourced languages such as Macedonian.

Keywords

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

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, Serbian and Croatian.

In this study, we build on our earlier work presented at the ITAT conference (WAFNL workshop) [10], 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 approaches [16, 6] to machine learning and deep learning models [15, 2]. Lexicon-based systems rely on dictionaries and modifiers such as 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, hybrid approaches combining lexicon features with statistical learning improve robustness [12, 18].

For Macedonian, Jovanoski et al. [9] compiled sentiment lexicons and manually annotated Twitter datasets, analyzing how seed lists affect induced lexicons. Uzunova and Kulakov [17] classified movie reviews, while Gajduk and Kocarev [4] achieved 92% accuracy on forum posts. The SADEmma 1.0 corpus [7] includes three-class news sentiment labels across languages, but the Macedonian portion has only 198 entries, limiting its usefulness for model training. Our previous work [10] introduced a curated lexicon of 4,000 words, later expanded to 8,000, evaluated on Macedonian Twitter data.

Despite its close relation to Bulgarian, Serbian, and Croatian, Macedonian sentiment analysis remains under-resourced. Comparable studies in Serbian and Slovenian report performance ranging from moderate to high, with F1 or accuracy scores around 76–83% depending on dataset and methodology [11, 8, 13, 3]. These findings indicate that our results align with trends observed in related South Slavic languages. This study extends prior work by integrating lexicon-based features into supervised classifiers, comparing Logistic Regression and SVMs for Macedonian sentiment classification, and, to our knowledge, represents the first combination of rule-based linguistic insights with statistical models for this language.

*These authors contributed equally.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia
© 2025 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2025.sikdd.16

3 Methodology

Our approach builds on the framework presented in Kochovska et al. [10], 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 manually-checked Macedonian lexicons:

- **Positive and Negative Lexicons:** Words indicating positive or negative sentiment.
- **Intensifiers and Diminishers:** Words that amplify or attenuate sentiment (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 words with minimal meaning, removed to improve feature quality.

Texts are preprocessed to normalize repetitions, remove URLs, mentions, punctuation, and stop-words. Each token is analyzed for sentiment considering intensifiers, diminishers, and negations. 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 sentiment suitable for supervised learning models.

3.2 Machine Learning Models

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

- **Logistic Regression (LR):** A linear classifier trained on the rule-based features. Hyperparameters for intensifier weight (1.5), diminisher weight (0.7), and negation window size (2) were adopted from our previous ITAT study, which tested 108 combinations to identify the optimal configuration.
- **Support Vector Machine (SVM):** A linear-kernel SVM trained on the same features. The C parameter was tuned via grid search (0.1–5), with the best performance at $C = 0.15$.

The selected rule-based configuration 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 [10]. 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

We evaluated the rule-based baseline and hybrid classifiers using stratified 5-fold cross-validation to ensure balanced sentiment class representation. For each fold, models were trained on 80% of the data and tested on 20%, repeating the process across five splits to obtain stable estimates.

Performance was measured primarily with F1 scores for positive and negative classes [10], enabling direct comparison with Jovanoski et al. [9]. Confusion matrices and full classification reports were also generated to evaluate performance on all three classes, including neutral, highlighting improvements in polarity detection and challenges in handling neutral sentiment.

Statistical significance of improvements was assessed using paired t-tests and Wilcoxon signed-rank tests on per-fold F1 scores.

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 [10], 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 hyperparameters 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).

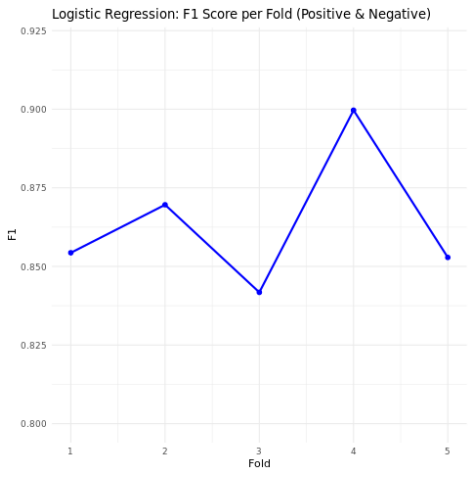


Figure 1: Logistic Regression: F1 score per fold for positive and negative classes.

