Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores

Maria Cristina Hinojosa Lee[1 [0009−0006−6924−5761]], Johan Braet[1 [0000−0001−5726−9543]], and Johan Springael[1 [0000−0003−1768−5598]]

Department of Engineering Management, University of Antwerp, Prinsstraat 13, 2000 Antwerp, Belgium
mariacristina.hinojosalee@uantwerpen.be

Original publication: Hinojosa-Lee, M.C., Braet, J., Springael, J. Applied Sciences **14**(21), 9863 (28 October 2024). https://doi.org/10.3390/app14219863

Abstract. We compare micro, macro and weighted F1-scores for multilabel emotion classification under class imbalance. By using distilled lexicons built from GoEmotions and XED, we compute the three variants at eleven thresholds and cross-evaluate them on the opposite dataset, as well as on aggregated versions of both datasets. The three measures offer a picture of the model performance: Weighted F1 offers a most balanced assessment; micro F1 overestimates performance, while macro F1 penalizes majority classes. We recommend reporting the three performance measures.

Keywords: Emotion analysis · Multilabel classification · F1-score · performance metrics

1 Introduction

Emotion analysis is used in diverse fields such as recommender systems, health, marketing, e-learning, and more. Because several emotions can co-occur in a single message, the task is framed as multilabel classification [3,1]. F-measure variants (micro, macro, and weighted) are popular to evaluate these tasks and compare the results of different studies [1,5,3]. However, current literature often fails to justify the selection of the variant they report. When class frequencies diverge, the three variants can rank the same system differently, which jeopardizes fair comparison between models and downstream decisions. [3] The contribution of this study is to provide insights over how different F1-score variants compare and how they could improve the evaluation of multilabel emotion classifiers. This is done in the context of class imbalance.

2 Data & Methodology

For this study, two popular corpora were used.

- GoEmotions: which consists of 58,000 manually emotion annotated English comments from Reddit and originally used 27 fine-grained emotion labels. The GoEmotions authors grouped the emotions hierarchically into the six basic emotions of Ekman: anger, disgust, fear, joy, sadness, and surprise.[2]
- XED 30,000 English emotion-annotated movie subtitles labelled with 8 emotions.[6]

During the preprocessing we used the hierarchical grouping used by the GoEmotions authors to turn the 27 emotions into the 6 emotions system from Ekman. To be able to compare the results of using both datasets, we dropped anticipation and trust from the XED dataset and the neutral category in both datasets. Table 1 shows the number of messages used for the experiments in the published paper.[3]

Table 1. Emotion distribution of messages used from each dataset.

Dataset	Total	Anger	Disgust	Fear	Joy	Sadness	Surprise
GoEmotions	38,237	7,021	1,012	929	21,730	4,030	6,668
XED	13,682	3,828	2,317	2,439	2,832	2,464	2,442

An alternative aggregated form was also applied, which merges anger, disgust, fear, and sadness into a single negative class, for a more balanced 3-label task. We created distilled emotion lexicons from the datasets and their aggregated versions by obtaining unigram lexicons from each corpus at different thresholds; 11 lexicons per source were produced. The diagrams explaining the process can be consulted in the published article in Sections 3.2 and 3.3. Each lexicon was then applied to the alternative dataset and the three F1-variants were calculated at different thresholds.

3 Results

Tables 2 and 3 show the results with the higher scores across the different F-scores. The comprehensive results can be consulted in Section 4 of the published paper.[3]

Table 2. XED distilled lexicon applied to GoEmotions (left) and XED aggregated applied to GoEmotions aggregated (right).

	6 emotion label			3 emotion label (aggregated)		
Threshold	Macro F1	Micro F1	Weighted F1	Macro F1	Micro F1	Weighted F1
0.000	0.271	0.310	0.504	0.508	0.530	0.588
0.100	0.277	0.320	0.510	0.514	0.538	0.593
0.200	0.290	0.346	0.522	0.524	0.556	0.601
0.300	0.308	0.384	0.530	0.525	0.569	0.603
0.400	0.317	0.416	0.528	0.516	0.569	0.594
0.500	0.317	0.426	0.519	0.509	0.563	0.584

Table 3. GoEmotions distilled lexicon applied to XED (left) and GoEmotions aggregated applied to XED aggregated (right).

	6 emotion label			3 emotion label (aggregated)		
Threshold	Macro F1	Micro F1	Weighted F1	Macro F1	Micro F1	Weighted F1
0.000	0.328	0.330	0.337	0.478	0.520	0.626
0.100	0.307	0.339	0.319	0.481	0.526	0.629
0.200	0.289	0.341	0.305	0.484	0.549	0.632
0.300	0.257	0.311	0.272	0.474	0.560	0.624
0.400	0.210	0.279	0.226	0.404	0.505	0.548

4 Conclusions and discussion

Following results in literature, we considered a benchmark of 0.53 for the F1-score. We found that this was achieved in the non-aggregated datasets by the weighted F1-score [3]. As conclusion we recommend to report all three F1-variants to offer a more nuanced image of the classifier. Together they show whether improvements stem from majority-label recall or balanced gains. We also consider that it is necessary to optimize on the weighted F1-score, and then audit the macro F1-score. This keeps overall performance high while protecting rare emotions without affecting the majority class. We also consider that it is relevant to adopt aggregated emotion schemes where robustness outweighs granularity. The study shows that metric choice alone can shift multilabel-emotion conclusions. The micro F1-score, though popular, is insufficient to evaluate a multilabel classifier; a triad of micro, macro and weighted scores is necessary for honest assessment. For future work we plan on repeating the experiment using the CancerEMO dataset (clinical forum posts)[4].

References

1. Ameer, I., Sidorov, G., Gómez-Adorno, H., Nawab, R.M.A.: Multi-label emotion classification on code-mixed text: Data and methods. IEEE Access **10**, 8779–8789 (2022). https://doi.org/10.1109/ACCESS.2022.3143819
2. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: Goemotions: A dataset of fine-grained emotions. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020). pp. 4040–4054 (2020). https://doi.org/10.48550/arXiv.2005.00547, https://aclanthology.org/2020.acl-main.372
3. Hinojosa Lee, M.C., Braet, J., Springael, J.: Performance metrics for multilabel emotion classification: Comparing micro, macro and weighted f1-scores. Applied Sciences **14**(21), 9863 (2024). https://doi.org/10.3390/app14219863
4. Jo, Y., Bang, S., Manzoor, E., Hovy, E., Reed, C.: Detecting attackable sentences in arguments. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1–23. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.1, https://aclanthology.org/2020.emnlp-main.1/
5. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4), 427–437 (2009)
6. Öhman, E., Pàmies, M., Kajava, K., Tiedemann, J.: XED: A multilingual dataset for sentiment analysis and emotion detection. arXiv (2020). https://doi.org/10.48550/arXiv.2011.01612

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.