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.

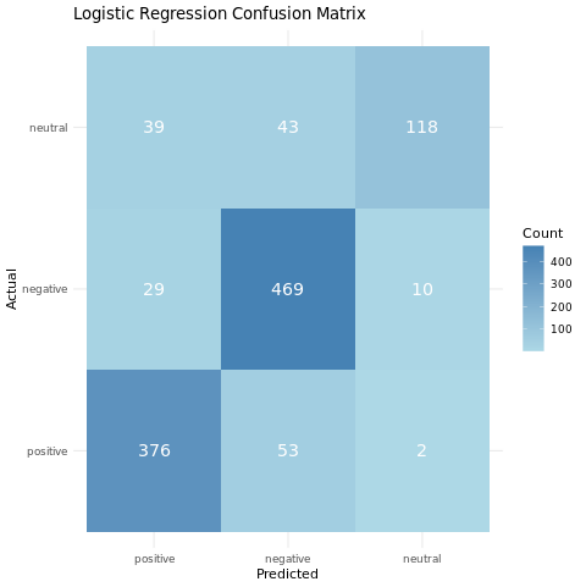


Figure 2: Logistic Regression confusion matrix for all classes.

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.

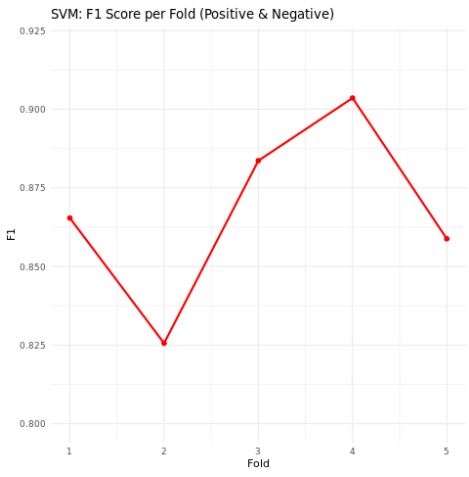


Figure 3: SVM: F1 score per fold for positive and negative classes.

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.

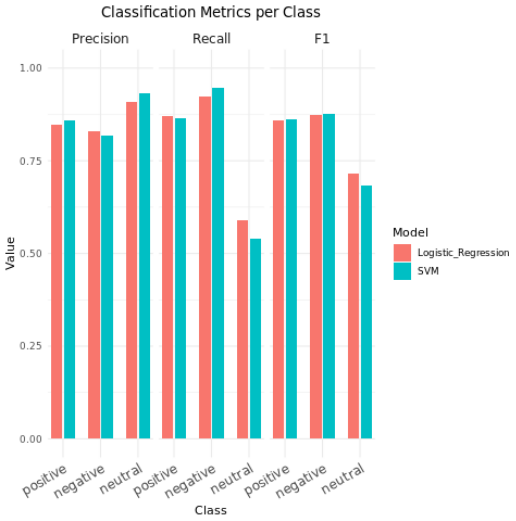


Figure 5: Overall precision, recall, and F1 scores for Logistic Regression and SVM.

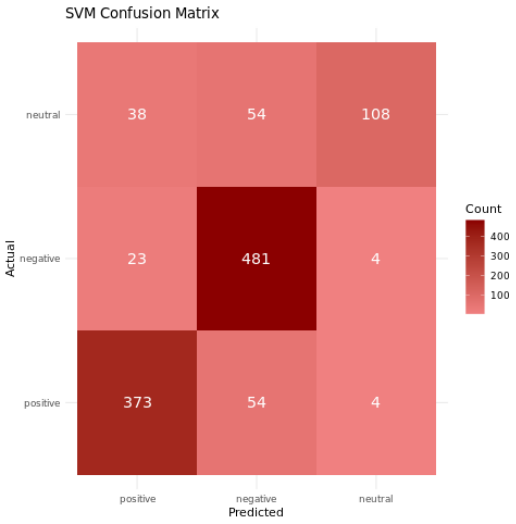


Figure 4: SVM confusion matrix for all classes.

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 demonstrates that our hybrid framework substantially improves over the purely rule-based approach. The baseline system reached a mean F1 score of 0.736 across folds, while Logistic Regression and SVM achieved 0.864 and 0.867, respectively. Paired t-tests confirmed that these improvements are statistically significant ($p < 0.001$). The Wilcoxon signed-rank test yielded $p = 0.0625$, slightly above the conventional threshold, likely due to the limited number of folds, but the performance trend remained consistent.

Most errors stem from the neutral class, where sentiment is often ambiguous or context-dependent, while positive and negative classes are reliably distinguished. This shows that leveraging lexicon-based features within machine learning models captures polarity effectively and generalizes well across folds. Overall, the results highlight the strength of hybrid models in combining the interpretability of rule-based systems with the adaptability of statistical learning. Future work should address the challenge of neutral sentiment and investigate richer contextual or semantic features.

5 Conclusion and Future Work

We presented a hybrid sentiment analysis framework for Macedonian, combining rule-based lexical features with Logistic Regression and Support Vector Machines. The hybrid models substantially outperformed the purely rule-based system, which achieved a mean F1 score of 73.6%. Both classifiers improved classification performance, particularly for polarized sentiment, while maintaining interpretability and robustness by relying exclusively on lexicon-derived features.

Our results demonstrate that integrating linguistic knowledge with statistical learning is effective for under-resourced languages like Macedonian, where annotated datasets are scarce. The rule-based component captures explicit, context-modified cues, while ML models generalize well across folds.

Future work includes:

- Incorporating syntactic and semantic embeddings to better capture context and subtle neutral sentiment.
- Experimenting with attention-based or transformer models for long-range dependencies.
- Expanding annotated datasets across social media, reviews, and user-generated content.
- Investigating domain adaptation to generalize across different text types.
- Integrating additional linguistic cues such as POS tags or dependency relations.
- Exploring multilingual transformers (e.g., mBERT, XLM-R) fine-tuned on Macedonian [2, 1].
- Using large language models to generate synthetic Macedonian training data [19, 14, 5].

This work provides a strong foundation for Macedonian sentiment analysis, highlighting the value of hybrid approaches and paving the way for richer linguistic feature integration and advanced modeling.

References

[1] Alexis Conneau et al. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors. Association for Computational Linguistics, Online, (July 2020), 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

[2] 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.

[3] Darja Fišer and Tomaž Erjavec. 2016. Analysis of sentiment labeling of slovene user-generated content. In Nasl. z nasl. zaslona. Znanstvena založba Filozofske fakultete, 22–25. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Fiser_Erjavec_Analysis-of-Sentiment-Labeling.pdf.

[4] 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].

[5] Nils Constantin Hellwig, Jakob Fehle, and Christian Wolff. 2024. Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings. *Expert Systems with Applications*, 261, (Oct. 2024), 125514. DOI: 10.1016/j.eswa.2024.125514.

[6] 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.

[7] 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.

[8] Danka Jokić, Ranka Stanković, and Branislava Šandrih Todorović. 2024. Abusive speech detection in Serbian using machine learning. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*. Ruslan Mitkov, Saad Ezzini, Tharindu Ranasinghe, Ignatius Ezeani, Nouran Khallaf, Cengiz Acarturk, Matthew Bradbury, Mo El-Haj, and Paul Rayson, editors. International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security, Lancaster, UK, (July 2024), 153–163. https://aclanthology.org/2024.nlpaics-1.18/.

[9] 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/.

[10] 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.

[11] Adela Ljajić, Ulfeta Marovac, and Aldina Avdic. 2017. Sentiment analysis of twitter for the serbian language. In (Mar. 2017).

[12] 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.

[13] Igor Mozetic, Miha Grcar, and Jasmina Smailovic. 2016. Multilingual twitter sentiment classification: the role of human annotators. In vol. 11. (Feb. 2016). DOI: 10.1371/journal.pone.0155036.

[14] Koena Ronny Mabokela, Mpho Primus, and Turgay Celik. 2025. Advancing sentiment analysis for low-resourced african languages using pre-trained language models. *PLOS ONE*, 20, 6, (June 2025), 1–37. DOI: 10.1371/journal.pone.0325102.

[15] 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/.

[16] 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.

[17] 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.

[18] 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.

[19] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: a reality check. https://arxiv.org/abs/2305.15005 arXiv: 2305.15005 [cs.CL].